Lecture 6: EE675A Introduction to Reinforcement Learning

25/01/23

Lecturer: Prof. Subrahmanya Swamy Peruru *Scribe: Akshat Sharma, Harsh Garg*

1 Recap

In the previous lecture we discussed about generalized Thompson Sampling ,which involved applying Gaussian TS with no knowledge of the underlying model dynamics. We also discussed the concept of bayesian regret, where environments less likely to be encountered are given less weight in the regret calculation. Recall that Bayesian regret $= E_{\text{env} \sim \text{prior}}[R(T|\text{env, algo})]$. In this lecture, we discuss the preliminaries required to derive the lower bound for regret for multi-armed bandit (MBA) problems.

2 Lower bound on regret

The lower bound on regret is motivated as follows: A lower bound on regret $f(T)$ is meaningful for any chosen algorithm/strategy (π), there exists an environment (env) on which the algorithm has regret of the order $f(T)$. There is no universal strategy, i.e., no π which performs well on all environments. We will proceed to show that the lower bound is of the form $\Omega(\sqrt{KT})$. Mathematically, for every algorithm π there exists an environment (env) such that,

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

The argument can also be written in the form of a min-max objective over the regret.

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

Maximum over (env) means selecting the worst-case environment for a certain algorithm and minimum over algorithm means finding the algorithm with the least such regret. This gives the lower bound on regret. Some results from hypothesis testing are required to prove this result.

3 Hypothesis Testing

3.1 Introduction

Let's define two distributions: $P \sim \mathcal{N}(0,1)$ and $Q \sim \mathcal{N}(\delta,1)$. The hypothesis which we wish to test is whether some observed data $(x_1, x_2, \cdots x_T)$ is drawn from the distribution P or from Q. A trivial approach could be to compare the sample mean with the mean of $P(0)$ and $Q(\delta)$ and return the distribution with the closer mean. However, this approach does not consider the uncertainty involved in the empirical mean due to finite number of samples (T). We may make an error if the empirical mean lies on the wrong side of $\delta/2$. It can be shown that for any T, there exists a small enough δ, such that there is a constant chance of making an error (picking the wrong distribution).

$$\text{if } \delta < \frac{1}{\sqrt{T}};\ P(\text{making an error}) \geq e^{\frac{-1}{8}} \tag{1}$$

3.2 Bretagnolle-Huber inequality

To help us hypothesize which distribution these samples belong to, let's define an event A, upon occurring which we predict that samples belong to distribution P.

$$\begin{aligned} A &= \{\text{sample mean} < a\}, \text{for some } a \\ A^c &= \{\text{sample mean} \geq a\} \end{aligned}$$

For e.g, a reasonable selection of a would be $\frac{\delta}{2}$ since we have no other information.
We now further define the probabilities of 2 event:

$$\begin{aligned} P(A^c) &= \mathbb{P}(\text{wrong predictions} \mid \text{samples from P}) \\ Q(A) &= \mathbb{P}(\text{wrong predictions} \mid \text{samples from Q}) \end{aligned}$$

If $P = Q$, then it is impossible to predict whether the sample came from P or Q since both the distributions are identical. Best strategy would be to randomly pick a distribution, i.e.

$$P(A^c) = \frac{1}{2} \text{ and } Q(A) = \frac{1}{2}$$

But when both the distributions are not identical, we will try to use the below inequality which lower bounds these probabilities using $KL(P||Q)$.

Theorem: *Bretagnolle-Huber inequality: Let P & Q be two probability distributions on the same sample space, then for any event A, we have*

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

4 Sketch of regret analysis

Let there be an environment (env) with two arms with the reward distributions being $D_{a_1} = \mathcal{N}(0,1)$ and $D_{a_2} = \mathcal{N}(\delta,1)$. Even after some 'free' trails (exploring without paying any regret) the algorithm may still get confused between the two arms with a constant probability as shown in the section above. Hence after exploring for 'T' rounds and playing another 'T' actual rounds, using equation (1) we can write the expected regret as

$$T \cdot \delta \cdot \mathbb{P}(error) \simeq T \cdot \frac{1}{\sqrt{T}} \cdot e^{-\frac{1}{8}} = O(\sqrt{T})$$

Note:

- The above argument is not rigorous, one can think of several issues such as why are we seeing only one of arm samples in "free" trials, etc.
- The above argument is just given to give a feel of where does the term $\sqrt{T}$ come in the picture of regret lower bound.

A refined argument: Since we are interested in bounds for the "worst-case" environment and showing that no algorithm can do better than $O(\sqrt{kT})$ in such an environment, we are going to do the following: We are going to construct two environments and say that in one of the two environments, every algorithm will incur $\Omega(\sqrt{kT})$ regret.
The two environments are:

$$Env1 := \{\mathrm{a}_1 \sim \mathcal{N}(\delta,1) \,\text{and}\, \mathrm{a}_2 \sim \mathcal{N}(0,1)\}$$
$$Env2 := \{\mathrm{a}_1 \sim \mathcal{N}(\delta,1) \,\text{and}\, \mathrm{a}_2 \sim \mathcal{N}(2\delta,1)\}$$

Observe that these two environments satisfy two properties:

- $Env1$ and $Env2$ are statistically similar (assuming δ is small)
- The best possible arm in $Env1$ is the worst arm in $Env2$.

Important Observations: The decision of what actions to take at time $'t'$ depends only on the actions & rewards played/observed till time '$(t-1)'$. i.e. $a(t)$ depends only on $(a_1, r_1, a_2, r_2, ..., a_{t-1}, r_{t-1})$.
As the environments are quite similar, one can expect the same series of actions and rewards to be encountered in both environments. Hence the number of times an action (a_1) is taken in (env1) and (env2) is the same.

$$i.e., \ E_{env1}[n_t(a_1)] \simeq E_{env2}[n_t(a_1)] = x$$

Hence the regret in environment 1 can be written as $E[R|\text{env1}] = \delta \cdot (T - E_{\text{env1}}[n_t(a_1)])$ and the regret in environment 2 can be written as $E[R|\text{env2}] = \delta \cdot (E_{\text{env2}}[n_t(a_1)])$. The worst-case regret could be the minimum of the two, i.e.

$$\text{Worst-case regret } = \inf_x \ \max \, (T - x, x) \cdot \delta$$

Therefore the worst-case regret occurs at $x = T/2$, and is equal to $\delta T/2 = \sqrt{T}/2 = O(\sqrt{T})$

Note:

- Again, these arguments are not very rigorous since we assumed $E_{env1}[n_T(a_1)] \simeq E_{env2}[n_T(a_2)]$
- In the next lecture, we will use the notion of KL divergence to make these arguments rigorous.

References

1. Chapter 13, 14 from “Bandit Algorithms” book by T. Lattimore, C. Szepesvari
2. Lectures notes by “Prof. Sanjay Shakkottai”