Bypassing the Impossibility of Online Learning Thresholds: Unbounded Losses and Transductive Priors

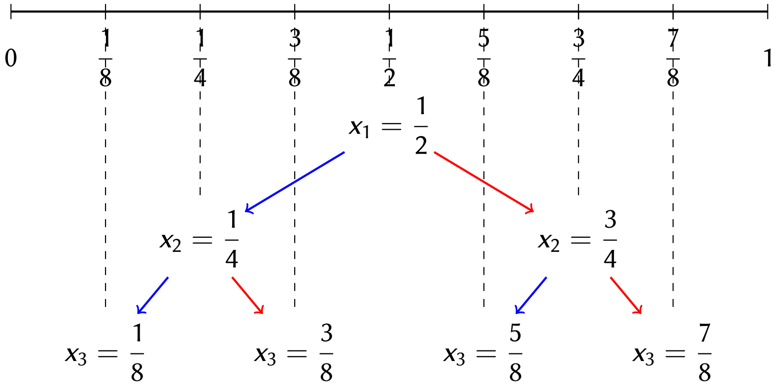
Nikita Zhivotovskiy[1]

[1]UC Berkeley, Department of Statistics

Based on the joint work with
Jian Qian and Sasha Rakhlin (MIT)

Online learning for thresholds is hard

Observe a sequence $x_1, \ldots, x_T$ of points in $[0, 1]$ labeled by a threshold as either $+1$ or -1.



This yields T mistakes after T rounds.

Thresholds $\mapsto$ classification with half-spaces (linear classification).

Question

How does the difficulty of online learning thresholds affect online learning with unbounded losses:

- *logistic regression loss* $-\log(\sigma(y\langle x, \theta\rangle))$,
- *hinge loss* $(1 - y\langle x, \theta\rangle)_+$,
- *regression with square loss* $(y - \langle x, \theta\rangle)^2$*?*

Sequential linear regression

We observe a sequence $(x_t, y_t)_{t=1}^T$, with $x_t \in \mathbb{R}^d$, $y_t \in \mathbb{R}$. At round t we observe x_t and $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and want to predict y_t.

$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \left(\sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + (\langle x_t, \theta \rangle)^2 + \lambda \|\theta\|^2 \right).$$

Theorem: Vovk, 1998

Assume that $\max_t \|x_t\|_2 \leq r$ *and* $\max_t |y_t| \leq m$. *The following holds for any* $\theta^\star \in \mathbb{R}^d$*:*

$$\sum_{t=1}^{T} (y_t - \langle x_t, \widehat{\theta}_t \rangle)^2 \leq \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + \lambda \|\theta^\star\|_2^2 + dm^2 \log\left(1 + \frac{Tr^2}{d\lambda}\right).$$

Back to binary loss and thresholds

There are ways to bypass the difficulty of the threshold example:

- Assuming that the sequence $x_1, \dots, x_T$ is i.i.d.
- Making the margin assumption as in the perceptron analysis.
- Smoothed online learning.
- Transductive setup: the set $\{x_1, \dots, x_T\}$ is known in advance.

If we are given the set $\{x_1, \dots, x_T\}$, we can limit ourselves to $T + 1$ predictors and make at most $\log_2(T + 1)$ mistakes.

- The transductive model in online learning provides a simple playground where the difficulty of learning thresholds is not present.
- Transductive online regret bounds imply statistical excess risk bounds!

Regression: Can we improve Vovk's bound?

Assume that $\{x_1, \dots, x_T\}$ is known in advance.

Initiated by Bartlett, Koolen, Malek, Takimoto, and Warmuth (2015): the minimax strategy for the regression problem is found.

Theorem: Gaillard, Gerchinovitz, Huard, Stoltz (2019)

$$\widetilde{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \left(\sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + (\langle x_t, \theta \rangle)^2 + \lambda \sum_{i=1}^{T} (\langle x_i, \theta \rangle)^2 \right).$$

The following holds for any $\theta^\star \in \mathbb{R}^d$,

$$\sum_{t=1}^{T} (y_t - \langle x_t, \widetilde{\theta}_t \rangle)^2 \le \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + \lambda T m^2 + d m^2 \log\left(1 + \frac{1}{\lambda}\right).$$

Implications

In particular, fixing $\lambda = \frac{d}{T}$, we obtain for $T > 2d$, for any sequence of x_t-s and for any $\theta^\star$,

$$\sum_{t=1}^{T}(y_t - \langle x_t, \widetilde{\theta}_t\rangle)^2 - \sum_{t=1}^{T}(y_t - \langle x_t, \theta^\star\rangle)^2 \lesssim dm^2 \log\left(T/d\right).$$

The loss is technically unbounded, we bound neither $\|x_t\|$, nor $\|\theta^\star\|$!
Since $x_1, \ldots, x_T$ are known, we may assume $\|x_t\|_2 \leq 1$.

We still might have to pay for $\|\theta^\star\|_2$.

Question

In the transductive setup, for which loss functions can we obtain the $d \log T$ regret bound independent of both $x_1, \ldots, x_T$ and $\theta^\star$?

An approach based on exponential weights

The upper bound of Vovk (1998), is usually proved by general results for FTRL predictors + linear algebra.

We return to the original approach: Vovk's predictor is an instance of the exponential weights predictor.

Let $\ell_\theta(\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a set of loss functions parameterized by some $\Theta \subseteq \mathbb{R}^d$.
Fix some prior μ over Θ. Define $\rho_1 = \mu$ and for $t \geq 2$:

$$\rho_t(\theta) \propto \exp\left(-\eta \sum_{i=1}^{t-1} \ell_\theta(x_i, y_i)\right) \mu(\theta).$$

Beyond FTRL/linear algebra: ExpWeights for Sparsity

Example: How to take sparsity of $\theta^\star$ into account? ($\|\theta^\star\|_0 \leq s$)

Choose the data dependent prior in $\mathbb{R}^d$, which is a product of d scaled densities in $\mathbb{R}$,

$$f(x) = \frac{3}{2(1+|x|)^4}.$$

$$\mu(\theta) = \prod_{j=1}^{d} \frac{3 \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2/\tau}}{2\left(1 + |\theta^{(j)}| \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2/\tau}\right)^4}.$$

An x_t-independent version of this prior has been used by Dalalyan and Tsybakov for denoising problems.

Sparsity

Define

$$L_t(\theta, x, y) = (y - \langle x, \theta \rangle)^2 + \sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2, \text{ — A quadratic form!}$$

$$\widehat{f}_t(x) = \frac{m}{2} \log \left(\frac{\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2m^2} L_t(\theta, x, m)\right) \mu(\theta) d\theta}{\underbrace{\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2m^2} L_t(\theta, x, -m)\right)}_{\text{Gaussian-type integral}} \mu(\theta) d\theta} \right).$$

Theorem: Qian, Rakhlin, Zh

Assume that $\max_t |y_t| \le m$ *and that the smallest scaled singular value condition (similar to the lower tail of the RIP condition) is satisfied with constant* κ_s. *For any s-sparse* $\theta^\star \in \mathbb{R}^d$,

$$\sum_{t=1}^{T} (y_t - \widehat{f}_t(x_t))^2 - \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 \lesssim s m^2 \log\left(\frac{dT}{\kappa_s^2 s}\right).$$

Logistic regression

Logistic regression with the logarithmic loss: $x \in \mathbb{R}^d$, $y \in \{1, -1\}$.
Our probability assignment for x is given by

$$\sigma(y\langle x, \theta\rangle) = \frac{1}{1 + \exp(-y\langle x, \theta\rangle)}.$$

We focus on the logarithmic/cross-entropy loss $-\log(\sigma(y\langle x, \theta\rangle))$.

$$\mathrm{Regret} = -\sum_{t=1}^{T} \log(\widehat{p}_t(x_t, y_t)) - \inf_{\theta} \left[-\sum_{t=1}^{T} \log(\sigma(y_t\langle x_t, \theta\rangle)) \right].$$

Probability assignments in logistic regression

What are the best known regret bounds?

$$\text{Regret} = -\sum_{t=1}^{T} \log(\widehat{p}_t(x_t, y_t)) - \inf_{\theta \in \mathbb{R}^d} -\sum_{t=1}^{T} \log(\sigma(y_t \langle x_t, \theta \rangle)).$$

- Online gradient descent: Regret $\lesssim \|\theta^\star\|\sqrt{T}$.
- Online Newton step: Regret $\lesssim d\exp(\|\theta^\star\|)\log(T)$.
- Exponential weights: Regret $\lesssim d\log(\|\theta^\star\| T)$ (Kakade and Ng, 2004, Cesa-Bianchi and Lugosi 2006, Foster, Kale, Luo, Mohri, and Sridharan 2018) — all related to (Vovk, 2001)'s work on sequential linear regression.

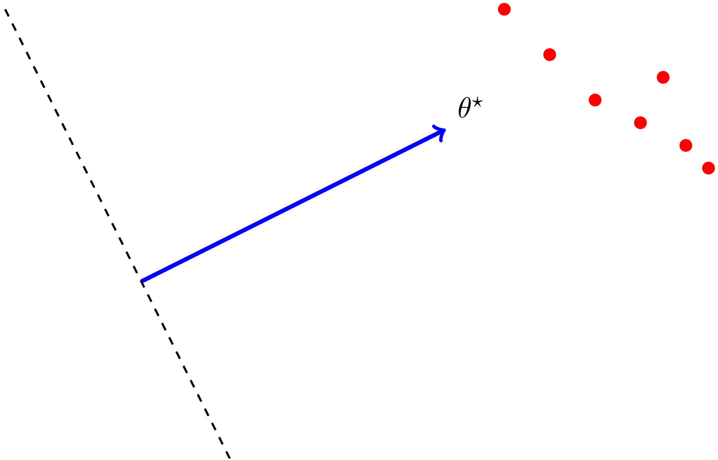
In fact, $\min\{d\log(\|\theta^\star\|), T\}$ cannot be improved! The example is based on the lower bound for classification of thresholds.

The hard case

Recall

$$\theta^\star = \arg\inf_{\theta \in \mathbb{R}^d} - \sum_{t=1}^{T} \log(\sigma(y_t \langle x_t, \theta \rangle)).$$

Do we really need to suffer from large $\|\theta^\star\|$?



Logistic Regression with known x_t-s

We focus on the sequential probability assignment where the covariates x_t (i.e., the set $\{x_1, \ldots, x_T\}$) are known in advance.

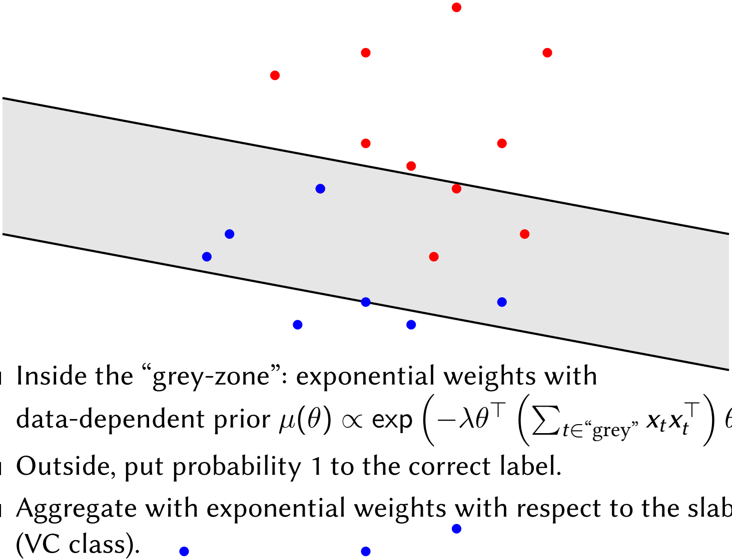
Theorem: Qian, Rakhlin, Zh

Given a known set of covariates $\{x_1, \ldots, x_T\}$, there exists an exp-weights-based sequence of probability assignments $\widehat{p}_t$ such that

$$\sum_{t=1}^{T} -\log(\widehat{p}_t(x_t, y_t)) - \inf_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} -\log(\sigma(y_t \langle x_t, \theta \rangle)) \lesssim d \log T.$$

Geometric ideas

The solution $\theta^\star$ classifies the sample as follows:



- Inside the "grey-zone": exponential weights with data-dependent prior $\mu(\theta) \propto \exp\left(-\lambda\theta^\top\left(\sum_{t\in\text{“grey”}} x_t x_t^\top\right)\theta\right)$.
- Outside, put probability 1 to the correct label.
- Aggregate with exponential weights with respect to the slabs (VC class).

Implications in the i.i.d. case

Observation: Regret bounds in online learning with known x_t-s imply excess risk bounds in the i.i.d. case without any assumptions on x_t.

"Fixed design" online prediction implies results for random design statistical setup!

If we observe an i.i.d. sample $(X_1, Y_1), \ldots, (X_T, Y_T)$, then there is a predictor $\widetilde{p}$ such that

$$\mathbb{E}(-\log(\widetilde{p}(X, Y))) - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}(-\log(\sigma(Y\langle X, \theta \rangle))) \lesssim \frac{d \log T}{T}.$$

Classification with hinge loss

$$\frac{(\gamma - y\widehat{f}(x))_+}{\gamma}.$$

First, using exponential weights with Gaussian prior with clipping:

Theorem: Qian, Rakhlin, Zh

Assume that $\|x_t\| \le 1$*. Then, for any* $\eta \in [0, 3/(10\gamma)]$*, there is a sequence of predictors* $\{\widehat{f}_t(\cdot)\}_{t=1}^T$ *such that*

$$\sum_{t=1}^{T} \frac{(\gamma - y_t\widehat{f}_t(x_t))_+}{\gamma}$$

$$\le (1 + 2\eta\gamma)\left(\sum_{t=1}^{T} \frac{(\gamma - y_t\langle x_t, \theta^\star\rangle)_+}{\gamma} + \frac{cd\log\left(1 + \eta^2 T^2\|\theta^\star\|^2\right)}{\eta\gamma}\right).$$

Back to transductive setting

When the set $\{x_1, \ldots, x_T\}$ is known, the dependence on both γ and $\theta^\star$ disappears under the logarithm:

Theorem: Qian, Rakhlin, Zh

Assume that $\|x_t\| \leq 1$. Then, in the transductive setting, for any $\eta \in [0, 3/(10\gamma)]$, there is a sequence of predictors $\widehat{f}(x_t)$ such that

$$\sum_{t=1}^{T} \frac{(\gamma - y_t\widehat{f}(x_t))_+}{\gamma} \leq (1 + 2\eta\gamma)\left(\sum_{t=1}^{T} \frac{(\gamma - y_t\langle x_t, \theta^\star\rangle)_+}{\gamma} + \frac{cd\log(T)}{\eta\gamma}\right).$$