A chaining algorithm for online nonparametric regression

Sébastien Gerchinovitz

Institut de Mathématiques de Toulouse, Université Toulouse III – Paul Sabatier

This is a joint work with Pierre Gaillard.

Introduction

We consider the problem of online nonparametric regression with individual sequences. We present an algorithm based on the chaining technique.

Outline of the talk:

1. The chaining technique in the stochastic setting
2. Our setting: online regression with individual sequences
3. Large (nonparametric) function sets
4. An algorithm based on the chaining technique

1. The chaining technique in the stochastic setting
2. Online regression with individual sequences
3. Large (nonparametric) function sets
4. An algorithm based on the chaining technique

Bounding the expected supremum of a stochastic process

Technique introduced by Dudley (1967). Let $(X_f)_{f\in\mathcal{F}}$ be a centered stochastic process (indexed by a finite metric space $(\mathcal{F}, d)$) with subgaussian increments:

$$\forall f, g \in \mathcal{F}, \quad \forall \lambda > 0, \quad \log \mathbb{E} e^{\lambda(X_f - X_g)} \leqslant \frac{\lambda^2}{2} d(f,g)^2 \ .$$

Goal: upper bound the quantity $\mathbb{E}\big[\sup_{f\in\mathcal{F}} X_f\big] = \mathbb{E}\big[\sup_{f\in\mathcal{F}}(X_f - X_{f_0})\big]$ for any $f_0 \in \mathcal{F}$.

Lemma (see, e.g., Boucheron et al. 2013)

Let $Z_1, \ldots, Z_N$ be such that $\log \mathbb{E} e^{\lambda Z_i} \leqslant \lambda^2 v/2$ for all $\lambda \in \mathbb{R}$ and $i \in [N]$. Then, $\mathbb{E} \max_{i=1,\ldots,N} Z_i \leqslant \sqrt{2v \log N}$.

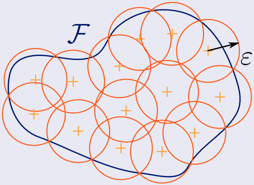
This lemma entails the pessimistic bound (correlations are not used):
$\mathbb{E}\big[\sup_{f\in\mathcal{F}}(X_f - X_{f_0})\big] \leqslant B\sqrt{2\log(\mathrm{card}\,\mathcal{F})}$ with $B = \sup_{f\in\mathcal{F}} d(f, f_0)$.

Discretizing the space $(\mathcal{F}, d)$ into small balls

Definition (metric entropy)

- Let $(\mathcal{F}, d)$ be a metric space of finite cardinality.
- ε-net: any subset $\mathcal{G} \subseteq \mathcal{F}$ such that

$$\forall f \in \mathcal{F},\ \exists g \in \mathcal{G} : \ d(f, g) \leqslant \varepsilon \quad \iff \quad \bigcup_{g \in \mathcal{G}} \bar{B}(g, \varepsilon) = \mathcal{F}$$



- $\mathcal{N}_d(\mathcal{F}, \varepsilon)$: smallest cardinality of an ε-net.
- metric entropy of $\mathcal{F}$ at scale ε: $\log \mathcal{N}_d(\mathcal{F}, \varepsilon)$.
  It measures the complexity (richness) of the space $(\mathcal{F}, d)$.

Successive refining discretizations:
Let $\mathcal{F}^{(0)} = \{f_0\}, \mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(K-1)}, \mathcal{F}^{(K)} = \mathcal{F}$ be minimal $B/2^k$-nets of $\mathcal{F}$:

$$\forall f \in \mathcal{F},\ \exists \pi_k(f) \in \mathcal{F}^{(k)},\ d\big(f, \pi_k(f)\big) \leqslant B/2^k \ .$$

Chaining argument: using the lemma at multiple scales, we get:

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}}(X_f - X_{f_0})\Big] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{k=1}^{K} \left(X_{\pi_k(f)} - X_{\pi_{k-1}(f)}\right)\right] \\
&\leqslant \sum_{k=1}^{K} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \Big(\underbrace{X_{\pi_k(f)} - X_{\pi_{k-1}(f)}}_{\text{small increments}}\Big)\right] \\
&\leqslant 6 \sum_{k=1}^{K} B 2^{-k} \sqrt{\log \mathcal{N}_d(\mathcal{F}, B/2^k)} \\
&\leqslant 12 \int_0^{B/2} \sqrt{\log \mathcal{N}_d(\mathcal{F}, \varepsilon)}\, d\varepsilon \ .
\end{aligned}$$

Multi-scale discretization to exploit the correlations

Successive refining discretizations:
Let $\mathcal{F}^{(0)} = \{f_0\}, \mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(K-1)}, \mathcal{F}^{(K)} = \mathcal{F}$ be minimal $B/2^k$-nets of $\mathcal{F}$:

$$\forall f \in \mathcal{F},\ \exists \pi_k(f) \in \mathcal{F}^{(k)},\ d\big(f, \pi_k(f)\big) \leqslant B/2^k \ .$$

Chaining argument: using the lemma at multiple scales, we get:

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}}(X_f - X_{f_0})\Big] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{k=1}^{K} \Big(X_{\pi_k(f)} - X_{\pi_{k-1}(f)}\Big)\right] \\
&\leqslant \sum_{k=1}^{K} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \Big(\underbrace{X_{\pi_k(f)} - X_{\pi_{k-1}(f)}}_{\text{small increments}}\Big)\right] \\
&\leqslant 6 \sum_{k=1}^{K} B 2^{-k} \sqrt{\log \mathcal{N}_d(\mathcal{F}, B/2^k)} \\
&\leqslant 12 \underbrace{\int_0^{B/2} \sqrt{\log \mathcal{N}_d(\mathcal{F}, \varepsilon)}\, d\varepsilon}_{\text{Dudley's entropy integral}} \ .
\end{aligned}$$

Setting: online regression with individual sequences

Prediction task: at each time $t \in \mathbb{N}^*$, predict the observation $y_t \in \mathbb{R}$ from the input $x_t \in \mathcal{X}$, on the basis of the past data $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$.

Initial step: the environment chooses arbitrary deterministic sequences $(y_t)_{t \geqslant 1}$ in $\mathbb{R}$ and $(x_t)_{t \geqslant 1}$ in $\mathcal{X}$ but the forecaster has not access to them.

At each time round $t \in \mathbb{N}^*$,

1. The environment reveals the input $x_t \in \mathcal{X}$.
2. The forecaster chooses a prediction $\widehat{y}_t \in \mathbb{R}$.
3. The environment reveals the observation $y_t \in \mathbb{R}$ and the forecaster incurs the loss $(y_t - \widehat{y}_t)^2$.

Goal: minimizing regret

Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a set of functions.

Goal of the forecaster: on the long run, to predict almost as well as the best function $f \in \mathcal{F}$ in hindsight, that is, to minimize the regret:

$$\mathrm{Reg}_T(\mathcal{F}) \triangleq \sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(y_t - f(x_t)\right)^2 .$$

Individual sequence setting: our goal is to minimize the regret $\mathrm{Reg}_T(\mathcal{F})$ uniformly over all sequences $(y_t)_{t \geqslant 1}$ in $[-B, B]$ and $(x_t)_{t \geqslant 1}$ in $\mathcal{X}$; typically:

$$\sup_{\substack{|y_t| \leqslant B \\ x_t \in \mathcal{X}}} \left\{ \frac{1}{T} \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} (y_t - f(x_t))^2 \right\} \leqslant o(1) \quad \text{when } T \to +\infty .$$

Particular case: finite $\mathcal{F}$

Assume that $\mathcal{F} = \{f_1, f_2, \ldots, f_N\} \subseteq \mathbb{R}^{\mathcal{X}}$ is finite. We can use a well-known algorithm studied, e.g., by Kivinen and Warmuth (1999) and Vovk (2001):

Algorithm (Exponentially Weighted Average forecaster (EWA))

Parameter: $\eta > 0$

At each round $t \geqslant 1$,

- *Using past data, compute the weight vector* $\widehat{\boldsymbol{w}}_t = (\widehat{w}_{t,1}, \ldots, \widehat{w}_{t,N})$ *as*

$$\widehat{w}_{t,j} \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} (y_s - f_j(x_s))^2\right)}{\sum_{j'=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} (y_s - f_{j'}(x_s))^2\right)} \,, \quad 1 \leqslant j \leqslant N \,;$$

- *Compute the convex combination (convex aggregate):*

$$\widehat{y}_t \triangleq \sum_{j=1}^{N} \widehat{w}_{t,j} \, f_j(x_t) \,.$$

Regret guarantee when $\mathcal{F}$ is finite

If $\mathcal{F}$ contains N functions, then we have a $\mathcal{O}(\log N)$ upper bound on the regret under the boundedness assumption:

$$|y_1|, \ldots, |y_T| \leqslant B \qquad \text{and} \qquad \|f_1\|_\infty, \ldots, \|f_N\|_\infty \leqslant B \ .$$

Theorem (Kivinen and Warmuth 1999)

Assume that $\mathcal{F} = \{f_1, f_2, \ldots, f_N\} \subseteq [-B, B]^{\mathcal{X}}$.

Then, the EWA algorithm tuned with $\eta = 1/(8B^2)$ satisfies: for all sequences $(y_t)_{t \geqslant 1}$ in $[-B, B]$ and $(x_t)_{t \geqslant 1}$ in $\mathcal{X}$, for all $T \geqslant 1$,

$$\sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 - \min_{1 \leqslant j \leqslant N} \sum_{t=1}^{T} \left(y_t - f_j(x_t)\right)^2 \leqslant 8B^2 \log N \ .$$

Remark 1: the requirement $\forall j, \|f_j\|_\infty \leqslant B$ can be removed via clipping.

Regret guarantee when $\mathcal{F}$ is finite

If $\mathcal{F}$ contains N functions, then we have a $\mathcal{O}(\log N)$ upper bound on the regret under the boundedness assumption:

$$|y_1|, \ldots, |y_T| \leqslant B \qquad \text{and} \qquad \|f_1\|_\infty, \ldots, \|f_N\|_\infty \leqslant B \ .$$

Theorem (Kivinen and Warmuth 1999)

Assume that $\mathcal{F} = \{f_1, f_2, \ldots, f_N\} \subseteq [-B, B]^{\mathcal{X}}$.

Then, the EWA algorithm tuned with $\eta = 1/(8B^2)$ satisfies: for all sequences $(y_t)_{t \geqslant 1}$ in $[-B, B]$ and $(x_t)_{t \geqslant 1}$ in $\mathcal{X}$, for all $T \geqslant 1$,

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \min_{1 \leqslant j \leqslant N} \sum_{t=1}^{T} (y_t - f_j(x_t))^2 \leqslant 8B^2 \log N \ .$$

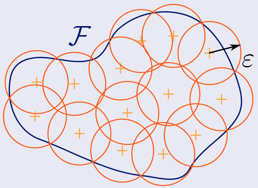
Remark 1: the requirement $\forall j, \|f_j\|_\infty \leqslant B$ can be removed via clipping.
Remark 2: we can obtain a similar bound if $B = \max_{1 \leqslant t \leqslant T} |y_t|$ is unknown.

Large function sets $\mathcal{F}$: finite approximation

Definition (metric entropy for sup norm)

- Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a set of bounded functions endowed with the sup norm $\|f\|_\infty \triangleq \sup_{x \in \mathcal{X}} |f(x)|$.
- ε-net: any subset $\mathcal{G} \subseteq \mathcal{F}$ such that

$$\forall f \in \mathcal{F},\ \exists g \in \mathcal{G} : \ \|f - g\|_\infty \leqslant \varepsilon \ .$$



- $\mathcal{N}_\infty(\mathcal{F}, \varepsilon)$: smallest cardinality of an ε-net.
- metric entropy of $\mathcal{F}$ at scale ε: $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)$.

Large function sets $\mathcal{F}$: finite approximation (2)

Assume that $\mathcal{F}$ is infinite (the EWA algorithm cannot be used). Small regret is still achievable if $\mathcal{F}$ can be well approximated by a finite set.

Discretizing $\mathcal{F}$ (Vovk, 2006): approximate $\mathcal{F}$ with a minimal ε-net and run the EWA algorithm on this finite subset:

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \min_{1 \leqslant j \leqslant \mathcal{N}_\infty(\mathcal{F},\varepsilon)} \sum_{t=1}^{T} \big(y_t - f_j(x_t)\big)^2 + 8B^2 \log \mathcal{N}_\infty(\mathcal{F},\varepsilon)$$

$$\leqslant \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \big(y_t - f(x_t)\big)^2 + T\varepsilon^2 + 4TB\varepsilon + 8B^2 \log \mathcal{N}_\infty(\mathcal{F},\varepsilon)$$

Finite-dimensional case: given $\varphi_j : \mathcal{X} \to [-B, B]$ and a compact set $\Theta \subseteq \mathbb{R}^d$, define

$$\mathcal{F} = \left\{ \sum_{j=1}^{d} \theta_j \varphi_j : \theta \in \Theta \right\} \subseteq \mathbb{R}^{\mathcal{X}} .$$

Note that $\mathcal{N}_\infty(\mathcal{F},\varepsilon) \lesssim (1/\varepsilon)^d$. Choosing $\varepsilon \approx 1/T$ yields a regret at most of the order of $d \log(T)$, which is optimal (parametric rate).

What if $\mathcal{F}$ is very large (nonparametric)?

Nonparametric set: assume that $\mathcal{F}$ is much larger than in the finite-dimensional case:

$$\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx (1/\varepsilon)^p \qquad \text{as} \qquad \varepsilon \to 0 \ .$$

Example: Hölder class $\mathcal{F} \subseteq \mathbb{R}^{[0,1]}$ of regularity $\beta = q + \alpha$:

$$\left| f^{(q)}(x) - f^{(q)}(y) \right| \leqslant \lambda |x - y|^\alpha \quad \text{and} \quad \forall k \leqslant q, \ \|f^{(k)}\|_\infty \leqslant B$$

In this case, $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx \varepsilon^{-1/\beta}$ so that $p = 1/\beta$.

EWA is suboptimal: the regret bound $T\varepsilon^2 + 4TB\varepsilon + 8B^2 \log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)$ becomes roughly $T\varepsilon + (1/\varepsilon)^p$. Optimizing in ε only yields:

$$\sum_{t=1}^{T} \big(y_t - \widehat{y}_t\big)^2 \leqslant \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \big(y_t - f(x_t)\big)^2 + \mathcal{O}\big(T^{p/(p+1)}\big) \ ,$$

which is worse than the optimal rate $\mathcal{O}\big(T^{p/(p+2)}\big)$.

We still assume that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx (1/\varepsilon)^p$ as $\varepsilon \to 0$.

Optimal regret: through a non-constructive approach (reduction to a stochastic problem via von Neumann minimax theorem), Rakhlin and Sridharan (2014) proved that, if $p \in (0, 2)$, then

$$
\begin{aligned}
\mathrm{Reg}_T(\mathcal{F}) &\leqslant c_1 B^2 \big(1 + \log \mathcal{N}_\infty(\mathcal{F}, \gamma)\big) + c_2 B \sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} d\varepsilon \\
&\lesssim \gamma^{-p} + \sqrt{T} \int_0^\gamma \varepsilon^{-p/2} d\varepsilon \\
&\lesssim T^{p/(p+2)} \qquad \text{for } \gamma = T^{-1/(p+2)}.
\end{aligned}
$$

The rate $T^{p/(p+2)}$ is better than $T^{p/(p+1)}$ obtained previously with EWA, and it is (in a sense) optimal.

We still assume that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx (1/\varepsilon)^p$ as $\varepsilon \to 0$.

Optimal regret: through a non-constructive approach (reduction to a stochastic problem via von Neumann minimax theorem), Rakhlin and Sridharan (2014) proved that, if $p \in (0, 2)$, then

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{F}) &\leqslant c_1 B^2 \big(1 + \log \mathcal{N}_\infty(\mathcal{F}, \gamma)\big) + c_2 B\sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} d\varepsilon \\
&\lesssim \gamma^{-p} + \sqrt{T} \int_0^\gamma \varepsilon^{-p/2} d\varepsilon \\
&\lesssim T^{p/(p+2)} \qquad \text{for } \gamma = T^{-1/(p+2)}.
\end{aligned}$$

Example (Hölder class with regularity β):
Since $p = 1/\beta$, we get $\mathrm{Reg}_T(\mathcal{F})/T = \mathcal{O}\big(T^{-2\beta/(2\beta+1)}\big)$ if $\beta > 1/2$.
Therefore, same rate as in the statistical setting (for $\beta > 1/2$).

Optimal rates by Rakhlin and Sridharan (2014)

We still assume that $\log \mathcal{N}_\infty(\mathcal{F},\varepsilon) \approx (1/\varepsilon)^p$ as $\varepsilon \to 0$.

Optimal regret: through a non-constructive approach (reduction to a stochastic problem via von Neumann minimax theorem), Rakhlin and Sridharan (2014) proved that, if $p \in (0,2)$, then

$$\begin{aligned}
\mathrm{Reg}_T(\mathcal{F}) &\leqslant c_1 B^2\big(1+\log \mathcal{N}_\infty(\mathcal{F},\gamma)\big) + c_2 B\sqrt{T}\int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)}\,d\varepsilon \\
&\lesssim \gamma^{-p} + \sqrt{T}\int_0^\gamma \varepsilon^{-p/2}\,d\varepsilon \\
&\lesssim T^{p/(p+2)} \qquad \text{for } \gamma = T^{-1/(p+2)}.
\end{aligned}$$

The above integral is a Dudley entropy integral.

- In statistical learning with i.i.d. data, useful to derive risk bounds for empirical risk minimizers (e.g., Massart 2007; Rakhlin et al. 2013).
- Also appears in online learning with individual sequences. Earlier appearances: Opper and Haussler (1997); Cesa-Bianchi and Lugosi (1999, 2001).

Our contributions

1. We provide an explicit algorithm that achieves the Dudley-type regret bound (when $p \in (0,2)$):

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant c_1 B^2 \big(1+\log \mathcal{N}_\infty(\mathcal{F},\gamma)\big) + c_2 B\sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)}\, d\varepsilon\ .$$

   Nota: contrary to Rakhlin and Sridharan (2014), our bounds are not in terms of the stronger notion of *sequential entropy*.

2. This algorithm uses ideas from the chaining technique, and relies on a new subroutine (Multi-variable Exponentiated Gradient algorithm) to perform optimization at different scales simultaneously.

3. We address computational issues by showing how to construct more efficient and quasi-optimal ε-nets (for Hölder classes).

Linearizing the square loss can help locally (1)

Suppose we play with loss functions $\boldsymbol{u} \mapsto \ell_t(\boldsymbol{u})$, $t \geqslant 1$, that are convex and differentiable over the simplex $\Delta_N = \left\{\boldsymbol{u} \in \mathbb{R}_+^N : \sum_{i=1}^N u_i = 1\right\}$.

Algorithm (Exponentiated Gradient—EG)

Parameter: $\eta > 0$

At each round $t \geqslant 1$, compute the weight vector $\widehat{\boldsymbol{u}}_t \in \Delta_N$ by

$$\widehat{u}_{t,j} \triangleq \frac{1}{Z_t} \exp\left(-\eta \sum_{s=1}^{t-1} \partial_{\widehat{u}_{s,j}} \ell_s(\widehat{\boldsymbol{u}}_s)\right) \,, \quad 1 \leqslant j \leqslant N \,.$$

Theorem (Kivinen and Warmuth 1999 and Cesa-Bianchi 1999)

Assume ℓ_t convex, diff, and $\|\nabla \ell_t\|_\infty \leqslant G$. For $\eta = G^{-1}\sqrt{2\log(N)/T}$,

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

Linearizing the square loss can help locally (1)

Suppose we play with loss functions $\boldsymbol{u} \mapsto \ell_t(\boldsymbol{u})$, $t \geqslant 1$, that are convex and differentiable over the simplex $\Delta_N = \left\{ \boldsymbol{u} \in \mathbb{R}_+^N : \sum_{i=1}^N u_i = 1 \right\}$.

Algorithm (Exponentiated Gradient—EG)

Parameter: $\eta > 0$
At each round $t \geqslant 1$, *compute the weight vector* $\widehat{\boldsymbol{u}}_t \in \Delta_N$ *by*

$$\widehat{u}_{t,j} \triangleq \frac{1}{Z_t} \exp\left(-\eta \sum_{s=1}^{t-1} \partial_{\widehat{u}_{s,j}} \ell_s(\widehat{\boldsymbol{u}}_s) \right) , \quad 1 \leqslant j \leqslant N .$$

Theorem (Kivinen and Warmuth 1999 and Cesa-Bianchi 1999)

Assume ℓ_t *convex, diff, and* $\|\nabla \ell_t\|_\infty \leqslant G$. *For* $\eta = G^{-1}\sqrt{2\log(N)/T}$,

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

Linearizing the square loss can help locally (2)

Application: we want to predict almost as well as the best function in $\mathcal{F} = \{f_0 + g_j : j = 1, \ldots, N\}$ with $\|g_j\|_\infty$ small (neighbors of f_0).

We use EG with $\ell_t(\boldsymbol{u}) = \Big(y_t - f_0(x_t) - \sum_{j=1}^N u_j g_j(x_t)\Big)^2$, $\boldsymbol{u} \in \Delta_N$.

Since $\|\nabla \ell_t\|_\infty \lesssim B \max_j \|g_j\|_\infty$, the EG algorithm satisfies:

$$\sum_{t=1}^T \Big(\underbrace{y_t - f_0(x_t) - \sum_{j=1}^N \widehat{u}_{t,j} g_j(x_t)}_{=\widehat{y}_t}\Big)^2 \leqslant \min_{1 \leqslant j \leqslant N} \sum_{t=1}^T (y_t - f_0(x_t) - g_j(x_t))^2 + \square B \max_{1 \leqslant j \leqslant N} \|g_j\|_\infty \sqrt{T \log N}$$

Advantage: the above regret bound $B \max_j \|g_j\|_\infty \sqrt{T \log N}$ improves on $B^2 \log N$ (obtained by EWA) when $\max_j \|g_j\|_\infty \ll B\sqrt{\log(N)/T}$.

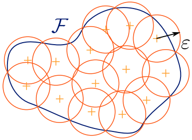
Thus, linearizing the square loss can help if the functions in $\mathcal{F}$ are close.

Turning the chaining technique into an online algorithm

We still assume that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \approx (1/\varepsilon)^p$ as $\varepsilon \to 0$. Recall that we want to prove a bound of the form:

$$\sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 \leqslant \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(y_t - f(x_t)\right)^2 + [\text{small term}]$$

Chaining principle: as previously, we discretize $\mathcal{F}$ and use projections $\pi_k(f)$ such that $\sup_f \|\pi_k(f) - f\|_\infty \leqslant \gamma/2^k$ for all $k \geqslant 0$.



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

Aggregation at two different levels

$$\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\big(y_t-f(x_t)\big)^2=\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\Bigg(y_t-\underbrace{\pi_0(f)}_{\in\mathcal{F}^{(0)}}(x_t)-\sum_{k=1}^{\infty}\underbrace{\big[\pi_k(f)-\pi_{k-1}(f)\big]}_{\in\mathcal{G}^{(k)}}(x_t)\Bigg)^2$$

Sufficient goal:

$$\sum_{t=1}^{T}\big(y_t-\widehat{y}_t\big)^2\leqslant\inf_{f_0,g_1,\ldots,g_K}\sum_{t=1}^{T}\big(y_t-(f_0+g_1+\ldots+g_K)(x_t)\big)^2+[\text{small term}]$$

Two aggregation levels:

$$\left.\begin{aligned} f_{0,1} &\xrightarrow[\text{}]{\text{low scale gradient descent}} \widehat{f}_{t,1}\\ f_{0,2} &\longrightarrow \widehat{f}_{t,2}\\ &\vdots\\ f_{0,N_0} &\longrightarrow \widehat{f}_{t,N_0}\end{aligned}\right\}\xrightarrow{\text{high scale EWA}}\widehat{y}_t=\sum_{j=1}^{N_0}\widehat{w}_{t,j}\widehat{f}_{t,j}(x_t)$$

Combining two regret guarantees

High-scale aggregation Using an Exponentially Weighted Average (EWA) forecaster $\widehat{f}_t = \sum_{j=1}^{N_0} \widehat{w}_{t,j} \widehat{f}_{t,j}$ yields

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \min_{1 \leqslant j \leqslant N_0} \sum_{t=1}^{T} \left(y_t - \widehat{f}_{t,j}(x_t) \right)^2 + \Box B^2 \log N_0$$

Low-scale aggregation Recall that $\mathcal{G}^{(k)} = \{\pi_k(f) - \pi_{k-1}(f) : \ f \in \mathcal{F}\}$. Denote $\mathcal{G}^{(k)} = \left\{g_1^{(k)}, \ldots, g_{N_k}^{(k)}\right\}$.

We designed a multi-variable extension of the Exponentiated Gradient algorithm:

$$\widehat{f}_{t,j} \triangleq f_{0,j} + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \widehat{u}_{t,i}^{(j,k)} g_i^{(k)}$$

which yields, for all $j = 1, \ldots, N_0$,

$$\sum_{t=1}^{T} \left(y_t - \widehat{f}_{t,j}(x_t) \right)^2 \leqslant \min_{g_1, \ldots, g_K} \sum_{t=1}^{T} \left(y_t - (f_{0,j} + g_1 + \ldots + g_K)(x_t) \right)^2 + 120 B \sqrt{T} \int_0^{\gamma/2} \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)} d\varepsilon \ .$$

Main result

The next theorem indicates that the Chaining Exponentially Weighted Average forecaster satisfies a Dudley-type regret bound.

Theorem (Gaillard and G., 2015)

Let $B > 0$, $T \geqslant 1$, and $\gamma \in \left(\frac{B}{T}, B\right)$.

- *Assume that $\max_{1 \leqslant t \leqslant T} |y_t| \leqslant B$ and that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leqslant B$.*
- *Assume that $(\mathcal{F}, \|\cdot\|_\infty)$ is totally bounded and define $\mathcal{F}^{(0)}$ and $\mathcal{G}^{(k)}$ as above.*

Then, the Chaining Exponentially Weighted Average forecaster (tuned with appropriate parameters) satisfies:

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant B^2\left(5 + 50 \log \mathcal{N}_\infty(\mathcal{F}, \gamma)\right) + 120 B\sqrt{T} \int_0^{\gamma/2} \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}\, d\varepsilon\ .$$

Computational issues: dyadic discretization

We assume that $\mathcal{F} = \{f : [0,1] \to [-B, B] : \ f \text{ is 1-Lipschitz}\}$.

Regret bound:
We know that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) = \mathcal{O}(\varepsilon^{-1})$.
Therefore, our algorithm obtains $\mathrm{Reg}_T(\mathcal{F}) = \mathcal{O}(T^{1/3})$, which is optimal.

Computational issue:
Our algorithm updates $\exp(\mathcal{O}(T))$ weights at every round t.
Hence very poor time and space computational complexities.

Solution:
$\mathcal{F}$ has a sufficiently nice structure that can be exploited to construct computationally manageable ε-nets with quasi-optimal cardinality.

For example: piecewise-constant approximations on a dyadic discretization lead to $\mathcal{O}(T^{1/3} \log T)$ regret and per-round time complexity.

Conclusion

- We designed an explicit algorithm with a Dudley-type regret bound for online nonparametric regression.
- We provided an efficient implementation for Hölder classes.

Thank you for your attention!

Advertisement: we organize a workshop about *Sequential learning and applications* in Toulouse on November 9-10, 2015.

http://www.irit.fr/cimi-machine-learning/node/8

Appendix

5 Computational issues: dyadic discretization

Lipschitz class $\mathcal{F}$: a computationally efficient discretization

We assume that $\mathcal{F} = \{ f : [0,1] \to [-B, B] : \ f \text{ is 1-Lipschitz} \}$.

Regret bound:
We know that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) = \mathcal{O}(\varepsilon^{-1})$.
Therefore, our algorithm obtains $\mathrm{Reg}_T(\mathcal{F}) = \mathcal{O}(T^{1/3})$, which is optimal.

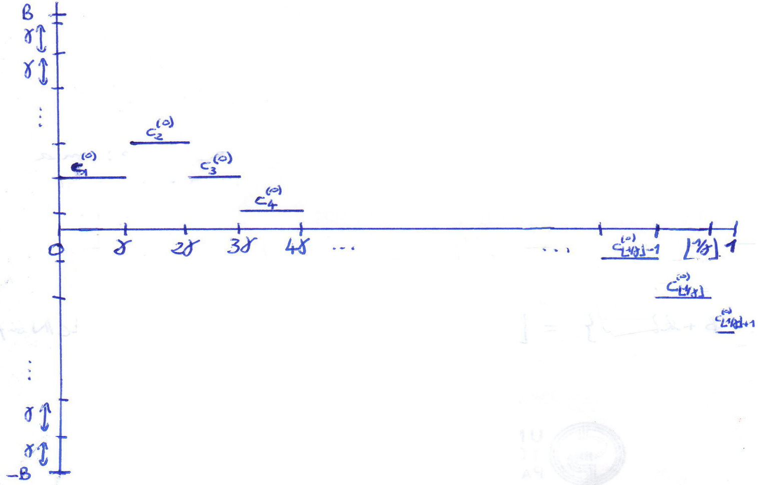
Computational issue:
Our algorithm updates exponentially many weights at every round t.
Hence poor time and space computational complexities.

Solution:
$\mathcal{F}$ has a sufficiently nice structure that can be exploited to construct computationally manageable ε-nets with quasi-optimal cardinality.

High-level discretization (piecewise-constant approximation)

- Partition the x-axis $[0,1]$: $I_a \triangleq \big[(a-1)\gamma, a\gamma\big)$, $a = 1, \ldots, \frac{1}{\gamma}$.
- Discretize the y-axis $[-B,B]$: $\mathcal{C}^{(0)} = \big\{-B + j\gamma : \ j = 0, \ldots, \frac{2B}{\gamma}\big\}$.

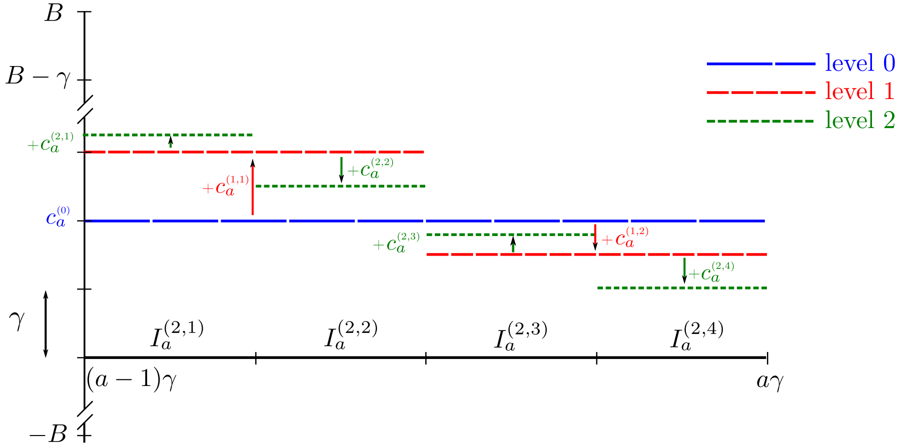
$\mathcal{F}^{(0)}$: set of piecewise-constant functions $f^{(0)}(x) = \sum_{a=1}^{1/\gamma} c_a^{(0)} \mathbb{I}_{x \in I_a}$, $c_a^{(0)} \in \mathcal{C}^{(0)}$.



Low-level discretization (dyadic approximation)

$\mathcal{F}^{(M)}$: set of all functions $f_c : [0,1] \to \mathbb{R}$ of the form

$$f_c(x) = \underbrace{\sum_{a=1}^{1/\gamma} c_a^{(0)} \mathbb{I}_{x \in I_a}}_{f^{(0)}(x)} + \sum_{m=1}^{M} \underbrace{\sum_{a=1}^{1/\gamma} \sum_{n=1}^{2^m} c_a^{(m,n)} \mathbb{I}_{x \in I_a^{(m,n)}}}_{g^{(m)}(x)} .$$



Regret and computational efficiency

Theorem (Gaillard and G., 2015)

Let $B > 0$, $T \geqslant 2$, and $\mathcal{F}$ be the set of all 1-Lipschitz functions from $[0, 1]$ to $[-B, B]$. Assume that $\max_{1 \leqslant t \leqslant T} |y_t| \leqslant B$.

Then, the Dyadic Chaining Algorithm (see preprint) satisfies, for some absolute constant $c > 0$,

$$\mathrm{Reg}_T(\mathcal{F}) \leqslant c \max\{B, B^2\}\, T^{1/3} \log T \ .$$

Remark: additional log factor, but computationally tractable:

- per-round time complexity: $\mathcal{O}\big(T^{1/3} \log T\big)$;
- space complexity: $\mathcal{O}\big(T^{4/3} \log T\big)$.

Bibliographie I

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.

N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *J. Comput. System Sci.*, 59(3):392–411, 1999.

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

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

R.M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167, 1999.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

M. Opper and D. Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction, and Distribution*. Spinger Verlag, 1997.

A. Rakhlin and K. Sridharan. Online nonparametric regression. *JMLR W&CP*, 35 (Proceedings of COLT 2014):1232–1264, 2014.

Bibliographie II

A. Rakhlin, K. Sridharan, and A.B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 2013. URL http://arxiv.org/abs/1308.1147. To appear.

V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69:213–248, 2001.

V. Vovk. Metric entropy in competitive on-line prediction. *arXiv*, 2006. URL http://arxiv.org/abs/cs.LG/0609045.