A chaining algorithm for online non parametric regression

Pierre Gaillard

October 27, 2017

INRIA Paris, ENS Paris

This is joint work with Nicolò Cesa-Bianchi, Claudio Gentile and Sebastien Gerchinovitz

Table of Contents

Online prediction of arbitrary sequences

The framework of this talk

Sequential prediction of arbitrary time-series[1]:

- a time-series $y_1, \dots, y_n \in \mathcal{Y} = [-B, B]$ is to be predicted step by step
- covariates $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$ are sequentially available

At each forecasting instance $t = 1, \dots, n$

- the environment reveals $\mathbf{x}_t \in \mathcal{X}$
- the player is ask to form a prediction $\widehat{y}_t$ of y_t based on
  - the past observations $y_1, \dots, y_{t-1}$
  - the current and past covariates $\mathbf{x}_1, \dots, \mathbf{x}_t$
- the environment reveals y_t

Goal: minimize the average loss: $\widehat{L}_n = \frac{1}{n} \sum_{t=1}^{n} (\widehat{y}_t - y_t)^2$.

Difficulty: no stochastic assumption on the time series

- neither on the observations (y_t)
- nor on the covariates $(\mathbf{x}_t)$

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. 2006.

The framework of this talk

Sequential prediction of arbitrary time-series:

- a time-series $y_1, \dots, y_n \in \mathcal{Y} = [-B, B]$ is to be predicted step by step
- covariates $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$ are sequentially available

At each forecasting instance $t = 1, \dots, n$

- the environment reveals $\mathbf{x}_t \in \mathcal{X}$
- solution: produce the prediction as a function of $\mathbf{x}_t$

$$\widehat{y}_t = \widehat{f}_t(\mathbf{x}_t)$$

- the environment reveals y_t

Goal: minimize our average regret against a reference function class $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$

$$\mathrm{Reg}_n(\mathcal{F}) \stackrel{\text{def}}{=} \underbrace{\frac{1}{n}\sum_{t=1}^{n}\big(\widehat{f}_t(\mathbf{x}_t) - y_t\big)^2}_{\text{our performance}} - \underbrace{\inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\big(f(\mathbf{x}_t) - y_t\big)^2}_{\text{reference performance}}$$

The framework of this talk

Sequential prediction of arbitrary time-series:

- a time-series $y_1, \dots, y_n \in \mathcal{Y} = [-B, B]$ is to be predicted step by step
- covariates $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$ are sequentially available

At each forecasting instance $t = 1, \dots, n$

- the environment reveals $\mathbf{x}_t \in \mathcal{X}$
- solution: produce the prediction as a function of $\mathbf{x}_t$

$$\widehat{y}_t = \widehat{f}_t(\mathbf{x}_t)$$

- the environment reveals y_t

Goal: minimize our average regret against a reference function class $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$

$$\mathrm{Reg}_n(\mathcal{F}) \stackrel{\text{def}}{=} \underbrace{\frac{1}{n}\sum_{t=1}^{n}\left(\widehat{f}_t(\mathbf{x}_t) - y_t\right)^2}_{\text{our performance}} - \underbrace{\inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\left(f(\mathbf{x}_t) - y_t\right)^2}_{\text{reference performance}} \underbrace{= o(1)}_{\text{Goal}}$$

Online to batch conversion

Online regret bound:

$$\mathrm{Reg}_n(\mathcal{F}) \stackrel{\text{def}}{=} \underbrace{\frac{1}{n}\sum_{t=1}^{n}\big(\widehat{f}_t(x_t) - y_t\big)^2}_{\text{our performance}} - \underbrace{\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\big(f(x_t) - y_t\big)^2}_{\text{reference performance}} \underbrace{= o(1)}_{\text{Goal}}$$

If the data (x_t, y_t) is i.i.d. we can bound the excess risk of $\bar{f}_n = \frac{1}{n}\sum_{t=1}^{n}\widehat{f}_t$:

$$\begin{aligned}\mathbb{E}\Big[\Big(\bar{f}_n(X) - Y\Big)^2\Big] - \inf_{f\in\mathcal{F}}\mathbb{E}\big[(f(X)-Y)^2\big] &\overset{\text{Convexity}}{\leqslant} \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\big[(\widehat{f}_t(X)-Y)^2\big] - \inf_{f\in\mathcal{F}}\mathbb{E}\big[(f(X)-Y)^2\big] \\ &\leqslant \mathbb{E}[\mathrm{Reg}_n(\mathcal{F})] \quad = o(1)\end{aligned}$$

Finite reference class: prediction with expert advice

A strategy for finite $\mathcal{F}$

Assumption: $\mathcal{F} = \{f_1, \dots, f_K\} \subset \mathcal{Y}^{\mathcal{X}}$ is finite

The exponentially weighted average forecaster (Hedge)[1]

At each forecasting instance t,

- assign to each function $f_k \in \mathcal{F}$ the weight

$$\widehat{p}_{k,t} = \frac{\exp\Big(-\eta \sum_{s=1}^{t-1} \big(f_k(x_s) - y_s\big)^2\Big)}{\sum_{j=1}^{K} \exp\Big(-\eta \sum_{s=1}^{t-1} \big(f_j(x_s) - y_s\big)^2\Big)}$$

- form function $\widehat{f}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} f_k$ and predict $\widehat{y}_t = \widehat{f}_t(x_t)$

Performance: if $\mathcal{Y} = [-B, B]$ and $\eta = 1/(8B^2)$

$$\mathrm{Reg}_n(\mathcal{F}) \overset{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \big(\widehat{f}(x_t) - y_t\big)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \big(f(x_t) - y_t\big)^2 \leqslant \frac{8B^2 \log K}{n}$$

If B is not known in advance, η can be tuned online (doubling trick).

[1] Littlestone and Warmuth (1994) and Vovk (1990)

Proof

1. Upper bound the instantaneous loss

$$\begin{aligned}
\left(y_t - \widehat{f}_t(x_t)\right)^2 &= \left(y_t - \textstyle\sum_{k=1}^K \widehat{p}_{k,t} f_k(x_t)\right)^2 \\
&\overset{\text{for } \eta \leqslant 1/(8B^2)}{\leqslant} -\frac{1}{\eta} \log\left(\sum_{k=1}^K \widehat{p}_{k,t} e^{-\eta\left(y_t - f_k(x_t)\right)^2}\right) \leftarrow \text{exp-concavity} \\
&\overset{\text{by definition of } \widehat{p}_{k,t+1}}{=} -\frac{1}{\eta} \log\left(\frac{\widehat{p}_{k,t}}{\widehat{p}_{k,t+1}} e^{-\eta\left(y_t - f_k(x_t)\right)^2}\right) \\
&= \left(y_t - f_k(x_t)\right)^2 + \frac{1}{\eta} \log \frac{\widehat{p}_{k,t+1}}{\widehat{p}_{k,t}}
\end{aligned}$$

2. Sum over all t, the sum telescopes

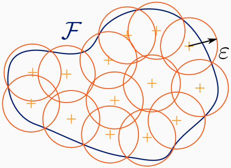
$$\sum_{t=1}^n \left(y_t - \widehat{f}_t(x_t)\right)^2 - \left(y_t - f_k(x_t)\right)^2 \leqslant \frac{1}{\eta} \log \frac{\widehat{p}_{k,n+1}}{\widehat{p}_{k,1}} \leqslant \frac{\log K}{\eta} = 8B^2 \log K$$

Large reference class

1. Approximate $\mathcal{F}$ by a finite set $\mathcal{F}_\varepsilon$ such that

$$\forall f \in \mathcal{F} \quad \exists f_\varepsilon \in \mathcal{F}_\varepsilon \quad \|f - f_\varepsilon\|_\infty \leqslant \varepsilon\,. \qquad (1)$$

   Such set $\mathcal{F}_\varepsilon$ is called an ε-net of $\mathcal{F}$
2. Run Hedge on $\mathcal{F}_\varepsilon$



Definition (metric entropy)

The cardinal of the smallest ε-net $\mathcal{F}_\varepsilon$ that satisfies (1) is denoted $\mathcal{N}_\infty(\mathcal{F},\varepsilon)$. The metric entropy of $\mathcal{F}$ is $\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)$.

Regret bound of order (forgetting constants):

$$\mathrm{Reg}_n(\mathcal{F}) = \mathrm{Reg}_n(\mathcal{F}_\varepsilon) + \left[\inf_{f_\varepsilon \in \mathcal{F}_\varepsilon} \sum_{t=1}^n (y_t - f_\varepsilon(x_t))^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^n (y_t - f(x_t))^2\right]$$

$$\lesssim \underbrace{\frac{\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)}{n}}_{\text{Regret of Hedge on } \mathcal{F}_\varepsilon} + \underbrace{\varepsilon}_{\text{Approximation of } \mathcal{F} \text{ by } \mathcal{F}_\varepsilon}$$

Examples of reference classes: the parametric case

If $\mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \varepsilon^{-p}$ for $p > 0$ as $\varepsilon \to 0$,

$$\mathrm{Reg}_n(\mathcal{F}) \lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)}{n} + \varepsilon$$

$$\approx \frac{\log(\varepsilon^{-p})}{n} + \varepsilon \overset{\varepsilon\approx 1/n}{\approx} \frac{p\log(n)}{n}$$

Example

Assume you have $d \geqslant 1$ black-box forecasters $\varphi_1,\dots,\varphi_d \in \mathcal{X}^{\mathcal{Y}}$

- linear regression in a compact ball

$$\mathcal{F} = \Big\{ \textstyle\sum_{j=1}^d u_j\varphi_j : \quad \text{for } \boldsymbol{u} \in \Theta \underset{\text{comp.}}{\subset} \mathbb{R}^d \Big\} \quad \to \quad \mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \varepsilon^{-d}$$

- sparse linear regression

$$\mathcal{F} = \Big\{ \textstyle\sum_{j=1}^d u_j\varphi_j : \quad \text{for } \boldsymbol{u} \in [0,1]^d \text{ s.t. } \|\boldsymbol{u}\|_1 = 1 \text{ and } \|\boldsymbol{u}\|_0 = s \Big\}$$

Then[2],

$$\log \mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \log \binom{d}{s} + s\log\big(1 + 1/(\varepsilon\sqrt{s})\big) \quad \to \quad \mathrm{Reg}_n(\mathcal{F}) \lesssim \frac{s\log(1+dn/s)}{n}$$

[2] F. Gao, C.-K. Ing, and Y. Yang. "Metric entropy and sparse linear approximation of ℓq-hulls for 0< q≤ 1". In: *Journal of Approximation Theory* (2013).

Examples of reference classes: the parametric case

If $\mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \varepsilon^{-p}$ for $p > 0$ as $\varepsilon \to 0$,

$$\mathrm{Reg}_n(\mathcal{F}) \lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)}{n} + \varepsilon$$

$$\approx \frac{\log(\varepsilon^{-p})}{n} + \varepsilon \overset{\varepsilon \approx 1/n}{\approx} \frac{p\log(n)}{n} \quad \to \text{ optimal}$$

Example

Assume you have $d \geqslant 1$ black-box forecasters $\varphi_1, \dots, \varphi_d \in \mathcal{X}^{\mathcal{Y}}$

- linear regression in a compact ball

$$\mathcal{F} = \left\{ \textstyle\sum_{j=1}^d u_j \varphi_j : \quad \text{for } \boldsymbol{u} \in \Theta \underset{\text{comp.}}{\subset} \mathbb{R}^d \right\} \quad \to \quad \mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \varepsilon^{-d}$$

- sparse linear regression

$$\mathcal{F} = \left\{ \textstyle\sum_{j=1}^d u_j \varphi_j : \quad \text{for } \boldsymbol{u} \in [0,1]^d \text{ s.t. } \|\boldsymbol{u}\|_1 = 1 \text{ and } \|\boldsymbol{u}\|_0 = s \right\}$$

Then[2],

$$\log \mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim \log \binom{d}{s} + s \log\left(1 + 1/(\varepsilon\sqrt{s})\right) \quad \to \quad \mathrm{Reg}_n(\mathcal{F}) \lesssim \frac{s\log(1+dn/s)}{n}$$

[2] F. Gao, C.-K. Ing, and Y. Yang. "Metric entropy and sparse linear approximation of ℓq-hulls for 0< q≤ 1". In: *Journal of Approximation Theory* (2013).

What if $\mathcal{F}$ is non parametric?

Non parametric class: if $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \lesssim \varepsilon^{-p}$ for $p > 0$ as $\varepsilon \to 0$.

$$\mathrm{Reg}_n(\mathcal{F}) \lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n} + \varepsilon$$

$$\lesssim \frac{\varepsilon^{-p}}{n} + \varepsilon \overset{\varepsilon = n^{-1/(p+1)}}{\approx} n^{-\frac{1}{p+1}}$$

Example

- **1-Lipschitz ball** on $[0,1]$

$$\mathcal{F} = \left\{ f \in \mathcal{Y}^{\mathcal{X}} : \quad \forall x, y \in \mathcal{X} \subset [0,1] \quad \|f(x) - f(y)\| \leqslant \|x - y\| \right\}$$

Then $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1} \quad \to \quad \mathrm{Reg}_n(\mathcal{F}) \lesssim n^{-1/2}$

- **Hölder ball** on $\mathcal{X} \subset [0,1]$ with regularity $\beta = q + \alpha > 1/2$

$$\mathcal{F} = \left\{ f \in \mathcal{Y}^{\mathcal{X}} : \forall x, y \in \mathcal{X} \ \left| f^{(q)}(x) - f^{(q)}(y) \right| \leqslant |x - y|^\alpha \text{ and } \forall k \leqslant q, \|f^{(k)}\|_\infty \leqslant B \right\}$$

Then[3] $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1/\beta} \quad \to \quad \mathrm{Reg}_n(\mathcal{F}) \lesssim n^{-\frac{\beta}{\beta+1}}$.

[3] G. Lorentz. “Metric Entropy, Widths, and Superpositions of Functions”. In: *Amer. Math. Monthly* 6 (1962).

What if $\mathcal{F}$ is non parametric?

Non parametric class: if $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \lesssim \varepsilon^{-p}$ for $p > 0$ as $\varepsilon \to 0$.

$$\mathrm{Reg}_n(\mathcal{F}) \lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n} + \varepsilon$$

$$\lesssim \frac{\varepsilon^{-p}}{n} + \varepsilon \overset{\varepsilon = n^{-1/(p+1)}}{\approx} n^{-\frac{1}{p+1}}$$

→ suboptimal:

$n^{-\frac{2}{p+2}}$ if $p < 2$

$n^{-\frac{1}{p}}$ if $p > 2$

Example

- **1-Lipschitz ball** on $[0, 1]$

$$\mathcal{F} = \left\{ f \in \mathcal{Y}^{\mathcal{X}} : \quad \forall x, y \in \mathcal{X} \subset [0, 1] \quad \big\| f(x) - f(y) \big\| \leqslant \|x - y\| \right\}$$

Then $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1} \quad \to \quad \mathrm{Reg}_n(\mathcal{F}) \lesssim n^{-1/2}$ → suboptimal: $n^{-\frac{2}{3}}$

- **Hölder ball** on $\mathcal{X} \subset [0, 1]$ with regularity $\beta = q + \alpha > 1/2$

$$\mathcal{F} = \left\{ f \in \mathcal{Y}^{\mathcal{X}} : \forall x, y \in \mathcal{X} \ \big| f^{(q)}(x) - f^{(q)}(y) \big| \leqslant |x - y|^\alpha \text{ and } \forall k \leqslant q, \|f^{(k)}\|_\infty \leqslant B \right\}$$

Then[3] $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1/\beta} \quad \to \quad \mathrm{Reg}_n(\mathcal{F}) \lesssim n^{-\frac{\beta}{\beta+1}}$ → suboptimal: $n^{-\frac{\beta}{\beta+1/2}}$.

[3] G. Lorentz. “Metric Entropy, Widths, and Superpositions of Functions”. In: *Amer. Math. Monthly* 6 (1962).

Minimax rates

Theorem (Rakhlin and Sridharan, 2014[4])

The minimax rate of the regret if of order

$$\inf_{\gamma \geqslant \varepsilon \geqslant 0} \left\{ \frac{\log \mathcal{N}^{\mathrm{seq}}(\mathcal{F},\gamma)}{n} + \int_{\varepsilon}^{\gamma} \sqrt{\frac{\log \mathcal{N}^{\mathrm{seq}}(\tau,\mathcal{F})}{n}}\, \mathrm{d}\tau + \varepsilon \right\}$$

where $\log \mathcal{N}^{\mathrm{seq}}(\mathcal{F},\varepsilon) \leqslant \log \mathcal{N}_\infty(\mathcal{F},\varepsilon)$ *is the sequential entropy of* $\mathcal{F}$.

$\frac{\log \mathcal{N}_\infty(\mathcal{F},\gamma)}{n}$: regret of Hedge against γ-net → crude approximation

n: approximation error of the ε-net → fine approximation

$\int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F},\tau)}{n}}\, \mathrm{d}\tau$: from large scale γ to small scale ε.

This term is a Dudley entropy integral that appears in

- Chaining to bound the supremum of a stochastic process (Dudley 1967)
- Statistical learning with i.i.d. data to derive risk bounds (e.g., Massart 2007; Rakhlin et al. 2013)
- Online learning with arbitrary sequences (Opper and Haussler 1997; Cesa-Bianchi and Lugosi 1999)

[4] A. Rakhlin and K. Sridharan. "Online Nonparametric Regression". In: *COLT* (2014).

Minimax rates

Theorem (Rakhlin and Sridharan, 2014[4])

The minimax rate of the regret if of order

$$\inf_{\gamma \geqslant \varepsilon \geqslant 0} \left\{ \frac{\log \mathcal{N}_\infty(\mathcal{F}, \gamma)}{n} + \int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\tau, \mathcal{F})}{n}}\, \mathrm{d}\tau + \varepsilon \right\}$$

if $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \log \mathcal{N}^{\mathrm{seq}}(\mathcal{F}, \varepsilon)$.

$\frac{\log \mathcal{N}_\infty(\mathcal{F},\gamma)}{n}$: regret of Hedge against γ-net → crude approximation

n: approximation error of the ε-net → fine approximation

$\int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F},\tau)}{n}}\, \mathrm{d}\tau$: from large scale γ to small scale ε.

This term is a Dudley entropy integral that appears in

- Chaining to bound the supremum of a stochastic process (Dudley 1967)
- Statistical learning with i.i.d. data to derive risk bounds (e.g., Massart 2007; Rakhlin et al. 2013)
- Online learning with arbitrary sequences (Opper and Haussler 1997; Cesa-Bianchi and Lugosi 1999)

[4] A. Rakhlin and K. Sridharan. "Online Nonparametric Regression". In: *COLT* (2014).

Theorem (Rakhlin and Sridharan, 2014[4])

The minimax rate of the regret if of order

$$\inf_{\gamma \geqslant \varepsilon \geqslant 0} \left\{ \frac{\log \mathcal{N}_\infty(\mathcal{F}, \gamma)}{n} + \int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\tau, \mathcal{F})}{n}} \, \mathrm{d}\tau + \varepsilon \right\}$$

if $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \log \mathcal{N}^{\mathrm{seq}}(\mathcal{F}, \varepsilon)$.

Example: let $p \in (0, 2)$ and $\mathcal{F}$ such that

$$\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-p} \qquad \text{as } \varepsilon \to \infty\,.$$

The minimax regret is then of order

$$\frac{\gamma^{-p}}{n} + \int_\varepsilon^\gamma \frac{\tau^{-p/2}}{\sqrt{n}} \, \mathrm{d}\tau + \varepsilon \quad = \quad \frac{\gamma^{-p}}{n} + \frac{\gamma^{1-p/2}}{n} + 0 \quad \approx \quad n^{-\frac{2}{p+2}}$$

for the optimal choices $\varepsilon = 0$ and $\gamma \approx n^{-1/(p+2)}$.

[4] A. Rakhlin and K. Sridharan. "Online Nonparametric Regression". In: *COLT* (2014).

Our contributions

Propose a constructive algorithm which:

- achieves the Dudley-type regret bound

$$\mathrm{Reg}_n \lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \gamma)}{n} + \int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}{n}}\, \mathrm{d}\tau + \varepsilon$$

- efficient version for Hölder class in $[0,1]$ (costs a log factor)

Key-subroutine (Multi-variable EG) to go from scale γ to scale ε.

Function class	Metric entropy	Regret of Hedge	Our Regret
	$\varepsilon^{-p} \quad p \in (0,2)$	$n^{-1/(p+1)}$	$n^{-2/(p+2)}$
Lipschitz on $[0,1]$	ε^{-1}	$n^{-1/2}$	$n^{-2/3}$
β-Hölder on $[0,1]$	$\varepsilon^{-1/\beta} \quad \beta > 1/2$	$n^{-\beta/(\beta+1)}$	$n^{-\beta/(\beta+1/2)}$
Sparse lin. reg.	$\log \binom{d}{s} + s \log\left(1 + 1/(\varepsilon\sqrt{s})\right)$	$\frac{s \log(1+dn/s)}{n}$	$\frac{s \log(1+dn/s)}{n}$

Suboptimality of the previous approach

Why was the previous approach suboptimal? We were treating the functions in the discretization as uncorrelated experts, which is too pessimistic and harmful when $\mathcal{F}$ is large.

To deal with it, we will need the following property for the regret bound:

"if all function in $\mathcal{F}$ are close from one another, the regret should be small"

Hedge achieves this!

Hedge with regret scaling with loss range

Assumption: $\mathcal{F} = \{f_1, \ldots, f_K\} \subset \mathcal{Y}^{\mathcal{X}}$ is finite such that

$$\forall f_i, f_j \in \mathcal{F}, \quad \|f_i - f_j\|_\infty \leqslant \Delta$$

The exponentially weighted average forecaster (Hedge)[5]

At each forecasting instance t,

- assign to each function $f_k \in \mathcal{F}$ the weight

$$\widehat{p}_{k,t} = \frac{\exp\Big(-\eta \sum_{s=1}^{t-1} \big(f_k(x_s) - y_s\big)^2\Big)}{\sum_{j=1}^{K} \exp\Big(-\eta \sum_{s=1}^{t-1} \big(f_j(x_s) - y_s\big)^2\Big)}$$

- form function $\widehat{f}_t = \sum_{k=1}^{K} \widehat{p}_{k,t} f_k$ and predict $\widehat{y}_t = \widehat{f}_t(x_t)$

Performance: if $\mathcal{Y} = [-B, B]$ and well-tuned η

$$\mathrm{Reg}_n(\mathcal{F}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n}\big(\widehat{f}(x_t) - y_t\big)^2 - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\big(f(x_t) - y_t\big)^2 \lesssim \begin{cases} \frac{B^2 \log K}{n} \\ B\Delta\sqrt{\frac{\log K}{n}} \end{cases}$$

[5] Littlestone and Warmuth (1994) and Vovk (1990)

Proof

We replace the exp-concavity property of the square loss with the Hoeffding's lemma.

Lemma (Hoeffding)

If X is a random variable with $|X| \leqslant B$. Then,

$$\forall \eta \in \mathbb{R}, \qquad \mathbb{E}[X] \leqslant -\frac{1}{\eta} \log \left(\mathbb{E}\left[e^{-\eta X}\right] \right) + \frac{\eta B}{4} .$$

1. Upper bound the instantaneous loss

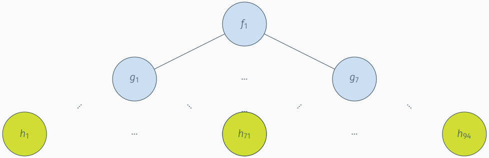
$$\left(y_t - \widehat{f}_t(\mathbf{x}_t)\right)^2 - \left(y_t - f_k(\mathbf{x}_t)\right)^2 \quad \leqslant \quad \log \frac{\widehat{p}_{k,t+1}}{\widehat{p}_{k,t}} + \frac{\eta B \Delta}{4}$$

2. Sum over all t, the sum telescopes

$$\sum_{t=1}^{n} \left(y_t - \widehat{f}_t(\mathbf{x}_t)\right)^2 - \left(y_t - f_k(\mathbf{x}_t)\right)^2 \leqslant \frac{1}{\eta} \log \frac{\widehat{p}_{k,n+1}}{\widehat{p}_{k,1}} + + \frac{\eta B \Delta n}{4} \lesssim B\Delta \sqrt{n \log K}$$

Build a hierarchy of discretizations:

- the level-m discretization approximates $\mathcal{F}$ with precision $\gamma 2^{-m}$;
- each level-m node is connected to its closest level-$(m-1)$ node;



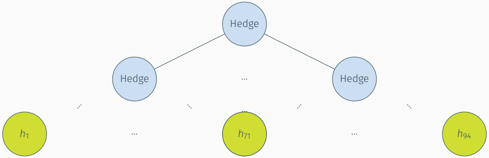
Hierarchical Hedge algorithm:

- each leaf h recommends its own (discretized) function $h(x_t)$;
- each internal node hosts an instance of Hedge using its children as experts; its regret is at most of order $\gamma 2^{-m}\sqrt{\frac{\ln N_{2^{-m}}}{n}}$ at level m since its children's losses are $\gamma 2^{-m}$-close.

Algorithmic chaining

Build a hierarchy of discretizations:

- the level-m discretization approximates $\mathcal{F}$ with precision $\gamma 2^{-m}$;
- each level-m node is connected to its closest level-$(m-1)$ node;



Hierarchical Hedge algorithm:

- each leaf h recommends its own (discretized) function $h(x_t)$;
- each internal node hosts an instance of Hedge using its children as experts; its regret is at most of order $\gamma 2^{-m}\sqrt{\frac{\ln N_{2^{-m}}}{n}}$ at level m since its children's losses are $\gamma 2^{-m}$-close.

Chaining (continued)

Summing the local regret bounds over any path in the tree, we obtain a regret bound of

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{F}) &\lesssim B^2 \frac{\log(N_\gamma)}{n} + B \sum_{m=0}^{M-1} \gamma 2^{-m} \sqrt{\frac{\ln N_{\gamma 2^{-m}}}{n}} + B 2^{-M} \\
&\lesssim B^2 \frac{\log \mathcal{N}_\infty(\mathcal{F}, \gamma))}{n} + B \int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n}} d\varepsilon + B\varepsilon
\end{aligned}$$

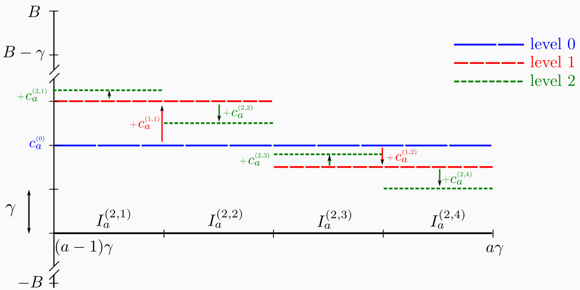
Remarks:

- Same upper bound as the one proven by Rakhlin, Sridharan, and Tewari, 2015 in a nonconstructive manner.
- Matches the lower bound of Hazan and Megiddo, 2007.

Efficient implementation for Lipschitz functions

The idea is to design **computationally manageable coverings** $\mathcal{F}^{(k)}$, $k \geqslant 0$:

- 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 $k \geqslant 1$).



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 $\mathcal{O}(n^{4/3})$.**

For Hölder functions: piecewise constant → piecewise polynomials

Extensions

Extension to general loss functions

Goal: minimize the regret

$$\mathrm{Reg}_n = \frac{1}{n}\sum_{t=1}^{n} \ell_t(\widehat{y}_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell_t(f(x_t))$$

for generic sequences of loss functions (ℓ_t).

If the loss functions ℓ_t are Lipschitz, we can achieve

$$\mathrm{Reg}_n(\mathcal{F}) \lesssim \underbrace{\cancel{\frac{\log \mathcal{N}_\infty(\mathcal{F},\gamma)}{n}}}_{} + \int_\varepsilon^1 \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F},\tau)}{n}}\,\mathrm{d}\tau + \varepsilon$$

Large scale term not possible
(was thanks to strong convexity)

Lipschitz class on $[0,1]^d$	Metric entropy	Hedge Regret	Our Regret
$d=1$	ε^{-1}	$n^{-1/3}$	$n^{-1/2}$
$d=2$	ε^{-2}	$n^{-1/4}$	$n^{-1/2}\log n$
$d\geqslant 3$	ε^{-d}	$n^{-1/(d+2)}$	$n^{-1/d}$

First constructive algorithm to achieve the optimal[6] rates.

The rate $n^{-1/(d+2)}$ was achieved by G. and Baudin, 2014 and Hazan and Megiddo, 2007.

[6] A. Rakhlin and K. Sridharan. "Online Nonparametric Regression with General Loss Functions". In: *arXiv* (2015).

Can we use chaining for other feedbacks?

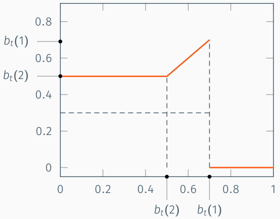
Bandit feedback: the learner only observes its loss $\ell_t(\widehat{y}_t)$ instead of ℓ_t

- Bad news: deriving regret bounds that scale as the effective range of the arms' losses, which was key for full information, is not possible in general for adversarial bandits (Gerchinovitz and Lattimore, 2016).
- Regret bounds : $T^{-1/(d+3)}$ for semi-Lipschitz losses or $T^{-1/(d+2)}$ for convex Lipschitz losses. See also the work of Slivkins (2014).

Can we use chaining for other feedbacks? One-sided feedback

One-sided full-information feedback: the learner obbserves $\ell_t(y)$ for all $y \geqslant \widehat{y}_t$.

Example of application: online auctions in web advertising.



- This stronger feedback, together with Lipschitzness of the losses, enables us to derive a regret bound for a variant of Exp4 that scales as the effective range of the arms' losses.
- Hierarchical algorithm: in the earlier tree, we replace Hedge with Exp4 (bandit algorithm). We obtain a regret of order $T^{-1/(d+1)}$ or even $T^{-1/(d+2/3)}$ with an additional hierarchical penalization trick.

Other related open questions

Get the sequential entropy $\mathcal{N}^{\text{seq}}(\mathcal{F}, \varepsilon)$ instead of the metric entropy $\mathcal{N}_\infty(\mathcal{F}, \varepsilon)$

Efficient version for other function classes

- step-wise Lipschitz functions → application to classification
- generalized additive models → useful to predict electricity consumption

Similar results with other algorithms (Kernel regression)

Thank you !

References

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

G. and P. Baudin. “A consistent deterministic regression tree for non-parametric prediction of time series”. http://arxiv.org/abs/1405.1533. 2014.

G. and S. Gerchinovitz. “A Chaining Algorithm for Online Nonparametric Regression”. In: *Proceedings of COLT’15*. Vol. 40. JMLR: Workshop and Conference Proceedings, 2015, pp. 764–796.

F. Gao, C.-K. Ing, and Y. Yang. “Metric entropy and sparse linear approximation of ℓq-hulls for $0< q\leq 1$”. In: *Journal of Approximation Theory* 166 (2013), pp. 42–55.

S. Gerchinovitz and T. Lattimore. “Refined Lower Bounds for Adversarial Bandits”. In: *Advances in Neural Information Processing Systems 29 (NIPS’16)*. 2016, pp. 1198–1206.

E. Hazan and N. Megiddo. “Online Learning with Prior Knowledge”. In: *Proceedings of the 20th Annual Conference on Learning Theory (COLT’07)*. Ed. by N. H. Bshouty and C. Gentile. Vol. 4539. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 499–513.

N. Littlestone and M. K. Warmuth. "The Weighted Majority Algorithm". In: *Information and Computation* 108.2 (1994), pp. 212–261.

G. Lorentz. "Metric Entropy, Widths, and Superpositions of Functions". In: *Amer. Math. Monthly* 69.6 (1962), pp. 469–485.

A. Rakhlin and K. Sridharan. "Online Nonparametric Regression". In: *COLT* 35 (2014), pp. 1232–1264.

A. Rakhlin and K. Sridharan. "Online Nonparametric Regression with General Loss Functions". In: *arXiv* (2015).

A. Rakhlin, K. Sridharan, and A. Tewari. "Online learning via sequential complexities.". In: *Journal of Machine Learning Research* 16 (2015), pp. 155–186.

A. Slivkins. "Contextual bandits with similarity information.". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 2533–2568.

V. Vovk. "Aggregating Strategies.". In: *Proceedings of the Third Workshop on Computational Learning Theory*. 1990, pp. 371–386.

V. Vovk. "Competitive on-line statistics". In: *International Statistical Review* 69.2 (2001), pp. 213–248.