10-716: Advanced Machine Learning	Spring 2019

Review Session: April 30

Lecturer: Pradeep Ravikumar *Scribes: Scribes: Raphael Olivier, Yuyan Wang*

25.1 An Upper Bound for Cumulative Regret

This is a review session. In this session, we discuss Learning with Expert Advice, which is not covered in the homework. We are going to discuss a special case where we develop upper bounds on the cumulative regret.

25.1.1 Recap of Learnning With Expert Advice

We use $R_{E,n}$ to denote the cumulative regret for an expert E.

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

The forecastor's goal is to minimize $R_{E,n}$. The instantaneous regret at time t is defined as:

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

As a result, we have $R_{E,n} = \sum_{t=1}^{n} r_{E,t}$.

Assume the forecastor uses *Weighted Average Prediction*, i.e., at every step, we let

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

As in class, We define a *Potential Function*: $\Phi : \mathbb{R}^N \mapsto \mathbb{R}$. Formally,

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

where $\mathbf{u}$ is a vector, $\phi(\cdot)$ is a non-negative, increasing, twice-differentiable function; $\psi(\cdot)$ is a non-negative, strictly increasing, concave, twice-differentiable function. Then we let $w_{j,t-1} = \nabla_j \Phi(R_{t-1})$. Then, we have the *Rockwell Condition*:

$$\sup_{y_t \in Y} \sum_{i=1}^{N} r_{i,t} \nabla_i \Phi(R_{t-1}) \leq 0$$

If this condition is satisfied, we have:

$$\Phi(R_n) \leq \Phi(0) + \frac{1}{2} \sum_{t=1}^{n} c(r_t)$$

with $c(r_t) = \sup_{\mathbf{u}\in\mathbb{R}^N} \psi'(\sum_{i=1}^N \phi(u_i)) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2$. And if $\phi(\cdot)$ is invertible, we have:

$$\max_i R_{i,n} \le \phi^{-1}\psi^{-1}(\Phi(R_n))$$

25.1.2 Developing the Upper Bound

In this subsection, we focus on the *Exponentially Weighted Average Forecastor*, which uses weight:

$$w_{i,t-1} = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^N e^{\eta R_{j,t-1}}}$$

Since $w_{i,t-1} = \nabla_i \Phi(R_{t-1})$, we set $\Phi_\eta(\mathbf{u}) = \frac{1}{\eta}\ln(\sum_{i=1}^N e^{\eta u_i})$. We want to show the following inequality:

$$\hat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \frac{\ln N}{\eta} + \frac{n\eta}{2}$$

Let $\phi(x) = e^{\eta x}$ and $\psi(x) = \frac{1}{\eta}\ln x$. Then,

$$\begin{aligned} c(r_t) &= \sup_{\mathbf{u}\in\mathbb{R}^N} \psi'(\sum_{i=1}^N \phi(u_i)) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2 \\ &= (\frac{1}{\eta} \cdot \frac{1}{\sum_{i=1}^N e^{\eta u_i}} \eta^2 \cdot \sum_{i=1}^N [e^{\eta u_i} \cdot r_{i,t}^2]) \end{aligned}$$

Apply Hlder's inequality and we get $\sum_{i=1}^N [e^{\eta u_i} \cdot r_{i,t}^2] \le \eta^2 (\sum_{i=1}^N e^{\eta u_i}) \max_{i=1,\dots,N} r_{i,t}^2$. Putting this together, we have $c(r_t) \le \eta \max_{i=1,\dots,N} r_{i,t}^2 \le \eta$ since we assume $r_{i,t} \in [0,1]$. Plugging $c(r_t)$ in, we reach the conclusion:

$$\Phi_\eta(R_n) \le \frac{\ln N}{\eta} + \frac{n\eta}{2}$$

25.2 A refined inequality

The Blackwell condition actually gives rarely an optimal result. Now we will use a different method to prove the following result :

Lemma 25.1 *Under the previous notations,*

$$\hat{L}_n - \min_{1\le i\le N} \frac{\ln(N)}{\eta} + \frac{n\eta}{8}$$

Proof: By simply dividing the weights by a constant factor, we can redefine

$$w_{i,t} = e^{-\eta L_{i,n}}$$

$$\hat{p}_{i,t} = \frac{\sum_{i=1}^N w_{i,t} R_{i,t}}{\sum_{j=1}^N w_{j,t}}$$

We also define $W_t = \sum_{i=1}^N w_{i,t}$. In particular, $W_0 = N$ and $W_n = \sum_{i=1}^N e^{-\eta L_{i,n}}$

On the one hand,

$$\begin{aligned}\ln(\frac{W_n}{W_0}) &= \ln(\sum_{i=1}^{N} e^{-\eta L_{i,n}}) - \ln(N) \\ &\geq \ln(\max_i e^{-\eta L_{i,n}} - \ln(N) \\ &= -\eta \min_i L_{i,n} - \ln(N)\end{aligned}$$

On the other hand,

$$\begin{aligned}\ln(\frac{W_t}{W_{t-1}}) &= \ln(\frac{\sum_{i=1}^{N} e^{-\eta L_{i,t}}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}) \\ &= \ln(\frac{\sum_{i=1}^{N} e^{-\eta(L_{i,t-1}+l(f_{i,t},y_t))}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}) \\ &= \ln(\frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1}} e^{-\eta l(f_{i,t},y_t)}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}) \\ &= \ln(\frac{\sum_{i=1}^{N} w_{i,t-1} e^{-\eta l(f_{i,t},y_t)}}{\sum_{j=1}^{N} w_{j,t-1}}) \\ &= \ln(\mathbb{E}[e^{-\eta X}])\end{aligned}$$

where X is a discrete random variable defined by

$$\mathbb{P}[X = l(f_{i,t}, y_t)] = \frac{w_{i,t-1}}{\sum_{j=1}^{N} w_{j,t-1}}$$

We can use Hoeffding's lemma : if $a \leq X \leq b$ then

$$\ln(\mathbb{E}[e^{sX}]) \leq s\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}$$

With $s = -\eta$, $a = 0$ and $b = 1$:

$$\begin{aligned}\ln(\frac{W_t}{W_{t-1}}) &\leq -\eta \frac{\sum_{i=1}^{N} w_{i,t-1} l(f_{i,t}, y_t)}{\sum_{i=1}^{N} w_{i,t-1}} + \frac{\eta^2}{8} \\ &\leq -\eta l(\frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{i=1}^{N} w_{i,t-1}}, y_t) + \frac{\eta^2}{8} = l(\hat{p}_t, y_t)\end{aligned}$$

using Jensen's inequality. By induction we get :

$$\ln(\frac{W_n}{W_0}) \leq -\eta \hat{L}_{n,t} + \frac{n\eta^2}{8}$$

We can combine our two inequalities involving $\ln(\frac{W_n}{W_0})$:

$$-\eta \min_i L_{i,n} - \ln(N) \leq -\eta \hat{L}_{n,t} + \frac{n\eta^2}{8}$$

Which can be rearanged into

$$\hat{L}_n - \min_{1 \leq i \leq N} \frac{\ln(N)}{\eta} + \frac{n\eta}{8}$$

∎