

Online nonparametric regression with adversarial data

Pierre Gaillard

January 22, 2018

INRIA Paris, ENS Paris

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

1. Online prediction of arbitrary sequences
2. Finite reference class: prediction with expert advice
3. Large reference class
4. Extensions

Online prediction of arbitrary sequences

The framework of this talk

Sequential prediction of arbitrary time-series¹:

- 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 asked to form a prediction \hat{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: $\hat{L}_n = \frac{1}{n} \sum_{t=1}^n (\hat{y}_t - y_t)^2$.

Difficulty: no stochastic assumption on the time series

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

¹N. 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

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

- the environment reveals y_t

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

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

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

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

- the environment reveals y_t

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

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

Online regret bound:

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

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

$$\begin{aligned} \mathbb{E} \left[(\bar{f}_n(X) - Y)^2 \right] - \inf_{f \in \mathcal{F}} \mathbb{E} [(f(X) - Y)^2] &\stackrel{\text{Convexity}}{\leq} \frac{1}{n} \sum_{t=1}^n \mathbb{E} [(\hat{f}_t(X) - Y)^2] - \inf_{f \in \mathcal{F}} \mathbb{E} [(f(X) - Y)^2] \\ &\leq \mathbb{E}[\text{Reg}_n(\mathcal{F})] = o(1) \end{aligned}$$

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

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

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

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

$$\text{Reg}_n(\mathcal{F}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^n (\hat{f}_t(x_t) - y_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n (f(x_t) - y_t)^2 \leq \frac{8B^2 \log K}{n}$$

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

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

1. Upper bound the instantaneous loss

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

2. Sum over all t , the sum telescopes

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

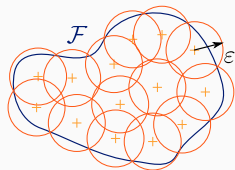
Large reference class

1. Approximate \mathcal{F} by a finite set \mathcal{F}_ε such that

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

Such set \mathcal{F}_ε is called an ε -net of \mathcal{F}

2. Run Hedge on \mathcal{F}_ε



Definition (metric entropy)

The cardinal of the smallest ε -net \mathcal{F}_ε 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):

$$\begin{aligned} \text{Reg}_n(\mathcal{F}) &= \text{Reg}_n(\mathcal{F}_\varepsilon) + \left[\inf_{f_\varepsilon \in \mathcal{F}_\varepsilon} \sum_{t=1}^n (y_t - f_\varepsilon(x_t))^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^n (y_t - f(x_t))^2 \right] \\ &\lesssim \underbrace{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n}}_{\text{Regret of Hedge on } \mathcal{F}_\varepsilon} + \underbrace{\varepsilon}_{\text{Approximation of } \mathcal{F} \text{ by } \mathcal{F}_\varepsilon} \end{aligned}$$

Examples of reference classes: the parametric case

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

$$\begin{aligned} \text{Reg}_n(\mathcal{F}) &\lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n} + \varepsilon \\ &\approx \frac{\log(\varepsilon^{-p})}{n} + \varepsilon \stackrel{\varepsilon \approx 1/n}{\approx} \frac{p \log(n)}{n} \end{aligned}$$

Example

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

- linear regression in a compact ball

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

- sparse linear regression

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

Then²,

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

²F. Gao, C.-K. Ing, and Y. Yang. "Metric entropy and sparse linear approximation of ℓ_q -hulls for $0 < q \leq 1$ ". 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 $\varepsilon \rightarrow 0$,

$$\begin{aligned} \text{Reg}_n(\mathcal{F}) &\lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n} + \varepsilon \\ &\approx \frac{\log(\varepsilon^{-p})}{n} + \varepsilon \stackrel{\varepsilon \approx 1/n}{\approx} \frac{p \log(n)}{n} \rightarrow \text{optimal} \end{aligned}$$

Example

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

- linear regression in a compact ball

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

- sparse linear regression

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

Then²,

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

²F. Gao, C.-K. Ing, and Y. Yang. "Metric entropy and sparse linear approximation of ℓ_q -hulls for $0 < q \leq 1$ ". 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 $\varepsilon \rightarrow 0$.

$$\begin{aligned} \text{Reg}_n(\mathcal{F}) &\lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n} + \varepsilon \\ &\lesssim \frac{\varepsilon^{-p}}{n} + \varepsilon \quad \varepsilon = n^{-1/(p+1)} \approx n^{-\frac{1}{p+1}} \end{aligned}$$

Example

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

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

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

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

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

Then³ $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1/\beta} \rightarrow \text{Reg}_n(\mathcal{F}) \lesssim n^{-\frac{\beta}{\beta+1}}$.

³G. 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 $\varepsilon \rightarrow 0$.

$$\begin{aligned} \text{Reg}_n(\mathcal{F}) &\lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}{n} + \varepsilon \\ &\lesssim \frac{\varepsilon^{-p}}{n} + \varepsilon \quad \varepsilon = n^{-1/(p+1)} \approx n^{-\frac{1}{p+1}} \end{aligned}$$

→ suboptimal:

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

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

Example

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

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

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

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

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

Then³ $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1/\beta} \rightarrow \text{Reg}_n(\mathcal{F}) \lesssim n^{-\frac{\beta}{\beta+1}} \rightarrow \text{suboptimal: } n^{-\frac{\beta}{\beta+1/2}}$.

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

Theorem (Rakhlin and Sridharan, 2014⁴)

The minimax rate of the regret if of order

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

where $\log \mathcal{N}^{\text{seq}}(\mathcal{F}, \epsilon) \leq \log \mathcal{N}_{\infty}(\mathcal{F}, \epsilon)$ 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

$\int_{\epsilon}^{\gamma} \sqrt{\frac{\log \mathcal{N}_{\infty}(\mathcal{F}, \tau)}{n}} d\tau$: 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 Hausler 1997; Cesa-Bianchi and Lugosi 1999)

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

Theorem (Rakhlin and Sridharan, 2014⁴)

The minimax rate of the regret if of order

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

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

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

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

$\int_\varepsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}{n}} d\tau$: 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 Hausler 1997; Cesa-Bianchi and Lugosi 1999)

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

Theorem (Rakhlin and Sridharan, 2014⁴)

The minimax rate of the regret if of order

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

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

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

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

The minimax regret is then of order

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

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

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

Our contributions

Propose a constructive algorithm which:

- achieves the Dudley-type regret bound

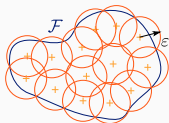
$$\text{Reg}_n \lesssim \frac{\log \mathcal{N}_\infty(\mathcal{F}, \gamma)}{n} + \int_\epsilon^\gamma \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}{n}} d\tau + \epsilon$$

- **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
	$\epsilon^{-p} \quad p \in (0, 2)$	$n^{-1/(p+1)}$	$n^{-2/(p+2)}$
Lipschitz on $[0, 1]$	ϵ^{-1}	$n^{-1/2}$	$n^{-2/3}$
β -Hölder on $[0, 1]$	$\epsilon^{-1/\beta} \quad \beta > 1/2$	$n^{-\beta/(\beta+1)}$	$n^{-\beta/(\beta+1/2)}$
Sparse lin. reg.	$\log \binom{d}{s} + s \log(1 + 1/(\epsilon\sqrt{s}))$	$\frac{s \log(1+dn/s)}{n}$	$\frac{s \log(1+dn/s)}{n}$

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

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

The exponentially weighted average forecaster (Hedge)⁵

At each forecasting instance t ,

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

$$\hat{p}_{k,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} (f_k(\mathbf{x}_s) - y_s)^2\right)}{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} (f_j(\mathbf{x}_s) - y_s)^2\right)}$$

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

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

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

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

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

Lemma (Hoeffding)

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

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

1. Upper bound the instantaneous loss

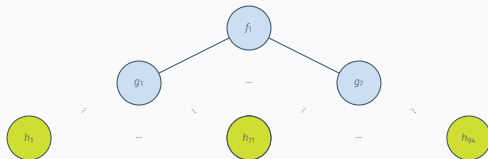
$$(y_t - \widehat{f}_t(\mathbf{x}_t))^2 - (y_t - f_k(\mathbf{x}_t))^2 \leq \log \frac{\widehat{p}_{k,t+1}}{\widehat{p}_{k,t}} + \frac{\eta B \Delta}{4}$$

2. Sum over all t , the sum telescopes

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

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;

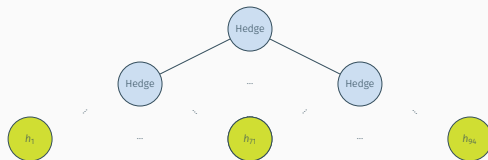


Hierarchical Hedge algorithm:

- 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.

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.

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

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

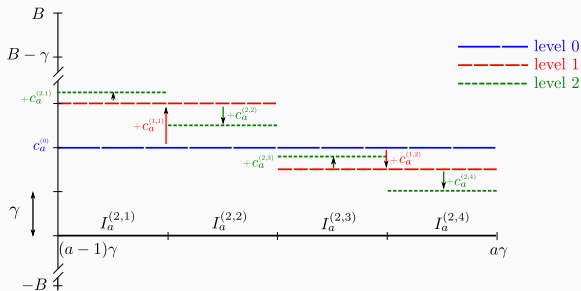
Remarks:

- 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** $\mathcal{F}^{(k)}$, $k \geq 0$:

- approximate any Lipschitz function $f \in [0, 1] \rightarrow [-B, B]$ with **piecewise constant** functions (level $k = 0$);
- refine the approximation via a **dyadic discretization** (levels $k \geq 1$).



At each round t , the point x_t falls into only one subinterval for each level k
 \Rightarrow No need to update all coefficients \Rightarrow **manageable complexity** $\mathcal{O}(n^{4/3})$.

For Hölder functions: piecewise constant \rightarrow **piecewise polynomials**

Extensions

Extension to general loss functions

Goal: minimize the regret

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

for generic sequences of loss functions (ℓ_t) .

If the loss functions ℓ_t are Lipschitz, we can achieve

$$\text{Reg}_n(\mathcal{F}) \lesssim \underbrace{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}{n}}_{\text{Large scale term not possible (was thanks to strong convexity)}} + \int_\varepsilon^1 \sqrt{\frac{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}{n}} d\tau + \varepsilon$$

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}	$n^{-1/4}$	$n^{-1/2} \log n$
$d \geq 3$	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.

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

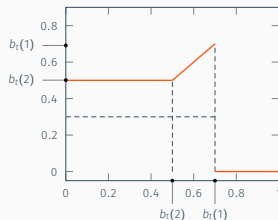
Bandit feedback: the learner only observes its loss $\ell_t(\hat{y}_t)$ instead of ℓ_t

- 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 observes $\ell_t(y)$ for all $y \geq \hat{y}_t$.

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.

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 \rightarrow application to classification
- generalized additive models \rightarrow useful to predict electricity consumption

Similar results with other algorithms (Kernel regression)

- recent results on online learning with kernels (Calandriello, Lazaric, and Valko, 2017)
- Can we get optimal rate together with optimal complexity?

THANK YOU !

References

-  D. Calandriello, A. Lazaric, and M. Valko. “Efficient second-order online kernel learning with adaptive embedding”. In: *Neural Information Processing Systems*. 2017.
-  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 ℓ_q -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.