S. Gerchinovitz

Chaining and prediction of individual sequences

(joint with N. Cesa-Bianchi, P. Gaillard, C. Gentile)

We will focus on how chaining (probabilistic tool) can be used to construct algorithms for nonparametric online learning of individual (non-stochastic) sequences.

1. "Classical chaining" : a brief recap

Technique to control $\max_{t \in \mathcal{T}} X_t$ where
- $(\mathcal{T}, d)$: metric space
- $\mathbb{E}(X_t) = 0$
- small increments $X_t - X_s$ when $d(t,s)$ small

1.1. Small $\mathcal{T}$

Lemma: Let $X_1, \dots, X_N$ be real r.v. such that: $\exists v \geq 0$,
$\forall i, \ \forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\frac{\lambda^2 v}{2}}$ (subgaussianity)
Then, $\mathbb{E}\left[\max_{1 \leq i \leq N} X_i\right] \leq \sqrt{2 v \log N}$

NB: for $X_i \overset{iid}{\sim} \mathcal{N}(0, v)$,
$\mathbb{E}\left[\max_{1 \leq i \leq N} X_i\right] \underset{N \to +\infty}{\sim} \sqrt{2 v \log N}$

Proof (Pisier-like argument):

$$\mathbb{E}\left[\max_{1 \leq i \leq N} X_i\right] = \frac{1}{\lambda} \log \exp\left(\lambda \, \mathbb{E}\left[\max_i X_i\right]\right) \quad \text{for any } \lambda > 0$$

$$\leq \frac{1}{\lambda} \log \mathbb{E}\left[\max_{1 \leq i \leq N} e^{\lambda X_i}\right] \quad \text{by Jensen's ineq.}$$

$$\leq \frac{1}{\lambda} \log\left(\sum_{i=1}^{N} \mathbb{E}\left[e^{\lambda X_i}\right]\right)$$

$$\leq \frac{\log N}{\lambda} + \frac{\lambda v}{2} = \sqrt{2 v \log N} \quad \text{for } \lambda := \sqrt{\frac{2 \log N}{v}}$$

1.2. What about $\mathbb{E}\left[\max_{t\in\mathcal{T}} X_t\right]$ for finite but large $\mathcal{T}$?

Assume that:

(i) $\mathcal{T}$ is finite (for simplicity)

(ii) $\mathbb{E}(X_t) = 0 \quad \forall t\in\mathcal{T}$

(iii) the increments are subgaussian:

$$\forall s,t\in\mathcal{T},\ \forall\lambda\in\mathbb{R},\ \mathbb{E}\left[e^{\lambda(X_t-X_s)}\right] \leq e^{\frac{\lambda^2}{2}d^2(s,t)}$$

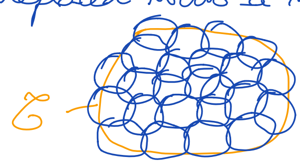
Letting $B := \max_{s,t\in\mathcal{T}} d(s,t)$, we can use the above lemma with $v = B^2$ and get

$$\mathbb{E}\left[\max_{t\in\mathcal{T}} X_t\right] = \mathbb{E}\left[\max_{t\in\mathcal{T}} (X_t - X_{t_0})\right] \leq \sqrt{2B^2\log|\mathcal{T}|}$$

Issue: This upper bound is usually very pessimistic!

Indeed, (iii) implies that $\mathrm{Var}(X_t - X_s) \leq d^2(t,s)$ so that X_t and X_s are highly correlated when $d(t,s)$ is small

$\rightsquigarrow$ the cardinality $|\mathcal{T}|$ can be replaced with a much smaller number?



$$\bigcup_{i=1}^{N} \overline{B}(t_i,\varepsilon) = \mathcal{T}$$

For $\varepsilon>0$, define

- ε-covering: any finite set $\{t_1,\dots,t_N\}$ s.t. the closed balls $\overline{B}(t_i,\varepsilon)$ cover $\mathcal{T}$
- $\mathcal{N}(\mathcal{T},d,\varepsilon)$: the smallest cardinality N of an ε-covering (or $+\infty$ if none)

$\log\mathcal{N}(\mathcal{T},d,\varepsilon)$ is called "metric entropy" at scale ε. Quantifies richness of $(\mathcal{T},d)$.

Ex $\log\mathcal{N}([0,1]^d, \|\cdot\|, \varepsilon) \approx \log\left(\frac{1}{\varepsilon}\right)^d \approx d\log\frac{1}{\varepsilon}$

Proposition (Dudley '67). NB: there exist several refinements/extensions.

Assume (i), (ii), and (iii) above. Then, setting $B := \max_{s,t\in\mathcal{T}} d(s,t)$,

$$\mathbb{E}\left[\max_{t\in\mathcal{T}} X_t\right] \leq 12 \int_0^{\frac{B}{2}} \sqrt{\log \mathcal{N}(\mathcal{T}, d, \varepsilon)}\, d\varepsilon$$

centers of balls of radius ε_m.

Proof: $\varepsilon_m := \frac{B}{2^m} \rightsquigarrow$ discretization $\mathcal{T}_m \subseteq \mathcal{T}$, $|\mathcal{T}_m| = \mathcal{N}(\mathcal{T}, d, \varepsilon_m)$

s.t. $\mathcal{T} = \bigcup_{t\in\mathcal{T}_m} \overline{B}(t, \varepsilon_m)$

$$\mathcal{T}_0 = \{t_0\} \subseteq \mathcal{T}_1 \subseteq \cdots \subseteq \mathcal{T}_M = \mathcal{T}.$$

We set $\pi_m(t) \in \operatorname{argmin}_{s\in\mathcal{T}_m} d(s,t)$. (approximation of t at scale ε_m)

$$\mathbb{E}\left[\max_{t\in\mathcal{T}} X_t\right] = \mathbb{E}\left[\max_{t\in\mathcal{T}} (X_t - X_{t_0})\right]$$

$$= \mathbb{E}\left[\max_{t\in\mathcal{T}} \sum_{m=0}^{M-1} \left(X_{\pi_{m+1}(t)} - X_{\pi_m(t)}\right)\right]$$

$$\leq \sum_{m=0}^{M-1} \mathbb{E}\left[\max_{t\in\mathcal{T}} \left(X_{\pi_{m+1}(t)} - X_{\pi_m(t)}\right)\right]$$

$$\overset{\text{lemma}}{\leq} \sum_{m=0}^{M-1} \sqrt{2 B_m^2 \log\left(|\mathcal{F}_m| \cdot |\mathcal{F}_{m+1}|\right)}$$

with $B_m = \max_{t\in\mathcal{T}} d\left(\pi_{m+1}(t), \pi_m(t)\right)$

$\leq \frac{B}{2^{m+1}} + \frac{B}{2^m} = \frac{3B}{2^{m+1}}$ by the triangle inequality

$$\leq 6 \sum_{m=0}^{M-1} \frac{B}{2^{m+1}} \sqrt{\log \mathcal{N}\left(\mathcal{T}, d, \frac{B}{2^{m+1}}\right)}$$

$$\lesssim 12 \sum_{m=0}^{M-1} \int_{\frac{B}{2^{m+2}}}^{\frac{B}{2^{m+1}}} \sqrt{\log \mathcal{N}(\hat{\mathcal{G}}, d, \varepsilon)}\, d\varepsilon$$

$$\lesssim 12 \int_0^{\frac{B}{2}} \sqrt{\log \mathcal{N}(\hat{\mathcal{G}}, d, \varepsilon)}\, d\varepsilon \qquad \blacksquare$$

2. Chaining in (nonparametric) online learning

2.1. Online learning problem:

Sequence $(x_1, y_1), (x_2, y_2), \ldots$ which is arbitrary.

At any round $t \geq 1$:

- We observe $x_t \in \mathcal{X}$
- We predict $\hat{y}_t \in \mathcal{Y}$
- We observe $y_t \in \mathcal{Y}$ and suffer loss $\ell(\hat{y}_t, y_t)$ for some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

Goal of the learner: minimize the regret

$$R_T(\mathcal{F}) := \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)$$

for some large (nonparametric) function set $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, and for all sequences $(x_t, y_t)_{1 \leq t \leq T}$.

We have an analogue to Section 1, as shown below.

2.2. Simple case: $\mathcal{F}$ is finite (and small)

Lemma: Assume
- $\mathcal{F}$ is finite
- $|\ell(f(x),y)-\ell(g(x),y)| \leq B \quad \forall f,g\in\mathcal{F}, \forall x,y$

Then there exists a randomized algorithm that chooses the $\hat{y}_t$ in such a way that

(or w.h.p.)

$$\forall (x_t,y_t)_{1\leq t\leq T} \in (\mathcal{X}\times\mathcal{Y})^T, \quad \mathbb{E}\left[R_T(\mathcal{F})\right] \leq B\sqrt{\frac{T}{2}\log|\mathcal{F}|}$$

- Also true if the elements of $\mathcal{F}$ are themselves randomized algorithms.
- A famous algorithm doing so is the "Exponential Weights" algorithm (or "Hedge"): $\hat{y}_t \sim \sum_{f\in\mathcal{F}} p_t(f)\,\delta_{f(x_t)}$

where
$$p_t(f) = \frac{\exp\left(-\eta\sum_{s=1}^{t-1}\ell(f(x_s),y_s)\right)}{\sum_{g\in\mathcal{F}}\exp\left(-\eta\sum_{s=1}^{t-1}\ell(g(x_s),y_s)\right)}$$

Proof also uses subgaussianity (of $\ell(\hat{y}_t,y)$ given the past).

- The lemma of Section 1 can also be used directly by reducing the learning problem to the computation of $\mathbb{E}\left[\max_f X_f\right]$ for some well-chosen $(X_f)_{f\in\mathcal{F}}$.

[see Rakhlin, Sridharan, and Tewari, JMLR 2015]

2.3. Nonparametric case: $\mathcal{F}$ is large

$(\mathcal{Y}, d) \rightsquigarrow (\mathcal{F}, d_\infty)$ where $d_\infty(f,g) := \sup_{x\in\mathcal{X}} d(f(x), g(x))$

$\mathcal{N}_\infty(\mathcal{F}, \varepsilon) := \mathcal{N}(\mathcal{F}, d_\infty, \varepsilon)$ = minimal number of d_∞-balls to cover $\mathcal{F}$ at scale ε.

Proposition: Assume that
- $\forall y \in \mathcal{Y}$, $\ell(\cdot, y)$ is L-Lipschitz
- $\mathcal{N}_\infty(\mathcal{F}, \varepsilon) < +\infty \quad \forall \varepsilon > 0$

Let $M \geq 1$. There exists a (computationally horrible) randomized algorithm such that, setting $B := \sup_{f,g\in\mathcal{F}} d_\infty(f,g)$,

$$\forall (x_t, y_t)_{1\le t\le T}, \quad \mathbb{E}\left[R_T(\mathcal{F})\right] \leq 6L\sqrt{2T}\int_{B2^{-M-1}}^{\frac{B}{2}} \sqrt{\log \mathcal{N}_\infty(\mathcal{F},\varepsilon)}\, d\varepsilon + TLB2^{-M}$$

Algo: uses chaining in its construction

Partially inspired from Cesa-Bianchi and Lugosi (1999) that was specific to $\ell(y,y') = |y-y'|$.

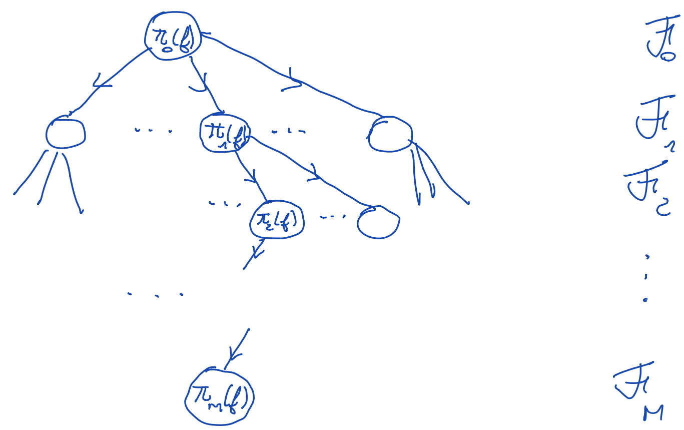
- Hierarchical discretization of $(\mathcal{F}, d_\infty)$:

$\mathcal{F}_0 = \{f_0\}$

$\mathcal{F}_m$ = smallest discretization of $\mathcal{F}$ at scale $B/2^m$, $m \geq 1$ with cardinality $N_m := \mathcal{N}(\mathcal{F}, d_\infty, B2^{-m})$

Level-m approximation: $\pi_m(f) \in \operatorname{argmin}_{g\in\mathcal{F}_m} d_\infty(g, f)$

We build a Directed Acyclic Graph (NB: we could also build a tree), where level-m nodes are labelled with elements of $\mathcal{F}_m$, and where we connect $\pi_m(f)$ with $\pi_{m+1}(f)$ for all $f \in \mathcal{F}$ and $m = 0, \dots, M-1$



→ At time t, each leaf $f \in \mathcal{F}_M$ recommends to play $f(x_t)$

→ On each internal node $f_m \in \mathcal{F}_m$, we place an instance $A_m(f_m)$ of the exponential weighting algorithm that learns the best of its children.

→ We play according to the algorithm sitting on the root.

Proof of regret bound:

Fix $f \in \mathcal{F}$. The regret against f can be rewritten as a sum of M regrets along the path $A_0(\pi_0(f)) \to A_1(\pi_1(f)) \to \dots \to A_M(\pi_M(f))$

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \sum_{t=1}^{T} \ell(f(x_t), y_t)\right]$$

$$\leq \sum_{m=0}^{M-1} \mathbb{E}\left[\sum_{t=1}^{T} \ell\big(A_{m,t}(\pi_m(f)), y_t\big) - \sum_{t=1}^{T} \ell\big(A_{m+1,t}(\pi_{m+1}(f)), y_t\big)\right] + TLB2^{-M}$$

$$\leq \sum_{m=0}^{M-1} B_m \sqrt{\frac{T}{2} \log N_{m+1}} + TLB2^{-M}$$

where B_m = range of the losses of children of level-m nodes $\leq L \times 2 \times 3B2^{-(m+1)}$

and $\log N_{m+1} = \log |\mathcal{F}_{m+1}| = \log \mathcal{N}_\infty(\mathcal{F}, B2^{-m-1})$

$$\leq 6L\sqrt{2T} \sum_{m=0}^{M-1} B2^{-(m+2)} \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, B2^{-(m+1)})} + TLB2^{-M}$$

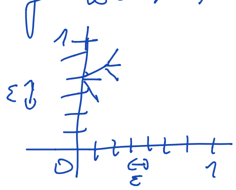
$$\leq 6L\sqrt{2T} \int_{\frac{B}{2^{M+1}}}^{\frac{B}{2}} \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}\, d\varepsilon + TLB2^{-M} \qquad \blacksquare$$

2.4. Example: $\mathcal{F} \equiv$ Lipschitz functions

$\mathcal{X} = [0,1]^d$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{F} = \left\{ f : [0,1]^d \to \mathbb{R} : |f(x) - f(x')| \leq \underset{\text{any norm}}{\|x - x'\|} \quad \forall x, y \in [0,1]^d \right\}$$

$$\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^d$$



$\frac{1}{\varepsilon} \times 3^{1/\varepsilon}$ functions $\rightsquigarrow \log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \asymp \frac{1}{\varepsilon}$ when $d = 1$.

Plugging $\log \mathcal{N}_\infty(\mathcal{F},\varepsilon) \simeq \left(\frac{1}{\varepsilon}\right)^d$ into the regret bound and taking $M = \left\lceil \frac{1}{d}\log_2 T \right\rceil$ (so that $2^{-M} \approx T^{-1/d}$), we get:

$$\mathbb{E}\left[R_T\right] \lesssim L\sqrt{T}\int_{BT^{-1/d}}^{B} \sqrt{\left(\frac{1}{\varepsilon}\right)^d}\, d\varepsilon + T^{1-1/d} \approx T^{1-1/d}$$

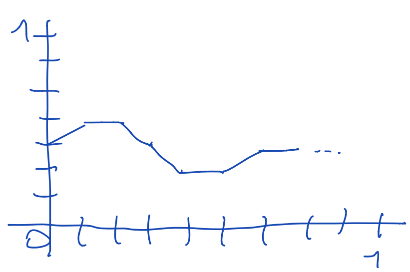
which matches the lower bound of Hazan and Megiddo (2007).

__NB__: faster rates are possible under stronger assumptions on $\ell(\cdot,y)$
(e.g., faster rates with square loss or any exp-concave loss)

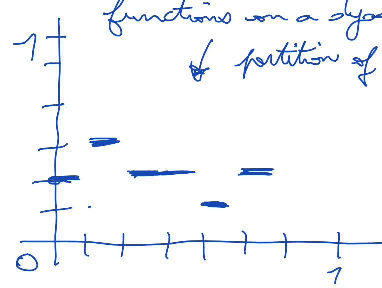
Efficient variant of the algorithm (when $\mathcal{F}$ = set of bounded and Lipschitz functions on $[0,1]^d$)

- The previous "algo" was super inefficient: requires $\exp(\mathrm{poly}(T))$ updates at any round $t \geq 1$.
- Fortunately, there exist coverings $\mathcal{F}_m$ that are much more managable from a computational viewpoint, while being almost optimal (up to log factors) from a statistical viewpoint.

Main idea in dimension $d=1$:



replaced with



piecewise-constant functions on a dyadic partition of $[0,1]^d$

optimal covering but computationally inefficient

additional log factor in regret but polynomial time algo.
independent of dimension d.