Introduction to Reinforcement Learning (RL) || EE675A || 2022-2023-II

Lecture - 7 (30th January 2023)
Formal proof for lower bound on regret for Bandit Problem

Lecturer: Prof. Subrahmanya Swamy Peruru Scribe: Nitish Kumar and Sahil Maurya

1 Recap

In the previous lecture, we saw the lower bound of regret for the multi-armed bandit problem as

$$\min_{algo} \max_{env} E[R(T|\text{env, algo})] \geq \Omega(\sqrt{KT})$$

We prove this regret using hypothesis testing where we assumed two very similar environments, and the best arm in one environment (system) is the worst possible arm in another environment (system).
Note:- Two environments being very similar means that they have very similar statistics.
After this, we saw Bretagnolle-Huber(B-H) inequality for two probability distributions as

$$P(A^c) + Q(A) \geq \frac{1}{2} exp(-\text{KL}(P, Q)), \forall \text{ A}$$

Where P & Q be two probability distributions on the same sample space.
Till now proven, the lower bound of regret for multi-armed bandit (MBA) problems using assumption and hypothesis. Now, in this lecture will see the formal proof of the lower bound on regret for multi-armed bandits (MBA). Then will also look at some theorem for finding the KL divergence between two distributions named divergence decomposition and also try to formalize how we find the lower bound regret for more than two arms.

2 Formal proof of lower bound on regret

We will start the proof with the same as the previous lecture, then try to find where the loose argument was made and work on that to prove it formally.
Let's say that we have two environments for 2-arms bandits. Each arm has a different Gaussian distribution. The best arm in the first environment will be similar to the worst in the second environment.

The two environments are:

$$Env1 := \{\text{a}_1 \sim \mathcal{N}(\Delta, 1) \text{ and } \text{a}_2 \sim \mathcal{N}(0, 1)\}$$

$$Env2 := \{\text{a}_1 \sim \mathcal{N}(\Delta, 1) \text{ and } \text{a}_2 \sim \mathcal{N}(2\Delta, 1)\}$$

Regrets in environments:
Let's say we have total T samples and in those we have to find regret of environment-1. We will look in those T samples to see how many times we have played arm 2, i.e. the times we didn't play the optimal arm.

$$R_{Env1} = \mathbb{E}_{env1}[(T - n_T(a_1))\Delta]$$

Similarly, in the environment-2 out of T sample, the regret will be whenever we played arm 2, i.e., not the optimal arm

$$R_{Env2} = \mathbb{E}_{env2}[n_T(a_1)\Delta]$$

We assume both environments are very similar when $\Delta \to 0$, and it's very likely to get the same samples from both environments, so
$\mathbb{E}_{env1}[n_T(a_1)] \simeq \mathbb{E}_{env2}[n_T(a_1)] = x \longmapsto$ **[This is the loose argument that we assumed]**
Now let's minimize the regret

$$\min_x \max\{(T - x)\Delta, \Delta x\}$$

$$x^* = \frac{T}{2}$$

$$\Delta \simeq \frac{1}{\sqrt{T}}$$

$$Regret \simeq \Omega(\sqrt{T})$$

2.1 Proving the regret preciously without making loose argument

Now, let's say a(t) is the arm played at t turn and r(t) is the reward received at turn t after playing an arm. The probability $P(a(1), r(1), a(2), r(2), ...a(T), r(T))$ of seeing this sequence is function of both **environment** and **algorithm** used.

Notations:

$$Environment1 := \nu_1$$

$$Environment2 := \nu_2$$

$$Algorithm := \pi$$

Probabilities of seeing the sample specified above in both environments:

$$\mathbb{P}_{\nu_1,\pi} = \pi(a(1)) \cdot \mathbb{P}_{\nu_1}(r(1)|a(1)) \cdot \pi(a(2)|a(1), r(1)) \cdot .....$$

$$\mathbb{P}_{\nu_2,\pi} = \pi(a(1)) \cdot \mathbb{P}_{\nu_2}(r(1)|a(1)) \cdot \pi(a(2)|a(1), r(1)) \cdot .....$$

To formally understand how close are the sample sequences observed in environment 1 and environment 2 with algorithm π, we use $KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi})$ and for finding this will use.

For continuous distribution, KL diversion of them is shown below:

$$KL(f(x), g(x)) = \int f(x) log \frac{f(x)}{g(x)} \, dx$$

But in our case, the distributions $\mathbb{P}_{\nu_1,\pi}$ and $\mathbb{P}_{\nu_2,\pi}$ are neither discrete nor continuous as it will be discrete if arms are discrete and continuous if the arms are gaussian distribution. So we can't apply the above KL divergence formula.

2.2 Divergence Decomposition Theorem

Theorem: Divergence Decomposition: : Let there be K arms: $A = \{1, 2, 3, ...k\}$

$$KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi}) = \sum_{a \in A} \mathbb{E}_{\nu_1,\pi}[n_T(a)] \cdot KL(D_a^{\nu_1}, D_a^{\nu_2})$$

where, $D_a^{\nu_i}$: Reward distribution of arm a in environment i. This theorem helps us compute $KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi})$ in terms of KL divergence between reward distributions.

2.3 Computing $KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi})$

Applying the divergence theorem for environment 1, environment 2:

$$\nu_1 = \{\mathcal{N}(\Delta, 1), \mathcal{N}(0, 1)\}$$

$$\nu_2 = \{\mathcal{N}(\Delta, 1), \mathcal{N}(2\Delta, 1)\}$$

To find the KL divergence between reward distributions will use the formula as they are continuous:

$$KL(D_{a_1}^{\nu_1}, D_{a_1}^{\nu_2}) = \frac{(\mu_a - \mu_b)^2}{2\sigma^2}$$

Here in our case, $\sigma = 1$. So,

$$KL(D_{a_1}^{\nu_1}, D_{a_1}^{\nu_2}) = 0$$

$$KL(D_{a_2}^{\nu_1}, D_{a_2}^{\nu_2}) = \frac{(2\Delta)^2}{2} = 2\Delta^2$$

Now applying the divergence decomposition theorem, we have:

$$KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi}) = \sum_{a \in A} \mathbb{E}_{\nu_1,\pi}[n_T(a)] \cdot KL(D_a^{\nu_1}, D_a^{\nu_2})$$

$$KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi}) = 0 + \mathbb{E}_{\nu_1,\pi}[n_T(a_2)](2\Delta^2)$$
$$KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi}) = \mathbb{E}_{\nu_1,\pi}[n_T(a_2)](2\Delta^2)$$

2.4 Calculating the lower bound of regret

Now, Recall **B-H inequality** from previous lecture:

$$P(A) + Q(A^c) \geq \frac{1}{2} exp(-KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi})) \quad \forall A$$

Let us define an event A, such that:

i. A is the bad event in environment 1
ii. A^c is the bad event in environment 2

$$A = \{n_T(a_1) \leq \frac{T}{2}\}$$
$$A^c = \{n_T(a_1) > \frac{T}{2}\}$$

$$\mathbb{P}_{\nu_1,\pi}(A) + \mathbb{P}_{\nu_2,\pi}(A^c) \geq \frac{1}{2} exp(-KL(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi})) = \frac{1}{2} exp(-\mathbb{E}_{\nu,\pi}[n_T(a_2)]2\Delta^2)$$

Now let us represent the regret in terms of these events A and A^c

$$Env1 := \mathbb{E}[R(T; \nu_1, \pi)] \geq \frac{\Delta T}{2} \mathbb{P}_{\nu_1,\pi}(A)$$

$$Env2 := \mathbb{E}[R(T; \nu_2, \pi)] \geq \frac{\Delta T}{2} \mathbb{P}_{\nu_2,\pi}(A^c)$$

$$\mathbb{E}[R(T; \nu_1, \pi)] + \mathbb{E}[R(T; \nu_2, \pi)] \geq \frac{\Delta T}{2} (\mathbb{P}_{\nu_1,\pi}(A) + \mathbb{P}_{\nu_2,\pi}(A^c))$$

$$\mathbb{E}[R(T; \nu_1, \pi)] + \mathbb{E}[R(T; \nu_2, \pi)] \geq \frac{\Delta T}{2} (\frac{1}{2} exp(-(\mathbb{E}_{\nu_1}[n_T(a_2)]2\Delta^2)))$$

A simple fact:

$$x + y \geq z$$
$$\implies \max\{x, y\} \geq \frac{z}{2}$$

Using this fact,

$$\max\{\mathbb{E}[R(T; \nu_1, \pi], \mathbb{E}[R(T; \nu_2, \pi]\} \geq \frac{\Delta T}{8} exp(-\mathbb{E}_{\nu_1}[n_T(a_2)]2\Delta^2)$$

Using that,

$$\mathbb{E}_{\nu_1}[n_T(a_2)] \leq T$$

Using the above inequality,

$$\max\{\mathbb{E}[R(T;\nu_1,\pi],\mathbb{E}[R(T;\nu_2,\pi]\} \geq \frac{\Delta}{8} exp(-T2\Delta^2)$$

Choosing $\Delta = \frac{1}{2\sqrt{T}}$

$$\max\{\mathbb{E}[R(T;\nu_1,\pi],\mathbb{E}[R(T;\nu_2,\pi]\} \geq \frac{T}{16\sqrt{T}} exp(\frac{-1}{2})$$

$$\max\{\mathbb{E}[R(T;\nu_1,\pi],\mathbb{E}[R(T;\nu_2,\pi]\} \geq \Omega(\sqrt{T})$$

Hence, Proved

2.5 Method for calculating lower bound regret for $K \geq 2$

In this case, we also will take the two environments:-

2.5.1 Environment-1:

In this environment, every arm has gaussian distribution with variance one, and one out of K arms has mean $\mu = \Delta$ and the other has mean $\mu = 0$

$$Env1 := (\mathcal{N}(\Delta,1),\mathcal{N}(0,1),\mathcal{N}(0,1),\mathcal{N}(0,1),\mathcal{N}(0,1),.....)$$

2.5.2 Environment-2:

In this environment, every arm has gaussian distribution with variance one, and one out of K arms has mean $\mu = 2\Delta$ and also, one arm has mean $\mu = \Delta$, and the remaining K-2 arms will have mean $\mu = 0$

$$Env2 := (\mathcal{N}(\Delta,1),\mathcal{N}(0,1),.....,\mathcal{N}(2\Delta,1),\mathcal{N}(0,1),\mathcal{N}(0,1),.....)$$

2.5.3 Key Idea for distributions in each environment

- Let's say arm $\hat{a}$ is the arm which is least played in env-1 while using policy π, make that the best arm in env-2.
- For exact details of the proof for $K \geq 2$ case, please read **Bandit Algorithm** book chapter 15.

References

1. Lectures notes by "Prof. Sanjay Shakkottai"
2. Chapters 14 and 15 from "Bandit Algorithms" book by T. Lattimore, C. Szepesvari