CS861: Theoretical Foundations of Machine Learning Lecture 21 - 10/21/2024

University of Wisconsin–Madison, Fall 2024

Lecture 21: Adversarial Bandits

Lecturer: Kirthevasan Kandasamy *Scribed by: Xinyu Li and Zhexuan Liu*

1 Adversarial Bandits

Adversarial bandits is a variant of the expert problem, where the learner only observes the loss for the action taken (called bandit feedback). It has the following components:

1. There are a set of K actions, denoted $[K]$.
2. On round t, learner chooses an action $A_t \in [K]$.
3. An adversary (environment) simultaneously picks a loss vector $\ell_t \in [0,1]^K$, where $\ell_t(i)$ is the loss for action i.
4. The learner incurs losses $\ell_t(A_t)$.
5. The learner observes *only* $\ell_t(A_t)$ (Bandit feedback).

Recall for a sequence of losses $\ell = (\ell_1, \ldots, \ell_T) \in [0,1]^{K \times T}$ and a sequence of actions $A = (A_1, \ldots, A_T) \in [K]^T$, define the regret as

$$R'_T(A, \ell) = \sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a).$$

For a randomized policy, similarly, we define the regret $R_T(\pi, \ell)$ as

$$R_T(\pi, \ell) = \mathbb{E}[R'_T(\pi, \ell)] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)\right]$$

where we take expectation w.r.t. the randomness of π, $A_t \sim \pi(\cdot | A_1, \ell_1, \ldots, A_{t-1}, \ell_{t-1})$. As before, we are interested in bounding $\sup_{\ell \in [0,1]^K} R_T(\pi, \ell)$.

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

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:

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

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

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

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 $L_0 \leftarrow \mathbf{0}_K$;
  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 execute A_t.
    Observe 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 analyzing the algorithm, we 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. $[\hat{\ell}_t^2(a) \mid p_t] = \frac{\ell_t^2(a)}{p_t(a)}$.

The expectation is w.r.t. the randomness of the algorithm.

Proof (proof of Lemma 1)
(i) For any $a \in [K]$, $\mathbb{E}[\hat{\ell}_t(a) \mid p_t] = p_t(a)\frac{\ell_t(a)}{p_t(a)} + (1 - p_t(a)) \cdot 0 = \ell_t(a)$.

(ii) Similarly, for any $a \in [K]$, $\mathbb{E}[\hat{\ell}_t^2(a) \mid p_t] = p_t(a)\frac{\ell_t^2(a)}{p_t^2(a)} + (1 - p_t(a)) \cdot 0 = \frac{\ell_t^2(a)}{p_t(a)}$. □

We have the theorem for the upper bound of the regret of EXP3 as follows.

Theorem 1 (EXP3)**.** *Suppose $\ell_t \in [0,1]^K$ for all t, and we choose $\eta = \sqrt{\frac{\log(K)}{KT}}$. Then for all T, the regret of EXP3 satisfies*

$$R_T(\pi, \ell) \leq 2\sqrt{KT \log(K)}.$$

Remark The regret of Hedge is $\mathcal{O}(\sqrt{T \log K})$ whereas for Hedge it is $\mathcal{O}(\sqrt{KT \log(K)})$. The additional $\sqrt{K}$ factor is due to reduced (limited) feedback.

Before getting into proving Theorem 1, recall the following lemma from the Hedge theorem.

Lemma 2. *(Hedge Lemma) Let $\lambda = (\lambda_1, \ldots, \lambda_T) \in \mathbb{R}_+^K$ be a sequence of losses. Let $p = (p_1, \ldots, p_T)$ be the sequence of probability vectors chosen by Hedge with learning rate $\eta \in [0, 1]$. For any $a \in [K]$, if $p_t^\top \lambda_t \leq 1$ for all t, we have*

$$\bar{R}_T(p, \lambda, a) \triangleq \sum_{t=1}^{T} p_t^\top \lambda_t - \sum_{t=1}^{T} \lambda_t(a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \lambda_t^2.$$

Proof (proof of Theorem 1)

Let $a_* = \arg\min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$ be the best fixed arm in hindsight. We will apply the Hedge lemma stated above with $\lambda_t \leftarrow \hat{\ell}_t$ and $a \leftarrow a_*$. The conditions of the lemma is satisfied:

$$\eta = \sqrt{\frac{\log(K)}{KT}} \leq 1, \quad \text{as } K \geq 2 \text{ and } T \geq 1.$$

$$p_t^\top \hat{\ell}_t = \sum_{a=1}^{K} p_t(a) \cdot \frac{\ell_t(a)}{p_t(a)} \mathbb{1}(A_t = a) = p_t(A_t) \cdot \frac{\ell_t(A_t)}{p_t(A_t)} = \ell_t(A_t) \leq 1.$$

Therefore, we have,

$$\sum_{t=1}^{T} p_t^\top \hat{\ell}_t - \sum_{t=1}^{T} \hat{\ell}_t(a_*) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \hat{\ell}_t^{\,2}. \tag{2}$$

Recall for any a, we have (i) $\mathbb{E}[\hat{\ell}_t(a)|p_t] = \ell_t(a)$, (ii) $\mathbb{E}[\hat{\ell}^2(a)|p_t] = \frac{\ell_t^2(a)}{p_t(a)}$. We then take expectations of Equation 2 on both sides,

$$\mathbb{E}[\text{LHS}] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[p_t^\top \hat{\ell}_t | p_t] - \sum_{t=1}^{T} \underbrace{\mathbb{E}[\hat{\ell}_t(a_*)|p_t]}_{=\ell_t(a_*) \text{ by (i)}}\right].$$

Further by (i) again, we have

$$\mathbb{E}[p_t^\top \hat{\ell}_t | p_t] = p_t^\top \mathbb{E}[\hat{\ell}_t | p_t] = p_t^\top \ell_t = p_t^\top \ell_t = \mathbb{E}[\ell_t(A_t)|p_t].$$

Therefore,

$$\mathbb{E}[\text{LHS}] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\ell_t(A_t) \mid p_t] - \sum_{t=1}^{T} \ell_t(a_*)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \sum_{t=1}^{T} \ell_t(a_*) = R_T(\pi, \ell).$$

For the RHS,

$$\mathbb{E}[\text{RHS}] = \frac{\log(K)}{\eta} + \eta \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[p_t^\top \hat{\ell}_t^{\,2} | p_t]\right].$$

By (ii) and the fact that the losses are bounded in $[0, 1]$,

$$\mathbb{E}[p_t^\top \hat{\ell}_t^{\,2} | p_t] = p_t^\top \mathbb{E}[\hat{\ell}_t^2 | p_t] = \sum_{a=1}^{K} p_t(a) \cdot \frac{\ell_t^2(a)}{p_t(a)} = \sum_{a=1}^{K} \ell_t^2(a) \leq K.$$

Therefore,

$$\mathbb{E}[\text{RHS}] \leq \frac{\log(K)}{\eta} + \eta KT.$$

Hence,

$$R_T(\pi, \ell) \leq \frac{\log(K)}{\eta} + \eta KT \leq 2\sqrt{KT \log(K)} \quad \text{as } \eta = \frac{\log(K)}{KT}.$$

□

2 Lower bounds for adversarial bandits

The following theorem provides a lower bound for the minimax rate of regret of the adversarial multi-armed bandit problem.

Theorem 2. *For the adversarial multi-armed bandit problem, the minimax regret satisfies,*

$$\inf_{\pi} \sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \in \Omega(\sqrt{KT}).$$

Remark Recall the minimax lower bound for stochastic bandits is $\inf_\pi \sup_{\nu \in \mathcal{P}} R_T(\pi, \nu) \in \Omega(\sqrt{KT})$. Note that the adversarial bandit problem is is applicable in more general settings than the stochastic bandit problem. Moreover, the regret definitions are different for the adversarial bandit and stochastic bandit problems. For the adversarial bandits, the regret depends on the best action in hindsight. While, the regret of stochastic bandits depends on the arm with the lowest expected mean value. Despite this, we find that the minimax regret is similar for both problems. This is because the hardest stochastic bandit problems are as hard as the hardest adversarial bandit problems. In fact, the proof of this lower bound will rely on similar techniques to the proof of the lower bound for stochastic bandits.

Our proof will consider stochastic losses and show that the expected regret is large. Then, there is at least one sequence of losses (drawn from this distribution) for which the regret should be large.

In the proof, we will use Bretagnolle-Huber inequality stated as follows. Let P_0, P_1 be any two distributions. For any event A,

$$P_0(A) + P_1\left(A^c\right) \geq \frac{1}{2} e^{-\mathrm{KL}(P_0, P_1)}.$$

Proof Let π be given. We will consider two stochastic bandit models $\nu^{(1)}$ and $\nu^{(2)}$, where $\nu^{(j)} = (\nu_1^{(j)}, \nu_2^{(j)}, \ldots, \nu_K^{(j)})$ and each $\nu_i^{(j)}$ has Bernoulli losses.

Let $P^{(1)}$ and $P^{(2)}$ denote the probability distributions of the action-loss sequence $A_1, \ell_1(A_1), ..., A_T, \ell_1(A_T)$ due to π's interaction with $\nu^{(1)}$ and $\nu^{(2)}$ respectively.

Let $\mathbb{E}^{(1)}$ and $\mathbb{E}^{(2)}$ denote the corresponding expectations.

Let $\mathbb{E}_\pi$ denote the expectation with respect to the randomness in the policy. Define

$$R'_T(A, \ell) \stackrel{\triangle}{=} \sum_{t=1}^{T} \ell(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell(a),$$

so that $R_T(\pi, \ell) = \mathbb{E}_\pi[R]_T(A, \ell)]$. We can now lower bound the worst case regret for π as follows,

$$\begin{aligned}
\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) &= \sup_{\ell \in [0,1]^{K \times T}} \mathbb{E}_\pi[R'_T(A, \ell)] \\
&\geq \mathbb{E}_{j \sim \mathrm{Unif}(\{1,2\})} \mathbb{E}_{\ell \sim \nu^{(j)}} \mathbb{E}_\pi[R'_T(\pi, \ell)] \\
&= \frac{1}{2} \mathbb{E}_\pi\left[\mathbb{E}_{\ell \sim \nu^{(1)}}\left[R'_T(A, \ell)\right]\right] + \frac{1}{2} \mathbb{E}_\pi\left[\mathbb{E}_{\ell \sim \nu^{(2)}}\left[R'_T(A, \ell)\right]\right].
\end{aligned}$$

The inequality uses max $\geq$ avg, noting that $j \sim \mathrm{Unif}(\{1, 2\})$ and then $\ell \sim \nu^{(j)}$ defines a distribution over $[0, 1]^{K \times T}$.

Noting that the pointwise minimum is concave, by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_{\ell \sim \nu^{(j)}}\left[R'_T(\pi, \ell)\right] &= \mathbb{E}_{\ell \sim \nu^{(j)}}\left[\sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)\right] \\
&\geq \mathbb{E}_{\ell \sim \nu^{(j)}}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{a \in [K]} \mathbb{E}_{\ell \sim \nu^{(j)}} \sum_{t=1}^{T} \ell_t(a) \\
&= \mathbb{E}_{\ell \sim \nu^{(j)}}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - T\mu_j^*,
\end{aligned}$$

where $\mu_j^* = \min_{a\in[K]} \mathbb{E}_{X\sim\nu_i^{(j)}}[X]$ (the optimal value in $\nu^{(j)}$).

Now we take $\mathbb{E}_\pi$ on both sides of Equation 2,

$$\mathbb{E}_\pi\left[\mathbb{E}_{\ell\sim\nu^{(j)}}\left[R'_T(\pi,\ell)\right]\right] \geq \mathbb{E}^{(j)}\left[\sum_{t=1}^{T}\ell_t(A_t)\right] - T\mu_j^* = R_T^{\text{stoc}}(\pi,\nu^{(j)}),$$

where $\mathbb{E}^{(j)}$ is expectation under π's interaction with $\nu^{(j)}$, and $R_T^{\text{stoc}}(\pi,\nu^{(j)})$ is the "stochastic bandit regret" of policy π on the stochastic bandit model $\nu^{(j)}$. Combine the equality above with Equation 2, we have

$$\sup_{\ell\in[0,1]^{K\times T}} R_T(\pi,\ell) \geq \frac{1}{2}\left(R_T^{\text{stoc}}(\pi,\nu^{(1)}) + R_T^{\text{stoc}}(\pi,\nu^{(2)})\right).$$

Denote $N_{a,T} = \sum_{t=1}^{T}\mathbb{1}(A_t = a)$. Let $\nu^{(1)}$ be defined as,

$$\nu_1^{(1)} = \text{Bern}\left(\frac{1}{2}-\delta\right) \text{ and } \nu_i^{(1)} = \text{Bern}\left(\frac{1}{2}\right), \forall \text{i} \in \{2,\ldots,\text{K}\},$$

where $\delta < 1/8$ is a parameter that we will specify later.

Since $\sum_{a=1}^{K}\mathbb{E}^{(1)}[N_{a,T}] = T$, there exists some $a' \in \{2,\ldots,K\}$ such that $\mathbb{E}^{(1)}[N_{a',T}] \leq T/(K-1)$. Define $\nu^{(2)}$ so that,

$$\nu_{a'}^{(2)} = \text{Bern}\left(\frac{1}{2}-2\delta\right) \text{ and } \nu_i^{(2)} = \nu_i^{(1)} \text{ for all } i \neq a'.$$

From our construction,

$$R_T^{\text{stoc}}\left(\pi,\nu^{(1)}\right) \geq \mathbb{P}^{(1)}\left(N_{1,T} \leq T/2\right)\frac{T\delta}{2}, \quad R_T^{\text{stoc}}\left(\pi,\nu^{(2)}\right) \geq \mathbb{P}^{(2)}\left(N_{1,T} > T/2\right)\frac{T\delta}{2}.$$

Therefore,

$$\begin{aligned}\sup_{\ell\in[0,1]^{K\times T}} R_T(\pi,\ell) &\geq \frac{T\delta}{4}\left(\mathbb{P}^{(1)}\left(N_{1,T}\leq T/2\right) + \mathbb{P}^{(2)}\left(N_{1,T} > T/2\right)\right)\\ &\geq \frac{T\delta}{8}\exp\left(-\text{KL}\left(\mathbb{P}^{(1)},\mathbb{P}^{(2)}\right)\right),\end{aligned}$$

where the last inequality is by Bretagnolle-Huber inequality.

Noting that $\mathbb{E}^{(1)}\left[N_{a',T}\right] \leq T/(K-1)$ by our construction, we have

$$\begin{aligned}\text{KL}\left(\mathbb{P}^{(1)},\mathbb{P}^{(2)}\right) &= \sum_{i=1}^{K}\mathbb{E}^{(1)}\left[N_{a,T}\right]\text{KL}\left(\nu_i^{(1)},\nu_i^{(2)}\right) = \mathbb{E}^{(1)}\left[N_{a',T}\right]\text{KL}\left(\nu_{a'}^{(1)},\nu_{a'}^{(2)}\right)\\ &\leq \frac{T}{K-1}\cdot C_1(2\delta)^2 = C_2\frac{T\delta^2}{K-1}.\end{aligned}$$

Therefore,

$$\sup_{\ell\in[0,1]^{K\times T}} R_T(\pi,\ell) \geq \frac{T\delta}{8}\exp\left(C_2\frac{T\delta^2}{K-1}\right).$$

Choosing $\delta = \sqrt{(K-1)/T}$, we have

$$\sup_{\ell\in[0,1]^{k\times T}} R_T(\pi,\ell) \geq C_3\sqrt{T(K-1)}.$$

□

Acknowledgements

The note is based on the scribed lecture materials prepared in Fall 2023 by Bo-Hsun Chen, Zexuan Sun and Congwei Yang.