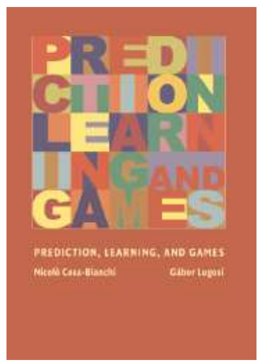
Week 1: Introduction to Online Learning

1 Introduction

This is written based on *Prediction, Learning, and Games (ISBN: 0521841089 / 0-521-84108-9) Cesa-Bianchi, Nicolo; Lugosi, Gabor*



1.1 A Gentle Start

Consider the problem of predicting an unknown sequence $y_1, y_2, ...$ of bits $y_t \in \{0, 1\}$. At each time t the forecaster first makes his guess $\hat{p}_t \in \{0, 1\}$ for y_t. Then the true bit y_t is revealed and the forecaster finds out whether his prediction was correct. To compute $\hat{p}_t$ the forecaster listens to the advice of N experts. This advice takes the form of a binary vector $(f_{1,t}, ..., f_{N,t})$, where $f_{i,t} \in \{0, 1\}$ is the prediction that expert i makes for the next bit y_t.

Goal Bound the number of time steps t in which $\hat{p}_t \neq y_t$, that is, to bound the number of mistakes made by the forecaster.

1.1.1 If we know 'god' exist among experts

If there is an expert that predicts y_t 100% correctly, you can define a ***forecaster*** (algorithm) that makes at most $\lfloor \log_2 N \rfloor$ mistakes.

Algorithm Do majority vote using experts that haven't made any mistakes yet.
1. Set $t = 1$. Define $w_k = 1$ for experts $k = 1..N$.
2. Predict $\hat{p}_t = 1$ if weighted average of $f_{1,t}, .., f_{N,t}$ is greater than 1/2, 0 otherwise.

3. y_t reveals. Set $w_k \leftarrow 0$ for those experts who made a mistake.
4. Increase $t \leftarrow t+1$ Go to 2.

Proof m: #mistakes that forecaster have made so far.
$w_k \in \{0,1\}$: weight for expert k.
W_m : Sum of weights of experts after making m mistakes.

1.1.2 If we don't know anything about experts?

Let's come up with a simple idea and bound the performance.

Algorithm Redefine $w_k \in [0,1]$. We do weighted majority vote with weight updates. At each round, apply $w_k \leftarrow \beta w_k$ for all experts who made a mistake ($\beta \in (0,1)$).

1. Set $t = 1$. Define $w_k = 1$ for experts $k = 1..N$.
2. Predict $\hat{p}_t = 1$ if weighted average of $f_{1,t}, .., f_{N,t}$ is greater than 1/2, 0 otherwise.
3. y_t reveals. Set $w_k \leftarrow \beta w_k$ for those experts who made a mistake.
4. Increase $t \leftarrow t+1$ Go to 2.

Analysis and Bound Consider we are at some time step where the forecaster made his m'th mistake.
k: expert index with the fewest mistakes so far.
m^*: mistakes made by expert k.

$$m \leq \left\lfloor \frac{\log N + m^* \log(\frac{1}{\beta})}{\log \frac{2}{1+\beta}} \right\rfloor \tag{1.1}$$

By choosing $\beta = 1/e$.

$$\begin{align} m &\leq \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}} m^* + \frac{1}{\frac{2}{1+\beta}} \ln N \tag{1.2} \\ &\leq 2.63 m^* + 2.63 \ln N \tag{1.3} \end{align}$$

2 Prediction with Expert Advice

2.1 Formal Protocol

Parameters: decision space $\mathcal{D}$ for $\hat{p}_t$ and $f_{i,t}$, outcome space $\mathcal{Y}$ for y_t, loss function $l : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, set $\mathcal{E}$ of expert indices.

For each round t = 1,2,...

1. the environment chooses the next outcome y_t and the expert advice $\{f_{E,t} \in \mathcal{D} : E \in \mathcal{E}\}$; the expert advice is revealed to the forecaster:
2. the forecaster chooses the prediction $\hat{p}_t \in \mathcal{D}$;
3. the environment reveals the next outcome $y_t \in \mathcal{Y}$;
4. the forecaster incurs loss $l(\hat{p}_t, y_t)$ and each expert E incurs loss $l(f_{E,t}, y_t)$.

2.2 Goal

The forecaster's goal is to keep as small as possible the ***cumulative regret*** (or simply ***regret***) with respect to each expert. Regret w.r.t expert E is defined by the sum

$$R_{E,n} = \sum_{t=1}^{n}(l(\hat{p}_t, y_t) - l(f_{E,t}, y_t)) = \hat{L}_n - L_{E,n} \tag{2.1}$$

where

$$\text{cumulative loss of forcaster} \ : \quad \hat{L}_n = \sum_{t=1}^{n} l(\hat{p}_t, y_t) \tag{2.2}$$

$$\text{cumulative loss of expert } E \ : \quad L_{E,n} = \sum_{t=1}^{n} l(f_{E,t}, y_t) \tag{2.3}$$

Sometimes it is also convenient to define ***instantaneous regret***

$$r_{E,t} = l(\hat{p}_t, y_t) - l(f_{E,t}, y_t) \tag{2.4}$$

The appropriate goal would be achieving ***vanishing per-round regret***, that is,

$$\max_{i=1..N} R_{i,n} = \hat{L}_n - \min_{i=1..N} L_{i,n} = o(n) \tag{2.5}$$

.

Note that this has to hold for outcome sequence $(y_t, t = 1..n)$ and any experts. In other words, we don't make any assumption about outcome sequence and experts. It is easy to notice that "per-round regret ($\frac{1}{n}\max_{i=1..N} R_{i,n}$)" converges to 0 as n goes to infinity.

* Don't be confused. $R_{i,t}$ is regret w.r.t expert i, and $\max_{i=1..N} R_{i,n}$ is regret of a forecaster.

3 Weighted Average Prediction

In this section we assume regression setting where $\mathcal{D}$ is continuous. We will introduce how to adjust these scheme to classification later. We now generalize weighted scheme that we introduced earlier.

$$\hat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N} w_{j,t-1}} \tag{3.1}$$

where $w_{i,t-1} \geq 0$ is the weight assigned to the expert $i = 1..N$.

Proper choice of w ?

Weight function We define $\phi : \mathbb{R} \rightarrow \mathbb{R}^+ \bigcup \{0\}$, which is nonnegative, convex, twice differentiable, and increasing function. Now we use this to define weight $w_{i,t-1} = \phi'(R_{i,t-1})$.

Sanity check for $\phi'(R_{i,t-1})$

3.1 Blackwell Condition

Lemma 2.1 *If the loss function l is convex in its first argument, then*

$$\sup_{y_t \in \mathcal{Y}} \sum_{i=1}^{N} r_{i,t}\phi'(R_{i,t-1}) \leq 0 \tag{3.2}$$

Proof Let's rephrase it.

$$\forall y_t \in \mathcal{Y}, \sum_{i=1}^{N} r_{i,t}\phi'(R_{i,t-1}) \leq 0 \tag{3.3}$$

Further rephrase using definition for $r_{i,t}$ and $\hat{p}_t$...

cf) Jensen's Inequality:
If a real function f is convex, $f\left(\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{j=1}^{n} w_j}\right) \leq \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{j=1}^{n} w_j}$, for $w_i < 0, i = 1..n$.

Scholars love new notations? If we further develop the framework in the following way, we can cast the above lemma as an interesting way. We introduce

$$\textbf{\textit{instantaneous regret vector}} \quad : \mathbf{r}_t \quad = (r_{1,t}, ..., r_{N,t}) \in \mathbb{R}^N \tag{3.4}$$

$$\textbf{\textit{regret vector}} \quad : \mathbf{R}_t \quad = \sum_{t=1}^{n} \mathbf{r}_t \tag{3.5}$$

It is also introduce a ***potential function*** $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}$ of the form

$$\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^{N} \phi(u_i)\right) \tag{3.6}$$

where $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is nonnegative, strictly increasing, concave, and twice differentiable function, and ϕ as defined before.

Now we can describe $\hat{p}_t$ using our new notations.

$$\hat{p}_t = \frac{\sum_{i=1}^{N} \nabla\Phi(\mathbf{R}_{t-1})_i f_{i,t}}{\sum_{j=1}^{N} \nabla\Phi(\mathbf{R}_{t-1})_j} \tag{3.7}$$

where $\nabla\Phi(\mathbf{R}_{t-1})_i = \partial\Phi(\mathbf{R}_{t-1})/\partial R_{i,t-1}$. It's exactly same definition of $\hat{p}_t$ as previous one since derivative of function ϕ will cancel out.

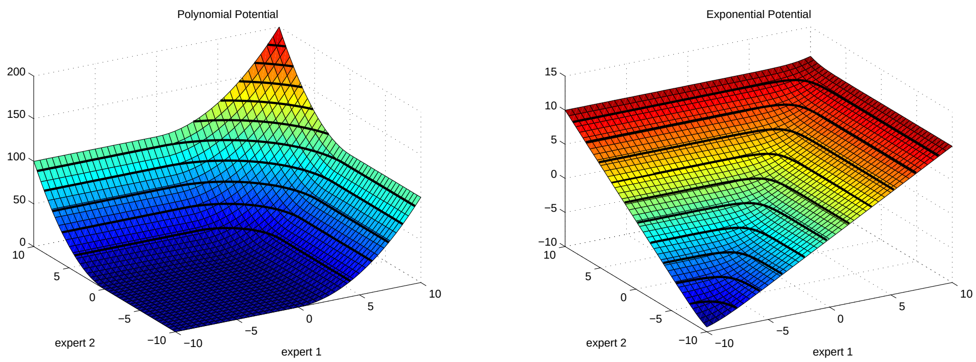


Figure 1: Examples of potential function. Left: $\Phi_p(\mathbf{R}_{t-1}) = (\sum_{i=1}^{N}(R_{i,t-1})_+^p)^{2/p}$. Right: $\Phi_\eta(\mathbf{R}_{t-1}) = (1/\eta)\ln(\sum_{i=1}^{N}\exp(\eta R_{i,t-1}))$. We chose $p = 2$ and $\eta = 1$ to plot.

Polynomially Weighted Average Forecaster What $\psi()$ and $\phi()$ are chosen?

$$\Phi_p(\mathbf{R}_{t-1}) = \left(\sum_{i=1}^{N}(R_{i,t-1})_+^p\right)^{2/p} = ||\mathbf{u}_+||_p^2 \tag{3.8}$$

See Figure 1.

Exponentially Weighted Average Forecaster What $\psi()$ and $\phi()$ are chosen?

$$\Phi_\eta(\mathbf{R}_{t-1}) = (1/\eta)\ln\left(\sum_{i=1}^{N} e^{\eta R_{i,t-1}}\right) \tag{3.9}$$

See Figure 1.

Scenario: Calculate $\mathbf{R}_{r-1}$, predict $\hat{p}_t$. Suppose we are at time $t-1$ and we've just received the outcome value y_{t-1}. $\mathbf{R}_{t-1} = (\mathbf{R}_{1,t-1}, ..., \mathbf{R}_{N,t-1})$ is calculated using convex loss function $l(\cdot,\cdot)$. After proceeding to time t with $\mathbf{R}_{t-1}$ and revealed expert values, we can calculate weights and do weighted sum to get prediction $\hat{p}_t$. We receive y_t and calculate $\mathbf{R}t$ again in the same way.

Blackwell condition Lemma 2.1 can be simply written as follows.

$$\sup_{y_t \in \mathcal{Y}} \mathbf{r}_t \cdot \nabla\Phi(\mathbf{R}_{t-1}) \leq 0 \tag{3.10}$$

It is called ***Blackwell condition***.

Interpretation - Regret & potential plot.

* Note that convex loss function is not the only way to achieve Blackwell condition.

Theorem 2.1 *Assume that a forecaster satisfies the Blackwell condition for a potential* $\Phi(\mathbf{u}) = \psi(\sum_{i=1}^{N} \phi(u_i))$. *Then, for all n=1,2,...*

$$\Phi(\mathbf{R}_n) \leq \Phi(\mathbf{0}) + \frac{1}{2}\sum_{t=1}^{n} C(\mathbf{r}_t), \tag{3.11}$$

where

$$C(\mathbf{r}_t) = \sup_{\mathbf{u}\in\mathbb{R}^N} \psi'\left(\sum_{i=1}^{N} \phi(u_i)\right)\sum_{i=1}^{N} \phi''(u_i) r_{i,t}^2 \tag{3.12}$$

We will use this to bound the regret of forecaster.

3.2 Exponentially Weighted Average Forecaster

Polynomially weighted average forecaster also gives a good bound too, but slightly worse.

Let's focus on exponentially weighted average forecaster for now. Recall that it is defined as

$$\Phi_\eta(\mathbf{R}_{t-1}) = (1/\eta)\ln\left(\sum_{i=1}^{N} e^{\eta R_{i,t-1}}\right) \tag{3.13}$$

Now the forecaster's prediction becomes

$$\hat{p}_t = \frac{\sum_{i=1}^{N}\exp(\eta(\hat{L}_{t-1} - L_{i,t-1}))f_{i,t}}{\sum_{j=1}^{N}\exp(\eta(\hat{L}_{t-1} - L_{j,t-1}))} = \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1}} f_{i,t}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}} \tag{3.14}$$

Corollary 2.2 *Assume that the loss function l is convex in its first argument and that it takes values in* $[0,1]$. *For any n and* $\eta > 0$, *and for all* $y_1, ..., y_n \in \mathcal{Y}$, *the regret of the exponentially weighted average forecaster satisfies*

$$\hat{L}_n - \min_{i=1..N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2} \tag{3.15}$$

By choosing $\eta = \sqrt{2\ln N/n}$, the upper bound becomes $\sqrt{2n\ln N}$

Proof

Compare this bound with one introduced earlier