Bandit Algorithm (Online Machine Learning)

Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 38
Proof Idea of Lower Bounds – I

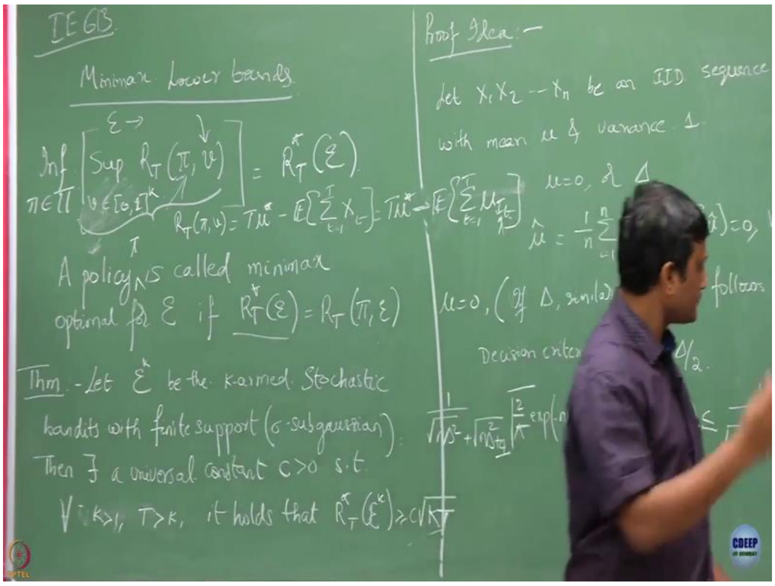
So, far we have looked into both this adversarial and stochastic sorting right. In both the cases, we; our focus so far has been only just coming up with the algorithm and bound its regret. So, now the question is whether whatever the bounds we were getting is that the best we could get or like whatever the performance we got this some algorithms; they were not optimal or they have been very suboptimal? So, we to need to characterize that, we need to also understand what is the best we can get in this set ups right?

So, what is the lower bound on this regret? Is that regret upper bounds we got for this different algorithm, how close they are to the best bounds we can get or are they matching the lower bounds? If they are matching the lower bounds, then our algorithms are already optimal.

Let us say, if matching up to a constant and matching we only look at in the terms that involve the number of rounds and the number of arms. So, for that we are now going to start looking into the lower bounds that one can get in this different settings and then we will discuss with our; what all the things we got they were optimal or not.

So, today we are only going to basically just briefly discuss the ideas and introduced some relevant results that need to prove the lower bound. And lower bound actual one we will prove to in the next class, but this class we will just discuss the broad idea, how we are going to derive this lower bounds.

(Refer Slide Time: 02:17)



IE 613

Minimax Lower bounds

$$\inf_{\pi \in \Pi} \left[\sup_{\nu \in [0,1]^K} R_T(\pi, \nu) \right] = R_T^*(\mathcal{E})$$

$R_T(\pi, \nu) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right]$

A policy π is called minimax optimal for $\mathcal{E}$ if $R_T^*(\mathcal{E}) = R_T(\pi, \mathcal{E})$

Thm. - Let $\mathcal{E}^K$ be the K-armed Stochastic bandits with finite support (σ-subgaussian). Then $\exists$ a universal constant $c > 0$ s.t $\forall K > 1$, $T > K$, it holds that $R_T^*(\mathcal{E}^K) \geq c\sqrt{KT}$

Proof Idea :-

Let $X_1, X_2 \cdots X_n$ be an IID sequence with mean μ & variance 1.

$\mu = 0$, or Δ

$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}$

$\mu = 0$, (If Δ, similar

follows

Decision criter

$\Delta/2$.

$\frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 4}}\sqrt{\frac{2}{\pi}}\exp($

So, we how defined for a; the expected regret of an algorithm pi for a given environment v; we have already defined, this is to be the expected regret of algorithm pi on my problem instance v. Now, nu is chosen by? The environment against whom you are learning, pi is the learner who is applying the policy; the policy we are denoting it as pi. So, there is an interaction going between learner and the environment.

So, now what we are interested is; what is the best we could do against the environment? So, environment can choose any probability distribution it likes right ok. So, now I am looking at; so the regret for a given policy pi, what is the worst regret I could get among all possible environment? So, I am assuming that this epsilon denotes the class of distributions that my environment choose, pick and assign the rewards to arms.

So, for example, this capital epsilon could be all possible distribution with bounded support. It can choose the environment from this set and assign rewards to arm, I am now looking at what is the worst regret I am going to get, but now what is your goal? Your goal is to minimize this and come up with a policy that is best. So, what you want to do? You want to minimize this, let us say pi is; capital pi is collections all policies that you can apply.

Now you want to; so this is the worst regret, if you are going to use a policy pi; since environment can choose any nu that would like from this set epsilon. You are; if I am going to apply policy pi, you are going to see what is the worst I am going to incur and now your goal is to come up with a pi which is going to minimize this.

So, that is why we are going to call this as minimax criteria and then we are going to denote it as; so notice that this depends on what environment you are looking at and also the time horizon over which you are looking at.

Now, a policy; so a policy pi, we are going to call it as minimax optimal; for this environment epsilon if the regret of that policy. So, is going to be whatever the optimal I can get. So, here it is not a nu, but this is function of the environment itself ok. So, we do not know such a policy exists or not right now because the right now do not know what is this quantity.

Student: (Refer Time: 06:23).

So, of course, we know already many policies for example, in the adversarial case we have already looked into EXP 3, EXP.IX, EXP3.P and all these are all different policies.

We do not know right now those policies are optimal because we do not know what is this value and in the stochastic case we know that you have already know the policy like UCB, KLUCB and Thompson sampling and another is MOSS policy; we have already discussed. We know there; what is the regret bound they are going to give but we do not know what is this. So, now, our goal is to see whether how to characterize this quantity R T pi star.

So, first I am going to just a state a result, give a bound on this minimax regret and after that we are just going to today; mostly we are going to talk heuristically, what line of proof we are going to follow to prove this theorem ok.

So, just first let me write this theorem; so this is square root KT; T is capital T. So, let us say epsilon K be a K armed stochastic bandit with finite support. Finite support like; let us say we will assume in this case it is between 0, 1 interval or we can also take it to be some sigma sub Gaussian for some known sigma value.

What we were then saying is; then there exists an universal constant C; when I say universal constant C, that is independent of my K and T such that for every K greater than 1, that is number of arms is greater than 1 and number of rounds greater than k, that is I have; I am round such that, it is larger; it allows me to pull each arm at least once.

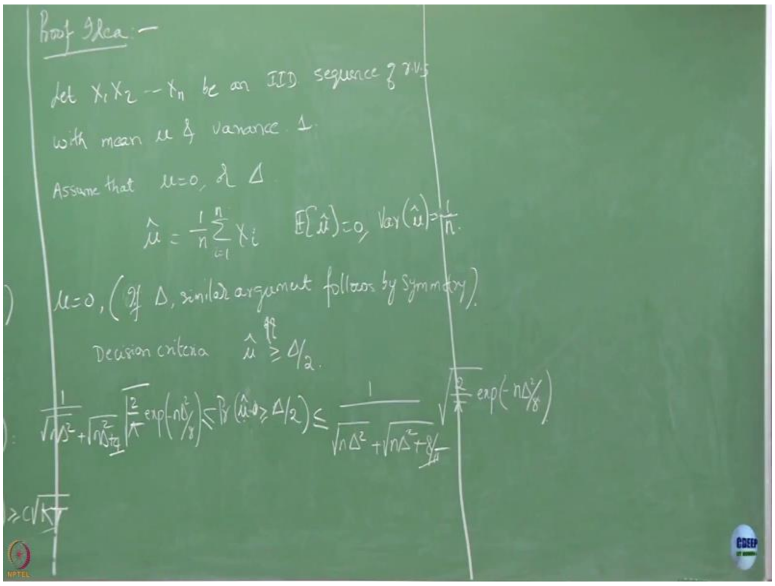
Then it holds that this minimax regret here is going to the lower bounded by C times square root of T K; that is it T is to be at least of order square root K T; that is if you are going to, it is; so you see that it is square root in K and also square root in T and this other parts is just a constant; it does not depend on any of this constant K and T.

So, as of now the way we have written this lower bound; we have taken; the way we have written this minimax regret, we have written it; we have taken over all possible problem instances right. So, this is we are kind of already made it independent of which problem is chosen from the set of problems we have. So, this is in a way we are looking for problem independent bound right.

So, it does not depend on which instance we have and now we are saying that if that; in that case the best you could expect for any policy to do is; is to get a have a bound that is of order square root T K ok; it cannot be smaller than this because we are already saying that it; this result is already saying that it has to be at least square root T K. Now, for our algorithms at least like for the UCB and MOSS algorithm, we have stated this bounds right which problem independent bounds.

What is that we have for UCB, the problem independent bound? Was it like a square order square root K T log T? So with there we had an extra factor of log square root log T; right ok?

(Refer Slide Time: 10:58)



Proof Idea:-

Let $X_1 X_2 -- X_n$ be an IID sequence of r.v.s

with mean μ & variance 1.

Assume that $\mu = 0$, & Δ

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E[\hat{\mu}] = 0, \; Var(\hat{\mu}) = \frac{1}{n}.$$

$\mu = 0$, (If Δ, similar argument follows by symmetry).

Decision criteria $\hat{\mu} \overset{?}{\geq} \Delta/2$.

$$\frac{1}{\sqrt{n\Delta^2}+\sqrt{n\Delta^2+4}}\sqrt{\frac{2}{\pi}}\exp(-n\Delta^2/8) \leq \Pr(\hat{\mu} \geq \Delta/2) \leq \frac{1}{\sqrt{n\Delta^2}+\sqrt{n\Delta^2+8/\pi}}\sqrt{\frac{2}{\pi}}\exp(-n\Delta^2/8)$$

$\gtrsim c\sqrt{KT}$

Let us have this, we will just or going to just discuss proof idea; many of the points we are going to have is pretty much hand waved, may not be that rigorous, but let us try to understand what is; what are the broad steps we are going to take and we will make them later formal.

So, first thing; so now, let us focus on the case of Gaussian bandits; when I say Gaussian bandits, let us assume the; all the arms have rewards that are Gaussian, but maybe with the different different means and variance. Or let us say, let us fix a variance and let us say that all are Gaussian distributed with a fixed variance, but they have different different means.

So, let be an IID with let us say mean mu and let us say variance is 1. Let us say I have I am just taking let us say one arm; let us say it has mean mu and variance 1. Now, suppose assume that mu takes value 0 or delta; only two values.

Now you have been told that like see, I have a given you a Gaussian; I have given you arm which has this Gaussian rewards generated according to Gaussian distributed. It has mean which could be either 0 or delta and it has a fixed variance 1, now how you are going to decide whether it has a mean 0 or mean delta?

Student: (Refer Time: 13:27).

So, let us say you have n samples; we are going to estimate the samples from this samples. So, this is your natural estimator here, now you are going to say its mean how you are going to conclude is; mean is 0 or delta?

Student: (Refer Time: 13:53) difference from 0 or delta.

Ok. So you have this, if it is closer to 0; then you are going to call the mean is 0, if it is closer to delta; we are going to call it delta. So, maybe you are; your decision boundary is delta by 2, if this mean happens to be less than delta by 2; you are going to call it as 0, if it is above delta by 2; you are going to call it as delta ok.

Student: Some error (Refer Time: 14:21).

There could be some error yes, so that is what you then you can potentially make a mistake with this finitely many samples let us see, what is that probability of making an error. So, this we have been told and then your decision was that; let us assume that now mu is 0, the true value happens to be 0. I mean you have been told either it is mu or delta; you have given that information, but the true value happens to be 0. Now, you have to figure out you have to figure out whether it is 0; from the information that it is 0 or delta.

If it is a delta, if it is delta; you can do similarly, argument follows by symmetry. Now, you have fixed mu to be 0; your estimator mu hat, what is the expected value of your estimator? It is going to be 0 right because the underlying; so its expected value is and what is its variance? Variance is going to be simply 1 by n right because these are all independent samples; variance of.

Now, your decision criteria is mu hat greater than delta by 2; you are going to always have this question whether mu hat is going to be delta by 2? If it is true, then you are going to say mean is delta otherwise you are going to say mean is 0. Now, probability that this indeed holds because your mu hat is a random variable, this may hold and what is the probability that this will hold?

So, we can compute upper bound is in a way; we did earlier by looking at the tail probabilities of the Gaussian distributions. So, mu hat here with this estimator; mu hat happens to be another Gaussian random variable with mean 0 and variance 1 by n.

So, for that Gaussian random variable, we are asking; what is the probability that it takes value larger than delta by 2? And based on the tail distribution of a Gaussian random variable, we can compute it as and this we did in our class when we and also we can actually lower bound it by a very similar quantity.

So, the lower bound; we did not discuss, but assume that this is indeed true, assume that like there is an upper bound and the corresponding lower bound on this probability; again you can show this. So, notice that upper bound and lower bound are essentially same, the only difference is happening at this constant; here it is 4 and here it is 8 by pi; everywhere it is almost same.

Now, you see that both terms have an exponential term and their decaying exponentially in the number of samples if we have. If n is large enough, this probability is very small that you are making such a mistake. So, I have basically this is nothing, but mu hat minus 0 here, that is your; your estimate deviates from the true value, but since I am assumed; I am just ignoring, I have just ignored that 0 there.

So, if n is large; upper bound and lower bound are going to be very small, but now the case where you see that all the places, this n is appearing with delta square; n delta square, n delta square, n delta square here.

If instead; if n is large, it is not the only thing here that determines this probability, but what is determining is; whether this n delta square is large or small. It may happen that delta is very small, even my n is very large; in that case n delta square is small and this probability may be random with not so small ok.

So, if this let us say for in our case, we have fixed delta right let us say fix a delta and if your n is not large enough such that, this n delta square is still significantly big than this probabilities may be not so small and then there is a significant probability that you make a made an error right.

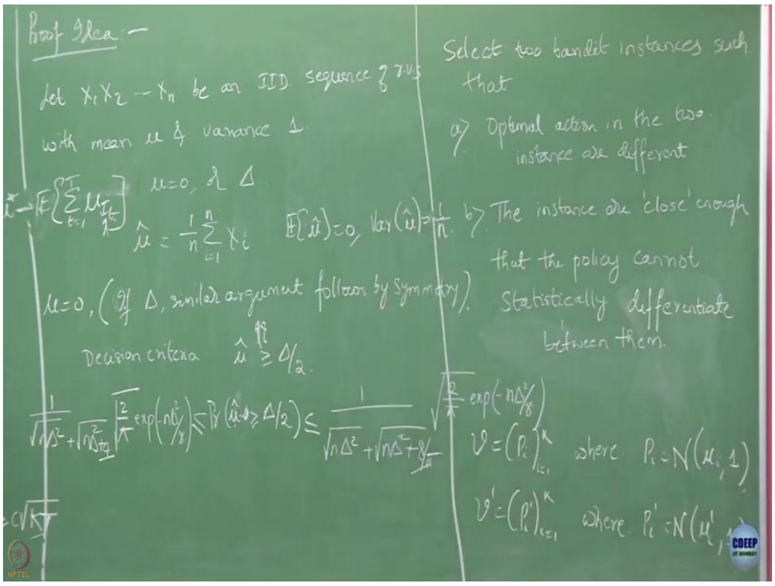
So, for fix a delta and if your n is not large enough; let us say this n delta square is relatively big or relatively small sorry; in that case this probability could be still not so small and there is a good amount of possibilities are your decisions would have been wrong. So, you see that in such cases; if your number of samples are not too many, you could potentially make mistakes with very good amount of probability.

Alternatively, if your delta is very very small; there is even if with you have large number of sample, it is potentially possible that you would have made and mistake in your judgment with good possibility ok. Now, so the general idea in coming up with this proof is to establish a scenario that you have two instance of bandits on which your algorithm cannot distinguish them, even though these two instances have different optimal actions.

So, let us say if we can construct such a scenario; then whatever algorithm you have it will be full right. So, what I am saying? Suppose, I can come up with two environments; both the environments are different may be slightly different but and both of them look very similar. They are different in the sense they have different optimal actions, but they look very similar in which case your algorithm may end up confusing one environment with the other and end up playing the optimal arm.

So, if you can construct such an environment; then you will be able to show that your algorithm is going to make mistakes on this case and then based on that, you will be able to have certain lower bound ok.

(Refer Slide Time: 24:00)



Proof Idea :-

Let $X_1 X_2 \dots X_n$ be an IID sequence of r.v.s with mean μ & variance 1.

$i^* \rightarrow \mathbb{E}\left\{\sum_{t=1}^{T} \mu_{I_t}\right\}$ $\mu = 0,$ & Δ

$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\quad \mathbb{E}[\hat{\mu}] = 0, \; Var(\hat{\mu}) = \frac{1}{n}$

$\mu = 0,$ (If Δ, similar argument follows by symmetry).

Decision criteria $\hat{\mu} \geq \Delta/2$.

$$\frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 4}}\sqrt{\frac{2}{\pi}}\exp\left(-n\Delta^2/8\right) \leq \Pr\left(\hat{\mu} - 0 \geq \Delta/2\right) \leq \frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 8/\pi}}\sqrt{\frac{2}{\pi}}\exp\left(-n\Delta^2/8\right)$$

$c\sqrt{KT}$

Select two bandit instances such that

a) Optimal action in the two instance are different

b) The instance are 'close' enough that the policy cannot statistically differentiate between them.

$\nu = (P_i)_{i=1}^{K}$ where $P_i = N(\mu_i, 1)$

$\nu' = (P_i')_{i=1}^{K}$ where $P_i' = N(\mu_i', 1)$

So, let us say I want to basically select two bandit instance such that. So, what we are saying? Suppose, let us say I can construct two instances of bandits; that means, I had to come up with two environments v and let us say v prime such that the two instances have two different optimal actions and also these two instances should be close enough, such that your policy cannot statistically differentiate between them.

So, then here it is a possibility that my algorithm will confuse one instance with the other and if it confuses one instance to the other, then the optimal actions; it thinks is actually not the optimal action right. So, right now I am not worried about the actual regret ok, yes if the gap between small; regret is going to be small.

What I am worried about is whether my algorithm is eventually identifying the current optimal action or not or like is it possible that with a finite T; it may happen that I can fool I give one instance for algorithm my algorithm, but my algorithm thinks it is the other instance and with the optimal action is the different one and it ends up playing up not the optimal action for that instance.

So, this is of two things two instance; so this all kind of what I am asking the two conditions I am asking they are somewhat conflicting right? I want dumb to be statistically close enough or not able to be differentiated; that means, they are kind of very identical right like very closely, if they are close maybe their optimal actions also need to be same.

Now, what I am asking is; if at all you can construct two instances which are statistically; we will try to make it more formal what we mean by looking statistically similar, why the optimal actions in them are different? If at all this can happen, now your it is likely that your algorithm is going to confuse between the two.

And if and in this cases if it happens to make a and it will be making bad choices and in that case it may be will show that in that case the regret is going to be of this sort. And notice that if I had to make them; even though it made a bad choice, like if I am able to successfully fool it and it got confused and for that particular instance; it picked an optimal arm it picked an arm which is not optimal for that action, but it in that case, we expect the regret to be linear right, if it is not able to pick the correct arm; that means, it is picking something else because of that regret, but we will see that.

To fool this algorithm, we have to set this delta very small and that is where what he was saying comes into picture; if delta is very small. Then it is not actually because the delta is again going to be a function of T itself, it is going to be not like growing like linearly in T, but it will be growing like square root T. Let us try to make this fine.

Student: Sir.

Ok.

Student: (Refer Time: 29:27) expectations and that is not happen.

Yes, it is the one with expectation.

Student: The expectation is will be on.

Expectation is on the randomness of the.

Student: Choice.

Rewards and also the choice of your arms.

Student: Sir, expectation on the rewards; then the it does not.

But my regret here actually depends on this instance itself rather it instance is what? It is nothing, but the mean values. The whatever the probability distributions finally, what effects my regret is that our corresponding mean values right. So, what it is; let me write this, what would we say this? This is nothing but X of; whatever the algorithms pulled over a time T. So, this is what your total reward; what you would have got.

Student: Right.

And this is what?

Student: (Refer Time: 30:30).

That is fine. So, in that case; let me write it as bit more new.

Student: Capital X; T, small x t also.

So, this is nothing, but t equals to; so expectation summation mu I t; t equals to 1 to T. Now expectation is there because this I t itself is random quantity; the arm you are going to pull, in round t that base is based on; the sample you have observed so far and that is the random samples; that is what I am going to worry.

See finally, what see like right now; I am only interested in the expected regret here right. So, in the expected regret as I have written here, what comes into picture is only the mean values. So, I am only interested in those parameters. So, for me this environment nu is nothing, but nu is or like; if I am going to define this epsilon to be some, let us say set of distributions that take value in bounded support ok; then I could just think of this nu to be.

Student: (Refer Time: 31:54).

$[0, 1]^k$, it is like you take any k vector where each component lies in 0, 1 that is going to define a environment for me. And this is what my (Refer Time: 32:11) is and actually this quantity here there exists is called minimax optimal, if there exists a policy pi. And actually this quantity here is this entire thing, this is for a particular nu and when I say it or a policy over this environment; this is the entire suit thing here ok.

So, now the question is can we construct up two problem instances which will satisfy these two conditions simultaneously? So, now let us say you are going to do this. So, let us take one problem instance where each P_i is Gaussian; mu i and variance, 1 and another problem instance. So, this is another mu with mu 1 right.