Exp3

Seminar Advanced Topics in Reinforcement Learning and Decision Making

Jakub Tłuczek
jakub.tluczek@unine.ch

Universite de Neuchâtel
Swiss Joint Master in Computer Science

March 18, 2025

Overview

1. Adversarial Bandits
2. Importance weighted estimators
3. Exp3
   - Proof outline
4. Adversarial linear bandits
5. Exp3 for adversarial linear bandits

Adversarial Bandits

- Reward sequence $(x_t)_{t=1}^n$, with $x_{t,i}$ reward of arm i at time t, fixed by an adversary.
- Agent's action is drawn from a distribution $P_t \in \mathcal{P}_{k-1}$, where $\mathcal{P}_d$ is a probability simplex over $d+1$ actions (i.e. $\mathcal{P}_d = \{p \in [0,1]^{d+1} : ||p||_1 = 1\} = \{p \in [0,1]^{d+1} : \sum_i p_i = 1\}$.
- The policy π maps from histories onto the distribution over action

$$\pi \; : \; ([k] \times [0,1])^* \to \mathcal{P}_{k-1}.$$

- At each timestep t:
  1. Agent chooses distribution P_t
  2. Action is drawn $A_t \sim P_t$
  3. Reward $X_t = x_{tA_t}$ is observed.

Adversarial Bandits regret

Regret in case of an adversarial bandit can be summarized as:

$$R_n(\pi, x) = \max_{i \in [k]} \sum_{t=1}^{n} x_{ti} - \mathbb{E}\left[\sum_{t=1}^{n} x_{tA_t}\right] \tag{1}$$

While Worst case regret for some policy π is:

$$R_n^*(\pi) = \sup_{x \in [0,1]^{n \times k}} R_n(\pi, x) \tag{2}$$

Importance weighted estimators

Before we go into the Exp3 algorithm, we have to introduce an unbiased estimator for reward $\hat{X}_{ti}$ at time t for arm i. It is defined as:

$$\hat{X}_{ti} = \frac{\mathbb{I}\{A_t = i\}X_t}{P_{ti}} \tag{3}$$

where P_{ti} is the probability of selecting arm i at time t. Why is $\hat{X}_{ti}$ unbiased?

$$\mathbb{E}\left[\hat{X}_{ti}\right] = \mathbb{E}\left[\frac{\mathbb{I}\{A_t = i\}X_t}{P_{ti}}\right] = \sum_j P_{tj}\hat{X}_{ti} = P_{ti}\hat{X}_{ti} + \sum_{j \neq i} P_{tj} \cdot 0 = P_{ti}\frac{X_t}{P_{ti}} = x_{ti} \tag{4}$$

Loss-based importance weighted estimator

Another unbiased estimator would be:

$$\hat{X}_{ti} = 1 - \frac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{ti}} \tag{5}$$

where we can prove it is unbiased by substituting $Y_t = 1 - X_t$. The advantage of loss based estimator is, that while in the case of estimator from previous slide its variance is:

$$\mathbb{V}\left[\frac{\mathbb{I}\{A_t = i\}X_t}{P_{ti}}\right] = x_{ti}^2 \frac{1 - P_{ti}}{P_{ti}} \tag{6}$$

while loss based estimator's variance is proportional to $(1 - x_{ti})^2$:

$$\mathbb{V}\left[1 - \frac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{ti}}\right] = (1 - x_{ti})^2 \frac{1 - P_{ti}}{P_{ti}} \tag{7}$$

Exp3

Exponential-weight algorithm for **exp**loration and **exp**loitation (hence **Exp**3) is based on changing the probabilities for each actions based on term $\hat{S}_t i = \sum_{s=1}^{t} \hat{X}_s i$, which is the sum of estimated rewards up until current round t.
Probability for each arm at each timestep is calculated using the following, softmax-like, formula:

$$P_{ti} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_j \exp(\eta \hat{S}_{t-1,j})} \tag{8}$$

Then, $A_t \sim P_t$ is played, reward X_t observed and estimated rewards adjusted:

$$\hat{S}_{ti} = \hat{S}_{t-1,i} + 1 - \frac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{ti}} \tag{9}$$

Exp3 expected regret

Regret for Exp3 is bounded from above by:

$$R_n \leq \frac{\log(k)}{\eta} + \eta nk \tag{10}$$

when we set learning rate $\eta = \sqrt{\log(k)/nk}$, then the regret bound would be optimized as follows:

$$R_n \leq 2\sqrt{nk\log(k)} \tag{11}$$

Define

$$\hat{S}_n = \sum_{t,i} P_{ti} \hat{X}_{ti}, \qquad W_t = \sum_{j=1}^{k} \exp\left(\eta \hat{S}_{tj}\right)$$

Define

$$\hat{S}_n = \sum_{t,i} P_{ti}\hat{X}_{ti}, \qquad W_t = \sum_{j=1}^{k} \exp\left(\eta \hat{S}_{tj}\right)$$

We now bound the exponential term

$$\exp(\eta \hat{S}_{ni}) \leq \sum_{j=1}^{k} \exp(\eta \hat{S}_{nj}) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = k \prod_{t=1}^{n} \frac{W_t}{W_{t-1}}$$

Define

$$\hat{S}_n = \sum_{t,i} P_{ti}\hat{X}_{ti}, \qquad W_t = \sum_{j=1}^{k} \exp\left(\eta \hat{S}_{tj}\right)$$

We now bound the exponential term

$$\exp(\eta \hat{S}_{ni}) \leq \sum_{j=1}^{k} \exp(\eta \hat{S}_{nj}) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = k \prod_{t=1}^{n} \frac{W_t}{W_{t-1}}$$

We now bound the ratio

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^{k} \frac{\exp(\eta \hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta \hat{X}_{t,j}) = \sum_{j=1}^{k} P_{tj} \exp(\eta \hat{X}_{t,j})$$

Define

$$\hat{S}_n = \sum_{t,i} P_{ti}\hat{X}_{ti}, \qquad W_t = \sum_{j=1}^{k} \exp\left(\eta \hat{S}_{tj}\right)$$

We now bound the exponential term

$$\exp(\eta \hat{S}_{ni}) \leq \sum_{j=1}^{k} \exp(\eta \hat{S}_{nj}) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = k \prod_{t=1}^{n} \frac{W_t}{W_{t-1}}$$

We now bound the ratio

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^{k} \frac{\exp(\eta \hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta \hat{X}_{t,j}) = \sum_{j=1}^{k} P_{tj} \exp(\eta \hat{X}_{t,j})$$

Using the inequalities

$$\exp(x) \geq 1 + x \forall x, \qquad \exp(x) \leq 1 + x + x^2 \forall x \leq 1$$

and some algebra...

$$\hat{S}_{ni} - \hat{S}_n \leq \frac{\log(k)}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{k} P_{t,j} \hat{X}_{tj}^2.$$

$$\hat{S}_{ni} - \hat{S}_n \leq \frac{\log(k)}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{k} P_{t,j} \hat{X}_{tj}^2.$$

Taking expectations gives R_{ni} (the expected regret of playing i) on the lhs and:

$$\sum_{j=1}^{k} P_{t,j} \hat{X}_{tj}^2 \leq k.$$

$$\hat{S}_{ni} - \hat{S}_n \leq \frac{\log(k)}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{k} P_{t,j} \hat{X}_{tj}^2.$$

Taking expectations gives R_{ni} (the expected regret of playing i) on the lhs and:

$$\sum_{j=1}^{k} P_{t,j} \hat{X}_{tj}^2 \leq k.$$

We can now sum over t to obtain

$$R_{ni} \leq \frac{\log(k)}{\eta} + \eta nk = 2\sqrt{nk \log(k)}$$

$\square$

Adversarial linear bandits

In adversarial linear settings, actions from action set $\mathcal{A} \subset \mathbb{R}^d$ are d-dimensional vectors, just as reward x_t at time t. Reward in this setting is given by inner product $\langle A_t, x_t \rangle$. Without loss of generality, we can switch to losses $y_t = 1 - x_t$. Therefore, if observed loss would be defined as $Y_t = \langle A_t, y_t \rangle$, then regret after n steps is defined as:

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} Y_t\right] - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \langle a, y_t \rangle \tag{12}$$

Exp3 for finite exponential weights

Probability distribution $P_t(a)$ is given by mixture distribution:

$$P_t(a) = (1-\gamma)\tilde{P}_t(a) + \gamma\pi(a) \tag{13}$$

where $\pi(a)$ is an exploration distribution mapping simplex $\mathcal{A} \to [0,1]; \sum_{a\in\mathcal{A}} \pi(a) = 1$, while $\tilde{P}_t(a)$ is a probability mass function:

$$\tilde{P}_t(a) \propto \exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a)\right) \tag{14}$$

Finally, loss estimate is estimated by $\hat{Y}_t = Q_t^{-1} A_t Y_t$, where Q_t is given by:

$$Q_t = \sum_{a\in\mathcal{A}} P_t(a) a a^\intercal \tag{15}$$

Exp3 for finite exponential weights

Distribution is calculated at each step by:

$$P_t(a) = \gamma\pi(a) + (1-\gamma)\frac{\exp\left(-\eta\sum_{s=1}^{t-1}\hat{Y}_s(a)\right)}{\sum_{a'\in\mathcal{A}}\exp\left(-\eta\sum_{s=1}^{t-1}\hat{Y}_s(a')\right)} \tag{16}$$

Action $A_t \sim P_t$ is sampled, loss $Y_t = \langle A_t, y_t\rangle$ is observed and loss estimate is updated using:

$$\hat{Y}_t = Q_t^{-1}A_tY_t \tag{17}$$

Exp3 regret for adversarial linear bandits

Exp3 regret is bounded from above by:

$$R_n \leq 2\sqrt{(2g(\pi) + d)n\log(k)} \tag{18}$$

where d is the dimension of $\mathcal{A}$, k is the number of arms and $g(\pi)$ equals:

$$g(\pi) = max_{a \in \mathcal{A}} ||a||^2_{Q^{-1}(\pi)} \tag{19}$$

Exp3 for continuous exponential weights

If the number of arms is big, or if it goes to ∞, then this algorithm becomes intractable. Instead of computing P_t for every arm, we can switch to continuous exponential weights. Assuming that $\mathcal{A}$ is convex, distribution is calculated by:

$$P_t(B) = \gamma\pi(B) + (1-\gamma)\frac{\int_B \exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a)\right) da}{\int_{\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a)\right) da} \tag{20}$$

Rest of the algorithm is analogous to the Exp3 for finite action sets.