APPENDIX

Contents

A Proof of Theorem 1

Let $1 \leqslant p, q \leqslant \infty, 0 < s < S$ and $f \in \arg\min_{f \in B^s_{pq}(\mathcal{X})} \sum_{t=1}^T \ell_t(f(x_t))$ the best function to fit the T data over $\mathcal{X} \times [-B, B]$. We start the proof with a decomposition of regret, with any oracle function $f^* \in \mathbb{R}^{\mathcal{X}}$, as

$$\begin{aligned} R_T(f) &= \sum_{t=1}^T \ell_t(\hat{f}_t(x_t)) - \ell_t(f(x_t)) \\ &= \sum_{t=1}^T \ell_t(\hat{f}_t(x_t)) - \ell_t(f^*(x_t)) \quad + \quad \sum_{t=1}^T \ell_t(f^*(x_t)) - \ell_t(f(x_t)) \\ &= \text{estimation regret} \quad + \quad \text{approximation regret.} \end{aligned}$$

Nonlinear oracle. Let $j_0 \in \mathbb{N}$ and $J \geqslant j_0$ to be optimized. We first recall that we use a wavelet development defined for any $J \geqslant j_0$ as

$$f_J(x) := \sum_{k \in \Lambda_{j_0}} \alpha_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{J} \sum_{k \in \Lambda_j} \beta_{j,k} \psi_{j,k}(x), \tag{12}$$

with $\alpha_{j_0,k} = \langle f, \phi_{j_0,k} \rangle$ and $\beta_{j,k} = \langle f, \psi_{j,k} \rangle, j \geqslant j_0$, is a truncated wavelet expansion up to level $J \geqslant j_0$. Using a truncated approximation f_J with a large value of J in (12) can lead to suboptimal regret performance, as it requires estimating a large number of wavelet coefficients, thereby incurring a high estimation error. To address this, we introduce a nonlinear oracle f^* that depends only on a selected subset of coefficients across the J levels. This approach, known as best-term or nonlinear approximation, is surveyed in the textbook [12], while constructions close to us in spirit can be found in [15, 16]. We now make this oracle explicit and show that it balances approximation and estimation errors, in particular achieving minimax optimality in our setting. We define the oracle as

$$f^* = f_{J^*} + f_{\Lambda^*}, \tag{13}$$

where f_{J^*} is a truncated wavelet expansion up to level $J^* \leqslant J$, as in (12) (i.e. we keep all the detail coefficients up to level J^*), and the nonlinear part

$$f_{\Lambda^*} = \sum_{(j,k) \in \Lambda^*} \beta_{j,k} \psi_{j,k}, \quad \Lambda^* \subset \{(j,k) : k \in \Lambda_j,\ J^* < j \leqslant J\},$$

uses only wavelet coefficients indexed by an oracle set Λ^* drawn from the finer scales $j \in (J^*, J]$. The cardinality of Λ^*, i.e., the number of retained coefficients, will be optimized in the analysis.

Intuition: The component f_{Λ^*} of the oracle consists of the $|\Lambda^*|$ largest coefficients chosen adaptively from the fine scales greater than J^*. The procedure is termed *nonlinear* because the choice of coefficients varies with the function f, rather than being a fixed linear rule, contrary to the first J^* levels which keep all coefficients independently of the function.

For any $k \in \Lambda_j, j > j_0$, define $v_{j,k} := \beta_{j,k} 2^{js'}$ with $s' = s + \frac{d}{2} - \frac{d}{p}$ as in (6). Observe that the definition of the Besov norm (6) allows a control over the set $\{v_{j,k} : k \in \Lambda_j, J^* < j \leqslant J\}$ in terms of ℓ^p-norm since

$$\sum_{J^* < j \leqslant J} \sum_{k \in \Lambda_j} |v_{j,k}|^p \leqslant (J - J^*)^{(1-\frac{p}{q})_+} \left(\sum_{J^* < j \leqslant J} \left(\sum_{k \in \Lambda_j} |v_{j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{p}{q}} \leqslant \left[(J - J^*)^{(\frac{1}{p} - \frac{1}{q})_+} \|f\|_{B^s_{pq}} \right]^p =: C_f^p, \tag{14}$$

by Hölder's inequality if $q > p$, else by convexity with $q \leqslant p$.

Let Λ^* denote the set of indices corresponding to the $|\Lambda^*|$ largest wavelet coefficients (in absolute value) among all $(v_{j,k})$ with $j \in [J^* + 1, J]$ and $k \in \Lambda_j$. The cardinality $|\Lambda^*|$ - that is, the number of wavelet coefficients retained in the nonlinear component of the oracle estimator (13)—will be selected later in the analysis as a tuning parameter. Let $j > J^*$. We have that

$$|\Lambda^*| \cdot \min_{(j,k) \in \Lambda^*} |v_{j,k}|^p \leqslant \sum_{(j,k) \in \Lambda^*} |v_{j,k}|^p \leqslant C_f^p < \infty,$$

and in particular since $\forall (j,k) \notin \Lambda^*, |v_{j,k}| \leqslant \min_{(j',k') \in \Lambda^*} |v_{j',k'}|$ one has

$$\forall (j,k) \notin \Lambda^*,\ |\Lambda^*| |v_{j,k}|^p \leqslant C_f^p \implies \forall (j,k) \notin \Lambda^*,\ |\beta_{j,k}| \leqslant C_f 2^{-js'} |\Lambda^*|^{-\frac{1}{p}}. \tag{15}$$

We are now ready to analyze the regret in two steps—that is an *estimation error* and an *approximation error*.

Step 1: Bounding the estimation regret. We set

$$R_1 := \sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t)) - \ell_t(f^*(x_t)) = \sum_{t=1}^{T} \ell_t\left(\sum_{(j,k)} c_{j,k,t}\varphi_{j,k}(x_t)\right) - \ell_t\left(\sum_{(j,k)} c_{j,k}\varphi_{j,k}(x_t)\right),$$

where the sum is over all scaling and detail coefficients with indices in $\{(j_0,k) : k \in \bar{\Lambda}_{j_0}\} \cup \{(j,k) : j \geqslant j_0,\ k \in \Lambda_j\}$, where $c_{j,k}$ stands for either the scaling coefficient $\alpha_{j_0,k}$ or the detail coefficient $\beta_{j,k}$ (and their sequential counterparts $c_{j,k,t}$ depend on t), and $\varphi_{j,k}$ denotes either the scaling function $\phi_{j_0,k}$ or the wavelet function $\psi_{j,k}$.
Since $\hat{y} \mapsto \ell_t(\hat{y})$ is convex and both $\hat{f}_t, f^*$ are linear in the $\{c_{j,k}\}$, then $\ell_t \circ \hat{f}$ and $\ell_t \circ f^*$ are convex in $\{c_{j,k}\}$ and we have by convexity:

$$R_1 \leqslant \sum_{t=1}^{T}\sum_{j,k} g_{j,k,t}(c_{j,k,t} - c_{j,k}),$$

where $g_{j,k,t} = \ell_t'(\hat{f}_t(x_t))\varphi_{j,k}(x_t)$ by Equation (5). Observe that $\max_t g_{j,k,t} \leqslant 2^{\frac{dj}{2}} G\|\varphi\|_\infty =: \hat{G}_j$ for any j,k. Then, first by Assumption 1, and second by the structure of the oracle (13) — namely, that $\forall j > J^*$ such that $(j,k) \notin \Lambda^*$, we have $c_{j,k} = 0$ — we get:

$$\begin{aligned}
R_1 &\leqslant \sum_{j,k}\sum_{t=1}^{T} g_{j,k,t}(c_{j,k,t} - c_{j,k}) \\
&\leqslant \sum_{j,k} |c_{j,k}| \left(C_1\sqrt{\textstyle\sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2\hat{G}_j\right) \\
&= \underbrace{\sum_{j\leqslant J^*,\, k} |c_{j,k}| \left(C_1\sqrt{\textstyle\sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2\hat{G}_j\right)}_{:=R_1(f_{J^*})} + \underbrace{\sum_{\substack{(j,k)\in\Lambda^* \\ j>J^*}} |c_{j,k}| \left(C_1\sqrt{\textstyle\sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2\hat{G}_j\right)}_{:=R_1(f_{\Lambda^*})} \qquad (16)
\end{aligned}$$

where $C_1, C_2 > 0$ are factors (possibly including $\log T$; see Assumption 1). The estimation regret R_1 is thus controlled in (16) by a sum of individual regrets over the nonzero coefficients $c_{j,k}$ that define $f^* = f_{J^*} + f_{\Lambda^*}$. The sum naturally splits into two parts: the linear part $R_1(f_{J^*})$ over $\{(j_0,k) : k \in \bar{\Lambda}_{j_0}\} \cup \{(j,k) : j_0 \leqslant j \leqslant J^*,\ k \in \Lambda_j\}$, and the nonlinear part $R_1(f_{\Lambda^*})$ over the indices in Λ^*.

- Linear part: bounding $R_1(f_{J^*})$. The wavelet basis $\{\phi_{j_0,k}, \psi_{j,k}\}$ is assumed to be S-regular with $S > s$, so we can invoke the characterization of Besov spaces with $\|f\|_{B^s_{pq}} < \infty$ (see Eq. (6)). Let $p, p' \geqslant 1$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Applying Hölder's inequality to the detail coefficients at levels $j \in [j_0, J^*]$, we obtain, with $\hat{G}_j = G2^{\frac{jd}{2}}\|\psi\|_\infty, j \geqslant j_0$:

$$\begin{aligned}
&\sum_{j_0\leqslant j\leqslant J^*}\sum_{k\in\Lambda_j} |\beta_{j,k}| \left(C_1\sqrt{\textstyle\sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2 G\|\psi\|_\infty 2^{\frac{jd}{2}}\right) \\
&\qquad \leqslant \sum_{j_0\leqslant j\leqslant J^*} \Big(\sum_{k\in\Lambda_j} |\beta_{j,k}|^p\Big)^{\frac{1}{p}} \left(C_1\Big(\sum_{k\in\Lambda_j} \big(\sqrt{\textstyle\sum_{t=1}^{T} |g_{j,k,t}|^2}\big)^{p'}\Big)^{\frac{1}{p'}} + C_2 G\|\psi\|_\infty 2^{\frac{jd}{2}} |\Lambda_j|^{\frac{1}{p'}}\right) \\
&\qquad = \sum_{j_0\leqslant j\leqslant J^*} \|\boldsymbol{\beta}_j\|_p \left(C_1\sqrt{\Big(\textstyle\sum_{k\in\Lambda_j} \big(\sum_{t=1}^{T} |g_{j,k,t}|^2\big)^{\frac{p'}{2}}\Big)^{\frac{2}{p'}}} + C_2 G\|\psi\|_\infty 2^{\frac{jd}{2}} |\Lambda_j|^{\frac{1}{p'}}\right) \\
&\qquad \leqslant \sum_{j_0\leqslant j\leqslant J^*} \|\boldsymbol{\beta}_j\|_p \left(C_1 |\Lambda_j|^{(\frac{1}{2}-\frac{1}{p})_+} \sqrt{\textstyle\sum_{k\in\Lambda_j}\sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2 G 2^{\frac{jd}{2}} |\Lambda_j|^{1-\frac{1}{p}}\right)
\end{aligned} \qquad (17)$$

  where the last inequality uses $\|x\|_{\frac{p'}{2}} \leqslant |\Lambda_j|^{(\frac{2}{p'}-1)_+}\|x\|_1$ for a vector x of dimension $|\Lambda_j|$ and $(\cdot)_+ := \max\{\cdot, 0\}$.
  Repeating for the scaling coefficients for $k \in \bar{\Lambda}_{j_0}$, summing in $j = j_0, \ldots, J^*$ and bounding $|\Lambda_j| \leqslant \lambda 2^{dj}$ and $|\bar{\Lambda}_{j_0}| \leqslant \lambda 2^{dj_0}$, we get:

$$\begin{aligned}
R_1(f_{J^*}) &\leqslant \|\boldsymbol{\alpha}_{j_0}\|_p \left(C_1\lambda 2^{dj_0(\frac{1}{2}-\frac{1}{p})_+} \sqrt{\textstyle\sum_{k\in\Lambda_j}\sum_{t=1}^{T} |g_{j_0,k,t}|^2} + C_2 G\|\phi\|_\infty 2^{\frac{j_0 d}{2}} \lambda 2^{dj_0(1-\frac{1}{p})}\right) \\
&\quad + \textstyle\sum_{j=j_0}^{J^*} \|\boldsymbol{\beta}_j\|_p \left(C_1\lambda 2^{dj(\frac{1}{2}-\frac{1}{p})_+} \sqrt{\sum_{k\in\Lambda_j}\sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2 G\|\psi\|_\infty 2^{\frac{jd}{2}} \lambda 2^{dj(1-\frac{1}{p})}\right) \qquad (18)
\end{aligned}$$

where we recall the scaling coefficients are $\boldsymbol{\alpha}_{j_0} = (\alpha_{j_0,k})$ and the detail coefficients at scale j are $\boldsymbol{\beta}_j = (\beta_{j,k})$.
On the other hand, over each level $j \geqslant j_0$, one has

$$\begin{aligned}\sqrt{\sum_{k\in\Lambda_j}\sum_{t=1}^{T}|g_{j,k,t}|^2} &= \sqrt{\sum_{k\in\Lambda_j}\sum_{t=1}^{T}|\ell_t'(\hat{f}_t(x_t))\psi_{j,k}(x_t)|^2} \\ &\leqslant G\sqrt{\sum_{k\in\Lambda_j}\sum_{t=1}^{T}|\psi_{j,k}(x_t)|^2} \\ &= G2^{\frac{dj}{2}}\sqrt{\sum_{t=1}^{T}\sum_{k\in\Lambda_j}|\psi(2^j x_t - k)|^2}\,, \end{aligned} \tag{19}$$

where we used the fact that $|\ell'(\hat{f}_t(x_t))| \leqslant G$ (since $\hat{y} \mapsto \ell_t(\hat{y})$ is G-Lipschitz), the definition of $\psi_{j,k}$ and we applied Jensen's inequality. Equation (19) also holds for the scaling level, replacing $\psi_{j,k}$ by $\phi_{j_0,k}$ over the index set $\bar{\Lambda}_{j_0}$.
By D.2, one has

$$\sup_x \textstyle\sum_k |\phi(x-k)|^2 \leqslant M_\phi\|\phi\|_\infty \quad \text{and} \quad \sup_x \textstyle\sum_k |\psi(x-k)|^2 \leqslant M_\psi\|\psi\|_\infty.$$

With $1 - \frac{1}{p} \leqslant \frac{1}{2} + (\frac{1}{2} - \frac{1}{p})_+$ we get from (18) and (19)

$$\begin{aligned}R_1(f_{J^*}) \leqslant \lambda G\Big[&\big(C_1(M_\phi\|\phi\|_\infty)^{\frac12}\sqrt{T} + C_2\|\phi\|_\infty 2^{\frac{d}{2}j_0}\big)\|\boldsymbol{\alpha}_{j_0}\|_p 2^{dj_0(\frac12+(\frac12-\frac1p)_+)} \\ &\big(C_1(M_\psi\|\psi\|_\infty)^{\frac12}\sqrt{T} + C_2\|\psi\|_\infty 2^{\frac{d}{2}J^*}\big)\sum_{j=j_0}^{J^*}\|\boldsymbol{\beta}_j\|_p 2^{dj(\frac12+(\frac12-\frac1p)_+)}\Big]. \end{aligned} \tag{20}$$

Then since $\|f\|_{B^s_{pq}} < \infty$ in (6), we apply Hölder's inequality with $q, q' \geqslant 1$ that entails

$$\begin{aligned}\sum_{j=j_0}^{J^*}\|\boldsymbol{\beta}_j\|_p 2^{jd(\frac12+(\frac12-\frac1p)_+)} &= \sum_{j=j_0}^{J} 2^{-j(s+\frac{d}{2}-\frac{d}{p})}2^{j(s+\frac{d}{2}-\frac{d}{p})}\|\boldsymbol{\beta}_j\|_p 2^{jd(\frac12+(\frac12-\frac1p)_+)} \\ &\leqslant \Big(\sum_{j=j_0}^{J^*} 2^{-jq'(s-\frac{d}{p}-d(\frac12-\frac1p)_+)}\Big)^{\frac{1}{q'}}\Big(\sum_{j=j_0}^{J} 2^{jq(s+\frac{d}{2}-\frac{d}{p})}\|\boldsymbol{\beta}_j\|_p^q\Big)^{\frac1q} \\ &\leqslant \|f\|_{B^s_{pq}}\sum_{j=0}^{J^*} 2^{-j(s-\frac{d}{p}-d(\frac12-\frac1p)_+)}, \quad \text{since } \|\cdot\|_{q'} \leqslant \|\cdot\|_1, q' \geqslant 1. \end{aligned}$$

Finally, we get from (20) with $\|\boldsymbol{\alpha}_{j_0}\|_p \leqslant \|f\|_{B^s_{pq}}$:

$$\begin{aligned}R_1(f_{J^*}) \leqslant \lambda G\|f\|_{B^s_{pq}} M\Big(&\big(C_1\sqrt{T} + C_2 2^{\frac{d}{2}j_0}\big)2^{dj_0(\frac12+(\frac12-\frac1p)_+)} \\ &+ \big(C_1\sqrt{T} + C_2 2^{\frac{d}{2}J^*}\big)\sum_{j=j_0}^{J^*} 2^{-j\beta}\Big), \end{aligned} \tag{21}$$

with $\beta := s - \frac{d}{p} - d(\frac12 - \frac1p)_+$ and $M := \big(\max(M_\phi\|\phi\|_\infty, M_\psi\|\psi\|_\infty, \|\phi\|_\infty^2, \|\psi\|_\infty^2)\big)^{\frac12} < \infty$

- Nonlinear part: bounding $R_1(f_{\Lambda^*})$. Let $\Lambda^* = \cup_{j=J^*+1}^{J}\Lambda_j^*$ where $|\Lambda_j^*| \leqslant |\Lambda_j|$ is now the oracle sparse set made of positions k at level j. One has by (17) on the levels $j = J^*+1, \dots, J$,

$$\begin{aligned}R_1(f_{\Lambda^*}) &\leqslant \sum_{J^*<j\leqslant J}\|\boldsymbol{\beta}_j\|_p\left(C_1|\Lambda_j^*|^{(\frac12-\frac1p)_+}\sqrt{\textstyle\sum_{k\in\Lambda_j^*}\sum_{t=1}^{T}|g_{j,k,t}|^2} + C_2G2^{\frac{jd}{2}}\|\psi\|_\infty|\Lambda_j^*|^{1-\frac1p}\right) \\ &\leqslant GM\sum_{J^*<j\leqslant J}\|\boldsymbol{\beta}_j\|_p\big(C_1|\Lambda_j^*|^{(\frac12-\frac1p)_+}2^{\frac{dj}{2}}\sqrt{T} + C_2 2^{\frac{jd}{2}}|\Lambda_j^*|^{1-\frac1p}\big) \end{aligned}$$

where second inequality follows from (19). Then, using Hölder's inequality with $q \geqslant 1$ one has

$$\sum_{J^*<j\leqslant J} 2^{\frac{dj}{2}} \|\boldsymbol{\beta}_j\|_p |\Lambda_j^*|^{(\frac{1}{2}-\frac{1}{p})_+} = \sum_{J^*<j\leqslant J} 2^{j(s+\frac{d}{2}-\frac{d}{p})} \|\boldsymbol{\beta}_j\|_p \cdot 2^{-j(s-\frac{d}{p})} |\Lambda_j^*|^{(\frac{1}{2}-\frac{1}{p})_+}$$
$$\leqslant \|f\|_{B^s_{pq}} \sum_{J^*<j\leqslant J} 2^{-j(s-\frac{d}{p})} |\Lambda_j^*|^{(\frac{1}{2}-\frac{1}{p})_+}.$$

Finally, since $\sum_{J^*<j\leqslant J} |\Lambda_j^*| = |\Lambda^*|$, one has

$$\sum_{J^*<j\leqslant J} 2^{-j(s-\frac{d}{p})} |\Lambda_j^*|^{(\frac{1}{2}-\frac{1}{p})_+} \leqslant |\Lambda^*|^{(\frac{1}{2}-\frac{1}{p})_+} \sum_{J^*<j\leqslant J} 2^{-j(s-\frac{d}{p})} \leqslant |\Lambda^*|^{(\frac{1}{2}-\frac{1}{p})_+} \frac{2^{-J^*(s-\frac{d}{p})}}{2^{s-\frac{d}{p}}-1},$$

by Hölder's inequality in the case $p \geqslant 2$ and since $s > \frac{d}{p}$. Similarly, for the second term we have

$$\sum_{J^*<j\leqslant J} 2^{\frac{dj}{2}} \|\boldsymbol{\beta}_j\|_p |\Lambda_j^*|^{1-\frac{1}{p}} \leqslant \|f\|_{B^s_{pq}} \frac{2^{-J^*(s-\frac{d}{p})}}{2^{s-\frac{d}{p}}-1} |\Lambda^*|^{1-\frac{1}{p}}.$$

All in one, with $|\Lambda^*|^{1-\frac{1}{p}} \leqslant |\Lambda^*|^{\frac{1}{2}+(\frac{1}{2}-\frac{1}{p})_+}$ one has

$$R_1(f_{\Lambda^*}) \leqslant GM\|f\|_{B^s_{pq}} \frac{C_1\sqrt{T} + C_2|\Lambda^*|^{\frac{1}{2}}}{2^{s-\frac{d}{p}}-1} \cdot 2^{-J^*(s-\frac{d}{p})} |\Lambda^*|^{(\frac{1}{2}-\frac{1}{p})_+}. \tag{22}$$

- Bound on R_1: We use (21) and (22) and we reach

$$\begin{aligned} R_1 &\leqslant R_1(f_{J^*}) + R_1(f_{\Lambda^*}) \\ &\leqslant G\|f\|_{B^s_{pq}} M \Big[\lambda\big(C_1\sqrt{T} + C_2 2^{\frac{d}{2}j_0}\big) 2^{dj_0(\frac{1}{2}+(\frac{1}{2}-\frac{1}{p})_+)} \\ &\quad + \lambda\big(C_1\sqrt{T} + C_2 2^{\frac{d}{2}J^*}\big) \sum_{j=j_0}^{J^*} 2^{-j\beta} \\ &\quad + \big(C_1\sqrt{T} + C_2|\Lambda^*|^{\frac{1}{2}}\big) |\Lambda^*|^{(\frac{1}{2}-\frac{1}{p})_+} \frac{2^{-J^*(s-\frac{d}{p})}}{2^{s-\frac{d}{p}}-1} \Big], \end{aligned} \tag{23}$$

  where we recall $\beta := s - \frac{d}{p} - d(\frac{1}{2} - \frac{1}{p})_+$ and $s' = s + \frac{d}{2} - \frac{d}{p}$ and $|\Lambda^*|$ is the number of 'non-linear' coefficients we keep below level J^*.

Step 2: Bounding the approximation regret. We now bound the term incurred by approximating f by its nonlinear wavelet approximation f^*. Using the G-Lipschitz property of each loss ℓ_t and the uniform bound on the approximation error, we obtain:

$$R_2 := \sum_{t=1}^{T} \big(\ell_t(f^*(x_t)) - \ell_t(f(x_t))\big) \leqslant G \sum_{t=1}^{T} |f^*(x_t) - f(x_t)| \leqslant GT\|f^* - f\|_\infty. \tag{24}$$

With $f^* = f_{J^*} + f_{\Lambda^*}$ and f_J the truncated wavelet expansion (12) at level $J \geqslant J^* \geqslant j_0$ we have with the triangle inequality

$$R_2 \leqslant GT(\|(f_{J^*} + f_{\Lambda^*}) - f_J\|_\infty + \|f_J - f\|_\infty)$$

First, since f_{J^*}, f_J are both wavelet expansion truncated respectively at level J^* and J, one has

$$\begin{aligned} \|(f_{J^*} + f_{\Lambda^*}) - f_J\|_\infty &= \Big\| \sum_{(j,k)\notin\Lambda^*} \beta_{j,k}\psi_{j,k} \Big\|_\infty \\ &\leqslant \sum_{(j,k)\notin\Lambda^*} \|\beta_{j,k}\psi_{j,k}\|_\infty \\ &\leqslant \sum_{j=J^*+1}^{J} 2^{j\frac{d}{2}} \sup_{k:(j,k)\notin\Lambda^*} |\beta_{j,k}| \cdot \|\textstyle\sum_{k\in\Lambda_j} |\psi(2^j\cdot - k)|\|_\infty \quad \leftarrow \text{by definition of } \psi_{j,k} \\ &\leqslant M_\psi C_f |\Lambda^*|^{-\frac{1}{p}} \sum_{j=J^*+1}^{J} 2^{j\frac{d}{2}} 2^{-js'} \qquad \leftarrow \text{by Definition D.2 and (15)} \\ &\leqslant M_\psi C_f |\Lambda^*|^{-\frac{1}{p}} \frac{2^{-J^*(s-\frac{d}{p})}}{2^{s-\frac{d}{p}}-1} \qquad \leftarrow \text{replacing } s' \text{ and with } s > \frac{d}{p}. \end{aligned} \tag{25}$$

Second, with $s > \frac{d}{p}$, using the characterizations of Besov spaces and classical results on Sobolev embeddings (see, e.g., [20, Prop. 4.3.8] or [8, 13]), $B^s_{pq}(\mathcal{X}) \subset B^{s-\frac{d}{p}}_{\infty\infty}(\mathcal{X})$ and one has

$$\|f_J - f\|_\infty \leqslant M_\psi \|f\|_{B^s_{pq}} \sum_{j>J} 2^{-j(s-\frac{d}{p})} \leqslant M_\psi \|f\|_{B^s_{pq}} \frac{2^{-J(s-\frac{d}{p})}}{2^{s-\frac{d}{p}} - 1}, \tag{26}$$

where $s > \frac{d}{p}$.
Finally, with (25) and (26) one has

$$R_2 \leqslant \frac{G M_\psi \|f\|_{B^s_{pq}}}{2^{s-\frac{d}{p}} - 1} \big(2^{-J(s-\frac{d}{p})} T + (J - J^*)^{(\frac{1}{p}-\frac{1}{q})_+} 2^{-J^*(s-\frac{d}{p})} |\Lambda^*|^{-\frac{1}{p}} T\big). \tag{27}$$

Step 3: Optimization on $J^*, J, |\Lambda^*|$ and conclusion. Let $j_0 = 0$. From (23) and (27) we reach the following regret bound

$$\begin{aligned} R_T(f) = R_1 + R_2 \leqslant CG\|f\|_{B^s_{pq}} \Bigg[&\left(C_1\sqrt{T} + C_2 2^{\frac{d}{2}J^*}\right)\left(1 + \sum_{j=0}^{J^*} 2^{-j\beta}\right) \\ &+ \left(C_1\sqrt{T} + C_2|\Lambda^*|^{\frac{1}{2}}\right) 2^{-J^*(s-\frac{d}{p})} |\Lambda^*|^{(\frac{1}{2}-\frac{1}{p})_+}\Bigg) \\ &+ 2^{-J(s-\frac{d}{p})} T + (J - J^*)^{(\frac{1}{p}-\frac{1}{q})_+} 2^{-J^*(s-\frac{d}{p})} |\Lambda^*|^{-\frac{1}{p}} T \Bigg], \end{aligned} \tag{28}$$

with C some constant that can change from a line to another (depending on $\lambda, \|\psi\|_\infty, M, M_\psi$,...), $\beta = s - \frac{d}{p} - d(\frac{1}{2} - \frac{1}{p})_+$. We keep the explicit dependence on J, J^*, and $|\Lambda^*|$, as we now aim to optimize the upper bound with respect to these parameters. We have three different regimes depending on the sign of β in (28). Observe that

$$\beta = \begin{cases} s - \frac{d}{2} & \text{if } p \geqslant 2, \\ s - \frac{d}{p} & \text{if } p < 2, \end{cases}$$

and we also have $s > \frac{d}{p}$.

Case 1: $\beta > 0$. This regime corresponds to sufficiently regular functions: since $s > \frac{d}{p}$, this corresponds to the case

$$p < 2, \qquad \text{or} \qquad s > \frac{d}{2}.$$

In this case, the geometric sum is bounded by

$$\sum_{j=0}^{J^*} 2^{-j\beta} \leqslant \frac{1}{1 - 2^{-\beta}}.$$

Choosing

$$\begin{cases} J^* = \lceil \frac{1}{d} \log_2(T) \rceil, \\ |\Lambda^*| = 2^{J^* d}, \\ J = \left\lceil \frac{\frac{S}{d}}{\varepsilon} \log_2(T) \right\rceil, \end{cases} \implies \begin{cases} 2^{-J^*(s-\frac{d}{p})} |\Lambda^*|^{-\frac{1}{p}} T = 2^{-sJ^*} T = T^{1-\frac{s}{d}}, \\ 2^{-J^*(s-\frac{d}{p})} |\Lambda^*|^{(\frac{1}{2}-\frac{1}{p})_+} \leqslant T^{\frac{1}{2}-\frac{s}{d}}, \\ 2^{-J(s-\frac{d}{p})} T \leqslant T^{1-\frac{s}{d}}, \\ |\Lambda^*|^{\frac{1}{2}} = 2^{\frac{d}{2}J^*} = \sqrt{T} \end{cases} \tag{29}$$

with $s - \frac{d}{p} > \varepsilon > 0$ and $S > s$, and this entails a total regret

$$\begin{aligned} R_T(f) \leqslant CG\|f\|_{B^s_{pq}} \Bigg[&(C_1 + C_2)\sqrt{T}\left(2 + \frac{1}{1 - 2^{-\beta}} + T^{\frac{1}{2}-\frac{s}{d}}\right) \\ &+ T^{1-\frac{s}{d}} \left(1 + \left(\frac{1}{d}\left(\frac{S}{\varepsilon} - 1\right)\log_2(T)\right)^{(\frac{1}{p}-\frac{1}{q})_+}\right)\Bigg]. \end{aligned} \tag{30}$$

With $s \geqslant d/2$, we have $T^{1-s/d} \leqslant \sqrt{T}$ and $R_T(f) = O(G\|f\|_{B^s_{pq}}\sqrt{T})$.
Remark. The notation $O(\cdot)$ here hides $\log_2(T)$ factors that appear when $p < q$. This originates from the nonlinear oracle construction in the analysis (see Inequality (14)). In addition, log terms may also be absorbed into the constants C_1, C_2 coming from the parameter-free subroutine (see Assumption 1). This remark also holds for the remaining cases.

Case 2: $\beta = 0$**.** This critical regime occurs when $p \geqslant 2$ and $s = \frac{d}{2}$. The sum becomes:

$$\sum_{j=0}^{J^*} 2^{-j\beta} = J^* + 1.$$

Choosing $J^*, |\Lambda^*|$ and J as in (29) yields the bound:

$$R_T(f) \leqslant CG\|f\|_{B^s_{pq}} \left[(C_1 + C_2)\sqrt{T}\left(2 + \frac{1}{d}\log_2 T + T^{\frac{1}{2}-\frac{s}{d}}\right) + \sqrt{T}\left(1 + \left(\frac{1}{d}\left(\frac{S}{\varepsilon} - 1\right)\log_2(T)\right)^{(\frac{1}{p}-\frac{1}{q})_+}\right)\right]. \quad (31)$$

That is $R_T(f) = O(G\|f\|_{B^s_{pq}} \log_2(T)\sqrt{T})$.

Case 3: $\beta < 0$**.** This corresponds to the low regularity case: $\beta = s - \frac{d}{2}$ and

$$p \geqslant 2 \qquad \text{and} \qquad \frac{d}{p} < s < \frac{d}{2}.$$

Here, the geometric sum is bounded as:

$$\sum_{j=0}^{J^*} 2^{-j\beta} \leqslant \frac{2^{-J^*\beta}}{2^{-\beta} - 1}.$$

With $J^*, |\Lambda^*|$ and J as in (29), the regret bound becomes:

$$R_T(f) \leqslant CG\|f\|_{B^s_{pq}} \left[(C_1 + C_2)\sqrt{T}\left(1 + \frac{T^{-\frac{\beta}{d}}}{2^{-\beta} - 1} + T^{\frac{1}{2}-\frac{s}{d}}\right) + T^{1-\frac{s}{d}}\left(1 + \left(\frac{1}{d}\left(\frac{S}{\varepsilon} - 1\right)\log_2(T)\right)^{(\frac{1}{p}-\frac{1}{q})_+}\right)\right]. \quad (32)$$

With $\sqrt{T}T^{-\frac{\beta}{d}} = \sqrt{T}T^{\frac{1}{2}-\frac{s}{d}} = T^{1-\frac{s}{d}}$, one has $R_T(f) = O(G\|f\|_{B^s_{pq}} T^{1-\frac{s}{d}})$.

B Proof of Corollary 1

The proof is based on that of Theorem 1, in Appendix A, case $p = q = \infty$.
Let $s > 0, f \in \arg\min_{f \in \mathscr{C}^s(\mathcal{X})} \sum_{t=1}^T \ell_t(f(x_t))$ the best function to fit the T data over $\mathcal{X} \times \mathbb{R}$ and $f^* = f_J$ defined as in (12). One key point is that in the case of Hölder-smooth function ($p = q = \infty$), the nonlinear set Λ^* of wavelet coefficients will not be needed to achieve optimal rates.
We start with a decomposition of regret with the oracle $f^* = f_J$ as in the proof of Theorem 1 in Appendix A and we have:

$$R_T(f) = \underbrace{\sum_{t=1}^T \ell_t(\hat{f}_t(x_t)) - \ell_t(f_J(x_t))}_{=:R_1} \quad + \quad \underbrace{\sum_{t=1}^T \ell_t(f_J(x_t)) - \ell_t(f(x_t))}_{=:R_2}$$

Step 1: Bounding estimation regret R_1. From (16), one has

$$R_1 \leqslant \sum_{k \in \bar{\Lambda}_{j_0}} |\alpha_{j_0,k}| \left(C_1 \sqrt{\textstyle\sum_{t=1}^T |g_{j,k,t}|^2} + C_2 G \|\phi\|_\infty 2^{\frac{j_0 d}{2}} \right) + \sum_{j=j_0}^J \sum_{k \in \Lambda_j} |\beta_{j,k}| \left(C_1 \sqrt{\textstyle\sum_{t=1}^T |g_{j,k,t}|^2} + C_2 G \|\psi\|_\infty 2^{\frac{jd}{2}} \right) \quad (33)$$

where $C_1, C_2 > 0$ are relative to Assumption 1, $\alpha_{j,k}$ refers to the scaling coefficients and $\beta_{j,k}$ the detail coefficients.
Since the wavelet basis $\{\phi_{j_0,k}, \psi_{j,k}\}$ is assumed to be S-regular with $S > s$ (Definition 2) and $f \in \mathscr{C}^s(\mathcal{X})$, by Proposition 1, the detail coefficients at every level $j \geqslant j_0$ are bounded as:

$$|\beta_{j,k}| = |\langle f, \psi_{j,k} \rangle| \leqslant C(\psi, s) |f|_s 2^{-j(s+d/2)} ,$$

where $C(\psi, s)$ is a positive constant that only depends on the S-regular wavelet basis and $|f|_s$ refers to the semi-norm of f defined in (8).
For the scaling level j_0, then for every k, one has:

$$|\alpha_{j_0,k}| = |\langle f, \phi_{j_0,k} \rangle| \leqslant \|f\|_\infty \cdot \|\phi_{j_0,k}\|_1 \leqslant 2^{-\frac{j_0 d}{2}} \|\phi\|_1 \|f\|_\infty ,$$

where we used

$$\|\phi_{j_0,k}\|_1 \leqslant \int_{\mathbb{R}^d} 2^{j_0 d/2} \phi(2^{j_0} x - k)\, dx \overset{u = 2^{j_0} x - k}{=} 2^{j_0 \frac{d}{2}} 2^{-j_0 d} \int_{\mathbb{R}^d} |\phi(u)|\, du = 2^{-\frac{j_0 d}{2}} \|\phi\|_1$$

and $\|\phi\|_1 < \infty$ since the scaling function ϕ is assumed to be localized (e.g. compactly supported).
Then, plugging the above upper bound, we get with $g_{j_0,k,t} \leqslant G \|\phi\|_\infty 2^{\frac{j_0 d}{2}}$:

$$R_1 \leqslant G \|\phi\|_1 \|f\|_\infty 2^{-\frac{j_0 d}{2}} \cdot |\bar{\Lambda}_{j_0}| \cdot \|\phi\|_\infty \big(C_1 2^{\frac{j_0 d}{2}} \sqrt{T} + C_2 2^{\frac{j_0 d}{2}} \big) + C(\psi, s) |f|_s \sum_{j=j_0}^J 2^{-j(s+d/2)} \left(C_1 \sum_{k \in \Lambda_j} \sqrt{\sum_{t=1}^T |g_{j,k,t}|^2} + C_2 G \|\psi\|_\infty 2^{\frac{jd}{2}} |\Lambda_j| \right) ,$$

Using Cauchy-Schwarz's inequality as long as the form of the gradients in (5) and the bound (19), we have over each level $j \in [j_0, J]$,

$$\sum_{k \in \Lambda_j} \sqrt{\sum_{t=1}^T |g_{j,k,t}|^2} \leqslant G \lambda^{\frac{1}{2}} 2^{dj} \sqrt{\sum_{t=1}^T \sum_{k \in \Lambda_j} |\psi(2^j x_t - k)|^2} \leqslant G (\lambda \|\psi\|_\infty M_\psi)^{\frac{1}{2}} 2^{dj} \sqrt{T}$$

where $\| \sum_{k \in \Lambda_j} |\psi(\cdot - k)|^2 \|_\infty \leqslant \|\psi\|_\infty M_\psi < \infty$ (see D.2) and $|\Lambda_j| = \lambda 2^{dj}$. Finally, with $M = \max(M_\psi \|\psi\|_\infty, \|\phi\|_\infty^2, \|\psi\|_\infty^2)^{\frac{1}{2}}$

$$R_1 \leqslant GM\lambda \left(\|f\|_\infty \|\phi\|_1\, 2^{j_0 d} \left(C_1 \sqrt{T} + C_2 \right) + C(\psi, s)\, |f|_s \left(C_1 \sqrt{T} + C_2 2^{\frac{d}{2} J} \right) \sum_{j=j_0}^J 2^{-j(s-d/2)} \right) \quad (34)$$

Setting $j_0 = 0$, the sum can be upper-bounded with 3 different cases as

$$\sum_{j=0}^J 2^{-j(s-\frac{d}{2})} \leqslant \begin{cases} (1 - 2^{-(s-\frac{d}{2})})^{-1} & \text{if } d < 2s\,, \\ J + 1 & \text{if } d = 2s\,, \\ 2^{-(J+1)(s-\frac{d}{2})} (2^{-(s-\frac{d}{2})} - 1)^{-1} & \text{if } d > 2s\,. \end{cases}$$

Step 2: Bounding the approximation regret. Following (27), one has:

$$R_2 := \sum_{t=1}^{T} \ell_t(\hat{f}^*(x_t)) - \ell_t(f(x_t)) \leqslant G \sum_{t=1}^{T} |\hat{f}^*(x_t) - f(x_t)| \leqslant GT \|K_J f - f\|_\infty \leqslant C_4 GT |f|_s 2^{-sJ}\,, \quad (35)$$

where $C_4 = C_4(\psi, s)$ - see [20, Prop. 4.1.5] for instance with assumption D.3.

Step 3: upper-bounding $R_T(f)$. We need to balance (34) and (35), and finding the optimal $J \geqslant 0$. Taking $j_0 = 0$ — i.e. $|\bar{\Lambda}_{j_0}| \leqslant \lambda$ — and $J = \lceil \frac{1}{d} \log_2(T) \rceil$ entails the desired bound in the 3 cases $d < 2s$, $d = 2s$ and $d > 2s$.

Remark. In the preceding proof we showed that a single (linear) global resolution level $J = \lceil \frac{1}{d} \log_2 T \rceil$ suffices to attain the minimax regret for Hölder-smooth competitors, in contrast to general competitors in $B^s_{pq}(\mathcal{X})$, which require a *nonlinear* mechanism (see Appendix A). Nevertheless, even in the Hölder-smooth case one may take a larger level $J = \lceil \frac{S}{d\varepsilon} \log_2 T \rceil \geqslant \lceil \frac{1}{d} \log_2 T \rceil$ as in Theorem 1. In the analysis, set the oracle coefficients $\beta_{j,k} = 0$ for levels $j > \lceil \frac{1}{d} \log_2 T \rceil$; the estimation regret, combined with Assumption 1, reduces to a sum over the remaining nonzero coefficients—namely, those in the linear part—and leads to the same rates.

C Proof of Theorem 2

The proof uses a first key result that we state and prove right after.

Theorem 3 (Local regret over Besov spaces)**.** *Let $T \geqslant 1, 1 \leqslant p, q \leqslant \infty, s > \frac{d}{p}$ and $f \in B^s_{pq}$. Under the same assumptions of Theorem 1 and Assumption 2, Algorithm 2 with $\|f\|_\infty \leqslant B$ has regret*

$$R_T(f) \lesssim G \inf_{\mathcal{T}'} \left\{ \sum_{n \in \mathcal{L}(\mathcal{T}')} B\sqrt{|T_n|} + \|f\|_{s_n} \cdot 2^{-\mathrm{l}(n) s_n} \cdot |T_n|^{r(s_n,p)} \right\},$$

and if (ℓ_t) are exp-concave:

$$R_T(f) \lesssim G \inf_{\mathcal{T}'} \left\{ B|\mathcal{L}(\mathcal{T}')| + \sum_{n \in \mathcal{L}(\mathcal{T}')} \|f\|_{s_n} \cdot 2^{-\mathrm{l}(n) s_n} \cdot |T_n|^{r(s_n,p)} \right\},$$

where $\lesssim$ hides logarithmic factors in T, and constants independent of f or T, $\mathcal{L}(\mathcal{T}')$ denotes the set of leaves in a pruning $\mathcal{T}' \subset \mathcal{T}$, $\|f\|_{s_n}$ are local Besov norms, $\mathrm{l}(n)$ is the level of node $n \in \mathcal{L}(\mathcal{T}')$, and the local rate exponent is given by

$$r(s_n, p) = \begin{cases} \frac{1}{2} & \text{if } s_n \geqslant \frac{d}{2} \text{ or } p < 2, \\ 1 - \frac{s_n}{d} & \text{otherwise.} \end{cases}$$

Remark. Theorem 3 holds for any pruning $\mathcal{T}'$ of $\mathcal{T}$. In particular, our procedure effectively competes against the best pruning with respect to the profile of the competitor f. Intuitively, Algorithm 2 achieves a spatial trade-off over the input space: it can refine locally by going deeper with high $\mathrm{l}(n)$ at the cost of increasing the number of leaves $|\mathcal{L}(\mathcal{T}')|$, while remaining coarser and less accurate in other regions, with fewer leaves to compete against. In particular, when applying the result to a specific pruning, we show in Theorem 4 that Algorithm 2 achieves minimax-optimal (local) regret when facing exp-concave losses.

Proof of Theorem 3. Let $1 \leqslant p, q \leqslant \infty, s > \frac{d}{p}, f \in B^s_{pq}$ such that $B := \|f\|_\infty < \infty$ - this is possible since f is continuous over $\mathcal{X}$ with the condition $s > d/p$ and embedding of B^s_{pq} in L^∞.

Grid for scaling coefficients at starting scale j_0. Observe that

$$|\alpha_{j_0,k}| = |\langle f, \phi_{j_0,k} \rangle| \leqslant \|f\|_\infty \cdot \|\phi_{j_0,k}\|_1 \leqslant 2^{-\frac{j_0 d}{2}} \|\phi\|_1 B.$$

Let $\varepsilon_{j_0} > 0$. We define the regular grid $\mathcal{A}_{j_0}$ of ε_{j_0}-precision, used to learn the scaling coefficients at level j_0, denoted $(\alpha_{j_0,k})$, with

$$|\mathcal{A}_{j_0}| = \left\lceil 2^{-j_0 \frac{d}{2}} \, 2B \, \|\phi\|_1 \, \varepsilon_{j_0}^{-1} \right\rceil$$

points, regularly spaced in the interval $[-B2^{-j_0 \frac{d}{2}} \|\phi\|_1, \; B2^{-j_0 \frac{d}{2}} \|\phi\|_1]$. In the following, we will use a local grid $\mathcal{A}_{\mathrm{l}(n)}$ to learn scaling coefficient $\alpha_{n,k}$ at a scale $j_0 = \mathrm{l}(n)$ locally over the space $\mathcal{X}$. In particular, we will carefully set the local precision $\varepsilon_{\mathrm{l}(n)}$ to handle regret terms.

Definition of the oracle associated to a pruning. Let $\mathcal{T}'$ be some pruning of $\mathcal{T}$ and $\mathcal{P}(\mathcal{T}') = (\mathcal{X}_n)_{n \in \mathcal{L}(\mathcal{T}')}$ be the associated partition of $\mathcal{X}$. Let $\mathcal{A}_{\mathrm{l}(n)}$ denote the grid of precision $\varepsilon_{\mathrm{l}(n)}$ as described above. We define the prediction function of pruning $\mathcal{T}'$, at any time $t \geqslant 1$

$$\hat{f}_{\mathcal{T}',t}(x) = \sum_{n \in \mathcal{L}(\mathcal{T}')} [\hat{f}_{n,\mathbf{a}_n,t}(x)]_B \,, \qquad x \in \mathcal{X},$$

where each $\hat{f}_{n,\mathbf{a}_n,t}$ is a sequential predictor of type (3), with starting scale $j_0 = \mathrm{l}(n)$, restricted to $\mathcal{X}_n$, and initialized at the oracle scaling coefficients

$$\mathbf{a}_n = (a_{n,k})_{k \in \bar{\Lambda}_{j_0,n}} = \operatorname*{arg\,min}_{\mathbf{a} \in \mathcal{A}_{\mathrm{l}(n)}} \|\mathbf{a} - \boldsymbol{\alpha}_{j_0,n}\|_\infty,$$

that is, the best approximating vector $\mathbf{a}$ in the grid $\mathcal{A}_{\mathrm{l}(n)}$ for the subset of scaling coefficients $\boldsymbol{\alpha}_{j_0,n,k}, k \in \bar{\Lambda}_{j_0,n}$, whose basis functions $\phi_{j_0,k}$ are supported on $\mathcal{X}_n$. For simplicity, we slightly abuse notation by writing $\mathbf{a} \in \mathcal{A}_{\mathrm{l}(n)}$, treating the grid as a tensor grid of the same dimension as $\mathbf{a}$. In particular, the number of coefficients in $\bar{\Lambda}_{j_0}$ whose supports intersect $\mathcal{X}_n$ satisfies

$$|\bar{\Lambda}_{j_0,n}| \leqslant |\bar{\Lambda}_{j_0}| \, 2^{-\mathrm{l}(n) d} \leqslant \lambda,$$

since $\mathrm{l}(n) = j_0$.

Decomposition of regret. We have a decomposition of regret as:

$$\mathrm{Reg}_T(f) = \underbrace{\sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t)) - \ell_t(\hat{f}_{\mathcal{T}',t}(x_t))}_{=:R_1} + \underbrace{\sum_{t=1}^{T} \ell_t(\hat{f}_{\mathcal{T}',t}(x_t)) - \ell_t(f(x_t))}_{=:R_2}, \tag{36}$$

R_1 is the regret related to the estimation error of the expert-aggregation algorithm compared to some oracle partition $\mathcal{P}(\mathcal{T}')$ associated to $\mathcal{T}'$, i.e. the error the algorithm commits while aiming the oracle partition $\mathcal{P}(\mathcal{T}')$. On the other hand, R_2 is related to the error of the model predicting over subregions in $\mathcal{P}(\mathcal{T}')$, against some function $f \in B^s_{pq}$ and corresponds to the (localised) regret discussed in Theorem 1.

Step 1: Upper-bounding R_2 as local regrets. Recall that $\mathcal{P}(\mathcal{T}')$ form a partition of $\mathcal{X}$. Hence, for any $x_t \in \mathcal{X}$, the prediction at time t is $\hat{f}_{\mathcal{T}',t}(x_t) = [\hat{f}_{j_0,n,\mathbf{a}_n,t}(x_t)]_B$ with $n \in \mathcal{N}(\mathcal{T}')$ the unique node such that $x_t \in \mathcal{X}_n$ at time t. Then, R_2 can be written as follows:

$$\begin{aligned} R_2 &= \sum_{t=1}^{T} \sum_{n \in \mathcal{L}(\mathcal{T}')} (\ell_t(\hat{f}_{\mathcal{T}',t}(x_t)) - \ell_t(f(x_t))) \mathbb{1}_{x_t \in \mathcal{X}_n} \\ &= \sum_{n \in \mathcal{L}(\mathcal{T}')} \sum_{t \in T_n} \ell_t([\hat{f}_{n,\mathbf{a}_n,t}(x_t)]_B) - \ell_t(f(x_t)) \\ &\leqslant \sum_{n \in \mathcal{L}(\mathcal{T}')} \sum_{t \in T_n} \ell_t(\hat{f}_{n,\mathbf{a}_n,t}(x_t)) - \ell_t(f(x_t)), \end{aligned} \tag{37}$$

where we set $T_n = \{1 \leqslant t \leqslant T : x_t \in \mathcal{X}_n\}, \mathcal{X}_n \subset \mathcal{X}, n \in \mathcal{L}(\mathcal{T}')$ and (37) is because $[\hat{f}_{n,\mathbf{a}_n,t}]_B \leqslant \hat{f}_{n,\mathbf{a}_n,t}$ and ℓ_t is convex and has minimum in $[-B, B]$ with $B \geqslant \|f\|_\infty$.
The decomposition in (37) represents a sum of *local* error approximations of the function f over the partition $\mathcal{P}(\mathcal{T}')$, using predictors $\hat{f}_{n,\mathbf{a}_n}, n \in \mathcal{L}(\mathcal{T}')$. Recall that for every $n \in \mathcal{L}(\mathcal{T}')$, $\hat{f}_{n,\mathbf{a}_n}$ is a prediction function associated to a wavelet decomposition (3), where the scaling coefficients start at $\mathbf{a}_n$ over $\mathcal{X}_n$ and with $j_0 = \mathrm{l}(n)$. In proof of Theorem 1 (Appendix A) we showed that any wavelet decomposition adapts to any regularity via $\|f\|_{B^s_{pq}}, s$ of f. Thus, the approximation error of $\hat{f}_{j_0,n,\mathbf{a}_n}$ with respect to f remains similar to that in (27), but now with regard to a Besov function with local smoothness s_n and norm $\|f\|_{s_n} := \|f\|_{B^{s_n}_{pq}(\mathcal{X}_n)}$ over $\mathcal{X}_n$ - see (10). Specifically, from (37), (23), (27), we get (without applying Hölder's inequality on the scaling coefficients):

$$\begin{aligned} R_2 \leqslant \sum_{n \in \mathcal{L}(\mathcal{T}')} \Bigg[& G\|\phi\|_\infty \sum_{k \in \bar{\Lambda}_{j_0,n}} |\alpha_{j_0,k} - a_{n,k}| \big(C_1 \sqrt{|T_n|} + C_2 2^{\frac{\mathrm{l}(n)d}{2}}\big) \\ & + \underbrace{\sum_{j=\mathrm{l}(n)}^{\mathrm{l}(n)+J_n^*} \lambda \|\boldsymbol{\beta}_j\|_p 2^{dj(\frac{1}{2}-\frac{1}{p})_+} \big(C_1 \sqrt{\textstyle\sum_{k \in \Lambda_{j,n}} \sum_{t=1}^{T} |g_{j,k,t}|^2} + C_2 G \|\psi\|_\infty 2^{\frac{jd}{2}} 2^{dj(1-\frac{1}{p})}\big)}_{\text{estimation error on wavelet coefficients as in (23)}} \\ & + \underbrace{2^{-(\mathrm{l}(n)+J_n^*)(s_n - \frac{d}{p})} |\Lambda_n^*|^{(\frac{1}{2}-\frac{1}{p})_+}}_{\text{estimation error on nonlinear wavelet coefficients}} \\ & + \underbrace{C_5 G \|f\|_{s_n} \big(2^{-(\mathrm{l}(n)+J_n^*)(s_n - \frac{d}{p})} |\Lambda_n^*|^{-\frac{1}{p}} |T_n| + 2^{-(\mathrm{l}(n)+J_n)(s_n - \frac{d}{p})} |T_n|\big)}_{\text{approximation error (27) over } \mathcal{X}_n \text{ at scale } j_0 + J_n} \Bigg], \end{aligned} \tag{38}$$

with C_1, C_2 as in Assumption 1 and C_5 a constant that can be deduced from (27) and $j_0 = \mathrm{l}(n)$ for each $n \in \mathcal{L}(\mathcal{T}')$. In particular, by definition of $\mathbf{a}_n = \arg\min_{\mathbf{a} \in \mathcal{A}_{\mathrm{l}(n)}} \|\boldsymbol{\alpha}_{j_0,n} - \mathbf{a}\|_\infty$, and given that $\mathcal{A}_{\mathrm{l}(n)}$ is a grid with precision $\varepsilon_{\mathrm{l}(n)} > 0$, one has

$$|\alpha_{j_0,k} - a_{n,k}| \leqslant \frac{\varepsilon_{\mathrm{l}(n)}}{2} \quad \text{for every } k \in \bar{\Lambda}_{j_0,n}.$$

From (38), one can bound the absolute values of the scaling terms by $\varepsilon_{\mathrm{l}(n)}/2$ and using $|\bar{\Lambda}_{j_0,n}| \leqslant \lambda$. For every $n \in \mathcal{L}(\mathcal{T}')$, let

$$\varepsilon_{\mathrm{l}(n)} = B\big(2^{\frac{\mathrm{l}(n)d}{2}} \sqrt{T}\big)^{-1}$$

that gives for every $n \in \mathcal{L}(\mathcal{T}')$

$$\sum_{k \in \bar{\Lambda}_{j_0,n}} \frac{\varepsilon_{\mathrm{l}(n)}}{2}\big(C_1\sqrt{|T_n|} + C_2 2^{\frac{\mathrm{l}(n)d}{2}}\big) \leqslant \frac{\lambda}{2}B\big(C_1 2^{-\frac{\mathrm{l}(n)d}{2}} + C_2 T^{-\frac{1}{2}}\big),$$

where we used $\sqrt{|T_n|}/\sqrt{T} \leqslant 1$. Then, one can factorize the sum in j and the approximation term by $2^{-\mathrm{l}(n)s_n}$ over each $n \in \mathcal{L}(\mathcal{T}')$. Finally, applying Hölder's inequality over the sum in j (see (21)) and following the same optimization steps in $J_n^*, J_n, |\Lambda_n^*|$ as in Proof of Theorem 1 we get, with M defined as in (21):

$$R_2 \leqslant \lambda GMB|\mathcal{L}(\mathcal{T}')|\big(C_1 + C_2 T^{-\frac{1}{2}}\big) + \lambda GM \sum_{n \in \mathcal{L}(\mathcal{T}')} C_n \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} \begin{cases} \sqrt{|T_n|} & \text{if } s_n \geqslant \frac{d}{2} \text{ or } p < 2 \\ |T_n|^{1-\frac{s_n}{d}} & \text{else,} \end{cases} \tag{39}$$

where $C_n = C_n(C_1, C_2, C_3, s_n, \psi, p)$ can be deduced from similar calculation as in (30), (31) and (32) and can include $\log T$ dependencies.

Step 2: Upper-bounding the estimation error R_1. R_1 is due to the error incurred by sequentially learning the prediction rule $\hat{f}_{\mathcal{T}'}$ associated with an oracle pruning $\mathcal{T}'$ of $\mathcal{T}$, along with the best scaling coefficients $(\mathbf{a}_n)_{n \in \mathcal{L}(\mathcal{T}')}$ selected from the grid $(\mathcal{A}_{\mathrm{l}(n)})_{n \in \mathcal{L}(\mathcal{T}')}$.
Note that at each time t, only a subset of nodes in $\mathcal{T}$ are active and output predictions. Specifically, for any time $t \geqslant 1$, we define in Algorithm 2 the set of active experts at round t as

$$\mathcal{E}_t = \{(n, \mathbf{a}_n) : x_t \in \mathcal{X}_n\}.$$

Moreover, we assume bounded gradients: for any time $t \geqslant 1$ and expert $e \in \mathcal{E}$,

$$|\nabla_{t,e}| = \big|\ell_t'(\hat{f}_t(x_t)) \cdot [\hat{f}_{e,t}(x_t)]_B\big| \leqslant GB,$$

which satisfies Assumption 2 with $\tilde{G} = BG$.
Using standard sleeping reduction, one can prove that, for any expert $(n, \mathbf{a}_n), n \in \mathcal{L}(\mathcal{T}'), t \geqslant 1$ - see Proof of Theorem 2 in [29] Eq. (31)-(35):

$$\begin{aligned}
(\ell_t(\hat{f}_t(x_t)) - \ell_t(\hat{f}_{n,\mathbf{a}_n,t}(x_t)))\mathbb{1}_{x_t \in \mathcal{X}_n} &\leqslant \ell_t'(\hat{f}_t(x_t))(\hat{f}_t(x_t) - \hat{f}_{n,\mathbf{a}_n,t}(x_t))\mathbb{1}_{x_t \in \mathcal{X}_n} \quad \leftarrow \text{by convexity of } \ell_t \\
&= (\nabla_t^\top \tilde{\mathbf{w}}_t - \nabla_{(n,\mathbf{a}_n),t})\mathbb{1}_{x_t \in \mathcal{X}_n} \qquad (40) \\
&= \nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}. \qquad (41)
\end{aligned}$$

Then, with $T_n = \{1 \leqslant t \leqslant T : x_t \in \mathcal{X}_n\}, n \in \mathcal{L}(\mathcal{T}')$:

$$\begin{aligned}
R_1 &= \sum_{t=1}^{T} \sum_{n \in \mathcal{L}(\mathcal{T}')} (\ell_t(\hat{f}_t(x_t)) - \ell_t(\hat{f}_{n,\mathbf{a}_n,t}(x_t)))\mathbb{1}_{x_t \in \mathcal{X}_n} \quad \leftarrow \{\mathcal{X}_n, n \in \mathcal{L}(\mathcal{T}')\} \text{ partition of } \mathcal{X} \\
&\leqslant \sum_{n \in \mathcal{L}(\mathcal{T}')} \sum_{t=1}^{T} (\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}) \quad \leftarrow \text{by (41)} \\
&\leqslant \sum_{n \in \mathcal{L}(\mathcal{T}')} \Big(C_3\sqrt{\log\big(|\mathcal{E}|\big)}\sqrt{\textstyle\sum_{t=1}^{T}(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t})^2} + C_4\tilde{G}\Big) \quad \leftarrow \text{by Assumption 2} \\
&= C_4 BG|\mathcal{L}(\mathcal{T}')| + C_3\sqrt{\log\big(|\mathcal{E}|\big)} \sum_{n \in \mathcal{L}(\mathcal{T}')} \sqrt{\sum_{t \in T_n} (\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t})^2}, \qquad (42)
\end{aligned}$$

where the last equality holds because for any $n \in \mathcal{L}(\mathcal{T}'), \nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t} = 0$ if $x_t \notin \mathcal{X}_n$ and $\tilde{G} = BG$.
The proof goes on with two different cases depending on the losses' convex properties:

- *Case 1: $(\ell_t)_{1 \leqslant t \leqslant T}$ convex.*
  Observe that at any time $t \in [T], \|\nabla_t\|_\infty \leqslant \tilde{G}$ and $\|\mathbf{w}_t\|_1 = 1$, which gives $|\nabla_t^\top \mathbf{w}_t| \leqslant \tilde{G} = BG$. Then, from (42)

$$R_1 \leqslant C_4 BG|\mathcal{L}(\mathcal{T}')| + 2C_3\sqrt{\log\big(|\mathcal{E}|\big)}BG \sum_{n \in \mathcal{L}(\mathcal{T}')} \sqrt{|T_n|} \tag{43}$$

In case of convex losses, we finally have by (36), (39) and (43):

$$\mathrm{Reg}_T(f) \leqslant 2C_3BG\sqrt{\log\left(|\mathcal{E}|\right)} \sum_{n\in\mathcal{L}(\mathcal{T}')} \sqrt{|T_n|} + \lambda GBM|\mathcal{L}(\mathcal{T}')|\left(C_1 + C_2T^{-\frac{1}{2}}\right) + \lambda GM \sum_{n\in\mathcal{L}(\mathcal{T}')} C_n\|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} \begin{cases} \sqrt{|T_n|} & \text{if } s_n \geqslant \frac{d}{2} \text{ or } p < 2, \\ |T_n|^{1-\frac{s_n}{d}} & \text{else,} \end{cases} \tag{44}$$

where $|\mathcal{E}| \leqslant |\mathcal{N}(\mathcal{T})|\left(2\|\phi\|_1 T^{\frac{1}{2}}\right)^\lambda$ since for every $n \in \mathcal{T}$ one has $|\bar{\Lambda}_{j_0,n}| \leqslant \lambda$ and

$$|\mathcal{A}_{\mathrm{l}(n)}| = \lceil 2B\|\phi\|_1/\varepsilon_{\mathrm{l}(n)}\rceil = \lceil 2\|\phi\|_1 T^{\frac{1}{2}}\rceil$$

by the choice of the precision $\varepsilon_{\mathrm{l}(n)} = B2^{-\frac{\mathrm{l}(n)d}{2}}T^{-\frac{1}{2}}$. In particular, the grids have a number of points that does not grow exponentially with T, making the construction computationally feasible. Finally, since (44) holds for all pruning $\mathcal{T}'$ of our main tree $\mathcal{T}$, one can take the infimum over all pruning to get the desired upper-bound.

- *Case 2: $(\ell_t)_{1\leqslant t\leqslant T}$ η-exp-concave.*
  If the sequence of loss functions (ℓ_t) is η-exp-concave for some $\eta > 0$, then thanks to [24, Lemma 4.3] we have for any $0 < \mu \leqslant \frac{1}{2}\min\{\frac{1}{G}, \eta\}$ and all $t \geqslant 1, n \in \mathcal{L}(\mathcal{T}')$, using (41):

$$(\ell_t(\hat{f}_t(x_t)) - \ell_t(\hat{f}_{n,\mathbf{a}_n,t}(x_t)))\mathbb{1}_{x_t\in\mathcal{X}_n} \leqslant \nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t} - \frac{\mu}{2}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2 \tag{45}$$

Summing (45) over $t \in [T]$ and $n \in \mathcal{L}(\mathcal{T}')$, we get:

$$\begin{aligned} R_1 &\leqslant \sum_{n\in\mathcal{L}(\mathcal{T}')}\sum_{t\in T_n} \nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t} - \frac{\mu}{2}\sum_{n\in\mathcal{L}(\mathcal{T}')}\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2 \\ &\leqslant C_4\tilde{G}|\mathcal{L}(\mathcal{T}')| + \tilde{C}_3 \sum_{n\in\mathcal{L}(\mathcal{T}')} \sqrt{\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2} - \frac{\mu}{2}\sum_{n\in\mathcal{L}(\mathcal{T}')}\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2, \end{aligned} \tag{46}$$

where last inequality is by (42) and we set $\tilde{C}_3 = C_3\sqrt{\log\left(|\mathcal{E}|\right)}$ and $\tilde{G} = BG$. Young's inequality gives, for any $\nu > 0$, the following upper-bound:

$$\sqrt{\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2} \leqslant \frac{1}{2\nu} + \frac{\nu}{2}\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2. \tag{47}$$

Finally, plugging (47) with $\nu = \mu/\tilde{C}_3 > 0$ in (46), we get

$$\begin{aligned} R_1 &\leqslant C_4BG|\mathcal{L}(\mathcal{T}')| + \tilde{C}_3 \sum_{n\in\mathcal{L}(\mathcal{T}')}\left(\frac{\tilde{C}_3}{2\mu} + \frac{\mu}{2\tilde{C}_3}\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2\right) \\ &\quad - \frac{\mu}{2}\sum_{n\in\mathcal{L}(\mathcal{T}')}\sum_{t\in T_n}\left(\nabla_t^\top \mathbf{w}_t - \nabla_{(n,\mathbf{a}_n),t}\right)^2 \\ &= \left(\frac{C_3^2\log\left(|\mathcal{E}|\right)}{2\mu} + C_4BG\right)|\mathcal{L}(\mathcal{T}')|, \end{aligned} \tag{48}$$

again with $|\mathcal{E}| \leqslant |\mathcal{N}(\mathcal{T})|\left(2\|\phi\|_1 T^{\frac{1}{2}}\right)^\lambda$. Then, one can deduce the final bound from equations (36), (39) and (48) and taking the infimum over the prunings $\mathcal{T}'$.

Worst case regret bound. Note that since we assume that $\|f\|_\infty \leqslant B$, and that all local predictors $\hat{f}_e, e \in \mathcal{E}$ in Algorithm 2 are clipped in $[-B, B]$, we first have for any $x \in \mathcal{X}$,

$$|\hat{f}_t(x)| = \sum_{e\in\mathcal{E}_t} w_{e,t}|[\hat{f}_{e,t}(x)]_B| \leqslant B\sum_{e\in\mathcal{E}_t} w_{e,t} = B.$$

Thus,

$$\begin{aligned}
\mathrm{Reg}_T(f) &= \sum_{t=1}^{T} \ell_t(\hat{f}_t(x_t)) - \ell_t(f(x_t)) \\
&\leqslant \sum_{t=1}^{T} G|\hat{f}_t(x_t) - f(x_t)| && \leftarrow \ell_t \text{ is } G\text{-Lipschitz} \\
&\leqslant G\sum_{t=1}^{T} |\hat{f}_t(x_t)| + |f(x_t)| \\
&= 2BGT
\end{aligned} \tag{49}$$

□

We now restate a complete version of Theorem 2 from the main text and provide its proof below.

Theorem 4. *Let $T \geqslant 1, 1 \leqslant p, q \leqslant \infty, s > \frac{d}{p}$. Let $f \in B^s_{pq}$ and $B \geqslant \|f\|_\infty$. Let $\mathcal{T}'$ be any pruning of $\mathcal{T}$, together with a collection of local smoothness indices $(s_n)_{n\in\mathcal{L}(\mathcal{T}')}$ defined as in (10) and local norms $\|f\|_{s_n}$. Then, under the same assumptions of Theorem 1 and Assumption 2, Algorithm 2 satisfies*

$$\begin{aligned}
R_T(f) \lesssim G \sum_{n\in\mathcal{L}(\mathcal{T}')} \Big(& B^{1-\frac{d}{2s_n}} \big(2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}\big)^{\frac{d}{2s_n}} \sqrt{|T_n|}\mathbb{1}_{s_n \geqslant \frac{d}{2}} \\
& + \big(2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}|T_n|^{1-\frac{s_n}{d}}\big)\mathbb{1}_{s_n<\frac{d}{2}} + B\sqrt{|T_n|}\Big)
\end{aligned}$$

and moreover we also have, if (ℓ_t) are exp-concave:

$$\begin{aligned}
R_T(f) \lesssim G \sum_{n\in\mathcal{L}(\mathcal{T}')} \Big(& B^{1-\frac{2d}{2s_n+d}} \big(2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}\big)^{\frac{2d}{2s_n+d}} |T_n|^{\frac{d}{2s_n+d}} \mathbb{1}_{s_n \geqslant \frac{d}{2}} \\
& + 2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}|T_n|^{1-\frac{s_n}{d}}\mathbb{1}_{s_n<\frac{d}{2}} + B\Big),
\end{aligned}$$

where $\lesssim$ hides logarithmic factors in T, and constants independent of f or T.

Proof of Theorem 4.
Let $\mathcal{T}' \subset \mathcal{T}$ be some pruning of $\mathcal{T}$. We define $\mathcal{T}'_{\text{ext}}$ the extension of $\mathcal{T}'$ such that all terminal nodes $n \in \mathcal{L}(\mathcal{T}')$ is extended with a tree $\mathcal{T}'_n$ of depth $h_n \in \mathbb{N}$. In particular, for any $n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}}), \mathrm{l}(n') = \mathrm{l}(n) + h_n$ with $n \in \mathcal{L}(\mathcal{T}')$. See Figure 4 for an illustrative example.

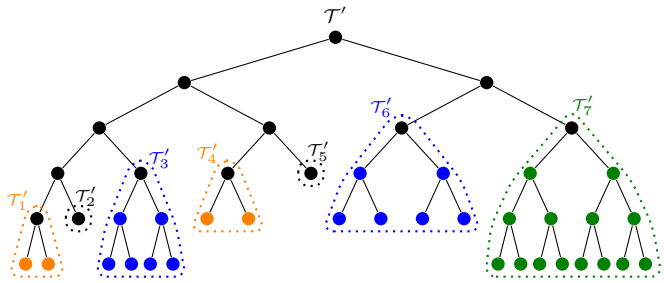

Figure 4: Example of an extended tree $\mathcal{T}'_{\text{ext}} = \mathcal{T}' \cup \mathcal{T}'_1 \cup \cdots \cup \mathcal{T}'_7$, formed by a subtree $\mathcal{T}'$ (black nodes) and its extensions (colored nodes). Each dotted set corresponds to a subtree $\mathcal{T}'_n$, rooted at a leaf $n \in \mathcal{L}(\mathcal{T}')$ and extended to depth h_n. The depths vary: $h_2 = h_5 = 0$ (black boxes), $h_1 = h_4 = 1$ (orange), $h_3 = h_6 = 2$ (blue), and $h_7 = 3$ (green). The leaves of $\mathcal{T}'_{\text{ext}}$ appear at different levels depending on the values of (h_n) and the level $\mathrm{l}(n)$ of the leaves $n \in \mathcal{L}(\mathcal{T}') = \{1, 2, 3, 4, 5, 6, 7\}$.

Observe that the total number of leaves in the extended pruning $\mathcal{T}'_{\text{ext}}$ is

$$|\mathcal{L}(\mathcal{T}'_{\text{ext}})| = \sum_{n\in\mathcal{L}(\mathcal{T}')} |\mathcal{L}(\mathcal{T}'_n)|. \tag{50}$$

Define

$$s_n := \min_{n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}})} s_{n'}, \quad n \in \mathcal{N}(\mathcal{T}'). \tag{51}$$

Remark also that by definition of the Besov norm in (6) (and via the usual embedding (p, q) fixed - see, e.g., [20]), one has for any $n \in \mathcal{L}(\mathcal{T}')$, $\|f\|_{s_n} \geqslant \|f\|_{s_{n'}}, n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}})$. In particular, every tree extension $\mathcal{T}'_n$ at node $n \in \mathcal{L}(\mathcal{T}')$ has $|\mathcal{L}(\mathcal{T}'_n)| = 2^{h_n d}$ leaves.

Case (ℓ_t) convex. Applying Theorem 3 in the convex case on the extended pruning $\mathcal{T}'_{\text{ext}}$, gives

$$R_T(f) \leqslant CG \sum_{n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}})} \left(B\sqrt{|T_{n'}|} + \|f\|_{s_{n'}} \cdot 2^{-\mathrm{l}(n')s_{n'}} \cdot |T_{n'}|^{r_{n'}} \right) \tag{52}$$

with C some constant that hides $\log T$ terms that can change from an inequality to another and $r_{n'} \in \{\frac{1}{2}, 1 - \frac{s_{n'}}{d}\}$ is the local rate described in Theorem 3. Note that by (51) one has $r'_n \leqslant r_n, n' \in \mathcal{L}(\mathcal{T}'_n)$. Recall that for every $n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}}), \mathrm{l}(n') = \mathrm{l}(n) + h_n$ for $n \in \mathcal{L}(\mathcal{T}')$ and since every leaves in $\mathcal{L}(\mathcal{T}'_n)$ is partitioning each terminal node $n \in \mathcal{T}'$, one has by Jensen's inequality:

$$\sum_{n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}})} \sqrt{|T_{n'}|} = \sum_{n \in \mathcal{L}(\mathcal{T}')} \sum_{n' \in \mathcal{L}(\mathcal{T}'_n)} \sqrt{|T_{n'}|} \leqslant \sum_{n \in \mathcal{L}(\mathcal{T}')} \sqrt{|\mathcal{L}(\mathcal{T}'_n)||T_n|}. \tag{53}$$

Then, by (51),(52) and (53) one gets (with $r_{n'} \leqslant r_n, n' \in \mathcal{L}(\mathcal{T}'_n)$)

$$\begin{aligned} R_T(f) &\leqslant CG \sum_{n \in \mathcal{L}(\mathcal{T}')} \sum_{n' \in \mathcal{L}(\mathcal{T}'_n)} \left(B\sqrt{|T_{n'}|} + \|f\|_{s_n} 2^{-(\mathrm{l}(n)+h_n)s_n} |T_{n'}|^{r_n} \right) \\ &\leqslant CG \sum_{n \in \mathcal{L}(\mathcal{T}')} \left(B\sqrt{|\mathcal{L}(\mathcal{T}'_n)||T_n|} + \|f\|_{s_n} 2^{-(\mathrm{l}(n)+h_n)s_n} \textstyle\sum_{n' \in \mathcal{L}(\mathcal{T}'_n)} |T_{n'}|^{r_n} \right). \end{aligned} \tag{54}$$

Further, applying Hölder's inequality over the sum over $n' \in \mathcal{L}(\mathcal{T}'_n)$ in (54) with $(1 - r_n) + r_n = 1$ (r_n is constant over the sum in n'):

$$\begin{aligned} R_T(f) &\leqslant CG \sum_{n \in \mathcal{L}(\mathcal{T}')} \left(B\sqrt{|\mathcal{L}(\mathcal{T}'_n)||T_n|} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{-h_n s_n} |\mathcal{L}(\mathcal{T}'_n)|^{1-r_n} \left(\textstyle\sum_{n' \in \mathcal{L}(\mathcal{T}'_n)} |T_{n'}| \right)^{r_n} \right) \\ &= CG \sum_{n \in \mathcal{L}(\mathcal{T}')} \left(B\sqrt{|T_n| 2^{h_n d}} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{dh_n(1 - \frac{s_n}{d} - r_n)} |T_n|^{r_n} \right) \end{aligned} \tag{55}$$

where we used

$$\sum_{n' \in \mathcal{L}(\mathcal{T}'_n)} |T_{n'}| = |T_n| \quad \text{and} \quad 2^{-h_n s_n} |\mathcal{L}(\mathcal{T}'_n)|^{1-r_n} = 2^{-h_n s_n} (2^{dh_n})^{1-r_n} = 2^{dh_n(1 - \frac{s_n}{d} - r_n)}.$$

Define the local regrets under the sum over $n \in \mathcal{N}(\mathcal{T}')$ by

$$R_n(f) := B\sqrt{|T_n| 2^{h_n d}} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{dh_n(1 - \frac{s_n}{d} - r_n)} |T_n|^{r_n},$$

that we now want to optimize in $h_n \in \mathbb{N}$. This leads to two different cases depending on the values of the local exponent r_n, defined in Theorem 3.

- *Case $s_n \geqslant \frac{d}{2}$ or $p < 2$: $r_n = \frac{1}{2}$*
  The local regret grows as:

  $$B\sqrt{|T_n| 2^{h_n d}} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{-h_n(s_n - \frac{d}{2})} \sqrt{|T_n|}.$$

  Therefore, setting $h_n = \max\left\{0, \left\lceil \frac{1}{s_n} \log_2 \left(2^{-\mathrm{l}(n)s_n} \|f\|_{s_n} B^{-1} \right) \right\rceil\right\}$ this entails

  $$R_n(f) \leqslant C \max\{B, B^{1 - \frac{d}{2s_n}} (2^{-\mathrm{l}(n)s_n} \|f\|_{s_n})^{\frac{d}{2s_n}}\} \sqrt{|T_n|}$$

- *Case $s_n < \frac{d}{2}$: $r_n = 1 - \frac{s_n}{d}$*
  The local regret grows as:

  $$R_n(f) = B\sqrt{|T_n| 2^{h_n d}} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} |T_n|^{1 - \frac{s_n}{d}},$$

  and the best choice is $h_n = 0$ in this case that entails

  $$R_n(f) = B\sqrt{|T_n|} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} |T_n|^{1 - \frac{s_n}{d}}$$

Finally, in the case (ℓ_t) are convex losses, we deduce that the regret is upper bounded as

$$\begin{aligned} R_T(f) &\leqslant CG \sum_{n\in\mathcal{L}(\mathcal{T}')} R_n(f) \\ &\leqslant CG \sum_{n\in\mathcal{L}(\mathcal{T}')} \Big(\big(\max\{B, B^{1-\frac{d}{2s_n}} (2^{-\mathrm{l}(n)s_n}\|f\|_{s_n})^{\frac{d}{2s_n}}\} \sqrt{|T_n|} \big) \mathbb{1}_{s_n \geqslant \frac{d}{2}} \\ &\qquad + \big(B\sqrt{T_n} + 2^{-\mathrm{l}(n)s_n}\|f\|_{s_n} |T_n|^{1-\frac{s_n}{d}} \big) \mathbb{1}_{s_n < \frac{d}{2}} \Big) \end{aligned} \tag{56}$$

Case (ℓ_t) exp-concave. Applying Theorem 3 in the exp-concave case on the extended pruning $\mathcal{T}'_{\text{ext}}$, gives

$$R_T(f) \leqslant CG \Big(B|\mathcal{L}(\mathcal{T}'_{\text{ext}})| + \sum_{n'\in\mathcal{L}(\mathcal{T}'_{\text{ext}})} \|f\|_{s_{n'}} \cdot 2^{-\mathrm{l}(n')s_{n'}} \cdot |T_{n'}|^{r_{n'}} \Big) \tag{57}$$

with C some constant that hides $\log T$ terms and that can change from an inequality to another and $r_{n'} \in \{\frac{1}{2}, 1-\frac{s_{n'}}{d}\}$ is the local rate described in Theorem 3. Note that $|\mathcal{L}(\mathcal{T}'_{\text{ext}})| = \sum_{n\in\mathcal{L}(\mathcal{T}')} |\mathcal{L}(\mathcal{T}'_n)|$ and again for every $n' \in \mathcal{L}(\mathcal{T}'_{\text{ext}}), \mathrm{l}(n') = \mathrm{l}(n) + h_n$ for $n \in \mathcal{L}(\mathcal{T}')$ and $r_{n'} \leqslant r_n, n' \in \mathcal{L}(\mathcal{T}'_n)$. We get

$$R_T(f) \leqslant CG \sum_{n\in\mathcal{L}(\mathcal{T}')} \Big(B|\mathcal{L}(\mathcal{T}'_n)| + \|f\|_{s_n} 2^{-(\mathrm{l}(n)+h_n)s_n} \textstyle\sum_{n'\in\mathcal{L}(\mathcal{T}'_n)} |T_{n'}|^{r_n} \Big). \tag{58}$$

Using $|\mathcal{L}(\mathcal{T}'_n)| = 2^{h_n d}$ and applying Hölder's inequality over the sum over $n' \in \mathcal{L}(\mathcal{T}'_n)$ in (58) with $(1-r_n) + r_n = 1$ as in (55) entails

$$R_T(f) \leqslant CG \sum_{n\in\mathcal{L}(\mathcal{T}')} \big(B2^{h_n d} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{dh_n(1-\frac{s_n}{d}-r_n)} |T_n|^{r_n} \big).$$

Again, we define the local regrets under the sum over $n \in \mathcal{N}(\mathcal{T}')$ as

$$R_n(f) := B2^{h_n d} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{dh_n(1-\frac{s_n}{d}-r_n)} |T_n|^{r_n},$$

that we optimize in $h_n \in \mathbb{N}$. The cases are the same as for the convex case, according to the values of the local exponent r_n, defined in Theorem 3.

- *Case $s_n \geqslant \frac{d}{2}$ or $p < 2$: $r_n = \frac{1}{2}$*
  The local regret grows as
  $$R_n(f) = B2^{h_n d} + \|f\|_{s_n} 2^{-\mathrm{l}(n)s_n} 2^{-h_n(s_n-\frac{d}{2})} \sqrt{|T_n|}.$$
  Afterwards, optimizing in h_n such that
  $$B2^{h_n d} = 2^{-\mathrm{l}(n)s_n} \|f\|_{s_n} 2^{-h_n(s_n-\frac{d}{2})} \sqrt{|T_n|}$$
  leads to $h_n = \max\big\{0, \lceil \frac{1}{2s_n+d} \log_2 \big((B^{-1} 2^{-\mathrm{l}(n)s_n} \|f\|_{s_n})^2 |T_n| \big) \rceil \big\}$, that entails
  $$R_n(f) \leqslant C \max\left\{ B, B^{1-\frac{2d}{2s_n+d}} \big(2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}\big)^{\frac{2d}{2s_n+d}} |T_n|^{\frac{d}{2s_n+d}} \right\}$$
- *Case $s_n < \frac{d}{2}$: $r_n = 1 - \frac{s_n}{d}$*
  The local regret grows as
  $$R_n(f) = B2^{h_n d} + 2^{-\mathrm{l}(n)s_n} \|f\|_{s_n} |T_n|^{1-\frac{s_n}{d}},$$
  and the best choice is $h_n = 0$ which entails
  $$R_n(f) = B + 2^{-\mathrm{l}(n)s_n} \|f\|_{s_n} |T_n|^{1-\frac{s_n}{d}},$$

Finally, with (ℓ_t) exp-concave losses, the regret is bounded as

$$\begin{aligned} R_T(f) &\leqslant CG \sum_{n\in\mathcal{L}(\mathcal{T}')} R_n(f) \\ &\leqslant CG \sum_{n\in\mathcal{L}(\mathcal{T}')} \Big(\max\big\{B, B^{1-\frac{2d}{2s_n+d}} \big(2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}\big)^{\frac{2d}{2s_n+d}} |T_n|^{\frac{d}{2s_n+d}} \big\} \mathbb{1}_{s_n \geqslant \frac{d}{2}} \qquad (59) \\ &\qquad + \big(B + 2^{-\mathrm{l}(n)s_n} \|f\|_{s_n} |T_n|^{1-\frac{s_n}{d}} \big) \mathbb{1}_{s_n < \frac{d}{2}} \Big). \qquad (60) \end{aligned}$$

Remark. Taking $\mathcal{T}'$ as the pruning associated to the root, this entails $O(T^{\frac{d}{2s+d}}) = O(T^{1-\frac{2s}{2s+d}})$ which is minimax-optimal for this case - see [35].

□

D Discussion on the unbounded case: $s < \frac{d}{p}$

As in most previous works in statistical learning, this paper primarily considers competitive functions $f \in B^s_{pq}(\mathcal{X})$ with $s > \frac{d}{p}$, which ensures that $f \in L^\infty(\mathcal{X})$ with $\|f\|_\infty < \infty$.
A natural question is whether our Algorithm 1 remains competitive - that is, achieves sublinear regret - in the more challenging regime where $s < \frac{d}{p}$. Indeed, in the case $s < \frac{d}{p}$, prediction rules may no longer be bounded in sup-norm. For example, the function $f(x) = x^{-1/2}\mathbf{1}_{x\in(0,1]}$ belongs to $L^p([0,1])$ for $p < 2$ but not to $L^\infty([0,1])$, illustrating the type of singularity permitted when $s < \frac{d}{p}$. In such cases, the boundedness condition $\|f_J - f\|_\infty < \infty$ required in (24) may fail, where f_J denotes the truncated J-level wavelet expansion defined in (2). Nevertheless, we discuss how Algorithm 1 can still offer performance guarantees in certain settings, particularly when the input data $\{x_t\}$ are well distributed over $\mathcal{X}$.
Indeed, by Hölder's inequality, (24) can be upper bounded as:

$$\sum_{t=1}^{T} |f_J(x_t) - f(x_t)| \leqslant T\left(\frac{1}{T}\sum_{t=1}^{T} |f_J(x_t) - f(x_t)|^p\right)^{\frac{1}{p}}, \tag{61}$$

where the sum on the right-hand side defines an empirical ℓ^p semi-metric over the input data $\{x_t\}_{t=1}^T$, denoted:

$$d_T^p(f_J, f) := \left(\frac{1}{T}\sum_{t=1}^{T} |f_J(x_t) - f(x_t)|^p\right)^{\frac{1}{p}}.$$

The upper bound (61) suggests that tighter control may be obtained by focusing on the empirical norm $d_T^p(f_J, f)$ rather than on the sup-norm, which may not be finite.

First case: the empirical semi-norm d_T^p approximates the L^p norm. Assume that the semi-norm $d_T^p(f_J, f)$ is close to the true L^p norm $\|f_J - f\|_{L^p}$. Such an equivalence is expected when the data $\{x_t\}$ are well distributed over $\mathcal{X}$, for example when $x_t \sim \mathcal{U}(\mathcal{X})$ i.i.d., or when x_t are equally spaced, such as $x_t = \frac{t}{T}$ for $t = 1, \dots, T$. By the law of large numbers or standard concentration arguments, one typically has $d_T^p(f_J, f) \approx \|f_J - f\|_{L^p}$ in expectation or with high probability.
Classical approximation results (e.g., [20, Prop. 4.3.8]) then yield $\|f_J - f\|_{L^p} \lesssim 2^{-Js}$ for $f \in B^s_{pq}(\mathcal{X}) \subset L^p(\mathcal{X})$. Optimizing over J to balance estimation and approximation regrets leads to a regret bound of $O(T^{1-\frac{s}{d}})$ - see the proof of Theorem A, last case $\beta < 0$. This regret is sublinear as soon as $s > 0$ and becomes linear when $s = 0$, as is typical for $f \in L^p$.
Nevertheless, minimax analysis from [34, 35] shows that a regret of $O(T^{1-1/p})$ is possible, which improves upon our bound whenever $s < \frac{d}{p}$. Whether a constructive algorithm achieving this minimax regret exists in the regime $s < \frac{d}{p}$ remains, to the best of our knowledge, an open and interesting question.

Second case: the semi-norm d_T^p fails to approximate the L^p norm. If the data points $\{x_t\}$ are concentrated near singularities (e.g., near 0 in the example above), the empirical norm $d_T^p(f_J, f)$ can differ significantly from the true norm $\|f_J - f\|_{L^p}$, making the latter less informative in practice.
In such adversarial or non-uniform settings, it seems preferable to control the empirical norm $d_T^p(f_J, f)$ directly, as it more accurately reflects the distribution of the observed data. Addressing this challenge may require adaptive sampling strategies, localization techniques, or alternative norms that account for the geometry or density of the input distribution.

E Review of multi-resolution analysis

In this section we present some of the basic ingredients of wavelet theory. Let's assume we have a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$.

Definition 1 (Scaling function)**.** *We say that a function $\phi \in L^2(\mathbb{R}^d)$ is the scaling function of a multiresolution analysis (MRA) if it satisfies the following conditions:*

1. *the family*
$$\{x \mapsto \phi(x-k) = \textstyle\prod_{i=1}^d \phi(x_i - n_i) : k \in \mathbb{Z}^d\}$$
is an ortho-normal basis, that is $\langle \phi(\cdot - k), \phi(\cdot - n)\rangle = \delta_{k,n}$;
2. *the linear spaces*
$$V_0 = \left\{f = \textstyle\sum_{k\in\mathbb{Z}^d} c_k \phi(\cdot - k), (c_k) : \sum_{k\in\mathbb{Z}^d} c_k^2 < \infty\right\}, \ldots, V_j = \{h = f(2^j \cdot) : f \in V_0\}, \ldots,$$
are nested, i.e. $V_{j-1} \subset V_j$ for all $j \geqslant 0$.

We note that under these two conditions, it is immediate that the functions

$$\{\phi_{j,k} = 2^{dj/2}\phi(2^j \cdot -k)\,, k \in \mathbb{Z}^d\}$$

form an ortho-normal basis of the space $V_j, j \in \mathbb{N}$. One can define the projection kernel of f over V_j (from here we also say kernel projection at scale or level j) as

$$K_j f(x) := \sum_{k\in\mathbb{Z}^d} \langle f, \phi_{j,k}\rangle \phi_{j,k}(x) = \int_{\mathbb{R}^d} K_j(x,y) f(y)\, dy\,, \tag{62}$$

with $K_j(x,y) = \sum_{n\in\mathbb{Z}^d} \phi_{j,k}(x)\phi_{j,k}(y) = \sum_{k\in\mathbb{Z}^d} 2^{dj}\phi(2^j x - n)\phi(2^j y - n)$ (which is not of convolution type) but has comparable approximation properties that we detail after.

Incremental construction via wavelets. Since the spaces (V_j) are nested, one can define nontrivial subspaces as the orthogonal complements $W_j := V_{j+1} \ominus V_j$. We can then telescope these orthogonal complements to see that each space $V_j, j \geqslant j_0$ can be written as

$$V_j = V_{j_0} \oplus \left(\bigoplus_{l=j_0}^{j} W_l\right) \quad \text{for any } j_0 \in \mathbb{N}.$$

Let ψ be a mother wavelet corresponding to the scaling function ϕ. The associated wavelets are defined as follows: for $E = \{0,1\}^d \setminus \{0\}$, we set

$$\psi^\varepsilon(x) = \psi^{\varepsilon_1}(x_1)\cdots\psi^{\varepsilon_d}(x_d), \quad \psi^\varepsilon_{j,n} = 2^{jd/2}\psi^\varepsilon(2^j x - n), \quad j \geqslant 0, \quad n \in \mathbb{Z}^d,$$

where $\psi^0 = \phi$, $\psi^1 = \psi$. For each j, these functions form an orthonormal basis of W_j.
Analogously, one can now observe that for every $j \geqslant j_0$,

$$K_j f = K_{j_0} f + \sum_{l=j_0}^{j-1} (K_{l+1} f - K_l f)\,, \tag{63}$$

where each increment in the sum can be written as

$$K_{j+1} f - K_j f = \sum_k \sum_\varepsilon \langle f, \psi^\varepsilon_{j,k}\rangle \psi^\varepsilon_{j,k},$$

where for each $j \geqslant 1$, the set

$$\left\{\psi^e_{j,k} = 2^{dj/2}\psi^\varepsilon(2^j x - k) : \varepsilon \in E,\, k \in \mathbb{Z}^d\right\}$$

forms a basis of W_j for some wavelet ψ, with $E := \{0,1\}^d \setminus \{0\}$. For simplicity, we include the index ε in the multi-index k. Finally, the set $\{\phi_{j_0,k}, \psi_{j,k}\}$ constitutes a *wavelet basis*.
For our results we will not be needing a particular wavelet basis, but any that satisfies the following key properties.

Definition 2 (S-regular wavelet basis)**.** *Let $S \in \mathbb{N}^*$ and $j_0 = 0$. The multiresolution wavelet basis*

$$\{\phi_k = \phi(\cdot - k), \psi_{j,k} = 2^{jd/2}\psi(2^d \cdot -k)\}$$

of $L^2(\mathbb{R}^d)$ with associated projection kernel $K(x,y) = \sum_k \phi_k(x)\phi_k(y)$ is said to be S-regular if the following conditions are satisfied:

(D.1) ***Vanishing moments and normalization:***

$$\int_{\mathbb{R}^d} \psi(x)\, x^\alpha \, dx = 0 \quad \textit{for all multi-indices } \alpha \textit{ with } |\alpha| < S, \qquad \int_{\mathbb{R}^d} \phi(x)\, dx = 1.$$

Moreover, for all $v \in \mathbb{R}^d$ *and* α *with* $1 \leqslant |\alpha| < S$,

$$\int_{\mathbb{R}^d} K(v, v+u)\, du = 1, \quad \int_{\mathbb{R}^d} K(v, v+u)\, u^\alpha \, du = 0.$$

(D.2) ***Bounded basis sums:***

$$M_\phi := \sup_{x\in\mathbb{R}^d} \sum_k |\phi(x-k)| < \infty, \qquad M_\psi := \sup_{x\in\mathbb{R}^d} \sum_k |\psi(x-k)| < \infty.$$

(D.3) ***Kernel decay:*** *For* $\kappa(x, y)$ *equal to* $K(x, y)$ *or* $\sum_k \psi(x-k)\psi(y-k)$, *there exist constants* $c_1, c_2 > 0$ *and a bounded integrable function* $\phi : [0, \infty) \to \mathbb{R}$ *such that*

$$\sup_{v\in\mathbb{R}^d} |\kappa(v, v-u)| \leqslant c_1 \phi(c_2\|u\|), \qquad C_S := \int_{\mathbb{R}^d} \|u\|^S \phi(\|u\|)\, du < \infty.$$

Case of a bounded compact $\mathcal{X} \subset \mathbb{R}^d$. The above definition applies to wavelet systems on $\mathbb{R}^d$, but can be extended to compact domains $\mathcal{X} \subset \mathbb{R}^d$ using standard boundary-corrected or periodized constructions. Notable examples include the compactly supported orthonormal wavelets of Daubechies [11, Chapter 7] and the biorthogonal, symmetric, and highly regular wavelet bases of Cohen et al. [9]. Just as in the case of $\mathbb{R}^d$, we can build a tensor-product wavelet basis $\{\phi_k, \psi_{j,k}\}$, for example using periodic or boundary-corrected Daubechies wavelets. At the j-th level, there are now $O(2^{jd})$ wavelets $\psi_{j,k}$, which we index by $k \in \Lambda_j$, the set of indices corresponding to wavelets at level j. This coincides with the expansion used in Equation (2).

Control of wavelet coefficients and characterization of Hölder spaces. Remarkably, the norm of the space $\mathscr{C}^s(\mathcal{X})$ has a useful characterisation by wavelet bases - see [31] or [20] for a review on the characterisation of smoothness according to wavelet basis.

Proposition 1. *Let* $s > 0$ *we thus have the following:*

$$f \in \mathscr{C}^s(\mathcal{X}) \implies \sup_k |\langle f, \psi_{j,k}\rangle| \leqslant C|f|_s 2^{-j(s+d/2)}, \tag{64}$$

where $C = C(\psi, S)$ *is some constant that depends only on the (*S*-regular) wavelet basis.*

Proof. Let ψ be a compactly supported wavelet in $\mathbb{R}^d$ with S vanishing moments, i.e.,

$$\int_{\mathbb{R}^d} x^\beta \psi(x)\, dx = 0 \quad \text{for all multi-indices } \beta \text{ with } |\beta| < S.$$

Assume that $f \in \mathscr{C}^s(\mathbb{R}^d)$ for some $s > 0$, with $s < S$, so the wavelet vanishing moments match the regularity of f. Let $\psi_{j,k}(x) := 2^{jd/2}\psi(2^j x - k)$ be the wavelet at scale j and location $k \in \mathbb{Z}^d$. The wavelet coefficient is given by

$$c_{j,k} := \langle f, \psi_{j,k}\rangle = \int_{\mathbb{R}^d} f(x)\psi_{j,k}(x)\, dx.$$

We define the center of the wavelet support as $x_{j,k} := 2^{-j}k$ and write a Taylor expansion of f at $x_{j,k}$:

$$f(x) = P_{x_{j,k}}(x) + R_{x_{j,k}}(x),$$

where $P_{x_{j,k}}$ is the Taylor polynomial of degree $\lfloor s \rfloor$ and $|R_{x_{j,k}}(x)| \leqslant |f|_s \|x - x_{j,k}\|_\infty^s$ for x near $x_{j,k}$ and where $|f|_s = \sup_{|m|=\lfloor s \rfloor} \|D^m f\|_{s-|m|}$.
Using the vanishing moments of ψ, we have

$$c_{j,k} = \int_{\mathbb{R}^d} R_{x_{j,k}}(x)\psi_{j,k}(x)\, dx.$$

Now perform the change of variables $u = 2^j x - k$, so $x = 2^{-j}u + x_{j,k}$ and $dx = 2^{-jd}du$:

$$c_{j,k} = 2^{-jd/2} \int_{\mathbb{R}^d} R_{x_{j,k}}(x_{j,k} + 2^{-j}u)\psi(u)\, du.$$

By the Hölder remainder estimate, we have

$$|R_{x_{j,k}}(x_{j,k} + 2^{-j}u)| \leqslant |f|_s \|2^{-j}u\|^s = |f|_s 2^{-js}\|u\|^s.$$

Therefore,

$$|c_{j,k}| \leqslant |f|_s 2^{-j(s+d/2)} \int_{\mathbb{R}^d} |\psi(u)| \|u\|^s\, du,$$

and since ψ is compactly supported and smooth, the integral is finite. Hence, defining $C(\psi, s) = \int_{\mathbb{R}^d} |\psi(u)|\|u\|^s\, du < \infty$ we get the result. □

Remark. The smoothness s of f translates into faster decay of the coefficients given sufficiently ($S > s$) regular wavelets.

F Summary of the results and comparison to the literature

Table 2: Regret rates, parameter requirements and time complexity for online regression algorithm with (ℓ_t) square losses and $s > d/p$.

Paper		Setting	Input Parameters	Regret Rate	Complexity
Vovk [39]		$f \in B^s_{pq},\ p, q \geqslant 1$	$s, p, B \geqslant \|f\|_\infty$	$T^{1-\frac{s}{s+d}}$	$\exp(T) + Td$
Vovk [40]		$f \in B^s_{pq},\ p \geqslant 2,\ q \in [\frac{p}{p-1}, p]$ $f \in \mathscr{C}^s,\ p = \infty,\ s \geqslant \frac{d}{2}$	$s, p, B \geqslant \|f\|_\infty$	$T^{1-\frac{1}{p}}$ $T^{1-\frac{s}{d}+\varepsilon}$	Not feasible
Gaillard and Gerchinovitz [17]		$f \in W^s_p,\ p \geqslant 2,\ s \geqslant \frac{d}{2}$ $f \in W^s_p,\ p > 2,\ s < \frac{d}{2}$ $f \in \mathscr{C}^s,\ p = \infty,\ d = 1,\ s > \frac{1}{2}$	$s, p, B \geqslant \|f\|_\infty$	$T^{1-\frac{2s}{2s+d}}$ $T^{1-\frac{s}{d}}$ $T^{1-\frac{2s}{2s+1}}$	$\exp(dT)$ $\exp(dT)$ $\mathrm{poly}(T)$
Liautaud et al. [29]		$f \in \mathscr{C}^s,\ p = \infty,\ s \in (1/2, 1],\ d = 1$	$B \geqslant \|f\|_\infty$	$T^{1-\frac{2s}{2s+1}}$	$\mathrm{poly}(T)$
Zadorozhnyi et al. [42]		$f \in W^s_p,\ p \geqslant 2,\ s \geqslant \frac{d}{2}$ $f \in W^s_p,\ p > 2,\ s < \frac{d}{2}$	s, p	$T^{1-\frac{2s}{2s+d}+\varepsilon}$ $T^{1-\frac{s}{d}\frac{p-d/s}{p-2}+\varepsilon}$	$\mathrm{poly}(T)d$
This work	Alg. 1	$f \in B^s_{pq},\ p, q \geqslant 1,\ s \geqslant \frac{d}{2}$ or $p \leqslant 2$ $f \in B^s_{pq},\ p > 2,\ q \geqslant 1,\ s < \frac{d}{2}$	$S \geqslant s, \varepsilon < s - \frac{d}{p}$	$\sqrt{T}$ $T^{1-\frac{s}{d}}$	$\mathrm{poly}(T)S^d$
	Alg. 2	$f \in B^s_{pq},\ p, q \geqslant 1,\ s \geqslant \frac{d}{2}$ or $p \leqslant 2$ $f \in B^s_{pq},\ p > 2,\ q \geqslant 1,\ s < \frac{d}{2}$	$S \geqslant s, \varepsilon < s - \frac{d}{p}, B \geqslant \|f\|_\infty$	$T^{1-\frac{2s}{2s+d}}$ $T^{1-\frac{s}{d}}$	$\mathrm{poly}(T)S^d$
Minimax rates Rakhlin and Sridharan [34, 35]		$f \in B^s_{pq},\ p, q \geqslant 1,\ s \geqslant \frac{d}{2}$ $f \in B^s_{pq},\ p > 2,\ q \geqslant 1,\ s < \frac{d}{2}$	Non constructive	$T^{1-\frac{2s}{2s+d}}$ $T^{1-\frac{s}{d}}$	Non constructive

Comparison to Vovk [40]. Vovk [40] provide a general analysis for prediction in Banach spaces, focusing on the regime $s > d/p$. They achieve regret rates of $O(T^{1-1/p})$ for certain Besov spaces B^s_{pq} with $p \geqslant 2$ and $q \in [p/(p-1), p]$. These rates are independent of the smoothness parameter s, except in the case $p = \infty$, where they obtain $O(T^{1-\frac{s}{d}})$. However, this remains suboptimal in their setting with square loss. In contrast, our analysis yields the minimax-optimal rate $O(T^{1-\frac{2s}{2s+d}})$ over a broader class of Besov spaces B^s_{pq} with arbitrary $p, q \in [1, \infty]$ and $s > d/p$.

Comparison to Vovk [39]. Vovk [39] investigates prediction under general metric entropy conditions, proposing algorithms that compete with a reference class of functions in terms of covering numbers. While their approach is highly general and applies to a broad range of normed spaces, the regret bounds they derive, of order $O(T^{1-\frac{s}{s+d}})$, still do not match the minimax-optimal rates known for functions in B^s_{pq}.

Comparison to Zadorozhnyi et al. [42]. Their approach focuses on Sobolev spaces $W^s_p(\mathcal{X})$ with $p \geqslant 2$ and $s > \frac{d}{p}$, and they obtain suboptimal rates, in the regime $s < \frac{d}{2}$, of $O(T^{1-\frac{s}{d}\cdot\frac{p-d/s}{p-2}+\varepsilon})$, for arbitrarily small ε. In comparison, our rates $O(T^{1-\frac{2s}{2s+d}})$ are minimax-optimal over a broader class of Besov spaces B^s_{pq} with arbitrary s, p, q satisfying $s > \frac{d}{p}$, which include the Sobolev balls considered in their work.

Computational complexity. Most existing work in online nonparametric regression over Besov spaces (including Sobolev spaces), such as Rakhlin and Sridharan [34, 35], Vovk [39, 40], does not provide efficient (i.e., polynomial-time) algorithms. The work by Rakhlin and Sridharan [34, 35] offers a minimax-optimal analysis, but does not yield constructive procedures - computing the offset Rademacher complexity, as required by their method, is numerically infeasible in practice. The approach of using the Exponentiated Weighted Average (EWA) algorithm in nonparametric settings, as proposed by Vovk [39], suffers from both suboptimal regret rates and prohibitive computational complexity, since it requires updating the weights of each expert in a covering net, leading to a total cost of $O(\exp(T))$. Vovk [40] introduce the defensive forecasting approach, which also avoids efficient implementation as it relies on the so-called Banach feature map - a representation that is typically inaccessible or intractable in practice. The Chaining EWA forecaster of Gaillard and Gerchinovitz [17] achieves optimal regret bounds in the online nonparametric setting. However, its algorithm is provably polynomial-time only in the case $p = \infty$ and $d = 1$; in general dimensions and p, its direct implementation requires $O(\exp(dT))$ operations. Zadorozhnyi et al. [42] propose an efficient algorithm with total computational complexity of order $O(T^3 + dT^2)$. We note that their algorithm has a linear cost in d, making it particularly suitable for high-dimensional settings with smooth competitors in $W^s_p(\mathcal{X})$ (with $s > \frac{d}{2}$).

Finally, our algorithms are both optimal and efficient, with computational costs (after T rounds) of

$$O(T\times J\times S^d) = O\left(T\frac{S}{d\varepsilon}\log_2(T)S^d\right) \quad \text{and} \quad O(T\times|\mathcal{A}|^\lambda\times J_0\times J\times S^d) = O\left(T^{1+\frac{\lambda}{2}}\frac{S^2}{d^2\varepsilon^2}\log_2(T)^2S^d\right)$$

for Algorithm 1 and Algorithm 2 respectively (taking a partitioning tree of maximum depth $J_0 = \lceil\frac{S}{d\varepsilon}\log_2 T\rceil$).

G Besov embeddings in usual functional spaces

We refer to [8, 13, 20] for precise statements of the classical embedding theorems. For convenience, we recall some of the most useful embeddings in Table 3.

Table 3: Classical embeddings of Besov spaces B^s_{pq}

Condition on (s,p,q)	**Target Space**	**Embedding Type**
$s > \frac{d}{p}$	L^∞	Continuous embedding
$s = \frac{d}{p}$, $q = 1$	L^∞	Critical embedding
$s > d\left(\frac{1}{p} - \frac{1}{r}\right)$, $p < r$	L^r	Continuous embedding
$s_1 > s_2$	$B^{s_2}_{pq}$	Continuous embedding
$s = s$, $p_1 \leqslant p_2$, $q_1 \leqslant q_2$	$B^s_{p_2q_2}$	Continuous embedding
B^s_{pp}	W^s_p	Equivalence (for $s \in \mathbb{N}$)
$B^s_{\infty\infty}$	$\mathscr{C}^s$	Norm equivalence with Hölder

H Summary of optimal regret in Online Nonparametric Regression

This section summarizes the results in [34, 35] for mimimax-optimal rate of regret in the adversarial online nonparametric regression setting.

Proposition 2 ([35]). *Assume the sequential entropy at scale $\varepsilon > 0$ is $O(\varepsilon^{-\alpha}), \alpha > 0$ for the target class function. Optimal regret is then summarized in the table:*

Table 4: Optimal regret for different loss functions

Loss Function	**Range on** α	**Optimal Regret**
Absolute loss	$\alpha \in (0,2]$	$T^{\frac{1}{2}}$
	$\alpha > 2$	$T^{1-\frac{1}{\alpha}}$
Square loss	$\alpha \in (0,2]$	$T^{1-\frac{2}{2+\alpha}}$
	$\alpha > 2$	$T^{1-\frac{1}{\alpha}}$

In particular, for Hölder functions $\mathscr{C}^s(\mathcal{X}), s > 0$, and $B^s_{pq}(\mathcal{X}), s > \frac{d}{p}$ one has $\alpha = \frac{d}{s}$.