EECS598: Prediction and Learning: It's Only a Game Fall 2013

Lecture 4: Action Setting of Online Learning

Prof. Jacob Abernethy *Scribe: S. Abu-El-Haija & C. Saint Croix TEX: E. Brinkman*

Announcements

- Lecture Videos are on CTools
- You can reach the GSI at email eecs598-plg@umich.edu

4.1 Recap: Online prediction and combining advice

- Here, we have N experts
- For every time step $t = 1, \dots, T$, each expert says $f_1^t, \dots f_N^T$
- An algorithm then combines the advice as a weighted average:

$$\hat{y}_t = \frac{\sum_{i=1}^N w_i^t f_i^T}{\sum_{i=1}^N w_i^t}$$

- Algorithm then observes outcome y_t, and suffers loss $\ell(\hat{y}_t, y_t)$

4.2 Sequential Decision Making / Game Playing (a.k.a *Hedge setting*)

- On each round, algorithm chooses action $i \in [N]$.
- Alternatively algorithm choses a distribution $p^t \in \triangle_N$
- Nature chooses $\ell^t \in [0,1]^N$
- Algorithm suffers expected loss $p^t \cdot \ell^t = \sum_{i=1}^N p_i^t \ell_i^t$

4.3 Predictive Setting vs. Action Setting

	Predictive Setting	Action Setting
Loss	$\ell(f_i^t, y_t)$	ℓ_i^t
Weight	$w_i^t = exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_i^s, y_s)\right)$	$w_i^t = exp\left(-\eta \sum_{s=1}^{t-1} \ell_i^s\right)$
Play	$\hat{y}_t = \frac{\sum_{i=1}^N w_i^t f_i^T}{\sum_{i=1}^N w_i^t}$	$p^t = \left\langle \frac{w_1^t}{\sum_j w_j^t}, \frac{w_2^t}{\sum_j w_j^t}, \dots, \frac{w_N^t}{\sum_j w_j^t} \right\rangle$
Algorithm Loss	$\ell(\hat{y}_t, y_t) = \ell\left(\frac{\sum_{i=1}^N w_i^t f_i^T}{\sum_{i=1}^N w_i^t}, y_t\right)$	$p_t \cdot \ell_t = \frac{\sum_i w_i^t \ell_i^t}{\sum_i w_i^t}$

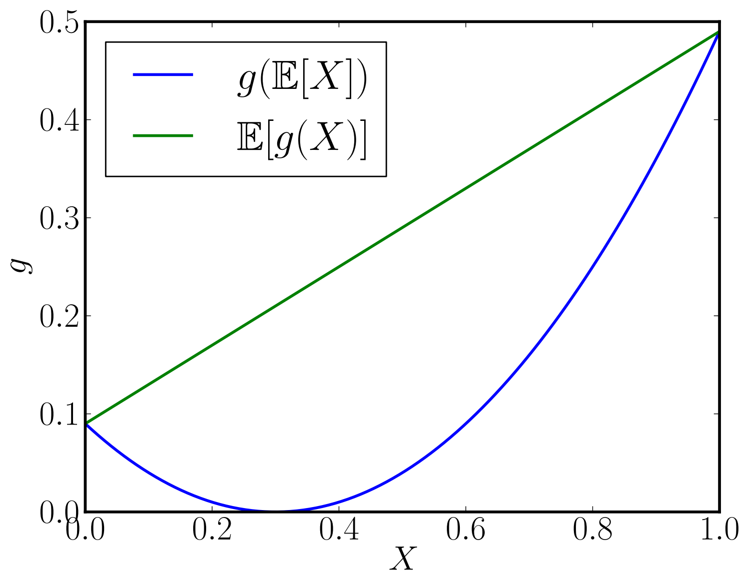

Figure 1: Jensen's Inequality

4.4 Finding a Bound on the Loss of EWA

We need the following preliminaries to do this derivation:

4.4.1 Jensen's Inequality

Jensen's inequality follows from (/is) the definition of convexity. For a convex function g:

$$\mathbb{E}\left[g(X)\right] \geq g\left(\mathbb{E}[X]\right).$$

Jensen's inequality can be visualized with the graph in Figure 1. Imagine a Bernoulli random variable with parameter p, and the convex function g. $\mathbb{E}[g(X)] = (1-p)g(0) + pg(1)$, which is the line between $g(0)$ and $g(1)$. $g(\mathbb{E}[X]) = g(p)$, which is just the convex function g in the range $[0, 1]$.

The bottom row of the Prediction vs. Action (4.3) table can be viewed with Jensen's inequality when

$$X = f_i^t \qquad\qquad i \sim \left\langle \frac{w_i^t}{\sum_j w_j^t} \right\rangle \qquad\qquad g(x) = \ell(x, y_t).$$

When phrased like this, the loss in the prediction setting is equivalent to $g(\mathbb{E}[X])$, and the loss in the action setting is equivalent to $\mathbb{E}[g(X)]$. When viewed in this light, the loss of the prediction setting is bounded by the loss in the action setting.

4.4.2 Other Useful Lemmas

The following identity can be used to find a tighter bound:

Lemma 4.1.

$$\log \mathbb{E}[e^s X] \leq (e^s - 1)\,\mathbb{E}[X]; \textit{ for } X \in [0, 1][1].$$

The following identity provides a less tight bound, but it is easier to work with (we will use this in the upcoming derivation)

Lemma 4.2.

$$\log \mathbb{E}[e^s X] \leq s\mathbb{E}[X] + \frac{s^2}{8}; \textit{for } X \in [0,1].$$

4.4.3 Loss Bound

First lets bound $-\log w^{t+1} + \log w^t$ where $w^t = \sum_j w_j^t$

$$\begin{aligned} -\log w^{t+1} + \log w^t &= -\log \frac{w^{t+1}}{w^t} \\ &= -\log \left(\frac{\sum_i w_i^t exp\left(-\eta \ell_i^t\right)}{\sum_i w_i^t} \right) \\ &\geq \frac{\eta \sum_i w_i^t \ell_i^t}{\sum_i w_i^t} - \frac{\eta^2}{8} \end{aligned}$$

The inequality at the end is a result of Lemma 4.2.

Finally, we'll bound the total loss (L_{MA}^T) with the loss of a single expert i (L_i^T).

$$\begin{aligned} \eta L_i^T + \log N &= -\log\left(e^{-\eta L_i^T}\right) + \log(N) \\ &\geq -\log\left(w^{T+1}\right) + \log(w^1) \\ &= \sum_{t=1}^{T} -\log(w^{t+1}) + \log(w^t) \\ &\geq \sum_{t=1}^{T} \eta \frac{\sum_i w_i^t \ell_i^t}{\sum_i w_i^t} - \frac{\eta^2}{8} \\ &= \eta L_{MA}^T - \frac{T\eta^2}{8} \\ L_{MA}^T &\leq L_i^T + \frac{\log N}{\eta} + \frac{T\eta}{8} \end{aligned}$$

4.4.4 Tuning η

Example:

$$\min_\eta \frac{A}{\eta} + B\eta = 2\sqrt{AB}$$

A general trick is that the minimum of additive convex functions is achieved when all of the terms are equal.

[1] This is bad notation for saying the support of X is $[0,1]$.

Lets tune our bound on loss.

$$L_{MA}^T \leq L_i^T + \frac{\log N}{\eta} + \frac{T\eta}{8} \qquad A = \log N$$

$$L_{MA}^T \leq L_i^T + \underbrace{O\left(\sqrt{T \log N}\right)}_{\text{Regret}_T}. \qquad B = \frac{T}{8}$$

Note, that our regret is sublinear in T. In other words

$$\frac{L_{MA}^T}{T} - \frac{L_i^T}{T} = o(1).$$

This is also known as the "no regret property" or hannan consistency[citation needed].

We can also use Lemma 4.1 to get a similar result for any expert i:

Theorem 4.3. *For any expert i*

$$L_{MA}^T \leq L_i^T + O\left(\sqrt{L_i^T \log N}\right)$$

sketch.

$$\begin{aligned} L_{MA}^T &\leq \frac{1}{1-e^{-\eta}}\left(\eta L_i^T + \log N\right) \\ &\approx (1+\eta)L_i^T + \frac{\log N}{\eta} \\ &= L_i^T + \underbrace{\eta L_i^T}_{\leq \eta T} + \frac{\log N}{\eta} \\ &= L_i^T + O\left(\sqrt{L_i^T \log N}\right) \end{aligned}$$

□

4.4.5 Non-uniform prior

Theorem 4.4. *For any expert i, and an initial prior over experts* $\mathbf{p} \in \triangle_N$

$$L_{MA}^T \leq \frac{1}{1-e^{-\eta}}\left(\eta L_i^T + \log \frac{1}{p_i}\right)$$

4.5 Hyperexperts

A hyper expert is a path through experts.

$$\mathcal{I} := [N]^T = [N] \times [N] \times \ldots \times [N]$$
$$\mathbf{i} \in \mathcal{I} \text{ is a hyper expert}$$

We define EWA variables on hyper experts as such:

$$\ell_{\mathbf{i}}^t := \ell_{i_t}^t,$$
$$w_{\mathbf{i}}^t := \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i_s}^s\right),$$
$$v_i^t := \sum_{\mathbf{i}\in\mathcal{I}:\mathbf{i}_t=i} w_i^t.$$

Apply the EWA bound to hyper experts:

$$L_{MA}^T \leq \frac{1}{1-e^{-\eta}}\left(\eta L_{\mathbf{i}}^T + T\log N\right).$$

This SUCKS! Scales linearly with T $\implies$ we're not learning.