Online Nonparametric Regression

Sasha Rakhlin and Karthik Sridharan

Department of Statistics, The Wharton School
University of Pennsylvania

June 14, 2014

Online Regression

individual sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$

At each time step $t = 1, \ldots, n$,

- x_t is revealed by Nature
- Prediction $\widehat{y}_t \in \mathcal{Y}$ is made by the forecaster
- $y_t \in \mathcal{Y}$ is revealed by Nature

Regret with respect to class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$:

$$\frac{1}{n} \sum_{t=1}^{n} (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} (f(x_t) - y_t)^2$$

Online Regression

individual sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$

At each time step $t = 1, \ldots, n$,

- x_t is revealed by Nature
- Prediction $\widehat{y}_t \in \mathcal{Y}$ is made by the forecaster
- $y_t \in \mathcal{Y}$ is revealed by Nature

Regret with respect to class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$:

$$V_n(\mathcal{F}) = \inf_{\text{Algo}} \sup_{\{(x_t, y_t)\}_{t=1}^n} \left\{ \frac{1}{n} \sum_{t=1}^n (\widehat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n (f(x_t) - y_t)^2 \right\}$$

Online Regression

What is known?

- Finite $\mathcal{F}$ – Foster '90
- $\mathcal{F}$ is a bounded subset of $\mathbb{R}^d$ – Vovk '98, '00; Azoury and Warmuth '01
- $\mathcal{F}$ is set of s-sparse vectors – Gerchinovitz '13
- $\mathcal{F}$ is a ball in Besov space – Vovk '06, but loose bounds. Vovk states open question.

Online Regression

What is known?

- Finite $\mathcal{F}$ – Foster '90
- $\mathcal{F}$ is a bounded subset of $\mathbb{R}^d$ – Vovk '98, '00; Azoury and Warmuth '01
- $\mathcal{F}$ is set of s-sparse vectors – Gerchinovitz '13
- $\mathcal{F}$ is a ball in Besov space – Vovk '06, but loose bounds. Vovk states open question.

Sad state of affairs as compared to Statistics and Statistical Learning!

Online Regression

What is known?

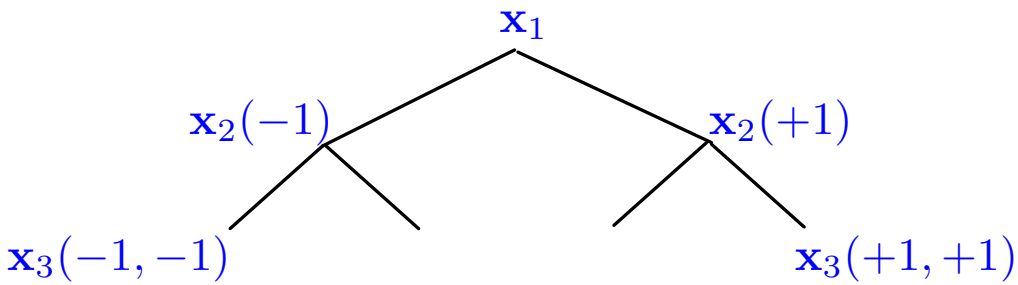
- Finite $\mathcal{F}$ – Foster '90
- $\mathcal{F}$ is a bounded subset of $\mathbb{R}^d$ – Vovk '98, '00; Azoury and Warmuth '01
- $\mathcal{F}$ is set of s-sparse vectors – Gerchinovitz '13
- $\mathcal{F}$ is a ball in Besov space – Vovk '06, but loose bounds. Vovk states open question.

Sad state of affairs as compared to Statistics and Statistical Learning!

Two claims:
- Behavior of regret $V_n(\mathcal{F})$ is characterized by "complexity" of $\mathcal{F}$
- There is a canonical learning algorithm

Consider a complete binary tree $\mathbf{x}$ of depth n labeled by elements of $\mathcal{X}$. Level $i \in \{1, \dots, n\}$:

$$\mathbf{x}_i : \{\pm 1\}^{i-1} \to \mathcal{X}$$



A *path* is a sequence $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \{\pm 1\}^n$.

- We write $\mathbf{x}_t(\epsilon)$ for $\mathbf{x}_t(\epsilon_{1:t-1})$.

Covering numbers

Definition (R., Sridharan, Tewari, 2010)**.**

A set V of $\mathbb{R}$-valued trees is an α-cover of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ on $\mathbf{x}$ if

$$\forall f \in \mathcal{F},\ \forall \epsilon \in \{\pm 1\}^n\ \exists \mathbf{v} \in V \ \text{ s.t. } \quad \left(\frac{1}{n}\sum_{t=1}^{n} \left|f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)\right|^2\right)^{\frac{1}{2}} \leq \alpha$$

Size of smallest cover is denoted by $\mathcal{N}_2(\alpha, \mathcal{F}, \mathbf{x})$.

Sequential entropy:

$$\sup_{\mathbf{x}} \log \mathcal{N}_2(\alpha, \mathcal{F}, \mathbf{x})$$

Given an $\mathcal{X}$-valued tree,

$$\mathscr{R}_n(\mathcal{F}; \mathbf{x}) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) \right]$$

is called *sequential Rademacher complexity* of $\mathcal{F}$ on $\mathbf{x}$.

(R., Sridharan, Tewari '10): Minimax regret for learning with *absolute loss* is given (to within constant 2) by sequential Rademacher complexity.

Leads to a martingale version of uniform laws of large numbers and sequential extension of classical complexity notions (covering numbers, scale-sensitive dimensions, etc).

Unfortunately, can only yield rates no better than $n^{-1/2}$.

Puzzle of faster rates for online learning

Statistical Learning: aggregation + localization

- In i.i.d. case, proper learning (e.g. ERM) can be suboptimal. So, empirical process story alone is not enough.
- (R., Sridharan, Tsybakov '13): aggregation (e.g. Exp. Weights) of proper learning algorithms on small data-dependent subsets.

Puzzle of faster rates for online learning

Statistical Learning: aggregation + localization

- In i.i.d. case, proper learning (e.g. ERM) can be suboptimal. So, empirical process story alone is not enough.
- (R., Sridharan, Tsybakov '13): aggregation (e.g. Exp. Weights) of proper learning algorithms on small data-dependent subsets.

Online Learning:

- Can't even define these subsets because data not available in advance.
- On the positive side, online protocol is inherently improper.
- Vovk outlined two very different algorithms for Besov spaces: based on uniform convexity and on metric entropy. What else is out there?
- Instead of looking for abstract algorithms, we start with value and later come back to get algorithms.

Handwaving argument

Regret against a single $f \in \mathcal{F}$:

$$(\hat{y}_t - y_t)^2 - (f(x_t) - y_t)^2$$

After applying minimax theorem, the best we can do is

$$(\mu_t - y_t)^2 - (f(x_t) - y_t)^2$$

where $\mu_t = \mathbb{E}[y_t]$.

Handwaving argument

Regret against a single $f \in \mathcal{F}$:

$$(\hat{y}_t - y_t)^2 - (f(x_t) - y_t)^2$$

After applying minimax theorem, the best we can do is

$$(\mu_t - y_t)^2 - (f(x_t) - y_t)^2 = 2(y_n - \mu_n)(f(x_n) - \mu_n) - (f(x_n) - \mu_n)^2$$

where $\mu_t = \mathbb{E}[y_t]$.

Theorem

For online regression with responses y_t in a bounded interval $[-1, 1]$:

$$V_n(\mathcal{F}) \leq \sup_{x,\mu} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n} 4\epsilon_t \Big(f(x_t(\epsilon)) - \mu_t(\epsilon) \Big) \right]$$

where x ranges over $\mathcal{X}$-valued trees, μ over $[-1, 1]$-valued trees of depth n.

- ⋆ *Lower bound is matching to within a constant.*

Theorem

For online regression with responses y_t *in a bounded interval* $[-1, 1]$*:*

$$V_n(\mathcal{F}) \leq \sup_{x,\mu} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n} 4\epsilon_t \Big(f(x_t(\epsilon)) - \mu_t(\epsilon) \Big) - \Big(f(x_t(\epsilon)) - \mu_t(\epsilon) \Big)^2 \right]$$

where x *ranges over* $\mathcal{X}$*-valued trees,* μ *over* $[-1, 1]$*-valued trees of depth* n*.*

- ⋆ *Lower bound is matching to within a constant.*

Theorem

For online regression with responses y_t in a bounded interval $[-1, 1]$:

$$V_n(\mathcal{F}) \le \sup_{x,\mu} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n} 4\epsilon_t \Big(f(x_t(\epsilon)) - \mu_t(\epsilon) \Big) - \Big(f(x_t(\epsilon)) - \mu_t(\epsilon) \Big)^2 \right]$$

where x ranges over $\mathcal{X}$-valued trees, μ over $[-1, 1]$-valued trees of depth n.

- *Lower bound is matching to within a constant.*

Lemma

For any $\mathcal{Z}$-valued tree z, class $\mathcal{G}$ of functions $\mathcal{Z} \to [-1, 1]$, and $\gamma \in (0, 1]$,

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^{n} 4\epsilon_t g(z_t(\epsilon)) - g(z_t(\epsilon))^2 \right] \le \frac{32 \log \mathcal{N}_2(\gamma, \mathcal{G}, z)}{n} + \inf_{\rho \in (0, \gamma]} \left\{ 4\rho + \frac{12}{\sqrt{n}} \int_\rho^\gamma \sqrt{\log \mathcal{N}_2(\delta, \mathcal{G}, z)}\, d\delta \right\}$$

Theorem

For $\mathcal{F}$ *with* $\log \mathcal{N}_2(\alpha, \mathcal{F}, n) \leq \alpha^{-p}$,

- $p > 2$: $V_n(\mathcal{F}) \leq C n^{-1/p}$
- $p \in (0, 2)$: $V_n(\mathcal{F}) \leq C n^{-2/(2+p)}$
- *Parametric case:* $V_n(\mathcal{F}) \leq C d n^{-1} \log(n)$
- *Finite set* $\mathcal{F}$: $V_n(\mathcal{F}) \leq C n^{-1} \log |\mathcal{F}|$

Lower bounds:

- $p \geq 2$: *for any* $\mathcal{F}$ *of uniformly bounded functions with a lower bound of* α^{-p} *on sequential entropy growth,* $V_n(\mathcal{F}) \geq \tilde{\Omega}(n^{-1/p})$
- $p \in (0, 2]$: *for any* $\mathcal{F}'$ *of uniformly bounded functions, there exists* $\mathcal{F}$ *with the same sequential entropy growth s.t.* $V_n(\mathcal{F}) \geq \tilde{\Omega}(n^{-2/(2+p)})$
- *There exists parametric class* $\mathcal{F}$, *s.t.* $V_n(\mathcal{F}) \geq \Omega(d n^{-1} \log(n))$

Remarks

For $p > 2$, curvature of loss does not help (consistent with iid result)

Online-to-batch implies a method with correct rates for distrib.-free *Statistical Learning*! (modulo differences between sequential and iid)

Online algorithm is using completely different techniques/language.

Purely empirical-process based story for improper learning. (can we do same for iid?)

Algorithms

Find a relaxation such that

$$\mathbf{Rel}_n\left(x_{1:n}, y_{1:n}\right) \geq -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \left(f(x_t) - y_t\right)^2$$

and for any $t \in [n]$ and any $x_t \in \mathcal{X}$

$$\inf_{\hat{y}_t \in [-B,B]} \sup_{y_t \in [-B,B]} \left\{\left(\hat{y}_t - y_t\right)^2 + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t-1}, y_t\right)\right\} \leq \mathbf{Rel}_n\left(x_{1:t-1}, y_{1:t-1}\right)$$

If $\left(\hat{y}_t - y_t\right)^2 + \mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right)$ is convex in y_t, then an algorithm is

$$\hat{y}_t = \mathrm{Clip}\left(\frac{\mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, B)\right) - \mathbf{Rel}_n\left(x_{1:t}, (y_{1:t-1}, -B)\right)}{4B}\right)$$

Canonical Algorithm

Lemma

The following relaxation is admissible:

$$\mathscr{R}_n(x_{1:t}, y_{1:t}) = \sup_{x,\mu} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left[\sum_{j=t+1}^{n} 4B\epsilon_j (f(x_j(\epsilon)) - \mu_j(\epsilon)) - (f(x_j(\epsilon)) - \mu_j(\epsilon))^2 - \sum_{j=1}^{t} (f(x_j) - y_j)^2 \right]$$

The prediction is given by

$$\hat{y}_t = (4B)^{-1} \left(\mathscr{R}_n(x_{1:t}, (y_{1:t-1}, B)) - \mathscr{R}_n(x_{1:t}, (y_{1:t-1}, -B)) \right)$$

This algorithm enjoys regret bound of offset Rademacher.

Example: Linear Regression

Corollary

For any $\lambda > 0$, $A_t = \sum_{j=1}^{t} z_j z_j^\top + \lambda I$ the following is an admissible relaxation

$$\mathbf{Rel}_n\left(x_{1:t}, y_{1:t}\right) = \left\| \sum_{j=1}^{t} y_j z_j \right\|_{A_t^{-1}}^2 + 4B^2 \log\left(\frac{(n/d)^d}{\Delta(A_t)} \right) - \sum_{j=1}^{t} y_j^2 \ .$$

It leads to the Vovk-Azoury-Warmuth forecaster:

$$\hat{y}_t = \mathrm{Clip}\left(x_t^\top \left(\sum_{j=1}^{t} x_j x_j^\top + \lambda I \right)^{-1} \left(\sum_{j=1}^{t-1} y_j x_j \right) \right)$$

and enjoys the regret bound

$$\frac{1}{n} \sum_{t=1}^{n} (\hat{y}_t - y_t)^2 \le \frac{1}{n} \sum_{t=1}^{n} (f^\top x_t - y_t)^2 + \frac{\lambda}{2n} \|f\|_2^2 + \frac{4dB^2 \log\left(\frac{n}{\lambda d}\right)}{n}$$

Statistical Estimation

Statistical Learning

Online Learning

Statistical Estimation

Statistical Learning

Online Learning

$$\log \mathcal{N}(\alpha, \mathcal{F}) \asymp \alpha^{-p}$$

$p \in (0,2)$	$n^{-\frac{2}{2+p}}$	$n^{-\frac{2}{2+p}}$	$n^{-\frac{2}{2+p}}$
$p \geq 2$	$n^{-\frac{2}{2+p}}$	$n^{-1/p}$	$n^{-1/p}$

Further Directions

- Extension to other losses (journal version soon)
- New online regression algorithms for interesting classes?
- Implications for statistical learning, especially in light of an algorithmic toolbox for improper learning?
- Understanding the phenomenon of no gaps between Statistical and Online Learning.