Talk at CREST : October 5, 2015

Title : A chaining algorithm for online nonparametric regression

→ Joint work with Pierre Gaillard

→ Paper: COLT 2015

I/ Setting : online (nonparametric) regression with individual sequences (7')

Online protocol : at each round $t \in \mathbb{N}^*$,

- get $x_t \in \mathcal{X}$
- predict $\hat{y}_t \in \mathbb{R}$
- observe $y_t \in \mathbb{R}$ ⇝ loss : $(y_t - \hat{y}_t)^2$

Goal : minimize the regret ($\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$) :

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2 =: \mathrm{Reg}_T(\mathcal{F})$$

Individual sequences : we want guarantees of the form

$$\sup_{(x_t)_{t\geq 1},\ (y_t)_{t\geq 1} : |y_t| \leq B} \mathrm{Reg}_T(\mathcal{F}) \leq o(T)$$

↑ depends on the size of $\mathcal{F}$

Ex : • finite $\mathcal{F}$: $\sup_{(x_t),(y_t)} \mathrm{Reg}_T(\mathcal{F}) \lesssim \log|\mathcal{F}|$

• $\mathcal{F} := \left\{ \sum_{j=1}^{d} u_j \varphi_j : u \in \Delta_d \right\}$: $\sup \mathrm{Reg}_T(\mathcal{F}) \lesssim d \log T$

• $\mathcal{F} := \{ f : [0,1] \to [-B,B]$ 1-Lip $\}$: $\sup \mathrm{Reg}_T(\mathcal{F}) \lesssim T^{1/3}$.

II/ Simplest case: finite $\mathcal{F}$ (6')

Algorithm: Exponentially Weighted Average Forecaster (EWA)

$\forall t \geq 1$

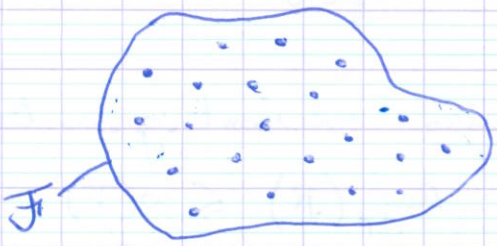
- $\hat{w}_{t,j} = \dfrac{\exp\left(-\eta \sum_{s=1}^{t-1} (y_s - f_j(x_s))^2\right)}{Z_t}$, $1 \leq j \leq |\mathcal{F}|$
- $\hat{y}_t := \sum_{j=1}^{|\mathcal{F}|} \hat{w}_{t,j} f_j(x_t)$

Thm (Kivinen and Warmuth 1999)

$\mathrm{Reg}_T(\mathcal{F}) \leq 8B^2 \log|\mathcal{F}|$ for all (x_t) and (y_t) s.t. $|y_t| \leq B$ and $|f_j(x_t)| \leq B$, if $\eta := \frac{1}{8B^2}$.

III/ What if $\mathcal{F}$ is (uncountably) infinite? (20')

1) Natural idea: discretize



$\mathcal{F}^{(\varepsilon)}$: finite set s.t. $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{F}^{(\varepsilon)}$: $\|f-g\|_\infty \leq \varepsilon$.

(we assume $\|f\|_\infty < B$ $\forall f \in \mathcal{F}$, and actually: $\mathcal{F}$ uniformly bounded).

Then, apply EWA to points in $\mathcal{F}^{(\varepsilon)}$:

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \min_{f \in \mathcal{F}^{(\varepsilon)}} \sum_{t=1}^{T} (y_t - f(x_t))^2 + 8B^2 \log|\mathcal{F}^{(\varepsilon)}|$$

$$\leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} (y_t - f(x_t))^2 + T\varepsilon^2 + 2T \cdot 2B\varepsilon + 8B^2 \log|\mathcal{F}^{(\varepsilon)}|$$

$$\rightsquigarrow \mathrm{Reg}_T(\mathcal{F}) \lesssim T\varepsilon + \log |\mathcal{F}^{(\varepsilon)}| \qquad (*)$$

[introduce metric entropy here]

Ex: $\mathcal{F} := \left\{ \sum_{j=1}^{d} u_j \varphi_j : u \in \Delta_d \right\} \rightsquigarrow \log |\mathcal{F}^{(\varepsilon)}| \lesssim d \log \frac{1}{\varepsilon}$

$$\mathrm{Reg}_T(\mathcal{F}) \lesssim T\varepsilon + d \log \frac{1}{\varepsilon}$$

$$\lesssim \underbrace{d \log T}_{\text{optimal!}} \quad \text{for } \varepsilon \approx \frac{1}{T}$$

2) Suboptimal bound if $\mathcal{F}$ is too large

Assume that $\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \simeq \left(\frac{1}{\varepsilon}\right)^p$ (instead of $d \log \frac{1}{\varepsilon}$)

- $\mathcal{N}_\infty(\mathcal{F}, \varepsilon)$: smallest possible $|\mathcal{F}^{(\varepsilon)}|$ = "metric entropy"
- $\left(\frac{1}{\varepsilon}\right)^p$: large! "nonparametric"

$(*)$ yields: $\mathrm{Reg}_T(\mathcal{F}) \lesssim T\varepsilon + \left(\frac{1}{\varepsilon}\right)^p$

$$\lesssim \underbrace{T^{\frac{p}{p+1}}}_{\text{suboptimal!}} \quad \text{for } \varepsilon \approx T^{-\frac{1}{p+1}}$$

Indeed:

Rakhlin and Sridharan (COLT'14) proved that $\mathrm{Reg}_T(\mathcal{F})$ can be as small as:

Dudley-type regret bound

$$\boxed{\mathrm{Reg}_T(\mathcal{F}) \lesssim \log \mathcal{N}_\infty(\mathcal{F}, \gamma) + \sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}\, d\varepsilon}$$

$$\lesssim \gamma^{-p} + \sqrt{T} \int_0^\gamma \varepsilon^{-p/2}\, d\varepsilon$$

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

$$\lesssim \gamma^{-p} + \sqrt{T} \cdot \gamma^{1-p/2} \qquad (\text{if } p < 2)$$

$$\lesssim \underbrace{T^{\frac{p}{p+2}}}_{\text{optimal!}} \quad \text{for } \gamma \approx T^{-\frac{1}{p+2}} \qquad (\ll T^{\frac{p}{p+1}})$$

(if $p \in [0,2)$)

($T^{1-1/p}$ if $p > 2$, with a slightly different Dudley bound)

Corollary : $\mathcal{F} := \{ f : [0,1] \to \mathbb{R} \ : \ \|f^{(k)}\|_\infty \le B \ \forall k = 0, \dots, p,$
$|f^{(p)}(x) - f^{(p)}(y)| \le L |x-y|^\alpha \}$

$p \in \mathbb{N}$, $\alpha \in (0,1]$.

Assume $\beta := p + \alpha > 1/2$

$\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^{1/\beta}$ $\rightsquigarrow$ $p \simeq \frac{1}{\beta}$

$\rightsquigarrow$ $\mathrm{Reg}_T(\mathcal{F}) \lesssim T^{\frac{p}{p+2}} = T^{\frac{1/\beta}{1/\beta+2}} = T^{\frac{1}{2\beta+1}}$

$\rightsquigarrow$ $\frac{\mathrm{Reg}_T(\mathcal{F})}{T} \lesssim T^{-\frac{2\beta}{2\beta+1}}$ as for iid data!

(or $T^{-\beta}$ if $\beta \le 2$, as for iid data)

BUT: Rakhlin et. al (2014): non-constructive regret bound (no explicit algo).

Our contribution : explicit algo. (with metric entropy instead of sequential entropy)

IV / Linearizing the square loss on help boldly (10')

EWA may be suboptimal ; instead : EG:

Algo : Exponentiated Gradient (EG)
* loss functions : $\ell_t(u)$, $u \in \Delta_N$, convex and diff.
* For all $t \ge 1$,

$$\hat{u}_{t,i} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \partial_{\hat{u}_{s,i}} \ell_s(\hat{u}_s)\right)}{Z_t}, \quad 1 \le i \le N$$

Regret bound (KW'93, CB'99)

Assume ℓ_t convex, diff, and $\|\nabla \ell_t\|_\infty \le G$

Then:

$$\sum_{t=1}^T \ell_t(\hat{u}_t) \le \min_{u \in \Delta_N} \sum_{t=1}^T \ell_t(u) + G\sqrt{2T \ln N}$$

for $\eta = \frac{1}{G}\sqrt{\frac{2 \log N}{T}}$

Application: $\ell_t(u) := \left(y_t - f_0(x_t) - \sum_{j=1}^N u_j g_j(x_t)\right)^2$, $u \in \Delta_N$

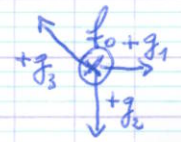
$$\sum_{t=1}^T \underbrace{\left(y_t - f_0(x_t) - \sum_{j=1}^N \hat{u}_{t,j} g_j(x_t)\right)^2}_{=: \hat{y}_t} \leq \min_{1\leq j\leq N} \sum_{t=1}^T \left(y_t - f_0(x_t) - g_j(x_t)\right)^2 + G\sqrt{2T \log N}$$

$\leq B \max_j \|g_j\|_\infty$

NB: $G\sqrt{2T\log N} \ll B^2 \log N \iff G \ll B^2 \sqrt{\frac{\log N}{T}}$

EG "better" than EWA if $\max_j \|g_j\|_\infty \ll B\sqrt{\frac{\log N}{T}}$

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



(10')

V/ Main algo: the Chaining Exponentially Weighted Average forecaster

Chaining technique: approximate any $f \in \mathcal{F}$ via $\pi_0(f) \in \mathcal{F}^{(0)}$, $\pi_1(f) \in \mathcal{F}^{(1)}, \dots$

$$\sup_f \|\pi_k(f) - f\|_\infty \leq \frac{\gamma}{2^k}, \quad |\mathcal{F}^{(k)}| = \mathcal{N}_\infty(\mathcal{F}, \gamma/2^k)$$

Rewrite:

$$\inf_{f\in\mathcal{F}} \sum_{t=1}^T (y_t - f(x_t))^2 = \inf_{f\in\mathcal{F}} \sum_{t=1}^T \left(y_t - \underbrace{\pi_0(f)}_{\in \mathcal{F}^{(0)}}(x_t) - \sum_{k=1}^{+\infty} \left(\pi_{k+1}(f) - \pi_k(f)\right)(x_t)\right)^2$$

Algo: two levels of aggregation in parallel:

- low-scale aggregation: new multi-scale version of EG to be competitive with all increments $\pi_{k+1}(f) - \pi_k(f)$, $k \geq 1$, simultaneously $\rightsquigarrow$ one instance around each $\pi_0(f) \in \mathcal{F}^{(0)}$.

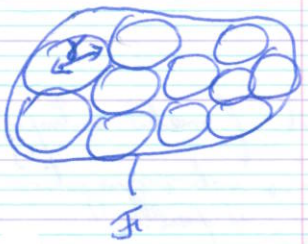
  $\rightsquigarrow$ regret contribution $\approx \sum_{k=1}^{+\infty} \frac{\gamma B}{2^k} \sqrt{T \log \mathcal{N}_\infty(\mathcal{F}, \gamma/2^k)}$

  $$\approx B\sqrt{T} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathcal{F}, \varepsilon)}\, d\varepsilon$$

  $\rightsquigarrow$ a Dudley type regret-bound

- high-scale aggregation: EWA to aggregate the resulting $\mathcal{N}_\infty(\mathcal{F}, \gamma)$ forecasters

  $\rightsquigarrow$ regret contribution $\approx B^2 \log \mathcal{N}_\infty(\mathcal{F}, \gamma)$.



References / previous work (10')

Cesa-Bianchi and Lugosi (2001) : two-side aggregation
"Worst-case bounds for the log loss of predictors" log loss, online setting
chaining in the analysis

Rakhlin et al. (to appear) : square loss, batch setting
"Empirical entropy, minimax regret and minimax risk" two-side aggregation
chaining in the analysis

Rakhlin et al. (COLT'14) : square loss, online
"Online nonparametric regression" chaining in the analysis
(but stronger notion of entropy)

Cesa-Bianchi (1999) : absolute loss, online
"On prediction of individual seq" chaining in the algo.

Second contribution : efficient algo for Hölder classes

function class	regret bound	time complexity	space complexity
Lipschitz on $[0,1]$	$T^{1/3} \log(T)$	$T^{4/3} \log T$	$T^{4/3} \log T$
β-Hölder on $[0,1]$	$T^{\frac{1}{2\beta+1}} \log^{3/2} T$	poly(T)	poly(T)

Main idea : dyadic discretization to only update few weights at each round.