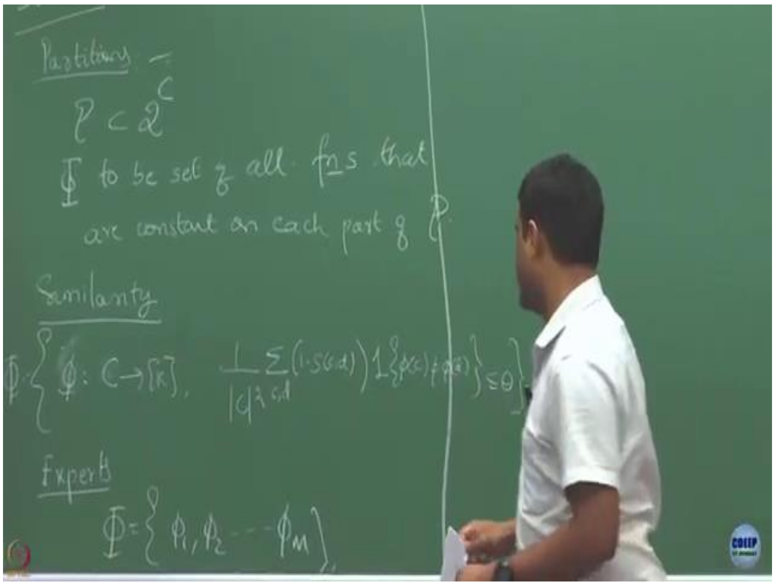
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 52
Exp4 Algorithm

So in the last class we said that our benchmark is instead of considering all possible functions that map by context to arm. I am going to restrict all possible such functions. My benchmark is going to be slightly weaker and the possibilities we considered is one is where my function assigns.

So, I will consider a partition and assume that my function going to select the same arm for all the context falling in that particular set. So, that is the one thing we had in terms of the partition.

(Refer Slide Time: 01:14)

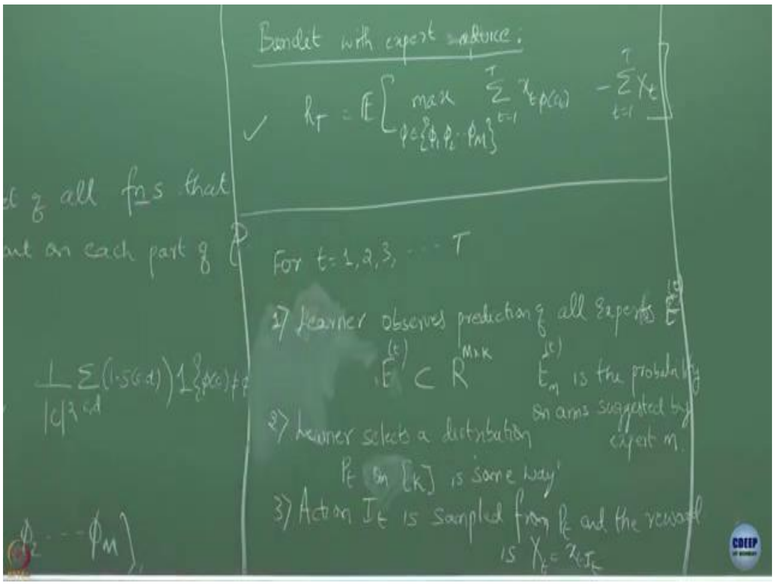

Partitions
$P \subset 2^C$
Φ to be set of all fns that are constant on each part of P
Similarity
$\Phi\{\phi : C \to [K], \frac{1}{|C|^2}\sum_{c,d}(1-S(c,d))1\{\phi(c)\neq\phi(d)\} \leq \theta\}$
Experts
$\Phi = \{\phi_1, \phi_2 \cdots \phi_M\}$

So, we said take a partition. So then we said that ϕ to be set of all functions that are constant on each part of P ok, this is one thing we said.

So, second thing was based on this similarity, for that we said that we will consider all ϕ such that. So, we said this is at the I want to consider all ϕ from C to k, such that this quantities less than let us say some θ and other possibly we said is we just consider some experts. And this experts are going to correspond to some finite number of functions ok. So now the benchmark we have relaxed by looking for such functions ok.

So, now we said that we are going to focus on this case. Now, for this case what is happening? We have these many functions and we want to see which is the function, I should be choosing want to find select the best function among this possible available functions right. And I am going to call this functions may be like I can treat them as some experts and the question is which is the expert I should be interested in.

(Refer Slide Time: 04:21)



Bandit with expert advice:

✓ $R_T = \mathbb{E}\left[\max_{\phi \in \{\phi_1, \phi_2, \dots \phi_M\}} \sum_{t=1}^{T} x_{t\phi(c_t)} - \sum_{t=1}^{T} X_t\right]$

For $t = 1, 2, 3, \dots T$

1) Learner observes prediction of all experts $E^{(t)}$

$E^{(t)} \subset R^{M \times K}$ $\quad E_m^{(t)}$ is the probability on arms suggested by expert m.

2) Learner selects a distribution P_t on $[K]$ in some way

3) Action I_t is sampled from P_t and the reward is $X_t = x_{tI_t}$

of all fns that
on each part of

$\frac{1}{|C|^{2}} \sum (1.5G(d)) 1\{\phi(c) \neq$

$\phi_2 \dots \phi_M\}$

So, now we are going to call this whatever we are going into call bandit with expert advice. So, as I said we are going to treat functions as like some experts and now my benchmark is defined in terms of this function. So, what my benchmark will now become? So, my benchmark will be so what is the regret I will be interested in that case? $R_T = E[\max_{\phi \in \{\phi_1, \phi_2, \dots, \phi_M\}} \sum_{\{t=1\}}^{T} x_{t\phi(c_t)} - \sum_{\{t=1\}}^{T} X_t]$.

So, over the time period t, I am just now going to see which is the best expert in hindsight and how I compare against that using whatever the policy or algorithm I am going to use ok. So, this is the total reward that you collected, this is what the best you could have gotten in hindsight. So now we have relax this benchmark here instead of considering all possible functions, we are only considering some finite functions. And now we are calling them as let us say experts and now we are trying to see how to solve this problem.

So, now I am now going to look into the setup, but with a slightly generalization of this, then we will come back to this like. So, what now I am going to assume is there are M experts ok. So, when you have this M functions right, what happens? You could when you observe a context in a round each of this functions would be as may be like pointing you to choose this arm this arm or the other.

Now, we can just assume that like now these functions are some M experts who would be playing an arm that is as prescribed by this function itself ok. So, if it is a function then it is a deterministic map right, if you see a context then you are going to you will be recommended to place a particular arm and that that expert will play that.

So, now let us just kind of go for a slight generalization of this, assume that this is an expert as such that instead of telling for this context which particular arm to play they will come up with the distribution on it. When you say context now you are going to give a distribution according to which you should be selecting a particular arm.

What I can say then in that case we can assume that the way we can setup this problem as, there are M experts in each round I am going to tell them the context I received and in turn each of this experts is going to tell me their corresponding distributions.

So, each experts will have what is the distribution with which you should be playing your arms, so they will reveal me. So, actually what the learner get is now he is going to he got one distribution from each of these experts, so that I can think of as a matrix. So, I have gotten a distribution from each of this experts.

Now, I have to decide according to which experts distribution I am going to pull an arm ok. So, what you are going to do is you yourself is going to come up with a distribution on the experts ok. And accordingly you are going to choose an expert and then whatever that expert suggested you the distribution. Then you are going to play the arm according to that distribution ok. So, earlier we in the EXP3 what you did there was only one expert like. So, you are the only expert there, so you just come up with the distribution in the arms and you pulled.

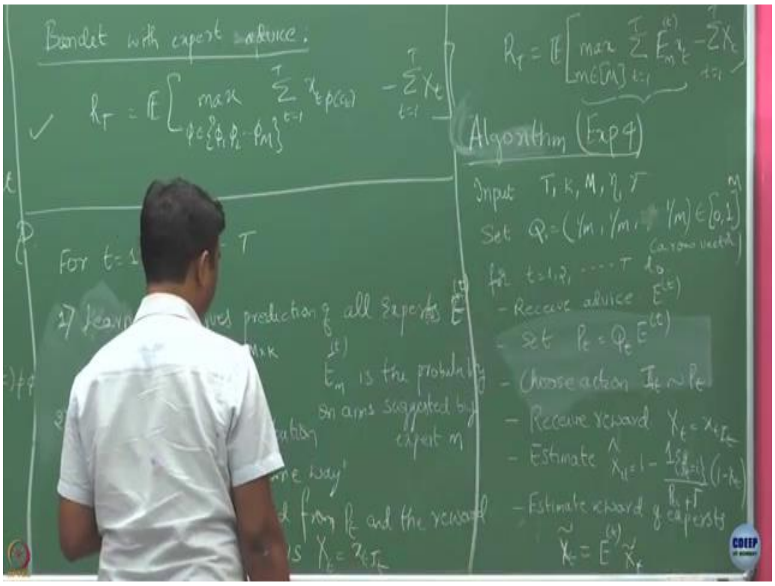
Now, there are multiple experts. So, you maintain a distribution on them and these experts are maintaining a distribution on the arms. So we can assume that in every round when the when a context comes right, so that context is visible to all the experts. So, they just see that and then they will tell what should be the arm, how should the arm should be picked according to their own distribution.

So, let us say that learner observes; so $E^{(t)}$. So, what is this $E^{(t)}$? this $E_t^{(t)} \subset R^{M \times K}$ is a; $E^{(t)}$ is here is a matrix, but what is this? So, we will assume that each row of this matrix correspond to a probability vector associated with that expert, so it is $M \times K$ matrix right. So, each row corresponding to one.

Then this is what he observes from all the experts, then the learner selects the distribution P_t on k is somewhere we will say this how. So, how this learner comes up with the distribution on k we will specify this, as of now just assume that the learner after getting this E of the will come up with a distribution according to which he want to play an arm.It is k we will may specify how it is on k, it is going to be through M he will come up with on k. And then he is going to action I_t is sampled from P_t and the reward is.

So, we recall that in every round t a reward vector is assigned to arms ok, which we denoted as that vector as x_t. So, x_{t1} corresponding to the reward assigned to r_1 in round t x_{t2} to corresponding to reward assigned to arm 2 in round t like that. So now, if you are going to play action I_t, the reward you gotten is x_{tI_t} this I_t is what you have played.

(Refer Slide Time: 14:28)



Bandit with expert advice:
$R_T = \mathbb{E}\left[\max_{\phi \in \{\phi_1, \phi_2 \cdots \phi_M\}} \sum_{t=1}^{T} x_{t\phi(t)} - \sum_{t=1}^{T} X_t\right]$
For t=1 ... T
1) Learner observes prediction of all experts $E^{(t)}$
$E_m^{(t)}$ is the probability on arms suggested by expert m
$X_t = x_{tI_t}$
$R_T = \mathbb{E}\left[\max_{m \in [M]} \sum_{t=1}^{T} E_m^{(t)} x_t - \sum_{t=1}^{T} X_t\right]$
Algorithm (Exp4)
Input T, k, M, η, γ
Set $Q_1 = (1/m, 1/m, \cdots 1/m) \in [0,1]^M$ (a row vector)
for t=1,2, ... T do
- Receive advice $E^{(t)}$
- Set $P_t = Q_t E^{(t)}$
- Choose action $I_t \sim P_t$
- Receive reward $X_t = x_{tI_t}$
- Estimate $\hat{X}_{ti} = 1 - \frac{1\{I_t = i\}}{P_{ti}}(1 - X_t)$
- Estimate reward of experts
$\tilde{X}_t = E^{(t)} \hat{X}_t$

Now, with this set up the regret you have observed is expected value of ok, let us understand this. So, this part is clear this is the total reward you have accumulated over a period of time. Now, what is this?

So, what is $E_M^{(t)}$?, what this gives you for. So, what is $E_M^{(t)}$ we said that $E_M^{(t)}$ is nothing but the probability distribution on the arm suggested by expert m and what is x_t? x_t is the reward vector ok. So, this x_t is a actually column vector for me and this is a row vector. What is this give you?

For expert M right, for expert M if you have gone with expert M this would this is what the expected reward you would have obtain in round t. Now, what I am looking it, what is the best reward I would have got in hindsight? If what is that expert I should have followed to get the best reward in hindsight? Ok. Now I am comparing with what I have gotten. So, with this set up we have we have now this is our problem regret minimization problem ok.

Now, let us focus on this and try to see what is the algorithm we should be using to get to solve this. So, already we discuss what is

(Refer Time: 17:01) right, exponentially weighted exploration and exportation algorithm with experts. I mean some version of that ok.

(Refer Slide Time: 20:45)

Bandit with expert advice.
Algorithm (Exp4)
Input T, K, M, η, γ
Set Q₁ = (1/M, 1/M, ..., 1/M) ∈ [0,1]^M
for t = 1, 2, ..., T do
- Receive advice E^(t)
- Set P_t = Q_t E^(t)
- Choose action I_t ~ P_t
- Receive reward X_t = x_{t I_t}
- Estimate
- Estimate reward of experts
- Update distribution Q_t
End for
observes prediction of all experts
is the probability on arms suggested by expert m
is sampled from P_t and the reward is X_t = x_{t I_t}

So, what we did with in EXP3? There also we only got to observe only for one arm, but we estimated reward for all arm using important sampling, we continue to do that here also. So, here what for all I, what is the range of i here?

Student: (Refer Time: 23:04).

1 to k for all arms you are going have this estimate. So, let us try to understand this algorithm now. What it is trying to do? As I said it is like it has two things to deal with experts and the distribution given by them. So, it is maintaining this algorithm is simply maintaining a distribution on the experts and once it selects an expert according to that distribution, it is simply going to follow the distribution given by that expert to pull an arm.

So, it is going to initially assume a uniform distribution on the experts, we do not know initially which one is good and then what is going to do? It is going to set $P_t = Q_t \times E^{(t)}$. What is $E^{(t)}$? It is a matrix right you have already described it, $E^{(t)}$ is a matrix where each row corresponding to the distribution given by the expert and now it is going to pull an arm according to P_t.

Now, the now if I do these is it same as first selecting an arm expert according to distribution Q_t? So, if I am going to play an arm I_t according to distribution P_t, is it same as first selecting an expert according to distribution Q_t and then selecting an arm given by that distribution so ok.

So, putting it in a different form suppose what I am going to do is I have distribution Q_t on the experts right. So, first I am going to select an expert according to distribution Q_t. So, when I select that particular expert what now I basically now what is a distribution vector a probability vector I got.

Now, I am going to select an arm according to that vector. Now, is it same as saying that I am going to pull an arm I t according to distribution given by P_t right. Then that because that P_t is nothing but the product of $Q_t \times E^{(t)}$.

So, now you have basically selected an arm and you are going to just receive the arm reward for that arm. Now, based on what all what is the thing you have observed, like you are just doing the important sampling here and then got estimates of all the arms ok.

And now what we are going to do? So, what is this X_{it}? This is now. So, $\widehat{X_{\iota t}}$ is the estimate you have for arm I in round t. Now, $\widehat{X_t}$ here is nothing but the vector of this guy. So, this is the ith component and now X_t is nothing but the vector of those estimates you have.

Now, what you are going to do is this is the estimate of the arms I am going to get. Now, then what is the average I would have obtain for an expert? So, now let us say so now let us focus on one. So, E of t is what it is a basically a matrix set let us focus on one row of it, corresponding to one particular expert let us call that expert M.

So, that take that expert M th probability vector and then multiply it with the column vector which are now the estimates of the rewards in that round. So, what that will give you? The expected estimated rewards you would have obtain from expert arm in that round right.

So, what this $\widetilde{X_t}$? It will give you it is a vector again it will give you each component give you will give you what is the reward expected reward that each of the expert would have obtained in that round.

Now, you take those values and then try to update the weights for each of this expert, in a way very similar to what you have did earlier we just give them the weights in this exponentially weighted form ok. So, what is this $\widetilde{X_{t_\iota}}$ is going to give you. What is the estimated reward that i' th expert would have obtained so far.

And now there are tuning parameters here and of course you need to know how many experts are there, how many arms are there and we have also this parameter γ here. So, if you have γ here like this, suppose if it is γ is greater than 0. What was so in the in the standard bandit setting? In the adversarial setting what this algorithm was when we have use the estimator like this? The experts are already fixed. So, they are using some policy ok, whenever you are going to give a context right in the for that context they are going to tell what is that they have would use. So, this experts are fixed ok, we are updating our weights on them, which context you have observed in that round t, because the expert is going to; for each context the expert may have different distributions ok. But those distributions are fixed for example if they have 10 context, for first context you may have one distribution second you may have another like that, but they are not going to change with time.

It may be like after two different point of time you observe the same context, for those two context the expert would written you the same distribution ok. And he is not going to adopt it based on what context you have seen.

In this our goal is just to identify the see experts they have already figured out, but for which context how should be arm should be selected. Now, our job is who has figured out it well. So, that is what like from the experts we want to identify the best one.