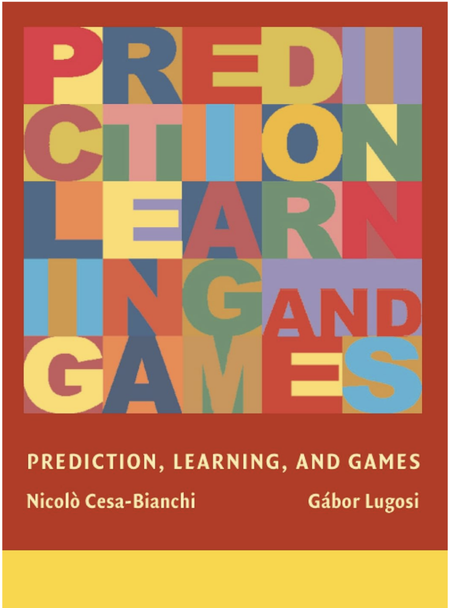
The Many Faces of Exponential Weighting

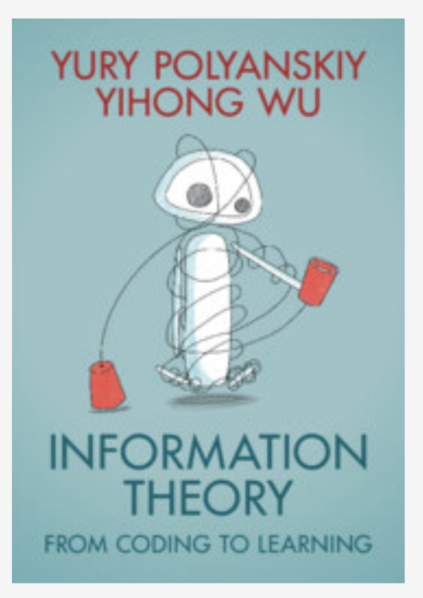
Nikita Zhivotovskiy[1]

[1]UC Berkeley, Department of Statistics

Winter School on Mathematical Aspects of Data Science,
Darwin, Australia, Summer 2024

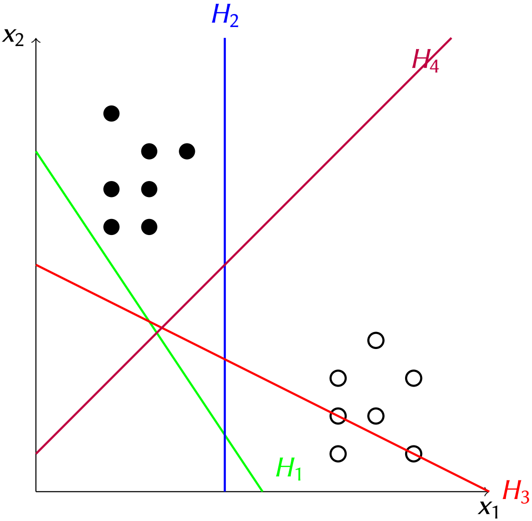
Where can one read about the topic?




YURY POLYANSKIY
YIHONG WU
INFORMATION THEORY
FROM CODING TO LEARNING

Additional Relevant Literature

- **Convex Optimization: Algorithms and Complexity** by Sebastien Bubeck
- **A Modern Introduction to Online Learning** by Francesco Orabona
- **The Multiplicative Weights Update Method: a Meta Algorithm and Applications** by Sanjeev Arora, Elad Hazan, and Satyen Kale
- **Introduction to Online Convex Optimization** by Elad Hazan
- **Bandit Algorithms** by Tor Lattimore and Csaba Szepesvari
- **Understanding Machine Learning: From Theory to Algorithms** by Shai Shalev-Shwartz and Shai Ben-David
- **The Many Faces of Exponential Weights in Online Learning** by Dirk van der Hoeven, Wouter M. Koolen, and Tim van Erven


H_2
x_2
H_4
H_1
H_3
x_1

Classification with margin

We work with $\{-1, 1\}$ labels.

We say that a set of labeled vectors S_N (in $\mathbb{R}^p$) is linearly separable with a margin γ if there is a vector $v \in \mathbb{R}^p \setminus \{0\}$ such that for any $(x, y) \in S_N$, where $x \in \mathbb{R}^p$ and $y \in \{1, -1\}$:

$$\frac{y\langle v, x\rangle}{\|v\|} \geq \gamma.$$

The distance between x and the hyperplane induced by v is

$$\frac{|\langle v, x\rangle|}{\|v\|}.$$

We consider the classifier of the form $x \mapsto \text{sign}(\langle x, w\rangle)$.

The point (x, y) is classified correctly if

$$y\,\mathsf{sign}(\langle x, w\rangle + b) > 0,$$

and is misclassified if

$$y\,\mathsf{sign}(\langle x, w\rangle + b) \leq 0.$$

We focus on $b = 0$ for simplicity.

Perceptron algorithm

Two classical papers:

- The Perceptron – A Perceiving and Recognizing Automaton (1957) by F. Rosenblatt.
- On convergence proofs on perceptrons (1962) by A.B. Novikoff.

In 1958 The New York Times reported the perceptron to be "the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

Perceptron algorithm

Perceptron Algorithm.

- Input: $S_N = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ (a linearly separable dataset with margin $\gamma > 0$)
- Set $w_1 = 0$. (Initialization)
- For $i = 1, \ldots, N$ do

```
If y_i⟨w_i, x_i⟩ ≤ 0
        w_{i+1} = w_i + y_i x_i,
Else
        w_{i+1} = w_i,
```

- Return: w_{N+1}.

Whenever w_i misclassifies x_i, we update it by using the rule $w_{i+1} = w_i + y_i x_i$. This implies that

$$y_i\langle w_{i+1}, x_i\rangle = y_i\langle w_i, x_i\rangle + \|x_i\|^2 \geq y_i\langle w_i, x_i\rangle.$$

Theorem of Novikoff

Theorem: A. Novikoff 1963

Assume that we are given a set of labeled vectors

$$S_N = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

in $\mathbb{R}^d$ *that is linearly separable with a margin* γ*. The number of updates (misclassifications) made by the Perceptron algorithm when processing* S_N *is bounded by*

$$M = \frac{\max\limits_{i=1,\ldots,N} \|x_i\|_2^2}{\gamma^2}.$$

Running through the data multiple times we make a pass with no errors and thus create a perfect separator.

Multiplicative updates

The update rule for Perceptron is $w_{i+1} = w_i + y_i x_i$.

Assume that $w \in \Delta^d$ – a probability simplex in $\mathbb{R}^d$.

For this w, the linear separation for all (x, y) with margin γ is

$$y\langle w, x\rangle \geq \gamma.$$

Idea: do the multiplicative updates of coordinates

$$w_{t+1,i} = w_{t,i} \cdot \alpha_{t,i}.$$

Additive to multiplicative updates: Winnow Algorithm

Winnow Algorithm (Littlestone, 1988)

- **Input:** $\eta > 0$ (learning rate), N (number of iterations)
- **Initialize:** $w_1 = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$.
- **For** $t = 1, \ldots, N$ **do**
  - Receive x_t
  - Compute $\hat{y}_t = \mathrm{sign}\langle w_t, x_t \rangle$
  - Receive y_t
  - **If** $\hat{y}_t \neq y_t$ **then**
    - Compute $Z_t = \sum_{i=1}^{d} w_{t,i} \exp(\eta y_t x_{t,i})$
    - **For** $i = 1, \ldots, d$ **do**
    - Update $w_{t+1,i} = \frac{w_{t,i} \exp(\eta y_t x_{t,i})}{Z_t}$
  - **Else** set $w_{t+1} = w_t$
- **Return:** w_{N+1}

Theorem: Littlestone, 1988

Assume that we are given a set of labeled vectors

$$S_N = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

in $\mathbb{R}^d$ *that is linearly separable with a margin* γ *by a vector in* Δ^d*. The number of updates (misclassifications) made by the Winnow algorithm with* $\eta = \frac{\gamma}{\max\limits_{i=1,\ldots,N} \|x_i\|_\infty^2}$ *when processing* S_N *is bounded by*

$$M = \frac{2 \max\limits_{i=1,\ldots,N} \|x_i\|_\infty^2 \log d}{\gamma^2}.$$

Winnow is the special case of the exponential weights/multiplicative weights/hedge algorithm we cover in this mini-course.

Preliminaries: Kullback-Leibler Divergence

- Let ρ, π be probability densities supported on $\Theta \subseteq \mathbb{R}^d$.
- The **Kullback-Leibler divergence** (**KL divergence**, also known as **relative entropy**), is

$$\mathcal{KL}(\rho \parallel \pi) = \int_{\Theta} \log\left(\frac{\rho(\theta)}{\pi(\theta)}\right) \rho(\theta) d\theta = \mathbb{E}_{\theta \sim \rho}\left[\log\left(\frac{\rho(\theta)}{\pi(\theta)}\right)\right].$$

Fact:

1. $\mathcal{KL}(\rho \parallel \pi) \geq 0$
2. $\mathcal{KL}(\rho \parallel \pi) = 0$ if and only if $\rho(\theta) = \pi(\theta)$ almost everywhere.

Preliminaries

Lemma: Donsker-Varadhan variational formula

Let π be a probability density supported on $\Theta \subseteq \mathbb{R}^d$, and let $h : \Theta \to \mathbb{R}$ be a function with $\mathbb{E}_{\theta\sim\pi} e^{h(\theta)} < \infty$. Then

$$\log \mathbb{E}_{\theta\sim\pi} e^{h(\theta)} = \sup_{\rho} \left\{ \mathbb{E}_{\theta\sim\rho} h(\theta) - \mathcal{KL}(\rho \parallel \pi) \right\},$$

where the supremum is taken over all probability densities ρ such that $\mathcal{KL}(\rho \parallel \pi) < \infty$.
Moreover, the supremum in r.h.s. is achieved by

$$\rho'(\theta) = \frac{e^{h(\theta)}\pi(\theta)}{\mathbb{E}_{\theta'\sim\pi}\, e^{h(\theta')}}$$

Works equally well for discrete distributions.

Going back to prediction

Consider a loss function $\ell_\theta(x, y)$ parametrized by $\theta \in \Theta$.

Example: Linear classification

$$\ell_\theta(x, y) = \mathbb{1}[\mathsf{sign}(\langle x, \theta \rangle) \neq y].$$

Example: Empirical loss so far by t-th round of prediction

$$\sum_{i=1}^{t-1} \mathbb{1}[\mathsf{sign}(\langle x_i, \theta \rangle) \neq y_i].$$

At round t we want to construct a distribution over Θ based on the data we have seen so far. Naive idea:

$$\widehat{\rho}_t = \arg\min_\rho \mathbb{E}_{\theta \sim \rho} \left[\sum_{i=1}^{t-1} \ell_\theta(x_i, y_i) \right].$$

Entropic regularization

Fix $\eta > 0$ and the prior π over Θ,

$$\widehat{\rho}_t = \arg\min_{\rho} \left[\mathbb{E}_{\theta \sim \rho} \sum_{i=1}^{t-1} \ell_\theta(x_i, y_i) + \frac{1}{\eta} \mathcal{KL}(\rho \parallel \pi) \right].$$

We can solve this explicitly using the Donsker-Varadhan formula. Taking

$$h(\theta) = -\eta \sum_{i=1}^{t-1} \ell_\theta(x_i, y_i),$$

we have

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

We also have that the minimized value of the regularized loss is

$$-\frac{1}{\eta} \log\left(\mathbb{E}_{\theta \sim \pi} \exp\left(-\eta \sum_{i=1}^{t-1} \ell_\theta(x_i, y_i) \right) \right).$$

From measures to prediction

Once we built $\widehat{\rho}_t$, we can construct the predictor.

Importantly, this depends on a particular loss function we are using.

Example: Absolute loss with $y \in \mathbb{R}, x \in \mathbb{R}^d$,

$$|y - f_\theta(x)|.$$

Standard approach: build some $\widehat{\theta}$ and suffer the loss

$$|y_t - f_{\widehat{\theta}}(x_t)|.$$

If we construct the measure $\widehat{\rho}_t$, our prediction is $\mathbb{E}_{\theta \sim \widehat{\rho}_t} f_\theta$ and the loss

$$\left| y_t - \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} f_\theta(x_t) \right|.$$

Mix-loss and its properties

Recall the following formula:

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

Definition: Mix-loss

Fix $\eta > 0$. Given a sequence $\widehat{\rho}_1, \ldots, \widehat{\rho}_T$ of distributions, define the mix-loss at round t as

$$-\frac{1}{\eta} \log\left(\mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} \exp\left(-\eta \ell_\theta(x_t, y_t)\right)\right).$$

From the Donsker-Varadhan identity we have that the mix-loss is equal to

$$\min_{\rho} \left\{ \mathop{\mathbb{E}}_{\theta \sim \rho} \ell_\theta(x_i, y_i) + \frac{1}{\eta} \mathcal{KL}(\rho \parallel \widehat{\rho}_t) \right\}.$$

Tensorization of mix-losses

Lemma: Sum of mix-losses

The following holds for the distributions $\widehat{\rho}_1, \ldots, \widehat{\rho}_T$ output by the exponential weights algorithm:

$$\sum_{t=1}^{T} -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \widehat{\rho}_t} \exp\left(-\eta \ell_\theta(x_t, y_t)\right)\right)$$

$$= -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \pi} \exp \left(-\eta \sum_{t=1}^{T} \ell_\theta(x_t, y_t) \right) \right).$$

Proof.
A direct computation based on the definition of $\widehat{\rho}_t$. $\square$

A general recipe for analyzing exponential weights

1. Use the specific properties of the loss function to make a prediction such that

$$\text{Loss of the prediction at round } t \leq \underbrace{-\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\widehat{\rho}_t}\exp\left(-\eta\ell_\theta(x_t, y_t)\right)\right)}_{\text{mix-loss}_t}.$$

2. Use the tensorization property to prove

$$\sum_{t=1}^{T}\text{mix-loss}_t = -\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\pi}\exp\left(-\eta\sum_{t=1}^{T}\ell_\theta(x_t, y_t)\right)\right).$$

3. Upper bound using direct computation or via the Donsker-Varadhan duality formula

$$-\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\pi}\exp\left(-\eta\sum_{t=1}^{T}\ell_\theta(x_t, y_t)\right)\right).$$

The logarithmic loss

1 Let f be a density. Then

$$\mathbb{E}_{X\sim f}[-\log(f(X))] = \mathbb{E}_{X\sim f}\log\left(\frac{1}{f(X)}\right)$$

is the entropy.

2 Consider a classification task, where $y \in \{0, 1\}$ and we predict the probability of a 'success' $\hat{p} \in (0, 1)$. Note that $-(y\log(\hat{p}) + (1-y)\log(1-\hat{p}))$ is equivalent to the cross-entropy loss.

3 Consider data points $Z_1, \dots, Z_n$ and density f_θ. The maximum likelihood procedure $\log(\prod_{i=1}^n f_\theta(Z_i)) = \sum_{i=1}^n \log(f_\theta(Z_i))$. Maximizing this quantity over $\theta \in \Theta$ is equivalent to minimizing

$$-\sum_{i=1}^n \log(f_\theta(Z_i)).$$

The logarithmic loss

For a pair of densities f, g, it holds that

$$\mathbb{E}_{X\sim f}[-\log(g(X)) - (-\log(f(X)))] = \mathcal{KL}(f \parallel g).$$

The excess risk with respect to the logarithmic loss corresponds to the $\mathcal{KL}$ divergence if the data is generated by the risk minimizer.

The logarithmic loss is the easiest to work with when considering the exponential weights algorithm.

Assume we have a family of densities $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. We observe $z_1, \ldots, z_T$. Consider the mix-loss at round t,

$$-\frac{1}{\eta}\log\left(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_t}\exp\left(-\eta(-\log(f_\theta(z_t)))\right)\right).$$

Density estimation and the logarithmic loss

Recall our general strategy:

$$\text{Loss at round } t \le -\frac{1}{\eta}\log\left(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_t}\exp\left(-\eta(-\log(f_\theta(z_t)))\right)\right).$$

Observe that for $\eta = 1$ we immediately have

$$-\log\left(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_t} f_\theta(z_t)\right) = -\log\left(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_t}\exp\left(-(-\log(f_\theta(z_t)))\right)\right).$$

The predicted density $\mathbb{E}_{\theta\sim\widehat{\rho}_t} f_\theta$ is exactly the Bayesian mixture. Moreover,

$$\widehat{\rho}_t(\theta) \propto \prod_{i=1}^{t-1} f_\theta(z_i)\pi(\theta).$$

Example: Regret for a finite family of densities

Consider the finite family of densities parametrized by Θ of size M. That is,

$$\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_M}\}.$$

No assumptions are made except for $f_\theta(x) \geq 0$ and $\int f_\theta(x)dx = 1$.

Theorem

Let π be the uniform prior over Θ. The exponential weights algorithm with $\eta = 1$ satisfies

$$\sum_{t=1}^{T} -\log\Big(\mathop{\mathbb{E}}_{\theta\sim\widehat{\rho}_t} f_\theta(z_t)\Big) - \min_{\theta\in\Theta} \sum_{t=1}^{T} -\log(f_\theta(z_t)) \leq \log(M).$$

Progressive mixture estimator

The same set of finite densities, but for $\theta^\star \in \Theta$ we observe the full sample i.i.d.

$$Z_1, \ldots, Z_T$$

sampled according to $f_{\theta^\star}$. Our aim is to estimate $\theta^\star$.

Theorem: A. Barron (1987)

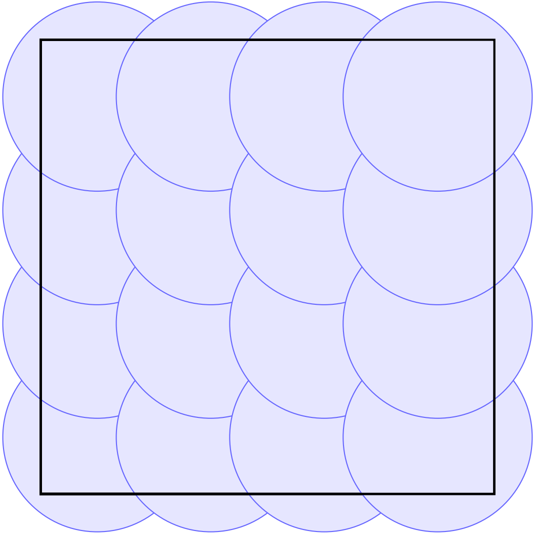
Consider the density predictor

$$\widehat{f} = \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} f_\theta.$$

The following bounds holds

$$\mathop{\mathbb{E}}_{Z_1, \ldots, Z_T} \mathcal{KL}(f_{\theta^\star} \,\|\, \widehat{f}) \le \frac{\log(M)}{T}.$$

Infinite Classes: Covering Numbers



Infinite Classes: Barron-Yang Construction

Let $\mathcal{F}$ be a collection of densities parametrized by Θ.

$$\mathcal{N}(\mathcal{F}, \mathcal{KL}, \varepsilon) = \min\{N \in \mathbb{N} : \exists q_1, \dots, q_N \text{ s. t. for all } \theta \in \Theta, \exists i \in [N] \text{ s.t. } \mathcal{KL}(f_\theta, q_i) \le \varepsilon^2\}.$$

Idea: Fix $\varepsilon > 0$ and let N_ε be the net corresponding to $\mathcal{N}(\mathcal{F}, \mathcal{KL}, \varepsilon)$. Let $\widehat{f}$ be a progressive mixture on $q_1, \dots, q_{N_\varepsilon}$ with the uniform prior on this set.

Theorem: Barron-Yang, 1999

Assume $Z_1, \dots, Z_T \sim f_{\theta^\star}$, with $f_{\theta^\star} \in \mathcal{F}$. Then there exists a $\widehat{f}$ which satisfies

$$\mathbb{E}_{Z_1,\dots,Z_T} \mathcal{KL}(f_{\theta^\star} \,\|\, \widehat{f}) \le \inf_{\varepsilon > 0} \left\{ \varepsilon^2 + \frac{\log \mathcal{N}(\mathcal{F}, \mathcal{KL}, \varepsilon)}{T} \right\}.$$

Example: Gaussian densities via Barron and Yang

Let $\mathcal{F} = \{\mathcal{N}(\theta, I_d) : \theta \in \Theta\}$, where $\Theta = B_2^d$.

We observe $Z_1, \ldots, Z_T \overset{\text{iid}}{\sim} \mathcal{N}(\theta^*, I_d)$, with $\theta^* \in \Theta$.

Note that

$$\mathcal{KL}\left(\mathcal{N}(\theta_1, I_d) \parallel \mathcal{N}(\theta_2, I_d)\right) = \frac{1}{2}\|\theta_1 - \theta_2\|_2^2.$$

By the volumetric argument:

$$\mathcal{N}(\mathcal{F}, \mathcal{KL}, \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^d.$$

Progressive mixture $\widehat{f}$ gives us the following bound:

$$\mathbb{E}_{Z_1,\ldots,Z_T}\mathcal{KL}(\mathcal{N}(\theta^*, I_d) \parallel \widehat{f}) \lesssim \inf_{\varepsilon>0}\left\{\varepsilon^2 + \frac{d\log(c/\varepsilon)}{T}\right\} \lesssim \frac{d\log T}{T}.$$

How to choose the optimal prior for exponential weights?

Clarke, Barron (1994), and Rissanen (1996) studied optimal prior distributions for exponential weights in the context of log-loss with asymptotic results, typically for well-specified i.i.d. data.

Heuristic derivation for the total loss ($\theta^\star$ is minimizer, $\ell_{t,\theta} := \ell_\theta(x_t, y_t)$):

$$
\begin{aligned}
&\mathbb{E}_{\theta\sim\pi} \exp\left(-\sum_{t=1}^{T} \eta\ell_{t,\theta}\right) \\
&\approx \int_{\mathbb{R}^d} \pi(\theta^\star) \exp\left(-\sum_{t=1}^{T} \eta\ell_{t,\theta^\star} - \frac{1}{2}(\theta-\theta^\star)^\top \mathrm{Hess}_t(\theta^\star)(\theta-\theta^\star)\right) d\theta \\
&= \pi(\theta^\star) \exp\left(-\sum_{t=1}^{T} \eta\ell_{t,\theta^\star}\right) \frac{(2\pi)^{d/2}}{\sqrt{\det\left(\sum_{t=1}^{T} \mathrm{Hess}_t(\theta^\star)\right)}}.
\end{aligned}
$$

Jeffreys prior for exponential weights

Applying $-\frac{1}{\eta}\log(\ldots)$ to the last expression, we get for (approximate) total error

$$\sum_{t=1}^{T} \ell_{t,\theta^\star} + \frac{d}{2\eta}\log\left(\frac{T}{2\pi}\right) + \frac{1}{\eta}\log\left(\frac{\sqrt{\det(\frac{1}{T}\sum_{t=1}^{T}\mathsf{Hess}_t(\theta^\star))}}{\pi(\theta^\star)}\right).$$

A natural idea to pick the Jeffreys prior:

$$\pi(\theta) \propto \sqrt{\det\left(\frac{1}{T}\sum_{t=1}^{T}\mathsf{Hess}_t(\theta)\right)}.$$

Idea: Find a prior using the above heuristic and then provide a finite sample regret bound with this prior.

Discrete probability assignments

We observe a sequence of bits $z_1, \dots, z_T$ (that is, $z_t \in \{0, 1\}$). Our aim is to assign probabilities sequentially such that the regret

$$\sum_{t=1}^{T} -\log(\widehat{p}(z_t)) - \inf_{p\in[0,1]} \sum_{t=1}^{T} (-\log(p)\mathbb{1}[z_t = 1] - \log(1-p)\mathbb{1}[z_t = 0])$$

Such a bound can immediately converted into a statistical bound

$$\mathop{\mathbb{E}}_{Z_1,\dots,Z_T} \mathcal{KL}(p \,\|\, \widetilde{p}) \leq \frac{\mathrm{Regret}}{T},$$

where we assume that $Z_t \sim \mathrm{Be}(p)$.

Discrete probability assignments

Let n_0 be the number of zeros and n_1 be the number of ones and define $p^\star = \frac{n_1}{n_0+n_1}$. We have

$$\inf_{p\in[0,1]} \sum_{t=1}^{T} (-\log(p)\mathbb{1}[z_t = 1] - \log(1-p)\mathbb{1}[z_t = 0])$$
$$= T(-p^\star \log(p^\star) - (1-p^\star)\log(1-p^\star)).$$

Compute the second derivative for Jeffreys prior:

$$\left|\frac{\partial^2}{\partial^2 p} \sum_{t=1}^{T} (-\log(p)\mathbb{1}[z_t = 1] - \log(1-p)\mathbb{1}[z_t = 0])\Big|_{p=p^\star}\right| \propto \frac{1}{p^\star(1-p^\star)}$$

Thus, the Jeffreys prior ($\propto \sqrt{\det(\mathrm{Hess}(\theta))}$) is the Beta$(1/2, 1/2)$ distribution

$$\pi(\theta) = \frac{1}{\pi\sqrt{p(1-p)}}.$$

Krichevsky-Trofimov estimator

Assume that before round t we observe n_0^t zeros and n_1^t ones, so that $n_0^t + n_1^t = t - 1$. Given the Beta$(1/2, 1/2)$ prior we note that

$$\widehat{\rho}_t \propto \frac{p^{n_1^t}(1-p)^{n_0^t}}{\pi\sqrt{p(1-p)}}.$$

And therefore,

$$\widehat{p}_t(1) = \frac{\int_0^1 \frac{p^{n_1^t+1}(1-p)^{n_0^t}}{\pi\sqrt{p(1-p)}}\,dp}{\int_0^1 \frac{p^{n_1^t}(1-p)^{n_0^t}}{\pi\sqrt{p(1-p)}}\,dp}$$

Furthermore, direct computations show that

$$\sum_{t=1}^{T} -\log(\widehat{p}(z_t)) - \inf_{p\in[0,1]}\sum_{t=1}^{T}\left(-\log(p)\mathbb{1}[z_t = 1] - \log(1-p)\mathbb{1}[z_t = 0]\right)$$
$$\leq \frac{1}{2}\log(T) + \log(2).$$

Square loss

Consider the square loss

$$(y - f_\theta(x))^2,$$

where f_θ is a class of functions parametrized by Θ.

Lemma: Mixability of the square loss (Vovk, 1990, 2001)

Assume that $|y| \leq m$ (no assumptions on f_θ). Consider the predictor

$$\widehat{f}_t(x) = \frac{m}{2}\log\left(\frac{\mathbb{E}_{\theta\sim\widehat{\rho}_t}\exp\left(-\frac{1}{2m^2}(m - f_\theta(x))^2\right)}{\mathbb{E}_{\theta\sim\widehat{\rho}_t}\exp\left(-\frac{1}{2m^2}(-m - f_\theta(x))^2\right)}\right).$$

Then

$$(y - \widehat{f}_t(x))^2 \leq \underbrace{-2m^2\log\left(\mathbb{E}_{\theta\sim\widehat{\rho}_t}\exp\left(-\frac{1}{2m^2}(y - f_\theta(x))^2\right)\right)}_{\textit{Mix-loss with } \eta=1/2m^2}.$$

Vovk's predictor

We are planning to interpret the following predictor:

$$\widehat{f}_t(x) = \frac{m}{2}\log\left(\frac{\mathbb{E}_{\theta\sim\widehat{\rho}_t}\exp\left(-\frac{1}{2m^2}(m - f_\theta(x))^2\right)}{\mathbb{E}_{\theta\sim\widehat{\rho}_t}\exp\left(-\frac{1}{2m^2}(-m - f_\theta(x))^2\right)}\right).$$

Fix $\lambda > 0$. Let us choose the Gaussian prior

$$\pi(\theta) \propto \exp\left(-\lambda\eta\|\theta\|_2^2\right).$$

Direct integration (only Gaussian integrals are involved) shows that

$$\widehat{f}_t(x_t) = \langle\widehat{\theta}_{t,x_t}, x_t\rangle,$$

where

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

Online linear regression

Due to Vovk's result relating predictions and mix-losses, we only have to bound the sum of mix-losses

$$-2m^2 \log\left(\mathbb{E}_{\theta\sim\pi}\exp\left(-\frac{1}{2m^2}\sum_{t=1}^{T}(y_t - \langle x_t, \theta\rangle)^2\right)\right).$$

Computations reduce to Gaussian integration. This leads to

Theorem: Vovk, 1998

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

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

Simplification of predictors: Exp-concavity

When both y and $f_\theta(x)$ are absolutely bounded by m we may use a different idea.

Let $\eta = \frac{1}{8m^2}$. Then for any distribution ρ,

$$\left(y - \mathop{\mathbb{E}}_{\theta\sim\rho} f_\theta\right)^2 \le \underbrace{-\frac{1}{\eta}\log\left(\mathop{\mathbb{E}}_{\theta\sim\rho} \exp(-\eta(y - f_\theta(x))^2)\right)}_{\text{mix-loss}}.$$

Simple bound for finite families

Assume that $|y_t| \le m$ and $|f_\theta(x_t)| \le m$ for all $\theta \in \Theta$ with $|\Theta| = M$.

Theorem

Under the boundedness assumptions introduced above, for any sequence $(x_t, y_t)_{t=1}^T$,

$$\sum_{t=1}^{T}\left(y_t - \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} f_\theta(x_t)\right)^2 - \inf_{\theta \in \Theta} \sum_{t=1}^{T}\left(y_t - f_\theta(x_t)\right)^2 \le 8m^2 \log M.$$

Progressive mixture for the square loss

Given a random pair (X, Y), define $R(f) = \mathbb{E}(f(X) - Y)^2$.

Theorem: Yang (2000), Catoni (1997)

Let $(X_t, Y_t)_{t=1}^T$ be an i.i.d. sample of copies of (X, Y). Assume that a.s. $|Y| \le m$ and $|f_\theta(X)| \le m$. Set

$$\widehat{f}^{pm} = \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} f_\theta.$$

The following bound holds for Θ of size M,

$$\mathbb{E}\, R(\widehat{f}^{pm}) - \min_{\theta \in \Theta} R(f_\theta) \le \frac{8m^2 \log(M)}{T}.$$

- Using Vovk's mixability result we can remove the assumption $|f_\theta(X)| \le m$.
- One can even replace $|Y| \le m$ by $\mathbb{E}[Y^2|X] \le m^2$ a.s.

Large variance of online-to-batch conversions

Progressive mixture rules do not give sharp high probability bounds in the random design setting

$$\mathbb{E}\, R(\widehat{f}^{\mathrm{pm}}) \quad \text{instead of} \quad R(\widehat{f}^{\mathrm{pm}})$$

Theorem: Audibert (2007)

With probability at least $1-\delta$*, over the realization of the training sample*

$$R(\widehat{f}^{pm}) - \min_{\theta\in\Theta} R(f_\theta) \lesssim \frac{\log(M)}{T} + \sqrt{\frac{\log(1/\delta)}{T}}.$$

*Most important*ly*, the term* $\frac{1}{\sqrt{T}}$ *cannot be improved in general!*

Variance reduction solution (Square loss)

Define the modified loss function at round t as follows:

$$\tilde{\ell}_t(f_\theta) = \left(\tfrac{1}{2}f_\theta(X_t) + \tfrac{1}{2}\hat{f}_t(X_t) - Y_t\right)^2 ,$$

We say that $\hat{f}_1, \ldots, \hat{f}_T$ satisfy the *bounded shifted regret* condition if

$$\sum_{t=1}^{T} \tilde{\ell}_t(\hat{f}_t) - \min_{\theta\in\Theta} \sum_{t=1}^{T} \tilde{\ell}_t(f_\theta) \le \mathcal{R}_T.$$

The regret bounds for the shifted regret are the same as for the original regret.

Variance reduction by shifted losses

Theorem: Cesa-Bianchi, Van der Hoeven, Zh. 2023

Assume that both $f_\theta(X)$ and Y are absolutely bounded by m. Let $\mathcal{R}_T$ be a bound on the shifted regret for $\hat{f}_1, \ldots, \hat{f}_T$ built sequentially using a random sample $(X_1, Y_1), \ldots, (X_T, Y_T)$. Define

$$\bar{f}_T = \frac{1}{T}\sum_{i=1}^{T}\hat{f}_i.$$

Then, with probability at least $1-\delta$,

$$R(\bar{f}_T) - \min_{\theta\in\Theta} R(f_\theta) \leq \frac{2\mathcal{R}_T}{T} + \frac{64m^2\log(1/\delta)}{T}.$$

The key aspect of this extension is its applicability to other loss functions, including logarithmic/cross entropy + we can accommodate an infinite Θ.

The proof is simple. High level ideas for the square loss:

$$
\begin{aligned}
&\left(\tfrac{1}{2}f_\theta(X_t) + \tfrac{1}{2}\hat{f}_t(X_t) - Y_t\right)^2 \\
&= \tfrac{1}{2}\left(f_\theta(X_t) - Y_t\right)^2 + \tfrac{1}{2}\left(\hat{f}_t(X_t) - Y_t\right)^2 - \tfrac{1}{4}\left(f_\theta(X_t) - \hat{f}_t(X_t)\right)^2.
\end{aligned}
$$

- Freedman's inequality (martingale counterpart to Bernstein's inequality) gives a variance term that may lead to an additional $\frac{1}{\sqrt{T}}$-factor.
- The negative term $-\frac{1}{4}\left(f_\theta(X_t) - \hat{f}_t(X_t)\right)^2$ compensates for this variance.

Extension to general loss functions (e.g., log-loss) is more involved but uses the same idea of variance compensation.

Bounded losses

The classical algorithm of Littlestone and Warmuth (1994) works with general bounded losses.

Theorem

Assume that $\ell_\theta(z_t) \in [0, m]$. Then for any $\eta > 0$, the exponential weights algorithm satisfies

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} \ell_\theta(z_t) \le \inf_{\gamma} \left\{ \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\theta \sim \gamma} \ell_\theta(z_t) + \frac{\mathcal{KL}(\gamma \,\|\, \pi)}{\eta} \right\} + \frac{T m^2 \eta}{8}.$$

Example: $|\Theta| = M$; π is a uniform measure, $m = 1$ imply after optimizing η,

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\theta \sim \widehat{\rho}_t} \ell_\theta(z_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_\theta(z_t) \le \sqrt{\frac{T \log(M)}{2}}.$$

Additional applications: Matrix multiplicative weights

We begin with the standard matrix concentration inequality.

Theorem: Matrix Bernstein inequality, Tropp (2011)

Assume that $X_1, \ldots, X_T$ are independent zero mean symmetric matrices such that $\|X_i\| \le L$ almost surely. The following holds

$$\mathbb{E}\,\lambda_{\max}\left(\sum_{i=1}^{T} X_i\right) \le \sqrt{2\left\|\sum_{i=1}^{T}\mathbb{E}\,X_i^2\right\|\log(d)} + \frac{1}{3}L\log(d).$$

As an exercise, we will try to think of this result as a corollary of the exponential weights regret bound.

From distributions to matrices

When working with Winnow, we played with the distribution simplex Δ^d.

Now we work with matrices. Let $\mathbb{D}_{d\times d}$ be the set of density matrices — the p.s.d. matrices with trace equal to 1.

An analog of inner products: $\langle A, B\rangle = \mathrm{Tr}(AB)$.

An analog of the $\mathcal{KL}$ divergence (A, B are p.s.d. but not always trace one):

$$\mathcal{KL}(A, B) = \langle A, \log A - \log B\rangle + \langle I, B - A\rangle.$$

It is easy to prove that for any $A \in \mathbb{D}_{d\times d}$,

$$\mathcal{KL}\left(A, \frac{1}{d}I\right) \le \log d.$$

Multiplicative weights on matrices

We are going to run the matrix multiplicative weights on the sequence $\left(-X_t + \eta X_t^2\right)_{t=1}^{T}$. Following our logic:

Fix $\eta \geq 0$ and consider the update rule (with identity prior)

$$\widetilde{\rho}_{t+1} = \operatorname*{arg\,min}_{\rho \succeq 0} \left\{ \langle \rho, -X_t + \eta X_t^2 \rangle + \frac{1}{\eta} \mathcal{KL}(\rho, \widehat{\rho}_t) \right\}.$$

We need to normalize these weights to make it a density matrix.

$$\widehat{\rho}_{t+1} = \operatorname*{arg\,min}_{\rho \in \mathbb{D}_{d\times d}} \mathcal{KL}(\rho, \widetilde{\rho}_{t+1}).$$

Following similar lines, we can show that for any $\rho \in \mathbb{D}_{d\times d}$,

$$\sum_{t=1}^{T} \langle \rho - \widehat{\rho}_t, X_t \rangle \le 2\sqrt{\log(d) \sum_{t=1}^{T} \langle \rho, X_t^2 \rangle + 4L\log(d)}.$$

One can easily show that

$$\sum_{t=1}^{T} \left\langle \rho, X_t^2 \right\rangle \le \left\| \sum_{t=1}^{T} X_t^2 \right\|.$$

Moreover,

$$\mathop{\mathbb{E}}_{X_t} \langle \widehat{\rho}_t, X_t \rangle = 0.$$

With some additional effort, this can be reduced to

$$\mathbb{E} \left\| \sum_{t=1}^{T} X_t \right\| \le 2\sqrt{\log(d) \left\| \sum_{t=1}^{T} \mathbb{E}\, X_t^2 \right\| + 4L\log(d)}.$$

Thank you!