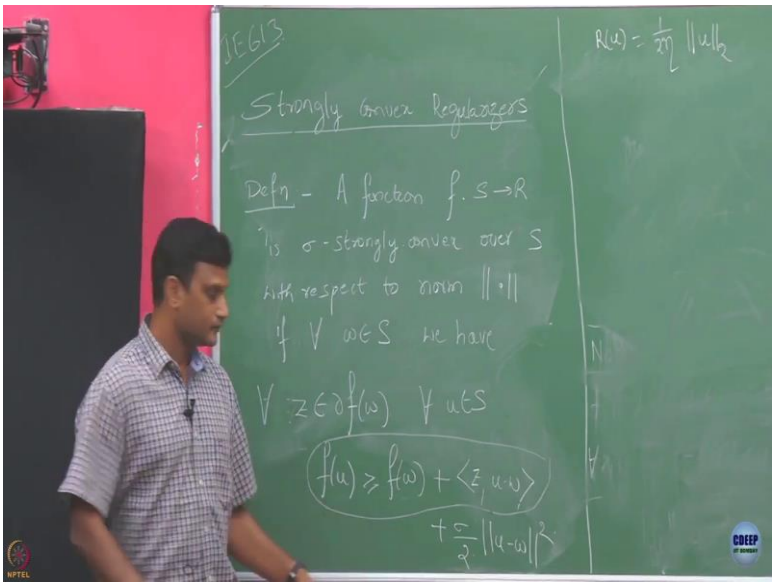
**Bandit Algorithm (Online Machine Learning)**
**Prof. Manjesh Hanawal**
**Industrial Engineering and Operations Research**
**Indian Institute of Technology, Bombay**

**Lecture - 23**
**FoReL with Strongly Convex Regulariser**

Now, we are going to see that by what we mean by Convex function. We defined we stated it in the last class that a convex function is the one which at any given point, I can get a lower bound on it right. Basically, that lower bound came from a tangent which pass through any given point and that tangent was defined in terms of the sub gradients at that point w that you are interested in.

So, that function, the tangent function that was like a lower bound. it was to my function all the time. But now, if we allow our self bit more flexibility that the gap between that convex function and my linear function, if it is like slightly good big enough, it grows well; then, we will have a better control over regret ok. So, let us first define them, what is that strongly convex functions.

(Refer Slide Time: 01:26)

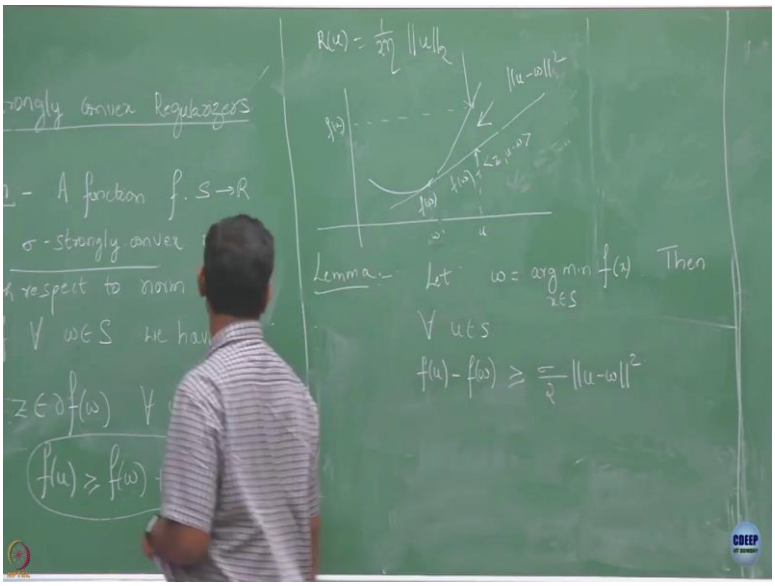



So, we are going to say, so this is the definition of my convex function; strongly convex function with parameter sigma. So, you see that this is the part which we have already defined. We know that a convex function this part already holds right. If you have a convex function f, in fact we have said that function f is convex if and only if this part

holds, where z is the sub gradient of your function f computed at point w and this part here, this is basically a tangent which is passing through your point w ok.

But now what we are saying is yes, this was my initially my lower bound of the function; but in addition to this, if I add this extra factor to this, that is I have increase my lower bound by this much amount; then, also this entire thing continues to hold as a lower bound on my function f. And notice that what is this, this part here? So, this is your this is at point w, we are computing the tangent. You should look into if you increase u, that is if you are moving away from omega, this guy is only increasing. So, basically if you look this difference between f (u) – f(w), this guy is like also growing.

(Refer Slide Time: 03:28)




So, in a way what we are doing is, so suppose like let us say this is my convex function and if I had taken some w. So, this will like f of w z minus v and also, what I want now is this term here sigma by 2 norm of u minus w. You see that this gap between this function and this at any point, let us take; so, let us say this is w, this is w and you take any other point at some point let us say this is to be u. If you now look into the difference of your function, this is let us say f(u) and this is like f (w).

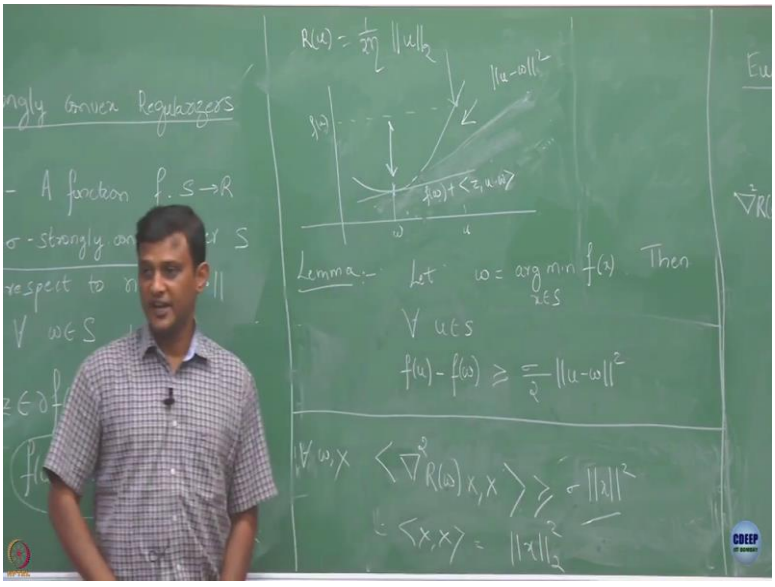
So, now, look into the difference between f(u) – f(w); earlier that difference was only this much, when it was simply a convex function. Now, it is also growing by this much. So, if you are going to take u away from omega, so you expect this gap to be also be enlarging right. So, in a way in that in this case the separation between this and this line, so maybe

I just un want this if you just take this line. So, this gap is also kind of growing very fast and growing means at least by this amount ok.

And this has to be; so, this is f and this is any point w. So, this kind of this portion is like norm of u minus w square ok. This is I am not saying like whatever the norm you have given; I want this to be take that and square it ok. So, this is whatever the norm. This is defined with respect to some norm and I want the square of that norm and this is what it is right. So, what we basically by bringing this sigma strong convexity, we are basically wanting this operation to be growing much faster. In this way, it will become easier for us to kind of possibly to identify this functions ok.

So, now let us have some more properties of this strongly convex function ok. First thing is what we call? So, this is another Lemma, I am going to call it as. Now, let instead of taking some of every point w, let us take a specific point w which is like a minimizer of this convex function. Then, for all u equals to S, we have this u of S minus f(w) is going to be sigma by 2 times norm of u. So, what we are now saying that instead of looking at this omega, where is let us say this function is getting minimized here ok.

(Refer Slide Time: 07:54)




Take this point and now, look at the tangent which is basically passing through this. Let us say like this ok. Now, we are saying that if you are going to look at the difference of your function like let us say this, basically this difference that is f(u) – f(w), this is growing like this. This will be at least this difference is going to be at least this amount

right. So, if so this property is one thing because at any when we are going to deal with our kind of minimize at every point, when we are finding $w_t$ in my algorithm right, we have been looking at the minimization of some convex function ok.

So, there this maybe result will be useful like, when I am going to look at this point, how was the value of the other points going to differ with respect to this value ok, fine. So, here we really do not need to assume that this result holds for a function f that is differentiable. This function need not be differentiable at any point ok; but let us try to understand this why this statement holds for the case, where my function is differentiable ok, at all the point. So, what we know about this? Suppose, if w is my minimum value of my function f, what will be the gradient at this point? It is going to be 0 right.

So, then in that case if I take this w to be the minimum value and in that case, this is a just like gradient at that point right and that will be 0. So, in that case this z is a zero vector. So, if you just plug in 0, what will be left it only?

Student: (Refer Time: 10:24).

This term which is like a lower bound in this case ok, but as I said it need not be just true for the case, where my function is differentiable. When it is not differentiable, (Refer Time: 10:38) only my sub gradients are there, this is fine. What is the; how to check that my function is then strongly convex? Ok, fine. What is how do you, how do you check my function is convex? That is the definition ok. You can directly see whether the definition holds, then you can check your. Are there any test for convexity?

Student: (Refer Time: 11:14).

Yeah, second derivative if it is a real function, I mean if its arguments is real, then you can differentiate with respect to that argument and see whether its second derivative is positive or not. But here, we are talking about convex function which was taking vectors as inputs right here, what is your test?

Student: (Refer Time: 11:35).

Yeah, Hessian addition right and what is that; what should be the property of that hessian?

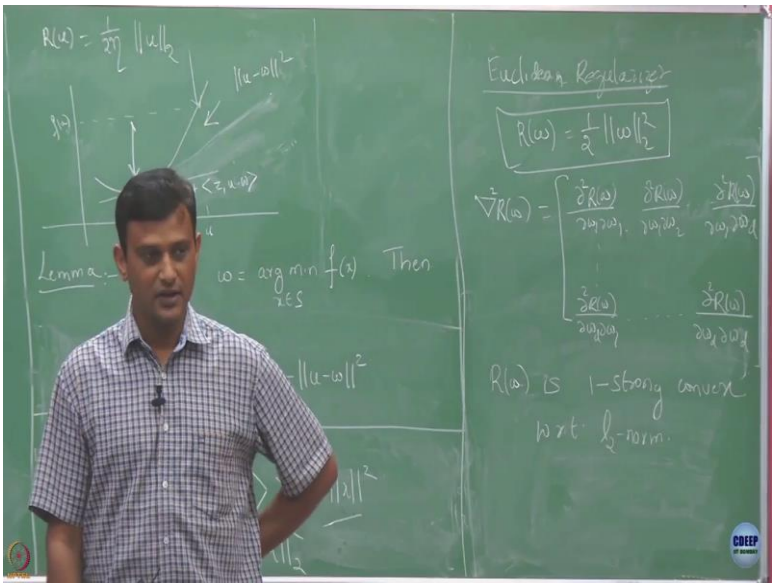
Student: (Refer Time: 11:49).

Yeah right. So, whatever like, what you want it should we want it to be the positive semi definite positive right? So, if it is strict fine, otherwise we are let us say we all we need is my hessian should be positive semi definite or positive definite, whatever you want.

Now, is there analog test for this? What is the test? So, the test for that is take the hessian and so take the hessian at omega and then, if it is going to be greater than or equals to sigma times this. Then, you can verify that your function is sigma strongly convex with respect to this norm. So, here I have defined this at this this is with respect to some norm right. If it satisfies this property with respect to some norm, then we are going to be and this is like a test.

In that case, we know that this definition holds. Because of that it is sigma strongly convex with respect to that norm ok. Now, let us see I think earlier, I made a mistake with respect to that Euclidean regularization ok. Now, let us see whether this property holds for Euclidean regularizer.

(Refer Slide Time: 14:08)




So, what is our Euclidean Regularizer? We said that R of omega is 1 by 2 and I am just keeping the eta term here, it is like Euclidean Regularizer. So, now, let us try to work out, what will be its second hessian looks like? What is its second hessian? So, what this matrix look like?

So, this is delta 2 R omega delta omega 1 delta omega 2 like this right. So, then this is delta 2 R(ω). So, maybe I will fix $w_1$. So, this is sorry, this is $w_1$ and this is $w_2$, all the way up to delta $\omega_1$ up to and delta let us say there are d components in this; this is like this and the last one is going to be delta of R(ω) times delta omega d delta $\omega_1$, all the way up to delta square.

So, can you tell me if compute like if I take this R(ω) to be like this, what is these values are going to be?

Student: (Refer Time: 16:01).

So, 2 is getting knocked off here right.

Student: Yeah.

It is just going to be identity function ok. Now, identity times a matrix, it is going to be X itself right and now, X into X and now, I am interested in I have already specified this function R with respect to Euclidean that is l-2 norm right. So, I would now see that with what whether this guy is again going to satisfy this property with some Euclidean, sorry l-2 norm? So, in this case, what I am going to get? This I am going to get X comma X and what is X comma X in a product?

Student: (Refer Time: 16:55).

It is?

Student: l 2 norm.

It is going to be right. Now, my Euclidean Regularizer here is strongly convex, but that is not a complete description right. It is strongly convex; what is what sigma?

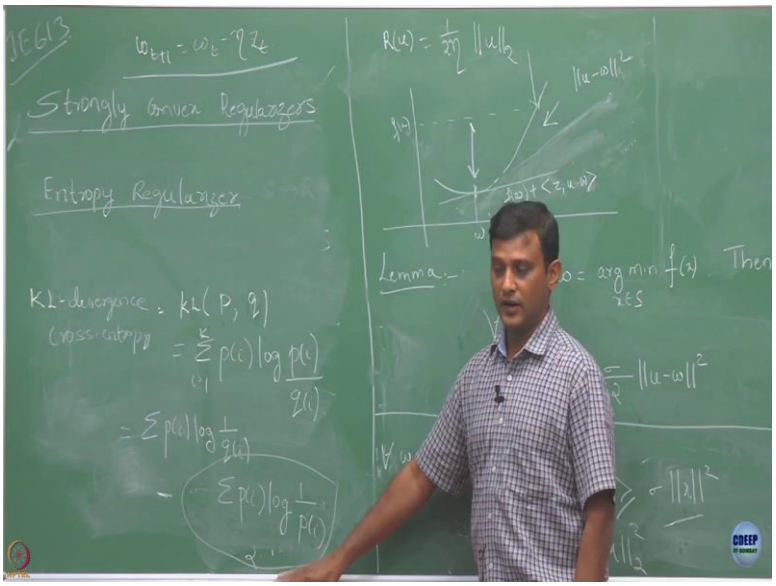
Student: (Refer Time: 17:23).

So, it is what we have shown is R(ω) is 1 strongly convex with respect to l-2 norm. So, if we have shown it to be strongly convex, 1 strongly convex maybe we missed up when you try to see that this guy was Lipschitz's or it should be still fine.

So, anyway, please do verify whether still like we have left it. We said that this is Lipschitz's, but we did not complete that right. So, just go and verify this, whether this is

a Lipschitz's? If yes, with what value of l? Ok fine. So, at least now, we know that this guy is any way convex; but it is more than that it is strictly convex with respect to the l-2 norm ok. This is one regularizer, we have. I am going to discuss with you today one more regularizer called Entropy regularizer.

(Refer Slide Time: 18:43)




So, how many of you heard about entropy function? You heard it, where?

Student: (Refer Time: 19:04).

Which machine learning? Is there any difference between entropy and cross entropy? We are talking about cross entropy, not entropy right. What is self-entropy?

Student: (Refer Time: 19:15).

So, we will revisit this a bit later; but let me write it cross entropy or and you are also using our scale divergence and so, you are just throwing many random terms or you know about them.

Student: Yeah and also (Refer Time: 19:32) from (Refer Time: 19:34).

We will discuss this later, but just like to be KL, it is called a KL divergence or like you also want to call it cross entropy ok. Before I write it, the way we are updating this w's right; w's are what they are coming from my set u. Because I am trying to minimize my function over u right. Suppose, let us say that is a convex function, when I consider this

linear function, when my losses are linear, what was my $w_t$ happened to be every time? It was just like a average of all the gradients right. If we are if I assume that this gradients are coming from a convex set; so, their average is also?

Student: Convex.

It is also some convex value right. But where they are w's; where they are probability vector? The w's the weights, I am getting in every round. $w_t$ how did I obtain? $w_t$ was like ok, so after simplification all the things $w_{t+1}$ was $w_t$ minus eta $z_t$ right. So, $z_t$ was not in my control that was chosen by the environment or the adversary. So, here $w_{t+1}$ is can you guarantee it to a probability vector?

Student: Yes.

It is need not be right. Why; why it has to be? It can be any value ok. But to if I want to cover the cases, where my w's has to be kind of probability vector right, then maybe like if I am going to use such an update rule, this $w_t$ will not be probability vector; continue to probability vector. But if you are now recall go back and recall your prediction with expert advice. What was $w_t$ 's there in every round? They are probability vectors right. You actually derive you made them to be probability vector in every rounds.

So, if you have this kind of regularizer which yielded in such kind of updates, you are not going to get probability vectors in each round and because of that, you will not be able to generalize or recover the prediction with expert setting with this convex functions. So, you remember how we are getting all these? We just pitched this convex of online optimization as a generalization of all the predictions with expert advice right.

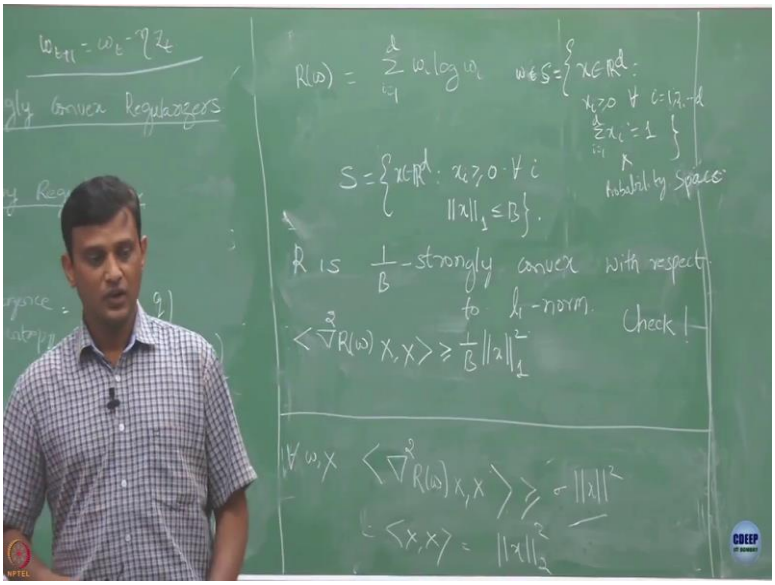
So, this online convex optimization that way should also cater to the case, where prediction with expert advice algorithm setting is recovered; but in so far we are not able to do that because this $w_t$ 's are not there. But suppose we want to restrict your wt's to be always to be probability vector, then what kind of regularizer's will be interested in applying? Then, comes all this scale divergence cross entropy because these are all makes sense, when you are dealing with probability vectors right ok.

Now, let us let us say this is between two vector, two probabilities; I am going to call this as p and q. So, p is one probability vector and q is another probability vector; both

defined on the same sample space ok. Then, I am going to define this to be $p_i$ log $p_i$ divided by $q_i$. Let us say my sample space has some n elements in it or maybe not n, maybe we will some k elements in this, it is fine.

And then, you should simplify this; I can write it as. So, can I simplify this in terms of this. So, notice that this part here only depends on the pi distribution; whereas, this part depends on both $p_i$ and $q_i$ here ok. So, this part here that $p_i$ log $p_i$, this part without this minus 1, we are going to call this function as entropy and we will discuss this later. This is kind of how much information is contained, when you generate a symbol according to this distribution ok. So, that is fine. Now, we are interested in this kind of regularizers and that is what I mean by entropy function; entropy regularizers.

(Refer Slide Time: 26:09)




So, I am going to define my R(ω) should be $w_i$ log $w_i$. Let us say it is over some d number of logs. (Refer Time: 26:32). So right now, I have not said this w has to be a probability vector, but let us say this is any vector w; then, I am going to define my entropy functions to be like this ok.

So, now let us say now we want to see is this what kind of function is this? Is this R(w)? Is it a is this strongly concave, sorry strongly convex? If yes, with what sigma and with what norm ok? So, can we see that? But now, let say I am taking this ω's to be coming from all my x in $R^d$ such that my $x_i$ 's are positive for all i 1, 2…d and my summation of

$x_i$ is equals to 1. What now I have made is what is this? Let me call this omega coming from belonging to S which is defined like this. So, what is this S? It is a?

Semicircle, is this a semicircle? I am what I am saying all the I take a if I take a vector, all its components are positive and they sum to 1. It is a simplex right; so, probability simplex right. So, more generally, what I can write instead of defining like this?

Student: (Refer Time: 29:11).

Probability simplex.

Student: Space.

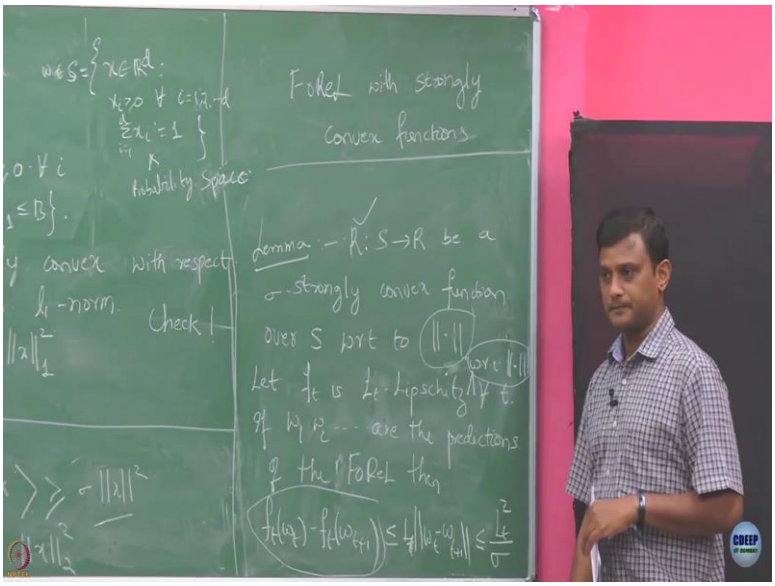
Yeah, probability space or you just call it probability ok, if you do not want let us call space ok. So, I want to just generalize this; instead of like this, I am going to define S to be all x; all $x_i$ components are positive and the l-1 norm of this is bounded by B. So, what I have done? This is like special case of this when I said B equals to 1 and what is l 1 norm of this? l-1 norm of x is what?

Summation of $x_i$ 's, but it is now simply summation of $x_i$, it is summation of mod of $x_i$'s. But we already ensured that $x_i$'s are positive. So, I do not need to put my mod there. Now, if I am going to define my $R(\omega)$ like this and where, my omegas are coming from this space, now what about this? Is it any strongly convex function or some parameters ok. So, it is so happens that R is 1/B - strongly convex.

So, I have specified what is sigma, then what is the norm. It is going to be strongly convex with respect to l-1 norm ok. I will leave this you to check this. So, how you are going to do this? Again, you just go and compute the hessian and see whether this guy is going to be greater than or equals to 1/B times l-1 norm squared fine. So, we have basically now defined 2 regularizers which are both strongly convex; but in two with respect to two different norms. Your Euclidean regularizer is strongly convex in l-2 norm; your Entropy regularizer is strongly convex in l-1 norm ok. So, that is fine.

We will see how considering this regularizer's in different with regularizers which are strongly convex in with respect to different norms; how does this effect? So, fine. Let us come back to just regularize having a strong convex functions ok. Now, what we will show is and also, we will be interested in functions which are Lipschitz's.

(Refer Slide Time: 32:59)




Now ok, now what we were interested in follow a regularized leaders with strongly convex functions ok. I will just state the result and we will look into this proof in the next time. So, this if you have this, if I going to apply my follow the regularize leader a strongly convex functions, this is what we will get. So, I am now just stating that what is the role of strong convexity of my regularizers in the bounds, we are going to get, when I going to apply my follow the regularize leader algorithm right.

So, remember we had terms like this, when we add follow the regularize regret. We needed to bound the difference of this function $f_t$ and round t at the point $w_t$ and $w_{t+1}$. Now, if my function $f_t$ is already Lipschitz, this holds already right. But now, further we are going to say that if my regularizer R is strongly convex, this can be upper further bounded as $L_t$ square by sigma. Sorry, this is also $L_t$ here; this is also $L_t$ here. So, this part is obvious the first one, the second one is coming from the strong convexity of my ok.

So, just what is the statement, we are saying you take a regular a function R which is strictly convex function, for with respect to some norm and now, take assume that this ft is to be Lipschitz's and that Lipschitz's constant is $L_t$ for function t again with respect to the same norm. So, you have to because we need to specify the norm, we are saying that these are specified with respect to same norm and if your algorithm is predicting this $w_2$, then this is the bound. So, in the next one, we will show this and see that how this helps in getting the bounds for both kind of regularizers.