CS 596: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Yikai Zhang

Lecture # 6
Oct 11, 2016

1 Online Learning in Bandit Setting

The Bandit setting can be described as:

- Algorithm has a list of N experts.
- On day t, algorithm will pick expert i_t.
- Loss of the expert $l^t_{i_t}$ is revealed.

While in previous learning problems algorithms have access to the loss of every expert, bandit setting only reveals loss of the expert that the algorithm chose. Although the 'game' becomes harder under this setting, it is still possible to achieve no regret.

It is not a bad first idea to just use RMW and feed 0 loss for the experts that we did not pick at each time step. However, this algorithm is not good as described. One can construct an adversarial sequence in which the the sub-optimal experts have excellent performance in several rounds at the beginning and the optimal one makes a lot of mistakes. From then point on, the optimal expert does really well but will have very low probability to be picked in the following rounds. This can be dealt with via a modified algorithm $EXP3$.

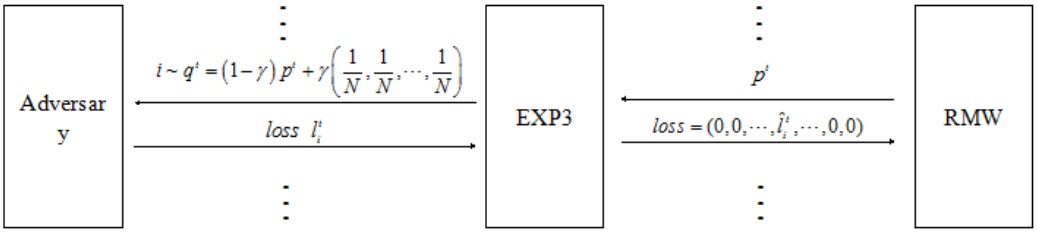

Figure 1: EXP3 algorithm

EXP3 Algorithm: The process of EXP3 algorithm is illustrated in figure 1. It is based on RMW but rather than picking experts using p_t, we 'smooth' the distribution of our experts by trading off between uniform distribution and the distribution output by RMW. Probability of picking each expert is $q^t = (1-\gamma)p^t + \gamma\, Unif(N)$. After receiving l^t_i, $\hat{l}^t_i = l^t_i/q^t_i$ (l^t_i is divided by $q^t_i$to make it unbiased) will be used as loss of expert i_t and 0 will be loss for experts not picked for RMW.

If we define $best\ loss = \frac{1}{T}\min_{j\in[N]} \sum_{t=1}^{T} l^t_j$ we have the following theorem.

Theorem 1. $Regret(EXP3) = O(\sqrt{\frac{NlogN}{T}})$

We will prove a weaker bound (but still no regret): $Regret(EXP3) \leq 3(\frac{NlogN}{T})^{\frac{1}{3}}$

Proof. By previous lecture we have following inequality for RMW.

$$E[M] \leq m + \epsilon T + \frac{\log(|H|)}{\epsilon} \tag{1}$$

In $EXP3$ we have $M = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t \hat{l}_t^t$ and $m = \frac{1}{T}\min_{j\in[N]}\sum_{t=1}^{T}\hat{l}_j^t$. Put them into (1) we have

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t \hat{l}_i^t \leq \frac{1}{T}\sum_{t=1}^{T}\hat{l}_j^t + \epsilon + \frac{N}{\gamma}\frac{logN}{\epsilon T} \qquad \forall j \in [N] \tag{2}$$

Here $\hat{l}_j^t$ is a random variable with expectation $E[\hat{l}_j^t] = E[\frac{l_i^t}{q_i^t}] = q_i^t(\frac{l_i^t}{q_i^t}) + 0(1-q_i^t) = l_i^t$. Taking expectation on both side of (2) we can get

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t l_i^t \leq \frac{1}{T}\sum_{t=1}^{T} l_j^t + \epsilon + \frac{N}{\gamma}\frac{logN}{\epsilon T} \qquad \forall j \in [N] \tag{3}$$

Thus we can have

$$\begin{aligned}
(1-\gamma)\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t l_i^t \leq \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t l_i^t &\leq \frac{1}{T}\sum_{t=1}^{T} l_j^t + \epsilon + \frac{N}{\gamma}\frac{logN}{\epsilon T} \\
\Rightarrow (1-\gamma)\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t l_i^t + \gamma\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} u_i^t l_i^t &\leq \frac{1}{T}\sum_{t=1}^{T} l_j^t + \epsilon + \frac{N}{\gamma}\frac{logN}{\epsilon T} + \gamma\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} u_i^t l_i^t \\
\Rightarrow \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} q_i^t l_i^t &\leq \frac{1}{T}\sum_{t=1}^{T} l_j^t + \epsilon + \frac{N}{\gamma}\frac{logN}{\epsilon T} + \gamma \\
\Rightarrow Regret &\leq \epsilon + \frac{N}{\gamma}\frac{logN}{\epsilon T} + \gamma
\end{aligned} \tag{4}$$

By picking $\epsilon = \gamma$ and setting $\gamma = (\frac{NlogN}{T})^{\frac{1}{3}}$ we have $Regret(EXP3) \leq 3(\frac{NlogN}{T})^{\frac{1}{3}}$.

□

2 Online Learning with large N

In many situations the number of experts could be very large or even $N \to \infty$. However, no regret is still possible if the 'game' is not too hard to play.

Theorem 2 (Informal)**.** *If the offline problem can be solved in polynomial time, the online problem can achieve no regret in polynomial time.*

Example 1. *Online regression*

- *S: set of experts = [0,1]*
- *On day t, pick* $y_t \in [0, 1]$.

- *Reveal* y_t^* *and loss* $= (y_t - y_t^*)^2$.

Where we define Best loss $= \min_{y\in[0,1]} \sum_{t=1}^{T}(y - y_t^*)^2$

Denote y_t^* as the revealed value of y^* in round t. We use the Follow the leader Algorithm: At time t, output $y_t = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i^*$.

Theorem 3. *For the regression problem above, Regret(FTL)* $= O(\frac{\log T}{T})$.

Proof. In the proof we will use another algorithm FTL^{++}:At time t, output $y_t = \frac{1}{t}\sum_{i=1}^{t} y_i^*$. (Note: We can not use this algorithm in practice since in Round t,y_t^* is unknown).
First we show that $Regret(FTL^{++}) \leq 0$.:
If we define $y_{best}^t = \underset{y\in[0,1]}{argmin} \sum_{t=1}^{t}(y - y_t^*)^2$, we have

$$\begin{aligned}
Loss(FTL^{++}) - Loss(Best) &= \frac{1}{T}\sum_{t=1}^{T}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T}(y_{best}^T - y_t^*)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T-1}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-1}(y_{best}^T - y_t^*)^2. \\
&\leq \frac{1}{T}\sum_{t=1}^{T-1}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-1}(y_{best}^{T-1} - y_t^*)^2. \\
&= \frac{1}{T}\sum_{t=1}^{T-2}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-2}(y_{best}^{T-1} - y_t^*)^2. \\
&\leq \frac{1}{T}\sum_{t=1}^{T-2}(y_t^{++} - y_t^*)^2 - \frac{1}{T}\sum_{t=1}^{T-2}(y_{best}^{T-2} - y_t^*)^2. \\
&\cdot \\
&\cdot \\
&\cdot \\
&\leq \frac{1}{T}(y_1^{++} - y_1^*)^2 - \frac{1}{T}(y_{best}^1 - y_1^*)^2. \\
&= 0.
\end{aligned} \tag{5}$$

Next we show that $Regret(FTL) - Regret(FTL^{++}) = O(\frac{logT}{T})$.

$$\begin{aligned}
Regret(FTL) - Regret(FTL^{++}) &= \frac{1}{T}\sum_{t=1}^{T}[(y_t - y_t^*)^2 - (y_t^{++} - y_t^*)^2] \\
&= \frac{1}{T}\sum_{t=1}^{T}|(y_t - y_t^{++})(y_t + y_t^{++} - 2y_t^*)| \\
&\leq \frac{2}{T}\sum_{1}^{T}|y_t - y_t^{++}| \\
&\leq \frac{2}{T}\sum_{1}^{T}\frac{1}{t} = O(\frac{logT}{T})
\end{aligned} \tag{6}$$

□

In general, the problem can be described as:

- S: set of experts $\subset R^d$
- On day t, pick $x_t \in S$.
- After picking experts, loss-function $l_t(\cdot)$ will be revealed($loss = l_t(x_t)$).

Where we define offline problem as: $Best\ loss = \min_{x \in S} \sum_{t=1}^{T} l_t(x)$.
The example above fits into this setting where $l_t(x) = (x_t - y_t^*)^2$ and $S = [0, 1]$. However, FTL can fail in general. Here is an example:

Example 2. *Suppose we have $S = \{+1, -1\}$. At time t, algorithm will receive loss function:loss $= c_t x$ where $c_t = (-1)^{t+1}$ for $t \geq 2$ and $c_1 = 0.5$. Then the answer given by algorithm will be $x_t = 1$ for odd t and $x_t = -1$ for even t. The number of mistakes will be proportional to T.*

In this example, FTL fails because it is too sensitive to input sequence thus not robust to 'bad' sequence. To deal with this problem, we can use regularized FTL called *Follow the perturbed leader* (FTPL) where $x_t = \underset{x \in S}{argmin} \sum_{i=1}^{t-1} l_i(x) + R(x)$. Note that $x_{best} = \underset{x \in S}{argmin} \sum_{t=1}^{T} l_t(x) + R(x)$.

Theorem 4. *If $D = \max_{x,y \in S} |R(x) - R(y)|$, $\|\nabla l_t(\cdot)\| \leq G$, we have $Regret(FTPL) \leq DG\sqrt{\frac{1}{T}}$*

3 Additional Readings

- Original EXP3 paper. `https://cseweb.ucsd.edu/~yfreund/papers/bandits.pdf`
- Survey by Elad Hazan. `http://ocobook.cs.princeton.edu/OCObook.pdf`