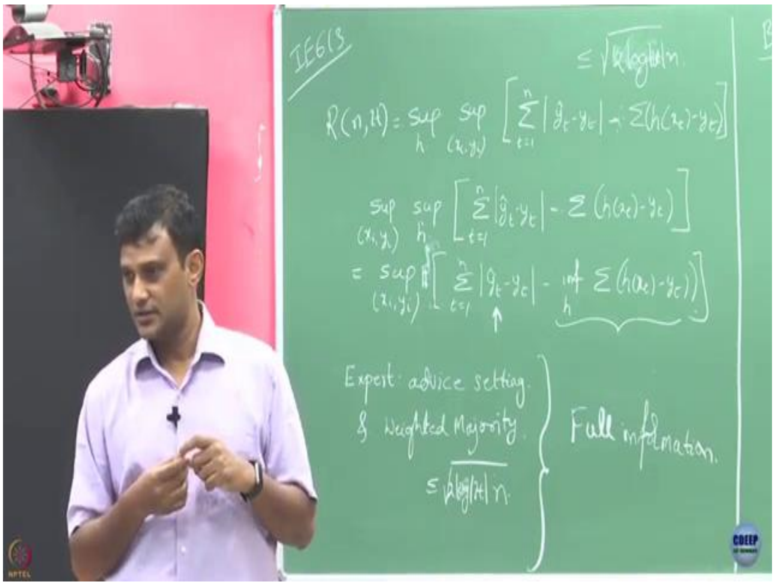
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 12
Full Information vs Bandit Setting

So, so, far we were studying what we call as learning with full information setup that is because in the case that we were interested in that is the online binary classification problem, we said that when an instance comes a sample comes we get to see the predictions of all the hypotheses that are available right; that means, you got to know what everybody would have said that was a full information setting.

(Refer Slide Time: 00:56).


IE613
$\leq \sqrt{\log|\mathcal{H}| n}$
$R(n,\mathcal{H}) = \sup_{h} \sup_{(x_t,y_t)} \left[\sum_{t=1}^{n} |\hat{y}_t - y_t| - \sum (h(x_t) - y_t)\right]$
$\sup_{(x_t,y_t)} \sup_{h} \left[\sum_{t=1}^{n} |\hat{y}_t - y_t| - \sum (h(x_t) - y_t)\right]$
$= \sup_{(x_t,y_t)} \left[\sum_{t=1}^{n} |\hat{y}_t - y_t| - \inf_{h} \sum (h(x_t) - y_t)\right]$
Expert advice setting
& Weighted Majority
$\leq \sqrt{2\log|\mathcal{H}| n}$
Full information

So, let me revisit the regret definitions we had in this case. So, how did we write the regret? Rn, we had another parameter here right?

Student: Hypothesis Class.

Hypothesis class. So, how did we write this was as supremum over H and sup over. So, this was over n rounds and then we wrote it as summation, what was the first term?

Student: $\hat{y}_t - y_t$

Then so, this is my; this is we defined as regret ok. So, just let revisit this definition a bit the way we defined what we said is whatever the sequence I am going to look, I am interested in the best. If I am going to apply single best hypothesis what is the how far I am from that hypothesis right. And, the kind of when we bound it is using our algorithms. It the bound hold independent of what was the hypothesis and what is the sequence, right.

So, now let me ask you. So, instead of defining regret like this, suppose if I had defined all my earlier bound still continue to hold, why? Yeah, but when I give the bound you took whatever H you took whatever sequence, I was able to bound this quantity right. So, earlier the bound cycle like if you recall they were like of the form $2^{|H|}$ something like this right which was independent.

So, even if you think it like this, so, if you are going to say that so, what is the difference between this metric and this metric here. So, here what you are doing if I right like this, you the hypothesis which is going to maximize this is selected based on the on the sequence right. It could change, but here what I am asking is here I am asking one hypothesis that maximizes irrespective of what is the overall possible the sequences.

But here this hypothesis could depend on what is the sequence you are looking at. That is the different, but still if I am if what are the bounds I gave there that bound is still valid for this case also right because that is independent of what is H and what is your sequence. So, if you rewrite this quantity a little bit right, I have just pulled this sup inside and it has become ∞ .

So, I mean we have already discussed this. This is saying that given a sequence, what is the best you could have given got an in handsight right. So, this is the loss you incur. Now, you are looking at what is among my hypothesis which is the best one gives and now you are comparing that against what you got; this is what you got. So, even though earlier we bounded for this quantity, but basically our all the bonds were valid for this ok.

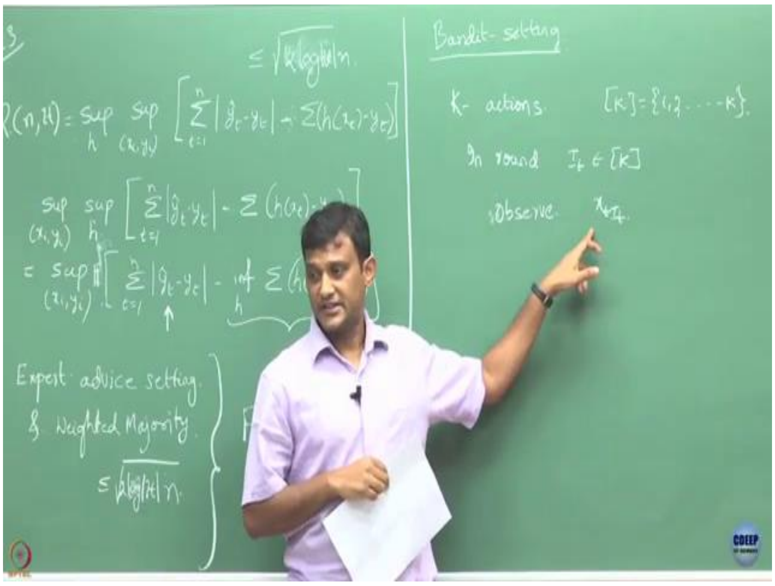
And another change we say this guy yt here did not be deterministic right. yt is what? yt is the prediction given by the earner in round t. We said that if he is gives a deterministic prediction then he can be made to add in every round. Then we allowed this $\hat{y}_t$ to be random ok. So, $\hat{y}_t$ so, then we said instead of this let us look at a expected regret right. When you allowed this $\hat{y}_t$ to be random then we considered the expected regret ok, fine.

So, this was the setup we have looked into the case when we dealt with what I what I called as full information setting right. In this setup we caught on this bound using weighted majority algorithm and in the weighted majority algorithm we said that we get to see the prediction made by all the hypotheses in every round.

Now, so, we had basically said that this using expert advice setting we and weighted majority we said that this one can achieve a regret bound of $\sqrt{\{2\ |H|n\}}$ for this setup . Now, slightly consider a case where my information structure is slightly limited ok. What I say my information? See when we are doing this learning right what is that information we are getting by taking an action is very important.

If you get more information maybe we can do we can learn faster right. If we less if we get less information maybe we will get you will be slower in learning. This setting we call it as full information we have. Now, I want to move to something called bandit setting.

(Refer Slide Time: 08:30)


$\leq \sqrt{k \log |H| n}$
$R(n,H) = \sup_h \sup_{(x_t,y_t)} \left[\sum_{t=1}^n |\hat{y}_t - y_t| - \sum (h(x_t) - y_t)\right]$
$\sup_{(x_t,y_t)} \sup_h \left[\sum_{t=1}^n |\hat{y}_t - y_t| - \sum (h(x_t) - y_t)\right]$
$= \sup_{(x_t,y_t)} \left[\sum_{t=1}^n |\hat{y}_t - y_t| - \inf_h \sum (h$
Expert advice setting
& Weighted Majority
$\leq \sqrt{2 \log |H| n}$
Bandit setting
K- actions $[K] = \{1, 2, \ldots, K\}$
In round $I_t \in [K]$
Observe $x_{t I_t}$

So, what I mean by bandit setting is when you play an action here you only get to see the loss you incur from playing that action and nobody else ok. Earlier in the full information case whatever action you played you incurred loss that action, but you also get to observe loss of everybody else, but here that is not there. Whatever you play you incur loss for that and you only observe the loss of that.

So, now, we are going to focus on this setting. So, can somebody tell I mentioned about this brief in the last class, can somebody tell example of full information setting and the bandit setting where this can arise? Previous yeah for whatever we have done, so, far for the actions we have played we can observe, but about the current setting I am asking when I say full information and bandit information right. This is about information what information you get in each round.

So, exactly, so, when you are thinking about a casino and playing a game as a player, you go and play one machine and you get to see whether you lose or win or whatever amount of money you make with that machine.

In that round you do not know what other machines would have given you the ones which have not played. You only get to see the one which you played that is exactly what we are going to call that bandit setting. Now, what is the example of full information setting? So, other example as I said also in the last class. If you have let us say, 10 row roads to take from your let us say home to office and each of them use different different travel time.

On a day if you take a particular route you only get to see what is the travel time on that route not about the other routes. But, maybe this could be also full information setting. Why? Because, let us say you are the learner, you have turned on your radio in your car when you are driving right and the usually right radio keep announcing right that much was the congestion traffic on that route was this much on this route was this much.

Even though you are experiencing the actual traffic on the route you are travelling, but through the radio announcement you are getting to know what is the traffic that is being happening or being experienced on the other roads even though you are not there right. If that if you do not have that radio who is announcing this you only know on that route in which you are travelling.

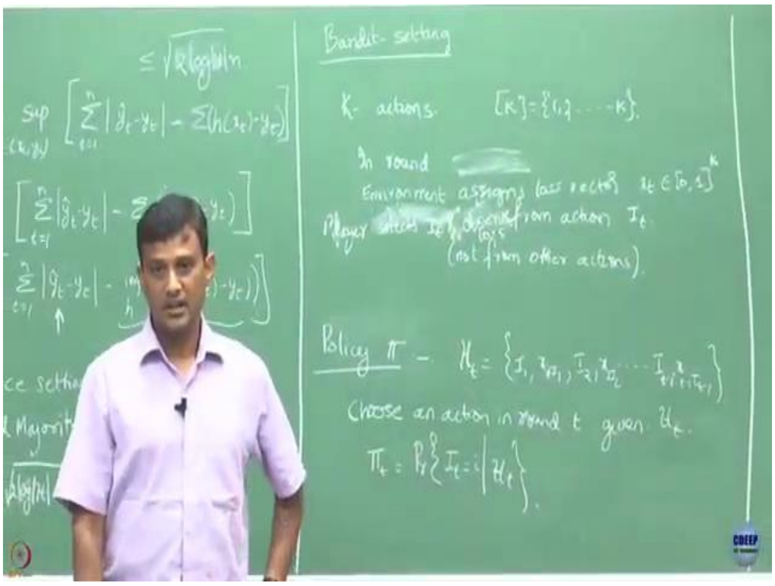
If somebody is giving you this information about other routes also then you have full information ok. And also the other example could be the share markets right; like on a particular day let us say, you have option to buy some 10 stocks or shares, you buy one share, on that share you get to see whether you made money or not. You may not get to see the stocks of the values of the other stocks or share, but maybe if you go and look into the newspaper they might have announced what is the loss that have been incurred by the other shares or stocks.

So, in so, far the setup is (Refer Time: 12:27) said to be somehow assume that I will get to observe the losses incurred by all the actions I have ok. Now, henceforth I will restrict that, I will get to see losses of only the actions I observe and not about others, because getting more information is always costly right here. Like in the road example you need to have a radio for that or in the share market like you may have to go to newspapers or some web portals and get all these details about the other ones ok.

Now, how we are going to how this set up is going to be different from the earlier setup? So, when it comes to regret definition that was our performance criteria will there be any difference in the definition of regret or do you think we should consider something else to evaluate performance with this setup ok? Let me make this bit most formal. Henceforth, I will only talk about actions. I will not talk about like hypothesis everything. Let us say there are K actions which I denote as ok.

In each round the learner is going to choose one of the action and when he chooses that action he is going to only observe the loss for that action. Now, I am going to denote x_{tIt} to be the loss incurred by playing action It, in round t. You understand this notation?

(Refer Slide Time: 14:56)



Bandit setting

K- actions. $[K]=\{1,2,\dots K\}$.

In round

Environment assigns loss vector $x_t \in [0,1]^K$

Player observes loss from action I_t

(not from other actions).

Policy π - $\mathcal{H}_t = \{I_1, x_{1I_1}, I_2, x_{2I_2} \dots I_{t-1}, x_{t-1,I_{t-1}}\}$

Choose an action in round t given $\mathcal{H}_t$.

$\pi_t = P_r\{I_t = i \mid \mathcal{H}_t\}$.

And observe may be like loss from action It, may be not from other actions ok. Now, let us specify the so, this is what the learner doing. So, here we will what is the environment doing? Environment will assign loss vector $x_t \in [0,1]^k$. So, in round t, environment is going to assign losses to this actions and player selects It and observe loss from that action whichever he has played ok. And this It here this It the learner is going to select like earlier it could be random.

He need not select this It according to a deterministic policy, he could randomize it ok. So, if this is the interaction between the learner and the environment, now we are going to define the regret of a policy π. Now, there is no hypothesis here right. It is just like the environment is assigning this loss vectors. And, let us say π is some policy that the learner is using. Let us precise let us make what is policy π. Policy of layer is to select an action in each round ok.

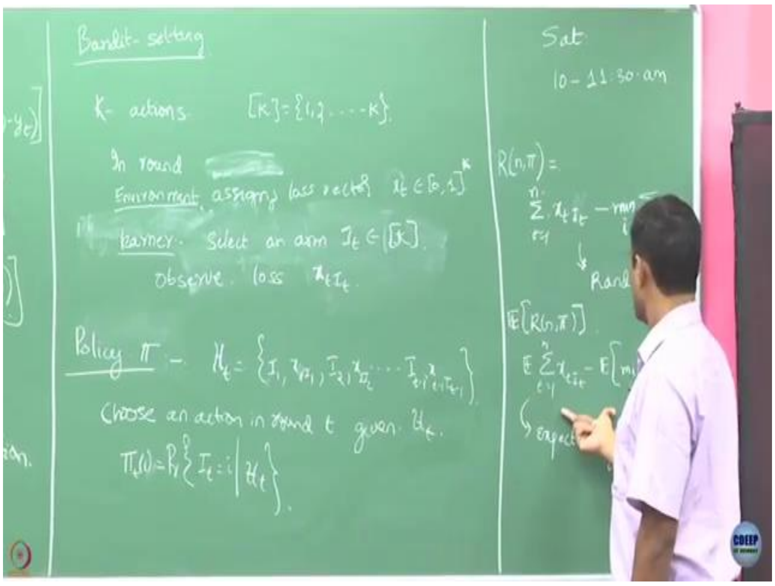
And how he can select an action in each round? He will select his action based on his past observations right. So, let us say Ht is the history he has observed till round t ok. What is that? It could be let us say in round 1 he observed I1 he played action I1 and based on that he observed this and then in the second round he observed this, all the way up to I_{t-1} with $x_{t-1I_{t-1}}$.

This is the history that the learner has right till round t. You understand this? I1 is the action played by player in first round and this is x, x1 is the vector in round 1, but x1s I1 component is the loss he actually observed and similarly till round t minus 1. Now, learner's policy is to choose an action given Ht ok. So, we are going to denote the policy to be in round t to be $\pi_t = \Pr\{I_t = i|H_t\}$.

So, the learner has access to this history based on that whatever he is going to choose an It in that round with what probability is going to choose that is his policy ok. So, π is this probability distributions that he comes up with in each round ok. Now, you notice that we are allowing It to be random and he is going to choose it to be small i with some probability ok.

Yeah, it could be arbitrary selected by the environment, like the case we did in the expert advice setting also right.

(Refer Slide Time: 20:15)



Bandit- setting

K- actions $[K] = \{1, 2, \dots K\}$.

In round

Environment assigns loss vector $x_t \in [0,1]^K$

Learner. Select an arm $I_t \in [K]$.

observe. loss x_{tI_t}.

Policy π - $\mathcal{H}_t = \{I_1, x_{1I_1}, I_2, x_{2I_2} \dots I_{t-1}, x_{t-1 I_{t-1}}\}$

Choose an action in round t given $\mathcal{H}_t$.

$\pi_t(i) = P_r\{I_t = i \mid \mathcal{H}_t\}$.

Sat

10 - 11:30 am

$R(n, \pi) =$

$\sum_{t=1}^{n} x_{tI_t} - \min$

Rand

$\mathbb{E}[R(n, \pi)]$

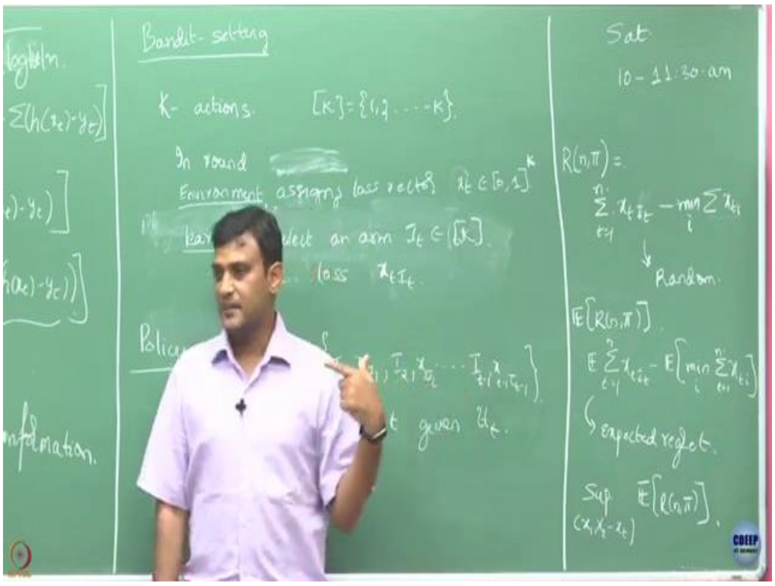
$\mathbb{E} \sum_{t=1}^{n} x_{tI_t} - \mathbb{E}[m$

expect

So, environment is just going to assign a loss vector to all the arms you have and that how he assigns? It could be totally arbitrary. And whether he will assign this loss as such that the components across this arms are correlated or uncorrelated they are not putting any constant, it could be arbitrarily selected.

And then this is my learner is going to select an arm It and he is going to observe loss x_{tIt} ok. So, in this case whatever he observes this is what he has actually incurred. Now, we are going to define the regret of the learner with this policy. So, depending on how he chooses this It that this probability that is going to define his policy right.

(Refer Slide Time: 21:29)


Bandit setting
K- actions [K] = {1,2, . . . K}
In round
Environment assigns loss vector $x_t \in [0,1]^K$
Learner select an arm $I_t \in [K]$
loss x_{tI_t}
Policy
given $\mathcal{H}_t$
Sat
10 - 11:30 am
$R(n,\pi) = \sum_{t=1}^{n} x_{tI_t} - \min_i \sum x_{ti}$
Random
$\mathbb{E}[R(n,\pi)]$
$\mathbb{E} \sum_{t=1}^{n} x_{tI_t} - \mathbb{E}[\min_i \sum_{t=1}^{n} x_{ti}]$
expected regret
$\sup_{(x_1, x_2 \cdots x_t)} \mathbb{E}[R(n,\pi)]$
information

Now, let us define this. We are going to define regret of policy to be the one which he incurred by playing the sequence of It's right. So, what is this? x_{tI_t}; t equals to 1 to n rounds. This is what he incurred minus what is the best he could have incurred? If he had seen this sequence for all the rounds, what is that? That is going to be $\sum_{\{t=1\}}^{n} x_{tI_t} - min_i \sum x_{t_i}$. So, if you knew all this loss vectors for all the actions you would have chosen the action which would have given you thus minimum sum right minimum loss.

So, that we are going to compare that with this is going to call as the regret. Is this regret or random quantity? Why? Because It is random right. So, this is a. So, what we may be interested in? Instead of this instead of bounding this random quantity, we may be interested in looking at the average performance of the learner. And what is that; that, we can take for that we can take to be the expected value of this.

This is why we wanted we would like to select an i which would have minimized this right. That is why this is what I get. This is what I would have like to do. I am comparing my with this and this is what I am going to call it as regret. We have discussed this right already for the full information case.

Now, I can look into the expected value of this quantity than this expectation is now governed by this probability vectors right. So, now what is this expected value is actually governed by my policy π here ok. And this is my expected regret ok, fine. If you notice that this regret here, whatever I have defined is actually same as this is just that this has been defined for the binary classification problem. But here we have instead of looking instead of thinking about xt yt's the value that is coming in round t and if I apply hypothesis class as this is the loss I am going to incur.

Instead of this that (Refer Time: 25:35) is just going to say that that value the loss is simply x t i ok. x t I in this language you can think of the loss I incur if I am going to apply ith hypothesis on the context xt, whatever the loss I am going to incur I could just write it that as xt i here fine. Now, we are going to see how to bound this, ok. What we would like to do is he want to bound this by taking supremum of this it is over all possible sequence like the way we have done it right.

So, we can make this to be this now quantity to be we can we want to bound with sup of whatever quantity and is exactly this (Refer Time: 26:42), ok. So, we have just written the whole thing that the regret now. When we wrote this things and these things we really did not worry about what is the information structure right. It is having full information and here we are saying this we wrote it for bandit information, but regret definition is still the same.

Where that picture comes into whether what kind of information I have is in your algorithm when you are going to update because here this policy is only observing the losses for the action it played not for everything else. It is not for everybody else ok. Now, let us try to see what is the algorithm we get.