A CHAINING ALGORITHM FOR ONLINE NON PARAMETRIC REGRESSION

Pierre Gaillard

December 2, 2015

University of Copenhagen

This is joint work with Sebastien Gerchinovitz

TABLE OF CONTENTS

ONLINE PREDICTION OF ARBITRARY SEQUENCES

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 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathcal{X}$ are sequentially available

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

- the environment reveals $\boldsymbol{x}_t \in \mathcal{X}$
- 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 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_t$
- the environment reveals y_t

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

Difficulty: no stochastic assumption on the time series

- neither on the observations (y_t)
- neither on the covariates $(\boldsymbol{x}_t)$

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

Sequential prediction of arbitrary time-series:

- a time-series $y_1, \ldots, y_n \in \mathcal{Y} = [-B, B]$ is to be predicted step by step
- covariates $x_1, \ldots, x_n \in \mathcal{X}$ are sequentially available

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

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

$$\widehat{y}_t = \widehat{f}_t(x_t)$$

- the environment reveals y_t

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

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

Sequential prediction of arbitrary time-series:

- a time-series $y_1, \ldots, y_n \in \mathcal{Y} = [-B, B]$ is to be predicted step by step
- covariates $x_1, \ldots, x_n \in \mathcal{X}$ are sequentially available

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

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

$$\widehat{y}_t = \widehat{f}_t(x_t)$$

- the environment reveals y_t

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

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

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 (EWA)[1]

At each forecasting instance t,

- assign to each function f_k the weight

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

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

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

$$\mathrm{Reg}_n(\mathcal{F}) \stackrel{\text{def}}{=} \sum_{t=1}^{n} \left(y_t - \widehat{f}(\boldsymbol{x}_t)\right)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \left(y_t - f(\boldsymbol{x}_t)\right)^2 \leqslant 8B^2 \log K$$

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

[1] Littlestone and M. K. Warmuth (1994) and Vovk (1990)

1. Upper bound the instantaneous loss

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

2. Sum over all t, the sum telescopes

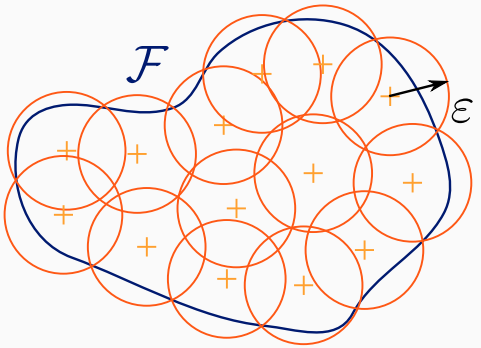
$$
\sum_{t=1}^n \left(y_t - \widehat{f}_t(x_t)\right)^2 - \left(y_t - f_k(x_t)\right)^2 \leqslant \frac{1}{\eta} \log \frac{\widehat{p}_{k,n+1}}{\widehat{p}_{k,1}} \leqslant \frac{\log K}{\eta} = 8B^2 \log K
$$

LARGE REFERENCE CLASS

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

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

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

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

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

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

Example

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

- **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],

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

[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).

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

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

Example

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

- **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],

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

[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?

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

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

Example

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

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

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

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

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

WHAT IF $\mathcal{F}$ IS NON PARAMETRIC?

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

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

→ suboptimal:

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

Example

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

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

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

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

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

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

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

Theorem (Rakhlin and Sridharan 2014[4])

The minimax rate of the regret if of order

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

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

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

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

$\sqrt{n} \int_{\varepsilon}^{\gamma} \sqrt{\log \mathcal{N}_{\infty}(\mathcal{F}, \tau)}\, \mathrm{d}\tau$: 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

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

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

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

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

$\sqrt{n} \int_\varepsilon^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}\, \mathrm{d}\tau$: 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

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

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

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

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).

Main algorithm which:

- achieves the Dudley-type regret bound

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

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

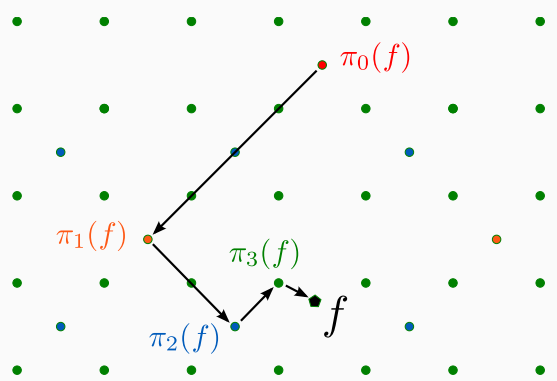
Instead of competing directly with $\mathcal{F}_\varepsilon$ for small ε (too many functions)

1. **create a "chain"** of refining approximations $\pi_0(f) \in \mathcal{F}^{(0)}, \pi_1(f) \in \mathcal{F}^{(1)}, \dots$ of any function $f \in \mathcal{F}_\varepsilon$ such that

$$\forall k \geqslant 0 \qquad \sup_{f \in \mathcal{F}} \|\pi_k(f) - f\|_\infty \leqslant \gamma/2^k$$

and

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



2. **compete with the chains**

$$\inf_{f \in \mathcal{F}_\varepsilon} \sum_{t=1}^{n} (y_t - f(x_t))^2 = \inf_{f \in \mathcal{F}_\varepsilon} \sum_{t=1}^{n} \left(y_t - \underbrace{\pi_0(f)}_{\in \mathcal{F}^{(0)}}(x_t) - \sum_{k=0}^{K_\varepsilon} \overbrace{\underbrace{[\pi_{k+1}(f) - \pi_k(f)]}_{\in \mathcal{G}^{(k)}}(x_t)}^{|\text{small increments}| \leqslant 3\gamma/2^{k+1}} \right)^2 .$$

where $\mathcal{G}^{(k)} \stackrel{\text{def}}{=} \{\pi_k(f) - \pi_{k-1}(f) : f \in \mathcal{F}\}$

COMPETE WITH THE CHAINS: TWO AGGREGATIONS LEVELS

$$\inf_{f\in\mathcal{F}_\varepsilon}\sum_{t=1}^{n}\big(y_t-f(x_t)\big)^2=\inf_{f\in\mathcal{F}_\varepsilon}\sum_{t=1}^{n}\Bigg(y_t-\underbrace{\pi_0(f)}_{\in\mathcal{F}^{(0)}}(x_t)-\sum_{k=0}^{K_\varepsilon}\overbrace{\underbrace{\big[\pi_{k+1}(f)-\pi_k(f)\big]}_{\in\mathcal{G}^{(k)}}(x_t)}^{|\text{small increments}|\leqslant 3\gamma/2^{k+1}}\Bigg)^2$$

It thus suffices to compete with

$$\inf_{f_0\in\mathcal{F}^{(0)}}\left\{\underbrace{\inf_{g_0\in\mathcal{G}^{(0)},\dots,g_{K_\varepsilon}\in\mathcal{G}^{(K_\varepsilon)}}\sum_{t=1}^{n}\Big(y_t-\big(f_0+g_0+\cdots+g_{K_\varepsilon}\big)(x_t)\Big)^2}_{\substack{\textbf{low-scale aggregation: gradient descents simultaneously at all } g_k\\ \text{Regret cost: } \sqrt{n}\int_\varepsilon^\gamma\sqrt{\log\mathcal{N}_\infty(\mathcal{F},\tau)}\,\mathrm{d}\tau}}\right\}$$

high-scale aggregation: run EWA to be competitive against all $f_0\in\mathcal{F}^{(0)}$

Regret cost: $\log(\mathrm{Card}\,\mathcal{F}^{(0)})=\log\mathcal{N}_\infty(\mathcal{F},\gamma)$

Let $\Delta_N \stackrel{\text{def}}{=} \{u \in \mathbb{R}_+^N : \sum_{i=1}^N u_i = 1\}$.

Setting: at each step $t \geqslant 1$, player plays $\widehat{u}_t \in \Delta_N$ and environment chooses convex and differentiable loss function $\ell_t : \Delta_N \to \mathbb{R}$.

The Exponentiated Gradient forecaster[5]

At each forecasting instance $t \geqslant 1$, define $\widehat{u}_t \in \Delta_N$ component-wise by

$$\widehat{u}_{k,t} \stackrel{\text{def}}{=} \frac{1}{Z_t} \exp\left(-\eta \sum_{s=1}^{t-1} \partial_{\widehat{u}_{k,s}} \ell_s(\widehat{u}_s) \right)$$

where Z_t is a normalization factor.

Regret bound: if $\|\nabla \ell_t\|_\infty \leqslant G$. For $\eta = G^{-1}\sqrt{2(\log N)/n}$

$$\sum_{t=1}^n \ell_t(\widehat{u}_t) \leqslant \min_{u \in \Delta_N} \sum_{t=1}^n \ell_t(u) + G\sqrt{2n \log N}$$

If G is small, $G\sqrt{n \log N}$ can be better than $B^2 \log N$.

[5] Kivinen and M. Warmuth (1997) and Cesa-Bianchi (1999)

EXPONENTIATED GRADIENT SIMULTANEOUSLY ON ALL CHAIN LINKS

Let Δ_N denote the simplex in $\mathbb{R}^N$.

Goal: minimize a sequence of multi-variable losses $(u^{(1)}, \ldots, u^{(K)}) \mapsto \ell_t(u^{(1)}, \ldots, u^{(K)})$ simultaneously over all variables $(u^{(1)}, \ldots, u^{(K)}) \in \Delta_{N_1} \times \ldots \times \Delta_{N_K}$.

The Multi-variable Exponentiated Gradient forecaster

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

initialization set $\widehat{u}_1^{(k)} \stackrel{\text{def}}{=} \left(\frac{1}{N_k}, \ldots, \frac{1}{N_k}\right) \in \Delta_{N_k}$ for all $k = 1, \ldots, K$.

:

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

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

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

end

Regret bound: If the ℓ_t are jointly convex and differentiable with $\|\nabla_{u^{(k)}} \ell_t\|_\infty \leqslant G^{(k)}$, then Multi-variable EG tuned with $\eta^{(k)} = \sqrt{2\log(N_k)/n}\,/G^{(k)}$ satisfies:

$$\sum_{t=1}^{n} \ell_t\left(\widehat{u}_t^{(1)}, \ldots, \widehat{u}_t^{(K)}\right) - \min_{u^{(1)}, \ldots, u^{(K)}} \sum_{t=1}^{n} \ell_t\left(u^{(1)}, \ldots, u^{(K)}\right) \leqslant \sqrt{2n} \sum_{k=1}^{K} G^{(k)} \sqrt{\log N_k}$$

HOW TO USE IT TO COMPETE WITH CHAINS?

Remember the goal:

$$\inf_{f_0 \in \mathcal{F}^{(0)}} \left\{ \underbrace{\inf_{g_0 \in \mathcal{G}^{(0)}, \ldots, g_{K_\varepsilon} \in \mathcal{G}^{(K_\varepsilon)}} \sum_{t=1}^{n} \Big(y_t - (f_0 + g_0 + \cdots + g_{K_\varepsilon})(x_t) \Big)^2}_{\text{low-scale aggregation: Multi-variable EG with loss functions:}} \right\}$$

$$\ell_t\big(u^{(1)}, \ldots, u^{(K_\varepsilon)}\big) = \big(y_t - f_0 - u^{(1)} \cdot g^{(0)} + \cdots + u^{(K_\varepsilon)} \cdot g^{(K_\varepsilon)}\big)^2$$

Regret cost: $\sqrt{n} \int_\varepsilon^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}\, \mathrm{d}\tau$

high-scale aggregation: run EWA to be competitive against all $f_0 \in \mathcal{F}^{(0)}$

Regret cost: $\log(\mathrm{Card}\mathcal{F}^{(0)}) = \log \mathcal{N}_\infty(\mathcal{F}, \gamma)$

Theorem (G. and Gerchinovitz 2015)

Let $B > 0$, $n \geqslant 1$, and $\gamma \in (B/n, B)$. Assume that $\max_{1 \leqslant t \leqslant n} |y_t| \leqslant B$ and that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leqslant B$. Then, Chaining EWA well-tuned (depending on γ, B, and n) satisfies:

$$\mathrm{Reg}_n(\mathcal{F}) \leqslant B^2\big(5 + 50 \log \mathcal{N}_\infty(\mathcal{F}, \gamma)\big) + 120 B \sqrt{n} \int_0^{\gamma/2} \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}\, d\varepsilon$$

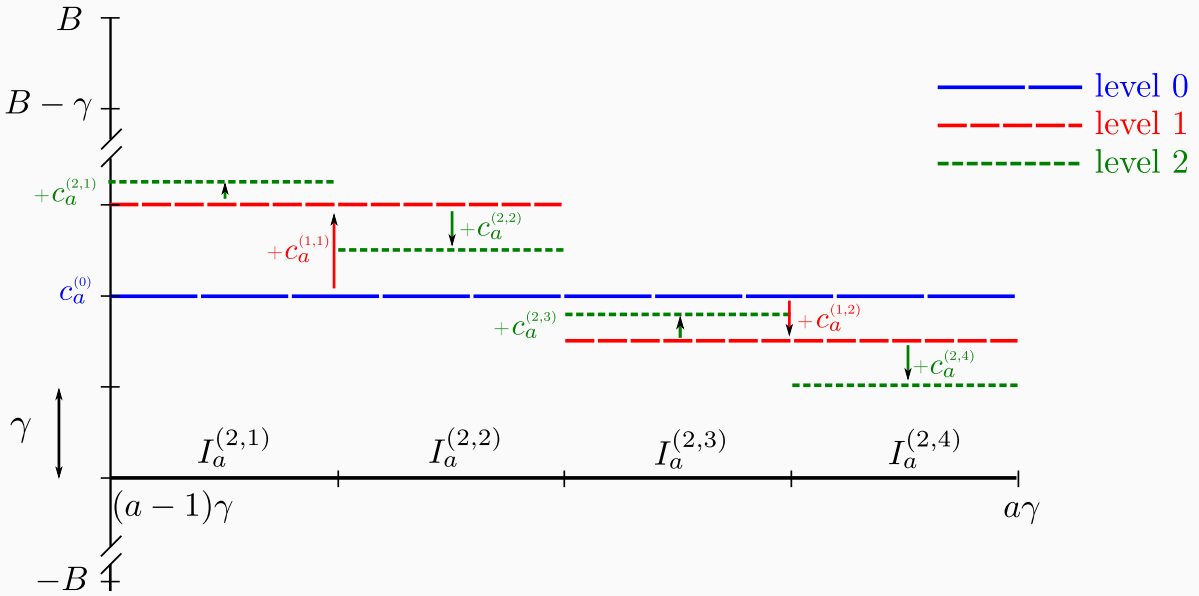
Remarks:

- the bound with $\varepsilon \neq 0$ and $\varepsilon\, n$ can also be obtained.
- calibrations in B and n should be possible by doubling trick and clipping

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.**

For Hölder functions: piecewise constant → piecewise polynomials

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

Can be extended to Lipschitz on $[0,1]^d$ functions at small cost.

We lose a log factor in the regret bound.

B. Guedj and his PhD student are implementing it for numerical experiments.

Extensions, current (and future) work

Goal: minimize the regret $\mathrm{Reg}_n = \sum_{t=1}^n \ell_t(\widehat{f}_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell_t(f)$ for generic sequences of loss functions (ℓ_t).

If the loss functions ℓ_t are convex and Lipschitz, we can achieve

$$\mathrm{Reg}_n(\mathcal{F}) \lesssim \underbrace{\cancel{\log \mathcal{N}_\infty(\mathcal{F}, \gamma)}}_{\text{Large scale term not possible (was thanks to exp-concavity)}} + \sqrt{n}\int_\varepsilon^1 \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \tau)}\,\mathrm{d}\tau + \varepsilon n$$

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

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

The rate $n^{(d+1)/(d+2)}$ 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).

EXTENSION TO LIPSCHITZ BANDITS?

Setting: at each step $t \geqslant 1$,

- simultaneously player plays $\widehat{x}_t \in \mathcal{F}$ (possibly random) and environment chooses ℓ_t convex Lipschitz (hidden)
- then player suffers and observes only $\ell_t(\widehat{x}_t)$

Goal: minimize the expected regret $\mathrm{Reg}_n(\mathcal{F}) \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^n \ell_t(\widehat{x}_t)\right] - \inf_{x\in\mathcal{F}} \sum_{t=1}^n \ell_t(x)$

	Full information	Bandits Feedback
$(\mathrm{Card}\mathcal{F} = K)$ + EWA	$\sqrt{n \log K}$	$\sqrt{nK}$
ε-net + EWA	$\sqrt{n \log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} + \varepsilon n$	$\sqrt{n\mathcal{N}_\infty(\mathcal{F}, \varepsilon)} + \varepsilon n$
Chaining	$\sqrt{n}\int_\varepsilon^1 \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, x)}\,\mathrm{d}x + \varepsilon n$	$\sqrt{n}\int_\varepsilon^1 \sqrt{\mathcal{N}_\infty(\mathcal{F}, x)}\,\mathrm{d}x + \varepsilon n$?

For $\mathcal{F} \subset [0,1]$ this would lead to $\mathcal{O}(\sqrt{n})$ regret $\rightarrow$ first constructive algorithm!

Difficulty: taking advantage of small loss ranges in bandits

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. “Analysis of two gradient-based algorithms for on-line regression”. In: *J. Comput. System Sci.* 59.3 (1999), pp. 392–411.

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≤ 1”. In: *Journal of Approximation Theory* 166 (2013), pp. 42–55.

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.


J. Kivinen and M.K. Warmuth. “Exponentiated Gradient Versus Gradient Descent for Linear Predictors”. In: *Information and Computation* 132.1 (1997), pp. 1–63.


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


G.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).


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.