CSC 588: Machine learning theory **Spring 2021**

Lecture 26: Finish UCB Analysis; Adversial multi-armed bandit and EXP3

Lecturer: Chicheng Zhang *Scribe: Bohan Li*

1 UCB Analysis

Theorem 1. *UCB satisfies two regret bounds dependent on action space K and time space T:*

1. *(gap dependent)*

$$R_T \leq \sum_{a:\Delta(a)>0} \frac{16 \ln T}{\Delta(a)} + 3K$$

Where $\Delta(a) = l(a) - l(a^*)$ is the suboptimality gap of action a.

2. *(gap independent)*

$$R_T \leq O(\sqrt{TK})$$

The gap dependent bound is loose for the case where the suboptimality gaps are small. In the extreme case where all $\Delta(a) = \frac{1}{T}$ for $a \neq a^*$, the regret bound would be on the order of $R_T \leq O(KT)$, which is clearly pessimistic. This motivates the gap independent bound, which we now prove.

Proof. We begin the proof by a useful lemma, which we proved last lecture.

Lemma 2. $\mathbb{E}[m_T(a)] \leq \frac{16 \ln T}{\Delta(a)^2} + 3$

Define a cutoff $\Delta > 0$. Group the arms by arms above and below the threshold such that:

$$R_T = \sum_a \mathbb{E}[m_T(a)]\Delta(a) = \sum_{a:\Delta(a)\in(0,\Delta]} \mathbb{E}[m_T(a)]\Delta(a) + \sum_{a:\Delta(a)>\Delta} \mathbb{E}[m_T(a)]\Delta(a)$$

We can use lemma 2 and the fact that we pull arms for at most T times:

$$\begin{aligned} R_T &\leq T\Delta + \sum_{a:\Delta(a)>\Delta} \left(\Delta(a) \frac{16 \ln T}{\Delta(a)^2} + 3\Delta(a) \right) \\ &\leq T\Delta + \frac{16K \ln T}{\Delta} + 3K \end{aligned}$$

Since each $\Delta(a) < 1$ and $\Delta > \Delta(a)$ in the summation. Here, Δ appears only in analysis, so we are free to choose Δ to minimize the lower bound. We therefore can equalize the two Δ terms to obtain that $\Delta = \sqrt{\frac{K \ln T}{T}}$ and that the corresponding bound is

$$R_T \leq O(\sqrt{TK \ln T})$$

This concludes the proof. □

Algorithm 1 Multi-armed bandit

```
for $t = 1, 2, \cdots, T$ do
    environment gives $l_t \in [0,1]^k$
    learner selects action $a_t \in \{1, \cdots, k\}$
    learner suffer loss $l_t(a_t)$
end for
```

2 Adversial multi-armed bandit

In cases where losses are non-stationary or do not come from a distribution, can we still give algorithms with regret guarantees? Recall the multi-armed bandit algorithm. We assume in this course that the l_t's are chosen before iteration 1. This is referred to as an oblivious adversary. We allow the learner to randomly select actions a_t based on a categorical distribution $p_t \in \Delta^{K-1}$. The goal is to minimize the regret

$$R_T = \mathbb{E}\Big[\sum_{t=1}^{T} l_t(a_t)\Big] - \min_a \sum_{t=1}^{T} l_t(a)$$

Alternatively, we can use the fact that

$$\mathbb{E}_{a_t \sim p_t}[l_t(a_t)] = \sum_{a=1}^{k} p_t(a) l_t(a) = \langle p_t, l_t \rangle$$
$$\min_a \sum_{t=1}^{T} l_t(a) = \min_{p \in \Delta^{K-1}} \langle p, \sum_{t=1}^{T} l_t \rangle$$

to write an alternative expression for the regret

$$R_T = \mathbb{E}\Big[\sum_{t=1}^{T} \langle p_t, l_t \rangle - \min_{p \in \Delta^{k-1}} \sum_{t=1}^{T} \langle p, l_t \rangle\Big]$$

Can we reuse the OMD algorithm to develop algorithms with low regret? At each time t, we are given $a_t \sim p_t$, as well as the loss $l_t(a_t)$.

Key Idea We can perform OMD on unbiased estimators of l_t's. Define $\hat{l}_t$ such that

$$\mathbb{E}_{a_t \sim p_t}[\hat{l}_t] = l_t$$

Here, low regret with respect to $\hat{l}_t$'s implies low regret with respect to l_t's. We construct $\hat{l}_t$ by the following method:

First, define

$$\hat{l}_t(a) = \begin{cases} 0 & a_t \neq a \\ x & a_t = a \end{cases}$$

We can then use the expectation above to solve for x.

$$\mathbb{E}_{a_t \sim p_t}[\hat{l}_t(a)] = (1 - p_t(a)) \cdot 0 + p_t(a) \cdot (x) = l_t(a)$$
$$x = \frac{l_t(a)}{p_t(a)}$$

We then write the estimator using an indicator

$$\hat{l}_t(a) = \frac{l_t(a)}{p_t(a)} I(a = a_t)$$

Using OMD with $\hat{l}_t$'s and $\Omega = \Delta^{K-1}$,

$$\psi(p) = \sum_{a=1}^{K} p(a) \ln p(a)$$

where $p_1 = (\frac{1}{K}, \cdots, \frac{1}{K})$.

$$p_{t+1}(a) \propto p_t(a) \exp(-\eta \hat{l}_t(a)) \propto \exp\bigg(-\eta \sum_{s=1}^{t} \hat{l}_s(a) \bigg)$$

This is referred to as the EXP3 algorithm. We would like to understand how this algorithm achieves the exploration/exploitation tradeoff.

1. (*exploitation*) When $\sum_{s=1}^{t} \hat{l}_s(a) \approx \sum_{s=1}^{t} l_s(a)$ is small, the probability $p_{t+1}(a)$ is greater, and thus action a that creates the small loss is more likely to be chosen.

2. (*exploration*) Since η is finite, this creates diversity amongst the possibilities of actions chosen. $\eta = \infty$ means that $p_{t+1}(a) = 0$ for all $a \neq a_t$, which becomes follow the leader. In addition, for $l_t(a_t) > 0$, the update step $p_{t+1}(a) \propto p_t(a) \exp(-\eta \hat{l}_t(a))$ skews p_{t+1} towards actions other than a_t. This effectively encourages taking other actions. It is possible to show that for $l_t(a_t) < 0$, EXP3 performs poorly.

3 Analysis of EXP3

Applying OMD guarantees naively yields

$$\sum_{t=1}^{T} \langle p_t, \hat{l}_t \rangle - \sum_{t=1}^{T} \langle p, \hat{l}_t \rangle \leq \frac{\ln d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} ||\hat{l}_t||_\infty^2, \forall p \in \Delta^{k-1}$$

We can then use properties of expectation to show that

$$\mathbb{E}\bigg[\sum_{t=1}^{T} \langle p_t, \hat{l}_t \rangle\bigg] = \mathbb{E}\bigg[\sum_{t=1}^{T} \langle p_t, l_t \rangle\bigg]$$
$$\mathbb{E}\bigg[\sum_{t=1}^{T} \langle p, \hat{l}_t \rangle\bigg] = \mathbb{E}\bigg[\sum_{t=1}^{T} \langle p, l_t \rangle\bigg]$$

This shows that the expectation of the left hand side of OMD guarantee is the regret quantity we want to bound. We then analyze the right hand side. Based on the definition of $\hat{l}_t$, it is clear that

$$||\hat{l}_t||_\infty = \frac{l_t(a_t)}{p_t(a_t)}$$

We would like to improve the regret bound for OMD with negative entropy regularizers. However, it is difficult to find a better bound on

$$\mathbb{E}_{a_t} \bigg(\frac{l_t(a_t)}{p_t(a_t)} \bigg)^2 \leq \sum_a \frac{1}{p_t(a)}$$

Since the $p_t(a)$'s can be heavily skewed towards good options, some $p_t(a)$'s can be arbitrarily small, and thus this value cannot be well controlled. This motivates the following stronger guarantee for OMD with negative entropy regularizer with positive loss vectors.

Lemma 3. *Given OMD with $\Omega = \Delta^{K-1}$, $\psi(w)$ as the negative entropy regularizer, and the learning rate η on $\{f_t(w) = \langle w, g_t \rangle\}_{t=1}^T$ where $g_t \in [0, \infty]^K$, we obtain the regret bound*

$$\begin{aligned}\sum_{t=1}^T \langle w_t, g_t \rangle - \sum_{t=1}^T \langle w, g_t \rangle &\leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T ||g_t||^2_{diag(w_t)} \\ &= \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{a=1}^K w_t(a) g_t(a)^2\end{aligned}$$

Where $diag(w_t)$ is a diagonal matrix with entries $w_t(i)$ for $i = 1, \cdots, K$.

Proof. From OMD:

$$\begin{aligned}\langle w_{t+1} - w, \eta g_t \rangle &\leq D_\psi(w, w_t) - D_\psi(w, w_{t+1}) - D_\psi(w_{t+1}, w_t) \\ &\leq \langle w_t - w_{t+1}, \eta g_t \rangle + D_\psi(w, w_t) - D_\psi(w, w_{t+1})\end{aligned}$$

The update step of OMD is that

$$w_{t+1}(a) = \frac{w_t(a) e^{-\eta g_t(a)}}{z_t} \geq w_t(a) e^{-\eta g_t(a)}$$

since $z_t \leq 1$. We apply this inequality to the stability term:

$$\langle w_t - w_{t+1}, \eta g_t \rangle \leq \sum_{a=1}^K w_t(a)(1 - e^{-\eta g_t(a)}) \eta g_t(a)$$

Fact 4. $e^x \geq 1 + x$

Applying fact 4,

$$\begin{aligned}\langle w_t - w_{t+1}, \eta g_t \rangle &\leq \sum_{a=1}^K w_t(a)(1 - (1 - \eta g_t(a))) \eta g_t(a) \\ &= \eta^2 \sum_{a=1}^K w_t(a) g_t(a)^2\end{aligned}$$

Summing over T and dividing by η, we obtain the above lemma. This concludes the proof. □

Applying lemma 3 to EXP3,

$$\begin{aligned}\text{RHS} &= \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{a=1}^K p_t(a) \hat{l}_t(a)^2 \\ &= \frac{\ln K}{\eta} + \eta \sum_{t=1}^T p_t(a_t) \left(\frac{l_t(a_t)}{p_t(a_t)} \right)^2 \\ &\leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \frac{1}{p_t(a_t)} \\ \mathbb{E}[\text{RHS}] &\leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \mathbb{E}\Big[\frac{1}{p_t(a_t)} \Big] \\ \mathbb{E}_{a_t}\Big[\frac{1}{p_t(a_t)} \Big] &= \sum_{a=1}^K p_t(a) \frac{1}{p_t(a)} = K \\ \mathbb{E}[\text{RHS}] &\leq \frac{\ln K}{\eta} + \eta T K\end{aligned}$$

To choose the learning rate to minize the regret bound, we balance the two terms so that

$$\eta = \sqrt{\frac{\ln K}{TK}}$$

Summary We have shown that the LHS in expectation is the expression for our desired regret, and have simplified bound for the RHS in expectation. This gives the following theorem:

Theorem 5. *EXP3 with* $\eta = \sqrt{\frac{\ln K}{TK}}$ *has regret* $R_T \leq O(\sqrt{TK \ln K})$.