CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Aditi Dudeja

Lecture # 11
17th October 2017

In the previous class, we had looked at the RMW algorithm. In this setting, the actions/decisions are referred to as experts. Assume that we have N experts. On any given day t, we choose an expert to use from among the N experts (like choosing a function $h \in H$). The adversary/world returns a loss vector $l^t = (l_1^t, \cdots l_N^t)$. If the expert i_t is chosen on day t, then the performance after T days is $\frac{1}{T}\sum_{t=1}^{T} l_{i_t}^t$. Whereas, the performance of the best expert is $\min_{j\in[N]} \frac{1}{T}\sum_{t=1}^{T} l_j^t$. We had proved the following theorem in the previous class:

Theorem 1. *Let M be the loss incurred by RMW at the end of day T and m be the loss of the best action, then:*

$$\frac{\mathbb{E}[M]}{T} \leq \frac{m}{T} + 2\gamma\sqrt{\frac{\log N}{T}} \tag{1}$$

where γ is the range of the loss. That is, for all $t \geq 0$ and $i \in [N]$, $l_i^t \in [0, \gamma]$.

We then discussed the **bandit setting**, which can be described as follows:

1. The algorithm has a list of N experts.
2. On day t, the algorithm will pick expert i_t.
3. Loss $l_{i_t}^t$ is revealed.

While in the previous learning problems, algorithms had access to the loss of every expert, bandit setting only reveals loss of the expert that the algorithm chose. We will show that even in this setting, it is possible to achieve no regret. We do this using the EXP3 algorithm that is described in Figure **??**.

EXP3 Algorithm

EXP3 can be seen as a modification of RMW. Instead of choosing experts according to the distribution output by RMW, we instead choose them using a perturbed distribution. Let $P^t = (p_1^t, \cdots, p_N^t)$ be the distribution output by RMW, and let $U_N = (\frac{1}{N}, \cdots, \frac{1}{N})$ be the uniform distribution. Let $Q^t = \gamma P^t + (1-\gamma)U_N$. The probability that EXP3 picks out expert i at time t is therefore, $q_i^t = (1-\gamma)p_i^t + \gamma(\frac{1}{N})$.

At time t, if the EXP3 algorithm picks out the expert i_t, then it gets back the loss $l_{i_t}^t$ of this expert. It then sends back the loss vector $\hat{l}^t$ to the RMW, where $\hat{l}^t = (\hat{l}_1^t, \cdots, \hat{l}_N^t)$. For $i \in [N]$, $\hat{l}_i^t$ is defined as follows:

$$\hat{l}_j^t = \left\{ \begin{array}{ll} 0, & \text{for } j \neq i_t \\ \frac{l_{i_t}^t}{q_{i_t}^t}, & \text{for } j = i_t \end{array} \right\}$$

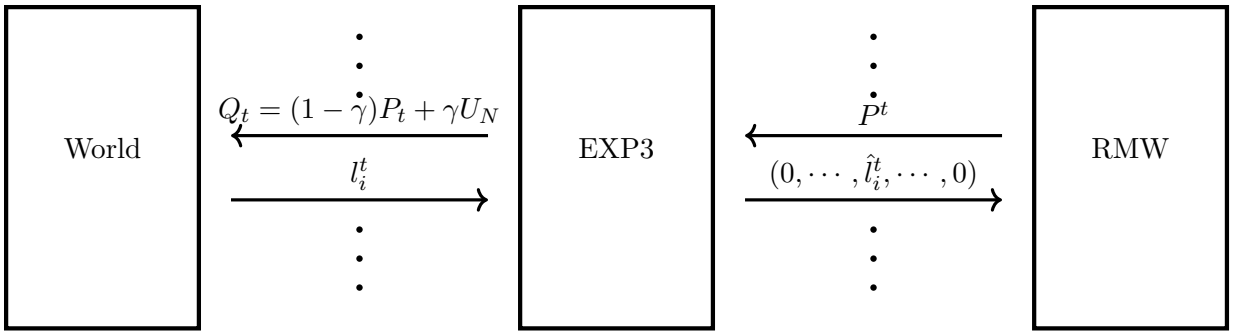

Figure 1: EXP3

Why Do We Scale?

At time t, $\hat{l}_j^t$ is a random variable that depends on the choices made by the algorithm EXP3. So, we need to correct for the fact that we are not choosing by the distribution output by RMW, but instead by Q_t. This is evident because:

$$\begin{aligned} \mathbb{E}[\hat{l}_j^t] &= 0(1 - q_j^t) + (\frac{q_j^t . l_j^t}{q_j^t}) \\ &= l_j^t \end{aligned} \tag{2}$$

In the last class, we had stated the following theorem about the bandit setting:

Theorem 2. *For any $T > 0$, assume $l_i^t \in [0, 1]$, then there is a perturbation γ for which $\mathbb{E}[Regret(EXP3)] \leq O(\sqrt{\frac{N \log N}{T}})$.*

However, to prove the no regret guarantee, it is sufficient to prove the following weaker statement:

Theorem 3. *For any $T > 0$, assume $l_i^t \in [0, 1]$, then there is a perturbation γ for which $\mathbb{E}[Regret(EXP3)] \leq 2(\frac{N^2 \log N}{T})^{\frac{1}{4}}$*

Proof. The proof strategy, also described in **??** will be as follows: We will think of EXP3 as the world. That is, RMW on day t chooses an expert with probability distribution $P_t = (p_1^t, \cdots, p_N^t)$. The EXP3 algorithm then returns it the loss vector $\hat{l}_t$. For RMW, we have the following guarantee as seen in the last lecture:

$$\frac{1}{T} \sum_{t=1}^{T} P_t . \hat{l}_t \leq \min_{j \in [N]} \sum_{t=0}^{T} \hat{l}_j^t + 2\gamma' \sqrt{\frac{\log N}{T}} \tag{3}$$

Here, γ' is the range of the loss $\hat{l}^t$. Further, we know that $\gamma' \leq \frac{N}{\gamma}$. Substituting this in the above equation, we have:

$$\frac{1}{T} \sum_{t=1}^{T} P_t . \hat{l}_t \leq \min_{j \in [N]} \sum_{t=0}^{T} \hat{l}_j^t + \frac{2N}{\gamma} \sqrt{\frac{\log N}{T}} \tag{4}$$

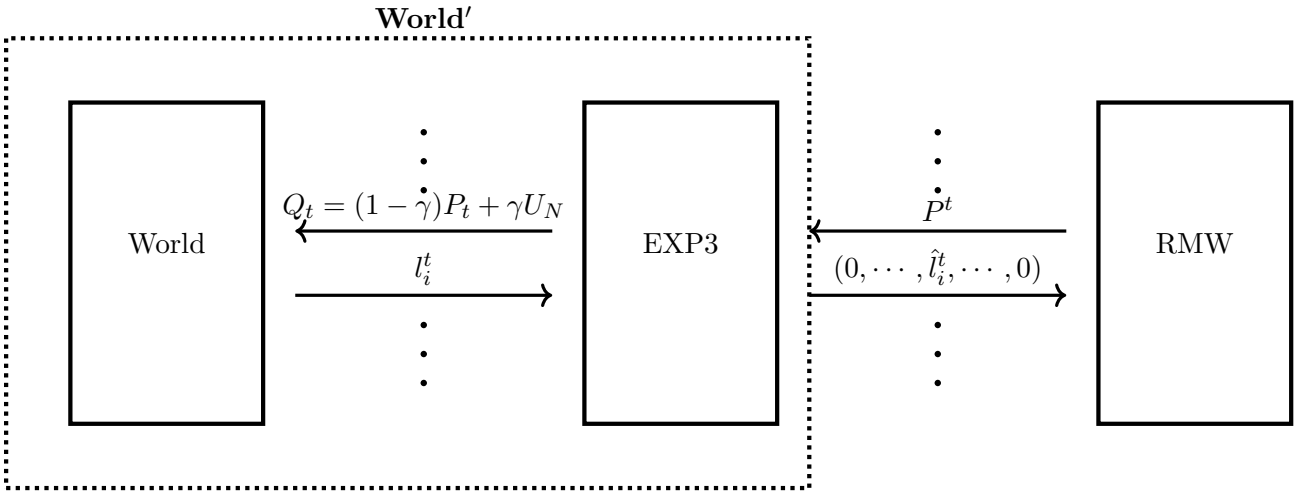

Figure 2: Analysis

Taking expectation on both sides, we have

$$\frac{1}{T}\sum_{t=0}^{T} P^t.l^t \leq \min_{j\in[N]} \sum_{t=0}^{T} l_j^t + \frac{2N}{\gamma}\sqrt{\frac{\log N}{T}} \tag{5}$$

It follows that:

$$\frac{1-\gamma}{T}\sum_{t=0}^{T} P^t.l^t \leq \min_{j\in[N]} \sum_{t=0}^{T} l_j^t + \frac{2N}{\gamma}\sqrt{\frac{\log N}{T}} \tag{6}$$

Adding $\frac{\gamma}{T}\sum_{t=1}^{T} U_N.l^t$ to both sides, we have:

$$\frac{1}{T}\sum_{t=0}^{T} Q^t.l^t \leq \min_{j\in[N]} \sum_{t=0}^{T} l_j^t + \frac{2N}{\gamma}\sqrt{\frac{\log N}{T}} + \frac{\gamma}{T}\sum_{t=1}^{T} U_N.l^t \tag{7}$$

Note that $\frac{\gamma}{T}\sum_{t=1}^{T} U_N.l^t \leq \gamma$. It follows that:

$$\frac{1}{T}\sum_{t=0}^{T} Q^t.l^t \leq \min_{j\in[N]} \sum_{t=0}^{T} l_j^t + \frac{2N}{\gamma}\sqrt{\frac{\log N}{T}} + \gamma \tag{8}$$

Choosing $\gamma = \sqrt{2N}(\frac{\log N}{T})^{\frac{1}{4}}$, our claim follows. □

Online Learning With Large N

The fact that N must be finite is a crucial assumption in the previous setting. However, in many practical scenarios, N can be very large or even infinite. Can we in such a scenario, still achieve a no-regret guarantee? The answer is yes.

Problem 1. *Imagine the following repeated game. In each round* $t = 1, 2, 3, \cdots, T$

1. *S is the set of N experts, which predict values between* 0 *and* 1.
2. *On day t, pick expert* i_t *which predicts the value* $y_{i_t} \in [0, 1]$.
3. *The adversary then reveals the true label* y_t^* *and the loss* $(y_{i_t} - y_t^*)^2$

For the above problem, we want to design a no-regret algorithm. Consider the offline version of this problem, wherein we knew the sequence $y_1, \cdots, y_T$ beforehand. In this case, we would choose a point that minimizes the total loss:

$$y_{best} = \text{argmin}_{y \in [0,1]} \sum_{t=1}^{T} (y - y_t^*)^2 \tag{9}$$

It is easy to see that in this case,

$$y_{best} = \frac{1}{T} \sum_{t=1}^{T} y_t^* \tag{10}$$

We suggest an algorithm for the online version, **Follow the Leader**. This algorithm at time t minimizes the loss on points $y_1, y_2, \cdots, y_{t-1}$ seen so far. That is at time t, predict:

$$y_{t-1}^b = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i^* \tag{11}$$

We make the following claim:

Theorem 4. *For the above regression problem, Regret(FTL)=* $O(\frac{\log T}{T})$

In order to prove the claim above, we will use a hypothetical algorithm, called **Be The Leader**. This algorithm at time t, predicts:

$$y_t^{++} = \frac{1}{t} \sum_{i=1}^{t} y_i^* \tag{12}$$

To prove Theorem **??**, we prove the following two claims:

Lemma 1. *For the above regression problem, Regret(FTL)−Regret(BTL)* $\leq O(\frac{\log T}{T})$.

Lemma 2. *For the above regression problem, Regret(BTL)* ≤ 0.

It is obvious that Theorem **??** follows from Lemma **??** and Lemma **??**.

Proof. (Lemma**??**) We prove it as follows:

$$\begin{aligned} Regret(FTL) - Regret(BTL) &= \frac{1}{T}\sum_{t=1}^{T}(y_t^b - y_t^*)^2 - (y_t^{++} - y_t^*)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}|(y_t^b - y_t^{++})(y_t^b + y_t^{++} - 2y_t^*)| \\ &\leq \frac{2}{T}\sum_{t=1}^{T}|(y_t^b - y_t^{++})| \\ &\leq \frac{2}{T}\sum_{t=1}^{T}\frac{1}{t} \\ &= O(\frac{\log T}{T}) \end{aligned} \tag{13}$$

The third inequality follows due to the convexity of the function. □

Proof. (Lemma **??**) Note that Regret(BTL)= Loss(BTL)-Best Loss. It follows that:

$$Regret(BTL) = \frac{1}{T}\sum_{t=1}^{T}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T}(y_{best}^T - y_t^*)^2 \tag{14}$$

Here, $y_{best}^k = \text{argmin}_{y\in[0,1]}\sum_{t=1}^{k}(y - y_t^*)^2$. Since $y_{best}^T = y_T^{++}$, it follows that:

$$\begin{aligned} Regret(BTL) &= \frac{1}{T}\sum_{t=1}^{T-1}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-1}(y_{best}^T - y_t^*)^2 \\ &\leq \frac{1}{T}\sum_{t=1}^{T-1}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-1}(y_{best}^{T-1} - y_t^*)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T-2}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-2}(y_{best}^{T-1} - y_t^*)^2 \\ &\leq \frac{1}{T}\sum_{t=1}^{T-2}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-2}(y_{best}^{T-2} - y_t^*)^2 \\ &\cdot \\ &\cdot \\ &\cdot \\ &\leq \frac{1}{T}(y_1^{++} - y_1^*)^2 - \frac{1}{T}(y_{best}^1 - y_1^*)^2 \\ &= 0 \end{aligned} \tag{15}$$

This proves our claim. □