Nonparametric Learning, Invariance, and Sketching Techniques for Online Algorithms

Nicolò Cesa-Bianchi

Università degli Studi di Milano

Joint work with:

- Rocco de Rosa (Roma) and Francesco Orabona (Yahoo NYC) — nonparametric learning
- Alekh Agarwal (MSR-NYC), John Langford (MSR-NYC) and Haipeng Luo (Princeton) — invariance and sketching

Online randomized classification in metric spaces

Setting

- Metric space $(\mathcal{X}, \rho)$
- Deterministic source $\boldsymbol{\sigma} = \big((x_1, y_1), (x_2, y_2), \dots\big)$
- Where $(x_t, y_t) \in \mathcal{X} \times \{0, 1\}$

For $t = 1, 2, \dots$

1. Receive next instance $x_t \in \mathcal{X}$
2. Output randomized prediction $\widehat{Y}_t \in \{0, 1\}$ where $\mathbb{P}\big(\widehat{Y}_t = 1\big) = p_t$
3. Observe true class label $y_t \in \{0, 1\}$
4. Pay loss $\mathbb{I}\{\widehat{Y}_t \neq y_t\}$ with probability $\mathbb{P}\big(\widehat{Y}_t \neq y_t\big) = \big|p_t - y_t\big|$

Nonparametric online learning

A nonparametric class

- $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz if there exists $0 \leqslant L < \infty$ such that
$$\left|f(x) - f(x')\right| \leqslant L\,\rho(x, x') \qquad x, x' \in \mathcal{X}$$
- $\mathrm{Lip} = \mathrm{Lip}(\mathcal{X})$ is the class of $[0,1]$-valued Lipschitz functions on $\mathcal{X}$
- How fast can we learn the best $f \in \mathrm{Lip}$ for an arbitrary σ?

Expected regret

$$R_T(\sigma) = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{f \in \mathrm{Lip}} \sum_{t=1}^{T} \ell_t\big(f(x_t)\big)$$

where $\ell_t(p) = \left|p - y_t\right|$

Doubling dimension of metric spaces

Doubling dimension

A metric space has doubling dimension d if $\Theta\left(\varepsilon^{-d}\right)$ nonoverlapping balls of radius ε can be packed in it

Nonparametric rates: k-NN

- Stochastic source: $X_1, X_2, \ldots \in \mathcal{X}$ i.i.d.
- $\mathbb{P}\left(Y_t = 1 \mid X_t = x_t\right) = f^*(x_t)$ for some $f^* \in \mathrm{Lip}$
- Then the excess expected mistake rate of k_T-NN w.r.t. $\mathbb{I}\{f^*(x) \geqslant \frac{1}{2}\}$ grows at optimal rate bounded by

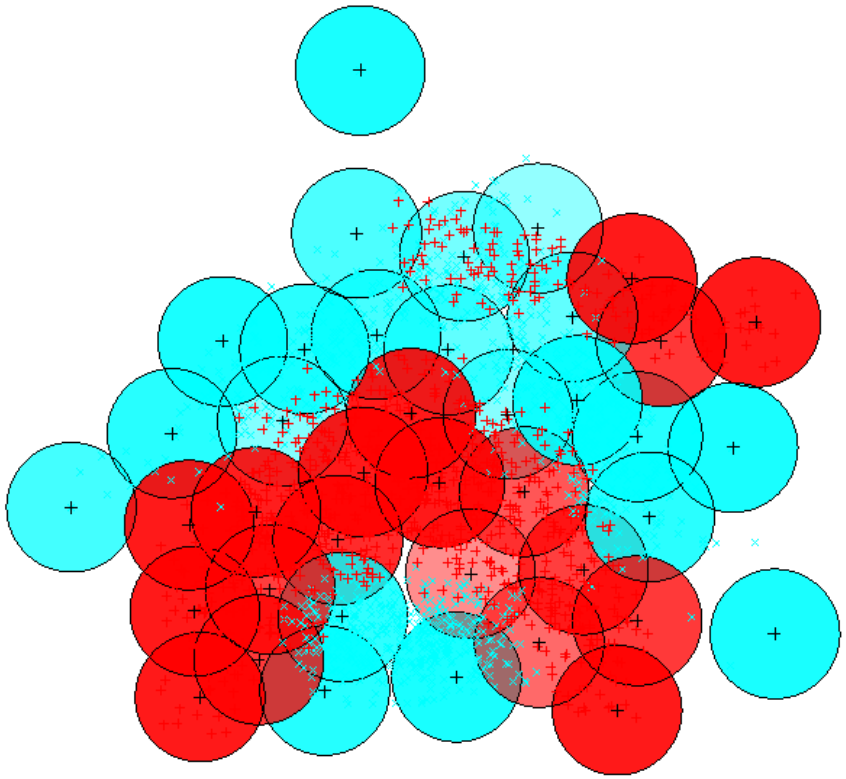
$$\mathcal{O}\left(T^{\frac{d+1}{d+2}}\right)$$

where d is the doubling dimension of $\mathcal{X}$

- Can we match this rate for any deterministic source?

A ball-covering approach [Hazan and Megiddo, 2007]

- The instance space is incrementally covered with "half-overlapping" balls
- Each ball has a local classifier sitting in it
- For prediction, local classifier is selected with a NN query
- For updating, ball radius determines relevant points
- Radius of each ball shrinks with time



Basic algorithm

Create a default cover $\mathcal{C}$ containing a single ball with a default local classifier

For $t = 1, 2, \ldots$

1. Receive next stream instance x_t and compute closest center $x_t^* \in \mathcal{C}$
2. Predict using local classifier sitting at x_t^*
3. Observe true class label y_t and update local classifier
4. If x_t is not contained in ball centered at x_t^* then add ball $B(x_t, \varepsilon_t)$ to $\mathcal{C}$
5. Shrink radii ε_t of all balls in $\mathcal{C}$ according to law $\varepsilon_t = t^{-\frac{1}{d+2}}$

We use majority voters as local classifiers

Analysis (part I)

- Fix any Lipschitz $f : \mathcal{X} \to [0,1]$ with Lipschitz constant L_f
- $p_s^{\text{opt}} \in \{0,1\}$ is best prediction for all (x_t, y_t) such that $x_t \in B(x_s, \varepsilon_t)$

$$
\begin{aligned}
R_T(\sigma) &= \sum_{x_s \in \mathcal{C}} \sum_{t\,:\,x_s = x_t^*} \Big(\ell_t(p_t) - \ell_t\big(f(x_t)\big) \Big) \\
&\leqslant \sum_{x_s \in \mathcal{C}} \sum_{t\,:\,x_s = x_t^*} \Big(\ell_t(p_t) - \ell_t(p_s^{\text{opt}}) \Big) + \sum_{x_s \in \mathcal{C}} \sum_{t\,:\,x_s = x_t^*} \Big(\ell_t\big(f(x_s)\big) - \ell_t\big(f(x_t)\big) \Big) \\
&\leqslant \sum_{x_s \in \mathcal{C}} \sqrt{T_s} + \sum_{x_s \in \mathcal{C}} \sum_{t\,:\,x_s = x_t^*} L_f \varepsilon_t \\
&\leqslant \sqrt{|\mathcal{C}|\, T} + L_f \sum_{t=1}^{T} t^{-\frac{1}{d+2}} \leqslant \sqrt{|\mathcal{C}|\, T} + L_f T^{\frac{d+1}{d+2}}
\end{aligned}
$$

Analysis (part II)

Since $\varepsilon_T = T^{-\frac{1}{d+2}}$, the packing number bound gives $|\mathcal{C}| \leqslant C_{\mathcal{X}} T^{\frac{d}{d+2}}$

Therefore

$$\begin{aligned} R_T(\boldsymbol{\sigma}) &\leqslant \sqrt{|\mathcal{C}|\, T} + L_f T^{\frac{d+1}{d+2}} \\ &\leqslant C_{\mathcal{X}} \sqrt{T^{\frac{d}{d+2}} T} + L_f T^{\frac{d+1}{d+2}} \\ &\leqslant (C_{\mathcal{X}} + L_f)\, T^{\frac{d+1}{d+2}} \end{aligned}$$

Regret bound

Theorem

For any deterministic source $\boldsymbol{\sigma} = \big((x_1, y_1), (x_2, y_2), \dots\big)$

$$\sum_{t=1}^{T} \ell_t(p_t) \leqslant \min_{f \in \mathrm{Lip}} \left(\sum_{t=1}^{T} \ell_t\big(f(x_t)\big) + (C_{\mathcal{X}} + L_f)\, T^{\frac{d+1}{d+2}} \right)$$

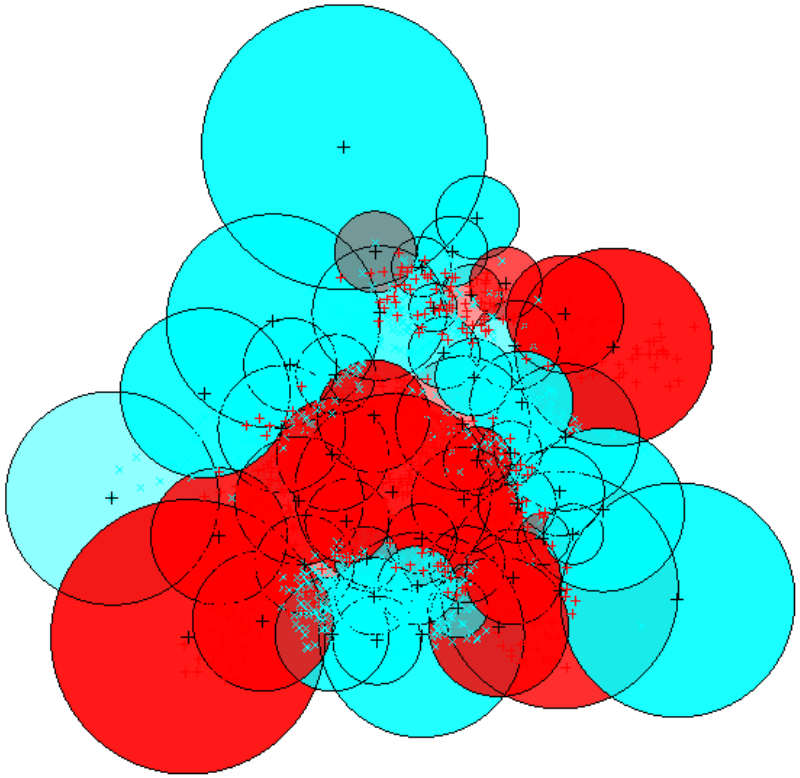
where d *is the doubling dimension of* $\mathcal{X}$

Remarks:

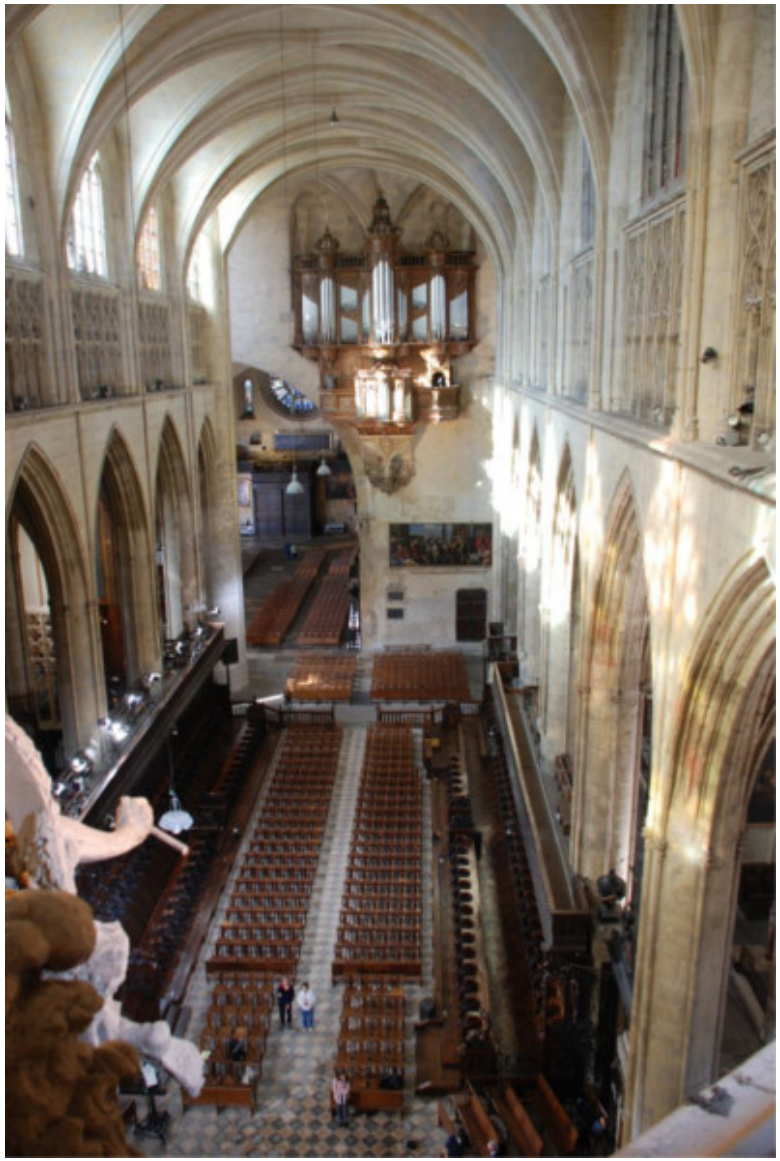
- Oracle inequality: best trade-off between fitness and complexity
- Doubling trick to learn the doubling dimension
- Easy to get parametric rate $\sqrt{T}$ against any f such that $L_f = 0$
- Scaling $L_f\, T^{\frac{d+1}{d+2}}$ possible, but not uniformly over Lip

Variant: radius shrinks with mistake rate

- $\varepsilon_t = m_t(s)^{-1/(d+2)}$ where $m_t(s)$ is number of mistakes of classifier sitting at $x_s \in \mathcal{C}$
- Cover adapts to local complexity of data source
- We lose "half-overlapping property" and packing argument is gone



Entr'Acte



Online linear prediction

Setting

- Data source $x_1, x_2, \ldots \in \mathbb{R}^d$
- Convex and differentiable loss sequence $f_1, f_2, \ldots$ on $\mathbb{R}$

For $t = 1, 2, \ldots$

1. Receive next instance $x_t \in \mathbb{R}^d$
2. Output linear prediction $p_t = \boldsymbol{w}_t^\top x_t$
3. Pay loss $\ell_t(\boldsymbol{w}_t) = f_t\left(\boldsymbol{w}_t^\top x_t\right)$
4. Observe gradient $g_t = \nabla \ell_t(\boldsymbol{w}_t) = f_t'(p_t) x_t$

Rates

Regret

$$R_T(w) = \sum_{t=1}^{T} \ell_t(w_t) - \sum_{t=1}^{T} \ell_t(w)$$

- We know that for arbitrary convex losses $R_T(w)$ scales with $\sqrt{T}$
- Dependence on w, $x_1, \ldots, x_T$, and $f_1, \ldots, f_T$?

First-order (Euclidean) methods

Online gradient descent [Zinkevich, 2003]

$w_{t+1} = w_t - \eta_t g_t$ followed by projection to ball of radius U

$$R_T(w) \leqslant (G_C U X)\sqrt{T}$$

for all

- w with $\|w\| \leqslant U$
- $x_1, \ldots, x_T$ with $\max_t \|x_t\| \leqslant X$
- $f_1, \ldots, f_T$ with $\max_t \max_{p\,:\,|p| \leqslant C} \left|f_t'(p)\right| \leqslant G_C$

Invariance to linear transformations

A natural comparison class

$$\mathcal{K}_C = \left\{ \boldsymbol{w} \in \mathbb{R}^d \,:\, \left| \boldsymbol{w}^\top \boldsymbol{x}_t \right| \leqslant C,\ t = 1, \ldots, T \right\}$$

- Invariant to any (invertible) linear transformation of data sequence
- More general than $\|\boldsymbol{x}_t\| \leqslant X$ and $\|\boldsymbol{w}\| \leqslant U$
- Can we design an invariant algorithm whose regret scales with C?

Second-order algorithms in a transductive setting

For any fixed $x_1, \dots, x_T$

$$w'_t = w_t - A^{-1} g_t \qquad w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{K}_C} \left\| w - w'_t \right\|_A$$

For any $w \in \mathcal{K}_C$

$$R_T(w) \leqslant \|w\|_A^2 + G_C^2 \sum_{t=1}^{T} \|x_t\|_{A^{-1}}^2$$

Now pick the best A for the worst-case $w \in \mathcal{K}_C$

Optimizing the bound

For any fixed $x_1, \ldots, x_T$

The solution to $\inf_{A \succ 0} \sup_{w \in \mathcal{K}'_C} \left(\|w\|_A^2 + G_C^2 \sum_{t=1}^{T} \|x_t\|_{A^{-1}}^2 \right)$

in the larger set $\mathcal{K}'_C = \left\{ w \in \mathbb{R}^d : \sum_{t=1}^{T} (w^\top x_t)^2 \leqslant TC^2 \right\}$

is $A = \frac{G_C}{C} \left(\sqrt{\frac{d}{T}} \right) \sum_{t=1}^{T} x_t x_t^\top$

This gives the bound $R_T(w) \leqslant C G_C \sqrt{dT}$

Can we get a similar bound in a non-transductive setting?

A nearly invariant online algorithm

Online Newton Step with projection

$$w'_t = w_t - A_t^{-1} g_t$$
$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{K}_{C,t+1}} \left\| w - w'_t \right\|_{A_t}$$
$$\text{where} \quad A_t = \alpha I + \sum_{s=1}^{t} \eta_s \, g_s g_s^\top$$
$$\text{and} \quad \mathcal{K}_{C,t+1} = \left\{ w \in \mathbb{R}^d \,:\, \left| w^\top x_{t+1} \right| \leqslant C \right\}$$

Regret bound

$$R_T(w) \leqslant \alpha \|w\|^2 + \left(\ln \frac{T}{\alpha} \right) C G_C \sqrt{dT} \quad \text{for all } w \in \mathcal{K}_{C,T}$$

A truly invariant online algorithm

Online Newton Step with projection and pseudoinverse

$$w'_t = w_t - A_t^+ g_t$$
$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{K}_{C,t+1}} \left\| w - w'_t \right\|_{A_t}$$
$$\text{where} \quad A_t = \sum_{s=1}^{t} \eta_s \, g_s g_s^\top$$
$$\text{and} \quad \mathcal{K}_{C,t+1} = \left\{ w \in \mathbb{R}^d \, : \, \left| w^\top x_{t+1} \right| \leqslant C \right\}$$

Regret bound

$$R_T(w) \leqslant C G_C \left(\sqrt{dT} + \left(R^2 \ln T \right) \sqrt{\frac{T}{d}} \right)$$

for all $w \in \mathcal{K}_{C,T}$ where R is the rank of the data matrix

Sketching

- Second-order algorithms require updating $A_t^{-1} = \alpha I + G_t G_t^\top$ where G_t is $d \times t$
- This is done in time $\Theta(d^2)$
- Want faster algorithms using $B_t = \alpha I + S_t S_t^\top$ where S_t is $d \times m$ for $m \ll d$
- Since $B_t^{-1} = \frac{1}{\alpha}\left(I + S_t \underbrace{\left(\alpha I + S_t S_t^\top\right)^{-1}}_{H} S_t^\top\right)$ where H is $m \times m$
- We can update B_t^{-1} in time $\mathcal{O}(d^2 + md) = \mathcal{O}(md)$

Sketching

Sketched Online Newton Step with projection

$$w'_t = w_t - A_t^{-1} g_t$$
$$w_{t+1} = \operatorname*{argmin}_{w \in \mathcal{K}_{C,t+1}} \left\| w - w'_t \right\|_{A_t}$$
$$\text{Replace} \quad A_t = \alpha I + \sum_{s=1}^{t} g_s g_s^\top$$
$$\text{with} \quad B_t = \alpha I + S_t S_t^\top$$

where S_t is $d \times m$

Regret bound for sketched ONS

$$R_T(w) \leqslant \alpha \|w\|^2 + \sum_{t=1}^{T} \|g_t\|^2_{B_t^{-1}} + \sum_{t=1}^{T} \|w_t - w\|^2_{B_t - B_{t-1}}$$

Gaussian Random Projection

$S_t = S_{t-1} + g_t Z_t^\top$ where Z_t are m i.i.d. draws from $\mathcal{N}\left(0, m^{-1/2}\right)$

1. $(1-\varepsilon)B_t \preceq A_t \preceq (1+\varepsilon)B_t$ w.h.p. for $m \geqslant \dfrac{R + \ln T}{\varepsilon^2}$
2. $\mathbb{E}_t\left[B_t - B_{t-1}\right] = g_t g_t^\top$

This leads to the same regret bound as before (but only in expectation)

$$R_T(w) \leqslant \alpha \|w\|^2 + \sum_{t=1}^{T} \underbrace{\|g_t\|^2_{B_t^{-1}}}_{B_t^{-1} \preceq \frac{1}{1-\varepsilon} A_t^{-1}} + \sum_{t=1}^{T} \underbrace{\|w_t - w\|^2_{B_t - B_{t-1}}}_{= g_t g_t^\top}$$

$\mathcal{O}(md)$ space, $\mathcal{O}(md)$ time per round

Frequent Directions [Ghashami, Liberty, Phillips, and Woodruff, 2014]

Item frequency approximation for streams

1. Keep m counters for d types of objects, $m \ll d$
2. If type of next incoming object has a counter, increase it, otherwise pick a counter at zero and set it to 1
3. If there are no counters at zero, then decrease all counters until the smallest hits zero

Matrix approximations

1. Keep m orthogonal vectors in $\mathbb{R}^d$, $m \ll d$
2. Add the next incoming vector and make the set orthogonal again via SVD
3. Shrink all vectors until the shortest hits zero

$\mathcal{O}(md)$ space, $\mathcal{O}(md)$ time per round

Frequent Directions

Regret bound

$$R_T(\boldsymbol{w}) \leqslant \alpha \|\boldsymbol{w}\|^2 + \left(\left(\ln \frac{T}{\alpha} \right) C G_C + \frac{\Gamma}{\alpha} \right) m \sqrt{\frac{T}{d}}$$

where $\Gamma_T = \sum_{i=m+1}^{d} \lambda_i \left(S_T S_T^\top \right)$ is the eigenvalue tail of the sketch

- When $m = d$ we recover the standard bound
- When data sequence has rank at most m we win

Additional results

- Faster rates for exp-concave losses
- Sketching using Oja's algorithm
- Sparsity-preserving sketching
- Experiments running