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

Lecture – 24
FoReL with Strongly Convex Regulariser (Cond.)

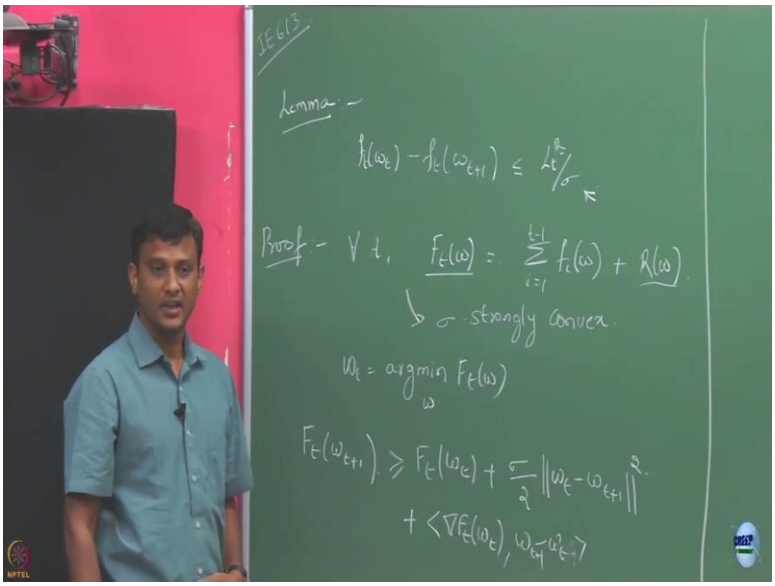
Yeah. So, let us just recap what we did in the last class. So, we introduced this notion of strongly convex function in the last class and then we derived some of its properties right. And, then we started talking about how to use different strongly convex functions. So, two candidates we discussed last time was the Euclidean regulariser and the entropic regulariser. Euclidean regulariser was strongly convex in which norm? It was in L-t norm and the entropy one was strongly convex in?

Student: (Refer Time: 01:05) L 1 norm.

L-1 norm. So, we will see how to use these two regularisers and we have also discussed that that the gradient being bounded, the norm of gradient being bounded is in a way in a sense equivalent to the Lipschitzness property of the convex functions, right. Now, continuing from our last class at the end we had shown that if I am going to use a strongly convex regulariser how my difference in the functions look like.

So, we are always interested in the regret bounds will be after form $f_t(w_t) - f_t(w_{t+1})$ right something out that is all and we are interested in bounding that.

(Refer Slide Time: 02:14)



So, let us rewrite that and try to complete the proof we discussed last time. So, we wanted to argue that $f_t(w_t) - f_t(w_{t+1})$ is upper bounded by L_t square by sigma. So, what is L_t here?

Student: (Refer Time: 02:45).

Yeah this is the Lipschitz constant for function f_t and what was sigma? It is the.

Student: (Refer Time: 02:57).

Regulariser, parameter it is sigma strongly convex that is sigma parameter. So, we say we are saying that now if I am going to use my follow the regularized leader, the regularizing being my sigma strongly convex function. And, each of my functions f_t is Lipschitz constant is Lipschitz with constant L_t; then if this w_t 's are can computed based on my follow the regularize reader then we are saying that this bound holds good ok.

So now, let us see why this is so ok. So, I made a formal statement of this in the last class, but I am just; I am just stating what is that we wanted to show in that claim. So, now its why this is true? Ok. So, for all t let us define $F_t(\omega)$ to be summation of. So, this is the function that follow the regularized leader trying to minimize in each round right. I am just calling that function to be F_t.

All these functions f_i 's are they are assumed to be a convex and also Lipschitz and this R(w) is assumed to be what strongly convex; strongly convex with parameter sigma. So, anyway R(w) is anyway convex and this is all the convex. So, this F_t is our convex function, but further we are saying that this R is sigma strongly convex.

Will this make this whole function also sigma strongly convex? So, you can verify this that is indeed true like if I add two convex functions it is going to remain convex function. But, if I add a strongly convex function to another convex function the sum becomes a strongly convex function and it continues to. So, this entire thing is going to be sigma strongly convex ok. Now, let us with this definition of F_t let us workout what happens.

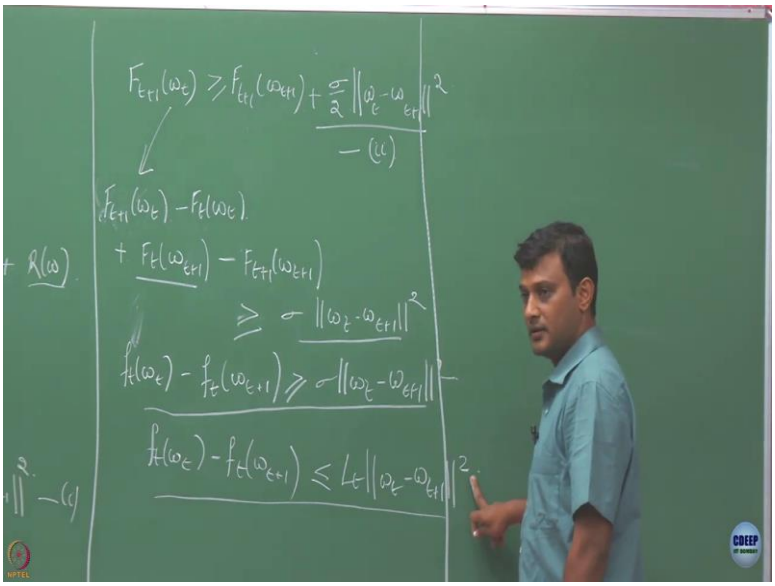
So, by our rule a follow the regularized leader what is F_t is going to be complete data device like argmin of F_t (ω), right. This is how my follow the regularized leader is going to work. Now, since this guy F_t is a sigma strongly convex function, I can write it as F_t (w_{t+1}) is upper bounded by F_t (w_t) sigma by 2 norm of w+1 whole square. So, now why this result is true? So, why this result is true? So, what I am doing here basically?

Student: (Refer Time: 07:24).

Yeah. So, F (w_t) here is what? w_t is the minimizer of F_t (w) that is by definition right and now I am applying I know that since F_t is a this convex function I have F sigma strongly convex function. It need to have this lower bound, but we had also something some linear term here right. What was that? That was the gradient of F_t (w_t) and something like w_{t+1} or maybe it was of t to plus 1 w_t ok. But, now that w_t is the minimizer of F_t (w_t) it was getting nullified right.

So, we have the given this as one of the properties of my sigma strongly convex function. So, once this is hold; so, here I am fixing w_t and now this is another w_{t+1} ok, this is a minimizer.

(Refer Slide Time: 08:48)



So, if this is the case I can also write it as, now I will do the same thing at F_{t+1}, now w_{t+1} is the minimizer of F_{t+1} ok. And, then this has to this will be plus sigma by 2 norm of w_t minus w_{t+1} norm whole square fine.

I have this two equations now. So, let us add these two together. So, if you add them; so, while adding I am also trying to do some simplifications w_{t+1} and this is greater than or equals to sigma times norm of w_t plus w_t plus whole square. This correct let if I.

So, right hand side by adding this and this they are the same I get this term and the other thing have simply how simply. So, F_{t+1} this corresponds to this and F_t (w_{t+1}) corresponds to this and this and this term I have just simply taken on the left hand side. So now, go back to this definition of F_t function here F_t function, F_{t+1} involves sum till t whereas, F_t involves sum till t minus 1 right and both of them computed at the same point w_t. What is this difference is going to be?

Student: (Refer Time: 10:54).

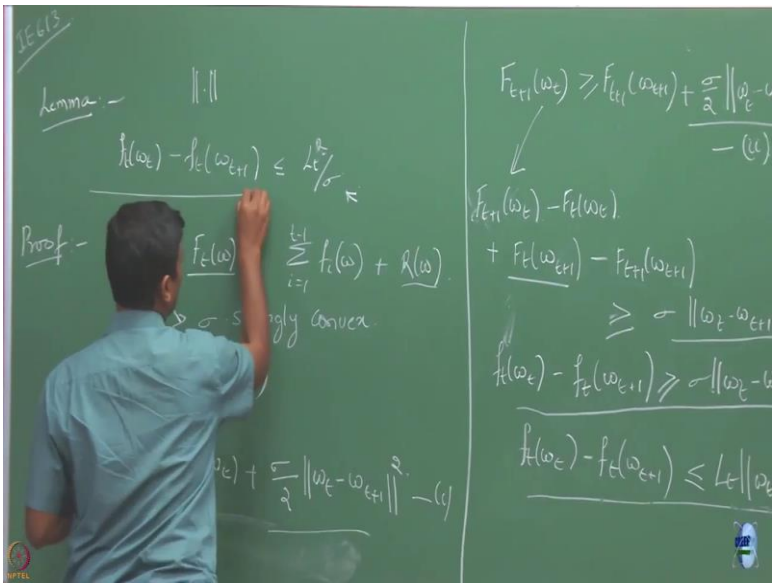
It is going to be f_t (w_t) ok. What about this difference? This is now computed at the point w_{t+1} for F_{t+1} and F_t and this is going to be what? Minus f_t.

Student: f (Refer Time: 11:27).

W_{t+1} and this is still sigma norm of w_t minus w_{t+1} whole square right. So, this so far I have only used the property of this function F and the sigma strong convexity of this. Now, I also want to bring in the Lipschitzness properties of my f_t functions ok. Now, by the Lipschitz property of my f_t function, how can I bound this?

This is upper bounded as L_t times norm of w_t minus w_{t+1} whole square, is this correct? This is just I am applying my definition of Lipschitzness of a f_t function.

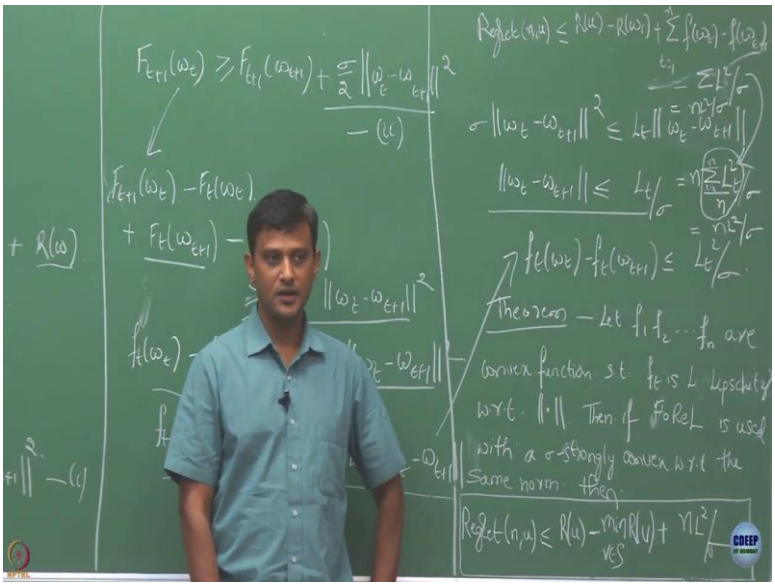
(Refer Slide Time: 12:35)



So, note that all I am doing first I am right now I have not specified any norm here, it could be L_1, L_2 whatever ok. So, if I am saying that my R(w) is strongly convex with respect to nor[m] some norm, this L_t that is a Lipschitzness is also with respect to the same norm. Now, with this I am going to know I have a lower bound on this through this and I have an upper sorry I have an upper bound through this and I have an lower bound on this.

So, if I am going to compare these two things. So, does this Lipschitzness says there is a square here? Just check I think there is no square here for Lipschitzness, it is just a norm ok. So, if I do this now what I am going to get from this? If I now this is a lower bound, this is an upper bound.

(Refer Slide Time: 13:48)



So, I now going to compare these two things through this I am going to get $w_t - w_{t+1}$ is upper bounded by my L_t times norm of w_{t+1} right. So, using this lower bound and upper bound I have this relation. But, now I just plugging this relation back here and that is what you wanted to show that is $f_t(\omega_t)$ minus $f_t(\omega_{t+1})$ that difference I have, I am just use it. So, this will using this I have whatever I want yeah. So, if modulus is there, but if remove modulus also the bound should hold right.

What is the Lipschitzness says? Absolute value of this should be upper bonded by this, yeah I found the left side if you remove the absolute value, this can be only smaller right. So, the upper bound still holds fine. So, with this lemma what we are finally, able to show is let me write it as a theorem now ok; is this clear why this is true? What we did? So, earlier we have already a result which says that regret is upper bounded by R(u) minus $R(w_1)$ plus summation $f(w_t)$ minus $f(w_{t+1})$.

We had this results right earlier, this is how we bounded the regret; now we are saying that. So, then this is simply follow the this is the bound for the follow the regularized leader with my regularizing function R. And what is w_1 here? What was w_1? w_1 is basically the what whatever your algorithm found in round 1 right, that was obtained by minimizing this R function. So, that is why this $R(w_1)$ I am simply writing as minimization over $R(\omega)$.

In the first round you do not have, this summation is empty; we are only minimizing the regularizing function ok. So, that is why instead of $R(w_1)$ I am writing. Now, for this part we have just demonstrated it to be this ok. So, every term here; so, this is for each term right ok. So, I have to make one more thing here. So, here what we are basically saying is this is like L_t square by sigma. So, what is now L here?

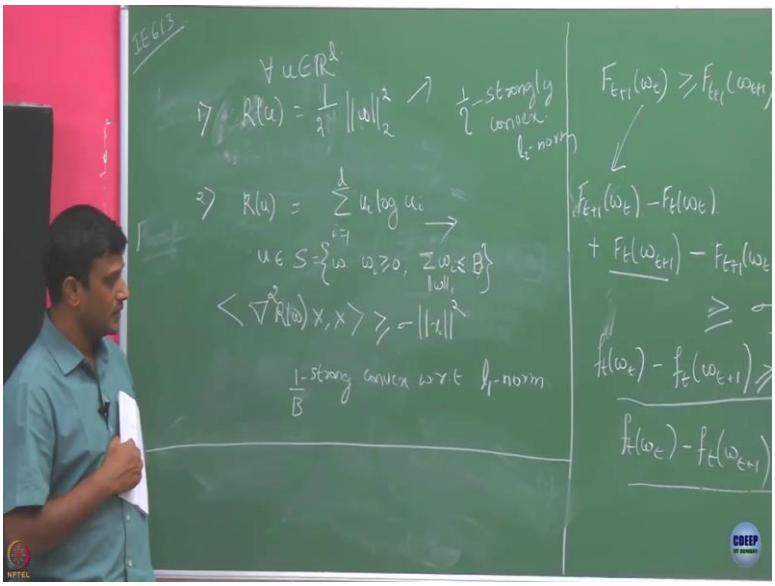
So, this is let us say they are all f_t is they are all L-Lipschitz, they have a same Lipschitz constant. In this case what this will turn out to be? This is going to turn out to be simply n L square by sigma and that is what we have written here. But what we really would; so, suppose this L_t 's are not the same and they are going to be different in each round. Let us say this is L_t square by sigma, I could I have just multiply n and n here.

So, here either I can assume that my Lipschitz constant L is same in for all the functions, in which case L_t will be replaced by L and this is simply n L square by sigma. Other possibility I can think is this is t equals to 1 to n, instead of assuming that all the functions are the same Lipschitz constant what I can assume is the average value of the Lipschitz constants. This is the average value of the Lipschitz constant right, square average value of the squared value of Lipschitz constant; I can say that that will be some L square.

And, in this case again I will get the same n L square sigma bound ok. So, what we are saying is as long as all your functions are convex with the same Lipschitz constant L then this bound holds. If you are going to use some regulariser, some sigma strongly convex regulariser with sigma here then this holds. Or, let them be all f_t in this f_t functions be convex and each f_t is L_t Lipschitz constant and the same bound holds provided I interpret this L square as the average value of all the L_t squares fine, fine.

So, now we have this nice simplified version on the regret bound provided my I use follow the regularized leader with a strongly convex function, sigma in this case sigma strongly convex function. So, now, let us go back and work out; if I use different different regularizers which we discussed like our Euclidean distance and entropy distance what is the bound we are going to get here you ok.

(Refer Slide Time: 20:52)



Let us go back, 1) regulariser is I am going to take to be R(u) is 1 by 2 eta is and 2) R(u) is we said summation $u_i \log u_i$; let say you are in some d dimensional space right. So, this guy is, is it a strongly convex function? We discussed this last term right, this is a strongly convex function with what sigma value?

Student: (Refer Time: 21:57).

1. No, I have written sigma also here, eta also here without eta it was ah.

Student: (Refer Time: 22:13).

If I have this is one strongly convex right. So, again what was the definite what was the property that we use to check strong convexity? The Hessian based condition we had right, it was del square R(ω) X into X if this is going to be greater than point of square. So, with respect to this norm done this guy this function R is a sigma strongly convex function. Now, can you check me, now I am already norm has been defined for you to be L-2 norm.

So, what is this? So, if you take this function and try to compute this condition for what sigma it holds? 1 by eta. So, let us say this is then 1 by eta strongly convex. And with respect to what norm? It is true with L-2 norm and what about this? So, for this we made some more assumptions right this u_i 's are; so, each of these u_i we also defined. So, this

is true for any u, this Euclidean distance or this Euclidean regulariser was defined for all u.

But, then we define when we defined this entropy regulariser we said that this is coming from a set S, where all my w, all my components are positive and also. So, this is what we call probability, since simplex which you did not like and we call it as probability space right; we all this u as that is coming from this space. Now, this function is it again; so, strongly convex.

And, now see when we said this thing right this was nothing, but L-1 norm. So, I do not we computed this last term or not you can verify that this one, if I am going to take this quantity to be 1 this is going to be 1 strongly convex with respect to L-1 norm yeah ok. This is the function right with respect to the L-1 norm and one can also show that if instead of this you should take it to be some B number; this can be shown to be like 1 by B strongly convex.

Or, like not exactly you can take it to be less than or equals to B (Refer Time: 26:12) ok; sorry like earlier it was equals to 1, it is anything less than or equals to 1 or less than or equals to B ok. So now, what is going to be this bound look like if I am going to apply this regularisers? Ok. So, can you now compute, if I take my Euclidean regulariser can you come to work out what is this going to be?