High-Probability Minimax Adaptive Estimation in Besov Spaces via Online-to-Batch

Paul Liautaud
Sorbonne Université, CNRS, LPSM
F-75005 Paris, France
paul.liautaud@sorbonne-universite.fr

Pierre Gaillard
Inria, École Normale Supérieure
PSL Research University
Paris, France
pierre.gaillard@inria.fr

Olivier Wintenberger
Sorbonne Université, CNRS, LPSM
F-75005 Paris, France
olivier.wintenberger@sorbonne-universite.fr

Abstract

We study nonparametric regression over Besov spaces from noisy observations under sub-exponential noise. Our goal is to obtain minimax-optimal high-probability bounds for the integrated squared error while adapting to the unknown noise level and the regularity parameters of the underlying Besov class. We introduce a wavelet-based online learning algorithm that sequentially processes noisy gradients and adapts to the gradient noise through an adaptive clipping rule, thus avoiding the need to tune parameters such as the noise variance or gradient bounds. As a by-product of our analysis, we derive high-probability adaptive regret bounds that scale with the ℓ_1-norm of the competitor. Finally, in the batch statistical setting, our method is the first to achieve high-probability minimax-optimal estimation rates over Besov spaces while adapting to all problem parameters, including the noise level. Our approach relies on a refined online-to-batch conversion and exploits the structure of the squared loss in combination with self-normalized concentration inequalities.

1 Introduction

A fundamental challenge in statistics and machine learning is the estimation of an unknown function f from noisy observations. Given a sample of $T \geqslant 1$ i.i.d. data points $(X_t, Y_t)_{t=1}^T$ following the regression model

$$Y_t = f(X_t) + \varepsilon_t, \quad t = 1, \dots, T,$$

where ε_t denotes a sub-exponential zero-mean noise, the goal is to construct an estimator $\bar{f}_T$ of the true function $f : \mathcal{X} \to \mathbb{R}$, supported on a compact domain $\mathcal{X} \subset \mathbb{R}^d$, $d \geqslant 1$. The quality of an estimator $\bar{f}_T$ is measured by the integrated squared error $\|\bar{f}_T - f\|_{L^2(\mathbb{P}_X)}$. When f is assumed to belong to a Besov space $\mathcal{B}^s_{pq}$ (see Section 3), the landmark work of [16] established the minimax-optimal rates for the mean integrated squared error, which are attained by wavelet shrinkage methods.

While this classical theory focuses on optimality in expectation, many modern learning applications—ranging from safety-critical systems to precision medicine—require guarantees that hold with *high probability*. This motivates the study of estimators that achieve minimax-optimal performance beyond the average-case regime. Our objective is therefore to design an estimator $\bar{f}_T$ that satisfies, with high probability, a minimax-adaptive risk bound of the form

$$\|\bar{f}_T - f\|_2^2 = O\left(\|f\|_{\mathcal{B}}^{\frac{2d}{2s+d}} (\sigma^2)^{\frac{2s}{2s+d}} T^{-\frac{2s}{2s+d}} \right), \tag{1}$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

for any continuous function $f \in \mathcal{B}^s_{pq}$, without prior knowledge of the Besov parameters (s, p, q), the Besov norm $\|f\|_{\mathcal{B}}$, or the noise variance σ^2. This bound is said to be *noise-level–aware* since it exhibits minimax-optimal dependence on the noise level σ. Importantly, unlike classical minimax analyses [e.g., 11, 16], which characterize rates asymptotically as $T \to \infty$ under known fixed noise level, our result adapts to the variance σ^2 and matches the noise-level-aware minimax rates recently derived in [12]. Achieving such guarantees simultaneously with high probability and full adaptivity remains challenging for standard wavelet shrinkage and aggregation techniques.

To address this challenge, we turn to the framework of *online learning*. Online algorithms are natural candidates for adaptive estimation, as they are designed to control regret without prior knowledge of the data-generating process. By defining the estimator $\bar{f}_T$ as the average of the online iterates, $\bar{f}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{f}_t$, one can relate the *excess risk* to the *stochastic regret* via Jensen's inequality:

$$\begin{aligned} \|\bar{f}_T - f\|_2^2 &= \mathbb{E}_{XY}[(\bar{f}_T(X) - Y)^2] - \mathbb{E}_{XY}[(f(X) - Y)^2] \\ &\leqslant \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{t-1}[(\hat{f}_t(X_t) - Y_t)^2] - \mathbb{E}_{t-1}[(f(X_t) - Y_t)^2]\,, \end{aligned} \tag{2}$$

where $\mathbb{E}_{t-1}$ denotes the expectation conditioned on the first $t-1$ samples and since $(X_t, Y_t), t \geqslant 1$ are i.i.d. This *online-to-batch conversion* principle [26, 4, 33] ensures that any regret bound for convex losses yields an equivalent excess risk bound *in expectation*. However, despite its popularity, standard online-to-batch conversion typically fails to provide sharp *high-probability* risk guarantees and adaptations of the online algorithms are required [37, 35]. While the expected regret may be optimal, fluctuations along the regret trajectory can lead to suboptimal concentration behavior when translated to the batch setting. This issue is particularly severe in nonparametric regimes, where the batch minimax rates can be as fast as $T^{-2s/(2s+d)}$, which is strictly faster than the online minimax rates $T^{-1/2}$ when $s \geqslant d/2$. In this regime, standard first-order concentration inequalities are insufficient to control these fluctuations at the optimal rate. Recently, [35] proposed a general reduction for deriving high-probability risk bounds from online learning guarantees. However, this approach is not sufficient to attain minimax-optimal rates in the low-regularity regime $s < d/2$: standard online-to-batch conversions based on known online regret bounds yield the slower rate $O(T^{-s/d})$ [31], whereas the optimal batch risk rate is $O(T^{-2s/(2s+d)})$. The central challenge is therefore to design online algorithms $(\hat{f}_t)$ whose adaptive nonparametric regret guarantees in (2) can be converted into minimax-optimal *high-probability* risk bounds of the form (1). Our analysis, inspired by [38], provides such a refined online-to-batch conversion by exploiting stochastic curvature and self-normalized concentration. Unlike standard batch analyses of nonparametric estimators, our proof derives minimax risk bounds from online regret guarantees via a refined online-to-batch conversion, which constitutes an interesting new route to minimax nonparametric rates.

Recently, [24] introduced a wavelet-based online algorithm that achieves adaptive minimax-optimal regret in the adversarial setting. Their approach relies in particular on so-called *comparator-adaptive* online algorithms, which guarantee regret bounds of order $\tilde{O}(\|\mathbf{c}\|_1 G\sqrt{T})$ for G-Lipschitz losses against any comparator $\mathbf{c} \in \mathbb{R}^N, N \geqslant 1$. However, in the unbounded stochastic setting considered here, the loss functions are no longer Lipschitz, and such guarantees can fail dramatically. To overcome this difficulty, we employ a *gradient clipping* strategy [39] with an adaptive online threshold. This modification allows us to recover high-probability control of the regret while preserving comparator adaptivity. As a by-product of our analysis, we obtain new results for stochastic online convex optimization, establishing regret bounds of the form $\tilde{O}(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T})$ holding with high probability for any comparator $\mathbf{c} \in \mathbb{R}^N$ with $\|\mathbf{c}\|_\infty < \infty$ and with unknown noise level σ, answering an open question in [39].

Notations. The notation $O(\cdot)$ hides universal constant factors, while $\tilde{O}(\cdot)$ additionally hides polylogarithmic factors. For any integer $k \geqslant 1$, we denote $[k] := \{1, \dots, k\}$. Boldface letters are used to denote multivariate quantities, whereas standard (non-bold) letters denote scalar quantities. We denote $a \vee b = \max\{a, b\}$ and, for any $x \in \mathbb{R}$ and threshold $\tau > 0$, the clipping operator of x in $[-\tau, \tau]$ is defined by $\mathrm{clip}(x, \tau) = x$ if $|x| \leqslant \tau$ and $\mathrm{sign}(x)\tau$ otherwise.

Contributions and outline of the paper. In Section 2, we first prove high-probability, comparator-adaptive regret bounds of order $\tilde{O}(\|\mathbf{c}\|_1\sqrt{T})$ for stochastic online convex optimization on $\|\mathbf{c}\|_\infty \leqslant C$, under sub-exponential gradient noise. In Section 2.2, we remove the need to know the noise level by introducing an adaptive gradient clipping procedure, which may be useful beyond the present nonparametric setting. Finally, in Section 3, we combine these tools with wavelet aggregation to

obtain a fully adaptive nonparametric estimator with high-probability, minimax, noise-level-aware risk bounds, without prior knowledge of σ or of the Besov regularity parameters. A key technical point is also that, in the low-regularity regime, a standard online-to-batch conversion would give the slower first-order rate; our analysis instead exploits the stochastic directional curvature condition together with self-normalized concentration to recover the sharp squared-risk rate.

1.1 Related work

Nonparametric regression over Besov spaces. The minimax theory for nonparametric regression over Besov spaces is well established, with classical results characterizing optimal rates in expectation; see, e.g., [11, 16]. More recently, [12] derived sharp upper and lower bounds that are explicitly *noise-level–aware* in the batch setting, under the assumption of known Gaussian noise variance. These results highlight the importance of adaptivity to the noise level and motivate the use of online learning techniques to achieve such adaptivity in a data-driven manner. Recently [24] proposed a polynomial-time online algorithm achieving minimax-optimal regret against adversarial sequences and Besov-smooth competitors. While their focus is on adversarial regret, our work leverages related online ideas to derive high-probability, noise-level–adaptive risk guarantees in the stochastic regression setting.

Online-to-Batch. Classical online-to-batch reductions [26, 4] relate online regret to batch risk, but are limited to guarantees in expectation and do not yield fast rates with high probability. The route to obtain high probability $O(\frac{1}{T})$ excess risk bounds via online to batch conversions was initiated by [37]. More recently, [35] proposed a general reduction for deriving high-probability risk bounds from sequential regret. However, this approach does not recover minimax-optimal rates in our setting, in particular in the low-regularity regime $s < d/2$, where the minimax-optimal online regret rate scales as $O(T^{-s/d})$ [31], which is strictly slower than the optimal batch rate $O(T^{-2s/(2s+d)})$. In contrast, our analysis departs from generic online-to-batch reductions and relies on a refined control of the stochastic regret in (2), under a stochastic directional derivative condition (Assumption 2). This approach is inspired by recent work on noisy online convex optimization [38], which established high-probability fast rates under sub-Gaussian perturbations, but extends it to a more general setting tailored to nonparametric regression and adaptive wavelet estimation.

Online convex optimization with stochastic unbounded gradients. Most online convex optimization algorithms require a priori gradient bounds, or incur exponential penalties when gradients are unbounded [8]. Recent works address unbounded gradients through stochastic assumptions and gradient clipping. In particular, [22] obtain in-expectation guarantees under sub-exponential noise, while [39] derive high-probability, parameter-free regret bounds using regularization techniques. As a by-product of our analysis, we complement this line of work by establishing high-probability ℓ_1 comparator-adaptive regret guarantees in a setting with bounded iterates and unbounded stochastic gradients.

2 High probability comparator-adaptive regret with constrained iterates and noisy unbounded gradients

We consider the following setting of stochastic online prediction. Let $N \geqslant 1$ and $(\ell_t : \mathbb{R}^N \to \mathbb{R}_+)_{t\geqslant 1}$ a collection of random convex differentiable loss functions, sequentially observed. At each time t, a learner forms a prediction $\mathbf{c}_t \in \mathbb{R}^N$ based on the past observations $\mathcal{F}_{t-1} = \{\mathbf{c}_1, \nabla\ell_1(\mathbf{c}_1), \dots, \mathbf{c}_{t-1}, \nabla\ell_{t-1}(\mathbf{c}_{t-1})\}$, where $\nabla\ell_s(\mathbf{c}_s)$ is the gradient of ℓ_s at $\mathbf{c}_s$. The learner aims at minimizing the cumulative stochastic regret

$$R_T(\mathbf{c}) := \sum_{t=1}^{T} \mathbb{E}_{t-1}[\ell_t(\mathbf{c}_t)] - \mathbb{E}_{t-1}[\ell_t(\mathbf{c})], \qquad \text{where} \quad \mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}],$$

with respect to all $\mathbf{c} \in \mathbb{R}^N$ such that $\|\mathbf{c}\|_\infty \leqslant C$ for some fixed $C > 0$. In particular, considering the Dirac masses, one obtains $\ell_t = \mathbb{E}_{t-1}[\ell_t]$ and the stochastic regret matches the regret more commonly used in the online (adversarial) learning literature. In this section, we present Algorithm 1, which achieves a stochastic regret bound of order $\tilde{O}(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T})$ with high probability with unbounded gradients of variance σ and expected bound $G \geqslant \left|\mathbb{E}_{t-1}[\nabla\ell_t(\mathbf{c}_t)]\right|$. In particular, we assume that gradients' noise admit sub-exponential tails, as formalized in Assumption 1.

Assumption 1 (Sub-exponential gradient noise)**.** *The gradient sequence $(\hat{\mathbf{g}}_t)_{t\geqslant 1}$ is said to have conditionally sub-exponential noise if there exist constants $\nu, \mu > 0$ such that, for all $t \geqslant 1$,*

conditionally on the past $\mathcal{F}_{t-1}$, the centered gradient noise $\boldsymbol{\xi}_t := \hat{\mathbf{g}}_t - \mathbb{E}_{t-1}[\hat{\mathbf{g}}_t]$ satisfies the following tail bound: for all $u \geqslant 0$,

$$\max_{1 \leqslant n \leqslant N} \mathbb{P}_{t-1}(|\xi_{n,t}| \geqslant u) \leqslant \exp\left(-\frac{1}{2}\min\left(\frac{u^2}{\nu^2}, \frac{u}{\mu}\right)\right),$$

where $\mathbb{P}_{t-1}$ denotes the conditional probability given the past observations $\mathcal{F}_{t-1}$.

Stochastic directional derivative condition. We introduce the following assumption on the random losses (ℓ_t) revealed by the environment.

Assumption 2 (Stochastic directional derivative condition)**.** *We say that the (random) losses ℓ_t satisfy the* stochastic directional derivative condition *for some $\alpha > 0$ if for every $\mathbf{c}_1, \mathbf{c}_2 \in [-C, C]^N, C > 0$*

$$\mathbb{E}_{t-1}\big[\ell_t(\mathbf{c}_1) - \ell_t(\mathbf{c}_2)\big] \leqslant \mathbb{E}_{t-1}\left[\nabla \ell_t(\mathbf{c}_1)^\top(\mathbf{c}_1 - \mathbf{c}_2) - \frac{\alpha}{2}\big(\nabla \ell_t(\mathbf{c}_1)^\top(\mathbf{c}_1 - \mathbf{c}_2)\big)^2\right].$$

Assumption 2 is a stochastic (risk-level) analogue of the directional curvature property implied by exp-concavity. In the deterministic case, exp-concavity implies a second-order lower bound in the direction of the gradient, which is the key property used in the analysis of Online Newton Step algorithms [20, see, e.g., Lemma 4.3]. Our assumption only requires this curvature inequality after conditioning on the past. It is therefore weaker than requiring each realized loss ℓ_t to be exp-concave, or even convex, pointwise. This type of stochastic relaxation is standard in fast-rate analyses; see, for instance, [36] and [27]. In online stochastic optimization, closely related assumptions appear in [38], where they are used to obtain logarithmic stochastic regret bounds for Online Newton-type algorithms. Finally, Assumption 2 is satisfied in the regression setting with squared loss under standard boundedness of the predictions and finite-variance assumptions; see Section 3.

2.1 Algorithm: Adaptive Learning with Unbounded Noisy Gradients via Clipping

We now introduce an online learning algorithm designed to operate in the presence of unbounded gradient observations, while retaining high-probability, comparator-adaptive regret guarantees of type $\tilde{O}(\|\mathbf{c}\|_1\sqrt{T})$. Our construction is modular: each coordinate is updated by an independent one-dimensional online subroutine, which receives clipped gradient feedback. This design allows us to leverage generic comparator-adaptive online algorithms (satisfying Assumption 3), while ensuring robustness in regimes where gradient magnitudes may be large.

Algorithm 1: Adaptive Learning with Unbounded Noisy Gradients via Clipping at time t

Input: $N \geqslant 1$ algorithms $\mathcal{A}_1, \dots \mathcal{A}_N$ satisfying Assumption 3, diameters (C_n) of feasible sets, bounds on gradients $|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c}_t)_n]| \leqslant G_{n,t}, n \in [N]$ and clipping margin $\Delta_t > 0$;
Predict $\mathbf{c}_t = (c_{n,t})_{n=1}^N$;
Receive noisy gradient $\hat{\mathbf{g}}_t = \nabla \ell_t(\mathbf{c}_t)$;
for $n = 1, \dots, N$ **do**
  Define clipping threshold as $\bar{G}_{n,t} = G_{n,t} + \Delta_t$;
  Clip gradient $\bar{g}_{n,t} = \max(-\bar{G}_{n,t}, \min(\bar{G}_{n,t}, \hat{g}_{n,t}))$ or $\bar{g}_{n,t} = 0$ if $G_{n,t} = 0$;
  Receive $c_{n,t+1} \in [-C_n, C_n]$ from $\mathcal{A}_n$ using $(\bar{g}_{n,t}, \bar{G}_{n,t})$;
Output: $\mathbf{c}_{t+1}$

Assumption on the coordinate subroutine. We assume that each coordinate subroutine $\mathcal{A}_n, n \in [N]$ satisfies the following assumption.

Assumption 3 (Comparator-adaptive algorithm)**.** *Let $T \geqslant 1, C > 0$ and let $g_1, \dots, g_T \in \mathbb{R}$ be a sequence of scalar gradients such that $|g_t| \leqslant G_t$ for all $t \leqslant T$. An online algorithm $\mathcal{A}$ starting at $c_1 \in [-C, C]$ and producing predictions $(c_t)_{t=1}^T \subset [-C, C]$ is said to be* comparator-adaptive *if for some factors $\Xi_1, \Xi_2 > 0$ and for any comparator $c \in [-C, C]$,*

$$\sum_{t=1}^T g_t(c_t - c) \;\leqslant\; |c - c_1|\left(\Xi_1\sqrt{\textstyle\sum_{t=1}^T g_t^2} + \Xi_2 \sup_{1 \leqslant t \leqslant T} G_t\right).$$

This assumption holds for a broad class of first-order online learning algorithms such as online mirror descent with self-tuned learning rates or coin-betting style updates [30, 9, 28, 21, 5]. These are referred to as *comparator-adaptive* algorithms as they provide optimal adaptivity to the magnitude of the comparator $|c|$ in the regret, typically at the cost of logarithmic factors absorbed into Ξ_1 and

Ξ_2. Algorithms satisfying the linear regret bound in Assumption 3 are primarily developed for the unconstrained setting ($C = \infty$). We here consider an extension in which the predictions c_t are constrained to $[-C, C]$, typically at the cost of a multiplicative constant in the regret, which we also absorb into the factors Ξ_1 and Ξ_2 [see, e.g., 9]. Importantly, these types of algorithms are not Lipschitz-adaptive: they require prior knowledge of a gradient bound $G_t \geqslant |g_t|$ at each time step $t \geqslant 1$, and failing to do so results in an exponential penalty in the regret bound [8, 28]. In our unbounded-gradient setting, this limitation can be overcome via gradient clipping (in Alg. 1), provided that the clipping level is carefully calibrated.

First result: high probability comparator-adaptive regret with known noise level. We first establish a guarantee for Algorithm 1 under the assumption that, at each time t, the algorithm is provided with the optimal clipping threshold Δ_t, which depends on the noise parameters (ν, μ) appearing in Assumption 1. This oracle assumption is subsequently removed in the following Section 2.2, where we design an adaptive gradient clipping strategy.

Theorem 1 (High-probability ℓ_1-comparator-adaptive regret)**.** *Let $G, C > 0$, $\delta \in (0, 1)$, and $N \in \mathbb{N}^*$. Assume $(\ell_t)_{t\geqslant 1}$ satisfy Assumption 2 and at each time step $t \geqslant 1$ we are given a stochastic gradient $\nabla \ell_t(\mathbf{c}_t) \in \mathbb{R}^N$ satisfying Assumption 1 with parameter $\nu, \mu > 0$, and*

$$\|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c}_t)]\|_\infty \leqslant G \quad \text{and} \quad \max_{1\leqslant n\leqslant N} \mathbb{E}_{t-1}\big[|\nabla \ell_t(c_t)_n - \mathbb{E}_{t-1}[\nabla \ell_t(c_t)_n]|^2\big] \leqslant \sigma^2.$$

Then Algorithm 1, run over T steps with $C_n = C$, $G_{n,t} = G$ for every $n \in [N], t \geqslant 1$ and clipping margins

$$\Delta_t = \left(\nu\sqrt{2\log\left(\tfrac{(\sqrt{\pi/2}\nu+2\mu)t}{\log(\delta^{-1})}\right)} \vee 2\mu\log\left(\tfrac{(\sqrt{\pi/2}\nu+2\mu)t}{\log(\delta^{-1})}\right)\right),\ t \geqslant 1$$

satisfies, with probability at least $1 - 2\delta$, for all comparators $\mathbf{c} \in \mathbb{R}^N$ such that $\|\mathbf{c}\|_\infty \leqslant C$,

$$R_T(\mathbf{c}) \leqslant O\left[\|\mathbf{c}\|_1(G+\sigma)\sqrt{T} + \left(\big(CN(G + (\nu \vee \mu)\log T)\big)^2 + \frac{1}{\alpha}\right)\log(\delta^{-1})\right].$$

The proof is postponed to Appendix A, where we state and prove a more general version of Theorem 1. This full result, stated as Theorem 4, will be instrumental for our regression setting in the next section, where we consider a multiscale approach with heterogeneous diameters.

Note that Algorithm 1 requires the noise parameters $\mu, \nu > 0$ to be specified in advance. In Section 2.2, we remove this requirement and develop an adaptive clipping procedure that achieves the same guarantees. In particular, the variance parameter $\sigma > 0$ is one dimensional and not passed to the algorithms while it appears optimally [22] in our regret bound $\tilde{O}(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T})$ that holds in high probability. Note also that for simplicity we stated Theorem 1 with uniform noise in (ν, μ, σ) over the coordinates. Defining a coordinate dependent clipping margin with $(\mu_n, \nu_n), n \in [N]$ would lead to an adaptive regret bound scaling as $O\big(\sum_n |c_n|(G+\sigma_n)\sqrt{T} + C^2(\nu_n \vee \mu_n)^2 \log(T)^2\big)$, with high probability.

Our Theorem 1 also parallels the results of [39]. As discussed in their work, even in the constrained setting where $\|\mathbf{c}_t\|_\infty \leqslant C, t \geqslant 1$, obtaining a high-probability regret bound of order $\tilde{O}(\|\mathbf{c}\|_1\sqrt{T})$ is nontrivial. Indeed, a direct application of standard sub-exponential martingale concentration inequalities typically introduces an additional deviation term of order $\tilde{O}(NC\sqrt{T})$, which is too coarse to preserve the desired dependence on $\|\mathbf{c}\|_1$. In contrast, under Assumption 2, we are able to leverage second-order concentration tools based on self-normalized martingale inequalities. This approach allows us to control the deviation term at the cost of only an additional $O(\log^2 T)$ factor, thereby maintaining the target high-probability regret bound $\tilde{O}(\|\mathbf{c}\|_1\sqrt{T})$ without resorting to regularized or surrogate loss functions. We also note that Theorem 5 of [39] does not yield regret bounds with ℓ_1-norm comparator dependence. Their high-probability guarantees are instead restricted to $\|\mathbf{c}\|_p$-dependence for some $p \in (1, 2]$. In contrast, maintaining an optimal dependence of order $\tilde{O}(\|\mathbf{c}\|_1\sqrt{T})$ with high probability is essential for our analysis. This ℓ_1-comparator adaptivity plays a crucial role in Section 3, where it enables the derivation of adaptive nonparametric risk bounds.

2.2 Adaptive clipping strategy

In Theorem 1, we showed that Algorithm 1, when run with an optimal clipping threshold $\bar{G}^* = G + \Delta^*$ that exceeds the true gradient bound $G \geqslant \|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c}_t)]\|_\infty$ by a logarithmic margin $\Delta^* \propto (\nu \vee \mu)\log T$, achieves a high-probability regret guarantee of $O\big(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T} + (CN)^2(G+\Delta^*)^2\big)$ for any $\mathbf{c} \in \mathbb{R}^N$, $\|\mathbf{c}\|_\infty \leqslant C$. In this section, we design a procedure that achieves similar comparator-

adaptive guarantees as in Theorem 1, *without* requiring prior knowledge of the clipping margin $\Delta^* > 0$, which depends on the unknown noise parameters (ν, μ). As a consequence, the proposed strategy adapts sequentially to the level of the stochastic noise.

Adaptive clipping via expert aggregation. Given $T \geqslant 1$, we define the following set of clipping margins to pass to Algorithm 1

$$\mathcal{D} := \{0\} \cup \left\{2^k : 0 \leqslant k \leqslant \lfloor \tfrac{1}{2} \log_2 T \rfloor + 1\right\}. \tag{3}$$

The procedure consists in running $|\mathcal{D}| = O(\log T)$ independent instances of Algorithm 1, with a bound on the true gradient for every time $t \geqslant 1$, $G \geqslant \|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c}_t)]\|_\infty, t \geqslant 1$ and with a clipping margin set at each $\Delta \in \mathcal{D}$. Each instance then produces a prediction $\mathbf{c}_t^\Delta$ at time t associated to each margin $\Delta \in \mathcal{D}$ and we leverage an expert algorithm satisfying Assumption 4 to sequentially aggregates the predictions $(\mathbf{c}_t^\Delta)_{\Delta \in \mathcal{D}}$ as, for every $t \geqslant 1$

$$\hat{\mathbf{c}}_t = \sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta, \qquad \text{where} \qquad w_t^\Delta \geqslant 0\,, \ \sum_{\Delta \in \mathcal{D}} w_t^\Delta = 1, \ t \geqslant 1. \tag{4}$$

Remark that given the bound of Theorem 1 scaling as $O\big(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T} + (CN(G+\Delta))^2\big)$, a clipping margin set at $\sup_{\Delta \in \mathcal{D}} \Delta = O(\sqrt{T})$ leads to a bound in $O(T)$ which is of the same order of the worst regret case bound $O(NCGT)$[1], which justifies that the range $\sqrt{T}$ of the grid $\mathcal{D}$ is sufficient.

Assumption on the expert aggregation algorithm. We introduce the following assumption for the expert aggregation subroutine we use in (4).

Assumption 4 (Second-order adaptive algorithm)**.** *Let $T, E \geqslant 1$ and let $\hat{\nabla}_1, \dots, \hat{\nabla}_T \in \mathbb{R}^E$ be any sequence of gradients. An expert aggregation algorithm is said to be* second-order adaptive *if, starting from the uniform prior $\mathbf{w}_1 = (\frac{1}{E})_{1 \leqslant e \leqslant E}$, it produces iterates $\mathbf{w}_t = (w_t^e)_{1 \leqslant e \leqslant E}$ such that $\sum_{e=1}^E w_t^e = 1$ and for every $e \in \{1, \dots, E\}$, the following regret bound holds:*

$$\sum_{t=1}^T \hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_{e,t} \leqslant \Xi_3 \sqrt{\log(E) \textstyle\sum_{t=1}^T (\hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_{e,t})^2} + \Xi_4 \log(E) \max_{1 \leqslant t \leqslant T} \|\hat{\nabla}_t\|_\infty,$$

where Ξ_3 and Ξ_4 are universal and independent of the gradient range and can hide $\log\log T$ terms.

Assuming prior knowledge of a bound on the gradients, `MLProd` [17] was the first algorithm to achieve this type of second-order regret guarantee. Building on this work, a substantial line of research has since focused on designing algorithms that enjoy such guarantees while being *scale-free* and *Lipschitz-adaptive* [see, e.g., 29, for a review]. Well-known instances of this class include `BOA` [37] and `Squint` [23]. Unlike standard methods that require an *a priori* bound on the sup-norm of the gradients (e.g., those satisfying Assumption 3), algorithms satisfying Assumption 4 adapt both to the observed gradient range $\max_{t \leqslant T} \|\hat{\nabla}_t\|_\infty$ and to the cumulative quadratic variation of the prediction errors.

Second result: high probability comparator-adaptive regret with adaptive clipping. We now establish a high-probability comparator-adaptive regret bound in Theorem 2, for the strategy discussed in (4), when the noise level is unknown. By introducing an adaptive clipping scheme, the algorithm automatically calibrates the clipping thresholds from the observed data, while retaining the same order of regret as in the oracle setting up to logarithmic factors.

Theorem 2. *Let $T, N \geqslant 1$. Assume $(\ell_t)_{t \geqslant 1}$ satisfy Assumption 2 for some $\alpha > 0$ and, for every $t \geqslant 1$ and $\mathbf{c} \in \mathbb{R}^N$ with $\|\mathbf{c}\|_\infty \leqslant C$, $\nabla \ell_t(\mathbf{c})$ satisfies Assumption 1,*

$$\|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c})]\|_\infty \leqslant G \quad \text{and} \quad \max_{1 \leqslant n \leqslant N} \mathbb{E}_{t-1}\big[|\nabla \ell_t(\mathbf{c})_n - \mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c})_n]|^2\big] \leqslant \sigma^2\,.$$

Consider the sequential predictions $\hat{\mathbf{c}}_t = \sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta$, $t \geqslant 1$, as in (4), with $(w_t^\Delta)_{\Delta \in \mathcal{D}}$ satisfying Assumption 4. Then, for every $\delta \in (0,1)$ and every $\mathbf{c} \in \mathbb{R}^N$ with $\|\mathbf{c}\|_\infty \leqslant C$, with probability at least $1 - 6\delta$,

$$R_T(\mathbf{c}) \leqslant O\left[\|\mathbf{c}\|_1(G+\sigma)\sqrt{T} + \left((\alpha+1)\Big(NC\Big(G + (\nu \vee \mu)\log(T)\Big)\Big)^2 + \frac{1}{\alpha}\right)\log(\delta^{-1})\right].$$

We provide a detailed version of Theorem 2 and a proof in Appendix B. To give some intuition, an adaptive expert algorithm of type Assumption 4 is key to derive this result: under Assumption 1, the maximum magnitude expert gradient $\sup_t \|\hat{\nabla}_t\|_\infty$ is a random variable that scales logarithmically

[1] For every $t \geqslant 1$, one has $\mathbb{E}_{t-1}[\ell_t(\mathbf{c}_t)] - \mathbb{E}_{t-1}[\ell_t(\mathbf{c})] \leqslant \|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c}_t)]\|_\infty \|\mathbf{c}_t - \mathbf{c}\|_1 \leqslant 2NCG$ by Assumption 2.

with T in high probability, and the variance (second order) term is managed with Assumption 2, enabling the recovery of high-probability regret bounds in Theorem 2, of type $\tilde{O}(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T})$ without assuming bounded gradient nor the knowledge of the noise parameters $\sigma, \nu, \mu > 0$. Interestingly, in the setting of sub-exponential noise and losses satisfying the stochastic directional derivative assumption, our result answers a challenging question raised by [39] as to whether it is possible to achieve a regret bound of order $\tilde{O}(\|\mathbf{c}\|_1(G+\sigma)\sqrt{T})$ while remaining adaptive to the noise amplitude. Importantly, we believe that our adaptive clipping procedure with unknown noise level can be applied to a broader range of learning problems. In particular, this result will play a central role in the learning of nonparametric functions with noisy observations in Section 3, and enables the derivation of minimax adaptive nonparametric risk guarantees with high probability.

3 High-Probability Regression over Besov Spaces via Online-to-Batch

We now turn to the main statistical contribution of the paper, where we leverage the online learning guarantees established in the previous sections to derive high-probability minimax risk bounds for nonparametric regression over Besov spaces.

Setting: statistical learning. We observe a sample of $T \geqslant 1$ i.i.d. pairs $(X_1, Y_1), \dots, (X_T, Y_T)$ drawn from an unknown distribution on $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact. The data follow the nonparametric regression model

$$Y_t = f(X_t) + \varepsilon_t, \quad t = 1, \dots, T,$$

where $f : \mathcal{X} \to \mathbb{R}$ is an unknown regression function and $(\varepsilon_t)_{t \geqslant 1}$ is a sequence of zero-mean noise variables with unknown finite variance $\mathbb{E}_{t-1}[\varepsilon_t^2 | X_t] = \sigma^2 > 0$. Our goal is to construct an estimator $\bar{f}_T$ that minimizes, in the minimax sense, the excess prediction risk

$$\mathbb{E}_{XY}\big[(\bar{f}_T(X) - Y)^2\big] - \mathbb{E}_{XY}\big[(f(X) - Y)^2\big],$$

uniformly over f belonging to a Besov space (defined below). The estimator $\bar{f}_T$ is obtained via an online-to-batch conversion and takes the form $\bar{f}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{f}_t$, where each $\hat{f}_t$ is a wavelet-based estimator updated sequentially. In particular, each predictor $\hat{f}_t$ is parameterized by a collection of wavelet coefficients, which are learned sequentially using the algorithms developed in Section 2.1 and Section 2.2. At each round t, the coefficients of the predictor $\hat{f}_t$ are updated based on the stochastic gradient $\hat{g}_t = \ell_t'(\hat{f}_t) = 2\big(\hat{f}_t(X_t) - Y_t\big)$, associated with the squared loss $\ell_t(\hat{f}_t) = (\hat{f}_t(X_t) - Y_t)^2$. This setting mirrors the stochastic online convex optimization framework analyzed in Section 2. Under the squared loss, Assumption 2 holds with $\alpha = [4((\sup_t \|\hat{f}_t\|_\infty + \|f\|_\infty)^2 + \sigma^2)]^{-1}$ (see Appendix C). Our goal is to control, for any comparator f, the stochastic regret $R_T(f) = \sum_{t=1}^{T} \mathbb{E}_{t-1}[\ell_t(\hat{f}_t)] - \mathbb{E}_{t-1}[\ell_t(f)]$ appearing in (2). Before describing our adaptive online wavelet regression procedure and deriving high-probability, noise-level–adaptive minimax rates matching recent optimal results [12], we briefly recall the necessary background on wavelet bases and Besov space approximation.

3.1 Besov functions and wavelets.

In the following, we provide a minimal introduction to wavelet approximation theory and Besov functions; see Appendix F for additional details. We refer the interested reader to [6, 18, 19] for a deeper introduction and comprehensive treatments of wavelets and Besov spaces.

Wavelet function representation. We consider compactly supported functions $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$, $d \geqslant 1$. To design our algorithm, we rely on a multiscale representation of f based on an orthonormal wavelet basis $\{\phi_{j_0,k}, \psi_{j,k}\}$. For a chosen starting scale $j_0 \in \mathbb{N}$, f admits the decomposition

$$f = \sum_{k \in \bar{\Lambda}_{j_0}} \alpha_{j_0,k}\, \phi_{j_0,k} + \sum_{j=j_0}^{\infty} \sum_{k \in \Lambda_j} \beta_{j,k}\, \psi_{j,k}. \tag{5}$$

Here, $\boldsymbol{\alpha}_{j_0} = (\alpha_{j_0,k})_{k \in \bar{\Lambda}_{j_0}}$ denotes the collection of *scaling coefficients*, with $|\bar{\Lambda}_{j_0}| \leqslant \lambda 2^{j_0 d}$, and for each $j \geqslant j_0$, $\boldsymbol{\beta}_j = (\beta_{j,k})_{k \in \Lambda_j}$ denotes the collection of *wavelet* (or *detail*) coefficients at scale j, with $|\Lambda_j| \leqslant \lambda 2^{jd}$. To lighten notation, we write $\varphi_{j,k}$ and $c_{j,k}$ for either a scaling function and its coefficient ($\varphi_{j_0,k} = \phi_{j_0,k}$, $c_{j_0,k} = \alpha_{j_0,k}$, $k \in \bar{\Lambda}_{j_0}$) or a wavelet and its detail coefficient ($\varphi_{j,k} = \psi_{j,k}$, $c_{j,k} = \beta_{j,k}$, $k \in \Lambda_j$).
When working with S-regular wavelet bases (with S sufficiently large, see Definition 2 in Appendix F), the expansion in (5) converges not only in $L^2(\mathcal{X})$ but also in $L^p(\mathcal{X})$ for $p \geqslant 1$ (or the space of

uniformly continuous functions) depending on whether $f \in L^p(\mathcal{X}), p \geqslant 1$. This broader convergence behavior is a key reason for adopting such regular bases.

Besov functions. Besov spaces $\mathcal{B}^s_{pq}$ constitute a classical family of function spaces indexed by three parameters: a smoothness parameter $s > 0$, an integrability parameter $p \in [1, \infty]$, and a summability parameter $q \in [1, \infty]$. This space can be intuitively viewed as the space of functions with $s > 0$ derivatives in $L^p(\mathcal{X})$, with $p \geqslant 1$, and parameter $q \geqslant 1$ allows for additional finer control of the regularity of the underlying functions. There exist several equivalent definitions of Besov spaces (e.g. using differences, or interpolation theory): we refer to [18, 19, 34] for detailed and general background on Besov spaces. In this work, we adopt the wavelet characterization, which is particularly well suited for the analysis of our wavelet-based Algorithm 2.
Let $s > 0$ and let $\{\phi_{j_0,k}, \psi_{j,k}\}$ be an orthonormal S-regular wavelet basis with $S > s$ (see Definition 2). We introduce the set of L^∞-bounded Besov functions over some compact $\mathcal{X} \subset \mathbb{R}^d$ of radius $B > 0$ and with regularity parameters $1 \leqslant p, q \leqslant \infty$ and smoothness $s > \frac{d}{p}$

$$\mathcal{B}^s_{pq}(B) = \Big\{ f \in L^\infty(\mathcal{X}) : \|f\|_{\mathcal{B}} := \|\boldsymbol{\alpha}_{j_0}\|_p + \Big(\sum_{j \geqslant j_0} \big(2^{j(s-\frac{d}{p}+\frac{d}{2})} \|\boldsymbol{\beta}_j\|_p\big)^q \Big)^{1/q} \leqslant B \Big\}. \quad (6)$$

Approximation with wavelets. The approximation properties of $\mathcal{B}^s_{pq}$ in wavelet bases [14, 6, 13] are used in the proof of Theorem 3 (Appendix D, Step 2).

3.2 Algorithm: Adaptive Online Wavelet Regression

In the adversarial setting, [24, Theorem 1] showed that using a single wavelet predictor at a fixed uniform resolution yields a suboptimal regret of order $O(\sqrt{T})$ in the high-regularity regime $s \geqslant d/2$. This phenomenon also appears in the batch setting, where the optimal rate is $O(T^{-2s/(2s+d)}) < O(T^{-\frac{1}{2}})$. A workaround is to rely on a multiscale aggregation strategy that averages several wavelet predictors associated with different resolutions. Intuitively, this approach aims at selecting the resolution that best balances approximation and variance terms. As shown in Theorem 3, such a strategy with Algorithm 1 achieves minimax optimal rates $O(T^{-2s/(2s+d)})$ in the batch setting, where the losses are random through the noise ε_t in the data. Remarkably, we show that our procedure also adapts to the noise level and attains an optimal dependence on the noise in the risk bound of Theorem 3, reaching the recent noise-level aware bound of [12], for any unknown variance σ. Our procedure builds directly on ideas of [24] and we summarize the resulting algorithm in Algorithm 2.

Procedure. Let $J_0 \geqslant 0$ be a maximum starting scale. At each time $t \geqslant 1$, we predict

$$\hat{f}_t(X_t) = \sum_{e \in \mathcal{E}} w_{e,t} \hat{f}_{e,t}(X_t) \qquad \text{with } w_{e,t} \geqslant 0\,, \ \sum_{e \in \mathcal{E}} w_{e,t} = 1, \quad (7)$$

where each $\hat{f}_{e,t}$ is a truncated wavelet estimator of type (5) at some maximum level $J \geqslant 0$, whose coefficients are learnt with Algorithm 1 and is associated to a triplet $e = (j_0, \boldsymbol{\alpha}_{j_0,1}, \Delta) \in [J_0] \times \mathcal{A}_{j_0} \times \mathcal{D} =: \mathcal{E}$: j_0 is the starting scale of the wavelet expansion; $\boldsymbol{\alpha}_{j_0,1}$ are the scaling coefficients at initial time $t = 1$ at level j_0 and set on a discretization grid $\mathcal{A}_{j_0} \subset [-2^{-j_0 d/2} \|f\|_\infty \|\phi\|_1; 2^{-j_0 d/2} \|f\|_\infty \|\phi\|_1]$ of precision $\epsilon_{j_0} = O(2^{-j_0 d/2}/\sqrt{T})$ and of size $|\mathcal{A}_{j_0}| = O(\sqrt{T})$ for all $j_0 \in [J_0]$; and Δ is some margin clipping belonging to $\mathcal{D}$ defined in (3). The total number of experts is given by $|\mathcal{E}| = |\mathcal{D}| \sum_{j_0=0}^{J_0} |\bar{\Lambda}_{j_0}|\, |\mathcal{A}_{j_0}| = O(T^{3/2} \log T)$, as for each level j_0, each scaling coefficient is discretized independently on the grid $\mathcal{A}_{j_0}$ of size $O(\sqrt{T})$, with $|\bar{\Lambda}_{j_0}| = O(2^{j_0 d})$, $J_0 = \lfloor d^{-1} \log_2 T \rfloor$ (see Theorem 3) and $|\mathcal{D}| = O(\log T)$. These experts are then aggregated according to (7) using a subroutine that satisfies Assumption 4. The complete procedure is summarized in Algorithm 2.

Boundedness of predictors. We enforce uniform boundedness through coefficient constraints. For the scaling coefficients, for all $k \in \bar{\Lambda}_{j_0}, |\alpha_{j_0,k}| \leqslant 2^{-j_0 \frac{d}{2}} \|\phi\|_1 \|f\|_\infty$, which defines the coarse grids $\mathcal{A}_{j_0}$. For wavelet coefficients, if $\|f\|_{\mathcal{B}} \leqslant B$, the Besov norm definition (6) and $s > d/p$ give

$$|\beta_{j,k}| \leqslant B\, 2^{-j(s-\frac{d}{p}+\frac{d}{2})} \leqslant B\, 2^{-j\frac{d}{2}} =: C_{j,k}.$$

Constraining coefficients to $[-C_{j,k}, C_{j,k}]$ ensures that all expert predictors $\hat{f}_{e,t}$ are uniformly bounded in sup-norm; see Lemma 4. This boundedness is crucial for the high-probability concentration arguments in Section 2. Importantly, the diameters $(C_{j,k})$ do not depend on the unknown parameters (s, p, q), at the cost of an additional $O(\log T)$ factor in the sup-norm bound on the predictors.

Algorithm 2: Adaptive Online Wavelet Regression

Input : Bound on expected derivative $\mathbb{E}_{t-1}[\ell_t'(\hat{f}_t(X_t))] \leqslant G, t \geqslant 1$; bound on Besov norm $B > 0$; set of wavelet experts $(\hat{f}_e)_{e\in\mathcal{E}}$;

Initialize : Diameters $(C_{j,k}) = (B2^{-jd/2})$; uniform weights $\tilde{\mathbf{w}}_1 = (\tilde{w}_{e,1})_{e\in\mathcal{E}}$; prediction functions $(\hat{f}_{e,1})_{e\in\mathcal{E}}$ with initial scaling coefficients set on the grids $(\mathcal{A}_{j_0})_{0\leqslant j_0\leqslant J_0}$ and detail coefficients at zero;

for $t = 1$ **to** T **do**
- Receive X_t and predict $\hat{f}_t(X_t) = \sum_{e\in\mathcal{E}} w_{e,t}\hat{f}_{e,t}(x_t)$;
- Reveal gradient $\hat{\nabla}_t = \nabla_{\mathbf{w}_t}\ell_t\big(\sum_{e\in\mathcal{E}} w_{e,t}\hat{f}_{e,t}(x_t)\big)$;
- Update $\mathbf{w}_{t+1} \leftarrow \texttt{weight}(\mathbf{w}_t, \hat{\nabla}_t)$ with `weight` satisfying Assumption 4 ;
- **for** $e \in \mathcal{E}$ **do**
  - Reveal gradient $\hat{\mathbf{g}}_{e,t} = (g_{j,k,t})$ of coefficients in $\hat{f}_{e,t}$ as in (8) ;
  - Update coefficients $(c_{j,k,t})$ of $\hat{f}_{e,t}$ using Algorithm 1 with input $\hat{\mathbf{g}}_{e,t}$, expected gradient bounds $(G_{j,k,t}) = G\cdot(|\varphi_{j,k}(X_t)|)$, diameters $(C_{j,k})$ and margin clipping Δ_e ;

Output : $(\hat{f}_t)_{1\leqslant t\leqslant T}$

Gradients and computational complexity. For each wavelet coefficient, the gradient is obtained by the chain rule:

$$\hat{g}_{j,k,t} = \Big[\partial_c\,\ell_t\Big(\hat{f}_t(X_t) - c_{j,k,t}\varphi_{j,k}(X_t) + c\varphi_{j,k}(X_t)\Big)\Big]_{c=c_{j,k,t}} = \ell_t'(\hat{f}_t(X_t))\varphi_{j,k}(X_t), \quad (8)$$

with $\ell_t'(\hat{f}_t(X_t)) = 2(\hat{f}_t(X_t) - Y_t)$ for the squared loss. Hence, using $\mathbb{E}_{t-1}[\varepsilon_t] = 0$, $|\mathbb{E}_{t-1}[\ell_t'(\hat{f}_t(X_t))]| \leqslant 2(\|\hat{f}_t\|_\infty + \|f\|_\infty)$, and we set $G = 2(\|\hat{f}_t\|_\infty + \|f\|_\infty)$ in Algorithm 1. Although there are $O(2^{jd})$ coefficients at level j, the update is sparse: using compactly supported wavelet, $\hat{g}_{j,k,t} = 0$ unless $\varphi_{j,k}(X_t) \neq 0$. Thus only the coefficients active at X_t are updated and clipped, with a deterministic sparsity pattern determined by X_t; see [25, 24] for related constructions and complexity details.

Third result: optimal noise level-aware risk bound in high probability. We now state our main nonparametric result, establishing a high-probability nearly minimax-optimal risk bound that is adaptive to the unknown noise level.

Theorem 3. *Let $T \geqslant 2$, $B > 0$, $p, q \in [1,\infty]$ and $s > d/p$. Let $(\hat{f}_e)_{e\in\mathcal{E}}$ be S-regular wavelet predictors, $S > s$, with $J = J_0 = \lfloor d^{-1}\log_2 T\rfloor$. Run Algorithm 2 with B and $G = 2(\hat{B}_\infty + B_\infty)$ where $\hat{B}_\infty$ and B_∞ are defined in Lemma 4, and set $\bar{f}_T = T^{-1}\sum_{t=1}^T \hat{f}_t$. Then, for every $\sigma \geqslant 0$ and every noise satisfying Assumption 1 with variance σ^2, Assumption 2 holds with $\alpha^{-1} = 4((\hat{B}_\infty + B_\infty)^2 + \sigma^2)$ (see Appendix C), and for every $\delta \in (0,1)$ and every $f \in \mathcal{B}^s_{pq}(B)$, with probability at least $1 - 6\delta$,*

$$\|\bar{f}_T - f\|_2^2 \leqslant \tilde{O}\left(\|f\|_{\mathcal{B}}^{\frac{2d}{2s+d}}(\sigma^2)^{\frac{2s}{2s+d}}T^{-\frac{2s}{2s+d}} + \|f\|_{\mathcal{B}}^2\,T^{-2(\frac{s}{d}-(\frac{1}{p}-\frac{1}{2})_+)} + (B^2+\sigma^2)\,\log(\delta^{-1})T^{-1}\right),$$

where $\tilde{O}$ hides universal constants and $\mathrm{polylog}(T)$ *factors.*

The proof is deferred to Appendix D. Several aspects of Theorem 3 are worth highlighting.

Noise-level-aware minimax rates and adaptivity. Our bound matches, up to logarithmic factors, the *Noise-Level-Aware (NLA)* minimax rates recently established in the batch setting by [12], who derive an explicit algorithm requiring full knowledge of (s, p, q, σ). To our knowledge, this is the first online procedure achieving NLA minimax-optimal rates *adaptively*, without knowledge of the noise level or the Besov parameters, at the sole cost of an additive $\tilde{O}((B^2+\sigma^2)T^{-1}\log\delta^{-1})$ term reflecting the price of the aggregation procedure. Note that the knowledge of B as input of Algorithm 2 could be relaxed using standard techniques, e.g., a doubling trick [3, 28], at the price of constant and logarithmic factors in the risk bound.

Low-noise and high-noise regimes. The bound in Theorem 3 reveals a sharp transition between two regimes, governed by the instance-dependent critical noise level $\sigma^* = \|f\|_{\mathcal{B}}\cdot$

$T^{-s/d+\frac{2s+d}{2s}(1/p-1/2)_+}$. In the low-noise regime $\sigma < \sigma^*$, the second term dominates and the bound yields $O(\|f\|_{\mathcal{B}}^2 T^{-2(s/d-(1/p-1/2)_+)})$ which is independent of σ and corresponds to the $\sigma \to 0$ regime of [12]. In the high-noise regime $\sigma \geqslant \sigma^*$, the first term dominates and recovers the classical minimax rate $O(T^{-2s/(2s+d)})$ with the optimal dependence on both σ and $\|f\|_{\mathcal{B}}$. Moreover, the leading terms depend on the true Besov norm $\|f\|_{\mathcal{B}}$, rather than only on the worst-case radius B, yielding sharper bounds for functions with smaller norm.

High-probability guarantee and deviation term. Remark that neither Algorithm 2 nor our clipping strategy (Section 2.2) depends on δ: the same run satisfies the bound in Theorem 3 for every $\delta \in (0,1)$. Indeed, as discussed in Section 2.2, the clipping grid $\mathcal{D}$ covers all relevant clipping levels, whatever the unknown noise and confidence levels. The level δ only enters through the additive deviation term $(B^2+\sigma^2)\log(\delta^{-1})/T$, which is dominated by the leading rate for every $\delta \geqslant \exp(-cT^{\frac{d}{2s+d}})$ in the high-noise regime $\sigma \geqslant \sigma^*$, where $c>0$ depends on $(\|f\|_{\mathcal{B}}, \sigma, B)$. Finally, note that the $\sigma^2\log(\delta^{-1})/T$ part is unimprovable, already for Gaussian mean estimation [2, 15], and the deviation term compares favorably with [12, Theorem 1] (as seen by inverting their tail bound (1.16) with $q=2$).

Comparison with online nonparametric regression and existing online-to-batch. Interestingly, the rates obtained in Theorem 3 can be compared with known bounds in online nonparametric regression. The rates in Theorem 3 contrast with known results in online nonparametric regression. For squared loss, [31] obtained minimax-optimal adversarial regret rates: when $s \geqslant d/2$, $R_T/T = O(T^{-2s/(2s+d)})$, matching the usual i.i.d. minimax risk rate, whereas for $s < d/2$ the rate deteriorates to $R_T/T = O(T^{-s/d})$, strictly slower than the batch rate $O(T^{-2s/(2s+d)})$; see also [32]. Thus, even with an online-optimal algorithm such as [24], a classical online-to-batch conversion does not yield optimal risk bounds in the low-regularity regime. The high-probability guarantees of Theorem 3 rely instead on a refined online-to-batch analysis: rather than applying first-order concentration to the regret trajectory as in existing reductions [e.g., 35], which is only tight when the regret concentrates at the $\sqrt{T}$ scale, we combine Assumption 2, a risk-level analogue of exp-concavity, with self-normalized martingale concentration to directly control the curvature of the squared loss. This recovers the sharp batch rate even when $s < d/2$, where the online regret rate is itself suboptimal. This sequential argument also differs from the batch analysis of [12], whose high-probability guarantee holds over the noise only, at a fixed grid design, and whose leading term depends on the worst-case radius B rather than on $\|f\|_{\mathcal{B}}$.

4 Conclusion

We proposed a sequential wavelet-based algorithm achieving noise-level-aware high-probability risk bounds that are minimax-optimal up to logarithmic factors over Besov spaces with $s > d/p$, without prior knowledge of the noise σ or the Besov parameters. One key technical ingredient is a refined online-to-batch analysis that goes beyond standard conversions where they fail to recover optimal rates: we combine a risk-level exp-concavity condition with self-normalized concentration to control the curvature of the loss. As by-products, we obtain ℓ_1-adaptive regret bounds and an adaptive clipping procedure of independent interest.
Our analysis focuses on the classical regime $s > d/p$, where Besov functions embed into L^∞. Natural future directions are to handle the boundary case $s = d/p$ and lower smoothness, to extend our results to $L^{p'}$ risks, $p' \neq 2$, and to investigate whether our online procedure also achieves the hybrid rates [19] known for Besov spaces.

References

[1] Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. 2008.

[2] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[3] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[4] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[5] Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pages 1216–1259. PMLR, 2021.

[6] Albert Cohen. *Numerical analysis of wavelet methods*, volume 32. Elsevier, 2003.

[7] Albert Cohen, Ingrid Daubechies, and J-C Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 45(5):485–560, 1992.

[8] Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on learning theory*, pages 643–677. PMLR, 2017.

[9] Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529. PMLR, 2018.

[10] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[11] Bernard Delyon and Anatoli Juditsky. On minimax wavelet estimators. *Applied and Computational Harmonic Analysis*, 3(3):215–228, 1996.

[12] Ronald DeVore, Robert D Nowak, Rahul Parhi, Guergana Petrova, and Jonathan W Siegel. Optimal recovery meets minimax estimation. *arXiv preprint arXiv:2502.17671*, 2025.

[13] Ronald A DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

[14] Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

[15] Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[16] David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921, 1998.

[17] Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

[18] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[19] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

[20] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[21] Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. In *Conference on Learning Theory*, pages 4160–4211. PMLR, 2022.

[22] Kwang-Sung Jun and Francesco Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Conference on learning theory*, pages 1802–1823. PMLR, 2019.

[23] Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175. PMLR, 2015.

[24] Paul Liautaud, Pierre Gaillard, and Olivier Wintenberger. Minimax adaptive online nonparametric regression over besov spaces. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[25] Paul Liautaud, Pierre Gaillard, and Olivier Wintenberger. Minimax-optimal and locally-adaptive online nonparametric regression. In *ALT 2025-36th International Conference on Algorithmic Learning Theory*, volume 272, pages 702–735, 2025.

[26] Nick Littlestone. From on-line to batch learning. In *Proceedings of the second annual workshop on Computational learning theory*, pages 269–284, 1989.

[27] Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1085–1093. PMLR, 2017.

[28] Zakaria Mhammedi and Wouter M Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, pages 2858–2887. PMLR, 2020.

[29] Zakaria Mhammedi, Wouter M Koolen, and Tim Van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory*, pages 2490–2511. PMLR, 2019.

[30] Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[31] Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

[32] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. 2017.

[33] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2025.

[34] Hans Triebel. *Theory of Function Spaces*. Birkhäuser, 2006.

[35] Dirk van der Hoeven, Nikita Zhivotovskiy, and Nicolò Cesa-Bianchi. High-probability risk bounds via sequential predictors. *arXiv preprint arXiv:2308.07588*, 2023.

[36] Tim van Erven, Peter Grünwald, Nishant Mehta, Mark Reid, and Robert Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[37] Olivier Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

[38] Olivier Wintenberger. Stochastic online convex optimization. application to probabilistic time series forecasting. *Electronic Journal of Statistics*, 18(1):429–464, 2024.

[39] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. *Advances in Neural Information Processing Systems*, 35:8000–8012, 2022.

APPENDIX

Contents

1 Introduction **1**
1.1 Related work 3

2 High probability comparator-adaptive regret with constrained iterates and noisy unbounded gradients **3**
2.1 Algorithm: Adaptive Learning with Unbounded Noisy Gradients via Clipping . . . 4
2.2 Adaptive clipping strategy 5

3 High-Probability Regression over Besov Spaces via Online-to-Batch **7**
3.1 Besov functions and wavelets. 7
3.2 Algorithm: Adaptive Online Wavelet Regression 8

4 Conclusion **10**

A Proof of Theorem 1 **14**

B Proof of Theorem 2 **17**

C Assumption 2 in Section 3 **21**

D Proof of Theorem 3 **21**

E Technical lemmas **29**

F Review of multi-resolution analysis **31**

A Proof of Theorem 1

Theorem 4 (High-probability parameter-free regret)**.** *Let $N \in \mathbb{N}^*, C_1, \dots, C_N > 0$ and $\delta \in (0,1)$. Assume we are given noisy gradients $\hat{\mathbf{g}}_1, \dots, \hat{\mathbf{g}}_T \in \mathbb{R}^N$ such that, for all $t \geqslant 1$, the gradient noise satisfies Assumption 1 with parameters $\nu, \mu > 0$, and*

$$|\mathbb{E}_{t-1}[\hat{g}_{n,t}]| \leqslant G_{n,t},\ n \in [N] \quad \textit{and} \quad \max_{1 \leqslant n \leqslant N} \mathbb{E}_{t-1}\left[|\hat{g}_{n,t} - \mathbb{E}_{t-1}[\hat{g}_{n,t}]|^2\right] \leqslant \sigma^2.$$

Then Algorithm 1, run with diameters (C_n), gradients bounds $(G_{n,t})$ and clipping margin

$$\Delta_t = \max\left\{\nu\sqrt{2\log\left(\tfrac{(\sqrt{0.5\pi}+2\mu)t}{\log(\delta^{-1})}\right)}, 2\mu\log\left(\tfrac{(\sqrt{0.5\pi}+2\mu)t}{\log(\delta^{-1})}\right)\right\}$$

satisfies, with probability at least $1-2\delta$, for all comparators $\mathbf{c} \in \mathbb{R}^N$ such that $|c_n| \leqslant C_n, n \in [N]$,

$$\begin{aligned} R_T(\mathbf{c}) \ \leqslant\ O\Bigg[\sum_{n=1}^N |c_n|\Big(\Xi_1\sqrt{\textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2]} + \big(\Xi_1\sqrt{\log(\delta^{-1})}+\Xi_2\big)\big(\sup_{1\leqslant t\leqslant T} G_{n,t} + \max(\nu,\mu)\log T\big)\Big) \\ + \left(\frac{1}{\alpha} + \left(\sup_{1\leqslant t\leqslant T}\sum_{n=1}^N C_n(G_{n,t}+\max(\nu,\mu)\log T)\mathbb{1}_{G_{n,t}>0}\right)^2\right)\log(\delta^{-1})\Bigg], \end{aligned}$$

where the notation $O[\cdot]$ hides multiplicative constants and $\log\log(\delta^{-1})$ terms.

Proof. Recall $\nabla L_t(\mathbf{c}_t) = \mathbb{E}_{t-1}[\hat{\mathbf{g}}_t]$ where we denote $\hat{\mathbf{g}}_t = (\hat{g}_{n,t})_{1\leqslant n\leqslant N} = \nabla \ell_t(\mathbf{c}_t)$ the noisy observed gradient and $\bar{\mathbf{g}}_t = (\mathrm{clip}(\hat{g}_{n,t}, \bar{G}_{n,t}))_{1\leqslant n\leqslant N}$ the clipped version of $\hat{\mathbf{g}}_t$ at time t. We set $Z_t = \hat{\mathbf{g}}_t^\top(\mathbf{c}_t - \mathbf{c})$ and $\bar{Z}_t = \bar{\mathbf{g}}_t^\top(\mathbf{c}_t - \mathbf{c})$ and one has by stochastic α-exp-concavity

$$\begin{aligned} \sum_{t=1}^T L_t(\mathbf{c}_t) - L_t(\mathbf{c}) &\leqslant \sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] - \frac{\alpha}{2}\mathbb{E}_{t-1}[Z_t^2] \\ &\leqslant \underbrace{\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t - \bar{Z}_t]}_{\text{clipping bias}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{t-1}[\bar{Z}_t] - \frac{\alpha}{2}\mathbb{E}_{t-1}[\bar{Z}_t^2]}_{\text{stochastic regret}} \end{aligned} \tag{9}$$

since $|\bar{Z}_t| \leqslant |Z_t|, t \geqslant 1$.

Bound on the stochastic regret. For every $t \geqslant 1$,

$$\sum_{n=1}^N |\bar{g}_{n,t}||c_{n,t} - c_n| \leqslant 2\sum_{n=1}^N \bar{G}_{n,t}C_n \leqslant 2\sup_{1\leqslant t\leqslant T}\sum_{n=1}^N \bar{G}_{n,t}C_n =: D.$$

Then, by the second inequality of Lemma 3, with $|\bar{Z}_t| \leqslant D, t \geqslant 1$ one has for every $\gamma > 0$, with probability $1-2\delta$

$$\sum_{t=1}^T \mathbb{E}_{t-1}[\bar{Z}_t] \leqslant \sum_{t=1}^T \bar{Z}_t + \frac{3\gamma}{2}\mathbb{E}_{t-1}[\bar{Z}_t^2] + \left(2D^2 + \frac{2}{\gamma}\right)\log(\delta^{-1}).$$

Taking $\gamma = \frac{\alpha}{3} > 0$ cancels the second order terms and we have a stochastic regret bounded as

$$\sum_{t=1}^T \mathbb{E}_{t-1}[\bar{Z}_t] - \frac{\alpha}{2}\mathbb{E}_{t-1}[\bar{Z}_t^2] \leqslant \sum_{t=1}^T \bar{Z}_t + \left(2D^2 + \frac{6}{\alpha}\right)\log(\delta^{-1}) \tag{10}$$

with probability $1-2\delta$. The problem is now reduced to bound the linear term $\sum_{t=1}^T \bar{Z}_t = \sum_{t=1}^T \bar{\mathbf{g}}_t^\top(\mathbf{c}_t - \mathbf{c})$ and one has

$$\begin{aligned}\sum_{t=1}^T \bar{\mathbf{g}}_t^\top(\mathbf{c}_t-\mathbf{c}) &= \sum_{n=1}^N \sum_{t=1}^T \bar{g}_{n,t}(c_{n,t}-c_n) \leqslant \sum_{n=1}^N |c_n| \left(\Xi_1 \sqrt{\textstyle\sum_{t=1}^T |\bar{g}_{n,t}|^2} + \Xi_2 \bar{G}_n \right) \\ &\leqslant \sum_{n=1}^N |c_n| \left(\Xi_1 \sqrt{\textstyle\sum_{t=1}^T 2\mathbb{E}_{t-1}[|\bar{g}_{n,t}|^2] + 2\bar{G}_n^2 \log(\delta^{-1})} + \Xi_2 \bar{G}_n \right) \end{aligned} \tag{11}$$

where first inequality is by Assumption 3 on iterates $(c_{n,t})$ and gradients $(\bar{g}_{n,t})$ with $|\bar{g}_{n,t}| \leqslant \bar{G}_n := \sup_{1\leqslant t\leqslant T} \bar{G}_{n,t}$; and last inequality holds with probability $1-\delta$, applying Lemma 2 on $(|\bar{g}_{n,t}|^2 - 2\mathbb{E}_{t-1}[|\bar{g}_{n,t}|^2])/(2\bar{G}_n^2)$.

Bound on the clipping bias. For every $t \geqslant 1$ one has for every coordinate $n \in [N], g_{n,t} = \mathbb{E}_{t-1}[\hat{g}_{n,t}], |g_{n,t}| \leqslant G_{n,t}$ and

$$|\hat{g}_{n,t} - \bar{g}_{n,t}| = (|\hat{g}_{n,t}| - \bar{G}_{n,t})_+ \leqslant (|\hat{g}_{n,t} - g_{n,t}| + |g_{n,t}| - \bar{G}_{n,t})_+ \leqslant (|\hat{g}_{n,t} - g_{n,t}| - (\bar{G}_{n,t} - G_{n,t}))_+.$$

Then, one has for every $n \in [N]$ with $\xi_{n,t} = \hat{g}_{n,t} - g_{n,t}$, $\Delta_{n,t} = \bar{G}_{n,t} - G_{n,t}$ and using Assumption 1, we have when $G_{n,t} > 0$ (else gradients are null and we take $\bar{G}_{n,t} = G_{n,t}$):

$$\begin{aligned}\mathbb{E}_{t-1}[|\xi_{n,t}|] &\leqslant \int_{\Delta_{n,t}}^\infty \mathbb{P}_{t-1}(|\xi_{n,t}| \geqslant u)\,\mathrm{d}u \\ &\leqslant \int_{\Delta_{n,t}}^\infty \exp\left(-\frac{1}{2}\min\left(\frac{u^2}{\nu^2}, \frac{u}{\mu}\right)\right)\mathrm{d}u \\ &= \int_{\Delta_{n,t}}^{\frac{\nu^2}{\mu}} \exp\left(-\frac{u^2}{2\nu^2}\right)\mathrm{d}u + \int_{\Delta_{n,t}}^\infty \exp\left(-\frac{u}{2\mu}\right)\mathrm{d}u \\ &\leqslant \sqrt{\frac{\pi}{2}}\nu \exp\left(-\frac{\Delta_{n,t}^2}{2\nu^2}\right) + 2\mu \exp\left(-\frac{\Delta_{n,t}}{2\mu}\right) \\ &\leqslant C(\nu,\mu)\exp\left(-\frac{1}{2}\min\left(\frac{\Delta_{n,t}^2}{\nu^2}, \frac{\Delta_{n,t}}{\mu}\right)\right), \end{aligned} \tag{12}$$

where $C(\nu,\mu) = \sqrt{0.5\pi}\nu + 2\mu$. Thus, the clipping bias term is bounded, with $|c_{n,t} - c_n| \leqslant 2C_n, n \in [N]$, as

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_{t-1}[\hat{\mathbf{g}}_t - \bar{\mathbf{g}}_t]^\top(\mathbf{c}_t - \mathbf{c}) &\leqslant \sum_{t=1}^T \sum_{n=1}^N 2C_n \mathbb{E}_{t-1}[|\hat{g}_{n,t} - \bar{g}_{n,t}|] \\ &\underset{(12)}{\leqslant} 2\sum_{n=1}^N \sum_{t=1}^T C_n C(\nu,\mu)\exp\left(-\frac{1}{2}\min\left(\frac{\Delta_{n,t}^2}{\nu^2}, \frac{\Delta_{n,t}}{\mu}\right)\right). \end{aligned} \tag{13}$$

Conclusion. Using a union bound, we finally get with probability at least $1-3\delta$, using (9), (10), (11), (13)

$$\begin{aligned}R_T(\mathbf{c}) \leqslant \sum_{n=1}^N |c_n| &\left(\Xi_1 \left(\sqrt{2\textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2} + \sqrt{2\bar{G}_n^2 \log(\delta^{-1})} \right) + \Xi_2 \sup_{1\leqslant t\leqslant T}(G_{n,t} + \Delta_{n,t}) \right) \\ &+ 2\sum_{n=1}^N \sum_{t=1}^T C_n C(\nu,\mu)\exp\left(-\frac{1}{2}\min\left(\frac{\Delta_{n,t}^2}{\nu^2}, \frac{\Delta_{n,t}}{\mu}\right)\right) \\ &+ 8\left(\sup_{1\leqslant t\leqslant T} \sum_{n=1}^N C_n(G_{n,t} + \Delta_{n,t}\mathbb{1}_{G_{n,t}>0})\right)^2 \log(\delta^{-1}) + \frac{6}{\alpha}\log(\delta^{-1}). \end{aligned} \tag{14}$$

By setting the clipping margin $\Delta_{n,t} = \max(\nu\sqrt{2L_t}, 2\mu L_t)$ with time-dependent $L_t = \log\left(\frac{t\,C(\nu,\mu)}{\log(\delta^{-1})}\right)$, the exponential term simplifies as follows:

$$\begin{aligned}\sum_{t=1}^{T} C(\nu,\mu)\exp\left(-\frac{1}{2}\min\left(\frac{\Delta_{n,t}^2}{\nu^2}, \frac{\Delta_{n,t}}{\mu}\right)\right) &= \sum_{t=1}^{T} C(\nu,\mu)e^{-L_t} \\ &= \sum_{t=1}^{T} C(\nu,\mu)\frac{\log(\delta^{-1})}{t\,C(\nu,\mu)} \\ &= \log(\delta^{-1})\sum_{t=1}^{T}\frac{1}{t} \\ &\leqslant \log(\delta^{-1})(\log(T)+1).\end{aligned}$$

Finally from (14) we get

$$\begin{aligned}R_T(\mathbf{c}) \leqslant \sum_{n=1}^{N}|c_n|\left(\Xi_1\sqrt{2\textstyle\sum_{t=1}^{T}\mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2}+\left(\Xi_1\sqrt{2\log(\delta^{-1})}+\Xi_2\right)(G_n+\Delta)\right) \\ +2\left[\sup_{1\leqslant t\leqslant T}\sum_{n=1}^{N}C_n(\log(T)+1)\mathbb{1}_{G_{n,t}>0}+4\left(\sup_{1\leqslant t\leqslant T}\sum_{n=1}^{N}C_n\left(G_{n,t}+\Delta\mathbb{1}_{G_{n,t}>0}\right)\right)^2+\frac{3}{\alpha}\right]\log(\delta^{-1})\end{aligned} \tag{15}$$

where $\Delta := \sup_{1\leqslant t\leqslant T}\Delta_{n,t} = \max\left\{\nu\sqrt{2\log\left(\frac{TC(\nu,\mu)}{\log(\delta^{-1})}\right)},\quad 2\mu\log\left(\frac{TC(\nu,\mu)}{\log(\delta^{-1})}\right)\right\} \leqslant O(\max(\nu,\mu)\log(T))$ and this completes the proof of Theorem 4. Theorem 1 is obtained by setting $C_1 = \cdots = C_N = C$ and $G_{1,t} = \cdots = G_{N,t} = G, t \geqslant 1$ and because for every $t \geqslant 1$, one has under the square root, with $g_{n,t} = \mathbb{E}_{t-1}[\hat{g}_{n,t}]$ and $|g_{n,t}| \leqslant G$

$$\begin{aligned}\mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2] &\leqslant \mathbb{E}_{t-1}[(|\hat{g}_{n,t}-g_{n,t}|+|g_{n,t}|)^2] \\ &\leqslant \mathbb{E}_{t-1}[|\hat{g}_{n,t}-g_{n,t}|^2]+\mathbb{E}_{t-1}[|g_{n,t}|^2] \\ &\leqslant \sigma^2+G^2.\end{aligned}$$

□

B Proof of Theorem 2

We restate and prove below a complete version of Theorem 2. We consider the general case with a coordinate- and time-dependent clipping rule where we have different expected gradient bounds $\mathbf{G}_t := (G_{n,t})_{1 \leqslant n \leqslant N}$ at each time $t \geqslant 1$.

Theorem 5. *Let $T, N \geqslant 1, (C_n) > 0$. Assume $(\ell_t)_{t \geqslant 1}$ satisfy Assumption 2 and for every $t \geqslant 1, \mathbf{c} \in \mathbb{R}^N, \|\mathbf{c}\|_\infty \leqslant C, \nabla \ell_t(\mathbf{c})$ satisfies Assumption 1 and*

$$|\mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c})_n]| \leqslant G_{n,t},\ n \in [N] \quad \text{and} \quad \max_{1 \leqslant n \leqslant N} \mathbb{E}_{t-1}[|\nabla \ell_t(\mathbf{c})_n - \mathbb{E}_{t-1}[\nabla \ell_t(\mathbf{c})_n]|^2] \leqslant \sigma^2 \,.$$

With sequential predictions $\hat{\mathbf{c}}_t = \sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta, t \geqslant 1$ as in (4) with $(w_t^\Delta)_{\Delta \in \mathcal{D}}$ satisfying Assumption 4, the regret is bounded, for every $\delta \in (0,1)$, with probability $1 - 6\delta$ and for every $\mathbf{c} \in \mathbb{R}^N, |c_n| \leqslant C_n$, as

$$\begin{aligned}
\sum_{t=1}^T L_t(\hat{\mathbf{c}}_t) - L_t(\mathbf{c}) \leqslant O\Bigg[\sum_{n=1}^N |c_n| \Bigg(\Xi_1 \sqrt{\textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2]} + \Big(\Xi_1 \sqrt{\log(\delta^{-1})} + \Xi_2 \Big) \Big(\sup_{1 \leqslant t \leqslant T} G_{n,t} + \mu \log T \Big) \Bigg) \\
+ \left(\frac{\Xi_3^2}{\alpha} + \Xi_4 \sum_{n=1}^N C_n \Big(\sup_{1 \leqslant t \leqslant T} G_{n,t} + 2\mu \log(TN) \mathbb{1}_{G_{n,t} > 0} \Big) \right) \log(\sqrt{T} + 1) \\
+ \left((\alpha + 1) \sup_{1 \leqslant t \leqslant T} \left(\sum_{n=1}^N C_n \Big(G_{n,t} + 2\mu \log(TN) \mathbb{1}_{G_{n,t} > 0} \Big) \right)^2 + \frac{1}{\alpha} \right) \log(\delta^{-1}) \Bigg] \,,
\end{aligned}$$

where the notation $O[\cdot]$ hides multiplicative constants.

Proof of Theorem 5. At each time $t \geqslant 1, \mathbf{c}_t^\Delta \in \mathbb{R}^N$ is the output of Algorithm 1 with clipping margins $\Delta_t = \Delta, t \geqslant 1$ and $\hat{\mathbf{c}}_t = \sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta$ is the average prediction over $\mathcal{D}$. One can decompose the regret, for every $\Delta^* \in \mathcal{D}$, as

$$\sum_{t=1}^T L_t\Big(\sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta \Big) - L_t(\mathbf{c}) = \underbrace{\sum_{t=1}^T L_t(\hat{\mathbf{c}}_t) - L_t(\mathbf{c}_t^{\Delta^*})}_{\text{expert aggregation regret}} + \underbrace{\sum_{t=1}^T L_t(\mathbf{c}_t^{\Delta^*}) - L_t(\mathbf{c})}_{\text{regret of Alg. 1 with } \Delta^* \quad := R_T(\mathbf{c})} \,.$$

Expert aggregation regret. Define $\mathbf{w}_t = (w_t^\Delta)_{\Delta \in \mathcal{D}}$ and one has

$$\hat{\nabla}_t = \nabla_{\mathbf{w}_t} \ell_t \Big(\sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta \Big) = \big(\hat{\mathbf{g}}_t^\top \mathbf{c}_t^\Delta \big)_{\Delta \in \mathcal{D}}$$

where in this part we denote $\hat{\mathbf{g}}_t = \nabla \ell_t(\hat{\mathbf{c}}_t) \in \mathbb{R}^N$ the gradient of ℓ_t evaluated at $\hat{\mathbf{c}}_t$, which satisfies Assumption 1 and $\mathbb{E}_{t-1}[\nabla \ell_t(\hat{\mathbf{c}}_t)_n]| \leqslant G_{n,t}$. Therefore, denoting $\hat{\nabla}_t = (\hat{\nabla}_t^\Delta)_{\Delta \in \mathcal{D}}$ and letting $Z_t = \hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_t^{\Delta^*}$, by Assumption 2 one has

$$\begin{aligned}
\sum_{t=1}^T L_t\Big(\sum_{\Delta \in \mathcal{D}} w_t^\Delta \mathbf{c}_t^\Delta \Big) - L_t(\mathbf{c}_t^{\Delta^*}) &\leqslant \sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] - \frac{\alpha}{2} \mathbb{E}_{t-1}[Z_t^2] \\
&\leqslant \sum_{t=1}^T Z_t + \frac{\gamma}{2} Z_t^2 + \frac{\gamma - \alpha}{2} \mathbb{E}_{t-1}[Z_t^2] + \frac{2}{\gamma} \log(\delta^{-1}) \qquad (16)
\end{aligned}$$

where last inequality holds with high probability $1 - \delta$ for any $\gamma > 0$ by Lemma 3. Moreover, one can bound the first order term by Assumption 4 with $Z_t = \hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_t^{\Delta^*}$

$$\sum_{t=1}^T Z_t \leqslant \Xi_3 \sqrt{\log(|\mathcal{D}|) \sum_{t=1}^T Z_t^2} + \Xi_4 \log(|\mathcal{D}|) \sup_{1 \leqslant t \leqslant T} \|\hat{\nabla}_t\|_\infty .$$

Further, by Young's inequality, we have for every $\eta > 0$

$$\Xi_3 \sqrt{\log(|\mathcal{D}|) \sum_{t=1}^{T} Z_t^2} \leqslant \frac{\Xi_3^2 \log(|\mathcal{D}|)}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} Z_t^2.$$

And in bound (16) we have

$$\sum_{t=1}^{T} Z_t + \frac{\gamma}{2} Z_t^2 \leqslant \frac{\Xi_3^2 \log(|\mathcal{D}|)}{2\eta} + \frac{\eta + \gamma}{2} \sum_{t=1}^{T} Z_t^2 + \Xi_4 \log(|\mathcal{D}|) \sup_{1 \leqslant t \leqslant T} \|\hat{\nabla}_t\|_\infty. \tag{17}$$

Now we want to relate $\sum_{t=1}^{T} Z_t^2$ to its clipped version using the fact that $\hat{\mathbf{g}}_t$ has coordinate with sub-exponential noise (Assumption 1). Indeed, remark that for every $t \geqslant 1$

$$Z_t = \hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_t^{\bar{G}^*} = \sum_{\Delta \in \mathcal{D}} \sum_{n=1}^{N} \hat{g}_{n,t} c_{n,t}^{\Delta} w_t^{\Delta} - \sum_{n=1}^{N} \hat{g}_{n,t} c_{n,t}^{\Delta} = \sum_{n=1}^{N} \hat{g}_{n,t} \left(\sum_{\Delta \in \mathcal{D}} w_t^{\Delta} c_{n,t}^{\Delta} - c_{n,t}^{\Delta^*} \right)$$

is still sub-exponential as a sum of sub-exponential random variables $\hat{g}_{n,t}$. We define a clipped version of Z_t, $\bar{Z}_t = \sum_{n=1}^{N} \bar{g}_{n,t} \left(\sum_{\Delta \in \mathcal{D}} w_t^{\Delta} c_{n,t}^{\Delta} - c_{n,t}^{\Delta^*} \right)$ where $\bar{g}_{n,t} = \mathrm{clip}(\hat{g}_{n,t}, \tau_{n,t}), n \in [N]$ for some $\tau_{n,t} > 0$ to be optimized in the proof and that only appears in the analysis. Then, we have

$$\sum_{t=1}^{T} Z_t^2 = \sum_{t=1}^{T} \bar{Z}_t^2 \tag{18}$$

with probability $\mathbb{P}(\forall 1 \leqslant t \leqslant T, Z_t = \bar{Z}_t)$. By the union bound,

$$\mathbb{P}\big(\forall\, 1 \leqslant t \leqslant T,\ Z_t = \bar{Z}_t\big) = 1 - \mathbb{P}\big(\exists\, 1 \leqslant t \leqslant T :\ Z_t \neq \bar{Z}_t\big) \ \geqslant\ 1 - \sum_{t=1}^{T} \mathbb{P}_{t-1}\big(Z_t \neq \bar{Z}_t\big).$$

In particular, for every $t \geqslant 1$,

$$\begin{aligned}
\mathbb{P}_{t-1}\big(Z_t \neq \bar{Z}_t\big) &= \mathbb{P}_{t-1}\left(\sum_{n=1}^{N} \hat{g}_{n,t} \Big(\sum_{\nu} w_t^{\nu} c_{n,t}^{\nu} - c_{n,t}^{\bar{G}^*} \Big) \neq \sum_{n=1}^{N} \bar{g}_{n,t} \Big(\sum_{\nu} w_t^{\nu} c_{n,t}^{\nu} - c_{n,t}^{\bar{G}^*} \Big) \right) \\
&\leqslant \sum_{n=1}^{N} \mathbb{P}_{t-1}\left(\hat{g}_{n,t} \Big(\sum_{\nu} w_t^{\nu} c_{n,t}^{\nu} - c_{n,t}^{\bar{G}^*} \Big) \neq \bar{g}_{n,t} \Big(\sum_{\nu} w_t^{\nu} c_{n,t}^{\nu} - c_{n,t}^{\bar{G}^*} \Big) \right) \\
&= \sum_{n=1}^{N} \mathbb{P}_{t-1}(\hat{g}_{n,t} \neq \bar{g}_{n,t}) \\
&= \sum_{n=1}^{N} \mathbb{P}_{t-1}(|\hat{g}_{n,t}| > \tau_{n,t}) \\
&\leqslant \sum_{n=1}^{N} \mathbb{P}_{t-1}(|\hat{g}_{n,t} - \mathbb{E}_{t-1}[\hat{g}_{n,t}]| > \tau_{n,t} - G_{n,t}) \\
&\leqslant \sum_{n=1}^{N} C(\nu, \mu) \exp\left(-\frac{1}{2} \min\left(\frac{(\tau_{n,t} - G_{n,t})^2}{\nu^2}, \frac{\tau_{n,t} - G_{n,t}}{\mu} \right) \right),
\end{aligned}$$

where second last inequality is because $|\mathbb{E}_{t-1}[\hat{g}_{n,t}]| \leqslant G_{n,t}, n \in [N]$, and last inequality is by Assumption 1 with $C(\nu, \mu) = \sqrt{0.5\pi}\nu + 2\mu$. Finally, taking

$$\begin{aligned}
\tau_{n,t} = G_{n,t} + \nu \sqrt{2 \log\left(\tfrac{NTC(\nu,\mu)}{\delta}\right)} \vee 2\mu \log\left(\tfrac{NTC(\nu,\mu)}{\delta}\right) \\
\implies C(\nu, \mu) \exp\left(-\frac{1}{2} \min\left(\frac{(\tau_{n,t} - G_{n,t})^2}{\nu^2}, \frac{\tau_{n,t} - G_{n,t}}{\mu} \right) \right) \leqslant \frac{\delta}{TN}
\end{aligned}$$

and Equality (18) holds with probability at least $1 - \sum_{t=1}^{T}\sum_{n=1}^{N} \delta \frac{1}{TN} = 1 - \delta$. With $\sum_{\Delta\in\mathcal{D}} w_t^\Delta = 1, t \geqslant 1$ and such clipping, we have for every $t \geqslant 1$

$$|\bar{Z}_t| = \left| \sum_{n=1}^{N} \bar{g}_{n,t} \Big(\sum_{\Delta\in\mathcal{D}} w_t^\Delta c_{n,t}^\Delta - c_{n,t}^{\Delta^*} \Big) \right| \leqslant \sup_{1\leqslant t\leqslant T} \sum_{n=1}^{N} 2C_n \big(G_{n,t} + \Delta(\tau_{n,t}) \big) =: D.$$

where $\Delta(\tau_{n,t}) := \nu\sqrt{2\log\left(\frac{NTC(\nu,\mu)}{\delta}\right)} \vee 2\mu\log\left(\frac{NTC(\nu,\mu)}{\delta}\right) \leqslant 2(\mu,\nu)\log\left(\frac{NTC(\nu,\mu)}{\delta}\right) = O(\log(TN))$ provided that $\log\left(\frac{NTC(\nu,\mu)}{\delta}\right) \geqslant \frac{1}{2}$, i.e. $T \geqslant \exp(1/2)\delta/(NC(\nu,\mu))$.
By Corollary 1, with $0 \leqslant \bar{Z}_t^2 \leqslant D^2$, one has with probability $1-\delta$

$$\sum_{t=1}^{T} \bar{Z}_t^2 \leqslant 2\sum_{t=1}^{T} \mathbb{E}_{t-1}[\bar{Z}_t^2] + 2D^2\log(\delta^{-1}).$$

Finally, with probability $1-2\delta$ (union bound), one has in Equation (17)

$$\frac{\eta+\gamma}{2}\sum_{t=1}^{T} Z_t^2 \leqslant (\eta+\gamma)\sum_{t=1}^{T} \mathbb{E}_{t-1}[\bar{Z}_t^2] + (\eta+\gamma)D^2\log(\delta^{-1}),$$

which gives after combining Equations (16), (17) and last inequality, since $\frac{\gamma-\alpha}{2}\mathbb{E}_{t-1}[Z_t^2] \leqslant \frac{\gamma-\alpha}{2}\mathbb{E}_{t-1}[\bar{Z}_t^2]$,$\gamma < \alpha$ and $|\bar{Z}_t| \leqslant |Z_t|$

$$\sum_{t=1}^{T} L_t(\hat{\mathbf{c}}_t) - L_t(\mathbf{c}_t^{\Delta^*}) \leqslant \frac{2\eta+3\gamma-\alpha}{2}\sum_{t=1}^{T}\mathbb{E}_{t-1}[\bar{Z}_t^2] + \frac{\Xi_3^2\log(|\mathcal{D}|)}{2\eta} + \Xi_4\log(|\mathcal{D}|)\sup_{1\leqslant t\leqslant T}\|\hat{\nabla}_t\|_\infty + \left((\eta+\gamma)D^2 + \frac{2}{\gamma}\right)\log(\delta^{-1}) \quad (19)$$

with probability $1-3\delta$. Finally, one has to control $\sup_{1\leqslant t\leqslant T}\|\hat{\nabla}_t\|_\infty$ with high probability. Indeed,

$$\|\hat{\nabla}_t\|_\infty = \sup_{\Delta\in\mathcal{D}} |\hat{\mathbf{g}}_t^\top \mathbf{c}_t^\Delta| \leqslant \sup_{1\leqslant t\leqslant T}\sum_{n=1}^{N} C_n|\hat{g}_{n,t}|$$

since $c_{n,t}^{\mathcal{D}} \in [-C_n, C_n]$ for every $\Delta\in\mathcal{D}$ and taking the sup over $1\leqslant t\leqslant T$, one has

$$\sup_{1\leqslant t\leqslant T}\|\hat{\nabla}_t\|_\infty \leqslant \sup_{1\leqslant t\leqslant T}\sum_{n=1}^{N} C_n|\hat{g}_{n,t}| = \sup_{1\leqslant t\leqslant T}\sum_{n=1}^{N} C_n|\bar{g}_{n,t}| \quad (20)$$

where last inequality holds probability at least $1-\delta$ with again $\bar{g}_{n,t} = \mathrm{clip}(\hat{g}_{n,t}, \tau_{n,t})$ using the same clipping threshold as before since

$$\mathbb{P}\left(\forall\, 1\leqslant t\leqslant T,\ \sum_{n=1}^{N} C_n|\hat{g}_{n,t}| = \sum_{n=1}^{N} C_n|\bar{g}_{n,t}|\right) \geqslant 1 - \sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{P}(|\hat{g}_{n,t}| > \tau_{n,t}) \geqslant 1-\delta.$$

Then, taking $\eta = \frac{\alpha-3\gamma}{2}, \gamma < \frac{\alpha}{3} < \alpha$ cancels the second order terms in (19) and gives, combined with (20)

$$\sum_{t=1}^{T} L_t(\hat{\mathbf{c}}_t) - L_t(\mathbf{c}_t^{\Delta^*}) \leqslant \frac{\Xi_3^2\log(|\mathcal{D}|)}{\alpha-3\gamma} + \Xi_4\log(|\mathcal{D}|)\sum_{n=1}^{N} C_n\tau_n + \left(\frac{\alpha-\gamma}{2}\left(2\sum_{n=1}^{N} C_n\tau_n\right)^2 + \frac{2}{\gamma}\right)\log(\delta^{-1}) \quad (21)$$

with probability $1-4\delta$ (union bound) and $\tau_n = \sup_{1\leqslant t\leqslant T} G_{n,t} + O(\max(\nu,\mu)\log\left(\frac{TN}{\delta}\right))$.

Regret $R_T(\mathbf{c})$ of Algorithm 1 with clipping margin Δ^*. Let

$$\Delta(\nu,\mu)=\max\left\{\nu\sqrt{2\log\left(\tfrac{(\sqrt{0.5\pi}+2\mu)T}{\log(\delta^{-1})}\right)},2\mu\log\left(\tfrac{(\sqrt{0.5\pi}+2\mu)T}{\log(\delta^{-1})}\right)\right\}$$

be the optimal clipping margin in Theorem 2 and define $\Delta^*=\arg\min_{\Delta>\Delta(\nu,\mu)\in\mathcal{D}}|\Delta-\Delta(\nu,\mu)|$ the closest upper margin in the grid $\mathcal{D}$.
From Equation (14) in Proof of Theorem 4 (Appendix A), with clipping margin Δ^*, $G_n=\sup_{1\leqslant t\leqslant T}G_{n,t}$ and $\hat{g}_{n,t}=\nabla\ell_t(\mathbf{c}_t^{\Delta^*}),|\mathbb{E}_{t-1}[\hat{g}_{n,t}]|\leqslant G_{n,t},n\in[N],t\geqslant 1$ (and assumptions of Theorem 4 are all satisfied), one has

$$\begin{aligned}R_T(\mathbf{c}^\mu)\leqslant\sum_{n=1}^N|c_n|\Bigg(&\Xi_1\sqrt{2\textstyle\sum_{t=1}^T\mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2}+\Big(\Xi_1\sqrt{2\log(\delta^{-1})}+\Xi_2\Big)(G_n+\Delta^*)\Bigg)\\&+2\sum_{n=1}^N\sum_{t=1}^T C_nC(\nu,\mu)\exp\left(-\frac{1}{2}\min\left(\frac{(\Delta^*)^2}{\nu^2},\frac{\Delta^*}{\mu}\right)\right)\\&+8\Bigg(\sup_{1\leqslant t\leqslant T}\sum_{n=1}^N C_n(G_{n,t}+\Delta^*\mathbb{1}_{G_{n,t}>0})\Bigg)^2\log(\delta^{-1})+\frac{6}{\alpha}\log(\delta^{-1})\end{aligned}\tag{22}$$

with probability $1-2\delta$. Since $\Delta^*>\Delta(\nu,\mu)$ we still have (see below (14))

$$\sum_{t=1}^T C(\nu,\mu)\exp\left(-\frac{1}{2}\min\left(\frac{(\Delta^*)^2}{\nu^2},\frac{\Delta^*}{\mu}\right)\right)\leqslant\log(\delta^{-1})(\log(T)+1)$$

and by the precision of the grid $\mathcal{D}$, one has $|\Delta^*-\Delta(\mu,\nu)|\leqslant 1,\Delta(\mu,\nu)=O((\nu\vee\mu)\log(T))$ and then the deviation term

$$O\Bigg(\sup_{1\leqslant t\leqslant T}\sum_{n=1}^N C_n(G_{n,t}+\Delta^*\mathbb{1}_{G_{n,t}>0})\Bigg)=O\Bigg(\sup_{1\leqslant t\leqslant T}\sum_{n=1}^N C_n(G_{n,t}+\Delta(\nu,\mu))\mathbb{1}_{G_{n,t}>0}\Bigg)$$

is still of the same order $O\Big(\log^2 T\log(\delta^{-1})\big(\sum_{n=1}^N C_n\big)^2\Big)$.

Conclusion. Combining (21) and (22) one has with probability $1-6\delta$ and $\gamma=\frac{\alpha}{4}<\frac{\alpha}{3}$

$$\begin{aligned}\sum_{t=1}^T L_t(\hat{\mathbf{c}}_t)-L_t(\mathbf{c}^\mu)\leqslant\sum_{n=1}^N|c_n|\Bigg[&\Xi_1\sqrt{2\textstyle\sum_{t=1}^T\mathbb{E}_{t-1}[|\hat{g}_{n,t}|^2}+\Big(\Xi_1\sqrt{2\log(\delta^{-1})}+\Xi_2\Big)(G_n+\tau)\Bigg]\\&+\frac{4\Xi_3^2\log(|\mathcal{D}|)}{\alpha}+\Xi_4\log(|\mathcal{D}|)\sum_{n=1}^N C_n\Big(G_n+\tau\Big)\\&+\Bigg[\frac{3\alpha}{8}\Bigg(2\sum_{n=1}^N C_n\Big(G_n+\tau\Big)^2+4(\log(T)+1)\sum_{n=1}^N C_n\\&\qquad+8\Bigg(\sum_{n=1}^N C_n(G_n+\tau)\Bigg)^2+\frac{14}{\alpha}\Bigg]\log(\delta^{-1}).\end{aligned}$$

where $\tau=O\big(\max(\nu,\mu)\log\left(\frac{TN}{\delta}\right)\big)$ and this concludes the proof with $|\mathcal{D}|\leqslant\sqrt{T}+1$. □

Remark. Considering sparse expected gradients $\mathbf{g}_t=\mathbb{E}_{t-1}[\hat{\mathbf{g}}_t]$ the sum $\sum_{n=1}^N$ in the deviation terms can be reduced, at each time as a sum over $n\in\operatorname{supp}(\mathbf{g}_t)$, provided that $G_{n,t}=0$ when $g_{n,t}=0$. This will be useful in particular in the non-parametric setting where N is exponential, but only a small fraction $|\operatorname{supp}(\mathbf{g}_t)|$ of coordinates are non zero (active) at time each time t, hence a much smaller diameter in the deviation terms.

C Assumption 2 in Section 3

In the squared loss regression setting of Section 3, the condition of Assumption 2 follows directly from boundedness of the predictions (see Lemma 4) and second order moment on $Y_t = f(X_t) + \varepsilon_t$. Denote $\hat{f}_t(\mathbf{c}) := \hat{f}_t(X_t, \mathbf{c})$ the prediction at time t, linear in $\mathbf{c}$. Indeed, for losses of the form

$$\ell_t(\mathbf{c}) = (\hat{f}_t(\mathbf{c}) - Y_t)^2,$$

one has

$$\ell_t(\mathbf{c}_1) - \ell_t(\mathbf{c}_2) = \nabla \ell_t(\mathbf{c}_1)^\top (\mathbf{c}_1 - \mathbf{c}_2) - \big(\hat{f}_t(\mathbf{c}_1) - \hat{f}_t(\mathbf{c}_2)\big)^2$$

Hence Assumption 2 holds for every $\alpha > 0$ such that

$$\frac{\alpha}{2}\mathbb{E}_{t-1}\Big[\big(\nabla \ell_t(\mathbf{c}_1)^\top (\mathbf{c}_1 - \mathbf{c}_2)\big)^2\Big] \leqslant \mathbb{E}_{t-1}\Big[\big(\hat{f}_t(\mathbf{c}_1) - \hat{f}_t(\mathbf{c}_2)\big)^2\Big].$$

With

$$\begin{aligned}(\nabla \ell_t(\mathbf{c}_1)^\top (\mathbf{c}_1 - \mathbf{c}_2))^2 &= 4\big(\hat{f}_t(\mathbf{c}_1) - Y_t\big)^2\big(\hat{f}_t(\mathbf{c}_1) - \hat{f}_t(\mathbf{c}_2)\big)^2 \\ &\leqslant 8\big((\hat{f}_t(\mathbf{c}_1) - f(X_t))^2 + \varepsilon_t^2\big)\big(\hat{f}_t(\mathbf{c}_1) - \hat{f}_t(\mathbf{c}_2)\big)^2,\end{aligned}$$

and

$$|\hat{f}_t(\mathbf{c})| \leqslant \hat{B}_\infty \text{ a.s. (Lemma 4)}, \quad |f(X_t)| \leqslant B_\infty \quad \text{and} \quad \mathbb{E}_{t-1}[\varepsilon_t^2 \mid X_t] \leqslant \sigma^2,$$

one has

$$\frac{\alpha}{2}\mathbb{E}_{t-1}\Big[\big(\nabla \ell_t(\mathbf{c}_1)^\top (\mathbf{c}_1 - \mathbf{c}_2)\big)^2\Big] \leqslant 4\alpha((\hat{B}_\infty + B_\infty)^2 + \sigma^2)\mathbb{E}_{t-1}[\big(\hat{f}_t(\mathbf{c}_1) - \hat{f}_t(\mathbf{c}_2)\big)^2]$$

and so one may take $\alpha = \frac{1}{4((\hat{B}_\infty + B_\infty)^2 + \sigma^2)}$ to satisfy Assumption 2.

D Proof of Theorem 3

Proof of Theorem 3. Let $f \in \mathcal{B}^s_{pq}(B)$ be fixed and $Y = f(X) + \varepsilon$ with $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] = \sigma^2 < \infty$. Recall that the training data set $((X_t, Y_t))_{1 \leqslant t \leqslant T} \sim (X, Y)$ is i.i.d. and with same law as the test pair (X, Y).

Decomposition of risk. For the square loss, one has for any predictor $\hat{f}$ this nice decomposition of the expected risk

$$\begin{aligned}\mathbb{E}_{XY}[(\hat{f}(X) - Y)^2] &= \mathbb{E}_X[\mathbb{E}[(\hat{f}(X) - Y)^2|X]] \\ &= \mathbb{E}_X[\mathbb{E}[(\hat{f}(X) - f(X) - \xi)^2|X]] \\ &= \mathbb{E}_X[\mathbb{E}[(\hat{f}(X) - f(X))^2|X]] + \mathbb{E}_X[\mathbb{E}[\xi^2|X]] - 2\mathbb{E}_X[\mathbb{E}[(\hat{f}(X) - f(X))\xi|X]] \\ &= \mathbb{E}_X[(\hat{f}(X) - f(X))^2] + \mathbb{E}[\xi^2] \\ &= \|\hat{f} - f\|_2^2 + \sigma^2. \end{aligned} \tag{23}$$

Going back to the excess risk and letting $\hat{f} = \bar{f}_T = \frac{1}{T}\sum_{t=1}^T \hat{f}_t$ in (23), after direct algebraic calculations one has

$$\begin{aligned}\|\bar{f}_T - f\|_2^2 &= \mathbb{E}_{XY}[(\bar{f}_T(X) - Y)^2] - \mathbb{E}_{XY}[(f(X) - Y)^2] \\ &\leqslant \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{XY}[(\hat{f}_t(X) - Y)^2] - \mathbb{E}_{XY}[(f(X) - Y)^2], \end{aligned} \tag{24}$$

by Jensen's inequality and convexity of the square loss. Further, since $(X_t, Y_t) \sim P_{(X,Y)}$ and is independent of $\mathcal{F}_{t-1}$ and $\hat{f}_t$ is $\mathcal{F}_{t-1}$ measurable, one has for every $t \geqslant 1$

$$\mathbb{E}_{XY}[(\hat{f}_t(X) - Y)^2] = \mathbb{E}_{t-1}[(\hat{f}_t(X_t) - Y_t)^2],$$

where we used the notation $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. And finally,

$$\|\bar{f}_T - f\|_2^2 \leqslant \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{t-1}[(\hat{f}_t(X_t) - Y_t)^2] - \mathbb{E}_{t-1}[(f(X_t) - Y_t)^2]}_{=:R_T(f)} \tag{25}$$

The problem now turns out to analyse the (stochastic) regret $R_T(f)$ and this situation mirrors the problem studied in Section 2 with losses $L_t(\hat{f}_t) = \mathbb{E}_{t-1}[\ell_t(\hat{f}_t(X_t))], t \geqslant 1$ where $\ell_t(\hat{f}_t(X_t)) = (\hat{f}_t(X_t) - Y_t)^2$ and ℓ_t's randomness is due to Y_t's one.
The quantity of interest $R_T(f)$ then decomposes for every oracle estimator $\hat{f}^*$, into an estimation regret term $R_T(\hat{f}^*)$ plus a bias (or approximation regret) term

$$R_T(f) = R_T(\hat{f}^*) + \sum_{t=1}^{T} \mathbb{E}_{t-1}[\ell_t(\hat{f}^*(X_t))] - \mathbb{E}_{t-1}[\ell_t(f(X_t))] \underset{(23)}{=} R_T(\hat{f}^*) + T\|\hat{f}^* - f\|_2^2,$$

with $\ell_t(\hat{f}) = (\hat{f} - Y_t)^2$.

Step 1: Bound on the estimation regret $R_T(\hat{f}^*)$. One has $\hat{f}_t(X_t) = \sum_e w_{e,t}\hat{f}_{e,t}(X_t)$ (average over all experts). Consider $e^* = (j_0^*, \Delta^*, \boldsymbol{\alpha}^*) \in [J_0] \times \mathcal{D} \times \mathcal{A}_{j_0^*}$ (j_0^* oracle starting scale, $\boldsymbol{\alpha}^*$ best scaling coefficients on the grid $\mathcal{A}_{j_0^*}$, and Δ^* the best clipping margin). We then define the expert with best parameters e^* at every time $t \geqslant 1$, $\hat{f}_{e^*,t}$ and the oracle regressor $\hat{f}^*_{e^*}(X_t)$ with oracle coefficients $(c_{j,k})$ and oracle parameters e^*. We have at every time $t \geqslant 1$, the following decomposition

$$\begin{aligned} R_T(\hat{f}^*) &= \sum_{t=1}^{T} L_t(\hat{f}_t(X_t)) - L_t(\hat{f}^*) \\ &= \underbrace{\sum_{t=1}^{T} L_t(\hat{f}_t) - L_t(\hat{f}_{e^*,t}(X_t))}_{\text{expert aggregation regret} \quad =:R_1} + \underbrace{\sum_{t=1}^{T} L_t(\hat{f}_{e^*,t}(X_t)) - L_t(\hat{f}^*_{e^*}(X_t))}_{\text{regret of Alg.1} \quad =:R_2}. \end{aligned}$$

1. **Bound on the expert aggregation regret R_1.** Let $\hat{g}_t = \ell_t'(\hat{f}_t(X_t)) = 2(\hat{f}_t(X_t) - Y_t)$ be the noisy observed gradient of $L_t'(\hat{f}_t(X_t))$. In particular, $\hat{g}_t$ satisfies Assumption 1 and

$$|\mathbb{E}_{t-1}[\hat{g}_t]| = 2\mathbb{E}_{t-1}[\hat{f}_t(X_t) - f(X_t) - \varepsilon_t]| \leqslant 2(\hat{B}_\infty + B) =: G.$$

The gradient experts $\hat{\nabla}_t = (\nabla_{e,t})_{e\in\mathcal{E}}$ writes $\hat{\nabla}_{e,t} = \hat{g}_t \cdot \hat{f}_{e,t}(X_t)$ and $|\mathbb{E}_{t-1}[\hat{\nabla}_{e,t}]| \leqslant G\hat{B}_\infty$ where $\hat{B}_\infty$ is the sup-norm on the experts, bounded as in Lemma 4. Therefore, letting $Z_{e^*,t} = \hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_{e^*,t}$, by Assumption 2 on (ℓ_t) one has

$$\begin{aligned} \sum_{t=1}^{T} L_t(\hat{f}_t) - L_t(\hat{f}_{e^*,t}) &\leqslant \sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_{e^*,t}] - \frac{\alpha}{2}\mathbb{E}_{t-1}[Z_{e^*,t}^2] \\ &\leqslant \sum_{t=1}^{T} Z_{e^*,t} + \frac{\gamma}{2}Z_{e^*,t}^2 + \frac{\gamma-\alpha}{2}\mathbb{E}_{t-1}[Z_{e^*,t}^2] + \frac{2}{\gamma}\log(\delta^{-1}) \end{aligned} \tag{26}$$

where last inequality holds with high probability $1-\delta$ for any $\gamma > 0$ by Lemma 3. Moreover, one can bound the first order term by Assumption 4 with $Z_{e^*,t} = \hat{\nabla}_t^\top \mathbf{w}_t - \hat{\nabla}_{e^*,t}$

$$\sum_{t=1}^{T} Z_{e^*,t} \leqslant \Xi_3\sqrt{\log(|\mathcal{E}|)\sum_{t=1}^{T} Z_{e^*,t}^2} + \Xi_4\log(|\mathcal{E}|)\sup_{1\leqslant t\leqslant T}\|\hat{\nabla}_t\|_\infty.$$

Further, by Young's inequality, we have for every $\eta > 0$

$$\Xi_3 \sqrt{\log(|\mathcal{E}|) \sum_{t=1}^{T} Z_{e^*,t}^2} \leqslant \frac{\Xi_3^2 \log(|\mathcal{E}|)}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} Z_{e^*,t}^2.$$

And in bound (26) we have

$$\sum_{t=1}^{T} Z_{e^*,t} + \frac{\gamma}{2} Z_{e^*,t}^2 \leqslant \frac{\Xi_3^2 \log(|\mathcal{E}|)}{2\eta} + \Xi_4 \log(|\mathcal{E}|) \sup_{1 \leqslant t \leqslant T} \|\hat{\nabla}_t\|_\infty + \frac{\eta + \gamma}{2} \sum_{e^* \in \mathcal{E}^*} \sum_{t=1}^{T} Z_{e^*,t}^2. \quad (27)$$

Further, since $\hat{g}_t$ satisfies Assumption 1 with some parameters (ν, μ), we have $\sup_{1 \leqslant t \leqslant T} \|\hat{\nabla}_t\|_\infty = \hat{B}_\infty \sup_{1 \leqslant t \leqslant T} |\hat{g}_t|$ that can be bounded, with probability $1 - \delta$, as

$$\sup_{1 \leqslant t \leqslant T} \|\hat{\nabla}_t\|_\infty \leqslant \tau \hat{B}_\infty \quad (28)$$

because $\forall 1 \leqslant t \leqslant T, g_t = \mathrm{clip}(g_t, \tau)$ with probability $1 - \delta$ as with $\tau = G + O((\nu \vee \mu) \log(T/\delta))$ under Assumption 1. Then, with $Z_{e^*,t} = \nabla_t^\top \mathbf{w}_t - \hat{\nabla}_{e^*,t} = \hat{g}_t(\hat{f}_t(X_t) - \hat{f}_{e^*,t}(X_t))$, one has also with high probability $1 - \delta$,

$$\sum_{t=1}^{T} Z_{e^*,t} = \sum_{t=1}^{T} \mathrm{clip}(g_t, \tau)(\hat{f}_t(X_t) - \hat{f}_{e^*,t}(X_t)),$$

Setting $\bar{Z}_{e^*,t} := \mathrm{clip}(g_t, \tau)(\hat{f}_t(X_t) - \hat{f}_{e^*,t}(X_t))$, we can use Corollary 1, with $0 \leqslant (\bar{Z}_{e^*,t})^2 \leqslant 4\tau^2 \hat{B}_\infty^2$ and we finally have with probability $1 - 2\delta$ (union bound)

$$\sum_{t=1}^{T} Z_{e^*,t}^2 \leqslant 2 \sum_{t=1}^{T} \mathbb{E}_{t-1}[\bar{Z}_{e^*,t}^2] + 8\tau^2 \hat{B}_\infty^2 \log(\delta^{-1}). \quad (29)$$

Plugging (27), (28) and (29) in (26), we get with probability $1 - 4\delta$ (union bound)

$$R_1 \leqslant \frac{2\eta + 3\gamma - \alpha}{2} \sum_{t=1}^{T} \mathbb{E}_{t-1}[\bar{Z}_{e^*,t}^2] + \log(|\mathcal{E}|) \left(\frac{\Xi_3^2}{2\eta} + \Xi_4 \tau \hat{B}_\infty \right) + 2 \left(4\tau^2 \hat{B}_\infty^2 + \frac{1}{\gamma} \right) \log(\delta^{-1}) \quad (30)$$

taking $0 < \gamma < \alpha$ so that $\frac{\gamma - \alpha}{2} \mathbb{E}_{t-1}[Z_{e^*,t}^2] \leqslant \frac{\gamma - \alpha}{2} \mathbb{E}_{t-1}[\bar{Z}_{e^*,t}^2]$. Finally, taking $0 < \gamma < \frac{\alpha}{3}, \eta = \frac{\alpha - 3\gamma}{2} > 0$ cancels the second order term and we have, with probability $1 - 4\delta$

$$R_1 \leqslant \log(|\mathcal{E}|) \left(\frac{\Xi_3^2}{\alpha - 3\gamma} + \Xi_4 \tau \hat{B}_\infty \right) + 2 \left(4\tau^2 \hat{B}_\infty^2 + \frac{1}{\gamma} \right) \log(\delta^{-1}). \quad (31)$$

with $\tau = G + O((\nu \vee \mu) \log(T/\delta))$

2. **Bound on R_2 via Theorem 1.** We can apply Theorem 1 to bound R_2. Recall that $\hat{f}_{e^*}^*$ is an oracle wavelet regressor (oracle coefficients, and oracle starting resolution) and

$$\hat{f}_{e^*,t} = \underbrace{\sum_{k \in \bar{\Lambda}_{j_0^*}} \alpha_{j_0^*,k,t}\, \phi_k}_{\text{scaling part}} + \underbrace{\sum_{j=j_0^*}^{j_0^*+J} \sum_{k \in \Lambda_j} \beta_{j,k,t}\, \psi_{j,k}}_{\text{detail part}}.$$

where $\alpha_{j_0^*,k,t}$ denotes scaling coefficients at time t, starting at $\alpha_{j,k,1}^* \in \mathcal{A}$, and $\beta_{j,k}, j \geqslant j_0^*$ are wavelet coefficients up to some optimized truncation level $J = \frac{S}{(2S+d)\kappa} \log(TB^2\sigma_0^{-2})$ (large

enough to kill the approximation term). Applying Theorem 1 (with Assumption 2 satisfied — see Appendix C) on coefficients $\{\alpha_{j,k}, \beta_{j,k}\}$, with oracle clipping margin $\Delta_{e^*} \propto \max(\nu,\mu)\log(T) \in \mathcal{D}$, with probability $1-2\delta$

$$\sum_{t=1}^{T} L_t(\hat{f}_{e^*,t}(X_t)) - L_t(\hat{f}^*_{e^*}(X_t)) \leqslant C \sum_{j,k} |c_{j,k} - c_{j,k,1}| \sqrt{\textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[\hat{g}^2_{j,k,t}]} + O\Big(\Big(\sup_{1\leqslant t\leqslant T} \sum_{j,k} C_{j,k}(G_{j,k,t} + \max(\nu,\mu)\log(T)) \mathbb{1}_{G_{j,k,t}>0} \Big)^2 \log(\delta^{-1}) \Big)$$

where $c_{j,k}$ stands for either $\alpha_{j,k}$ or $\beta_{j,k}$ and C is deduced from Theorem 4 and can include $\log T$ terms. We study the bound in 4 parts:

- Scaling coefficients: since the scaling coefficients at level j_0^* starts on a grid of precision $\epsilon_{j_0^*} = B_\infty \|\phi\|_1 (2^{j_0^* d/2}\sqrt{T})^{-1}$ (the grid is of size $O(\sqrt{T})$ for every level j_0^*)

$$\sum_{k\in\bar{\Lambda}_{j_0^*}} \underbrace{|\alpha_{j_0^*,k} - \alpha_{j_0^*,k,1}|}_{\leqslant \frac{\epsilon_{j_0^*}}{2}} \sqrt{\sum_{t=1}^T \mathbb{E}_{t-1}[\hat{g}^2_{j,k,t}]} \leqslant \epsilon_{j_0^*} |\bar{\Lambda}_{j_0^*}|^{\frac{1}{2}} \sqrt{\sum_{k\in\bar{\Lambda}_{j_0^*}} \sum_{t=1}^T \mathbb{E}_{t-1}[\hat{g}^2_{j,k,t}]}$$
$$\leqslant \frac{B_\infty \|\phi\|_1 \lambda^{\frac{1}{2}}}{\sqrt{T}} \sqrt{M_\phi \|\phi\|_\infty T} = (\lambda M_\phi \|\phi\|_\infty)^{\frac{1}{2}}$$

  where we used Jensen's inequality, $|\bar{\Lambda}_{j_0^*}| \leqslant \lambda 2^{j_0^* d}$, the form of the gradients $\hat{g}_{j,k,t} = \hat{g}_t \cdot \phi_k(X_t)$ and the fact that $\sum_k |\phi_k(X_t)|^2 \leqslant \|\phi\|_\infty \cdot \|\sum_k |\phi(\cdot - k)|\|_\infty = \|\phi\|_\infty \cdot M_\phi < \infty$.

- Detail coefficients: We use the same oracle as in [24, Theorem 1] that was designed for a control on the L^∞ bias and works also for the L^2-norm. We have a linear part and non linear part in the wavelet oracle: we keep all the detail coefficients in between $[j_0^*, j_0^* + J^*]$ up to some level $j_0^* + J^*$ and then retain only greatest coefficients $\beta_{j,k}$ from level $[j_0^* + J^*, J]$.

  - Linear part: $j = j_0^*, \dots, J^* + j_0^*$ For the linear part, denoting $\boldsymbol{\beta}_j = (\beta_{j,k})_{k\in\Lambda_j}, j = j_0^*, \dots j_0^* + J$, one has

$$\sum_{j=j_0^*}^{j_0^*+J^*} \sum_{k\in\Lambda_j} |\beta_{j,k}| \sqrt{\textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_{j,k,t}|^2]}$$
$$\leqslant \sum_{j=j_0^*}^{j_0^*+J^*} \|\boldsymbol{\beta}_j\|_p |\Lambda_j|^{(\frac{1}{2}-\frac{1}{p})_+} \sqrt{\textstyle\sum_{t=1}^T \sum_{k\in\Lambda_j(e^*)} \mathbb{E}_{t-1}[|\hat{g}_{j,k,t}|^2]}$$
$$\leqslant \sum_{j=j_0^*}^{j_0^*+J^*} \|\boldsymbol{\beta}_j\|_p |\Lambda_j|^{(\frac{1}{2}-\frac{1}{p})_+} \sqrt{\textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_t|^2] \| \sum_{k\in\Lambda_j} |\psi_{j,k}(\cdot)|^2 \|_\infty}$$
$$\leqslant \sum_{j=j_0^*}^{j_0^*+J^*} \|\boldsymbol{\beta}_j\|_p |\Lambda_j|^{(\frac{1}{2}-\frac{1}{p})_+} 2^{\frac{jd}{2}} \sqrt{M_\psi \|\psi\|_\infty \textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_t|^2]}$$
$$\leqslant \lambda^{\frac{1}{2}} \|f\|_{\mathcal{B}} \sqrt{M_\psi \|\psi\|_\infty \textstyle\sum_{t=1}^T \mathbb{E}_{t-1}[|\hat{g}_t|^2]} \sum_{j=j_0^*}^{j_0^*+J^*} 2^{-j\beta}$$

    where first inequality is by Hölder's inequality (with $p \geqslant 1$) and convexity on $\sum_k$, and last inequality applies Hölder's inequality on $\sum_j$ (with $q \geqslant 1$), $|\Lambda_j| \leqslant \lambda 2^{jd}$ and $M_\psi = \|\sum_k |\psi(\cdot - k)|\|_\infty < \infty$ and $\|f\|_{\mathcal{B}}$ is the Besov norm of the function f defined in (6) and we defined $\beta = s - \frac{d}{p} - (\frac{d}{2} - \frac{d}{p})_+$. Further, we have for every $t \geqslant 1$

$$\mathbb{E}_{t-1}[\hat{g}_t^2] = 4(\mathbb{E}_{t-1}[(\hat{f}_t(X_t) - f(X_t))^2] + \sigma^2) = 4(L_t(\hat{f}_t(X_t)) - L_t(f(X_t)) + \sigma^2)$$

and plugging this in the previous bound gives

$$C\|f\|_{\mathcal{B}}\Big(\sqrt{\textstyle\sum_{t=1}^{T}(L_t(\hat{f}_t(X_t))-L_t(f(X_t)))}+\sigma\sqrt{T}\Big)\textstyle\sum_{j=j_0^*}^{j_0^*+J}2^{-j\beta}.$$

where C can be deduced from the last lines.

– Detail coefficients in the nonlinear part: we consider only a sparse representation of the function in levels $[j_0^*+J^*,J]$, and we keep only an oracle set of coefficients $\Lambda^*=\{(j,k),J^*\leqslant j\leqslant J,\text{greatest coefficients }|\beta_{j,k}|\}$. Similarly to [24, Theorem 1], retaining only $|\Lambda^*|=2^{(j_0^*+J^*)d}$ gives a regret of the same order of that of the linear part above. Indeed, with $\Lambda^*=\cup_{j=J^*+j_0^*+1}^{J}\Lambda_j^*$ where $|\Lambda_j^*|\leqslant|\Lambda_j|$ is now the oracle sparse set made of positions k at level j. One has on the levels $j=J^*+j_0^*+1,\dots,J$,

$$C\sum_{J^*+j_0^*<j\leqslant J}\|\boldsymbol{\beta}_j\|_p\,|\Lambda_j^*|^{(\frac12-\frac1p)_+}\sqrt{\textstyle\sum_{k\in\Lambda_j^*}\sum_{t=1}^{T}\mathbb{E}_{t-1}[|\hat{g}_{j,k,t}|^2]}$$
$$\leqslant C\Big(\sqrt{\textstyle\sum_{t=1}^{T}(L_t(\hat{f}_t(X_t))-L_t(f(X_t)))}+\sigma\sqrt{T}\Big)\textstyle\sum_{J^*+j_0^*<j\leqslant J}\|\boldsymbol{\beta}_j\|_p|\Lambda_j^*|^{(\frac12-\frac1p)_+}2^{\frac{dj}{2}}$$

Then, using Hölder's inequality with $q\geqslant1$ one has

$$\sum_{J^*+j_0^*<j\leqslant J}2^{\frac{dj}{2}}\|\boldsymbol{\beta}_j\|_p|\Lambda_j^*|^{(\frac12-\frac1p)_+}=\sum_{J^*+j_0^*<j\leqslant J}2^{j(s+\frac d2-\frac dp)}\|\boldsymbol{\beta}_j\|_p\cdot2^{-j(s-\frac dp)}|\Lambda_j^*|^{(\frac12-\frac1p)_+}$$
$$\leqslant\|f\|_{\mathcal{B}}\sum_{J^*+j_0^*<j\leqslant J}2^{-j(s-\frac dp)}|\Lambda_j^*|^{(\frac12-\frac1p)_+}.$$

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

$$\sum_{J^*+j_0^*<j\leqslant J}2^{-j(s-\frac dp)}|\Lambda_j^*|^{(\frac12-\frac1p)_+}\leqslant|\Lambda^*|^{(\frac12-\frac1p)_+}\sum_{J^*+j_0^*<j\leqslant J}2^{-j(s-\frac dp)}$$
$$\leqslant|\Lambda^*|^{(\frac12-\frac1p)_+}\frac{2^{-(J^*+j_0^*)(s-\frac dp)}}{2^{s-\frac dp}-1},$$

by Hölder's inequality in the case $p\geqslant2$ and since $s>\frac dp$ and taking $|\Lambda^*|=2^{d(J^*+j_0^*)}$ entails a nonlinear regret of

$$C\|f\|_{\mathcal{B}}\Big(\sqrt{\textstyle\sum_{t=1}^{T}(L_t(\hat{f}_t(X_t))-L_t(f(X_t)))}+\sigma\sqrt{T}\Big)2^{-(J^*+j_0^*)(s-\frac dp-(\frac d2-\frac dp)_+)}$$

which is of the same order of the regret for the linear part.

- Deviation term: Let φ denote either the scaling basis function ϕ or the wavelet ψ. With $G_{j,k,t}=G\,|\varphi_{j,k}(X_t)|=G2^{\frac{jd}{2}}|\varphi(2^jX_t-k)|$ and $C_{j,k}=B2^{-\frac{jd}{2}}$ at any time $t\geqslant1$, one has

$$\sup_{1\leqslant t\leqslant T}\sum_{j,k}|C_{j,k}G_{j,k,t}|\leqslant GB\sup_x\sum_{j,k}|\varphi(x-k)|\leqslant\sum_{j=j_0^*}^{j_0^*+J}\sup_x\sum_{k\in\Lambda_j}|\varphi(x-k)|\leqslant GBM_\varphi(J+1),$$

and setting $N_\varphi=\sup_x\sum_k\mathbb{1}_{|\varphi(x-k)|>0}<\infty$ the number of wavelets involved at each scale (of order S^d)

$$\sup_{1\leqslant t\leqslant T}\sum_{n\in\Lambda_{e^*}}C_n\mathbb{1}_{G_{n,t}>0}\leqslant B\sum_{j=j_0(e^*)}^{j_0(e^*)+J}2^{-\frac{jd}{2}}\sup_x\sum_{k\in\Lambda_{e^*,j}}\mathbb{1}_{|\varphi(x-k)|>0}\leqslant\frac{BN_\varphi}{1-2^{-d/2}}.$$

Finally the deviation term is of order

$$D:=O\bigg(\Big(GBM_\varphi(J+1)+\frac{BN_\varphi}{1-2^{-d/2}}\max(\nu,\mu)\log(T)\Big)^2\log(\delta^{-1})\bigg),$$

where J will be of order $\log T$.

- Conclusion on R_2: One has with probability $1-2\delta$, summing all the previous bounds:

$$R_2 \leqslant C\|f\|_{\mathcal{B}}\Big(\sqrt{\textstyle\sum_{t=1}^{T}(L_t(\hat{f}_t(X_t)) - L_t(f(X_t)))} + \sigma\sqrt{T}\Big)\textstyle\sum_{j=j_0^*}^{j_0^*+J} 2^{-j\beta} + D$$

Step 2: Bound on the approximation term. Recall Λ^* denotes the set of indices corresponding to the $|\Lambda^*|$ largest wavelet coefficients (in absolute value) among all $v_{j,k} = \beta_{j,k}2^{j(s+d/2-d/p)}$ with $j \in (M^*, J]$ and $k \in \Lambda_j$, where $M^* = j_0^* + J^*$. We set $s' := s + d/2 - d/p > 0$ (since $s > d/p$).

Control of the ℓ^p norm. By the Besov norm characterization and Hölder's inequality over levels:

$$\sum_{M^*<j\leqslant J}\sum_{k\in\Lambda_j}|v_{j,k}|^p \;\leqslant\; (J-M^*)^{(1-p/q)_+}\Bigg(\sum_{M^*<j\leqslant J}\Big(\sum_{k\in\Lambda_j}|v_{j,k}|^p\Big)^{q/p}\Bigg)^{p/q}$$
$$\leqslant\; \Big[(J-M^*)^{(1/p-1/q)_+}\|f\|_{\mathcal{B}}\Big]^p \;=:\; C_f^p.$$

Since Λ^* collects the largest $|\Lambda^*|$ values of $|v_{j,k}|$:

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

so that for all $(j,k)\notin\Lambda^*$ with $j\in(M^*,J]$:

$$|\Lambda^*|\,|v_{j,k}|^p \leqslant C_f^p \quad\Longrightarrow\quad |\beta_{j,k}| \leqslant C_f\,|\Lambda^*|^{-1/p}\,2^{-js'}. \tag{32}$$

Parseval decomposition. Since $\hat{f}^*$ retains only the indices in Λ^* among levels $(M^*, J]$ and discards everything above J, we have:

$$\|\hat{f}^*_{e^*} - f\|_2^2 \;=\; \underbrace{\sum_{\substack{M^*<j\leqslant J\\(j,k)\notin\Lambda^*}}|\beta_{j,k}|^2}_{\text{(I) unselected fine-scale}} \;+\; \underbrace{\sum_{j>J}\sum_{k}|\beta_{j,k}|^2}_{\text{(II) tail beyond } J}. \tag{33}$$

Term (I): nonlinear approximation error. Applying (32) and summing over k at each level j (there are $|\Lambda_j| = 2^{jd}$ wavelet atoms at level j):

$$\text{(I)} \leqslant C_f^2\,|\Lambda^*|^{-2/p}\sum_{j=M^*+1}^{J}2^{jd}\cdot 2^{-2js'} = C_f^2\,|\Lambda^*|^{-2/p}\sum_{j=M^*+1}^{J}2^{-2j(s-d/p)}$$
$$\leqslant \frac{C_f^2\,|\Lambda^*|^{-2/p}\,2^{-2M^*(s-d/p)}}{2^{2(s-d/p)}-1} = \frac{C_f^2\,2^{-2M^*s}}{2^{2(s-d/p)}-1}, \tag{34}$$

where in the last step we used $|\Lambda^*| = 2^{M^*d}$, so that $|\Lambda^*|^{-2/p}\cdot 2^{-2M^*(s-d/p)} = 2^{-2M^*s}$.

Term (II): tail beyond J. Here we do not have a sparsity argument, but we can exploit the full ℓ^p structure of the Besov space. For each level $j > J$, the Besov norm gives $\|v_j\|_{\ell^p(\Lambda_j)} \lesssim \|f\|_{\mathcal{B}}$. We bound $\|v_j\|^2_{\ell^2(\Lambda_j)}$ according to the position of p relative to 2:

- *Case $p \leqslant 2$*: since $p \leqslant 2$, one has $\|v\|_{\ell^2} \leqslant \|v\|_{\ell^p}$ for any sequence, hence:

$$\sum_k|\beta_{j,k}|^2 = 2^{-2js'}\|v_j\|^2_{\ell^2} \;\leqslant\; 2^{-2js'}\|v_j\|^2_{\ell^p} \;\leqslant\; C^2\,\|f\|^2_{\mathcal{B}}\cdot 2^{-2js'},$$

  where the second inequality uses the Besov wavelet characterization $\|v_j\|_{\ell^p} \leqslant C\|f\|_{\mathcal{B}}$, with $C > 0$ depending only on s, p, d and the choice of wavelet basis.
- *Case $p \geqslant 2$*: Hölder's inequality applied to the finite sum over $|\Lambda_j| = 2^{jd}$ indices, with exponents $p/2$ and $(p/2)' = p/(p-2)$, gives

$$\|v_j\|^2_{\ell^2} = \sum_k|v_{j,k}|^2 \leqslant \|v_j\|^2_{\ell^p}\cdot 2^{jd(1-2/p)} \leqslant C^2\,\|f\|^2_{\mathcal{B}}\cdot 2^{jd(1-2/p)},$$

where the second inequality uses $\|v_j\|_{\ell^p} \leqslant C\|f\|_{\mathcal{B}}$, with $C > 0$ depending only on s, p, d and the wavelet basis. Therefore:

$$\sum_k |\beta_{j,k}|^2 = 2^{-2js'}\|v_j\|_{\ell^2}^2 \;\leqslant\; C^2\,\|f\|_{\mathcal{B}}^2 \cdot 2^{-2js'} \cdot 2^{jd(1-2/p)} = C^2\,\|f\|_{\mathcal{B}}^2 \cdot 2^{-2js},$$

where we used $-2s' + d(1-2/p) = -2(s + d/2 - d/p) + d(1-2/p) = -2s$.

In both cases, writing $\gamma := s - d(1/p - 1/2)_+$, we have the unified bound:

$$\sum_k |\beta_{j,k}|^2 \;\leqslant\; C^2\|f\|_{\mathcal{B}}^2 \cdot 2^{-2j\gamma}.$$

Note that $\gamma = s'$ for $p \leqslant 2$ and $\gamma = s$ for $p \geqslant 2$, so $\gamma > 0$ in both cases (since $s > d/p \geqslant d/p - d/2$). Summing over $j > J$:

$$\text{(II)} \;\leqslant\; C^2\|f\|_{\mathcal{B}}^2 \sum_{j>J} 2^{-2j\gamma} \;=\; \frac{C^2\|f\|_{\mathcal{B}}^2\, 2^{-2J\gamma}}{2^{2\gamma}-1} \;=\; \frac{C^2\|f\|_{\mathcal{B}}^2\, 2^{-2J(s-d(1/p-1/2)_+)}}{2^{2\gamma}-1}. \tag{35}$$

Final bound on R_3. Combining (34) and (35) into (33), with $C' = (J - M^*)^{2(1/p-1/q)_+}/(2^{2(s-d/p)} - 1)$ gives

$$R_3 = T\|\hat{f}^*_{e^*} - f\|_2^2 \;\leqslant\; \underbrace{C'^2\,\|f\|_{\mathcal{B}}^2\, T\, 2^{-2M^*s}}_{\text{nonlinear sparse bias}} + \underbrace{C^2\,\|f\|_{\mathcal{B}}^2\, T\, 2^{-2J(s-d(1/p-1/2)_+)}}_{\text{tail bias}}. \tag{36}$$

As a remark, the tail term is identified as the *optimal recovery rate (OR)*: fixing $J = \lfloor \frac{1}{d}\log_2 T \rfloor$ (so $2^J = O(T^{1/d})$), it satisfies

$$\|f\|_{\mathcal{B}}^2\, T\, 2^{-2J\gamma} \;=\; \|f\|_{\mathcal{B}}^2\, T^{1-2\gamma/d} \;=\; \|f\|_{\mathcal{B}}^2\, T^{1-2(s/d-(1/p-1/2)_+)},$$

which matches the known OR rate for B^s_{pq} [12, Eq. (1.13)] for L_2-norm. In the next Step 3, only $M^* \leqslant J$ remains to be optimized to balance the remaining terms.

Step 3: Optimization of the resolutions M^* and j_0^*. Combining the previous steps, with probability at least $1 - 6\delta$:

$$R_T(f) \;\leqslant\; R_1 + R_2 + R_3 + D,$$

where $R_1 = O(\log|\mathcal{E}| + \log^2 T \log\delta^{-1})$ and $D = O(\log^2 T \log\delta^{-1})$. We fix $J = \lfloor \frac{1}{d}\log_2 T \rfloor$, so that $2^{Jd} = O(T)$. By (36), the bound reads for some generic constant $C > 0$ (depending only on s, p, d and the wavelet basis) whose value may change from line to line:

$$R_T(f) \;\leqslant\; A\Big(\sqrt{R_T(f)} + \sigma\sqrt{T}\Big) + C\,\|f\|_{\mathcal{B}}^2\, T\, 2^{-2M^*s} + \underbrace{C\,\|f\|_{\mathcal{B}}^2\, T^{1-2(s/d-(1/p-1/2)_+)}}_{\text{OR term, fixed}} + (R_1 + D), \tag{37}$$

where $M^* = j_0^* + J^* \leqslant J$ and A depends on the smoothness regime:

$$A \;=\; C\|f\|_{\mathcal{B}} \cdot \begin{cases} 2^{-M^*(s-d/2)} & \beta = s - d/2 < 0, \\ J^* & \beta = s - d/2 = 0, \\ 2^{-j_0^*(s-d/2)} & \beta = s - d/2 > 0,\ p \geqslant 2, \\ 2^{-j_0^*(s-d/p)} & \beta = s - d/p > 0,\ p < 2. \end{cases}$$

The OR term is independent of M^* and is carried as a permanent additive contribution. We optimize $M^* \leqslant J$ solely to balance the remaining terms.

Resolving the quadratic inequality. Setting $U = \sqrt{R_T(f)}$ and $R'_3 = C\|f\|_{\mathcal{B}}^2 T 2^{-2M^*s}$, equation (37) gives $U^2 \leqslant AU + A\sigma\sqrt{T} + R'_3 + \dots$. Applying Young's inequality $AU \leqslant \frac{1}{2}U^2 + \frac{1}{2}A^2$ yields:

$$R_T(f) \;\leqslant\; A^2 + 2A\sigma\sqrt{T} + 2R'_3 \;+\; C\,\|f\|_{\mathcal{B}}^2\, T^{1-2(s/d-(1/p-1/2)_+)} \;+\; 2(R_1 + D). \tag{38}$$

In each regime below, we balance the variance term $A^2 + 2A\sigma\sqrt{T}$ (controlled by M^* and j_0^*) against the nonlinear bias $2R'_3$.

Case $\beta < 0$. Here $A = C\|f\|_{\mathcal{B}} 2^{-M^*(s-d/2)}$ increases with M^* since $s - d/2 < 0$. Denoting by $\rho := \|f\|_{\mathcal{B}}/\sigma$ the signal-to-noise ratio, balancing $\|f\|_{\mathcal{B}}\sigma\sqrt{T}\, 2^{-(s-d/2)M^*} = \|f\|_{\mathcal{B}}^2 T\, 2^{-2sM^*}$ gives $2^{(s+d/2)M^*} = \|f\|_{\mathcal{B}}\sigma^{-1}\sqrt{T}$, i.e.:

$$M^* \;=\; \min\Big(\Big\lceil \tfrac{1}{2s+d}\log_2(\rho^2 T)\Big\rceil,\; J\Big)\,,$$

valid when $\rho^2 T > 1$. The cap $M^* = J$ applies when $\rho > T^{s/d}$, in which case the nonlinear bias $\|f\|_{\mathcal{B}}^2 T\, 2^{-2Js} = \|f\|_{\mathcal{B}}^2 T^{1-2s/d}$ is absorbed into the OR term. Otherwise $M^* \leqslant J$ since $\frac{1}{2s+d} < \frac{1}{d}$. Substituting gives the stochastic rate $O(\|f\|_{\mathcal{B}}^{\frac{2d}{2s+d}} \sigma^{\frac{4s}{2s+d}} T^{\frac{d}{2s+d}})$, with $A^2 = o(T^{d/(2s+d)})$ negligible.

Case $\beta = 0$. Here $A = C\|f\|_{\mathcal{B}} J^*$; we set $J^* = \lceil \frac{1}{d}\log_2 T\rceil \leqslant J$, incurring an extra $\log T$ factor in the stochastic rate.

Case $\beta > 0$. Here $A = C\|f\|_{\mathcal{B}}\, 2^{-j_0^*(s-d/2)}$ decreases with j_0^*. Setting $J^* = 0$ (i.e. $j_0^* = M^*$) reduces to the same balance as Case $\beta < 0$, giving the same M^* and the same stochastic rate. Since $s > d/2$, A now decreases with M^* so the balance is a minimum rather than a saddle, and the condition $M^* \leqslant J$ holds since $\frac{1}{2s+d} < \frac{1}{d}$ for $s > 0$. The case $p < 2$ follows with $d/2$ replaced by d/p.

Weak signal. If $\rho^2 T \leqslant 1$, the trivial predictor $\hat{f}_t = 0$ achieves $R_T(f) \leqslant \|f\|_{\mathcal{B}}^2 T \leqslant \sigma^2$, which is dominated by the OR term. No wavelet resolution is needed.

Conclusion. In all regimes, dividing by T:

$$\frac{R_T(f)}{T} \;\leqslant\; O\Bigg(\underbrace{\|f\|_{\mathcal{B}}^{\frac{2d}{2s+d}}\, \sigma^{\frac{4s}{2s+d}}\, T^{-\frac{2s}{2s+d}}}_{\text{stochastic minimax term}} \;+\; \underbrace{\|f\|_{\mathcal{B}}^2\, T^{-2(s/d-(1/p-1/2)_+)}}_{\text{OR term}} \;+\; \frac{(B^2+\sigma^2)\log^2 T\log\delta^{-1}}{T}\Bigg). \tag{39}$$

This matches the noise-level-aware minimax rate of [12] (Theorem 3, Eq. (1.18)) with $q = 2$, up to logarithmic factors and while being fully adaptive to the function parameters and noise level. In particular:

- as $\sigma \to 0$, the stochastic minimax term vanishes and (39) recovers the optimal recovery rate $T^{-2(s/d-(1/p-1/2)_+)}$;
- for fixed $\sigma > 0$, the stochastic term dominates at large T and (39) recovers the classical minimax rate $T^{-2s/(2s+d)}$, with exact dependence on both σ and $\|f\|_{\mathcal{B}}$.

Transition regime. The two leading terms in (39) are balanced at the critical noise level σ^* defined by:

$$\|f\|_{\mathcal{B}}^{\frac{2d}{2s+d}}\, (\sigma^*)^{\frac{4s}{2s+d}}\, T^{-\frac{2s}{2s+d}} \;=\; \|f\|_{\mathcal{B}}^2\, T^{-2(s/d-(1/p-1/2)_+)}.$$

Solving explicitly yields:

$$\sigma^* \;=\; \|f\|_{\mathcal{B}} \cdot T^{-\frac{s}{d}+\frac{2s+d}{2s}(1/p-1/2)_+}, \qquad \rho^* := \frac{\|f\|_{\mathcal{B}}}{\sigma^*} \;=\; T^{\frac{s}{d}-\frac{2s+d}{2s}(1/p-1/2)_+}.$$

This matches the transition noise level of [12] (their Eq. (1.21) with $q = 2$). For $\sigma \geqslant \sigma^*$, the first term dominates and (39) recovers the minimax rate (first term) $\tilde{O}(\|f\|_{\mathcal{B}}^{\frac{2d}{2s+d}} \sigma^{\frac{4s}{2s+d}} T^{-\frac{2s}{2s+d}})$; for $\sigma \leqslant \sigma^*$, the deterministic term dominates and (39) recovers the optimal recovery rate (second term) $O(\|f\|_{\mathcal{B}}^2 T^{-2(s/d-(1/p-1/2)_+)})$, independent of σ. □

E Technical lemmas

We recall several standard concentration inequalities and technical lemmas that are used in the proofs.

Lemma 1 (Exponential inequality)**.** *Let Z be a non-negative and bounded random variable such that $0 \leqslant Z \leqslant D$ for some $D > 0$. Then, for every $0 \leqslant \eta < 1/D$ one has*

$$\mathbb{E}\left[\exp\left(\eta Z - \frac{\eta \mathbb{E}[Z]}{1-\eta D}\right)\right] \leqslant 1.$$

Proof. Let $\eta > 0$. Expanding the exponential, one has by linearity of the expectation

$$\mathbb{E}[\exp(\eta Z)] = \mathbb{E}\left[\sum_{k \geqslant 0} \frac{\eta^k Z^k}{k!}\right] = 1 + \sum_{k \geqslant 1} \frac{\eta^k}{k!}\mathbb{E}[Z^k] \underset{0 \leqslant Z \leqslant D}{\leqslant} 1 + \eta \mathbb{E}[Z] \sum_{k \geqslant 0} \frac{(\eta D)^k}{k!} \underset{\eta < 1/D}{\leqslant} 1 + \frac{\eta \mathbb{E}[Z]}{1-\eta D}.$$

Then, using $1 + x \leqslant \exp(x)$ and dividing by $\exp(\eta \mathbb{E}[Z]/(1-\eta D))$ we get the desired result. □

Lemma 2. *Let (Z_t) be a process adapted to the filtration $(\mathcal{F}_t)$ such that $\mathbb{E}_{t-1}[\exp(Z_t)] \leqslant 1$. Then, one has for any $\delta \in (0,1)$*

$$\mathbb{P}\big(\exists T \geqslant 1 : \textstyle\sum_{t=1}^T Z_t > \log(\delta^{-1})\big) \leqslant \delta.$$

Corollary 1. *Let $0 \leqslant Z_t \leqslant D$. Then $\tilde{Z}_t = \frac{Z_t}{2D}$ satisfies $\mathbb{E}_{t-1}[\exp(\tilde{Z}_t - 2\mathbb{E}_{t-1}[\tilde{Z}_t])] \leqslant 1$ and with probability $1-\delta$*

$$\sum_{t=1}^T Z_t \leqslant \sum_{t=1}^T 2\mathbb{E}_{t-1} Z_t + 2D\log(\delta^{-1}).$$

Lemma 3 (Second order concentration)**.** *Let $\delta \in (0,1)$. For any random variables $(Z_t) \subset \mathbb{R}$ adapted to $(\mathcal{F}_t)$ and any $\gamma > 0$, one has with probability $1-\delta$*

$$\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] \leqslant \sum_{t=1}^T Z_t + \frac{\gamma}{2}\big(\mathbb{E}_{t-1}[Z_t^2] + Z_t^2\big) + \frac{2}{\gamma}\log(\delta^{-1}).$$

Moreover, if (Z_t) are bounded as $|Z_t| \leqslant D, D > 0$ one has with probability $1-2\delta$

$$\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] \leqslant \sum_{t=1}^T Z_t + \frac{3\gamma}{2}\mathbb{E}_{t-1}[Z_t^2] + \left(2D^2 + \frac{2}{\gamma}\right)\log(\delta^{-1}).$$

Proof.

- First inequality: see Bercu and Touati [1] or Wintenberger [38, Prop. 5].
- Second inequality: Since $0 \leqslant Z_t^2 \leqslant D$, one has for every $t \geqslant 1$ by Lemma 1 on Z_t^2,

  $$\mathbb{E}_{t-1}[\exp(Z_t^2/(2D^2) - 2\mathbb{E}_{t-1}[Z_t^2/(2D^2)])] = \mathbb{E}_{t-1}[\exp(Z_t^2/(2D^2) - \mathbb{E}_{t-1}[Z_t^2/D^2])] \leqslant 1.$$

  Then, by Lemma 2 on $(Z_t^2/(2D^2) - \mathbb{E}[Z_t^2/D^2])$, one has

  $$\sum_{t=1}^T Z_t^2 \leqslant \sum_{t=1}^T 2\mathbb{E}[Z_t^2] + 2D^2\log(\delta^{-1}),$$

  which gives the desired bound with a union bound.

□

Lemma 4 (Sup-norm of functions and predictors.)**.** *Let $B > 0$, $f \in \mathcal{B}^s_{pq}(B)$, $s > \frac{d}{p}$ and diameters $C_{j,k} = B2^{-j\frac{d}{2}}$. Then, both f and predictors $\hat{f} = \sum_{j,k} c_{j,k}\varphi_{j,k}, |c_{j,k}| \leqslant C_{j,k}$ are bounded in sup-norm as follows:*

$$\|f\|_\infty \leqslant B_\infty \qquad \text{and} \qquad \|\hat{f}\|_\infty \leqslant \hat{B}_\infty\,,$$

with $B_\infty = \frac{BM_\varphi}{1-2^{-(s-\frac{d}{p})}}$, $\hat{B}_\infty \leqslant (J+1)BM_\varphi$ where $M_\varphi = \|\sum_k |\phi(\cdot - k)|\|_\infty \vee \|\sum_k |\psi(\cdot - k)|\|_\infty < \infty$.

Proof. Since $f \in \mathcal{B}^s_{pq}(B)$ one has for every j, $\mathbf{c}_j = (c_{j,k})_{k \in \Lambda_j}$

$$\|\mathbf{c}_j\|_\infty \leqslant \|\mathbf{c}_j\|_p \leqslant B2^{-j(s-\frac{d}{p}+\frac{d}{2})}.$$

This gives

$$\|f\|_\infty \leqslant \sum_{j=j_0}^{j_0+J} \|\mathbf{c}_j\|_\infty 2^{-j(s-\frac{d}{p})} \Big\| \sum_k |\varphi(\cdot - k)| \Big\|_\infty \leqslant BM_\varphi \sum_{j=j_0}^{j_0+J} 2^{-j(s-\frac{d}{p})} \leqslant \frac{BM_\varphi}{1 - 2^{-(s-\frac{d}{p})}}$$

and

$$\|\hat{f}\|_\infty \leqslant \sum_{j=j_0}^{j_0+J} |c_{j,k}| 2^{j\frac{d}{2}} \Big\| \sum_k |\varphi(\cdot - k)| \Big\|_\infty \leqslant BM_\varphi \sum_{j=j_0}^{j_0+J} 1 \leqslant BM_\varphi (J+1).$$

□

F 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*
$$V_0 = \left\{f = \textstyle\sum_{k\in\mathbb{Z}^d} c_k \phi(\cdot - k), (c_k) : \sum_{k\in\mathbb{Z}^d} c_k^2 < \infty\right\}, \dots, V_j = \{h = f(2^j \cdot) : f \in V_0\}, \dots,$$
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

$$K_j f(x) := \sum_{k\in\mathbb{Z}^d} \langle f, \phi_{j,k}\rangle \phi_{j,k}(x) = \int_{\mathbb{R}^d} K_j(x,y) f(y)\, dy\,, \tag{40}$$

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

$$\psi^\epsilon(x) = \psi^{\epsilon_1}(x_1)\cdots\psi^{\epsilon_d}(x_d), \quad \psi^\epsilon_{j,n} = 2^{jd/2}\psi^\epsilon(2^j x - n), \quad j \geqslant 0, \quad n \in \mathbb{Z}^d,$$

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 $j \geqslant j_0$,

$$K_j f = K_{j_0} f + \sum_{l=j_0}^{j-1} (K_{l+1} f - K_l f)\,, \tag{41}$$

where each increment in the sum can be written as

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

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

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

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 [10, Chapter 7] and the biorthogonal, symmetric, and highly regular wavelet bases of Cohen et al. [7]. 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 (5).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the abstract and introduction clearly state the main contributions: high-probability NLA minimax risk bounds for nonparametric regression, comparator-adaptive regret bounds, and adaptive gradient clipping. All claims are supported by formal theorems in Sections 2–3 with proofs in the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: all assumptions are stated explicitly before each theorem. Complete proofs are provided in Appendix and the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: This paper is purely theoretical and contains no experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification: This paper is purely theoretical and contains no experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This paper is purely theoretical and contains no experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper is purely theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This is foundational theoretical research in statistical learning theory with no direct negative societal impact. The work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational theoretical research on minimax estimation rates. It has no foreseeable direct negative societal impact, and broader impact discussion is not applicable.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release any model, dataset, or artifact that could be misused.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use any external code, dataset, or model asset. All cited works are properly referenced.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release any new dataset, code, or model asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as part of the core methodology. Any use was limited to writing assistance, which does not require declaration per NeurIPS policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.