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 $(C, \alpha)$-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,

$$m(f, Y) = \frac{1}{|Y|} \sum_{x \in Y} f(x)$$
$$\sigma^2(f, Y) = \frac{1}{|Y|} \sum_{x \in Y} (f(x) - m(f, Y))^2. \quad (16)(17)$$

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

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

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

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

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

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

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

Hence

$$\int_{B(x, r)} d(x, y)^s \, dv(y) \leq \sum_{k=K}^{\infty} \int_{B(x, 2^{-k}) \setminus B(x, 2^{-(k+1)})} d(x, y)^s \, dv(y)$$

$$\leq \sum_{k=K}^{\infty} 2^{-ks} \, dv(y)$$

$$\leq \sum_{k=K}^{\infty} 2^{-ks} \cdot \nu \left( B \left( x, 2^{-k} \right) \setminus B \left( x, 2^{-(k+1)} \right) \right)$$

$$\leq \sum_{k=K}^{\infty} \left[ 2^{-ks} \left( 2^{-k} - 2^{-k} \cdot 2^{-(k+1)} \right) \right],$$

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

$$\int_{B(x, r)} d(x, y)^s \, dv(y) \leq \left( 1 - \frac{1}{2} \cdot B \right) \cdot \sum_{k=K}^{\infty} \left( \frac{1}{2^{s+1}} \right)^k$$

$$\leq 2^{s+1} \left( 1 - \frac{1}{2} \cdot B \right) \left( 2^{-K} \right)^{s+1} \leq 2^{s+1} \left( 1 - \frac{1}{2} \cdot B \right) r^{s+1}. \square$$

Before proving theorem 1, we first introduce an alternative definition of function smoothness:
Therefore, the Cauchy–Schwartz inequality now yields

\[\sigma(f, B(x, r)) \leq C \cdot \nu(B(x, r))^{\alpha}. \tag{19}\]

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 \((C, \alpha)\)-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 \(\varepsilon \geq 0\) small enough, we have \(B = B(x, r)\) for \(r = \nu(B) + \varepsilon\). Now,

\[
\int_B (f(x) - m(f, B))^2 \, d\nu(x) = \int_B \left( f(x) - \frac{1}{\nu(B)} \int_B f(y) \, d\nu(y) \right)^2 \, d\nu(x)
\]

\[
= \frac{1}{\nu^2(B)} \int_B \left( \int_B f(x) - f(y) \, d\nu(y) \right)^2 \, d\nu(x)
\]

\[
\leq \frac{1}{\nu^2(B)} \int_B \left( \int_B |f(x) - f(y)| \, d\nu(y) \right)^2 \, d\nu(x).
\]

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

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

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

\[
\int_B (f(x) - m(f, B))^2 \, d\nu(x) \leq \left( \frac{C}{\nu(B)} \right)^2 \int_B (2^{\alpha+1}r^{\alpha+1})^2 \, d\nu(x)
\]

\[
\leq \left( \frac{2^{\alpha+1}C}{\nu(B)} \right)^2 \nu(B) \cdot 2^{2\alpha+2}
\]

\[
\leq \left( \frac{2^{\alpha+1}C}{\nu(B)} \right)^2 \nu(B) \cdot (\nu(B) + \varepsilon)^{2\alpha+2}.
\]

Since \(\varepsilon\) can be arbitrarily small, we conclude that

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

and therefore

\[
\sigma(f, B) = \frac{1}{\nu(B)} \int_B (f(x) - m(f, B))^2 \, d\nu(x) \leq C 2^{\alpha+1}\nu(B)^{\alpha+1/2}. \tag{20}
\]

Since \(\nu(B) \leq 1\), the theorem follows. \(\square\)

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

\[
\langle f, \psi_{\ell,k,j} \rangle = \int_{X^k_{\ell}} f(x) \psi_{\ell,k,j}(x) \, d\nu(x) = \int_{X^k_{\ell}} (f(x) - m(f, X^k_{\ell})) \psi_{\ell,k,j}(x) \, d\nu(x).
\]

The Cauchy–Schwartz inequality now yields

\[
|\langle f, \psi_{\ell,k,j} \rangle| \leq \sqrt{\int_{X^k_{\ell}} (f(x) - m(f, X^k_{\ell}))^2 \, d\nu(x)} \cdot \sqrt{\int_{X^k_{\ell}} (\psi_{\ell,k,j}(x))^2 \, d\nu(x)}
\]

\[
= \sigma(f, X^k_{\ell}).
\]
According to Lemma 2, if \( f \) is \((C, \alpha)\) Hölder, it is \((C2^{\alpha+1}, \alpha)\) mean-Hölder. In particular, Eq. (20) implies that
\[
|\langle f, \psi_{\ell,k,j}\rangle| \leq C2^{\alpha+1} \cdot \nu \left( \frac{X_{\lambda}^\ell}{C} \right)^{\alpha + \frac{1}{2}}.
\]
\[Q.E.D.

**Proof of Theorem 2:** Let \( x, y \in X \) and let \( \kappa \) and \( \lambda \) be such that \( \text{folder}(x,y) = X^\lambda_\kappa \). Our aim is to show that
\[
|f(x) - f(y)| \leq C' \cdot \nu \left( X^\lambda_\kappa \right)^\alpha
\]
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,
\[
|\langle f, \psi_{\ell,k,j} \rangle| \leq \frac{1}{\sqrt{\nu(X_{\lambda}^\ell)}} \leq \frac{1}{\sqrt{B \nu(X_{\lambda}^\ell)}}.
\]
and so
\[
|\psi_{\ell,k,j}(x)| \leq \frac{1}{\sqrt{\nu(X_{\lambda}^\ell)}} \leq \frac{1}{\sqrt{B \nu(X_{\lambda}^\ell)}}.
\]
Combining the bound on \( |\psi_{\ell,k,j}| \) with the bound on the coefficient decay of \( f \) gives that
\[
|f(x) - f(y)| \leq \frac{C}{\sqrt{B}} \sum_{\ell=\lambda}^{L} \sum_{j \in \text{sub}(\ell, \tau(\ell, x))} \nu(X_{\tau(\ell, x)}^\ell)^{\alpha + 1/2} \left( \frac{1}{\sqrt{\nu(X_{\lambda}^\ell)}} \right)
\]
and
\[
|f(x) - f(y)| \leq \frac{C}{\sqrt{B}} \sum_{\ell=\lambda}^{L} \sum_{j \in \text{sub}(\ell, \tau(\ell, y))} \nu(X_{\tau(\ell, y)}^\ell)^{\alpha + 1/2} \left( \frac{1}{\sqrt{\nu(X_{\lambda}^\ell)}} \right)
\]
Finally, since the tree is balanced, \( \nu(X_{\tau(\ell, x)}^\ell) \leq B^{1-\lambda} \nu(X_{\lambda}^\kappa) \), and \( |\text{sub}(\ell, k)| \leq \frac{1}{B} - 1 \). Thus,
\[
|f(x) - f(y)| \leq \frac{2C(1-B)}{B^{3/2}} \sum_{\ell=\lambda}^{L} \left( \frac{B^{\alpha}}{\nu(X_{\lambda}^\kappa)} \right)^{\ell - \lambda} \nu(X_{\lambda}^\kappa)^{\alpha}
\]
\[
< \frac{2C}{B^{3/2}} \frac{1}{1-B} \nu(X_{\lambda}^\kappa)^{\alpha} = C' \nu(X_{\lambda}^\kappa)^{\alpha}.
\]
\[Q.E.D.

**Proof of Theorem 3:** Let \( \tilde{f} = \sum_{|I| > \epsilon} a_I h_I(x) \). Then
\[
\|f - \tilde{f}\|_1 = \sum_x |f(x) - \tilde{f}(x)| = \sum_x \left( \sum_{|I| > \epsilon} a_I h_I(x) \right) \leq \sum_{|I| < \epsilon} |a_I| \sum_{x \in I} |h_I(x)|
\]
\[
\leq \sum_{|I| < \epsilon} |a_I| \sum_{x \in I} |h_I(x)|
\]
\[
\leq \sum_{|I| < \epsilon} |a_I| \sum_{x \in I} |h_I(x)|
\]
but according to the assumptions of the theorem, \( |h_I(x)| \leq 1/|I|^{1/2} \) and \( \text{supp}(h_I) = |I| \). Hence, \( \sum_{x \in I} |h_I(x)| < \epsilon/\sqrt{r} = \sqrt{r} \). Combining this with the entropy condition on the coefficients, \( \sum_I |a_I| \leq C \) the theorem follows.

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

\[
\Pr[R] \leq \sum_{i \in \text{sub}(\ell,k)} \Pr[|S \cap X^\ell_i| = 0] = \sum_{i \in \text{sub}(\ell,k)} (1 - \nu(X^\ell_i))^{|S|}
\]

\[
\leq \sum_{i \in \text{sub}(\ell,k)} e^{-|S\nu(X^\ell_i)} \leq \frac{1}{B} e^{-|S|B\nu(X^\ell_i)}
\]

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

\[
\text{var}[a_{\ell,k,j} | R^c] = \sum_{i \in \text{sub}(\ell,k)} \nu^2(X^\ell_i)\psi^2_{\ell,k,j}(X^\ell_i)|S|\] \( \mathbb{E}[a_{\ell,k,j}] \) \( \sqrt{|S\cap X^\ell_i|} \)

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

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

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

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

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

\[
\text{var}[a_{\ell,k,j}] = \mathbb{E}[\text{var} \left[ \hat{a}_{\ell,k,j} \bigg| Z \right]] + \text{var} \left[ \mathbb{E} \left[ \hat{a}_{\ell,k,j} \bigg| Z \right] \right]
\]

Now, by (24),

\[
\mathbb{E}[\text{var} \left[ \hat{a}_{\ell,k,j} \bigg| Z \right]] = \Pr[R^c] \sum_{i \in \text{sub}(\ell,k)} \nu^2(X^\ell_i)\psi^2_{\ell,k,j}(X^\ell_i)|S|\] \( \mathbb{E}[a_{\ell,k,j}] \) \( \sqrt{|S\cap X^\ell_i|} \)

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

\[
\mathbb{E}[\text{var} \left[ \hat{a}_{\ell,k,j} \bigg| Z \right]] \approx \Pr[R^c] \sum_{i \in \text{sub}(\ell,k)} \nu^2(X^\ell_i)\psi^2_{\ell,k,j}(X^\ell_i)|S|\] \( \mathbb{E}[a_{\ell,k,j}] \) \( \frac{1}{|S\cap X^\ell_i|} \)

where

\[
A_i = \begin{cases} \frac{1}{|S\cap X^\ell_i|} & |S \cap X^\ell_i| > 0 \\ 0 & |S \cap X^\ell_i| = 0 \end{cases}
\]

The quantity \( \mathbb{E}[A_i] \) is known as the first inverse moment of the Binomial distribution \( \text{Bin}(|S|, \nu(X^\ell_i)) \). Asymptotic expansions of this quantity have been studied extensively. In Rempala (2003), it was proved that

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

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

\[
\mathbb{E}[\text{var} \left[ \hat{a}_{\ell,k,j} \bigg| Z \right]] \approx \frac{\Pr[R^c]}{|S|} \sum_{i \in \text{sub}(\ell,k)} \nu(X^\ell_i)\psi^2_{\ell,k,j}(X^\ell_i)|S|\] \( \mathbb{E}[a_{\ell,k,j}] \) \( \frac{1}{|S\cap X^\ell_i|} \)

Wavelets on trees, graphs and high dimensional data
As \( f = (C, \alpha) \)-Hölder, according to Lemma 2 it is \((C_1, \alpha)\) mean-Hölder with \( C_1 = 2^{n+1}C \). Thus, \( \sigma^2(f, X_i^{t+1}) \leq C_2^2 \nu(X_i^{t+1})^{2\alpha} \). Since the tree is balanced, \( \nu(X_i^{t+1}) \leq 2^\nu(X_i^k) \). In addition,

\[
\frac{1}{\Pr[|S \cap X_i^{t+1}| > 0]} \leq \frac{1}{1 - e^{-|S|\nu(X_i^{t+1})}} \leq \frac{1}{1 - e^{-|S|2\nu(X_i^k)}}
\]

Therefore,

\[
\mathbb{E}\left[ \text{var}\left[ \hat{a}_{\ell,k,j} \mid Z \right] \right] \leq \frac{1}{|S|} \frac{C_2^2 B^{2\alpha}}{1 - e^{-|S|2\nu(X_i^k)}} \sum_{i \in \text{sub}(\ell,k)} \nu(X_i^{t+1}) \| \psi_{\ell,k,j} \|^2 (X_i^{t+1})
\]

\[
= \frac{1}{|S|} \frac{C_2^2 B^{2\alpha}}{1 - e^{-|S|2\nu(X_i^k)}}
\]

(29)

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

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

\[
\mathbb{E} [\hat{a}_{\ell,k,j} \mid Z] = \begin{cases} a_{\ell,k,j} & \text{under } R^c \\ 0 & \text{under } R \end{cases}
\]

(30)

Therefore,

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

(31)

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

\[
\mathbb{E}[\hat{a}_{\ell,k,j} - a_{\ell,k,j}]^2 \leq \frac{1}{|S|} \frac{C_2^2 B^{2\alpha}}{1 - e^{-|S|2\nu(X_i^k)}} + a_{\ell,k,j}^2 (1 - \Pr[R]) \Pr[R] + a_{\ell,k,j}^2 \Pr[R]^2
\]

\[
\leq \frac{1}{|S|} \frac{C_2^2 B^{2\alpha}}{1 - e^{-|S|2\nu(X_i^k)}} + \frac{1}{B} e^{-|S|2\nu(X_i^k)} a_{\ell,k,j}^2.
\]

(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,

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

\[
= \sum_{\ell,k,j} \mathbb{E} [a_{\ell,k,j} - \hat{a}_{\ell,k,j}]^2.
\]

Hence,

\[
\mathbb{E} \left\| f - \hat{f} \right\|^2 \leq \frac{C_2^2 B^{2\alpha}}{|S|} \sum_{\ell,k,j} \nu(X_i^{t+1})^{2\alpha} + \frac{1}{B} \sum_{\ell,k,j} e^{-|S|2\nu(X_i^k)} \| \psi_{\ell,k,j} \|^2
\]

\[
\leq \frac{C_2^2 B^{2\alpha}}{|S|} \sum_{\ell,k,j} \left( \frac{B^{2\alpha}}{1 - e^{-|S|B^\ell-1}} \right) + \frac{2\alpha + 1}{B} \sum_{\ell,k,j} e^{-|S|B^\ell} \left( B^{2\alpha + 1} \right)^{\ell-1}
\]

(33)