CS861: Theoretical Foundations of Machine Learning Lecture 23 - 10/27/2023

University of Wi[illegible] Fall 2023

Lecture 2[illegible]em (continued), Adversarial bandits

Lecturer: Kirthe[illegible] *Scribed by: Congwei Yang and Bo-Hsun Chen*

Disclaimer: *These n[illegible]jected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we will continue the discussion on Hedge algorithm, and then start the topic of adversarial bandits.

1 Experts problem (continued)

Consider the hedge algorithm introduced in last lecture. For any policy π, which samples action according to P_t on round t, define

$$\bar{R}_T(\pi, \ell, a) = \sum_{t=1}^{T} p_t^T \ell_t - \sum_{t=1}^{T} \ell_t(a)$$

Let $a^* = \arg\min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$. We have,

$$\begin{aligned} R_T(\pi, \underline{\ell}) &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\ell_t(A_t) \mid p_t\right]\right] - \sum_{t=1}^{T} \ell_t(a^*) \\ &= \mathbb{E}[\bar{R}(\pi, \underline{\ell}, a^*)] \end{aligned}$$

If we bound $\bar{R}(\pi, \underline{\ell}, a^*)$, then we have a bound for $R(\pi, \underline{\ell})$.

Theorem 1 (Hedge)**.** *Let the loss vector on round t be $\ell_t \in \mathbb{R}_+^K$ $\forall t$. Let $\ell_t^2 \in \mathbb{R}_+^K$ such that $\ell_t^2(i) = (\ell_t(i))^2$. Then, for $\eta \leq 1$, the Hedge algorithm satisfies*

(i) Let $\underline{l} = (\ell_1, \cdots, \ell_T)$ be an arbitrary sequence of losses and let $a \in [K]$. Then, if $p_t^\top \ell_t \leq 1$ for all t, we have

$$\bar{R}_T(\pi, \underline{\ell}, a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^T \ell_t^2$$

(ii) If $l_t \in [0,1]^K$ $\forall t$, then

$$\bar{R_T}(\underline{\ell}, a) \leq \frac{\log(K)}{\eta} + \eta$$

(iii) If we choose $\eta = \sqrt{\frac{\log(K)}{T}}$, then $\forall a \in [K]$, and all loss vector $\underline{\ell}$,

$$\bar{R}_T(\pi, \underline{\ell}, a) \leq 2\sqrt{Tlog(K)}$$

Proof Define $\Phi_t = \frac{1}{\eta}\log\left(\sum_{a=1}^K e^{-\eta L_t(a)}\right)$. Consider

$$
\begin{aligned}
\Phi_t - \Phi_{t-1} &= [\text{illegible}]\Big) \\
&= [\text{illegible}] \cdot e^{-\eta \ell_t(a)} \,/\, [\text{illegible}]_{-1}(a)\Big) \\
&= [\text{illegible}]\Big) \\
&\le \frac{1}{\eta}\log\left(\sum_{a=1}^K p_t(a)(1-\eta\ell_t(a)+\eta^2\ell_t^2(a))\right) \qquad (\text{As } e^{-y}\le 1-y+y^2\ \forall y\ge -1) \\
&= \frac{1}{\eta}\log(1-\eta p_t^T\ell_t+\eta^2 p_t^T\ell_t^2) \\
&\le \frac{1}{\eta}(-\eta p_t^T\ell_t+\eta^2 p_t^T\ell_t^2) \qquad (\text{As } \log(1+y)\le y\ \forall y\ge -1 \text{ and since } \eta p_t^\top \ell_t \le 1) \\
&= -p_t^T\ell_t+\eta p_t^T\ell_t^2
\end{aligned}
$$

We have $\Phi_t - \Phi_{t-1} \le -p_t^T\ell_t + \eta p_t^T\ell_t^2$, so $\Phi_T - \Phi_0 \le -\sum_{t=1}^T p_t^T\ell_t + \eta\sum_{t=1}^T p_t^T\ell_t^2$. Also

$$\Phi_0 = \frac{1}{\eta}\log\left(\sum_{i=1}^K e^{-\eta L_0(i)}\right) = \frac{\log(K)}{\eta}$$

$$
\begin{aligned}
\Phi_T &= \frac{1}{\eta}\log\left(\sum_{i=1}^K e^{-\eta L_T(i)}\right) \ge \frac{1}{\eta}\log(e^{-\eta L_T(a)}) = -L_T(a) \\
&= -\sum_{t=1}^T \ell_t(a)
\end{aligned}
$$

Thus

$$-\sum_{t=1}^T \ell_t(a) - \frac{\log(K)}{\eta} \le -\sum_{t=1}^T p_t^T\ell_t + \eta\sum_{t=1}^T p_t^T\ell_t^2$$

so

$$\bar{R}_T(\pi,\underline{\ell},a) = \sum_{t=1}^T p_t^T\ell_t - \sum_{t=1}^T \ell_t(a) \le \frac{\log(K)}{\eta} + \eta\sum_{t=1}^T p_t^T\ell_t^2$$

The proof for (i) is complete. To prove (ii), we note that $\ell_t \in [0,1]$. So $\ell_t^2(a) \le 1\ \forall a \Rightarrow p_t^T\ell_t^2 \le 1$. so $R_T(\pi,\underline{\ell},a) \le \frac{\log(K)}{\eta} + \eta T$. Statement (iii) Follows by optimizing over η.

□

2 Adversarial Bandits

Adversarial bandits is a variant of the expert problem, but the learner only observes the loss for the action taken. It has the following components:

1. Oneach round, learner chooses $A_t \in [K]$. Simultaneously, the environment picks $\ell_t \in [0,1]^K$.
2. The learner incur[illegible]
3. The learner obser[illegible] feedback).

The regret $R_T(\pi, \underline{\ell})$ [illegible]ame as the expert problem:

$$[illegible] \sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$$

$$R_T(\pi, \underline{\ell}) = \mathbb{E}[R'_T(\pi, \underline{\ell})]$$

As before, we are interested in bounding $\sup_{\underline{\ell}} R_T(\pi, \underline{\ell})$.

Here, the main challenge, when compared to full information feedback, is in balancing between exploration and exploitation.

2.1 The EXP-3 Algorithm

The main idea of EXP-3 algorithm is built on Hedge. We will estimate ℓ_t by only observing $\ell_t(A_t)$. For this, we will use the following inverse probability weighted estimator:

$$\hat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)} \mathbb{1}(a = A_t) = \begin{cases} \frac{\ell_t(A_t)}{p_t(A_t)} & \text{, if } a = A_t \\ 0 & \text{, otherwise} \end{cases} \tag{1}$$

Here, $p_t(a)$ is the probability of choosing action a in Hedge. So, $\hat{\ell}_t(a)$ would look as follows:

$$\hat{\ell}_t(a) = \begin{bmatrix} 0 & \dots & 0 & \frac{\ell_t(A_t)}{p_t(A_t)} & 0 & \dots & 0 \end{bmatrix}^T$$

We will show that $\hat{\ell}_t$ is an unbiased estimator of ℓ_t, i.e., $\mathbb{E}[\hat{\ell}_t | p_t] = \ell_t$.

The EXP3 algorithm is stated below.

Algorithm 1 EXP-3 (Exponential weights for exploration and exploitation)

Require: time horizon T, learning rate η
  Set $L_0 \leftarrow \underline{0}_K$;
  for $t = 1, 2, ..., T$ **do**
    Set $p_t(a) \leftarrow \frac{\exp(-\eta L_{t-1}(a))}{\sum_{j=1}^{K} \exp(-\eta L_{t-1}(j))}$;
    Sample $A_t \sim p_t$, and incur loss $\ell_t(A_t)$;
    Update $L_t(A_t) \leftarrow L_{t-1}(A_t) + \frac{\ell_t(A_t)}{p_t(A_t)}$;
    Update $L_t(a) \leftarrow L_{t-1}(a), \forall a \neq A_t$;
  end for

Intuitively, the exploitation for EXP3 comes from the fact that arms with large losses are discounted more in the losses. The exploration comes from the fact that we only discount arms that were pulled, so arms that are pulled less frequently are more likely to be pulled in future rounds.

Before, we analyze the algorithm, we will state the following lemma.

Lemma 1. *If $\hat{\ell}_t$ is chosen as in Eq. (1), the followings are true for all $a \in \mathcal{A}$:*

1. $\mathbb{E}[\hat{\ell}_t(a) \mid p_t] = \ell_t$

2. $[\hat{\ell}_t^2(a) \mid p_t] = \frac{\ell_t^2(a)}{p_t(a)}$

We will now state t[illegible]P3. We will prove this theorem in the next class.

Proof (proof of lemm[illegible]

(i) $\mathbb{E}[\hat{\ell}_t(a) \mid p_t] = p_t(a)$ [illegible] $\ell_t(a)$

(ii) $\mathbb{E}[\hat{\ell}_t^2(a) \mid p_t] = p_t(a)$ [illegible] $\frac{\ell_t^2(a)}{p_t(a)}$ □

We can get a theorem for the upper bound of the regret of EXP-3 as follows.

Theorem 2 (EXP-3)**.** *Assume the loss vectors on each round t satisfy $\ell_t \in [0,1]^K$. Then, EXP-3 satisfies:* $\forall \underline{\ell} = [\ell_1, ..., \ell_T]$, $R_T(\pi, \underline{\ell}) \le \frac{\log(K)}{\eta} + \eta KT$. *If we choose* $\eta = \sqrt{\frac{\log(K)}{KT}}$*, then* $R_T(\pi, \underline{\ell}) \le 2\sqrt{KT\log(K)}$.

Remark The upper bounds of some strategies that we discussed in class are compared here.

- Hedge: $\tilde{O}(\sqrt{T})$ (experts problem)
- EXP-3: $\tilde{O}(\sqrt{KT})$ (adversarial bandits)
- UCB: $\tilde{O}(\sqrt{KT})$ (stochastic bandits, minimax regret)

Hedge has a better regret than EXP3 since it is an easier problem. Interestingly, even though the adversarial bandit problem subsumes the stochastic bandit problem, the worst-case regret is the same. When we prove lower bounds for the adversarial bandit problems in the next class, we will see that the hardest stochastic bandit problems are as hard as the hardest adversarial bandit problems.