Multiscale Wavelets on Trees, Graphs and High Dimensional Data: Theory and Applications to Semi Supervised Learning

Matan Gavish[1] GAVISH@STANFORD.EDU
Boaz Nadler BOAZ.NADLER@WEIZMANN.AC.IL
Weizmann Institute of Science, P.O. Box 26, Rehovot, 76100, Israel

Ronald R. Coifman RONALD.COIFMAN@YALE.EDU
Yale University, New Haven, CT, 06520, USA

Abstract

Harmonic analysis, and in particular the relation between function smoothness and approximate sparsity of its wavelet coefficients, has played a key role in signal processing and statistical inference for low dimensional data. In contrast, harmonic analysis has thus far had little impact in modern problems involving high dimensional data, or data encoded as graphs or networks. The main contribution of this paper is the development of a harmonic analysis approach, including both learning algorithms and supporting theory, applicable to these more general settings. Given data (be it high dimensional, graph or network) that is represented by one or more hierarchical trees, we first construct multiscale wavelet-like orthonormal bases on it. Second, we prove that in analogy to the Euclidean case, function smoothness with respect to a specific metric induced by the tree is equivalent to exponential rate of coefficient decay, that is, to approximate sparsity. These results readily translate to simple practical algorithms for various learning tasks. We present an application to transductive semi-supervised learning.

1. Introduction

In recent years, vast data sets in the form of (i) graphs or networks, and (ii) data in high dimensional Euclidean space, are routinely collected in many areas. Analysis of these types of data is a major challenge for the statistics and machine learning communities.

The statistical theory underlying data analysis has been studied extensively in both communities. The traditional statistics literature has, by and large, focused on the setting of data in (low dimensional) Euclidean space (Lehmann & Casella, 2003; Hardle et al., 1998). In contrast, the theoretical machine learning community has developed a theory of learning from abstract hypothesis classes (Vapnik, 1998), where notions of function smoothness and the geometry of the ambient space are often absent.

Neither of these traditional approaches is particularly well suited to the graph or high-dimensional Euclidean settings. Many traditional statistical inference approaches are inapplicable for high dimensional Euclidean data and become meaningless on non-Euclidean data such as a graph. Nonetheless, both graph data and high dimensional data typically have a *rich geometrical structure*, which is often not directly exploited in the abstract machine learning framework. New statistical learning tools are thus needed, which would capitalize on the geometrical structure available in these settings. These tools should be accompanied by an underlying theory, specifying the conditions under which they can be expected to be useful.

In this work we propose a harmonic analysis framework and develop such tools and supporting theory, under the assumption that the geometry of a graph or a high-dimensional Euclidean data set is captured by a *hierarchical tree* of increasingly refined partitions. We present two main contributions. First, given data encoded as a hierarchical tree, we build data adaptive wavelet-like orthonormal bases for the space of functions over the data set, in the spirit of the Haar basis on the unit interval $[0, 1]$ (see Fig. 1). Second, we

[1]Currently at Stanford University, Stanford, CA.

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

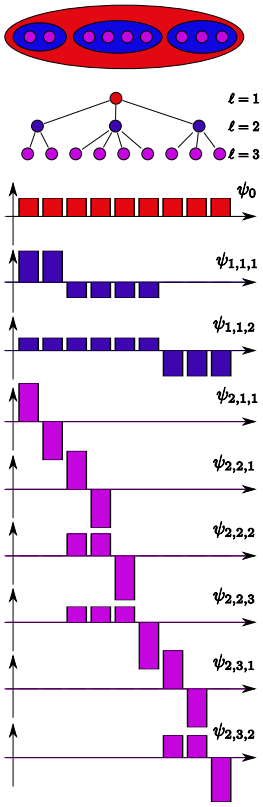

Figure 1. An illustration of a Haar-like basis.

prove that for these bases, function smoothness with respect to a certain tree metric can be measured by the rate of function coefficient decay. In particular, for a balanced tree (defined below), smooth functions have coefficient decay *exponential* in the tree level.

As fast coefficient decay, or approximate sparsity, is the key principle underlying many inference techniques in the Euclidean case, our results readily translate into new simple and practical algorithms for various learning tasks in the more general setting of graphs and high dimensional data. As a particular application of our approach, we present a novel semi-supervised learning (SSL) scheme. Preliminary results on various datasets show that our scheme is competitive with other approaches, often achieving lower prediction errors.

2. Problem Setting and Related Work

Let $X = \{x_1, \ldots, x_N\}$ be the dataset we wish to analyze. The samples x_i may be points in a high dimensional space, or nodes in a weighted graph or network. Most supervised learning problems in this setting involve inference on an unknown target function f defined on X. Examples include i) *function denoising*: given values $y_i = f(x_i) + \epsilon_i$, where ϵ_i are i.i.d. noise variables, estimate the underlying f, ii) *semi-supervised learning*: given $f|_S$ where S is a subset of X, estimate $f|_{X \setminus S}$, iii) *active learning*: given $f|_S$ choose the next unlabeled point $x \in X \setminus S$ to query, to "optimally" estimate f with as few queries as possible.

The key question is thus what are good methods to represent, process and learn functions on general datasets X having the form of a weighted graph or points in a high dimensional space.

In the ML community, a common approach to handle high dimensional data is to represent it as a symmetric weighted graph with the aid of a similarity kernel. Common methods to process functions defined on graphs, in turn, are based on the graph Laplacian matrix L, and its variants (Chapelle et al., 2006). For example, in (Zhu et al., 2003) global function smoothness w.r.t. the graph is measured by $f^T L f$. A related approach involves representing the target function in the eigenbasis of the graph Laplacian (Belkin & Niyogi, 2003). Here learning is performed by estimating the first few eigenbasis coefficients from the labeled data.

Despite their empirical success on various datasets, these global methods have several limitations. First, as described in (Nadler et al., 2009), measuring function smoothness via the graph Laplacian ($f^T L f$) leads to ill-posed problems for semi-supervised learning with high dimensional data as the number of unlabeled data grows to infinity. Second, eigenvector based methods are not always best suited for representing functions on a graph. Since these basis vectors are eigenvectors of a symmetric matrix, in general they have *global* support and become increasingly *oscillatory*. This limits the number of coefficients that can be robustly estimated. On the theoretical side, there is still no sufficient justification for these methods for data that is not necessarily sampled from a low dimensional manifold. Furthermore, to date, even in the manifold setting there is no well understood theory for how many coefficients to estimate. For example, Belkin & Niyogi (2003), heuristically propose to estimate $n/5$ coefficients where n is the total number of labeled points.

Inspired by the classical Euclidean setting, where similar limitations of the Fourier basis are alleviated by wavelet bases, in this paper we propose a *multiscale* harmonic analysis approach to learning from graph or high dimensional data and develop both new algorithms as well as supporting theory. Our key assumption is that the geometry and structures of the input graph or high dimensional data are captured by one or several (possibly random) hierarchical trees. First,

we remark that trees are ubiquitous in statistics and computer science. Given a dataset X, there are many methods to construct a hierarchical tree, including deterministic, random, agglomerative and divisive. Furthermore, in a Euclidean setting, tree-based classifiers are highly successful (Breiman et al., 1984; Breiman, 2001; Binev et al., 2005). Note, however, that our setting is different as we do not necessarily assume a Euclidean structure. In this paper we do not focus on the exact method of tree construction, but rather assume that the input data X is equipped with a hierarchical tree, either already given or constructed by some method. Our main results are a sensible definition of function smoothness with respect to this tree, a corresponding multiscale learning approach and its supporting theory, assuming this function smoothness.

As mentioned above, harmonic analysis has had thus far little impact on learning from graphs or from high dimensional data. As such, there are relatively few works suggesting multiscale representations for learning in these settings (Coifman & Maggioni, 2006; Mahadevan & Maggioni, 2006; Jansen et al., 2009). Jansen et al. (2009) explicitly state the lack of supporting theory for their methods: "The main disadvantage is that, apart from analogies with regular wavelets, there is currently no substantial body of theory behind our methods". This work provides a step forward towards the development of such a theory.

3. Main Results

Let X be the given dataset, and let $f : X \to \mathbb{R}$ be the unknown target function to be learned. Further, denote by $V = \{f \,|\, f : X \to \mathbb{R}\}$ the space of all functions on the dataset, with the inner product

$$\langle f, g\rangle = \frac{1}{N}\sum_{j=1}^{N} f(x_j)g(x_j). \tag{1}$$

Inspired by the success of multiscale wavelet decompositions for 1-d signals and 2-d images (Mallat, 1999; Hardle et al., 1998), our harmonic analysis approach consists of constructing a multiscale basis $\{\phi_j\}_{j=1}^{N}$ for the space V, such that the target function f admits an efficient sparse representation in this basis. Accurate approximation of f is then possible by estimating only a few of its coefficients $\langle f, \phi_j\rangle$.

3.1. Trees and Multi-Resulution Analysis

The starting point for constructing a basis for V is a *hierarchical tree* representation of the data X. We denote by $\ell = 1, \ldots, L$ the level in the tree with $\ell = 1$ being the root and $\ell = L$ the lowest level, where each sample x_j is a single leaf. We further denote by X_k^ℓ the set of all leaves of the k-th folder (subtree) of the tree at level ℓ, and by $sub(\ell, k)$ the number of subfolders of X_k^ℓ at the next level $\ell + 1$.

The crucial property we require of the given dataset X is that the resulting tree is *balanced*: that is, for all parent folders in the tree,

$$0 < \underline{B} \leq \frac{|\text{offspring folder}|}{|\text{parent folder}|} \leq \overline{B} < 1 \tag{2}$$

This property implies that $L = O(\log N)$. Note that $\underline{B} = \overline{B} = 1/2$ gives a perfectly balanced binary tree.

Our next step is based on the following simple observation (see also Donoho (1997); Murtagh (2007); Lee et al. (2008)): *A tree representation of data naturally induces a multi-resolution analysis (MRA) with an associated Haar-like wavelet basis.* In more detail, for each level ℓ denote by V^ℓ the space of functions constant on all folders (subtrees) at level ℓ,

$$V^\ell = \{f \,|\, f : X \to \mathbb{R}, f \text{ constant on all folders } X_j^\ell\}.$$

For example, V^1 is the one-dimensional space of constant functions on the dataset, $V^1 = \text{Span}_{\mathbb{R}}\{\mathbf{1}_X\}$, whereas $V^L = V$. By construction,

$$V^1 \subset \ldots \subset V^\ell \subset V^{\ell+1} \subset \ldots \subset V^L = V. \tag{3}$$

This sequence of subspaces resembles a multiresolution analysis of V, a key property for the development of wavelets on Euclidean spaces (Mallat, 1999). As in classical MRA, let W^ℓ $(1 \leqslant \ell < L)$ be the orthogonal complement of V^ℓ in $V^{\ell+1}$ ($V^{\ell+1} = W^\ell \bigoplus V^\ell$). The space of all functions V can then be decomposed as

$$V = V^L = [\bigoplus_{\ell=1}^{L-1} W^\ell] \bigoplus V^1 . \tag{4}$$

3.2. Haar-like bases on the dataset

Eq. (4) is the key to construct multiscale, localized orthonormal bases for the space V. Consider a folder X_k^ℓ at level ℓ that is split into two subfolders $X_i^{\ell+1}$ and $X_j^{\ell+1}$. Then, there is a zero mean Haar-like function $\psi_{\ell,k,1}$ supported only on these two subfolders, which is piecewise constant on each of them (see for example $\psi_{2,1,1}$ in Fig. 1). If a folder X_k^ℓ is split into three or more subfolders, then $sub(\ell, k) - 1$ Haar-like orthonormal functions $\{\psi_{\ell,k,j}\}_{j=1}^{sub(\ell,k)-1}$ need to be constructed.

The collection of all these functions, augmented by the constant function on X, forms an orthonormal basis of V, that we term a *Haar-like* basis, $\mathfrak{B} = \{\psi_{\ell,k,j}\}$. To clarify notation, ℓ denotes the level of

the tree, k is the index of folder X_k^ℓ at level ℓ, and $j = 1, \ldots, sub(\ell, k) - 1$. For binary trees the third index is not required, and we recover the standard double index wavelet notation $\psi_{\ell,k}$. Figure 1 illustrates some Haar-like wavelet functions for a simple tree. These functions resemble the classical Haar functions in the following sense: i) Since $W^\ell \subset V^{\ell+1}$, each $\psi_{\ell,k,j}$ is piecewise constant on folders at level $\ell+1$; ii) Since $\psi_{\ell,k,j}$ is supported on the folder X_k^ℓ, it is nonzero only on folders at level $\ell+1$ which are subfolders of X_k^ℓ; iii) Since $W^\ell \perp V^\ell$, each $\psi_{\ell,k,j}$ is orthogonal to the constant function on X_k^ℓ, $\langle \psi_{\ell,k,j}\,,\, \mathbf{1}_{X_k^\ell} \rangle = 0$.

3.3. Function Smoothness, Exponential Coefficient Decay and Learnability

Our main theoretical result is an explicit relation between the following concepts: the *geometry* of the data, as represented by the tree structure; the *smoothness* of a function with respect to the tree; and the *decay or sparsity* of its coefficients in a Haar-like expansion. The practical implications of these theorems to function learnability, e.g., to accurate approximation of f from partially labeled data are considered in section 4.

To relate the geometry of the data to the tree structure, we first define a probability measure ν on the tree as follows: For each set $S \subset X$, define $\nu(S) = |S|/|X|$. Next, consider the following tree ultrametric, also popular for metric approximations (Bartal, 1996; 1998):

Definition 3.1 *Tree Metric: For any $x, y \in X$, define*

$$d(x,y) = \begin{cases} \nu(folder(x,y)) & x \neq y \\ 0 & x = y \end{cases} \qquad (5)$$

where $folder(x,y)$ is the smallest folder in the tree containing both x, y.

Given the tree metric, the following definition of smoothness is a straightforward analogue of Hölder smoothness in the Euclidean setting[1]:

Definition 3.2 *A function f is (C,α)-Hölder w.r.t. the tree (with $0 < \alpha < 1$) if*

$$|f(x) - f(y)| \leq C\, d(x,y)^\alpha, \qquad \forall x, y \in X. \qquad (6)$$

With these definitions, the following theorems relate function smoothness, the geometry of the data and fast coefficient decay:

Theorem 1 *Let $f : X \to \mathbb{R}$ be (C,α)-Hölder and let $\psi_{\ell,k,j}$ be a Haar-like basis function supported on the folder X_k^ℓ. Then*

$$|\langle f\,,\, \psi_{\ell,k,j} \rangle| \leqslant C 2^{\alpha+1} \cdot \nu(X_k^\ell)^{\alpha+1/2}\,. \qquad (7)$$

The tree-balance requirement (2) implies that $\nu(X_k^\ell) \leqslant \overline{B}^{\ell-1}$ for all ℓ, k. Therefore, smoothness of f implies exponential decay rate of its wavelet coefficients as a function of the tree level ℓ,

$$|\langle f\,,\, \psi_{\ell,k,j} \rangle| \leqslant 2^{\alpha+1} C \cdot \overline{B}^{(\ell-1)(\alpha+1/2)}\,. \qquad (8)$$

In particular, for a perfectly balanced binary tree with $\overline{B} = 1/2$ we recover the familiar wavelet coefficient exponential rate of decay for smooth functions, see for example (Mallat, 1999), theorem 6.3.

Theorem 2 *Let $f : X \to \mathbb{R}$. Suppose that for all functions $\psi_{\ell,k,j}$ in some Haar-like basis*

$$|\langle f, \psi_{\ell,k,j} \rangle| \leqslant C \cdot \nu(X_k^\ell)^{\alpha+1/2}$$

with some $\alpha > 0$, and some constant C. Then f is (C',α)-Hölder, with $C' = \frac{2C}{\overline{B}^{3/2}} \frac{1}{1-\overline{B}^\alpha}$.

Finally, the following theorem, which seems new even in the Euclidean setting, shows an interesting relation between L_1 sparsity in a multiscale Haar-like basis, and function approximation:

Theorem 3 *Let $h_I(x)$ be a Haar-like basis, where each function h_I is supported on a set $I \subset X$ and such that $|h_I(x)| \leq 1/|I|^{1/2}$, and let $f = \sum_I a_I h_I(x)$. Assume $\sum_I |a_I| \leq C$, and for any $\epsilon > 0$ consider the approximation $\hat{f} = \sum_{|I|>\epsilon} a_I h_I$. Then*

$$\|f - \hat{f}\|_1 = \sum_{x \in X} |f(x) - \hat{f}(x)| \leq C\sqrt{\epsilon}. \qquad (9)$$

Proofs of these theorems appear in the supplementary material. The key point of these theorems is that the multiscale representation of a weighted graph via a hierarchical tree allows for the development of a theory of harmonic analysis on graphs. Theorems 1-2 provide a clear connection between the notion of function smoothness w.r.t. the tree and fast coefficient decay in a Haar-like expansion. These theorems have well known analogues in the Euclidean setting. Theorem 3 has interesting implications to statistical inference as it shows that under an L_1 bound on the coefficients in a Haar-like expansion, for a reconstruction error of $O(\sqrt{\epsilon})$ it is sufficient to estimate only coefficients corresponding to Haar-like functions with support larger than ϵ, that is, at most $O(1/\epsilon)$ coarse-scale coefficients. Eq. (20) in supplementary material shows that our

[1] For trees constructed on Euclidean spaces, Hölder w.r.t. the tree is not equivalent to Hölder in the Euclidean space. This issue is beyond the scope of this paper.

Haar-like functions satisfy the condition of Theorem 3, $|\psi_{\ell,k,j}(x)| \leq 1/(\underline{B}\nu(X_k^\ell))^{1/2}$.

We emphasize that a balanced tree is a key assumption in our method. Empirically, on many datasets with given affinities, various graph-to-tree constructions indeed resulted in balanced trees. We note that the ability to represent data by a balanced tree is related to some notion of "intrinsic low dimensionality" of the data. Theoretical conditions on graph affinities, on graph-to-tree constructions that provide balanced trees, and on the relations between function smoothness w.r.t. graph metrics and smoothness w.r.t. tree metrics are all subjects for further research.

From a computational view, Haar-like functions are piecewise constant and thus simple to handle. Furthermore, expanding a function or estimating its coefficients in the Haar-like domain admit fast algorithms resembling the fast wavelet transform. One limitation of Haar-like functions is their lack of smoothness. Due to the arbitrary locations of their discontinuities, this may introduce artifacts when processing functions in the coefficient domain. In the context of signal denoising, (Coifman & Donoho, 1995) suggested to remove such artifacts by averaging shifted Haar bases. Similar to their approach and to random forest (Breiman, 2001), we also average functions learned using several different randomly constructed trees.

Finally, from a mathematical viewpoint, the above construction of a probability measure and metric on a tree leads to a particular instance of an abstract object known in harmonic analysis as a *Space of Homogeneous Type* (Coifman & Weiss, 1977; Deng & Han, 2009). To the best of our knowledge, our work presents one of the first applications of these theoretical concepts to practical problems in statistics and machine learning. In section 4 we present an application of these theoretical results to semi-supervised learning.

4. Semi-Supervised Learning

The problem of transducive learning can be phrased in our setting as follows: An unknown target function $f : X \to \mathbb{R}$ is to be inferred, from its (possibly noisy) values $f\big|_S$ on a given subset $S = \{s_1, \ldots, s_n\} \subset X$. For example, when the values of f are class labels, the problem is to classify the remaining points $X \setminus S$, using the observed class labels $f(s_1), \ldots, f(s_n)$.

SSL is an active area of research with many different algorithms suggested over the past few years (Chapelle et al., 2006; Zhu, 2008). Given a basis $\{\phi_1, \ldots, \phi_N\}$ for V, a natural approach is to decompose the unknown function in a series $f = \sum_{i=1}^N \langle f, \phi_i \rangle \phi_i$ and then estimate the unknown coefficients $\langle f, \phi_i \rangle$ using the available values of f. For a dataset X augmented by pairwise affinities, this is the approach suggested by Belkin & Niyogi (2003) with ϕ_i the eigenvectors of the corresponding graph Laplacian. Building on Theorems 1-2, we now propose to use a Haar-like basis that is *designed* by some balanced tree, as an alternative to the Laplacian eigenbasis, that is *generated* by matrix diagonalization. We remark that (Herbster et al., 2009; Kemp et al., 2004; Neal & Zhang, 2006) also suggested SSL algorithms using trees, albeit by different approaches.

Our approach is as follows: Given a hierarchical tree representation of X and a labeled set $S \subset X$, (typically with $|S| \ll |X|$) only relevant Haar-like coefficients are estimated. For a folder X_k^ℓ with no labeled points, or with labeled points only on some of its subfolders, we set the respective wavelet coefficient to zero. Wavelet coefficients are estimated only on sufficiently large folders where labeled data is available in all their subfolders. Theorems 1-2 provide the theoretical support for this approach, provided the target function f is smooth w.r.t. the tree. We emphasize that even though the tree must be balanced (see Eq. (2)), the class labels themselves may be highly unbalanced.

To derive the formula for the estimated coefficients at sufficiently large folders it is instructive to first recall the formula for the actual wavelet coefficient $a_{\ell,k,j}$.

Definition 4.1 *The coefficient $a_{\ell,k,j}$ of $f : X \to \mathbb{R}$ corresponding to the basis function $\psi_{\ell,k,j}$ is given by*

$$\begin{aligned} a_{\ell,k,j} &= \langle f\,,\, \psi_{\ell,k,j} \rangle = \frac{1}{N} \sum_{x \in X_k^\ell} f(x) \psi_{\ell,k,j}(x) \qquad (10) \\ &= \sum_{i \in sub(\ell,k)} \nu(X_i^{\ell+1}) \psi_{\ell,k,j}(X_i^{\ell+1}) m(f, X_i^{\ell+1}) \end{aligned}$$

where $m(f, X_i^{\ell+1}) = \frac{1}{|X_i^{\ell+1}|} \sum_{x \in X_i^{\ell+1}} f(x)$ is the mean of f on the folder $X_i^{\ell+1}$, and $\psi_{\ell,k,j}(X_i^{\ell+1})$ is the constant value of $\psi_{\ell,k,j}$ on the folder $X_i^{\ell+1}$.

Given partial labeled data, Eq (10) suggests the following estimator for $a_{\ell,k,j}$: For folders with at least one empty subfolder, we set $\hat{a}_{\ell,k,j} = 0$. For folders whose subfolders each contain at least one labeled point,

$$\hat{a}_{\ell,k,j} = \sum_{i \in sub(\ell,k)} \nu(X_i^{\ell+1}) \psi_{\ell,k,j}(X_i^{\ell+1}) \hat{m}(f, X_i^{\ell+1}) \quad (11)$$

where $\hat{m}(f, X_i^{\ell+1}) = \frac{1}{|S \cap X_i^{\ell+1}|} \sum_{x \in S \cap X_i^{\ell+1}} f(x)$ is the empirical mean of f on $X_i^{\ell+1}$ using only the labeled data of the set S.

For a regression problem, our estimator for f is

$$\hat{f}(x) = \sum_{\ell,k,j} \hat{a}_{\ell,k,j}\, \psi_{\ell,k,j}(x) \qquad (12)$$

whereas for binary classification we output $sign(\hat{f})$.

Note that conditional on all subfolders of X_k^ℓ having at least one labeled point, $\hat{a}_{\ell,k,j}$ is *unbiased*, $\mathbb{E}[\hat{a}_{\ell,k,j}] = a_{\ell,k,j}$. For small folders there is a non-negligible probability of having empty subfolders, so overall $\hat{a}_{\ell,k,j}$ is biased. However, by Theorem 1, for smooth functions these coefficients are exponentially small in ℓ. The following theorem quantifies the expected L_2 error of both the estimate $\hat{a}_{\ell,k,j}$, and the function estimate $\hat{f}$. Its proof is in the supplementary material.

Theorem 4 *Let f be (C,α) Hölder, and define $C_1 = C2^{\alpha+1}$. Assume that the labeled samples $s_i \in S \subset X$ were randomly chosen from the uniform distribution on X with replacement. Let $\hat{f}$ be the estimator (12) with coefficients estimated via Eq. (11). Up to $o(1/|S|)$ terms, the mean squared error of coefficient estimates is bounded by*

$$\begin{aligned}\mathbb{E}[\hat{a}_{\ell,k,j} - a_{\ell,k,j}]^2 &\lessapprox \frac{1}{|S|}\frac{C_1^2 \overline{B}^{2\alpha} \nu(X_k^\ell)^{2\alpha}}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}} \\ &\quad + \frac{1}{\underline{B}} e^{-|S|\underline{B}\nu(X_k^\ell)} \cdot a_{\ell,k,j}^2 \end{aligned} \qquad (13)$$

The resulting overall MSE is bounded by

$$\begin{aligned}\mathbb{E}\,\|f-\hat{f}\|^2 &= \frac{1}{N}\sum_i (f(x_i)-\hat{f}(x_i))^2 \\ &\le \frac{C_1^2\overline{B}^{2\alpha}}{|S|}\sum_{\ell,k,j}\frac{\overline{B}^{2\alpha(\ell-1)}}{1-e^{-|S|\underline{B}^\ell}} \\ &\quad + \frac{2^{2\alpha+1}C_1^2}{\underline{B}}\sum_{\ell,k,j} e^{-|S|\underline{B}^\ell}(\overline{B}^{2\alpha+1})^{\ell-1}\end{aligned} \qquad (14)$$

The first term in (13) is the estimation error whereas the second term is the approximation error, e.g. the bias-variance decomposition. For sufficiently large folders, with $|S|\underline{B}\nu(X_k^\ell) \gg 1$, the estimation error decays with the number of labeled points as $|S|^{-1}$, and is smaller for smoother functions (larger α). The approximation error, due to folders empty of labeled points, decays exponentially with $|S|$ and with folder size.

The values $\overline{B}$ and $\underline{B}$ can be easily extracted from a given tree. Theorem 4 thus provides a non-parametric risk analysis that depends on a single parameter, the assumed smoothness class α of the target function.

5. Numerical Results

We present preliminary numerical results of our SSL scheme on several datasets. More results and Matlab code appear in supplementary material. We focus on two well-known handwritten digit data sets, MNIST and USPS. These are natural choices due to the inherent multiscale structures present in handwritten digits.

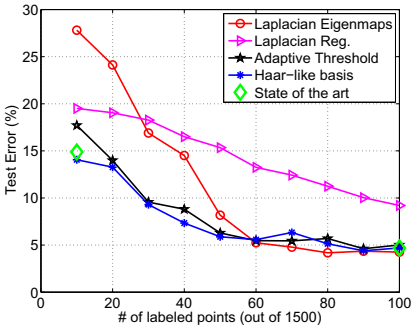

Figure 2. Results on the USPS benchmark.

Given a dataset X of N digits, of which only a small subset S is labeled, we first use all samples in X to construct an affinity matrix $W_{i,j}$ described below. A tree is constructed as follows: At the finest level, $\ell = L$, we have N singleton folders: $X_i^L = \{x_i\}$. Each coarse level is constructed from a finer level as follows: Random (centroid) points are selected s.t. no two are connected by an edge of weight larger than a "radius" parameter. This yields a partition of the current level according to the nearest centroid. The partition elements constitute the points of the coarser level. A coarse affinity matrix is constructed, where the edge weight between two partition elements C and D is $\sum_{i\in C, j\in D} W_{ij}^2$ where W is the affinity matrix of the finer level graph. The motivation for squaring the affinities at each new coarse level is to capture structures at different scales. As the choice of centroids is (pseudo) random, so is the resulting partition tree. With the partition tree at hand, we construct a Haar-like basis induced by the tree and estimate the coefficients of the target label function as described in Section 4.

We compare our method to Laplacian Eigenmaps (Belkin & Niyogi, 2003), with $|S|/5$ eigenfunctions, as suggested by the authors, and to the Laplacian Regularization approach of (Zhu et al., 2003). For the latter, we also consider an *adaptive* threshold for classification ($sign(y > q_{th})$), with q_{th} chosen such that the proportion of test labeled points of each class is equal to its value in the training set[2].

[2]Note that this method is *different* from the class mass normalization approach of (Zhu et al., 2003).

Table 1. Test classification errors for USPS benchmark

METHOD	10 LABELED	100 LABELED
1-NN	19.82	7.64
SVM	20.03	9.75
MVU + 1-NN	14.88	6.09
LEM + 1-NN	19.14	6.09
QC + CMN	**13.61**	6.36
DISCRETE REG.	16.07	**4.68**
TSVM	25.20	9.77
SGT	25.36	6.80
CLUSTER-KERNEL	19.41	9.68
DATA-DEP. REG.	17.96	5.10
LDS	17.57	4.96
LAPLACIAN RLS	18.99	**4.68**
CHM (NORMED)	20.53	7.65
Haar-like	**14.01**	**4.70**

5.1. The USPS benchmark

This benchmark (Chapelle et al., 2006) contains 1500 grey scale 16x16 images of the digits $\{0,\ldots,9\}$. The task is to distinguish the digits $\{2,5\}$ from the rest. The pixels are shuffled and some are missing, so each image is viewed as a vector $x_i \in \mathbb{R}^{241}$. The original benchmark consists of 10 training sets each with 10 labeled digits, and 10 training sets each with 100 labeled digits. The affinity matrix is $W(i,j) = \exp(-\|x_i - x_j\|^2/\varepsilon)$ with $\varepsilon = 30$. Table 1 shows reported results on this benchmark; The bottom row is the test error of our classifier, constructed by averaging over 10 trees, generated using different random seeds. Note that unlike other SSL methods reported in this benchmark, our Haar-like approach achieves competitive results for both few and many labeled points.

For a more careful comparison, we generated 10 random training sets for each of the sizes, $|S| = 20, 30, \ldots, 80, 90$. Fig. 2 shows that for this data set, the Haar-like classifier dominates the Laplacian Eigenfunction when labeled points are few and is comparable when they are many. When just a few labeled points are available, the oscillatory nature of the Laplacian eigenfunctions limits robust estimation, whereas the Haar-like multiscale approach allows us to capitalize on the multiscale structure of the data set.

Further insight into the fundamental difference between Laplacian Eigenmaps (which can be viewed as an extension of Fourier basis) and a Haar-like wavelet basis, corresponding to a tree constructed from the same affinity graph, can be gained by plotting the expansion coefficients of the original label function in these two bases. Fig 4 shows in log-scale an exponential rate of decay of the Haar-like coefficients (consistent with Theorem 1), in contrast to a polynomial rate of decay of Laplacian Eigenfunctions coefficients. This dramatic difference in the efficiency of representation of the target function may explain the difference in classification accuracy.

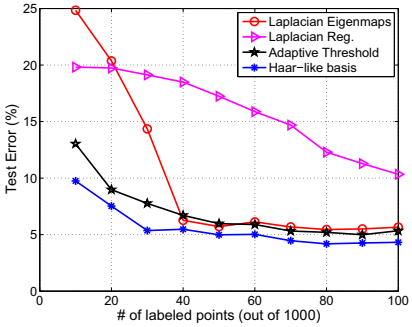

Figure 3. Results on the MNIST subset.

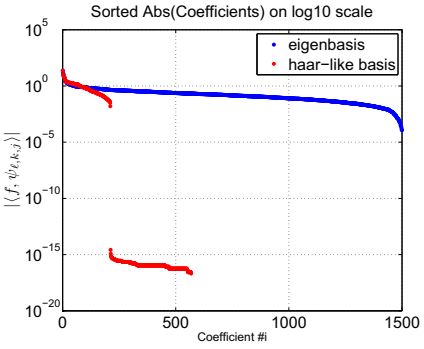

Figure 4. Coefficient decay of target function of USPS benchmark, in the Laplacian eigenbasis and in a Haar-like basis. Note that in the Haar-like basis 930 coefficients out of 1500 are identically zero.

5.2. A subset of the MNIST dataset

A similar comparison was done on small subsets of the MNIST handwritten digits[3]. For each of the digits $\{8,3,4,5,7\}$, 200 samples were selected at random. Digits 8 were labeled as $+1$, the rest as -1. The affinity matrix was $W_{i,j} = \exp(-(1-\rho_{ij})/\sigma_W)$ with $\sigma_W = 0.2$ and ρ_{ij} the maximal correlation between two images, up to a global up/down or left/right shift by one pixel. For each labeled set size $|S| = 10, 20, \ldots, 100$, classification results over 50 random sets were recorded. Fig. 3 shows average test errors, exhibiting a similar phenomena - a clear advantage of the Haar-basis approach for small labeled sets.

[3]available at http://yann.lecun.com/exdb/mnist/

6. Summary and Discussion

Multiscale representations of data and graphs via Haar-like bases have various potential applications, ranging from signal de-noising to density estimation. In this paper we considered their application to semi-supervised learning. Our approach raises many theoretical questions for further research, in particular regarding construction of trees that best capture the geometry of these challenging datasets.

Acknowledgments. The authors thank the anonymous referees for valuable suggestions. BN was supported by Israel Science Foundation grant 432/06. MG is supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship.

References

Bartal, Y. Probabilistic approximation of metric spaces and its algorithmic applications. In *Proc. of the 37th Annual IEEE Symp. on Foundations of Computer Science*, 1996.

Bartal, Y. On approximating arbitrary metrics by tree metrics. In *Proc. of the 30th Annual Symp. on Theory of Computing*, 1998.

Belkin, M. and Niyogi, P. Using manifold structure for partially labelled classification. In *Advances in Neural Information Processing Systems, Vol. 13*, 2003.

Binev, P., Cohen, A., Dahmen, W., DeVore, R., and Temlyakov, V. Universal algorithms for learning theory, part I: Piecewise constant functions. *JMLR*, pp. 1297–1321, 2005.

Breiman, L. Random forests. *Machine Learning*, 45: 5–32, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and regression trees*. Wadsworth International, Belmont, CA, 1984.

Chapelle, O., Schölkopf, B., and Zien, A. *Semi-Supervised Learning*. MIT Press, 2006.

Coifman, R.R. and Donoho, D.L. Translation-invariant de-noising. In *Wavelets and Statistics*, pp. 125–150, NY, 1995. Springer-Verlag.

Coifman, R.R. and Maggioni, M. Diffusion wavelets. *Appl. Comp. Harm. Anal.*, 21(1):53–94, 2006.

Coifman, R.R. and Weiss, G. Extensions of Hardy spaces and their use in analysis. *Bulletin of the AMS*, 83(4), 1977.

Deng, D. and Han, Y. *Harmonic Analysis on Spaces of Homogeneous Type*. Springer, Berlin, 2009.

Donoho, D.L. CART and best-ortho-basis: a connection. *Annals of Statistics*, 25:1870–1911, 1997.

Hardle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A.B. *Wavelets, Approximation and Statistical Applications*. Springer, NY, 1998.

Herbster, M., Pontil, M., and Rojas-Galeano, S. Fast prediction on a tree. In *Advances in Neural Information Processing Systems, Vol. 21*, 2009.

Jansen, M., Nason, G.P., and Silverman, B.W. Multiscale methods for data on graphs and irregular multidimensional situations. *J. Royal Stat. Soc. B*, 71: 97–125, 2009.

Kemp, C.C., Griffiths, T.L., Stromsten, S., and Tenenbaum, J.B. Semi-supervised learning with trees. In *Advances in Neural Information Processing Systems, Vol. 14*. MIT Press, 2004.

Lee, A.B., Nadler, B., and Wasserman, L. Treelets: an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, 2(2):437–471, 2008.

Lehmann, E.L. and Casella, G. *Theory of Point Estimation*. Springer, 2nd edition, 2003.

Mahadevan, S. and Maggioni, M. Value function approximation with diffusion wavelets and Laplacian eigenfunctions. In *Advances in Neural Information Processing Systems, Vol. 18*, 2006.

Mallat, S. *A wavelet tour of signal processing*. Academic Press, 2nd edition, 1999.

Murtagh, F. The Haar wavelet transform of a dendrogram. *J. Classification*, 24:3–32, 2007.

Nadler, B., Srebro, N., and Zhou, X. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabeled data. In *Advances in Neural Information Processing Systems, Vol. 21*, 2009.

Neal, R.M. and Zhang, J. High dimensional classification with Bayesian neural networks and Dirichlet diffusion trees. In *Feature Extraction: Foundations and Applications*, pp. 265–295. 2006.

Vapnik, V.N. *Statistical Learning Theory*. Wiley, NY, 1998.

Zhu, X. Semi-supervised learning literature review. Technical report, Computer Science Department, University of Wisconsin, 2008.

Zhu, X., Ghahramani, Z., and Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning, Vol. 13*, 2003.

A. Supplementary Material

In this section we present proofs of the various theorems in the paper. Recall that given a dataset X and its representation by a hierarchical tree, Eq. (5) defined a tree metric $d(x,y)$, whereas Eq. (6) defined (C,α)-Hölder smooth functions with respect to the tree metric. Let $f: X \to \mathbb{R}$. For any subset $Y \subset X$ we denote the mean and variance of f on Y as follows,

$$m(f,Y) = \frac{1}{|Y|}\sum_{x\in Y} f(x) \tag{16}$$

$$\sigma^2(f,Y) = \frac{1}{|Y|}\sum_{x\in Y}(f(x)-m(f,Y))^2. \tag{17}$$

Next, given the tree metric we denote by $B(x,r)$ the ball of radius r around x, that is

$$B(x,r) = \{y \in X \,|\, d(x,y) \le r\}$$

Observe that by definition, these balls are exactly the different folders of the tree that contain the node x.

The following lemma, standard in the theory of spaces of homogeneous type, will be useful in our proofs.

Lemma 1 *For any $x \in X$, $s > 0$ and $r > 0$ we have*

$$\int_{B(x,r)} d(x,y)^s d\nu(y) = \frac{1}{|X|}\sum_{y\in B(x,r)} d(x,y)^s \le C_s r^{s+1} \tag{18}$$

with $C_s = 2^{s+1}\left(1-\frac{1}{2}\underline{B}\right) \le 2^{s+1}$.

Proof: Recall that by the definition of the tree metric, $d(x,y) \leqslant 1$, for any $x,y \in X$. Let $K \in \mathbb{N}$ be such that $2^{-K-1} < r \leqslant 2^{-K}$. Then

$$B(x,r) \subset \biguplus_{k=K}^{\infty}\left[B\left(x,2^{-k}\right) \setminus B\left(x,2^{-(k+1)}\right)\right]$$

Hence

$$\begin{aligned}
\int_{B(x,r)} d(x,y)^s d\nu(y) &\leqslant \sum_{k=K}^{\infty} \int_{B(x,2^{-k})\setminus B\left(x,2^{-(k+1)}\right)} d(x,y)^s d\nu(y) \\
&\leqslant \sum_{k=K}^{\infty} \int_{B(x,2^{-k})\setminus B\left(x,2^{-(k+1)}\right)} 2^{-ks} d\nu(y) \\
&\leqslant \sum_{k=K}^{\infty} 2^{-ks}\cdot \nu\left(B\left(x,2^{-k}\right)\setminus B\left(x,2^{-(k+1)}\right)\right) \\
&\leqslant \sum_{k=K}^{\infty}\left[2^{-ks}\left(2^{-k}-\underline{B}\cdot 2^{-(k+1)}\right)\right],
\end{aligned}$$

where the last inequality follows from the tree balance condition, Eq. (2). This gives

$$\begin{aligned}
\int_{B(x,r)} d(x,y)^s d\nu(y) &\leqslant \left(1-\frac{1}{2}\cdot\underline{B}\right)\cdot\sum_{k=K}^{\infty}\left(\frac{1}{2^{s+1}}\right)^k \\
&\leqslant 2^{s+1}\left(1-\frac{1}{2}\cdot\underline{B}\right)\left(2^{-K}\right)^{s+1} \leqslant 2^{s+1}\left(1-\frac{1}{2}\cdot\underline{B}\right) r^{s+1}.
\end{aligned}$$

□.

Before proving theorem 1, we first introduce an alternative definition of function smoothness:

Definition A.1 *A function $f : X \to \mathbb{R}$ is (C, α)-Mean Hölder (w.r.t. the tree metric d) if for all $x \in X$ and any ball $B(x, r)$,*

$$\sigma\left(f, B(x,r)\right) \leqslant C \cdot \nu\left(B(x,r)\right)^{\alpha} . \tag{19}$$

where $\sigma(f, B(x, r))$ is defined in Eq. (17).

The following lemma shows that the two definitions of function smoothness w.r.t. the tree metric are related.

Lemma 2 *Let $f : X \to \mathbb{R}$ be (C, α)-Hölder with respect to the tree. Then f is $(2^{\alpha+1}C, \alpha)$ mean-Hölder.*

Proof of Lemma: Let $x \in X$ and let B be any ball around x. Since X is finite, for any $\varepsilon \geqslant 0$ small enough, we have $B = B(x, r)$ for $r = \nu(B) + \varepsilon$. Now,

$$\begin{aligned}
\int_B \left(f(x) - m(f,B)\right)^2 d\nu(x) &= \int_B \left(f(x) - \frac{1}{\nu(B)} \int_B f(y)\, d\nu(y)\right)^2 d\nu(x) \\
&= \frac{1}{\nu^2(B)} \int_B \left(\int_B f(x) - f(y)\, d\nu(y)\right)^2 d\nu(x) \leqslant \\
&\leqslant \frac{1}{\nu^2(B)} \int_B \left(\int_B |f(x) - f(y)|\, d\nu(y)\right)^2 d\nu(x) .
\end{aligned}$$

As $f : X \to \mathbb{R}$ is (C, α)-Hölder, this gives

$$\int_B \left(f(x) - m(f\,,\,B)\right)^2 d\nu(x) \leqslant \left(\frac{C}{\nu(B)}\right)^2 \int_B \left(\int_B d(x,y)^{\alpha}\, d\nu(y)\right)^2 d\nu(x) .$$

We now substitute $s = \alpha$ in Lemma 1 to obtain

$$\begin{aligned}
\int_B \left(f(x) - m(f\,,\,B)\right)^2 d\nu(x) &\leqslant \left(\frac{C}{\nu(B)}\right)^2 \int_B \left(2^{\alpha+1} r^{\alpha+1}\right)^2 d\nu(x) \\
&\leqslant \left(\frac{2^{\alpha+1}C}{\nu(B)}\right)^2 \nu(B)\, r^{2\alpha+2} \\
&\leqslant \left(\frac{2^{\alpha+1}C}{\nu(B)}\right)^2 \nu(B)\, (\nu(B) + \varepsilon)^{2\alpha+2} .
\end{aligned}$$

Since ε can be arbitrarily small, we conclude that

$$\int_B \left(f(x) - m(f\,,\,B)\right)^2 d\nu(x) \leqslant \left(\frac{2^{\alpha+1}C}{\nu(B)}\right)^2 \nu(B)^{2\alpha+3} = \left(2^{\alpha+1}C\right)^2 \nu(B)^{2\alpha+1}$$

and therefore

$$\sigma\left(f,\, B\right) = \sqrt{\frac{1}{\nu(B)} \int_B \left(f(x) - m(f\,,\,B)\right)^2 \nu(x)} \leqslant C 2^{\alpha+1} \nu(B)^{\alpha+1/2} . \tag{20}$$

Since $\nu(B) \leqslant 1$, the theorem follows. □.

Proof of Theorem 1: Recall that by definition, each Haar-like basis function $\psi_{\ell,k,j}$ is supported on the folder X_k^{ℓ}. It also has zero mean, namely $\int_{X_k^{\ell}} \psi_{\ell,k,j}(x)\, d\nu(x) = 0$, and unit norm, namely $\int_{X_k^{\ell}} \psi_{\ell,k,j}^2(x)\, d\nu(x) = 1$. Therefore,

$$\langle f\,,\, \psi_{\ell,k,j} \rangle = \int_{X_k^{\ell}} f(x) \psi_{\ell,k,j}(x)\, d\nu(x) = \int_{X_k^{\ell}} \left(f(x) - m(f, X_k^{\ell})\right) \psi_{\ell,k,j}(x)\, d\nu(x) .$$

The Cauchy–Schwartz inequality now yields

$$\begin{aligned}
|\langle f\,,\, \psi_{\ell,k,j} \rangle| &\leqslant \sqrt{\int_{X_k^{\ell}} \left(f(x) - m\left(f\,,\, X_k^{\ell}\right)\right)^2 d\nu(x)} \cdot \sqrt{\int_{X_k^{\ell}} \left(\psi_{\ell,k,j}(x)\right)^2 d\nu(x)} \\
&= \sigma\left(f, X_k^{\ell}\right) .
\end{aligned}$$

According to Lemma 2, if f is (C,α) Hölder, it is $(C2^{\alpha+1},\alpha)$ mean-Hölder. In particular, Eq. (20) implies that

$$|\langle f\,\psi_{\ell,k,j}\rangle| \leqslant C2^{\alpha+1}\cdot\nu\left(X_k^\ell\right)^{\alpha+\frac{1}{2}}\ .$$

□.

Proof of Theorem 2: Let $x,y\in X$ and let κ and λ be such that $folder(x,y)=X_\kappa^\lambda$. Our aim is to show that $|f(x)-f(y)|\leqslant C'\cdot\nu(X_\kappa^\lambda)^\alpha$ with C' given by Eq. (9).

To this end, we use the decomposition

$$f(x)=\sum_{\ell,k,j}\langle f,\psi_{\ell,k,j}\rangle\psi_{\ell,k,j}(x).$$

Note that by definition, for any coarse level $\ell<\lambda$ the samples x,y belong to the same folders, and thus $\psi_{\ell,k,j}(x)=\psi_{\ell,k,j}(y)$ for any k,j. Hence, the only terms contributing to the difference $f(x)-f(y)$ are those in the finer folders at levels $\ell=\lambda,\ldots,L$, where x,y belong to *different* folders. That is,

$$\begin{aligned}f(x)-f(y) &= \sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,x))}\left\langle f,\psi_{\ell,\tau(\ell,x),j}\right\rangle\cdot\psi_{\ell,\tau(\ell,x),j}(x)\\ &\quad-\sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,y))}\left\langle f,\psi_{\ell,\tau(\ell,y),j}\right\rangle\cdot\psi_{\ell,\tau(\ell,y),j}(y)\end{aligned}$$

where $\tau(\ell,x)$ is the folder at level ℓ that contains x, $x\in X_{\tau(\ell,x)}^\ell$. Next, recall that by definition the functions $\psi_{\ell,k,j}$ are all normalized, and they are constant on all subfolders of X_k^ℓ. Thus,

$$\|\psi_{\ell,k,j}\|^2=\sum_{i\in sub(\ell,k)}\nu(X_i^{\ell+1})\psi_{\ell,k}^2(X_i^{\ell+1})=1$$

and so

$$|\psi_{\ell,k,j}(x)|\leq\frac{1}{\sqrt{\nu(X_i^{\ell+1})}}\leq\frac{1}{\sqrt{\underline{B}\,\nu(X_k^\ell)}}\,. \tag{21}$$

Combining the bound on $|\psi_{\ell,k,j}|$ with the bound on the coefficient decay of f gives that

$$\begin{aligned}|f(x)-f(y)| &\leqslant \frac{C}{\sqrt{\underline{B}}}\sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,x))}\nu(X_{\tau(\ell,x)}^\ell)^{\alpha+1/2}\frac{1}{\sqrt{\nu(X_{\tau(\ell,x)}^\ell)}}\\ &+ \frac{C}{\sqrt{\underline{B}}}\sum_{\ell=\lambda}^{L}\sum_{j\in sub(\ell,\tau(\ell,y))}\nu(X_{\tau(\ell,y)}^\ell)^{\alpha+1/2}\frac{1}{\sqrt{\nu(X_{\tau(\ell,y)}^\ell)}}\end{aligned} \tag{22}$$

Finally, since the tree is balanced, $\nu(X_{\tau(\ell,x)}^\ell)\leq\overline{B}^{\ell-\lambda}\nu(X_\kappa^\lambda)$, and $|sub(\ell,k)|\leq\frac{1}{\underline{B}}-1$. Thus,

$$\begin{aligned}|f(x)-f(y)| &\leqslant \frac{2C(1-\underline{B})}{\underline{B}^{3/2}}\sum_{\ell=\lambda}^{L}\left(\overline{B}^\alpha\right)^{\ell-\lambda}\nu(X_\kappa^\lambda)^\alpha\\ &\leqslant \frac{2C}{\underline{B}^{3/2}}\frac{1}{1-\overline{B}^\alpha}\nu(X_\kappa^\lambda)^\alpha=C'\nu(X_\kappa^\lambda)^\alpha\,.\end{aligned}$$

□.

Proof of Theorem 3: Let $\hat{f}=\sum_{|I|>\epsilon}a_Ih_I(x)$. Then

$$\begin{aligned}\|f-\hat{f}\|_1 &= \sum_x|f(x)-\hat{f}(x)|=\sum_x\left|\sum_{|I|<\epsilon}a_Ih_I(x)\right|\\ &\leqslant \sum_{|I|<\epsilon}|a_I|\sum_{x\in I}|h_I(x)|\end{aligned} \tag{23}$$

but according to the assumptions of the theorem, $|h_I(x)| \leqslant 1/|I|^{1/2}$ and $supp(h_I) = |I|$. Hence, $\sum_{x\in I} |h_I(x)| < \epsilon/\sqrt{\epsilon} = \sqrt{\epsilon}$. Combining this with the entropy condition on the coefficients, $\sum_I |a_I| \leqslant C$ the theorem follows. □

Proof of Theorem 4: Recall that the coefficient $\hat{a}_{\ell,k,j}$ is given by Eq. (12) if all subfolders of X_k^ℓ at level $\ell+1$ each contain at least one labeled point. Otherwise, $\hat{a}_{\ell,k,j}$ is set to zero. Denote by R the event that at least one of the subfolders of X_k^ℓ does not contain labeled points. First of all,

$$\begin{aligned}\Pr[R] &\leqslant \sum_{i\in sub(\ell,k)} \Pr\left[\,|S\cap X_i^{\ell+1}|=0\right] = \sum_{i\in sub(\ell,k)} (1-\nu(X_i^{\ell+1}))^{|S|} \\ &\leqslant \sum_{i\in sub(\ell,k)} e^{-|S|\nu(X_i^{\ell+1})} \leqslant \frac{1}{\underline{B}} e^{-|S|\underline{B}\nu(X_k^\ell)}\end{aligned}$$

Conditional on the event R, we have $\mathbb{E}[\hat{a}_{\ell,k,j}] = var[\hat{a}_{\ell,k,j}] = 0$, whereas under R^c, we have that $\mathbb{E}[\hat{a}_{\ell,k,j}] = a_{\ell,k,j}$, and after some algebraic manipulations,

$$var[\hat{a}_{\ell,k,j}\,|\,R^c] = \sum_{i\in sub(\ell,k)} \nu^2(X_i^{\ell+1})\psi_{\ell,k,j}^2(X_i^{\ell+1})\frac{\sigma^2(f,X_i^{\ell+1})}{|S\cap X_i^{\ell+1}|} \tag{24}$$

To compute the mean squared error of the estimator $\hat{a}_{\ell,k,j}$ we use the identity

$$\mathbb{E}\left[\hat{a}_{\ell,k,j}-a_{\ell,k,j}\right]^2 = var[\hat{a}_{\ell,k,j}] + \left(\mathbb{E}[\hat{a}_{\ell,k,j}]-a_{\ell,k,j}\right)^2 . \tag{25}$$

Regarding the second term in (25), we have that $\mathbb{E}[\hat{a}_{\ell,k,j}] = a_{\ell,k,j}\,(1-\Pr[R])$. Thus,

$$\left(\mathbb{E}[\hat{a}_{\ell,k,j}]-a_{\ell,k,j}\right)^2 = a_{\ell,k,j}^2\Pr[R]^2 . \tag{26}$$

As for the first term in (25), let Z be the random variable defined as the indicator function of the event R, $Z=\mathbf{1}_R$. By the variance decomposition formula

$$var\left[\hat{a}_{\ell,k,j}\right] = \mathbb{E}\left[var\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] + var\left[\mathbb{E}\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] \tag{27}$$

Now, by (24),

$$\mathbb{E}\left[var[\hat{a}_{\ell,k,j}\middle|Z]\right] = \Pr[R^c]\sum_{i\in sub(\ell,k)} \nu^2(X_i^{\ell+1})\psi_{\ell,k,j}^2(X_i^{\ell+1})\sigma^2(f,X_i^{\ell+1})\mathbb{E}\left[\frac{1}{|S\cap X_i^{\ell+1}|}\,\middle|\,R^c\right]$$

For $|S|\gg 1$, we approximate the conditioning on R^c by the (simpler) conditioning on $\{|S\cap X_i^{\ell+1}|>0\}$. This gives

$$\mathbb{E}\left[var[\hat{a}_{\ell,k,j}\middle|Z]\right] = \Pr[R^c]\sum_{i\in sub(\ell,k)} \nu^2(X_i^{\ell+1})\psi_{\ell,k,j}^2(X_i^{\ell+1})\sigma^2\left(f,X_i^{\ell+1}\right)\frac{\mathbb{E}\left[A_i\right]}{\Pr[\,|S\cap X_i^{\ell+1}|>0\,]} \tag{28}$$

where

$$A_i = \begin{cases} \frac{1}{|S\cap X_i^{\ell+1}|} & |S\cap X_i^{\ell+1}|>0 \\ 0 & |S\cap X_i^{\ell+1}|=0 \end{cases} .$$

The quantity $\mathbb{E}\left[A_i\right]$ is known as the first *inverse moment* of the Binomial distribution $Bin\left(|S|\,,\,\nu(X_i^{\ell+1})\right)$. Asymptotic expansions of this quantity have been studied extensively. In Rempala (2003), it was proved that

$$\mathbb{E}\left[A_i\right] = \frac{1}{|S|\cdot\nu(X_i^{\ell+1})} + o\left(\frac{1}{|S|}\right) .$$

Using this approximation in (28) gives, up to an $o(1/|S|)$ error

$$\mathbb{E}\left[var\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] \approx \frac{\Pr[R^c]}{|S|}\sum_{i\in sub(\ell,k)} \nu(X_i^{\ell+1})\psi_{\ell,k,j}^2(X_i^{\ell+1})\frac{\sigma^2\left(f,X_i^{\ell+1}\right)}{\Pr[\,|S\cap X_i^{\ell+1}|>0]} .$$

As f is (C,α)-Hölder, according to Lemma 2 it is (C_1,α) mean-Hölder with $C_1 = 2^{\alpha+1}C$. Thus, $\sigma^2(f, X_i^{\ell+1}) \leqslant C_1^2 \nu(X_i^{\ell+1})^{2\alpha}$. Since the tree is balanced, $\nu(X_i^{\ell+1}) \leq \overline{B}\nu(X_k^\ell)$. In addition,

$$\frac{1}{\Pr[\,|S \cap X_i^{\ell+1}| > 0]} \leqslant \frac{1}{1-e^{-|S|\nu(X_i^{\ell+1})}} \leqslant \frac{1}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[var\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] &\leqslant \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}}\sum_{i\in sub(\ell,k)}\nu(X_i^{\ell+1})\psi_{\ell,k,j}^2(X_i^{\ell+1}) \\ &= \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}}\,. \end{aligned} \tag{29}$$

where the summation is simply $\|\psi_{\ell,k,j}\|^2 = 1$.

For the second term in Eq. (27), note that

$$\mathbb{E}\left[\hat{a}_{\ell,k,j}\,|\,Z\right] = \begin{cases} a_{\ell,k,j} & \text{under } R^c \\ 0 & \text{under } R \end{cases} \tag{30}$$

Therefore,

$$var\left[\mathbb{E}\left[\hat{a}_{\ell,k,j}\middle|Z\right]\right] = a_{\ell,k,j}^2\,(1-\Pr[R])\Pr[R]. \tag{31}$$

Combining (29), (31) into (25) gives that

$$\begin{aligned}\mathbb{E}[\hat{a}_{\ell,k,j}-a_{\ell,k,j}]^2 &\leqslant \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}} + a_{\ell,k,j}^2\,(1-\Pr[R])\Pr[R] + a_{\ell,k,j}^2\Pr[R]^2 \\ &\leqslant \frac{1}{|S|}\frac{C_1^2\overline{B}^{2\alpha}\nu^{2\alpha}(X_k^\ell)}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}} + \frac{1}{\underline{B}}e^{-|S|\underline{B}\nu(X_k^\ell)}\cdot a_{\ell,k,j}^2\,. \end{aligned} \tag{32}$$

Finally, to prove the formula for the mean squared error in estimating f we note that due to the orthogonality of the Haar-like basis functions,

$$\begin{aligned}\mathbb{E}\left\|f-\hat{f}\right\|^2 &= \mathbb{E}\left[\left\|\sum_{\ell,k,j}(a_{\ell,k,j}-\hat{a}_{\ell,k,j})\,\psi_{\ell,k,j}\right\|^2\right] = \mathbb{E}\left[\sum_{\ell,k,j}(a_{\ell,k,j}-\hat{a}_{\ell,k,j})^2\right] \\ &= \sum_{\ell,k,j}\mathbb{E}\left[a_{\ell,k,j}-\hat{a}_{\ell,k,j}\right]^2.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\,\|f-\hat{f}\|^2 &\leqslant \frac{C_1^2\overline{B}^{2\alpha}}{|S|}\sum_{\ell,k,j}\frac{\nu(X_k^\ell)^{2\alpha}}{1-e^{-|S|\underline{B}\nu(X_k^\ell)}} + \frac{1}{\underline{B}}\sum_{\ell,k,j}e^{-|S|\underline{B}\nu(X_k^\ell)}a_{\ell,k,j}^2 \\ &\leqslant \frac{C_1^2\overline{B}^{2\alpha}}{|S|}\sum_{\ell,k,j}\frac{\left(\overline{B}^{2\alpha}\right)^{\ell-1}}{1-e^{-|S|\underline{B}^\ell}} + \frac{2^{2\alpha+1}C_1^2}{\underline{B}}\sum_{\ell,k,j}e^{-|S|\underline{B}^\ell}\left(\overline{B}^{2\alpha+1}\right)^{\ell-1} \end{aligned} \tag{33}$$

□