Minimax Adaptive Online Nonparametric Regression over Besov spaces

Paul Liautaud[1], Pierre Gaillard[2], Olivier Wintenberger[1]

[1]Sorbonne Université, CNRS, LPSM, Paris, France ; [2]Université Grenoble Alpes, INRIA, CNRS, Grenoble INP, LJK, Grenoble, France

Setting: online nonparametric regression

Context & data: data arrives **sequentially** as a stream $x_1, \dots, x_t$ and we want to predict each response as follows:

Learning scenario

For each round $t = 1, \dots, T$, the learner or algorithm

- observes an input $x_t \in \mathcal{X} \subset \mathbb{R}^d$
- makes a prediction $\hat{f}_t(x_t) \in \mathbb{R}$ — *Choose $\hat{f}_t$ before observing ℓ_t*
- suffers a loss $\ell_t(\hat{f}_t(x_t))$ and observes gradients — *No assumptions on how ℓ_t is generated*
- updates his rule prediction $\hat{f}_t \to \hat{f}_{t+1}$ — *Based on observed gradients*

Goal: given some large (nonparametric) function set $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ we want to minimize the regret against any competitor $f \in \mathcal{F}$

$$\mathrm{Reg}_T(f) = \underbrace{\sum_{t=1}^T \ell_t(\hat{f}_t(x_t))}_{\text{our performance}} - \underbrace{\sum_{t=1}^T \ell_t(f(x_t))}_{\text{reference performance}} \underbrace{= o(T)}_{\text{goal}}$$

No stochastic assumption on data (x_t, ℓ_t): $(\hat{f}_t)$ have to perform well on **abitrary** and possibly **adversarial** data.

Assumptions:

- (ℓ_t) are general G-Lipschitz convex losses, $G > 0$ known;
- $\mathcal{X} \subset \mathbb{R}^d$ bounded compact subset;
- $\mathcal{F} \subset B^s_{pq}(\mathcal{X})$ the set of Besov continuous functions over $\mathcal{X}$, with **unknown** $p, q \in [1, \infty]$, $s > \frac{d}{p}$:

Multiscale representation: starting scale $j_0 \geq 0$, $\{\phi_{j_0,k}, \psi_{j,k}\}$ wavelet L^2-basis

$$\forall f \in L^p(\mathcal{X}), \quad f = \sum_{k \in \bar{\Lambda}_{j_0}} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0} \beta_{j,k}\psi_{j,k}, \quad \text{with } |\Lambda_j| = O(2^{dj}).$$

Wavelet characterization: $f \in B^s_{pq}(\mathcal{X})$ if

$$\|f\|_{B^s_{pq}} = \|\alpha_{j_0,\cdot}\|_{\ell^p} + \left(\sum_{j \geq j_0} 2^{j(s+\frac{d}{2}-\frac{d}{p})q} \|\beta_{j,\cdot}\|^q_{\ell^p}\right)^{\frac{1}{q}} < \infty.$$

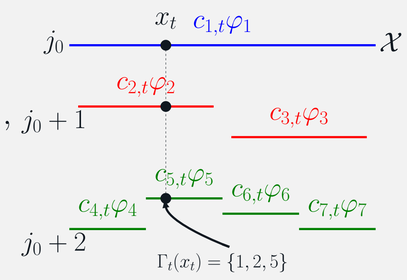
Contribution 1: Online Wavelet Regression

Wavelet regressor: $\{\varphi_{j,k}\}$ an S-regular wavelet basis with $S > s$,

$$\hat{f}_t(x_t) = \sum_{j=j_0}^J \sum_{k \in \Lambda_j} c_{j,k,t}\varphi_{j,k}(x_t) = \sum_{(j,k) \in \Gamma_t} c_{j,k,t}\varphi_{j,k}(x_t),$$

where the active wavelet coefficients at time t

$$\Gamma_t(x_t) := \{(j,k) : \varphi_{j,k}(x_t) \neq 0\} \quad \text{with} \quad |\Gamma_t| = O(S^d).$$



Algorithm:

Algorithm 1: Online wavelet regression at time t, $\hat{f}_t \to \hat{f}_{t+1}$

Input : $\hat{f}_t$ and bound $G > 0$, $0 < \varepsilon < s - \frac{d}{p}$

for $(j,k) \in \Gamma_t$ **do**

Predict $\hat{f}_t(x_t) = \sum_{(j,k) \in \Gamma_t} c_{j,k}\varphi_{j,k}(x_t)$;

Find $c_{j,k,t+1} \in \mathbb{R}$ to minimize $c_{j,k} \mapsto \ell_t(\hat{f}_t(x_t))$ using gradient

$g_{j,k,t} = \left[\partial_{c_{j,k}} \ell_t(\hat{f}_t(x_t))\right]_{c_{j,k}=c_{j,k,t}} = \ell_t'(\hat{f}_t(x_t))\varphi_{j,k}(x_t)$, G provided to subroutine [4, 1].

Output: $\hat{f}_{t+1}$

Analysis: Linear–Nonlinear Decomposition and Coefficient Decay

We design an oracle (truncated) wavelet expansion of f:

$$\hat{f}_J(x) = \underbrace{\sum_{j \leq J^*} \sum_{k \in \Lambda_j} c_{j,k}\,\varphi_{j,k}(x)}_{\textbf{linear}\text{, fully kept: } O(2^{J^* d}) \text{ coeff}} + \underbrace{\sum_{J^* < j \leq J} \sum_{k \in \Lambda_j} c_{j,k}\,\varphi_{j,k}(x)}_{\textbf{nonlinear}\text{, sparsified: } O(2^{J^* d}) \text{ biggest coeff}} \quad \text{with} \quad \begin{cases} J^* = \frac{1}{d}\log_2 T, \\ J = \frac{S}{d\varepsilon}\log_2 T. \end{cases}$$

Besov regularity $f \in B^s_{p,q}$ implies an L^∞ **control** of the error and a **geometric decay** of the coeff.:

$$\|\hat{f}_J - f\|_\infty \lesssim 2^{-J^* s} \quad \text{and} \quad |c_{j,k}| \lesssim 2^{-j\,(s+\frac{d}{2}-\frac{d}{p})}$$

Parameter-free [1, 4] subroutine has regret scaling with the norm of the $(c_{j,k})$

$$\sum_{j,k} |c_{j,k}| \sqrt{\sum_{t=1}^T |g_{j,k,t}|^2} \lesssim G\|f\|_{B^s_{pq}} \sqrt{T} \sum_j 2^{-j(s-\frac{d}{p}-(\frac{d}{2}-\frac{d}{p})_+)}$$

Main contributions

❶ An online learning algorithm that leverages a **wavelet** structure and achieves optimal regret for **convex losses** over **Besov-smooth** functions B^s_{pq} while adapting to s, p, q:

$$\forall f \in B^s_{pq}(\mathcal{X}), \quad \mathrm{Reg}_T(f) \lesssim G\|f\|_{B^s_{pq}} \begin{cases} \sqrt{T}, & \text{if } d \leq 2s \text{ or } p < 2, \\ T^{1-\frac{s}{d}}, & \text{if } d > 2s. \end{cases}$$

❷ An algorithm that handles both **convex** and **exp-concave losses** and achieves **optimal** (and **local**) **regret** against any $f \in B^s_{pq}(\mathcal{X})$ (with (ℓ_t) exp-concave)

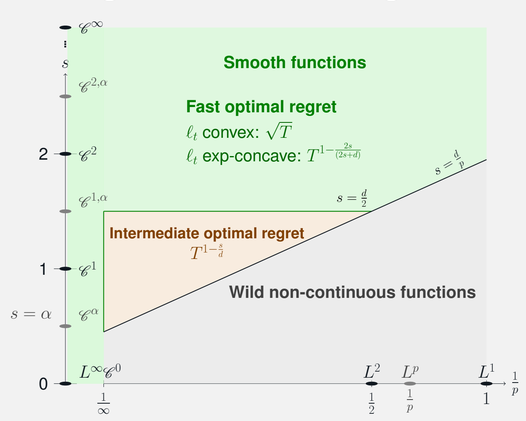
$$\mathrm{Reg}_T(f) \lesssim G \begin{cases} B^{1-\frac{2d}{2s+d}} \|f\|^{\frac{2d}{2s+d}}_{B^s_{pq}} T^{1-\frac{2s}{2s+d}}, & \text{if } s \geq \frac{d}{2}, \\ \|f\|_{B^s_{pq}} T^{1-\frac{s}{d}}, & \text{if } s < \frac{d}{2}. \end{cases}$$

Our algorithms are **constructive** and **polynomial-time**.

Comparison with previous constructive methods:

Paper	Setting	Input Parameters	Regret Rate	Complexity
[7]	$f \in B^s_{pq},\ p, q \geq 1$	$s, p, B \geq \|f\|_\infty$	$T^{1-\frac{s}{s+d}}$	$\exp(T) + Td$
[6]	$f \in B^s_{pq},\ p \geq 2,\ q \in [\frac{p}{p-1}, p]$ $f \in \mathscr{C}^s,\ p = \infty,\ s \geq \frac{d}{2}$	$s, p, B \geq \|f\|_\infty$	$T^{1-\frac{1}{p}}$ $T^{1-\frac{s}{d}+\varepsilon}$	Not feasible
[2]	$f \in W^s_p,\ p \geq 2,\ s \geq \frac{d}{2}$ $f \in W^s_p,\ p > 2,\ s < \frac{d}{2}$ $f \in \mathscr{C}^s,\ p = \infty,\ d = 1,\ s > \frac{1}{2}$	$s, p, B \geq \|f\|_\infty$	$T^{1-\frac{2s}{2s+d}}$ $T^{1-\frac{s}{d}}$ $T^{1-\frac{2s}{2s+1}}$	$\exp(dT)$ $\exp(dT)$ $\mathrm{poly}(T)$
[3]	$f \in \mathscr{C}^s,\ p = \infty,\ s \in (\frac{1}{2}, 1],\ d = 1$	$B \geq \|f\|_\infty$	$T^{1-\frac{2s}{2s+1}}$	$\mathrm{poly}(T)$
[8]	$f \in W^s_p,\ p \geq 2,\ s \geq \frac{d}{2}$ $f \in W^s_p,\ p > 2,\ s < \frac{d}{2}$	s, p	$T^{1-\frac{2s}{2s+d}+\varepsilon}$ $T^{1-\frac{s\,p-\frac{d}{2}}{d\,p-2}+\varepsilon}$	$\mathrm{poly}(T)d$
This work Alg. 1	$f \in B^s_{pq},\ p, q \geq 1,\ s \geq \frac{d}{2}$ or $p \leq 2$ $f \in B^s_{pq},\ p > 2,\ q \geq 1,\ s < \frac{d}{2}$	$S \geq s, \varepsilon < s - \frac{d}{p}$	$\sqrt{T}$ $T^{1-\frac{s}{d}}$	$\mathrm{poly}(T)S^d$
This work Alg. 2	$f \in B^s_{pq},\ p, q \geq 1,\ s \geq \frac{d}{2}$ or $p \leq 2$ $f \in B^s_{pq},\ p > 2,\ q \geq 1,\ s < \frac{d}{2}$	$S \geq s, \varepsilon < s - \frac{d}{p}, B \geq \|f\|_\infty$	$T^{1-\frac{2s}{2s+d}}$ $T^{1-\frac{s}{d}}$	$\mathrm{poly}(T)S^d$

Diagram of minimax regret



References

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

[2] Pierre Gaillard and Sébastien Gerchinovitz. "A chaining algorithm for online nonparametric regression". In: *Conference on Learning Theory*. PMLR. 2015, pp. 764–796.

[3] Paul Liautaud, Pierre Gaillard, and Olivier Wintenberger. "Minimax-optimal and Locally-adaptive Online Nonparametric Regression". In: *Algorithmic Learning Theory*. PMLR. 2025, pp. 702–735.

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

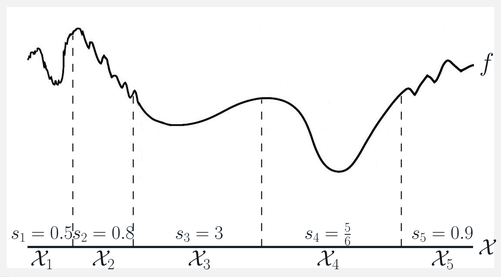
[5] Alexander Rakhlin and Karthik Sridharan. "Online non-parametric regression". In: *Conference on Learning Theory*. PMLR. 2014, pp. 1232–1264.

[6] Vladimir Vovk. "Competing with wild prediction rules". In: *Machine Learning* 69 (2007), pp. 193–212.

[7] Vladimir Vovk. "Metric entropy in competitive on-line prediction". In: *arXiv preprint cs/0609045* (2006).

[8] Oleksandr Zadorozhnyi et al. "Online nonparametric regression with Sobolev kernels". In: *arXiv preprint arXiv:2102.03594* (2021).

Contribution 2: Fast and Local Wavelet Regression



- $(\mathcal{X}_n)$ dyadic subsets of $\mathcal{X}$, forming partitions at different scales
- $(\hat{f}_n)$ wavelet regressors with different starting scale adapted to $(\mathcal{X}_n)$
- $s_n := \sup\{\alpha : f_{|\mathcal{X}_n} \in B^\alpha_{pq}(\mathcal{X}_n)\} \geq s$ local smoothness.

Results: for any $f \in B^s_{pq}(\mathcal{X})$ and any dyadic partition $\mathcal{P} = (\mathcal{X}_n)$ of $\mathcal{X}$,

$$\mathrm{Reg}_T(f) \lesssim \begin{cases} \sum_{n \in \mathcal{P}} B\sqrt{T_n} + 2^{-s_n \mathrm{l}(n)}\|f\|_{s_n}\sqrt{T_n}, & \text{if } (\ell_t) \text{ convex}, \\ B|\mathcal{P}| + \sum_{n \in \mathcal{P}} 2^{-s_n \mathrm{l}(n)}\|f\|_{s_n}\sqrt{T_n}, & \text{if } (\ell_t) \text{ exp-concave}. \end{cases}$$

where $\|f\|_{s_n}$ local Besov norm over $\mathcal{X}_n$ and $T_n = |\{t : x_t \in \mathcal{X}_n\}|$.

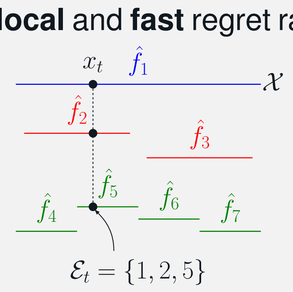
Curvature adaptivity: fast local rates when (ℓ_t) exp-concave:

$$\mathrm{Reg}_T(f) \lesssim \sum_n \left(B^{1-\frac{2d}{2s_n+d}} \left(2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}\right)^{\frac{2d}{2s_n+d}} |T_n|^{\frac{d}{2s_n+d}} \mathbb{1}_{s_n \geq \frac{d}{2}} + 2^{-\mathrm{l}(n)s_n}\|f\|_{s_n}|T_n|^{1-\frac{s_n}{d}} \mathbb{1}_{s_n < \frac{d}{2}} + B \right)$$

Local-adaptivity: our bound decomposes as a sum of **local** and **fast** regret rates.

Algorithm: given $B \geq \|f\|_\infty$, we predict at each time using experts awake at time t

$$\mathcal{E}_t = \{n : x_t \in \mathcal{X}_n\} \text{ and } \hat{f}_t(x_t) = \sum_{n \in \mathcal{E}_t} w_{n,t}[\hat{f}_{n,t}(x_t)]_B.$$



Sleeping expert reduction: no global expert cost, we pay local complexity only!

Algorithm 2: Online Local Wavelet Regression at time t

Input : Bounds $G > 0, 0 < \varepsilon < s - \frac{d}{p}, B \geq \|f\|_\infty$, wavelet regressors $(\hat{f}_{n,t})$, weights $(w_{n,t})$

Receive x_t, define active experts $\mathcal{E}_t = \{n : x_t \in \mathcal{X}_n\}$;

Reduce weights $w_{n,t} \leftarrow w_{n,t}/\sum_{n \in \mathcal{E}_t} w_{n,t}$ if $n \in \mathcal{E}_t$ and $w_{n,t} = 0$ else;

Predict with $\mathcal{E}_t$, $\hat{f}_t(x_t) = \sum_{n \in \mathcal{E}_t} w_{n,k,t}[\hat{f}_{n,k,t}(x_t)]_B$;

Reveal gradient $(\partial_{w_{n,t}} \ell_t(\sum_n w_{n,t}[\hat{f}_{n,t}(x_t)]_B))$ and update weights $(w_{n,t})$ with expert aggregation procedure;

for $e \in \mathcal{E}_t$ **do**

Reveal gradient $g_{n,t} = \ell_t'(\hat{f}_{n,t}(x_t))$;

Update wavelet regressor $\hat{f}_{n,t}$ using Algorithm 1 with $g_{n,t}$.

Output: Weights $(w_{n,t+1})$ and predictors $(\hat{f}_{n,t+1})$

Efficient and computationally feasible, per round complexity: $O(S^d\sqrt{T}\log(T))$.