Minimax Adaptive Online Nonparametric Regression over Besov Spaces

NeurIPS@Paris 2025, Paris

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

November 25th, 2025

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

Setting: online regression with individual sequences (1/2)

Online prediction scenario: at each round $t \in \mathbb{N}^*$, the forecaster

1. observes an input $x_t \in \mathcal{X}$;
2. chooses a prediction $\widehat{f}_t(x_t) \in \mathbb{R}$; — Choose $\widehat{f}_t$ before observing ℓ_t
3. incurs a loss $\ell_t(\widehat{f}_t(x_t))$ — No assumptions on how ℓ_t is generated
4. updates his prediction function $\widehat{f}_t \to \widehat{f}_{t+1}$ — Based on observed gradients

$\cdots \to x_t =$  $\to$  $\to$ 'zebra' $\to$  $\to x_{t+1} =$  $\to \cdots$

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

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

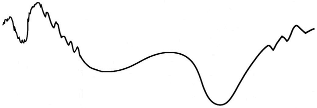
Setting: online regression with individual sequences (2/2)

Individual sequences: no stochastic assumption on data (x_t, ℓ_t)!
$\widehat{f}_1, \ldots, \widehat{f}_T$ have to perform **well** with all **arbitrary** and possibly **adversarial** sequences.

Assumptions:

- $\ell_1, \ldots, \ell_T$ are G-Lipschitz convex losses, with $G > 0$;
- $\mathcal{X} \subset \mathbb{R}^d$ bounded compact subset with dimension $d \geq 1$;
- $\mathcal{F} = B^s_{pq}$ the set of *Besov functions*, with $s > \frac{d}{p}, 1 \leqslant p, q \leqslant \infty$.

$\rightarrow$ Multiscale representation: $\{\varphi_{j_0,k}, \psi_{j,k}\}$ an orthonormal basis of $L^2(\mathcal{X})$, for any $f \in L^p(\mathcal{X})$



$$f = \sum_{k \in \bar{\Lambda}_{j_0}} \alpha_{j_0,k} \varphi_{j_0,k} + \sum_{j \geq j_0} \sum_{k \in \Lambda_j} \beta_{j,k} \psi_{j,k}. \qquad (1)$$

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

Contribution 1: online wavelet regression

We present an online learning method leveraging a **wavelet** structure, at $t \geq 1$:

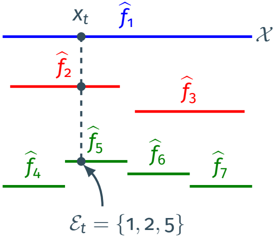
1. Receive input x_t, predict $\widehat{f}_t(x_t) = \sum_{j,k} c_{j,k,t}\, \psi_{j,k}(x_t)$, and obtain gradients $(g_{j,k,t})_{j,k}$
2. Update each wavelet coefficient $c_{j,k,t}$ using associated gradient $g_{j,k,t}$ and a parameter-free subroutine (Mhammedi et al., 2020).

Our algorithm achieves a regret against any function $f \in B^s_{pq}$

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

- **Adaptivity** to s, p, q and $\|f\|_{B^s_{pq}}$!
- **Automatic** thresholding: no need of an explicit threshold!
- Our rates are **minimax** over B^s_{pq} for general **convex** losses (Rakhlin et al., 2015).
- Our algorithm is **computationally tractable**.

Contribution 2: (fast) learning in heterogeneous environment



- $(\mathcal{X}_n)$ dyadic subsets of $\mathcal{X}$, forming partitions at different scales
- $(\widehat{f}_n)$ wavelet regressors with different starting scale adapted to $(\mathcal{X}_n)$

We present a **constructive** algorithm based on **expert aggregation**, at each time $t \geq 1$, given $B \geq \|f\|_\infty$:

1. Receive input x_t, define active expert set $\mathcal{E}_t = \{n : x_t \in \mathcal{X}_n\}$
2. **Predict** $\widehat{f}_t(x_t) = \sum_{n \in \mathcal{E}_t} w_{n,t} [\widehat{f}_{n,t}(x_t)]_B$, and obtain gradients with respect to $(w_{n,t})$ and wavelet coefficients
3. **Update** active weights $(w_{n,t})_{n \in \mathcal{E}_t}$ using a second-order aggregation algorithm (Gaillard, Stoltz, and Van Erven; Wintenberger, 2014; 2017) and active wavelet predictors $(\widehat{f}_{n,t})_{n \in \mathcal{E}_t}$ using associated gradients.

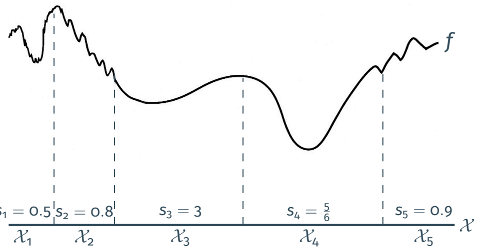
Contribution 2: (fast) learning in heterogeneous environment

We achieve **local** and **optimal** regret against any competitor $f \in B^s_{pq}$, $B \geq \|f\|_\infty$ with exp-concave losses (i.e., $\widehat{y} \mapsto \exp(-\eta \ell_t(\widehat{y}))$ concave)

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

where $\sum_n$ is over any partition $(\mathcal{X}_n)$, $T_n = \{1 \leqslant t \leqslant T : x_t \in \mathcal{X}_n\}$, s_n regularity over $\mathcal{X}_n$.

- ★ **Adaptivity** to local regularities ($\|f\|_{s_n}$) with respect to the partition;
- ★ **Adaptivity** to the loss curvature;
- 🏆 **Minimax** and low regret in highly-regular regions ($s_n \gg 1$);
- 💻 Our algorithm is **computationally tractable**.



Conclusion

- **First constructive algorithm** achieving **optimal and locally adaptive regret** against Besov-smooth competitors;
- A **single algorithm** that adapts simultaneously to the competitor's **local regularity** and to the **curvature** of the sequential losses;
- **Local adaptivity yields global gains**: our method is locally finer, hence globally superior to global approaches.

Thank you and see you at the poster!

Questions?

References

- Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, Learning, and Games*. Cambridge University Press.
- Cesa-Bianchi, Nicolò et al. (2017). "Algorithmic chaining and the role of partial feedback in online nonparametric learning". In: *Conference on Learning Theory*. PMLR, pp. 465–481.
- Cutkosky, Ashok and Francesco Orabona (2018). "Black-box reductions for parameter-free online learning in banach spaces". In: *Conference On Learning Theory*. PMLR, pp. 1493–1529.
- Gaillard, Pierre and Sebastien Gerchinovitz (2015). "A Chaining Algorithm for Online Nonparametric Regression". In: *COLT*.
- Gaillard, Pierre, Gilles Stoltz, and Tim Van Erven (2014). "A second-order bound with excess losses". In: *Conference on Learning Theory*. PMLR, pp. 176–196.
- Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2, pp. 169–192.
- Kuzborskij, Ilja and Nicolo Cesa-Bianchi (2020). "Locally-adaptive nonparametric online learning". In: *Advances in Neural Information Processing Systems* 33, pp. 1679–1689.

Liautaud, Paul, Pierre Gaillard, and Olivier Wintenberger (2024). "Minimax-optimal and Locally-adaptive Online Nonparametric Regression". In: *arXiv preprint arXiv:2410.03363*.

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

Orabona, Francesco and Dávid Pál (2016). "Coin betting and parameter-free online learning". In: *Advances in Neural Information Processing Systems* 29.

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

— (2015). "Online nonparametric regression with general loss functions". In: *arXiv preprint arXiv:1501.06598*.

Wintenberger, Olivier (2017). "Optimal learning with Bernstein online aggregation". In: *Machine Learning* 106.1, pp. 119–141.

Zinkevich, Martin (2003). "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936.