A Chaining Algorithm for Online Nonparametric Regression

Pierre Gaillard pierre-p.gaillard@edf.fr
Sébastien Gerchinovitz sebastien.gerchinovitz@math.univ-toulouse.fr

Problem: online nonparametric regression

For each round $t = 1, \dots, T$,
- Environment chooses $x_t \in \mathcal{X}$
- Forecaster predicts $\widehat{y}_t \in \mathbb{R}$
- Forecaster observes $y_t \in \mathbb{R}$ and suffers square loss $(\widehat{y}_t - y_t)^2$

Goal: minimize the regret

$$\mathrm{Reg}_T(\mathcal{F}) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \big(y_t - \widehat{y}_t\big)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \big(y_t - f(x_t)\big)^2$$

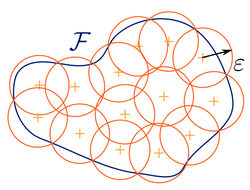
over nonparametric function classes $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Main contributions

We design an algorithm that achieves a Dudley-type regret bound (in the same vein as in Rakhlin and Sridharan [3] but in a constructive fashion):

$$\mathrm{Reg}_T(\mathcal{F}) \le \square B^2 \big(1 + \log \mathcal{N}_\infty(\mathcal{F}, \gamma)\big) + \square B \sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}\, \mathrm{d}\varepsilon \,,$$

where $\gamma \in \big(\frac{B}{T}, B\big)$ is a parameter of the algorithm,
B is an upper bound on $\max_{1 \le t \le T} |y_t|$,
$\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)$ denotes the metric entropy of $\mathcal{F}$.



Function class	Metric entropy	Regret bound
	ε^{-p} $\quad p \in (0,2)$	$T^{p/(p+2)}$
Lipschitz on $[0,1]$	ε^{-1}	$T^{1/3}$
β-Hölder on $[0,1]$	$\varepsilon^{-1/\beta}$ $\quad \beta > 1/2$	$T^{1/(2\beta+1)}$
Sparse lin. reg.	$\log \binom{d}{s} + s \log\big(1 + 1/(\varepsilon\sqrt{s})\big)$	$s \log(1 + d\,T/s)$

Second contribution: efficient version for Hölder classes (costs a log factor).

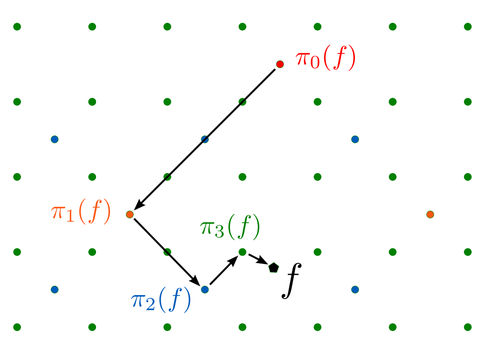
Chaining algorithm: main ideas

Chaining approximation: we approximate any function $f \in \mathcal{F}$ by a sequence of refining approximations $\pi_0(f) \in \mathcal{F}^{(0)}, \pi_1(f) \in \mathcal{F}^{(1)}, \dots$, such that for all $k \ge 0$,

$$\sup_f \|\pi_k(f) - f\|_\infty \le \gamma/2^k$$

and

$$\mathrm{Card}\, \mathcal{F}^{(k)} = \mathcal{N}_\infty(\mathcal{F}, \gamma/2^k)\,,$$



so that:

$$\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \big(y_t - f(x_t)\big)^2 = \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \Bigg(y_t - \pi_0(f)(x_t) - \sum_{k=0}^{\infty} \underbrace{\big[\pi_{k+1}(f) - \pi_k(f)\big](x_t)}_{|\text{small increments}| \le 3\gamma/2^{k+1}} \Bigg)^2 .$$

Our algorithm relies on a two-scale aggregation:

- **high-scale aggregation:** run an Exponentially Weighted Average forecaster to be competitive against every function $\pi_0(f)$ in the coarsest set $\mathcal{F}^{(0)}$;

  Small $\mathrm{Card}\,\mathcal{F}^{(0)}$ ➟ small regret

- **low-scale aggregation:** run many instances of (an extension of) Exponentiated Gradient so as to be competitive against all increments $\pi_{k+1}(f) - \pi_k(f)$.

  Small $\|\pi_{k+1}(f) - \pi_k(f)\|_\infty$ ➟ small regret

  The Multi-variable Exponentiated Gradient algorithm (defined below) achieves this guarantee at all small scales $\gamma/2^k$, $k \ge 0$, simultaneously.

Key subroutine: Multi-variable Exponentiated Gradient

Let Δ_N denote the simplex in $\mathbb{R}^N$. We design an extension of the Exponentiated Gradient algorithm to minimize a sequence of multi-variable loss functions $\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}\big) \mapsto \ell_t\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}\big)$ simultaneously over all the variables $\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}\big) \in \Delta_{N_1} \times \dots \times \Delta_{N_K}$.

input : tuning parameters $\eta^{(1)}, \dots, \eta^{(K)} > 0$.

initialization: set $\widehat{\mathbf{u}}_1^{(k)} \stackrel{\text{def}}{=} \big(\frac{1}{N_k}, \dots, \frac{1}{N_k}\big) \in \Delta_{N_k}$ for all $k = 1, \dots, K$.

for *each round* $t = 2, 3, \dots$ **do**

Compute the weight vectors $\big(\widehat{\mathbf{u}}_t^{(1)}, \dots, \widehat{\mathbf{u}}_t^{(K)}\big) \in \Delta_{N_1} \times \dots \times \Delta_{N_K}$ as follows ($Z_t^{(k)}$ is a normalization factor):

$$\widehat{\mathbf{u}}_{t,i}^{(k)} \stackrel{\text{def}}{=} \frac{\exp\left(-\eta^{(k)} \sum_{s=1}^{t-1} \partial_{\widehat{u}_{s,i}^{(k)}} \ell_s\left(\widehat{\mathbf{u}}_s^{(1)}, \dots, \widehat{\mathbf{u}}_s^{(K)}\right)\right)}{Z_t^{(k)}}\,, \quad i \in \{1, \dots, N_k\}\,.$$

end

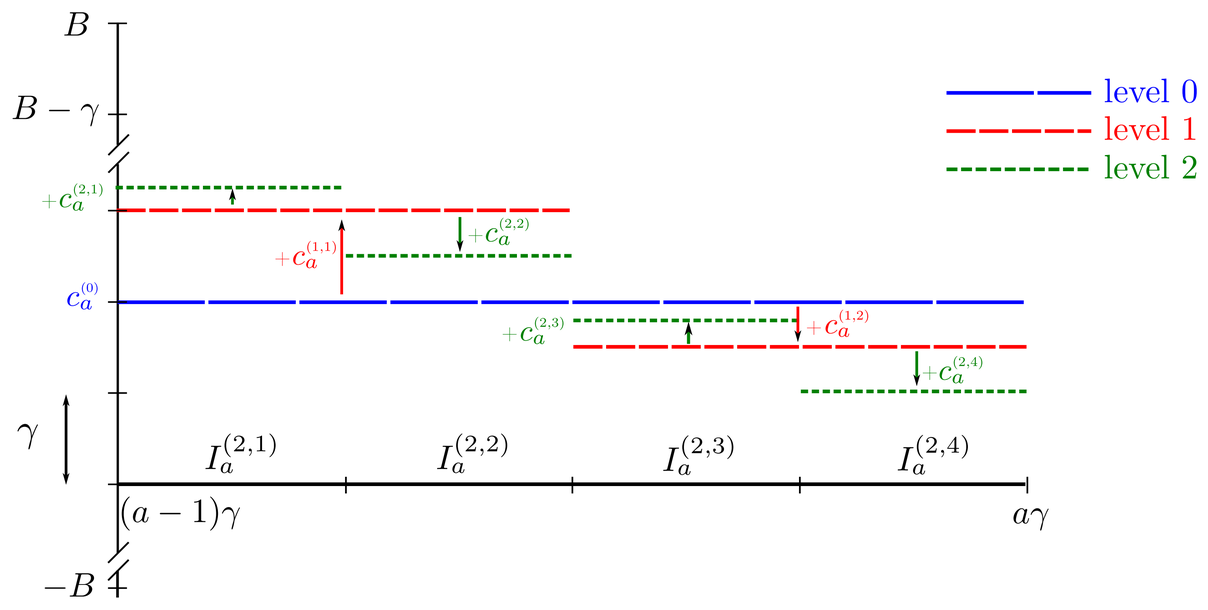
Regret bound: Assume that the ℓ_t are jointly convex and differentiable with bound $\|\nabla_{\mathbf{u}^{(k)}} \ell_t\|_\infty \le G^{(k)}$. Then, the Multi-variable Exponentiated Gradient algorithm tuned with the parameters $\eta^{(k)} = \sqrt{2\log(N_k)/T}\,/G^{(k)}$ satisfies:

$$\sum_{t=1}^{T} \ell_t\left(\widehat{\mathbf{u}}_t^{(1)}, \dots, \widehat{\mathbf{u}}_t^{(K)}\right) - \min_{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}} \sum_{t=1}^{T} \ell_t\left(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}\right) \le \sqrt{2T} \sum_{k=1}^{K} G^{(k)} \sqrt{\log N_k}\,,$$

where the minimum is taken over all $\big(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}\big) \in \Delta_{N_1} \times \dots \times \Delta_{N_K}$.

An efficient version for Hölder classes

The idea is to design **computationally manageable coverings** $\mathcal{F}^{(k)}$, $k \ge 0$:
- approximate any Lipschitz function $f \in [0,1] \to [-B, B]$ with piecewise constant functions (level $k = 0$);
- refine the approximation via a dyadic discretization (levels $k \ge 1$).



At each round t, the point x_t falls into only one subinterval for each level k
➟ No need to update all coefficients ➟ **manageable complexity.**

For Hölder functions: replace piecewise constant functions with piecewise polynomials.

Function class	Time complexity	Space complexity
Lipschitz on $[0,1]$	$T^{4/3} \log T$	$T^{4/3} \log T$
β-Hölder on $[0,1]$	$\mathrm{poly}(T)$	$\mathrm{poly}(T)$

References

[1] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Statist.*, 27:1865–1895, 1999.

[2] N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Mach. Learn.*, 43:247–264, 2001.

[3] A. Rakhlin and K. Sridharan. Online nonparametric regression. In *Proceedings of COLT'14*, volume 35, pages 1232–1264. JMLR W&CP, 2014.