E1 245: Online Prediction & Learning	Fall 2015
Lecture 5 — August 18	
Lecturer: Aditya Gopalan	*Scribe: Ravi Ranjan*

5.1 Recap

In the last lecture, we looked at general problem of 1-Bit prediction. We studied weighted majority algorithm for this problem and derived its worst case performance. We setup the problem of prediction-with-expert-advice and defined following:

- Decison space : $\mathscr{D}$
- Outcome space : $\mathscr{Y}$
- Loss function : $l : \mathscr{D} \times \mathscr{Y} \to \mathbb{R}^+$
- Experts : $\mathscr{E}$.

We saw some examples of the problem and defined Regret.

5.2 Experts game with convex losses

For this problme, we'll consider that

1. $\mathscr{D}$ is a convex set in $\mathbb{R}^d$
2. $l(.,y)$ is convex on $\mathscr{D}, \forall y \in \mathscr{Y}$.

Examples of some convex loss function are:

1. $l(p,y) = (p-y)^2$, for $\mathscr{D} = \mathscr{Y} = \mathbb{R}$
2. $l(p,y) = |p-y|$, for $\mathscr{D} = \mathscr{Y} = \mathbb{R}$
3. $l(p,y) = ||p-y||_q$, for $\mathscr{D} = \mathscr{Y} = \mathbb{R}^d$ and $q \geq 1$
4. $l(\pi,y) = log\left(\frac{1}{\pi(y)}\right)$, for $\mathscr{D} = \left\{\pi \in \mathbb{R}^d : \pi_i \geq 0 \; \forall i, \sum_{i=1}^d \pi_i = 1\right\}$, $\mathscr{Y} = \{1,2,3,\ldots,d\}$.

Example of non-convex loss function is:

$$l(p,y) = \mathbb{1}_{p \neq y}(p,y).$$

5.2.1 Exponentially weighted average forecaster

Exponantially weighted average forecaster(EXPTWTS) algorithm is shown in algorithm 1. It is also known as HEDGE or multiplicative weight algorithm. The algorithm takes learning-rate ($\eta \geq 0$) as parameter.

Algorithm 1 EXPWTS(η)

procedure
Parameter :
  $\eta \geq 0$
Initialize:
  $w_{i,1} \leftarrow 1, \forall i \in [N]$
  $t \leftarrow 1$
loop:
  $\hat{p}_t \leftarrow \frac{\sum_{i \in [N]} w_{i,t} f_{i,t}}{\sum_{i \in [N]} w_{i,t}}$
  See y_t
  $w_{i,t+1} \leftarrow w_{i,t} e^{-\eta l(f_{i,t}, y_t)}$

For analysis of algorithm 1, let's define the following functions

$$\hat{L}_t = \sum_{t=1}^{T} l(\hat{p}_t, y_t)$$

$$L_{i,T} = \sum_{t=1}^{T} l(f_{i,t}, y_t) \forall i \in \mathscr{E}.$$

Theorem 5.1. *If $\mathscr{D}$ is convex, $l(p,y)$ is convex on $\mathscr{D}$ and $l : \mathscr{D} \times \mathscr{Y} \to [0,1]$, then regret of algorithm 1 can be bounded by*

$$R_T(EXPWTS(\eta)) \leq \frac{\log|\eta|}{\eta} + \frac{\eta T}{8}$$

Proof: Let $|\mathscr{E}| = N$. Define the potential function

$$\phi_t = \frac{1}{\eta} \log w_t = \frac{1}{\eta} \log \sum_{i=1}^{N} w_{i,t} = \frac{1}{\eta} \log \sum_{i=1}^{N} e^{-\eta L_{i,t-1}}.$$

We have,

$$\begin{aligned}\phi_{T+1} - \phi_1 &= \frac{1}{\eta}\log\left(\frac{w_{T+1}}{w_1}\right)\\ &= \frac{1}{\eta}\log\left(\frac{\sum_{i=1}^{N} e^{-\eta L_{i,T}}}{N}\right)\\ &\geq \frac{1}{\eta}\log\left(\frac{\max_{i\in[N]} e^{-\eta L_{i,T}}}{N}\right)\end{aligned}$$

$$= -\min_{i\in N} L_{i,T} - \frac{\log N}{\eta}. \tag{5.1}$$

On the other hand, the per step change in potential is,

$$\begin{aligned}\phi_t - \phi_{t-1} &= \frac{1}{\eta}\log\frac{w_t}{w_{t-1}}\\ &= \frac{1}{\eta}\log\left(\frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-2}} e^{-\eta l(f_{i,t-1},y_{t-1})}}{\sum_{i=1}^{N} e^{-\eta L_{i,t-2}}}\right)\end{aligned}$$

$$= \frac{1}{\eta}\log\left(\sum_{i=1}^{N} q_i e^{-\eta l(f_{i,t-1},y_{t-1})}\right) \tag{5.2}$$

where,

$$q_i = \frac{e^{-\eta L_{i,t-2}}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-2}}} \geq 0.$$

So,

$$\sum_{i=1}^{N} q_i = 1.$$

Equation 5.2 can also be written in terms of expectation,

$$\phi_t - \phi_{t-1} = \frac{1}{\eta}\log\mathbb{E}\left[e^{-\eta l(f_{I,t-1},y_{t-1})}\right] \tag{5.3}$$

where, I is a random variable realization of i and q_i can be thought of as $\mathbb{P}(I = i)$. Applying Hoeffding's Lemma, stated in appendix A, on 5.3,

$$\phi_t - \phi_{t-1} \leq -\mathbb{E}\left[l(f_{I,t}, y_{t-1})\right] + \frac{\eta}{8}$$

$$\leq -l(\underbrace{\mathbb{E}[f_{I,t-1}]}_{\sum_{i=1}^{N} q_i f_{i,t-1} = \hat{p_{t-1}}}, y_{t-1}) + \frac{\eta}{8} \tag{5.4}$$

$$= -l(\hat{p_{t-1}}, y_{t-1}) + \frac{\eta}{8}. \tag{5.5}$$

The inequality 5.4 is due to Jensen's inequality, since $l(p,y)$ is a convex funtion by assumption of theorem. Summing equation 5.5 across $t = 2,3\ldots T+1$,

$$\phi_{T+1} - \phi_1 \leq -\sum_{t=1}^{T} l(\hat{p}_t, y_t) + \frac{\eta T}{8}. \tag{5.6}$$

Putting 5.6 and 5.1 together,

$$\hat{L_T} - \min_{i \in N} L_{i,T} \leq \frac{\eta T}{8} + \frac{\log N}{\eta}.$$

□

Note:

1. If $\eta = \sqrt{\frac{8 \log |\mathscr{E}|}{T}}$, then bound on regret is

$$R_T(EXPWTS) \leq \sqrt{\frac{T}{2} \log |\mathscr{E}|}. \tag{5.7}$$

   Bound 5.7 is tight.

2. Optimal value of η requires knowing T in advance. But algorithm 1 can be tweaked to get bound that holds uniformly over time. This is also called the 'doubling trick'. The bound in this case will be,

$$R(EXPWTS\prime) \leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\frac{T}{2} \log |\mathscr{E}|}.$$

Appendix

A Hoeffding's Lemma

Let X be a random variable with $a \leq X \leq b$, then $\forall z \in \mathbb{R}$

$$\log \mathbb{E}\left[e^{zx}\right] \leq z\mathbb{E}[X] + \frac{z^2}{8}(b-a)^2.$$

B Jensen's inequality

Let K be a convex set and X be a random variable, which always takes values from K. If $f : k \to \mathbb{R}$ is a convex function, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}f(X).$$

Bibliography

[1] Nicoló Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning and Games*. Cambridge University Press. 2006