Multiscale Density Estimation

R. M. Willett, *Student Member, IEEE*, and R. D. Nowak, *Member, IEEE*

July 4, 2003

Abstract

The nonparametric density estimation method proposed in this paper is computationally fast, capable of detecting density discontinuities and singularities at a very high resolution, spatially adaptive, and offers near minimax convergence rates for broad classes of densities including Besov spaces. At the heart of this new method lie multiscale signal decompositions based on piecewise-polynomial functions and penalized likelihood estimation. Upper bounds on the estimation error are derived using an information-theoretic risk bound based on squared Hellinger loss. The method and theory share many of the desirable features associated with wavelet-based density estimators, but also offers several advantages including guaranteed non-negativity, bounds on the L_1 error, small-sample quantification of the estimation errors, and additional flexibility and adaptability. In particular, the method proposed here can adapt the degrees as well as the locations of the polynomial pieces. For a certain class of densities, the error of the variable degree estimator converges at nearly the parametric rate. Experimental results demonstrate the advantages of the new approach compared to traditional density estimators and wavelet-based estimators.

Keywords: nonparametric estimation

I. Density Estimation

Accurate and efficient density estimation is often a fundamental first step in many areas, including source coding, data compression, statistical learning, and signal processing. The approach presented in this paper involves using penalized likelihood estimation on recursive dyadic partitions in order to produce near-optimal, piecewise polynomial density estimates. This results in a multiscale method that provides spatial adaptivity similar to wavelet-based techniques [1]. Like wavelet-based density estimators, the new method admits fast estimation algorithms and estimators that exhibit near minimax optimal rates of convergence in many function spaces. The proposed method has several additional advantages: estimates are guaranteed to be positive,

*Corresponding author: R. Willett. The authors are with the Department of Electrical and Computer Engineering, Rice University MS 366, Houston, TX 77251-1892 USA (e-mail: willett@ece.rice.edu, phone: 713 348 3230, fax: 713 348 6196). R. Willett was partially supported by the National Science Foundation Graduate Student Fellowship. R. Nowak was partially supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0349, the Office of Naval Research, grant no. N00014-00-1-0390.

theoretical bounds provide an indication of performance even for small sample sizes, and the method can be extended to free-degree piecewise polynomial estimation. Free-degree estimation exhibits rates of convergence within a logarithmic factor of the parametric rate for some classes of densities. We elaborate on these points below.

A. Relation to Classical and Wavelet Density Estimators

Classical nonparametric density estimation techniques, *e.g.* kernel or histogram methods, have been thoroughly explored in the density estimation literature [2–7]. Most of the theoretical analysis associated with these methods pertains to linear estimators, which are known to be quite sub-optimal (in the sense of rates of convergence) for many classes of densities, e.g., Besov spaces [8]. Because linear estimators do not adapt to spatial changes in the structure of the data, their density estimates are in practice frequently oversmoothed where the density is changing rapidly or undersmoothed where the density is changing more slowly. Such estimators do not preserve singularities or sharp changes in the underlying density. Spatially adaptive kernel methods have been proposed to overcome such limitations [9], but these methods have been outshined by recently devised wavelet-based density estimation techniques [8].

Wavelet-based techniques implicitly overcome this lack of spatial adaptivity because wavelets are well localized in both time and frequency and hence can provide good local estimates of the density. The estimation scheme presented by Donoho, *et al* [8], is representative of many wavelet-based density estimators and summarized here in order to highlight its similarities to and differences from the method proposed in this paper.

Any piecewise smooth density, $f(\cdot)$, such as one in a Besov space, can be represented in terms of scaling and wavelet coefficients:

$$f(t) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)$$

In an orthogonal system, each wavelet coefficient is the inner product of the density and the wavelet function at a particular scale and shift, and hence the wavelet coefficients are the expectation of the wavelet function. If $\boldsymbol{Y}$ is a random variable with density f, then we can write the

coefficients as:

$$d_{j,k} = \int f(t)\psi_{j,k}(t)dt = \mathbb{E}\left[\psi_{j,k}(\boldsymbol{Y})\right].$$

Thus a Monte Carlo estimate of each wavelet coefficient can be computed as

$$\widehat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(y_i).$$

Assuming that there are enough observations falling in the support of $\psi_{j,k}$, the central limit theorem can be invoked and $\widehat{d}_{j,k}$ can be assumed to be approximately Gaussian distributed with mean $d_{j,k}$ and some variance. In wavelet-based density estimation, these coefficient estimates are improved using a hard or soft thresholding scheme based on the Gaussianity of the coefficients, and then the thresholded coefficients are used to synthesize the final density estimate. To guarantee that (on average) a sufficient number of samples fall within the support of each wavelet basis function to justify the Gaussian approximation, wavelet-based density estimates are restricted to scales less than or equal to $2^{-\log(n/\log n)}$. This effectively erects a resolution limit for wavelet-based density estimation that is insensitive to the particular distribution of the data.

B. Improved Multiscale Density Estimation

Wavelet-based techniques are advantageous for both their near minimax convergence rates and the computational simplicity of filter-bank implementations. Little *a priori* knowledge of the density smoothness is required as long as the wavelet function $\psi(\cdot)$ is smooth enough. Similarly, the method introduced in this paper admits a computational efficient analysis, is spatially adaptive, and exhibits the same convergence rates as wavelet-based techniques. The new method has several additional benefits. First, the estimator always results in *bona fide* (*i.e.* non-negative) density estimates. Next, the error bounds developed in this paper provide an indication of small-sample performance as well as asymptotic rates of convergence. The risk analysis here produces a bound on the expected error between the true histogram (the quantized density with bin width inversely proportional to the sample size) and our density estimator. Asymptotically, the bound produces the rates of convergence known for wavelet-based density estimators. These bounds do not rely on the central limit theorem, and thus place no fundamental limit on the resolution

of the estimate; the resolution limit of the new estimator is dictated by the specific distribution of the data, not just the total number of observations. In particular, it may recover finer scale structure in regions of high density than traditional wavelet-based estimators are capable. Finally, the proposed method is easily extendable to free-degree piecewise polynomial estimation. It has been demonstrated for certain classes of densities that free-knot, free-degree estimation can achieve exponential rates of decay in approximation error [10]. This fact is exploited here to demonstrate near parametric rates of convergence for our density estimator.

Our approach, based on complexity-regularization, is similar in spirit to the seminal work of Barron and Cover [11]. In the proposed method, either fixed- or free-degree polynomials are fitted to a recursive dyadic partition (RDP) of the density's support (herein assumed to be the interval $[0, 1]$). Section II describes the basic formulation of our method. The RDP leads to a model selection problem that can be solved by binary tree pruning process. Appropriate pruning of this tree results in a penalized likelihood estimate of the density, and computationally efficient algorithms for computing either the fixed- or free-degree density estimates are proposed in Section III. The main convergence results are summarized in IV. Upper bounds on the estimation error (expected squared Hellinger distance) are established using several recent information-theoretic results, most notably the Li-Barron bound [12, 13] and Nowak and Kolaczyk's generalization of this bound [14]. Experimental results demonstrate the advantages of the new approach compared to traditional wavelet-based estimators in Section VI. Section VII discusses some of the implications of our results and directions for future work.

II. Problem Formulation

The basic set-up considered in this paper is as follows. A series of n independent and identically distributed observations, y_j, $j = 1, \ldots, n$ of a random variable, $\boldsymbol{Y}$, can be represented as a histogram with arbitrarily small bin counts. In this setup, the collection of N bin counts, $\{x_i\}_{i=1}^N$, can be modeled using a multinomial distribution, and density estimation can be reframed as a problem of multinomial parameter estimation. The multiscale method presented here finds the optimal piecewise polynomial fit to the bin counts using penalized likelihood estimation. The resulting multinomial bin probabilities form the equivalent of the probability mass function generating the discretized random variable observations.

The proposed method calculates density estimates by first determining the ideal partition of

the range of observations (assumed to be $[0, 1]$) and then using maximum likelihood estimation to fit a polynomial to each interval in the optimal partition. The space of possible partitions is a nested hierarchy of partitions defined through a recursive dyadic partition (RDP) of the unit interval, and the optimal partition is selected by optimally pruning a binary tree representation of the complete RDP of the data range. The effect of polynomial estimation on dyadic intervals is essentially an estimator with the same approximation capabilities as a wavelet-based estimator; this is established using approximation theoretic bounds in [15] for fixed-degree polynomials and [10] for free-degree polynomials. Thus there is neither an advantage nor a disadvantage (in an approximation-theoretic sense) in using a piecewise polynomial basis instead of a wavelet basis. It is the method (described below) by which the basis coefficients are computed that makes the proposed technique better suited for density estimation. Specifically, wavelet-based techniques typically only analyze the data up to some fixed resolution in order to ensure that there are approximately $\log n$ observations within the support of each wavelet at the finest resolution. This places a fundamental limit on the resolution achievable by such methods. The proposed method, in contrast, does not rely on any such assumption and hence could result in a final estimate with a higher resolution dictated by the data.

The goal of density estimation is to estimate the true density f using several observations drawn from f. Assume that either by choice or perhaps the limitations of measuring instruments, each of the observations has been discretized so that it can only be resolved up to $1/N$, where $N = 2^{\lceil \log_2 n \rceil}$. (Our approach is nonparametric, since the resolution increases like 1/n as the number of samples increases.) It is assumed that the effect of the quantization is to yield a vector of count measurements $\boldsymbol{x} \equiv \{x_i\}_{i=0}^{N-1}$, where each x_i is simply the number of events in the interval $I_i \equiv [i/N, (i+1)/N)$. The density f can be similarly discretized by defining $\boldsymbol{f} \equiv \{f_i\}_{i=0}^{N-1}$, where $f_i \equiv \int_{I_i} f(t)dt$. Note that $\boldsymbol{f}$ is the probability mass function (pmf) of the discretized random variables. The likelihood of observing $\boldsymbol{x}$, given the probabilities $\boldsymbol{f}$, is multinomial and denoted by $p(\boldsymbol{x}|\boldsymbol{f})$.

As mentioned above, the piecewise polynomial multiscale analysis presented here is performed on recursive dyadic partitions (RDPs) of the unit interval. An RDP of the count data is obtained by associating a count statistic $x_{I_{k,j}} \equiv \sum_{i:(\frac{i}{N}) \in I_{k,j}} x_k$ with each dyadic interval $I_{k,j} \equiv [k/2^j, (k+1)/2^j)$, $j = 0, \ldots, J-1$, $k = 0, \ldots, 2^j - 1$, and $J = \log_2(N)$. The set

of all dyadic intervals $\{I_{k,j}\}$ corresponds to a *complete* recursive dyadic partition (C-RDP) of $[0, 1]$. There also exists a C-RDP of the pmf $\boldsymbol{f}$ which is defined analogously. This RDP is called complete because all terminal nodes in the partition are intervals of width $1/N$ at the finest scale.

For a piecewise polynomial multiresolution analysis we consider an incomplete RDP that would contain larger terminal intervals which could correspond to intervals of homogeneous or smoothly varying densities. Such a partition can be obtained by merging (*i.e.* pruning) a C-RDP to form an optimal RDP $\mathcal{P}$ and using the polynomial coefficients $\boldsymbol{\theta}$ on the terminal intervals of $\mathcal{P}$. Thus the density estimate is completely described by $\mathcal{P}$ and $\boldsymbol{\theta}$; i.e. $\boldsymbol{f} = \boldsymbol{f}(\mathcal{P}, \boldsymbol{\theta})$. For a given RDP $\mathcal{P}$, the likelihood $p(\boldsymbol{x}|\boldsymbol{f})$ may be factorized as

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{f}(\mathcal{P}, \boldsymbol{\theta})) = \\ \prod_{I \in NT(\mathcal{P})} p(\{x_{lch(I)}\}|x_I, \rho_I) \\ \times \prod_{I \in T(\mathcal{P})} p(\{x_i\}_{i/N \in I}|x_I, \theta_I), \end{aligned} \tag{1}$$

where $NT(\mathcal{P})$ is the set of all non-terminal intervals in $\mathcal{P}$, and $T(\mathcal{P})$ is the set of all terminal intervals in $\mathcal{P}$. The binomial splitting probability ρ_I is simply the ratio of the probability mass in the left child of node I to the probability mass in node I, *i.e.* $\rho_I \equiv f_{lch(I)}/f_I$. The probability term $p(\{x_{lch(I)}\}|x_I, \rho_I)$ is the probability of observing $x_{lch(I)}$ counts in the left half of the interval I given that a total of x_I counts were observed in the entire interval I and given the probability mass ratio ρ_I. The model for the density of the process on each disjoint terminal interval $I \in T(\mathcal{P})$ is constrained such that $f_I = T \cdot \theta_I$, where T is a known Vandermonde matrix transforming the vector θ_I, the polynomial coefficients to be estimated, to the polynomial signal, f_I. The polynomial coefficients θ_I are further subject to the constraints that $\int_I f = 1$ (since ρ_I ensures correct normalization) and $f_I \geq 0$. These constraint results in the desired piecewise polynomial density estimate $\boldsymbol{f}(\mathcal{P}, \boldsymbol{\theta})$, where the breakpoints between the polynomial pieces are determined by the partition $\mathcal{P}$ and the polynomial coefficients on each interval in the partition are contained in the coefficient vector $\boldsymbol{\theta}$. The terminal node likelihood factors $p(\{x_i\}_{i/N \in I}|x_I, \theta_I)$ are the multinomial likelihoods of the observations in I given a polynomial coefficients $\boldsymbol{\theta}_I$ of the density f_I.

III. Multiscale Density Estimation

Piecewise polynomial approximation provides good approximations to piecewise smooth functions, and in fact exhibits the same approximate error decay as wavelets [15]. As discussed in the introduction, however, wavelet analysis is difficult to analyze for the multinomial data encountered in density estimation and can lead to some undesirable artifacts in practice (*e.g.* negativity). The alternative method proposed in this section can be computed as efficiently as a wavelet analysis, and yields demonstrably nearly optimal density estimates.

The proposed method operates by determining the optimal partition of the interval $[0, 1]$ based on the observations and then finding the optimal polynomial fit to the density on each interval in the partition. This provides for a very simple framework for penalized likelihood estimation, wherein the penalization is based on the complexity of the underlying partition. The complexity of a given partition is proportional to the total number of intervals, m. The goal here is to minimize the penalized likelihood function

$$L(\boldsymbol{f}) \;\equiv\; -\log p(\boldsymbol{x} \mid \boldsymbol{f}) \;+\; \text{pen}(\boldsymbol{f}), \tag{2}$$

where $p(\boldsymbol{x} \mid \boldsymbol{f})$ denotes a likelihood of the form (1) and $\text{pen}(\boldsymbol{f}(m, r))$ is the penalty associated with the estimate $\boldsymbol{f}$. We penalize the piecewise polynomial estimates according to a codelength required to uniquely describe each such model (*i.e.*, codes which satisfy the Kraft inequality). These codelengths will lead to near-minimax optimal estimators, as discussed in the next section. The codelengths are proportional to the size (number of intervals) in the partition associated with each model, and thus penalization leads to estimates that favor smaller partitions. In particular, for fixed-degree polynomial estimation, if $\boldsymbol{f}$ consists of order-r polynomials on m intervals, then

$$\text{pen}(\boldsymbol{f}(m, r)) \equiv (2m - 1) \log_e 2 + \frac{mr}{2} \log_e n. \tag{3}$$

For free-degree polynomial estimation, if $\boldsymbol{f}$ is defined by a total of d coefficients on m intervals, with the i^{th} interval having length ℓ_i, $i = 1, \ldots, m$, then

$$\text{pen}(\boldsymbol{f}(m, d, \{\ell_i\})) \equiv (2m - 1) \log_e 2 + \frac{d}{2} \log_e n + \sum_{i=1}^{m} \log_e \ell_i. \tag{4}$$

The penalties can be interpreted as a negative log-prior on the space of estimators. The penalties are designed to give good guaranteed performance by balancing between fidelity to the data (likelihood) and the estimate's complexity (penalty), which effectively controls the bias-variance trade-off.

The solution of

$$\begin{aligned} (\widehat{\mathcal{P}}, \widehat{\boldsymbol{\theta}}) &\equiv \arg\min_{\mathcal{P},\boldsymbol{\theta}} L(\boldsymbol{f}(\mathcal{P}, \boldsymbol{\theta})) \\ \widehat{\boldsymbol{f}} &\equiv \boldsymbol{f}(\widehat{\mathcal{P}}, \widehat{\boldsymbol{\theta}}) \end{aligned} \tag{5}$$

is called a penalized likelihood estimator (PLE). Section IV demonstrates that this form of penalization results in near minimax optimal density estimates. Minimizing (2) involves adaptively pruning the complete RDP based on the data. This pruning can be performed optimally and very efficiently. The pruning process is akin to a "keep or kill" wavelet thresholding rule. The PLE provides higher resolution and detail in areas of the signal where there are dominant edges or singularities with higher count levels (higher SNR). The partition underlying the PLE is pruned to a coarser scale (lower resolution) in areas with lower count levels (low SNR) and where the data suggest that the density is fairly smooth.

Observe that the structure of the penalized likelihood criterion stated in (2) and the likelihood factorization described in Section II allow an optimal density estimate to be computed quickly. The likelihood factorization allows the likelihood of the entire signal to be represented in a tree structure in which both likelihoods and parameter vectors are passed from the children to the parents [14]. Using this, it is possible to optimally prune an RDP of the data using a fast algorithm reminiscent of dynamic programming and the CART algorithm [16, 17].

The goal is to estimate the density $\boldsymbol{f}$ according to (5). In order to perform the fixed-degree estimation, the algorithm iterates from bottom to top through each level $j = \log_2 N, \ldots, 0$ of the C-RDP the observations. At level j, a binary hypothesis test is conducted for each of the 2^j nodes at that level. The hypotheses for each node are as follows:

- $\boldsymbol{H}_0$: Order r polynomially varying density segment (terminal node)
- $\boldsymbol{H}_1$: Keep optimal estimates of both children (non-terminal node)

The free-degree polynomial algorithm is similar, except a $(N/2^j + 1)$-ary hypothesis test is

Initialize:	$j = J - 1$
Loop:	for each node $I_{k,j}$ at level j
Calculate:	$L(\boldsymbol{\theta}_{\boldsymbol{H}_0}; I_{k,j})$
	$L(\boldsymbol{\theta}_{H_1}; I_{k,j}) = \sum_{I' \in ch(I_{k,j})} L_{min}(I')$
Save:	$L_{min}(I_{k,j}) = \min_{i \in \{0,1\}} L(\boldsymbol{\theta}_{\boldsymbol{H}_i}; I_{k,j})$
	$\boldsymbol{\theta}_{min}(I_{k,j}) = \arg \min_{i \in \{0,1\}} L(\boldsymbol{\theta}_{\boldsymbol{H}_i}; I_{k,j})$
Coarsen:	Scale $j = j - 1$
Goto Loop:	if $j \geq 0$
Estimate:	$\widehat{\boldsymbol{f}} = \boldsymbol{f}(\boldsymbol{\theta}_{min}(I_{0,0}))$

TABLE I
FIXED-DEGREE ALGORITHM PSEUDOCODE

conducted for each of the 2^j nodes at level j:

- $\boldsymbol{H}_d$: Order d ($d = 1, 2, \ldots, N/2^j$) polynomially varying density segment (terminal node)
- $\boldsymbol{H}_{N/2^j+1}$: Keep optimal estimates of both children (non-terminal node)

In the fixed-degree case, when the maximum polynomial degree is 0 ($r = 1$), the algorithm coincides with Haar analysis with a hereditary constraint [14]. The algorithm begins one scale above the leaf nodes in the binary tree and traverses upwards, performing a tree-pruning operation at each stage. For each node (i.e., dyadic interval) at a particular scale, the maximum likelihood parameter vector is optimally determined for each hypothesis using the Vandermonde matrix described in Section II and the penalized log likelihoods for each hypothesis are calculated. In particular, the penalized log likelihood for the split is computed using the optimal penalized log likelihoods computed at the previous, finer scale for both of the two children. The fixed-degree algorithm pseudocode is in Table I. In the pseudocode, $L(\boldsymbol{\theta}_{\boldsymbol{H}_0}; I_{k,j})$ denotes the penalized log likelihood term for segment $I_{k,j}$ under hypothesis $\boldsymbol{H}_0$. The free-degree algorithm pseudocode is in Table II. Here, $L(\boldsymbol{\theta}_{\boldsymbol{H}_d}; I_{k,j})$ denotes the penalized log likelihood term for segment $I_{k,j}$ under hypothesis $\boldsymbol{H}_d$.

IV. UPPER BOUNDS ON ESTIMATION ERROR

While the proposed polynomial pruning algorithm described above yields the optimal PLE, the analysis up to this point does not quantify the effectiveness of this method. The analysis in [14] established statistical risk bounds associated with estimating a density, $\boldsymbol{f}$, with a piecewise

Initialize:	$j = J - 1$
Loop:	for each node $I_{k,j}$ at level j
Calculate:	$L(\boldsymbol{\theta}_{\boldsymbol{H}_d}; I_{k,j})$ for $d = 1 \ldots N/2^j$
	$L(\boldsymbol{\theta}_{H_{N/2^j+1}}; I_{k,j}) = \sum_{I' \in ch(I_{k,j})} L_{min}(I')$
Save:	$L_{min}(I_{k,j}) = \min_{1 \le d \le N/2^j+1} L(\boldsymbol{\theta}_{\boldsymbol{H}_d}; I_{k,j})$
	$\boldsymbol{\theta}_{min}(I_{k,j}) = \arg \min_{1 \le d \le N/2^j+1} L(\boldsymbol{\theta}_{\boldsymbol{H}_d}; I_{k,j} I_{k,j})$
Coarsen:	Scale $j = j - 1$
Goto Loop:	if $j \ge 0$
Estimate:	$\widehat{\boldsymbol{f}} = \boldsymbol{f}(\boldsymbol{\theta}_{min}(I_{0,0}))$

TABLE II
FREE-DEGREE ALGORITHM PSEUDOCODE

constant estimator $\widehat{\boldsymbol{f}}$. In this section, their analysis is generalized for the case of piecewise polynomial estimation, as described above, and the resulting bound is used to establish the near-optimality of the proposed estimation method.

In this paper risk is defined to be proportional to the expected squared Hellinger distance between the true and estimated densities as in [11, 13]; that is,

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \equiv \frac{1}{N} \mathbb{E}_{\boldsymbol{f}} \left[L(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \right] \tag{6}$$

where $\mathbb{E}_{\boldsymbol{f}}$ denotes expectation with respect to the distribution $\boldsymbol{f}$ and

$$L(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \equiv H^2(p_{\widehat{\boldsymbol{f}}}, p_{\boldsymbol{f}}) = \int \left[\sqrt{p(\boldsymbol{x}|\widehat{\boldsymbol{f}})} - \sqrt{p(\boldsymbol{x}|\boldsymbol{f})} \right]^2 \nu(\boldsymbol{x}) \tag{7}$$

is the loss and ν is the dominating measure. The squared Hellinger distance is an appropriate error metric here for several reasons. First, it is a general non-parametric measure appropriate for any density. In addition, the Hellinger distance provides an upper and lower bound on the L_1 error because of the relation $H^2(p_1, p_2) \le \int |p_1 - p_2| \le 2H(p_1, p_2)$ for all distributions p_1 and p_2 [2]. The L_1 metric is particularly useful for density estimation because of Scheffé's identity [2], which states that if $\mathcal{B}$ is the class of all Borel sets of $[0, 1]$, then

$$\sup_{B \in \mathcal{B}} \left| \int_B \widehat{\boldsymbol{f}} - \int_B \boldsymbol{f} \right| = \frac{1}{2} \int \left| \widehat{\boldsymbol{f}} - \boldsymbol{f} \right|.$$

Scheffe's identity shows that a bound on the L_1 error provides a bound on difference between the true probability measure and the density estimator's measure on every event of interest. Similar bounds cannot be derived from bounds on the L_2 norm or the KL divergence.

Finally, using the squared Hellinger distance allows us to take advantage of a key information-theoretic inequality derived by Li and Barron [12, 13] to prove the following main theorem:

Theorem 1 *Let $\boldsymbol{f}$ be a vector consisting of n samples of a density, $f(\cdot)$, which is a member of the Besov space $B^{\alpha}_{\tau}(L_{\tau}([0,1]))$ where $0 < \alpha \leq r$, $1/\tau = r + 1/2$ and $L_2([0,1])$ is the approximation space. Further assume $0 < C_{\ell} \leq f(\cdot) C_u$. Let $\widehat{\boldsymbol{f}}$ be the fixed-degree order r penalized likelihood estimator satisfying (5) using the penalty in (3). Then*

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \leq C \left(\frac{\log^2 n}{n} \right)^{\frac{2\alpha}{2\alpha+1}} \tag{8}$$

for some constant C.

The penalization structure employed here minimizes the upper bound on the risk. Furthermore, this upper bound is within a logarithmic factor of the lower bound on the minimax risk, demonstrating the near-optimality of the proposed method. In particular, the minimax risk for all estimators is lower bounded by $n^{-\frac{2\alpha}{2\alpha+1}}$ [8]. The upper bound derived here is also within a logarithmic factor of the lower bound on the L_1 minimax error, $n^{-\frac{\alpha}{2\alpha+1}}$ [2], as stated in the following corollary:

Corollary 1 *Let $\boldsymbol{f}$ and $\widehat{\boldsymbol{f}}$ be defined as in Theorem 1. Then*

$$\mathbb{E}\left[\| \widehat{\boldsymbol{f}} - \boldsymbol{f} \|_{\ell_1} \right] \leq C \left(\frac{\log^2 n}{n} \right)^{\frac{\alpha}{2\alpha+1}}$$

for some constant C.

These results demonstrate the near-optimality of the penalization structure $\text{pen}(\boldsymbol{f}) = (2m - 1) \log_e 2 + \frac{mr}{2} \log_e N$ for fixed-degree estimation when the system resolution is such that there is an average of one observation per observation interval (*i.e.* $n \approx N$). A similar bound is possible for free-degree estimation:

Corollary 2 *Let $\boldsymbol{f}$ be a vector consisting of n samples of a density of the form $f(t) = \beta_0 t^{\beta_1} + \beta_2$ for some constants β_i, $i = 1, 2, 3$ such that $0 < C_\ell \leq f(\cdot) \leq C_u$. Let $\widehat{\boldsymbol{f}}$ be the free-degree penalized likelihood estimator satisfying (5) using the penalty in (4). Then*

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \leq C \, \frac{\log^4 n}{n} \tag{9}$$

for some constant C.

Note that this is within a logarithmic factor of the parametric rate of convergence, $1/n$. This result demonstrates the near-optimality of the penalization structure $\text{pen}(\boldsymbol{f}'; m, d, \{\ell_i\}) \equiv \frac{d}{2} \log_e N + (2m-1) \log_e 2 + \sum_{i=1}^{m} \log_e \ell_i$. Furthermore, it is within a logarithmic factor of the fixed-degree rate taken as $r \to \infty$. Also observe that the penalty structure for free-degree estimation is very similar to that of fixed-degree estimation. The only difference is that the lengths of the intervals must be penalized for free-degree estimation; this choice will be explained in detail in the proof of the theorem. Details on the risk bounds for both fixed and free-degree estimation are in Appendix A.

V. Computational Complexity

The previous section established the near-optimality of the proposed method using information theoretic arguments to bound the statistical risk. This section demonstrates that the proposed method is as computationally efficient as traditional wavelet analysis in addition to having the theoretical advantages discussed in the previous sections.

The proposed method's overall computational complexity depends on the complexity of the polynomial fitting operation on each interval in the recursive dyadic partition. Consider a terminal node in (1): $p(\{x_i\}_{\frac{i}{N} \in I} | x_I, \theta_I)$. There is no closed-form solution to the MLE of the polynomial coefficients with respect to the multinomial likelihood; however, they can be computed numerically, for example using the algorithm for the closely-related Poisson case by Unser and Eden [18]. The following lemma ensures that the polynomial coefficients can be computed quickly:

Lemma 1 *Let a mulitnomial likelihood be written $p(\boldsymbol{x}|\boldsymbol{f}, \sum_i x_i)$ and let $\boldsymbol{f}$ obey a polynomial model of the form $\boldsymbol{f} = T\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector containing the polynomial coefficients and T is a*

known Vandermonde matrix relating the polynomial coefficients to the multinomial probabilities. The multinomial likelihood can then be written as $p_x(\boldsymbol{\theta}) \equiv p(\boldsymbol{x}|T\boldsymbol{\theta}, \sum_i x_i)$*. Let* Θ *denote the set of all coefficient vectors* $\boldsymbol{\theta}$ *which result in a* bona fide *density. Then* $p_x(\boldsymbol{\theta})$ *is a convex function on* Θ*, which is a convex set.*

Lemma 1 is proved in Appendix C. Because the likelihood in the factorization is twice continuously differentiable and convex in the polynomial coefficients and the set of all admissible polynomial coefficients is convex, a numerical optimization technique such as Newton's method or gradient descent can find the optimal parameter values with quadratic or linear convergence rates, respectively.

This lemma is a key component in the computational complexity analysis of the proposed method. The theorem below is also proved in Appendix C. An "approximate" estimate refers to the result of the proposed method where the polynomial coefficients are computed using least squares, and an "exact" estimate refers to the result of the proposed method where the polynomial coefficients are computed by minimizing the multinomial log likelihood with a numerical minimization method.

Theorem 2 *A fixed-degree piecewise polynomial exact PLE can be computed in* $O(n)$ *calls to a convex minimization routine and* $O(n)$ *comparisons of the resulting (penalized) likelihood values, where* n *is the number of observation intervals. An approximate estimate can be computed with* $O(n \log n)$ *operations. A free-degree piecewise polynomial exact PLE can be computed in* $O(n \log n)$ *calls to a convex minimization routine and* $O(n)$ *comparisons of the resulting (penalized) likelihood values, and an approximate estimate can be computed in* $O(n^3)$ *operations.*

Note that the theorem does not imply that approximate estimates are more computationally demanding than exact estimates. The computational complexities for approximate estimates are based on the assumption that a least squares fit of m data points to d polynomial coefficients can be completed in $O(md)$ operations, whereas the order of operations for the exact estimate can vary with the choice of optimization method.

VI. Applications and Simulations

The analysis of the previous section demonstrates the strong theoretical arguments for using optimal tree pruning for multiscale density estimation. These findings are supported by nu-

merical experiments which consisted of comparing the polynomial density estimation technique presented here with a wavelet-based method. Seven test densities were generated using the well known test functions 'HeaviSine', 'Blocks', and 'Bumps' [19] and four of the Marron Gaussian mixtures: normal (#1), kurtotic unimodal (#4), separated bimodal (#7), and claw (#10) [20]. Probability mass functions (pmfs) of length 1024 were constructed from these test signals by shifting them to be strictly positive and normalizing them to one. This mimics a density estimation problem in which the measurement system has an accuracy of 10 bits. To simulate a set of observations from each density, 1024 iid samples from each pmf were generated by a random number generator. Notice that the total number of samples is approximately the same as the dimension of the pmfs, simulating the ideal situation in which the data are not binned.

The densities were estimated with the fixed- and free-degree PLE methods described in this paper and the wavelet hard- and soft-thresholding methods described in [8]. Ten estimations were performed using the wavelet thresholding method (with D8 wavelets and either hard or soft threshold levels) and the proposed PLE (with either piecewise cubic or free-degree polynomial fits). Cubic fits were chosen to provide the best comparison to D8 wavelets, which have three vanishing moments. The hard and soft wavelet threshold levels were chosen to minimize the average ℓ_1 estimation error over the seven distributions. In the wavelet thresholding case, these "clairvoyant" thresholds could not be obtained in practice, but here they provide an empirical lower bound on the achievable MSE performance for any practical hard-thresholding scheme. The MSE of these estimates are displayed in Figure 1. Clearly, even without the benefit of setting the penalization factor clairvoyantly or data adaptively, the multiscale PLE yields comparable errors to wavelet-based techniques for both smooth and spiky densities. When the proposed PLE penalties are similarly weighted to minimize the average ℓ_1 estimation error over the seven distributions, the polynomial techniques often outperform wavelet-based techniques, as shown in Figure 2. Notably, unlike wavelet-based techniques, the polynomial technique is guaranteed to result in a non-negative density estimate. Density estimates can be viewed in Figures 3 and 4.

VII. Conclusions and Ongoing Work

This paper presents a new method for density estimation based on fixed- and free-degree piecewise polynomial approximations of functions at multiple scales. Like wavelet-based estimators, this method can outperform some linear estimators because of its ability to isolate

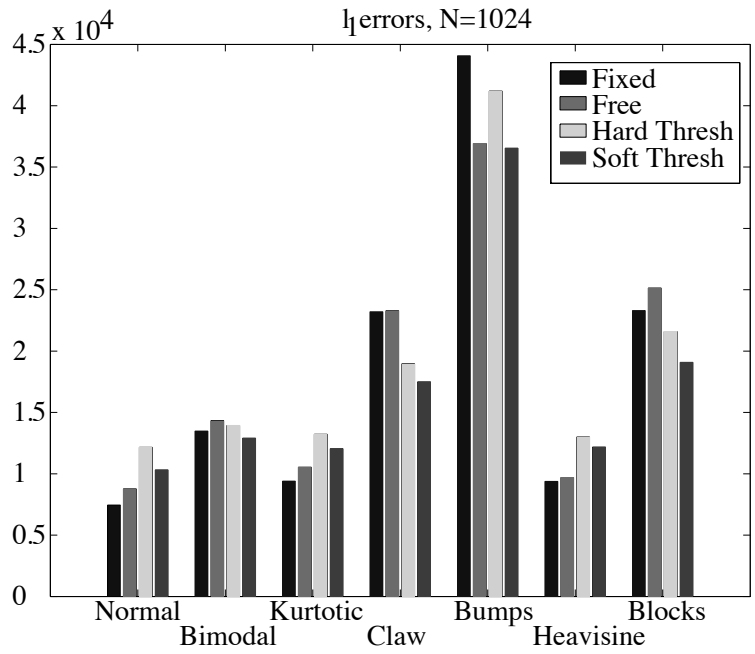

Fig. 1. Density estimation using theoretical penalties.

discontinuities or singularities in the density function. It can also produce higher-resolution estimates with small sample sizes than wavelet-based estimators. Experimental results support this claim, and risk analysis demonstrates the minimax near-optimality of the proposed method.

The method presented and analyzed in this paper demonstrates the power of multiscale analysis in a more general framework than that of traditional wavelet-based methods. Conventional wavelets are effective primarily because of two key features: (1) adaptive recursive partitioning of the data space to allow analysis at multiple resolutions, and (2) wavelet basis functions that are blind to polynomials according to their numbers of vanishing moments. The alternative method presented here is designed to exhibit these same properties without retaining other wavelet properties which are significantly more difficult to analyze in the case of non-Gaussian data. Furthermore, this new method achieves the same minimax near-optimality as traditional wavelet-based estimators for Gaussian observations [1]. Finally, these penalized likelihood estimators can be computed with the same computational complexity as traditional wavelet-based estimators. This paper also presents a computationally efficient method for free-degree polynomial density estimation. This new method allows the data to adaptively determine the smoothness of the underlying basis function instead of forcing the user to select a polynomial order

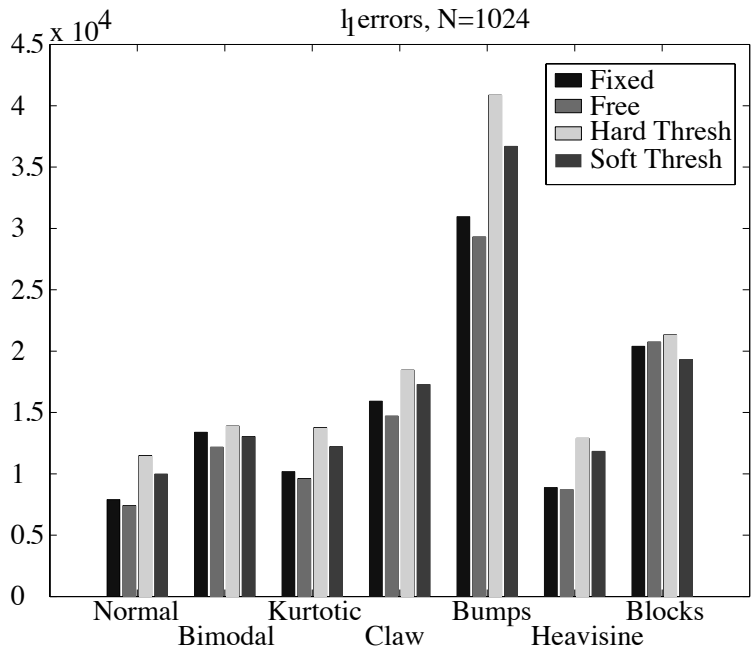

Fig. 2. Density Estimation using clairvoyant penalties.

or wavelet smoothness. These estimators have errors that converge nearly as quickly as the parametric rate in some cases.

As with wavelet-based and most other forms of multiscale analysis, the estimates produced by this proposed PLE method commonly exhibit change-points on the boundaries of the underlying recursive dyadic partition. Smoother estimates with the same theoretical advantages can be obtained through the use of Alpert bases [21] for moment interpolation as described by Donoho [22]. Future work in multiscale density estimation includes the investigation of translation invariant methods for improved change-point localization.

The methods presented here are applicable to a wide range of applications due to the general nature of the penalized likelihood objective function. For example, estimating the time-varying intensity of a Poisson process is a problem closely linked to density estimation. In density estimation, $\boldsymbol{f}$ represents the pmf and is constrained such that $\sum_i f_i = 1$, while in the Poisson case, $\boldsymbol{f}$ represents the Poisson intensity vector with no constraint on the total magnitude of the intensity. This distinction has a trivial impact on the risk bounds established in Section IV. Poisson intensity estimation is an important problem with several applications in astronomy and medicine.

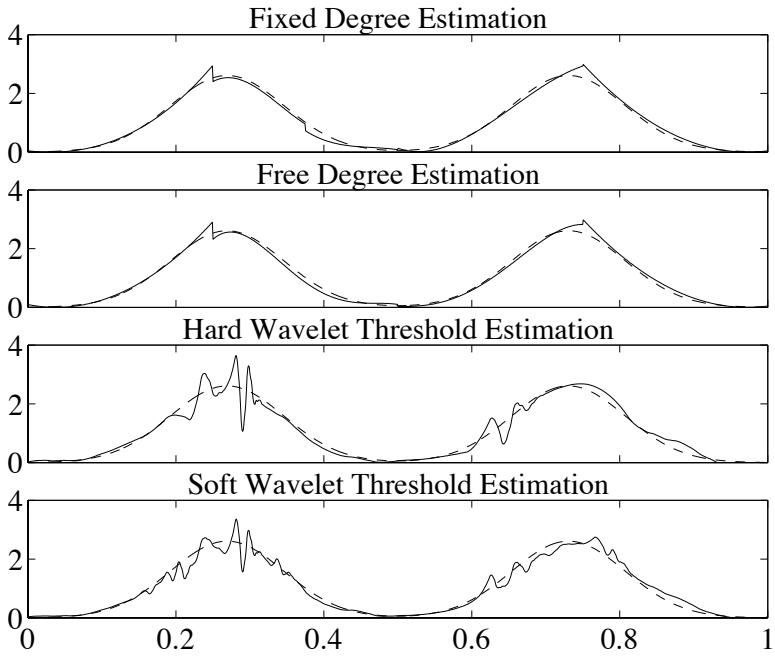

Fig. 3. Density estimation results for the bimodal density.

Finally, we mention possible extensions to multivariate density estimation. It is straightforward to demonstrate, using arguments parallel to the ones presented in this paper, that the near-minimax optimal convergence rates hold for $f : \mathbb{R}^d \to \mathbb{R}$ when f is in a Hölder-α smoothness space. Extensions to multivariate Besov spaces are not straightforward using our piecewise polynomial framework. On the other hand, wavelet methods are easily extended to deal with densities belonging to multivariate Besov spaces, due to the simple characterizations of such spaces in terms of wavelet expansions. However, the issue of characterizing spaces of inhomogeneous functions in multiple dimensions is an open problem in approximation theory. Besov spaces are not necessarily the appropriate framework for dealing with multivariate densities, due to the fact that singularities in multiple dimensions (*e.g.*, ridges) have a much richer structure than singularities in one dimension.

APPENDIX

I. PROOF OF RISK BOUND THEOREM

Proof of Theorem 1 A key step in proving this theorem and corollary involves bounding the expected loss, $\mathbb{E}_{\boldsymbol{f}}\left[L(\widehat{\boldsymbol{f}}, \boldsymbol{f})\right]$, using the Kullback-Leibler (KL) divergence. Kolaczyk and Nowak [14] accomplished this by proving the following theorem:

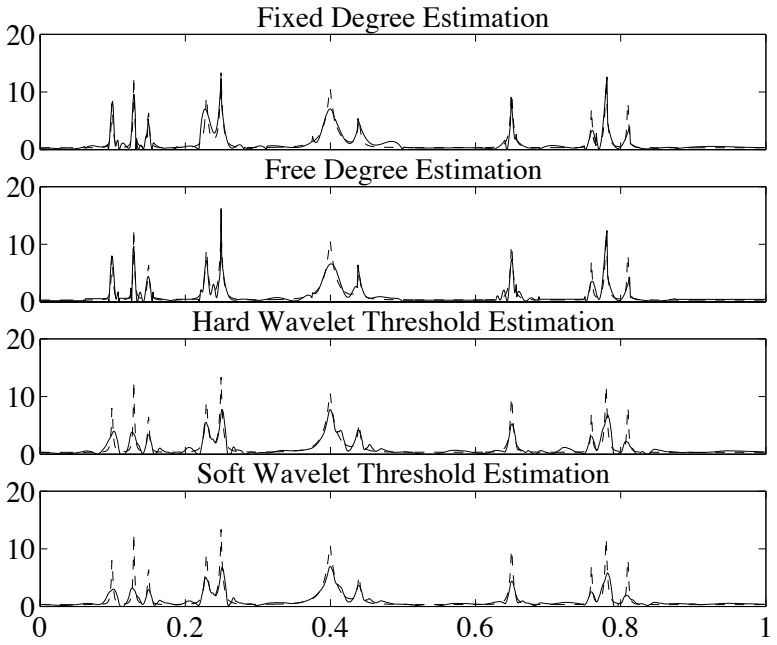

Fig. 4. Density estimation results for the bumps density.

Theorem 3 *Let Γ_N be a finite collection of estimators $\boldsymbol{f}'$ for $\boldsymbol{f}$, and pen$(\cdot)$ a function on Γ_N satisfying the condition*

$$\sum_{\boldsymbol{f}' \in \Gamma_N} e^{-\textit{pen}(\boldsymbol{f}')} \leq 1 \quad . \tag{10}$$

Let $\widehat{\boldsymbol{f}}$ be a penalized likelihood estimator given by

$$\widehat{\boldsymbol{f}}(\boldsymbol{x}) \equiv \arg \min_{\boldsymbol{f}' \in \Gamma_N} \{ -\log p\left(\boldsymbol{x} \mid \boldsymbol{f}'\right) + 2\textit{pen}\left(\boldsymbol{f}'\right) \} \quad . \tag{11}$$

Then

$$\mathbb{E}\left[H^2\left(p_{\widehat{\boldsymbol{f}}}, p_{\boldsymbol{f}}\right)\right] \leq \min_{\boldsymbol{f}' \in \Gamma_N} \{ K(p_{\boldsymbol{f}}, p_{\boldsymbol{f}'}) + 2\textit{pen}(\boldsymbol{f}') \} \quad . \tag{12}$$

Using the penalty function pen$(\boldsymbol{f}')$ as in (14) above, it is straightforward to demonstrate that the first condition (10) holds when Γ_N is defined as in the following lemma:

Lemma 2 *Let Γ_N be the collection of all N-length piecewise-polynomial vectors $\boldsymbol{f}'$, where each of m polynomial segments has at most r coefficients $a_{j,k} \in D_N[R_1, R_2]$, for some $R_1 < R_2$ for $k = 1, \ldots, r$ and $j = 1, \ldots, m$, and where $D_N[R_1, R_2]$ denotes a uniform discretization of the interval $[R_1, R_2]$ into $N^{1/2}$ equispaced values. Let pen$(\boldsymbol{f}'; m, r) \equiv \frac{mr}{2} \log_e N + (2m-1) \log_e 2$.*

Then

$$\sum_{\boldsymbol{f}' \in \Gamma_N} e^{-\mathit{pen}(\boldsymbol{f}')} \leq 1. \tag{13}$$

In other words, let Γ_N be a finite collection of order r piecewise polynomial estimators defined on a recursive dyadic partition of the observations. Using this, the estimate $\widehat{\boldsymbol{f}}$, defined by (2), can also be expressed as

$$\widehat{\boldsymbol{f}}(\boldsymbol{x}) \equiv \arg \min_{\boldsymbol{f}' \in \Gamma_N} \left[-\log p\left(\boldsymbol{x} \mid \boldsymbol{f}'\right) + \text{pen}(\boldsymbol{f}') \right] \quad . \tag{14}$$

In this expression, minimizing over a finite collection of estimators, Γ_N, is equivalent to the minimization over the finite collection of recursive partitions, $\mathcal{P}$, and coefficients, $\{a_{j,k}\}$, described in Section III.

Proof of Lemma 2

Consider constructing a unique code for every $f' \in \Gamma_N$. If f' consists of order-r polynomials on each of m dyadic intervals, then both the locations of the m intervals and the $m \cdot r$ coefficients need to be encoded. The m intervals can be encoded using $2m-1$ bits. To see this, note that dyadic intervals can be represented as leaf nodes of a binary tree, and a binary tree with m leaf nodes has a total of $2m-1$ nodes. Thus each node could be represented by one bit–a 0 for an internal node and a 1 for a leaf node. This can easily be verified with an inductive argument. Next the $m \cdot r$ coefficients need to be encoded. Since each coefficient has been quantized to one of $N^{1/2}$ levels, this requires a total of $\frac{mr}{2} \log_2 N$ bits. Thus each $\boldsymbol{f}' \in \Gamma_N$ can be uniquely encoded with a total of $2m - 1 + \frac{mr}{2} \log_2 N$ bits.

From the Kraft inequality, we know that the existence of this uniquely decodable scheme guarantees that

$$\sum_{\boldsymbol{f}' \in \Gamma_N} 2^{-\left(2m-1+\frac{mr}{2}\log_2 N\right)} \leq 1.$$

Therefore, if $\text{pen}(\boldsymbol{f}') = \frac{mr}{2}\log_e N + (2m-1)\log_e 2$,

$$\begin{aligned}\sum_{\boldsymbol{f}'\in\Gamma_N} e^{-\text{pen}(\boldsymbol{f}')} &= \sum_{\boldsymbol{f}'\in\Gamma_N} 2^{-\log_2(e)\left(\frac{mr}{2}\log_e N+(2m-1)\log_e 2\right)} \\ &= \sum_{\boldsymbol{f}'\in\Gamma_N} 2^{-\left(\frac{mr}{2}\log_2 N+2m-1\right)} \\ &\leq 1,\end{aligned}$$

as desired.

■

The next step in bounding the risk is to bound the KL divergence in (12). Kolaczyk and Nowak demonstrated that if $\boldsymbol{f}$ is a vector of multinomial probabilities, then

$$\frac{1}{N}K(p_{\boldsymbol{f}}, p_{\tilde{\boldsymbol{f}}'}) \leq \frac{2n}{C_\ell}\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 \tag{15}$$

where $n = \sum_i x_i$ is the total number of observations and C_ℓ is the lower bound on the true density $f(t)$.

The following construction can be used to bound the approximation error $\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2$: First let $\tilde{f}_a(\cdot)$ be the best m-piece order-r free-knot piecewise polynomial approximation of $f(\cdot)$. Next consider that the $m-1$ knots in $\tilde{f}_a(\cdot)$ will not normally lie on interval endpoints, but rather within one of the discrete intervals. Define $\tilde{f}_p(\cdot)$ to consist of the same m polynomials as $\tilde{f}_a(\cdot)$ but with each of the knots shifted to the nearest discrete interval endpoint. Note that, by construction $\tilde{f}_a(\cdot) = \tilde{f}_p(\cdot)$ on all but $m-1$ of the N discrete intervals. Finally, $\tilde{f}_p$ must be discretized to obtain a finite collection of models. This is accomplished by quantizing each of the r polynomial coefficients in each of the m polynomials to one of $N^{1/2}$ levels to yield the approximation $\tilde{f}'(\cdot)$. (Recall that $N = 2^{\lceil\log_2 n\rceil} \geq n$.) Now recall that $\boldsymbol{f}$ was defined on N discrete intervals as follows: $f_i \equiv \int_{I_i} f(t)dt$ for $i = 0 \ldots N-1$ and $I_i \equiv [i/N, (i+1)/N)$. Similarly, $\tilde{\boldsymbol{f}}'$ is defined by integrating $\tilde{f}'(\cdot)$ over the discrete intervals. This construction leads to the following bound on the ℓ_2 distance between the pmf $\boldsymbol{f}$ and the approximation $\tilde{\boldsymbol{f}}'$ (proved in Appendix B):

Lemma 3 *Let $\boldsymbol{f}$ be the sampled and quantized version of the true density $f(\cdot) \in B_\tau^\alpha$ using N*

integration samples, as defined above, and let $\tilde{\boldsymbol{f}}'$ *be a similarly sampled and quantized version of the best* m*-piece order* r *piecewise polynomial approximation of* $\boldsymbol{f}$*. Then*

$$\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 \leq C \left(\frac{m^{-r}}{N^{1/2}} + \frac{m^{1/2}}{N} + \frac{r}{N} \right)^2 \tag{16}$$

for some constant C*.*

Recall that $\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2$ was bounded for piecewise polynomial approximation in (16) above. This can be combined with (12) and (15) above to yield the bound

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \leq \min_{\boldsymbol{f}' \in \Gamma_N} \left\{ \frac{1}{N} \|\boldsymbol{f} - \boldsymbol{f}'\|_{\ell_2}^2 + \frac{2\gamma \log_2(N)}{N} \{\#\boldsymbol{\theta}_{\boldsymbol{f}'}\} \right\}.$$

This in turn is bounded above by $C \left(\frac{\log^2 N}{N} \right)^{\frac{2\alpha}{2\alpha+1}}$, as desired.

It has been established that

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \leq \min_{\boldsymbol{f}' \in \Gamma_N} \left\{ \frac{1}{N} \|\boldsymbol{f} - \boldsymbol{f}'\|_{\ell_2}^2 + \frac{2}{N} \text{pen}(\boldsymbol{f}') \right\}.$$

Note that the proposed method of estimating densities on recursive dyadic partitions typically requires a larger number of polynomial pieces than free-knot approximation would require. The term $\|\boldsymbol{f} - \boldsymbol{f}'\|_{\ell_2}^2$ was bounded assuming polynomial approximation was conducted on m (not necessarily dyadic) intervals. In practice, however, the proposed method would construct this approximation out of $m \log_2 N$ dyadic intervals. For example, if $\tilde{\boldsymbol{f}}'$ had one breakpoint at $N-1$, this would encoded using $\log_2(N)$ dyadic intervals: $\left[0, \frac{N}{2}\right)$, $\left[\frac{N}{2}, \frac{3N}{4}\right)$, etc. In general, any of the m polynomial segments represented by $\tilde{\boldsymbol{f}}'$ that do not lie on a dyadic partition need to be repartitioned a maximum of $\log_2(N)$ times. This combined with Theorem 3 yields:

$$\min_{\boldsymbol{f}' \in \Gamma_N^{(m)}} \left\{ n\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 + \frac{2}{N}\text{pen}(\tilde{\boldsymbol{f}}'; m\log_2 N, r) \right\} \leq$$

$$C_a m^{-2\alpha} + C_b\frac{m}{N} + C_c\frac{r^2}{N}$$
$$+C_{ab}\frac{m^{-\alpha+1/2}}{N^{1/2}} + C_{ac}\frac{m^{-\alpha}r}{N^{1/2}} + C_{bc}\frac{m^{1/2}r}{N}$$
$$+\frac{2}{N}\left(\frac{rm\log_2 N}{2}\log_e N + (2m\log_2 N - 1)\log_e 2\right) \tag{17}$$

This expression is minimized for $m \sim \left(\frac{\log^2(N)}{N}\right)^{\frac{-1}{2\alpha+1}}$. Substitution then yields that $R(\widehat{\boldsymbol{f}}, \boldsymbol{f})$ is bounded above by $C\left(\left(\frac{\log^2(N)}{N}\right)^{\frac{2\alpha}{2\alpha+1}}\right)$ for some constant C.

■

Proof of Corollary 1

The risk bound of Theorem 1 can be translated into an upper bound on the L_1 error between $\boldsymbol{f}$ and $\widehat{\boldsymbol{f}}$ as follows. First note that $\mathbb{E}\left[\sum_i |\widehat{f}_i - f_i|\right] \leq 2\mathbb{E}\left[H(\widehat{\boldsymbol{f}}, \boldsymbol{f})\right]$ because, for two densities p and q, $H^2(p,q) \leq \int |p-q| \leq 2H(p,q)$ [2]. This in turn is bounded by $2\mathbb{E}\left[\left(-2\log\mathcal{A}(\widehat{\boldsymbol{f}}, \boldsymbol{f})\right)^{1/2}\right]$, which follows from the analysis in [14]. This is equivalent to $2\mathbb{E}\left[\left(-(2/n)\log\mathcal{A}(p_{\widehat{\boldsymbol{f}}}, p_{\boldsymbol{f}})\right)^{1/2}\right]$, which can be derived by noting that $\mathcal{A}(\widehat{\boldsymbol{f}}, \boldsymbol{f}) = \sum_i (\widehat{f}_i f_i)^{1/2}$ and $\mathcal{A}(p_{\widehat{\boldsymbol{f}}}, p_{\boldsymbol{f}}) = \left(\sum_i (\widehat{f}_i f_i)^{1/2}\right)^n$. An application of Jensen's inequality leads to the bound $2\left(\mathbb{E}\left[-(2/n)\log\mathcal{A}(p_{\widehat{\boldsymbol{f}}}, p_{\boldsymbol{f}})\right]\right)^{1/2}$. Finally, this can be bounded by $C\left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$, for some constant C, which can be derived by noting from the analysis in [14] that the bound on $R(\widehat{\boldsymbol{f}}, \boldsymbol{f})$ is Theorem 1 is also a bound on $(1/N)\mathbb{E}[-2\log\mathcal{A}(p_{\widehat{\boldsymbol{f}}}, p_{\boldsymbol{f}})]$ and by recalling that $n \approx N$.

■

Proof of Corollary 2 The proof of the risk bound theorem for free-degree estimation requires another Kraft inequality:

Lemma 4 *Let Γ_N be the collection of all N-length free-degree piecewise-polynomial vectors defined on a recursive dyadic partition of the observations, and let d count the number of polynomial coefficients where the polynomial coefficients $a_k \in D_{N^2}[R_1, R_2]$, for some a $R_1 < R_2$ for $k = 1, \ldots, d$ and where $D_{N^2}[R_1, R_2]$ denotes a uniform discretization of the interval $[R_1, R_2]$*

into $N^{1/2}$ equispaced values. If $\boldsymbol{f}'$ has a total of m dyadic intervals, with the i^{th} interval having length $\ell_i \equiv N|I_i|$, and pen$(\boldsymbol{f}'; m, d, \{\ell_i\}) \equiv \frac{d}{2}\log_e N + (2m-1)\log_e 2 + \sum_{i=1}^m \log_e \ell_i$, then

$$\sum_{\boldsymbol{f}' \in \Gamma_N} e^{-\text{pen}(\boldsymbol{f}')} \leq 1. \tag{18}$$

Proof of Lemma 4 As in the fixed-degree case, this bound can be derived by considering a uniquely decodable scheme for encoding each $\boldsymbol{f}' \in \Gamma_N$. As before, the locations of the m dyadic intervals can be encoded using $2m-1$ bits. Now each of these intervals has length ℓ_i and represents a polynomial of order r_i, $i = 1, \ldots, m$, where $r_i \in \{1, \ldots, \ell_i\}$. Thus $\log_2 \ell_i$ bits are needed to encode the polynomial order, and $\frac{r_i}{2}\log_2 N$ bits are needed to encode each coefficient for the i^{th} interval. Thus the total number of bits needed to uniquely represent each $\boldsymbol{f}' \in \Gamma_N$ is $2m - 1 + \sum_{i=1}^m \left(\log_2 \ell_i + \frac{r_i}{2}\log_2 N\right) = 2m - 1 + \frac{d}{2}\log_2 N + \sum_{i=1}^m \log_2 \ell_i$.

The remainder of the proof follows that of Lemma 2. ■

As in the fixed-degree case, the next step in bounding the risk is to use the bound

$$\frac{1}{N}K(p_{\boldsymbol{f}}, p_{\tilde{\boldsymbol{f}}'}) \leq \frac{2n}{C_\ell}\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2.$$

Similar to Lemma 3 in the fixed-degree case, the ℓ_2 estimation error must be bounded for free-degree estimation, which is accomplished with the following lemma (proved in Appendix B):

Lemma 5 *Let $\boldsymbol{f}$ be the sampled and quantized version of the true density $f(t) = \beta_0 t^{\beta_1} + \beta_2$ using N integration samples for some constants β_i, $i = 1, 2, 3$ such that $0 < C_\ell \leq f(\cdot) \leq C_u$. Let $\tilde{\boldsymbol{f}}'$ be as above but with free-degree polynomial approximation on each partition interval, represented by a total of d coefficients. Then*

$$\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 \leq C\left(\frac{e^{-C_\beta\sqrt{d}}}{N^{1/2}} + \frac{d^{1/2}}{N} + \frac{d}{N}\right)^2 \tag{19}$$

for some constant C.

This bound quantifies the impact of shifting the knot locations and discretizing the polynomial coefficients. For a large number of discrete intervals, N, this impact is small relative to the free-knot piecewise polynomial approximation of the true density f.

In the case of free-degree estimation, equation (17) becomes

$$\begin{aligned}\min_{\boldsymbol{f}'\in\Gamma_N^{(m)}}&\left\{n\|\boldsymbol{f}-\tilde{\boldsymbol{f}}'\|_{\ell_2}^2+\frac{2}{N}\text{pen}(\boldsymbol{f}';m,d,\{\ell_i\})\right\}\leq\\&C_ae^{-2C_\beta\sqrt{d}}+C_b\frac{m}{N}+C_c\frac{d^2}{N}\\&+C_{ab}\frac{e^{-C_\beta\sqrt{d}}m^{1/2}}{N^{1/2}}+C_{ac}\frac{e^{-C_\beta\sqrt{d}}d}{N^{1/2}}+C_{bc}\frac{m^{1/2}d}{N}\\&+\frac{2}{N}\left(\frac{d}{2}\log_e^2N+(2m\log_2N-1)\log_e2+\sum_{i=1}^{m\log_2N}\log_e\ell_i\right)\end{aligned}$$

Note that $m\leq d$ and that since $\ell_i\leq\frac{N}{2}$ for $m>1$, $\sum_{i=1}^m\log_e\ell_i\leq m\log_eN-m\log_e2$. This expression is minimized for $d\sim\frac{\log^2(N)}{C_\beta^2}$. Substitution then yields that $R(\widehat{\boldsymbol{f}},\boldsymbol{f})$ is bounded above by $C\left(\frac{\log^4N}{N}\right)$ for some constant C.

■

II. Discrete Approximation Error Bounds

We begin with the proof of the corollary because of its relative complexity. The proof of the theorem will then follow with only minor modifications to the proof of the corollary.

Proof of Lemma 5 Define d as the total number of coefficients required to represent this polynomial, and let m be the number of polynomial pieces in $\tilde{f}_a(\cdot)$. Using the construction of $\tilde{\boldsymbol{f}}'$ outlined above and the triangle inequality,

$$\begin{aligned}\|\boldsymbol{f}-\tilde{\boldsymbol{f}}'\|_{\ell_2}^2&=\|(\boldsymbol{f}-\tilde{\boldsymbol{f}}_a)+(\tilde{\boldsymbol{f}}_a-\tilde{\boldsymbol{f}}_p)+(\tilde{\boldsymbol{f}}_p-\tilde{\boldsymbol{f}}')\|_{\ell_2}^2\\&\leq\|\boldsymbol{f}-\tilde{\boldsymbol{f}}_a\|_{\ell_2}^2+\|\tilde{\boldsymbol{f}}_a-\tilde{\boldsymbol{f}}_p\|_{\ell_2}^2+\|\tilde{\boldsymbol{f}}_p-\tilde{\boldsymbol{f}}'\|_{\ell_2}^2\\&\quad+2\|\boldsymbol{f}-\tilde{\boldsymbol{f}}_a\|_{\ell_2}\|\tilde{\boldsymbol{f}}_a-\tilde{\boldsymbol{f}}_p\|_{\ell_2}+2\|\boldsymbol{f}-\tilde{\boldsymbol{f}}_a\|_{\ell_2}\|\tilde{\boldsymbol{f}}_p-\tilde{\boldsymbol{f}}'\|_{\ell_2}\\&\quad+2\|\tilde{\boldsymbol{f}}_a-\tilde{\boldsymbol{f}}_p\|_{\ell_2}\|\tilde{\boldsymbol{f}}_p-\tilde{\boldsymbol{f}}'\|_{\ell_2}.\end{aligned}\tag{20}$$

Note that the last three terms are combinations of the square roots of the first three terms. Thus, after bounding the first three terms, a total bound can be easily calculated. The first three

terms of (20) can each be bounded as follows:

$$\begin{aligned} \|\boldsymbol{f} - \tilde{\boldsymbol{f}}_a\|_{\ell_2}^2 &\leq C_a \frac{e^{-2C_\beta\sqrt{d}}}{N}, \\ \|\tilde{\boldsymbol{f}}_a - \tilde{\boldsymbol{f}}_p\|_{\ell_2}^2 &\leq C_b \frac{m}{N^2}, \text{ and} \\ \|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 &\leq C_c \frac{d^2}{N^2}. \end{aligned}$$

Consider each of the three bounds separately.

Term a: As in Kolaczyk and Nowak's work, the Haar basis is an effective tool for bounding $\|\boldsymbol{f} - \tilde{\boldsymbol{f}}_a\|_{\ell_2}^2$ [14]. Specifically, let $h_{j,k}(i)$ be the (j,k)-th Haar function on the discrete space $\{0, 1, \ldots, N-1\}$; in other words, $h_{j,k}(i) \equiv (\chi_{j+1,2k+1}(i) - \chi_{j+1,2k}(i))/N_{j,k}^{1/2}$, where $\chi_{j,k}$ is the characteristic function for the discrete analogue of the interval $I_{j,k} \equiv [k/2^j, (k+1)/2^j)$, for $j = 0, \ldots, J-1, k = 0, \ldots, 2^j - 1$, and $J = \log_2(N)$; $N_{j,k} = N/2^j$ is the cardinality of this set. Similarly, let $h_{j,k}^c(t)$ be the continuous analog of $h_{j,k}(i)$ on the interval $[0,1]$, or $h_{j,k}^c(t) \equiv 2^{j/2}(\chi_{j+1,2k+1}^c(t) - \chi_{j_1,2k}^c(t))$. It then follows that $\langle \boldsymbol{f}, h_{j,k}\rangle_{\ell_2} = N^{-1/2}\langle f, h_{j,k}^c\rangle_{L_2}$; combining this with Parseval's relation yields

$$\begin{aligned} \|\boldsymbol{f} - \tilde{\boldsymbol{f}}_a\|_{\ell_2}^2 &= \sum_{(j,k)\in\mathcal{J}} \left(\langle \boldsymbol{f}, h_{j,k}\rangle_{\ell_2} - \langle \tilde{\boldsymbol{f}}_a, h_{j,k}\rangle_{\ell_2}\right)^2 \\ &= \frac{1}{N} \sum_{(j,k)\in\mathcal{J}} \left(\langle f, h_{j,k}^c\rangle_{L_2} - \langle \tilde{f}_a, h_{j,k}^c\rangle_{L_2}\right)^2 \end{aligned} \tag{21}$$

where $\mathcal{J}$ is the set of all (j,k) with $j = 0, 1, \ldots, J-1$ and $k = 0, 1, \ldots, 2^j - 1$. The expression (21) is bounded above by a similar sum over all (j,k), which is equal to the squared L_2 approximation error, $\|f - \tilde{f}_a\|_{L_2([0,1])}^2$. The analysis in [10] establishes that $\|f - \tilde{f}_a\|_{L_2}^2 = C_a e^{-2C_\beta d}$, when $f(t) = \beta_0 t^{\beta_1} + \beta_2$ for some constants β_i, $i = 1, 2, 3$ such that $0 < C_\ell \leq f(\cdot) \leq C_u$. In the L_2 bound, d is the total number of polynomial coefficients in $\tilde{f}_a(\cdot)$ and C_a is a constant.

Term b: Because $0 < C_\ell \leq f(\cdot) C_u$, the analysis in [15] implies $\|f - \tilde{f}_a\|_{L_\infty} \leq C_\infty < \infty$. By construction, $\tilde{f}_a(\cdot)$ has $m-1$ breakpoints. Thus for all but $m-1$ of these intervals, $\tilde{f}_{f,i} = \tilde{f}_{p,i}$.

For the remaining $m-1$ intervals,

$$\begin{aligned} |\tilde{f}_{f,i} - \tilde{f}_{p,i}| &\leq \int_{I_i} \left| \tilde{f}_a(t) - \tilde{f}_p(t) \right| dt \\ &\leq \frac{C_u - C_\ell + 2C_\infty}{N} \end{aligned}$$

which leads to the final bound

$$\|\tilde{\boldsymbol{f}}_a - \tilde{\boldsymbol{f}}_p\|_{\ell_2}^2 \leq C_b \, \frac{m}{N^2} \tag{22}$$

where C_b is a constant independent of m and N.

Term *c*: Quantization of each of the d polynomial coefficients produces the final error term. The polynomials can be expressed in terms of a shifted orthogonal Chebyshev polynomial basis, which allows the magnitudes of the coefficients to be bounded and hence quantized. Let r_j be the polynomial order on the j^{th} interval I_j, $j = 1, \ldots, m$. Note that $\sum_{j=1}^{m} r_j = d$ and $r_j \leq \ell_j \equiv N|I_j|$. The basis representation can then be expressed as follows:

$$\tilde{f}_p(t) = \sum_{j=1}^{m} \sum_{k=0}^{r_j - 1} a_{j,k} \widetilde{T}_{I_j,k}(t) \chi_{I_j}(t)$$

where $a_{j,k}$ is the basis coefficient, χ is the indicator function, and $\widetilde{T}_{I_j,k}(t)$ is the *normalized* k^{th} Chebyshev polynomial with support shifted and scaled to the interval I_j. Specifically, $\widetilde{T}_{[a,b],k}(t) \equiv \left(\frac{s}{\pi}\right)^{1/2} \left(\frac{2}{b-a}\right)^{1/2} T_k\left(\frac{2t-(b+a)}{b-a}\right)$, where $s \equiv (k > 0) + 1$ and T_k is the k^{th} Chebyshev polynomial with the property that $|T_k(t)| \leq 1$. Note that $\{\widetilde{T}_{[a,b],k}\}_k$ form an orthogonal (but not orthonormal) basis for $[a, b]$. In other words, $\int_a^b \frac{\left[\widetilde{T}_{[a,b],k}(t)\right]^2}{\sqrt{1-t^2}} dt = 1$. This means that the coefficient $a_{j,k}$ can be expressed as the weighted inner product of $\tilde{f}_p$ and $\widetilde{T}_{I_j,k}$. The magnitudes of these coefficients can be bounded above using an $L_1 - L_\infty$ bound argument; specifically, if

$I_j = [a, b)$,

$$\begin{aligned}
|a_{j,k}| &= \left| \int_{I_j} \frac{\tilde{f}_p(t)\widetilde{T}_{I_j,k}(t)}{\sqrt{1-\left(\frac{2t-(a+b)}{b-a}\right)^2}} dt \right| \\
&\leq (C_u + C_\infty)\left(\frac{2}{|I_j|}\right)^{1/2} \int_{I_j} \frac{1}{\sqrt{1-\left(\frac{2t-(a+b)}{b-a}\right)^2}} dt \\
&\leq C'''\left(\frac{2}{|I_j|}\right)^{1/2} \int_{-1}^{1} \frac{1}{\sqrt{1-y^2}} \frac{|I_j|}{2} dy \\
&= C''|I_j|^{1/2} .
\end{aligned}$$

The second inequality relies on the initial assumption that $\tilde{f}_p$ is bounded. It is now possible to quantize this coefficient to one of $N^{1/2}$ levels in $\left[-|I_j|^{1/2}C'', |I_j|^{1/2}C''\right]$. Let the quantized version of coefficient $a_{j,k}$ be denoted $[a_{j,k}]$. This quantization induces the following error for a given $t \in I_j$:

$$\begin{aligned}
|\tilde{f}_p(t) - \tilde{f}'(t)| &= \left| \sum_{k=0}^{r_j-1} (a_{j,k} - [a_{j,k}])\widetilde{T}_{I_j,k}(t) \right| \\
&\leq \frac{C''}{N^{1/2}} \sum_{k=0}^{r_j-1} \left(\frac{|I_j|}{|I_j|}\right)^{1/2} \\
&= C''\frac{r_j}{N^{1/2}}
\end{aligned}$$

The above bound leads to a bound on the difference between vector elements where $I_i \subset I_j$ (recall $|I_i| = 1/N$):

$$\begin{aligned}
|\tilde{f}_{p,i} - \tilde{f}'_i| &\leq \int_{I_i} |\tilde{f}_p(t) - \tilde{f}'(t)| dt \\
&\leq C''\frac{r_j}{N^{3/2}}
\end{aligned}$$

This yields the final bound:

$$\begin{aligned}\|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 &\leq C_c \sum_{j=1}^{m} \ell_j \left(\frac{r_j}{N^{3/2}}\right)^2 \\ &\leq \frac{C_c}{N^3} \sum_{j=1}^{m} \ell_j (r_j)^2 \\ &\leq \frac{C_c d^2}{N^2}\end{aligned}$$

for some constant C_c. The last inequality holds because $\ell_j \leq N$ and because $\sum_j r_j^2 \leq \left(\sum_j r_j\right)^2 = d^2$.

The three bounds in the lemma can now be used to bound the remaining terms in (20). For example, the fourth term is simply twice the square root of term a times the square root of term b, etc. Specifically,

$$\begin{aligned}2\|\boldsymbol{f} - \tilde{\boldsymbol{f}}_a\|_{\ell_2}\|\tilde{\boldsymbol{f}}_a - \tilde{\boldsymbol{f}}_p\|_{\ell_2} &\leq C_{ab}\, \frac{e^{-C_\beta\sqrt{d}} m^{1/2}}{N^{3/2}}, \\ 2\|\boldsymbol{f} - \tilde{\boldsymbol{f}}_a\|_{\ell_2}\|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}'\|_{\ell_2} &\leq C_{ac}\, \frac{e^{-C_\beta\sqrt{d}} d}{N^{3/2}}, \text{ and} \\ 2\|\tilde{\boldsymbol{f}}_a - \tilde{\boldsymbol{f}}_p\|_{\ell_2}\|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}'\|_{\ell_2} &\leq C_{bc}\, \frac{m^{1/2} d}{N^2}\end{aligned}$$

where C_{ab}, C_{ac}, and C_{bc} are constants. Now the entire quantity $\|\boldsymbol{f} - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2$ can be bounded by sum of the six bounds just derived.

■

Proof of Lemma 3

Define $\boldsymbol{f}$ as in the corollary, and let $\tilde{f}_a(\cdot)$ be the best m-piece free-knot piecewise polynomial approximation of $f(\cdot)$. Define the vectors $\tilde{\boldsymbol{f}}_a$, $\tilde{\boldsymbol{f}}_p$, and $\tilde{\boldsymbol{f}}'$ as in the proof of the corollary. Now the first three terms of (20) can each be bounded as follows:

$$\begin{aligned}\|\boldsymbol{f} - \tilde{\boldsymbol{f}}_a\|_{\ell_2}^2 &\leq C_a \frac{m^{-2\alpha}}{N}, \\ \|\tilde{\boldsymbol{f}}_a - \tilde{\boldsymbol{f}}_p\|_{\ell_2}^2 &\leq C_b\, \frac{m}{N^2}, \text{ and} \\ \|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 &\leq C_c\, \frac{r^2}{N^2}.\end{aligned}$$

Consider each of the three bounds separately.

Term a: This term is bounded in a manner parallel to that in the corollary with the exception that a different L_2 bound is is applied. The L_2 approximation error for either an m-term wavelet approximation or an m-piece piecewise polynomial approximation of order r (degree $r-1$) decays faster than $C_r m^{-r}$ for some constant C_r depending only on r when $f \in B^{\alpha}_{\tau}(L_{\tau}([0,1]))$ [15].

Term b: This term is bounded in a manner identical to that in the corollary.

Term c: This term is also bounded in a manner parallel to that in the corollary, but in this case $r_j = r$. This results in the bound:

$$\|\tilde{\boldsymbol{f}}_p - \tilde{\boldsymbol{f}}'\|_{\ell_2}^2 \quad \leq \quad C_c \frac{r^2}{N^2}$$

for some constant C_c.

The remainder of the proof follows the proof of lemma 5. ■

III. Proof of Computational Complexity Lemma and Theorem

Proof of Lemma 1 Because the multinomial probabilities must sum to one, there are only $M-1$ degrees of freedom in choosing $\boldsymbol{\theta} \in \mathbb{R}^M$. Let $\boldsymbol{\theta}_{M-1}$ be $M-1$ of the M polynomial coefficients. It is easy to check that the log multinomial likelihood function is convex in the multinomial probabilities $\boldsymbol{f} = T\,\boldsymbol{\theta}_{M-1}$. To prove the lemma, we refer to Theorem 5.7 in [23], which states that if T is a linear transformation from $\mathbb{R}^N$ to $\mathbb{R}^{M-1}$, then, for each convex function g on $\mathbb{R}^{M-1}$, the function $h \equiv gT$ defined by

$$(gT)(\boldsymbol{\theta}) \equiv g(T\boldsymbol{\theta})$$

is convex in $\boldsymbol{\theta}$ on $\mathbb{R}^N$. This and the concavity of the multinomial log likelihood shows that $h(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$.

To see that Θ is a convex set, consider two admissible coefficient vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ defining two *bona fide* densities $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, respectively. Then for any $\lambda < 1$ the density $\boldsymbol{f}_3 = \lambda \boldsymbol{f}_1 + (1-\lambda)\boldsymbol{f}_2$ is also a *bona fide* density, and can be described by the coefficient vector $\boldsymbol{\theta}_3 = \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2$ is also admissible. As a result, the set is convex. ■

Proof of Theorem 2 Recall $n \approx N$.

- *Free-degree, exact:* At level j in the binary tree, there are 2^j nodes and $N/2^j - 1$ possible polynomial orders for each node. This means that all the polynomial coefficients at level j can be computed with $O(N)$ calls to a convex optimization program. Thus the entire program requires $O(N \log N)$ calls to a convex optimization program and $O(N)$ comparisons between different polynomial models.
- *Free-degree, approximate:* Note that there are $N/2^j$ counts under each node at level j in the binary tree; if a least-squares fit of d coefficients to $N/2^j$ data points takes $O(dN/2^j)$ operations, then the total computational complexity can be expressed as

$$\sum_{j=0}^{\log_2 N-1} 2^j \sum_{d=1}^{N/2^j-1} O(dN/2^j) = O(N^3).$$

- *Fixed-degree, exact:* At level j in the binary tree, all the order-r polynomial coefficients can be computed with $O(2^j)$ calls to a convex optimization program. Thus the entire program requires $\sum_{j=0}^{\log_2 N-1} O(2^j) = O(N)$ calls to a convex optimization program and $O(N)$ comparisons between different polynomial models.
- *Fixed-degree, approximate:* Similar to the fixed-degree exact case, the fixed-degree approximate case total computational complexity can be expressed as

$$\sum_{j=0}^{\log_2 N-1} 2^j O(rN/2^j) = O(N \log N).$$

■

REFERENCES

[1] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?," *Journal of the Royal Statistical Society*, pp. 301–337, 1995.

[2] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, Ann Arbor, MI, 2001.

[3] B. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[4] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York, 1992.

[5] B. Prakasa-Rao, *Nonparametric Functional Estimation*, Academic Press, Orlando, 1983.

[6] R. Eubank, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York, 1988.

[7] R. Tapia and J. Thompson, *Nonparametric Probability Density Estimation*, John Hopkins University Press, Baltimore, 1978.

[8] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Statist.*, vol. 24, pp. 508–539, 1996.

[9] J. Fan, I. Gijbels, T. Hu, and L. Huang, "A study of variable bandwidth selection for local polynomial regression," *Statistica Sinica*, vol. 6, pp. 113–127, 1996.

[10] R. DeVore and K. Scherer, *Quantitative Approximation*, chapter Variable Knot, Variable Degree Spline Approximation to X^β, Academic Press, New York, 1980.

[11] A. Barron and T. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.

[12] Q. Li, *Estimation of Mixture Models*, Ph.D. thesis, Yale University, 1999.

[13] Q. Li and A. Barron, *Advances in Neural Information Processing Systems 12*, chapter Mixture Density Estimation, MIT Press, 2000.

[14] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," submitted to Annals of Stat. August, 2001. Available at http://cmc.rice.edu/docs/info.pl?doc=Kol2001Aug1Multiscale.

[15] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[16] D. Donoho, "Cart and best-ortho-basis selection: A connection," *Annals of Stat.*, vol. 25, pp. 1870–1911, 1997.

[17] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1983.

[18] M. Unser and M. Eden, "Maximum likelihood estimation of linear signal parameters for poisson processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 942–5, 1988.

[19] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[20] J. Marron and M. Wand, "Exact mean integrated squared error," *Annals of Statistics*, vol. 20, no. 2, pp. 712–736, 1992.

[21] B. Alpert, "A class of bases in l2 for sparse representation of integral operators," *SIAM Journal of Mathematical Analysis*, vol. 24, pp. 246–262, 1993.

[22] D. Donoho, N. Dyn, D. Levin, and T. Yu, "Smooth multiwavelet duals of alpert bases by moment-interpolation, with applications to recursive partitioning," Tech. Rep., Department of Statistics, Stanford University, 1996.

[23] T. R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1972.