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

Lecture 5: Fixed Share Forecaster

Prof. Jacob Abernethy *Scribe: Catherine Saint Croix, TEX: Erik Brinkman*

Announcements

- Office Hours Friday 1:30 - 3:15
- Bug in Homework #1 fixed

5.1 Regret

In a prediction setting we defined regret as the difference between the loss of the learning algorithm, and the loss of the best expert

$$\text{Prediction Regret} = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \min_i \sum_{t=1}^{T} \ell(f_i^t, y_t).$$

We can define a similar notion of regret for the action setting where in each round we pick a mixture over N actions $\mathbf{p}^t \in \triangle_N$ and each action has an associated cost for that round $\boldsymbol{\ell}^t \in [0,1]^N$. We can define the regret in this setting as the difference between our algorithm and the best action as

$$\text{Action Regret} = \sum_{t=1}^{T} \mathbf{p}^t \cdot \boldsymbol{\ell}^t - \underbrace{\min_i \sum_{t=1}^{T} \ell_i^t}_{\text{comparator}}.$$

Note that these settings are really two different ways too look at the same problem.

Question: Why should we care about a fixed comparator?

What if instead we tried to minimize

$$\text{Better Regret?} = \sum_{t=1}^{T} \mathbf{p}^t \cdot \boldsymbol{\ell}^t - \underbrace{\min_{i_1,\ldots,i_T} \sum_{t=1}^{T} \ell_i^t}_{\sum_{t=1}^{T} \min_i \ell_i^t}.$$

5.2 Exponential Weights Algorithm with Hyper Experts

Define an expert for every possible sequence of experts

$$\mathcal{I} = [N] \times [N] \times \ldots \times [N] = [N]^T$$
$$\mathbf{i} \in \mathcal{I}, \mathbf{i} = (i_1, \ldots, i_T).$$

On round t, hyper expert $\mathbf{i}$'s advice is to follow the advice of expert i_t.

Definition 5.1. $\tilde{\ell}_{\mathbf{i}}^t := \ell_{i_t}^t$, *the loss of a hyper expert on round t is the same as the recommended normal expert.*

Definition 5.2. $\tilde{w}_{\mathbf{i}}^t := \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{\mathbf{i}}^t\right)$, *the weight on a hyper expert follows the EWA.*

Definition 5.3. $v_j^t := \sum_{\mathbf{i} \in \mathcal{I}: i_t = j} \tilde{w}_{\mathbf{i}}^t$, *the total weight on a piece of advice at time t is the sum of all hyper expert weights that chose that advice.*

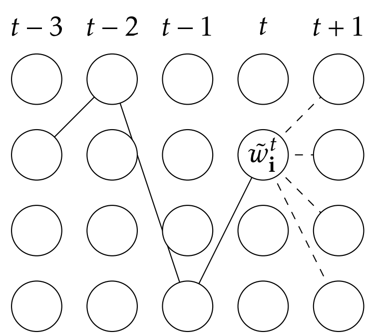
Exponential Weights on Hyper Experts

Play mixture

$$\mathbf{p}^t := \left\langle \frac{v_1^t}{\sum_j v_j^t}, \ldots, \frac{v_N^t}{\sum_j v_j^t} \right\rangle$$

Claim 5.4. $\mathbf{p}^t$ *is uniform for all t!*

Proof. Proof by picture. Each row represents a different expert, and the path represents a hyper expert. by time t each hyper expert that took the solid path has weight $\tilde{w}_{\mathbf{i}}^t$. When trying to predict for time $t+1$, each normal expert gains weight $\tilde{w}_{\mathbf{i}}^t$ from the hyper experts that shared a common history up to time t. Therefore very expert at time $t+1$ has the same weight, implying $\mathbf{p}^t$ is uniform.



□

Question: Does this hold with a prior? Yes, for every time but the initial one.

We can still use our regret bound to calculate the regret from using this algorithm.

$$\begin{aligned}
\text{Regret} &\le c\sqrt{T \log(\#\text{ of hyper experts})} \\
&= c\sqrt{T \log(N^T)} \\
&= c\sqrt{T T \log N} \\
&= cT\sqrt{\log N}
\end{aligned}$$

Because loss is bounded at 1 for each round, total loss is bounded by T, and therefore from the problem definition

$$\text{Regret} \le T.$$

Therefore the above bound on regret is meaningless.

5.2.1 Better Prior on Hyper Experts

Generate a random hyper expert sequence with parameter $\alpha \in (0,1)$.

(1) Sample i_1 u.a.r. (uniformly at random)

(2) For $t = 2,\ldots,T$

$$i_t = \begin{cases} i_{t-1} & \text{w.p. } 1-\alpha \\ j & \text{w.p. } \frac{\alpha}{N-1} \forall j \neq i_{t-1} \end{cases}$$

Thus, the prior over all of these sequences is

$$\Pi(\mathbf{i}) = \frac{1}{N}\left(\frac{\alpha}{N}\right)^{\sigma(\mathbf{i})}(1-\alpha)^{T-\sigma(\mathbf{i})-1}$$

where

$$\begin{aligned}\sigma(\mathbf{i}) &= \#\text{switchs in } \mathbf{i} \\ &= |\{t \in [T] : i_{t+1} \neq i_t\}|\end{aligned}$$

Straightforward Exercise: Show $\sum_{\mathbf{i}} \Pi(\mathbf{i}) = 1$.

5.2.2 Fixed Share Forecaster

$$\tilde{w}_{\mathbf{i}}^t = \underbrace{\Pi(\mathbf{i})}_{\tilde{w}_{\mathbf{i}}^1} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i_s}^s\right)$$

Bound:

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

where

$$\begin{aligned}\log\frac{1}{\Pi(\mathbf{i})} &= \log N + \sigma(\mathbf{i})\log\frac{N-1}{\alpha} + (T-\sigma(\mathbf{i})-1)\log\frac{1}{1-\alpha} \\ &\leq (\sigma(\mathbf{i})+1)\left(\log N + \log\frac{1}{\alpha}\right) + T\log\frac{1}{1-\alpha} \\ \alpha &= \frac{1}{T} \quad \text{*Jake Magic*} \\ \log\frac{1}{\Pi(\mathbf{i})} &\leq (\sigma(\mathbf{i})+1)(\log N + \log T) + T\log\frac{1}{1-\frac{1}{T}}\end{aligned}$$ [1]

If we limit the number of switches an expert makes to k then the regret bound becomes

$$\text{Regret}_T \leq \frac{\eta T + (k+1)(\log N + \log T) + O(1)}{\eta}.$$

After tuning η we get

$$\text{Regret}_T \leq O\left(\sqrt{T(k+1)(\log N + \log T)}\right)$$

Compare this bound to if we cut the time up into k segments and ran standard exponential weights over each segment ($T' := \frac{T}{k}$).

$$\begin{aligned}\text{Regret}_T &\leq \sum_{i=1}^{k} \sqrt{T' \log N} \\ &= k\sqrt{\frac{T}{k}\log N} \\ &= \sqrt{Tk\log N}\end{aligned}$$

This bound is very close to the bound on our modified hyper expert algorithm, except our hyper expert algorithm allows k switches anywhere, not just at predefined points.

5.3 Efficient Modified EWA

$$\begin{aligned}\hat{w}_i^{t+1} &= w_i^t \exp\left(-\eta \ell_i^t\right) \\ w_i^{t+1} &= (1-\beta)\frac{\hat{w}_i^{t+1}}{\sum_j w_j^{t+1}} + \beta\frac{1}{N}\end{aligned}$$

Weight on an expert never gets below $\frac{\beta}{N}$, so it's easier for individual experts to recover if they begin to make good predictions.

Easy Challenge: Show that this algorithm is the same as EWA on hyper experts with prior Π. What is the value of $\beta \in (0,1)$?

weight on expert i: $$v_j^t = \sum_{\mathbf{i} \in \mathcal{I}: i_t = j} \Pi(\mathbf{i}) \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i_s}^s\right)$$

key idea: $$v_j^{t+1} = \left(1 - \alpha - \frac{\alpha}{N}\right) v_j^t \exp\left(-\eta \ell_j^t\right) + \frac{\alpha}{N-1}\sum_i v_i^t \exp\left(-\eta \ell_i^t\right)$$

[1] $\left(\frac{1}{1-\frac{1}{T}}\right)^{-T} \approx e$