

Lecture 1: January 1, 2024

Lecturer: Yishay Mansour

Scribe¹: Amit Amram & Dean Oren & Yarden Rashti

1 Administration

The course website is in the Moodle. The final grade consists of,

1. Lecture scribe summary (in groups of 2-3) 30%.
2. Homework (~ 3 problem sets) 30%.
3. Research project 40%.

The topic of the course is *Online learning and Multi-Arm Bandit*, with an emphasis on stochastic models.

Recommended book: *Introduction to multi-armed bandits [Slivkins, 2019]*. Most of the material will be based on this book, and at the end of every lecture there will be specific references.

Related courses in TAU: Tomer Koren's course focuses on optimization, online optimization and concentrates mainly on adversarial models.

2 Introduction

The framework we study in this course models *Decision making under uncertainty*. Here are a few motivating examples:

- **Web advertising:** Websites need to choose what ad to present to a user. Get reward of 1 if there the user clicks and 0 otherwise. We don't know in advance what is the probability that the user would click on an ad.
- **Online pricing:** A seller proposes price p to a buyer; if the buyer buys the product, the seller gets the reward of p . Otherwise, they get no reward.
- **Stock market:** We need to choose the portfolio, and the reward is the change in stock price.

In Multi-Arm Bandit (MAB), we get feedback only on the actions we choose. The main challenge is finding the right balance between exploration (gathering new information on the different actions) and exploitation (exploiting the current information we have so far). We'll see many variations MAB models:

- **Additional information regarding other actions:** in MAB, we only receive information regarding the action we chose (e.g., in web advertising, we don't know if a user would have clicked on an ad we didn't present). There are model variations in which we receive partial feedback (or infer some information) regarding actions that were not taken (e.g., in online pricing, if the user buys at a price p , we can assume she will buy at any price $p' < p$; If the user does not buy at a price p we can assume that she will not buy at any price $p' > p$. In some variations, we receive full information regarding any action (e.g., in the stock market, we have complete information since we know all stock prices).
- **Different types of rewards:** *Stochastic reward* (i.i.d samples) will be our main focus throughout most of the course, and *Adversarial reward*, in which we have no statistical assumption on the rewards. The benchmark would be the best action in hindsight.

¹Based on the scribe notes of Uri Sherman & Tal Lancewicki from 2021/22

- **Contextual bandits:** We observe some information (an attribute vector) before taking an action. For example, we may have some information on a user, such as browser cookies for web advertising.
- **Bayesian Prior:** prior belief on how the user behaves - we'll update our beliefs as we get more information.

2.1 Applications

- **Medical trials:** How do doctors decide what treatment or procedure will best benefit the patient? We want to know the best treatment without testing in advance but dynamically test and converge towards the best treatment.
- **Web design:** A/B testing. For example, the color of a button may have a small impact on whether a user clicks on it and we want to choose the best color. We want to be able to do an online experiment and learn the color that would give us the highest profit.
- **Recommendations:** Websites aim to provide good recommendations for the user (music, products, etc.) in order to increase the amount of time the user spends on the website. Usually, we get feedback on the chosen recommendation but don't observe feedback on recommendations we did not choose.
- **Job scheduling:** assigning jobs to machines in a data center.
- **cellular networks:** Allocation of cellular devices to frequencies.
- **Robot control** (for which Reinforcement Learning is more relevant).
- **Sales optimization:** Setting prices to maximize revenue.

2.2 History

- Thompson [1933] present the first bandit algorithm. Introduce the topic of decision-making but without formulation.
- Robbins [1952] investigate the problem in the context of medical trials design, such that better treatments would be chosen more times than treatments not as good.
- Gittins [1979] - Operations Research. A Markov model and optimal solution for some settings.
- Auer et al. [2002a] present the UCB algorithm (stochastic MAB).
- Auer et al. [2002b] present EXP3 algorithm (adversarial MAB).

3 Concentration bounds

We start by deriving the concentration bounds, which we will extensively use. We have a random variable $X \in [0, 1]$. To estimate $\mu = \mathbb{E}[X]$ we can sample it a few times, $x_i \sim X$, and take the empirical average $\bar{\mu} = \frac{1}{m} \sum_{i=1}^m x_i$. How good is this estimation? The Law of Large Numbers (LLN) guarantees that the empirical average converges to μ as $m \rightarrow \infty$. But we want to bound the error of our estimation using a finite number of samples. For that, we'll use Hoeffding's inequality. Informally:

$$\mathbb{E} \left[\left| \frac{1}{m} \sum_{i=1}^m x_i - p \right| \right] \approx \frac{1}{\sqrt{m}},$$

and we would expect that,

$$\Pr \left[\left| \frac{1}{m} \sum_{i=1}^m x_i - p \right| \geq \frac{\lambda}{\sqrt{m}} \right] \leq e^{-\lambda^2}.$$

We want to prove Hoeffding's Theorem.

3.1 Sub-gaussian random variables

Definition X is σ^2 -sub-gaussian random variable if,

$$\forall \lambda \in \mathbb{R} : \mathbb{E} [e^{\lambda X}] \leq e^{\sigma^2 \frac{\lambda^2}{2}}.$$

Examples:

1. If $X \sim \mathcal{N}(0, \sigma^2)$ then X is σ^2 -sub-gaussian.
2. If $\mathbb{E}[X] = 0$ and $|X| \leq B$ then X is B^2 -sub-gaussian (will be proved later).

Lemma 1 (Properties of Sub-Gaussians) *Let X be a σ^2 -sub-gaussian random variable. The following properties hold:*

1. $\mathbb{E}[X] = 0$ and $\text{Var}(X) \leq \sigma^2$.
2. cX is $c^2\sigma^2$ -sub-gaussian.
3. If X_1, \dots, X_m are σ^2 -sub-gaussians (i.i.d), then $S = \sum X_i$ is $m\sigma^2$ -sub-gaussian.
4. If X_1, \dots, X_m are σ^2 -sub-gaussians (i.i.d), then $\frac{S}{m} = \frac{1}{m} \sum_{i=1}^m X_i$ is (σ^2/m) -sub-gaussian.

Proof:

1. For $\lambda \rightarrow 0$ we have,

$$e^{\lambda x} = 1 + \lambda x + \frac{\lambda^2 x^2}{2} + o(\lambda^3).$$

Ignoring $o(\lambda^3)$ terms,

$$\begin{aligned} 1 + \frac{\lambda^2 \sigma^2}{2} &\geq e^{\frac{\lambda^2 \sigma^2}{2}} \stackrel{\text{(sub-gaus)}}{\geq} \mathbb{E} [e^{\lambda x}] = 1 + \lambda \mathbb{E}[X] + \frac{\lambda^2 \mathbb{E}[X^2]}{2} \\ \implies \mathbb{E}[X] + \frac{\lambda \mathbb{E}[X^2]}{2} &\leq \frac{\lambda \sigma^2}{2}. \end{aligned}$$

By taking $\lambda \rightarrow 0$, we get $\mathbb{E}[X] \leq 0$. By applying the same argument on $-X$ we get $\mathbb{E}[-X] \leq 0$ and therefore $\mathbb{E}[X] = 0$.

Now, from the above we have, since $\mathbb{E}[X] = 0$,

$$\frac{\lambda \mathbb{E}[X^2]}{2} \leq \frac{\lambda \sigma^2}{2} \implies \text{Var}(X) = \mathbb{E}[X^2] \leq \sigma^2.$$

- 2.

$$\forall \lambda \in \mathbb{R} : \mathbb{E} [e^{\lambda(cX)}] = \mathbb{E} [e^{(c\lambda)X}] \leq e^{\frac{1}{2}(c\lambda)^2 \sigma^2} = e^{\frac{1}{2}\lambda^2 (c\sigma)^2},$$

where the inequality since X is σ^2 -sub-gaussian.

- 3.

$$\forall \lambda \in \mathbb{R} : \mathbb{E} [e^{\lambda S}] = \mathbb{E} \left[\prod_{i=1}^m e^{\lambda X_i} \right] = \prod_{i=1}^m \mathbb{E} [e^{\lambda X_i}] \leq \prod_{i=1}^m e^{\frac{\lambda^2 \sigma^2}{2}} = e^{\frac{\lambda^2 \sigma^2 m}{2}},$$

where the second equality is since the variables are independent.

4. Follows immediately from properties 2 and 3.

□

Theorem 2 *If X is σ^2 -sub-gaussian then for any ϵ ,*

$$\Pr[X \geq \epsilon] \leq e^{-\frac{\epsilon^2}{2\sigma^2}}$$

Proof:

$$\begin{aligned} \Pr[X \geq \epsilon] &= \Pr[e^{\lambda X} \geq e^{\lambda\epsilon}] \\ &\leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\epsilon}} && \text{(Markov)} \\ &\leq e^{\sigma^2 \frac{\lambda^2}{2} - \lambda\epsilon} && \text{(sub-gaus)} \end{aligned}$$

Choosing $\lambda = \frac{\epsilon}{\sigma^2}$ gives us the result. \square

Lemma 3 *If $\mathbb{E}[X] = 0$ and $|X| \leq B$ then X is B^2 -sub-gaussian.*

Proof: Define $F(\lambda) = \log \mathbb{E}[e^{\lambda X}]$. We have,

$$F'(\lambda) = \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]},$$

Note that the derivative and the expectancy are interchangeable. We take the second derivative:

$$F''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \right)^2.$$

Now, define a new distribution for X

$$dQ = \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]} dP.$$

And now we can bound $F''(\lambda)$ since $F''(\lambda) = \mathbb{E}_Q[X^2] - \mathbb{E}_Q^2[X] = \text{Var}_Q(X) \leq B^2$. Using the fundamental theorem of calculus,

$$\begin{aligned} \log \mathbb{E}[e^{\lambda X}] &= F(\lambda) - F(0) && (F(0) = 0) \\ &= \int_0^\lambda F'(y) dy \\ &= \int_0^\lambda F'(y) - F'(0) dy && (F'(0) = \mathbb{E}[X] = 0) \\ &= \int_0^\lambda \int_0^y F''(z) dz dy \\ &\leq \int_0^\lambda \int_0^y B^2 dz dy \\ &= B^2 \frac{\lambda^2}{2}. \end{aligned}$$

Taking the exponent on both sides completes the proof. \square

Theorem 4 (Hoeffding's inequality) *Let X_1, \dots, X_m independent r.vs where $X_i \in [0, 1]$, and let $\mu_i = \mathbb{E}[X_i]$. We have that,*

$$\Pr \left[\frac{1}{m} \sum_{i=1}^m (X_i - \mu_i) \geq \epsilon \right] \leq \exp \left(-\frac{\epsilon^2 m}{2} \right)$$

Proof: Define $\bar{X}_i = X_i - \mu_i$. Clearly, $\mathbb{E}[\bar{X}_i] = 0$ and $|\bar{X}_i| \leq 1$. By Lemma 3, \bar{X}_i is 1-sub-gaussian, and by Lemma 1, $\frac{1}{m} \sum_{i=1}^m (X_i - \mu_i)$ is $(1/m)$ -sub-gaussian. Applying Theorem 2 gives us the desired bound. \square

4 Stochastic Bandits

4.1 Model

There are K possible actions a_i for $1 \leq i \leq K$, and every action a_i has a probability distribution $D_a^{[0,1]}$ such that $\mu_a = E_{r \sim D_a}[r]$

4.2 Protocol

There are T time stamps $t \in [T]$. The algorithm selects an action $a_t \in A$ at time t and receives a reward $r_t \in [0, 1]$ (where $r_t \sim D_{a_t}$).

4.3 Performance

$$a^* = \operatorname{argmax}_{a \in A}(\mu_a), \mu^* = \max_{a \in A}(\mu_a) \quad (1)$$

$$\begin{aligned} \text{Regret} &= \max_{a \in A} \sum_{t=1}^T r_t(a) - \sum_{t=1}^T r_t(a_t) \\ \text{Pseudo Regret} &= \max_{a \in A} E \left[\sum_{t=1}^T r_t(a) \right] - E \left[\sum_{t=1}^T r_t(a_t) \right] \\ &= \mu^* T - \sum_{t=1}^T \mu_{a_t}(a_t) \\ &= \sum_{a \in A} (\mu^* - \mu_a) E |T_a| \quad (\text{where } T_a = \{t : a_t = a\}) \end{aligned}$$

5 Advanced Stochastic Multi-Armed bandit Algorithms

5.1 Explore-Then-Exploit Method

The explore-exploit algorithm aims to balance between exploring unknown arms to gather information and exploiting the currently best arm to maximize cumulative rewards. The algorithm starts with the exploration, where arms are selected randomly or based on uncertainty estimates to learn their reward distribution. As more data is collected, the algorithm gradually shifts towards exploitation by favoring arms with higher expected rewards. The balance between exploration and exploitation is crucial to achieve the best long-term performance in stochastic arm bandit scenarios.

Algorithm 1 Explore-then-exploit with parameter M .

```

for every action  $a \in A$  do
  for  $M$  times do
    Sample action  $a_t = a$ 
    Get reward  $r_t(a)$ 
    Calculate the  $\bar{\mu}_a = \frac{1}{M} \sum_{t \in T_a} r_t(a)$ , where  $T_a = \{t : a_t = a\}$ .
  end for
end for
Execute  $\bar{a}^* = \operatorname{argmax}_{a \in A} \bar{\mu}_a$  until time  $T$ .

```

Analysis:

Definitions:

$$\bar{\mu}_a = \frac{1}{M} \sum_{t \in T_a} r_t(a) \quad , \quad \Delta_a = \mu^* - \mu_a \quad (2)$$

Figure 1: Confidence bounds on μ .

$$E[\text{Pseudo Regret}] = \sum_{a \in A} \Delta_a M + (T - MK) \sum_{a \in A} \Delta_a \Pr[\bar{a}^* = a] \quad (3)$$

for $\lambda = \sqrt{\frac{8 \log(T)}{M}}$, we get that: $\Pr[|\Delta|_a \geq \lambda] \leq 2e^{-\frac{\lambda^2 m}{2}} = \frac{2}{T^4}$ union bound $\rightarrow \Pr[\exists a : |\Delta|_a \geq \lambda] \leq \frac{2K}{T^4} \leq \frac{2}{T^3}$

If B didn't occur $\mu_{\bar{a}^*} + \lambda \geq \bar{\mu}_{\bar{a}^*} \geq \bar{\mu}_{a^*} \geq \mu_{a^*} - \lambda \rightarrow 2\lambda \geq \mu_{a^*} - \mu_{\bar{a}^*} = \Delta_{\bar{a}^*}$

Then we get: $E[\text{PR}] \leq \underbrace{\sum \Delta_a M}_{[\text{explore}]} + \underbrace{(T - KM) \cdot 2\lambda}_{[\text{B didn't occur}]} + \underbrace{\frac{2}{T^3} \cdot T}_{[\text{B did occur}]} \leq KM + \frac{2T\sqrt{\log(T)}}{M} + \frac{2}{T^2}$

Let's choose $M = T^{\frac{2}{3}} K^{-\frac{2}{3}} \rightarrow E[\text{PR}] \leq K^{\frac{2}{3}} T^{\frac{2}{3}} + K^{\frac{2}{3}} T^{\frac{2}{3}} \sqrt{8 \log(T)} + \frac{2}{T^2}$

- This upper bound is a sub-linear bound unlike the stochastic upper bound which is \sqrt{T}

5.2 Disadvantage of The Explore then Exploit Method

One disadvantage of this method lies in the exploration phase. This phase involves sampling actions to gather information about their reward distributions. During this exploration, the algorithm may spend time sampling actions that have been previously observed to yield low rewards. This is known as the exploration-exploitation trade-off.

In an ideal scenario, the algorithm should use the information gained during exploration to bias its choices towards actions that have shown higher rewards in the past. And the number of samples per action $N_t(a) | a \in A$ should be higher for actions with higher rewards.

$$\text{for } \bar{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{t=1}^T r_t(a) \text{ and } \lambda_t(a) = \sqrt{\frac{8 \log(T)}{n_t(a)}}$$

We would like to show that: $\Pr[|\bar{\mu}_t - \mu_a| \geq \lambda_t(a)] \leq \frac{2}{T^4}$

The only problem is that $n_t(a)$ is a random variable and not a number. To bypass that, we can make a $K \times T$ table. We will sample all values in advanced and fill the table with the values s.t

Table(m, τ) = $R(m, \tau)$ (the reward of taking action m at time t).

$$\hat{V}_m(a) = \frac{1}{m} \sum_{t=1}^m R(t, a)$$

Now that m is a constant, we can calculate an upper bound:

$$\bar{\mu}_t(a) = \bar{V}_{n_t(a)}(a)$$

$$\forall a \forall m \Pr \left[|\bar{V}_m - \mu(a)| \geq \sqrt{\frac{8 \log(T)}{m}} \right] \leq \frac{2}{T^4}$$

A good event G : $\forall a \forall t |\bar{\mu}_t - \mu_a| \leq \lambda_t(a)$, then $\Pr[G] \geq 1 - \frac{2}{T^2}$

+To help balance the exploration and exploitation the of different actions we can make use of Confidence Bounds. The Confidence Bounds algorithm addresses this by considering not just the average rewards observed for each arm but also the uncertainty associated with those averages. It does this by calculating confidence intervals, which provide a range within which the true mean reward is likely to be.

Confidence bounds:

$$UCB_t(a) = \bar{\mu}_t(a) + \lambda_t(a), \quad LCB_t(a) = \bar{\mu}_t(a) - \lambda_t(a)$$

If we assume that event G has occurred, then $\forall a \forall t \mu_a \in [LCB_t(a), UCB_t(a)]$, where $P[G] = 1 - \frac{2}{T^2}$.

5.2.1 Successive Elimination

The idea of this algorithm is to execute all arms in a round-robin fashion, and eliminate sub-optimal arms. The goal is to remain only with candidates for the optimal arm, which include with high probability the optimal arm.

Algorithm 2 Successive Elimination

Initialize: $S \leftarrow A$

while $t \leq T$ **do**

 For each $a \in S$, pull a , i.e. execute a once

 For each $a' \in S$: if $\exists a'' \in S$ s.t. $UCB_t(a') < LCB_t(a'')$, Update $S \leftarrow S \setminus \{a'\}$.

end while

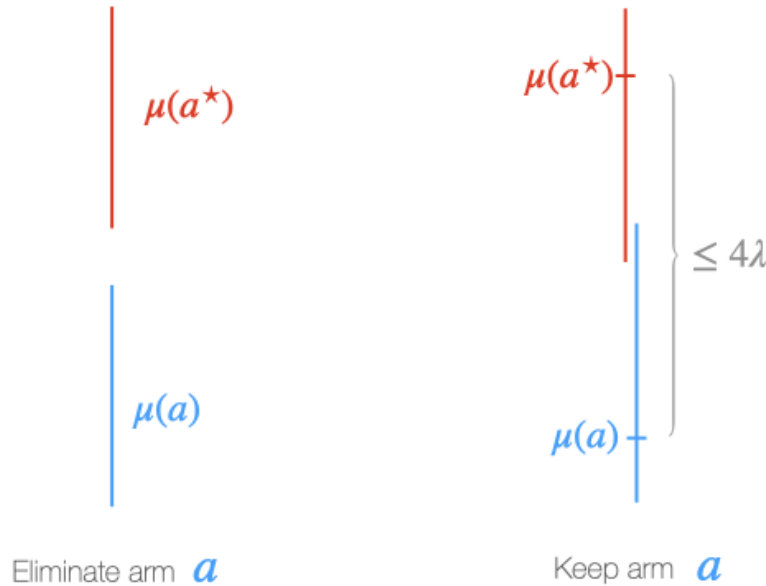


Figure 2: Successive Elimination illustration.

As illustrated in figure above, when we are in the good event G , the intervals $\{[LCB_t(a), UCB_t(a)] : a \in A\}$ are disjoint, hence: if $a', a'' \in A$ are arms such that a' obtains higher

expected value from a'' , then each value for the empirical mean value of the former interval is higher from each empirical mean value of the latter, hence the latter could be eliminated from the set of candidates for the optimal arm.

Two observations are in order;

1. If the good event G holds, then $a^* \in S$ at all times.
2. Any two arms a, a' that have not been disqualified, share the same number of pulls at all times; $n_t(a) = n_t(a')$.

Theorem 5 *The successive elimination algorithm obtains (pseudo) regret of $O(K\sqrt{T} \log T)$.*

Proof: Assume that G occurs. If we pull an active arm a_i , then

$$LCB_t(a^*) \leq UCB_t(a_i),$$

otherwise we would have eliminated a_i . Since G occurs, $LCB_t(a^*) < UCB_t(a_i)$, so:

$$\mu_t(a^*) - 2\lambda_t(a^*) \leq \hat{\mu}_t(a^*) - \lambda_t(a^*) = LCB_t(a^*) < UCB_t(a_i) = \hat{\mu}_t(a_i) + \lambda_t(a_i) \leq \mu_t(a_i) + 2\lambda_t(a_i).$$

Hence, since $\lambda_t := \lambda_t(a^*) = \lambda_t(a_i)$ (this property follows from observation 2 and the definition of $\lambda_t(a)$):

$$\begin{aligned} \Delta(a_i) &:= \mu^* - \mu(a_i) \leq 4\lambda_t = 4\sqrt{\frac{8 \log T}{n_t(a_i)}} \\ \implies n_T(a_i) &\leq \frac{c}{\Delta^2(a_i)} \log T, \end{aligned}$$

for some constant $c \in \mathbb{R}$. Given the above, we obtain

$$\mathbb{E}[\text{P.R.}] = \sum_{a \in A} \Delta(a) \mathbb{E}[n_T(a)] \leq \sum_{a \in A} \frac{c}{\Delta(a)} \log T + \frac{2}{T^2} T,$$

where in the last inequality we have bounded the bad event case with the maximal possible regret T . To derive the final regret bound, consider the two sets

$$\begin{aligned} A_1 &= \{a \mid \Delta(a) \geq 1/\sqrt{T}\}, \text{ and} \\ A_2 &= \{a \mid \Delta(a) < 1/\sqrt{T}\}. \end{aligned}$$

Now

$$\begin{aligned} \mathbb{E}[\text{P.R.}] &\leq \sum_{a \in A_1} \Delta(a) \mathbb{E}[n_T(a)] + \sum_{a \in A_2} \Delta(a) \mathbb{E}[n_T(a)] + \frac{2}{T} \\ &\leq \sum_{a \in A_1} \frac{c}{\Delta(a)} \log T + \frac{1}{\sqrt{T}} T + \frac{2}{T}, & (\sum_{a \in A_2} n_T(a) \leq T) \\ &= O(K\sqrt{T} \log T), & (\forall a \in A_1, \frac{1}{\Delta(a)} \leq \sqrt{T}) \end{aligned}$$

and the proof is complete. \square

5.2.2 The Upper Confidence Bound (UCB) Algorithm

The high level idea here is to assume under-explored arms will produce high reward. This principle is well known and shows up frequently, it is known as “optimism under uncertainty”.

Algorithm 3 UCB

Pull each arm once

On each round t , pull $a_t = \arg \max \text{UCB}_t(a)$.

Theorem 6 *The UCB algorithm obtains (pseudo) regret of $O(K\sqrt{T} \log T)$.*

Proof: As in the proof for the successive elimination algorithm, consider the good event G and assume it occurs. Then for whichever arm a_i we have chosen to pull on round t , we have that

$$\mu(a_i) + 2\lambda_t(a_i) \geq \hat{\mu}_t(a_i) + \lambda_t(a_i) = \text{UCB}_t(a_i) \geq \text{UCB}_t(a^*) \geq \mu^*.$$

This implies

$$\begin{aligned} 2\lambda_t(a_i) &\geq \mu(a_i) - \mu^* = \Delta(a_i) \\ \implies \frac{c \log T}{\Delta^2(a_i)} &\geq n_t(a_i), \end{aligned}$$

where $c \in \mathbb{R}$ is some constant. The rest of the proof proceeds as in the one for successive elimination. \square

It is instructive to ask what is the difference between the SE and UCB algorithms. The main difference is that in SE, exploration is done in the beginning of each phase, in a dedicated manner. In UCB however, exploration is intrinsic in the choice of each action.

6 Sleeping Actions

In this model, at each time step t , a limited set of possible actions $A_t \subset A$ is considered, unlike other models where all possible actions A may be utilized.

Algorithm 4 Sleeping Actions

Create a permutation π_t of A using UCB

Pick the first action in π_t that is also in A_t

$a_t = \operatorname{argmax}_{a \in A_t} (\text{UCB}(a))$

6.0.1 Complexity analysis

Definitions:

$$\Delta_{i,j} = \mu(a_i) - \mu(a_j)$$

let $N_{i,j}$ be the number of times we have selected action j while OPT selected action i

We know that for every time step t that we have selected action j while OPT selected action i . Then we know that $\text{UCB}_t(a_j) \geq \text{UCB}_t(a_i)$ (otherwise we would have chosen a_i)

And we get that

$$\mu(a_j) + 2\lambda_t(a_j) \geq \text{UCB}_t(a_j) \geq \text{UCB}_t(a_i) \geq \mu(a_i) \tag{4}$$

$$2\lambda_t(a_j) = 2\sqrt{\frac{8 \log(T)}{n_t(a_j)}} \geq \mu(a_i) - \mu(a_j) = \Delta_{i,j} \tag{5}$$

$$\frac{c}{\Delta_{i,j}^2 \log(T)} \geq n_t(a_j) \geq N_{i,j} \quad (6)$$

$$E[\text{PR}] = \sum_{i < j} \Delta_{i,j} N_{i,j} \leq \sum_{\substack{i < j \\ [K^2 \text{ pairs}]}} \frac{c}{\Delta_{i,j} \log(T)} + \frac{2}{T} \quad (7)$$

$E[\text{PR}]$ is bounded by $O(k^2)$. Let's try to improve this –

let $M_{i,j}$ be the number of times we played a_j instead of $\{a_1, \dots, a_i\}$. Notice that $N_{i,j} = M_{i,j} - M_{i-1,j}$. Define $M_{0,j} = 0$. and note that $\Delta_{j,j} = 0$.² Then,

$$\text{PR} = \sum_{j=2}^k \sum_{i=1}^{j-1} (M_{i,j} - M_{i-1,j}) \Delta_{i,j} = \sum_{j=2}^k \sum_{i=1}^{j-1} M_{i,j} (\Delta_{i,j} - \Delta_{i+1,j}) = \sum_{j=2}^k \sum_{i=1}^{j-1} M_{i,j} \Delta_{i,i+1}$$

For $1 \leq k \leq i$:

$$\mu(a_j) + 2\lambda_t(a_j) \geq UCB_t(a_j) \geq UCB(a_k) \geq \mu(a_k) \geq \mu(a_i)$$

$$\frac{c}{\Delta_{i,j}^2 \log(T)} \geq n_t(a_j) \geq M_{i,j}$$

New upper bound:

$$E[\text{PR}] \leq \sum_{j=2}^k \sum_{i=1}^{j-1} \frac{c}{\Delta_{i,j}^2 \log(T)} \Delta_{i,i+1} \quad (8)$$

²This implies that for $i = 1 : M_{i-1,j} \Delta_{i,j} = M_{0,j} \Delta_{i,j} = 0$ and for $i = j - 1 : M_{i,j} \Delta_{i+1,j} = M_{j-1,j} \Delta_{j,j} = 0$

Lemma 7

$$\sum_{j=2}^k \sum_{i=1}^{j-1} \frac{\Delta_{i,i+1}}{\Delta_{i,j}^2} \leq 2 \sum_{j=2}^k \frac{1}{\Delta_{j-1,j}}$$

Proof:

we'll define $f(x) = \frac{1}{(x - \mu_j)^2}$ and find an upper bound for $\sum_{i=1}^{j-1} f(x_i)(x_i - x_{i+1})$

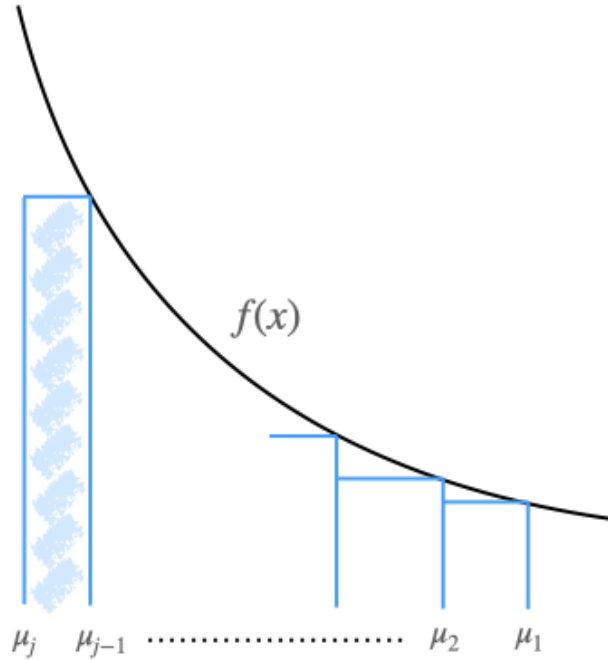


Figure 3: Bounding $f(x)$.

As illustrated above: since $f(x)$ is strictly decreasing in the interval $(u_j, u_1]$ then the integral of $f(x)$ in the interval $[u_{j-1}, u_1]$, which is the size of the area under $f(x)$ in the latter interval, is at least the sum of the areas of the rectangles of sizes $\{f(x_i) \times (x_i - x_{i+1}) : i \in [j - 2]\}$ where $\forall i \in [j - 1].x_i = \mu_i$. Thus:

$$\sum_{i=1}^{j-1} \frac{\Delta_{i,i+1}}{\Delta_{i,j}^2} \leq \int_{\mu_{j-1}}^{\mu_1} f(x)dx + f(\mu_{j-1})(\mu_{j-1} - \mu_j) \leq \frac{1}{\mu_{j-1} - \mu_j} + \frac{1}{\mu_{j-1} - \mu_j} = \frac{2}{\mu_{j-1} - \mu_j}$$

where

$$\int_{\mu_{j-1}}^{\mu_1} f(x)dx = \left[-\frac{1}{x - \mu_j} \right]_{\mu_{j-1}}^{\mu_1} = \frac{1}{\mu_{j-1} - \mu_j} - \frac{1}{\mu_1 - \mu_j} \leq \frac{1}{\mu_{j-1} - \mu_j}$$

□

References

- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.
- J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.
- H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
- A. Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.
- W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.