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

Lecture - 22
Strongly Convex Function

So, we have been discussing this online convex optimization problem, but then we saw that if we just do follow the regularize leader in some cases we may end up with bad regret, we showed with an example when my function was linear. Then we showed that if you somehow regularize it things can be better, and but we have to choose our regularizes appropriately ok.

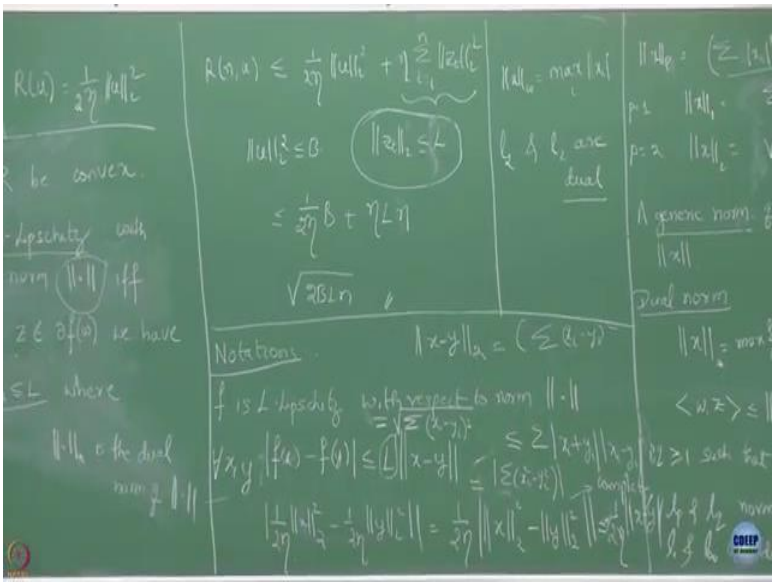
(Refer Slide Time: 00:56)

So, for the case when we have my loss functions defined to be like this, and when we use our regularizer to be worth 1 by eta times norm of think there was a 2 here right. We showed that we get a simple update mechanism which was basically a gradient descent that you update your weights taking into account the gradients in each step, and you will account your gradients in such a way that your weights will decrease according to the gradients, not simply the gradients of eta times the gradient.

So, based on that, we had this simple online gradient descent algorithm which was like start with parameter eta positive, then initialize with $w_1 = 0$ vector, then your update rule was w of less than u is w_t minus eta time Z_t. And what is this Z_t here, the Z_t is the you

say that this is the sub radiant computed at w_t ok. And if the function f was differentiable, then we say this is simply the derivative at that point, in which case the derivative here is simply this Z_t here right fine.

(Refer Slide Time: 03:26)



And for this algorithm what did the bound we showed finally, we had for this setup, we had shown that the regret with respect to and was upper bounded by. So, we had something like what was that 2 eta norm of u square plus summation 2 equals to 1 to t, then norm of Z_2 square. Is that correct?

Student: (Refer Time: 03:56).

There is an eta here ok. So, then we made an assumption here, we said that let say norm of this is upper bounded by beta, and also said that let say this upgradients are upper bounded by this L, then this upper bound turned out to be two eta b plus eta times L n times. After this what we did? We optimize to this over eta, because that eta was an input parameter. And by its appropriate choice of eta, what is the final bound we got – the best upper bound?√(2BLn) ok.

Now, from this expression as you see, the regret is dependent on the size of this gradients here right. Somehow, if whatever you are choosing if this gradients have to be very large then your regret will also going to be large, so that is why we assumed that the gradients could be anything, but we ensure that they are bounded, adversary cannot choose any

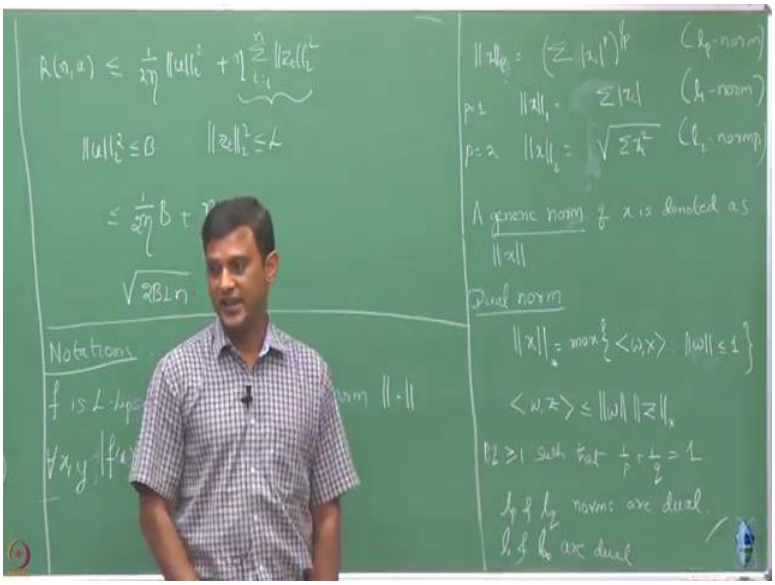
arbitrary gradient, it has to be related like this. And with that we got this bound. So, it looks like to get a good bound, one has to control these gradient levels.

Now we will see that this gradients can also be thought of course, this is a property of a your function f, but we can now going to connect it to the Lipschitz property of a of a function. So, now, we are going to make a bit little bit detour and do some notations. So, we are going to call f is l Lipschitz if you take, so I am going to assume this functions f for all real functions real valued functions.

So, if you take the difference between this function f at the points x and y, this should be upper bounded by L times this ok. And now you see that I have taken this to be real valued function, but this real valued function takes argument to be vectors x and y could be any vector. And if shift, if this is like for all x y right. And this L is you know constant, this does not depend on what is the value of x and y here.

And now you see that I have written this x minus y, and I am interested in some norm of this right. What is this, how is this norm defined here? Right now I am going to say f is L-Lipschitz with respect to norm. Now, I am going to define this whatever this operator now, right now just think it as an operator. This operator is this. If you tell me this norm and you are going to share your function is L-Lipschitz with respect to that norm then this is what I mean, this is the definition ok.

(Refer Slide Time: 08:18)



Now, what is the norms we are interested in? In general, if you have a x of P norm is defined as summation $|x_i|$ to the power P this whole to the power 1 by P ok. So, what is if P is equals to 1, we are going to call it as l 1 norm. How does l 1 norm look like? So, l 1 norm is like where it is simply going to be summation of norm absolute value of individual components ok.

Now you should take suppose P equals to 2, this is called l_2 norm. And how does this look like? This is the one and this is square of this, sorry just this right. And this is the one we are which we are most familiar with right, this is called l_2 norm. And in general any norm with this parameter P is defined to be this. And I mean we are going to denote a generic norm. So, I am when I say this, I have here I am just intend to say this is a norm of vector x, but I have not clearly specified whether this is a l-1 norm, l-2 norm or P-th norm ok.

So, this is just to denote this is a norm with respect to some P, and depending on what is that we are interested in we are going to say this is going to be l-1, norm or l-2 norm that is basically in this we have not specified the value of P depending on what is value of our P we will say P is 1 or 2. So, there is a notion of dual norm. So, this is a generic norm.

So, if you give me a norm whatever, x is dual norm I am going to define it as, and I am going to substitute it as like this. So, take any generic norm that has been given to you. Now, I am going to define its dual norm. How I am going to define? Its value is going to be max of this inner product, this is clear, what we mean by this over all w such that this happens.

Suppose, I am interested in finding the dual norm of l-1, then this guy is going to be simply l-1 here ok. And now if I am going to be interested in dual norm of l-2, in that case we have to take this to be with l-2 ok. So, if this is the generic definition of dual norm, then you have given a general norm whatever it is.

So, you can verify I am just leaving you if I have W, Z, two things this can be always upper bounded as norm of W and norm of its dual norm. Suppose, if you take inner product of two quantities W and Z, the first whatever the generic norm you have let say this is l-1 norm, then the second term is the dual of that l-1 norm. So, here that dual corresponds to what are the associated generic norm.

So, if this guy is l-2 norm, here when I say this is associated dual norm of l-2 ok. So, here you see that I can interchange W and Z write. So, here I could write it as Z and here I could write it as W in which case I am taken the generic norm with an W here and its dual norm on Z here ok. So, in most of our analysis we will be only interested in mostly P-1 and P-2 ok.

But it so happens that if you have p, q greater than equals to 1 such that 1 by p plus 1 by q is equals to 1. If this happens, then so I am just going to call this as like l-1 norm, and this is like l-2 norm here, and this is like l-p norm. Then l-p and l-q norms are dual ok. So, if you are going to take l-q norm, its dual norm will be given by q; if 1 by p plus 1 by q satisfies this ok.

So, suppose now let us do this. Suppose, I want to take p to be 1, so that is l-1 and then what is the q that is going to that satisfy this equation, if I take p equals to 1, what is the value of q that satisfy this equation? Infinity right. So, so, dual of l-1 norm is what, l infinity norm ok.

Now, let us define what is l infinity norm. So, by definition l infinity norm is this is $|x_i|$ to the power infinity and then take 1 by infinity power when the way to interpret that is simply max of i if the maximum component among them ok. I in this think of intuitively like if you are raising this to power infinity and then taking this to be 1 by that power.

So, if letting if you let p to go to infinity, what is this power you converging to 0 right. And then in this case the one with the largest value dominate, so that is why we can and you can argue that this definition just answer to be this max of this component ok, then by letting p going tending to infinity ok.

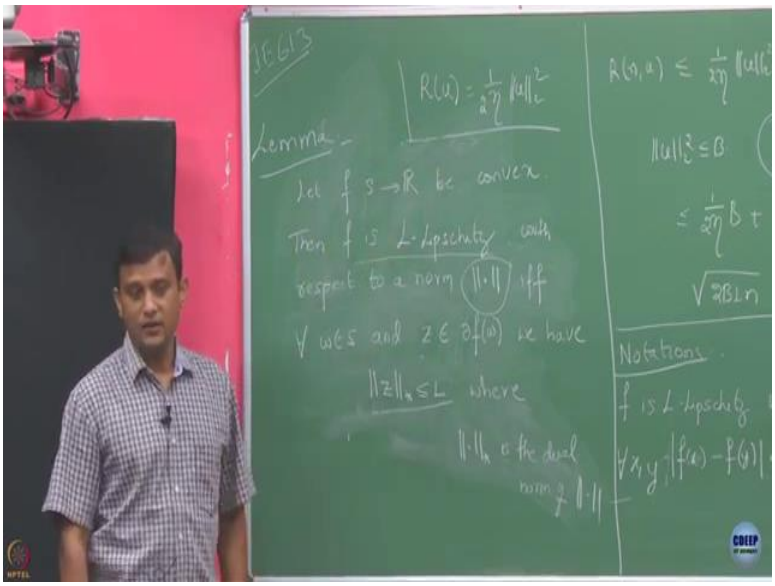
The other thing is so now let us take l-2. If I take p equals to 2 what is the value of q that satisfy this equation? 2. So, if my generic norm is l-2, what is this dual norm is again l-2. So, l-2 norm is dual of itself ok. So, that is why when I am, so l-2 norm is the most simpler to handle. I really do not need to if I just say l-2 norm, I really need not worry whether I am working in the original norm space or I am in the dual space ok, ok. With this that is why l-2.

So, here that is why it is important that whenever I am going to define my Lipschitzness with respect to a norm that matters, because it could be l-1, l-2, l-3 whatever and be

depending on that this l could change. This may satisfy this for some l-1, if it is l-1 norm; and if I am going to take this to be a l -2 norm, this l could be different right fine ok.

So, this with this notation now let us move back to what we are interested in. We here we were interested in the fact that this Z_t 's here which are sub gradients, we wanted them to be bounded. Now, is this sub gradients somehow related to the Lipschitzness of this function, and if that is the case with what parameter? So, here I to make all this things work to get this sub linear bound, I needed this condition right that all the sub gradients are upper bounded by that. Is it that equivalent to saying my function is L-Lipschitz, is that true? So, we will see that yes that is true.

(Refer Slide Time: 19:39)



So, here is the result I am just going to state it, we will skip the proof. So, let my f is a rear valued function from S to R be convex, then they are going to say that then f is L-Lipschitz. So, this basically states that if I have a convex function and real function, and this is Lipschitz's with respect to some norm ok. And this is, if and only if for every point in S and all the associated gradients – sub gradients this satisfies, that is in this grade sub gradients in the dual space, there also upper bonded by L. So, if my, so in a way what we are saying is if my function is L-Lipschitz, then my sub gradients are also upper bounded by L ok.

So, now, let us take this norm to be l-2 norm ok, then this is also we know that this is also l-2 norm. If my function here let say is L-Lipschitz with respect to l-2 norm, then all

the sub gradients should be also uniformly bounded by that same constant L. So, in a way what all the stuff we did here everything here works out instead of saying that my gradients here sub gradients are the dual norms of sub gradients are upper bounded by L. Instead of this if I say my function f by f_1, f_2 that I am going to see they are all L-Lipschitz then everything goes through here right ok.

So, henceforth that is why now instead of worrying about my gradients are the sub gradients are bounded in in the dual space, I just worry about whether my function is Lipschitz in the generic norm ok. So, hence for that is why I am only interest now henceforth I will focus on whether my function is Lipschitz or I will assume that my functions are Lipschitz with some constant that automatically implies that my sub gradients are also Lipschitz.

So, just notice here one small thing like when I derive out all these things, I assumed that the squared norm are upper bounded by L. But here when I converted to this, this is just dual norm so there is no square here. So, basically then we henceforth we have to replace our L by square root L that is the translation we have to do. Or maybe I what I should have done is when I wrote all this things instead of that, maybe I should have considered this value and put all the things, but anyway let us not reverse. So, we will just keep in mind that there is a square root l translation we have to do when we are to compare our bounds with whatever I achieved here.

So far what we did we all these things nicely worked out when we have a specific regularizer right. What was that regularizer? The regularizer we defined was R of u was 1 by eta.

Student: 2 eta.

2 eta ok. So, this is the specific regularizer we took, and this regularizer was defined. So, is this regularizer here was it a convex function? With respect to, so like when I when you is of course, it is a convex function, was it a Lipschitz? So, just see if I take f equals to R here. Is it a Lipschitz? And if it is Lipschitz you have to tell me for what L and with what norm ok. So, just plug in if you are just going to plug R here, R here, and just simplify what is that you are going to get ok.

Let us do that. So, this is I am going to get it as 1 upon 2 eta norm of 2 of x here minus 1 upon 2 eta ok. So, anyway this is already out right like this is 1 upon 2 eta is a constant, this is right. So, can you simplify this any further? Ok, what is the definition of this norm of x whole square? So, norm of summation of x_i square, and this is summation of y_i square. This is written as less than.

Student: (Refer Time: 27:42).

Plus and minus.

Student: (Refer Time: 27:57).

Why is this true?

Student: Higher.

So, this term here we actually get. So, if I just take it directly, so this term is nothing but whatever the components right.

Student: (Refer Time: 28:31).

Because of this ok, you have split this. But what I want is here this. So, what is does this definition here, if I am going to take it to be in l-2 norm, what I want this, this is nothing but summation x_i y_i square under root of this whole square. But what I have here is only x square minus y square ok. Is x square minus y square is upper bonded by $x_i - y_i$ whole square?

So, check out that this is indeed you can simplify this and we are you can get it as whatever this 1 by 2 eta norm of x minus y this whole square. Just simplify, expand this, try to get this $x_i - y_i$ square stuff in terms of x minus y whole square summation. Then you will end up something which you can write in terms of the norm of difference between x minus y ok.

So, what we have actually is this function here is indeed Lipschitz, and whatever the with whatever the constant we have in front of it that is 1 by 2 eta here ok. This one, so this is oh you just apply the definition whatever like what is this what we want? We want norm of x minus y right. And you want it to be l-2. What is this is going to be?

Student: (Refer Time: 30:41).

I am saying you just have to simplify it.

Student: (Refer Time: 30:47).

Right. So, you have to manipulate that and apply some inequality to get that right. This is not a finance, so I am skipping thus to work it out for you ok. So, you complete this. So, we have one regularizing function here which is convex. We can look it other regularizes as well right, which we want anyway this regularizing function was convex that is a good, because I was adding this regularization function to the sum of convex functions right. And I was minimizing it. So, if I am adding this regularizing function to the sum of convex function if this regularizer is also a convex function, then the whole function is convex for me. And I it is easier for me to minimize this, fine that property I desire.

Now, what we are saying is in addition to this, we would also like this all of this f_t functions to be Lipschitz right then we are going to get this. But R the regularize this function we have here also thought it as convex function which I got it in the zeroth round right, the way we treated R is like f_0. So, we also want it to be also called Lipschitz with some constant right.

So, now, let us look at if you are going to consider other regularizing functions maybe let say both convex and Lipschitz's, what kind of bound one can expect and what are the other possibilities we have? Ok, so we would be interested in something like more what is going to call as some functions, which are strongly convex, not just convex, but something more than that; and for that we will be able to derive some better bounds.