

**Deep Learning**  
**Prof. Mitesh M. Khapra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Module - 8.4**  
**Lecture – 8**  
 **$l_2$  regularization**

(Refer Slide Time: 00:13)

Different forms of regularization

- $l_2$  regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

The slide includes the NPTEL logo, the name 'Mitesh M. Khapra', and the course information 'CS7015 (Deep Learning) : Lecture 8'. A small video inset shows the professor speaking.

So, let us start with  $l_2$  regularization. So, I have seen this before.

(Refer Slide Time: 00:15)

• For  $l_2$  regularization we have,

$$\tilde{L}(w) = L(w) + \frac{\alpha}{2} \|w\|^2$$

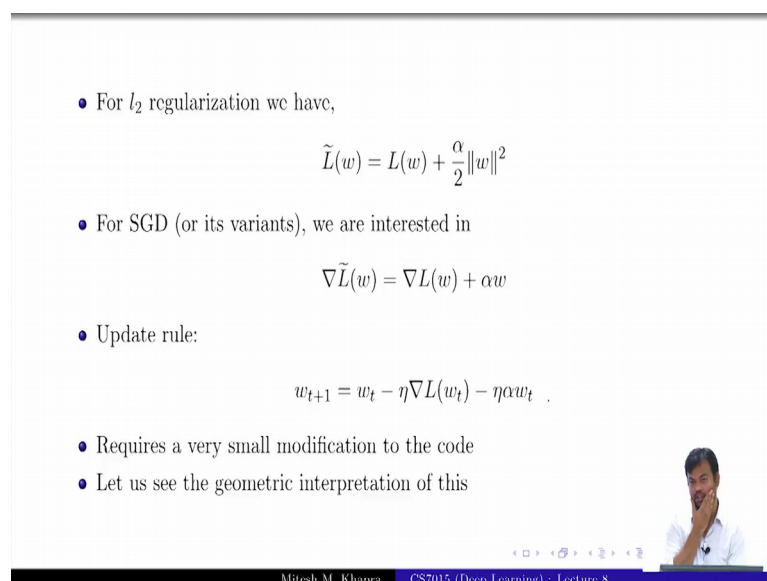
The equation is annotated with red circles around  $\tilde{L}(w)$ ,  $L(w)$ , and  $\frac{\alpha}{2} \|w\|^2$ . A red arrow points from the text 'we have,' to the equation. A red box labeled  $\mathcal{R}(w)$  is drawn around the regularization term, with an arrow pointing to it from the text above. The slide also features the NPTEL logo, the name 'Mitesh M. Khapra', and the course information 'CS7015 (Deep Learning) : Lecture 8'. A small video inset shows the professor speaking.

So, all of you see that this is  $l_2$  regularization right. What does  $l_2$  regularization does? Now tell me in the context of things that we have discussed today what is this? Empirical estimate of the train error and what is this. Is that fine right? So, everything that we are going to write is  $l_2$ , because of its  $w$ , but fine right ok. Now why does this relate to model complexity, what am I doing here actually by adding this?

So, they are going to see a very detailed analysis of this, but I just want to see first whether you get an intuition behind this. So, by doing that what you are trying to do? Not allow the model to become very complex right. You do not want a model where your weights can take any possible value; you just want the weights to be small. So, you are reducing the freedom on the model right, less freedom less complex, you get the intuition at least. We will see this in more detail, but at least you get the intuition why we are doing this.

So, we are using  $\omega$ . Remember that we are using this  $\omega$  theta as a surrogate for model complexity. So, if you add something in all  $\omega$  theta, just make sure you understand that this relates to model complexity ok fine. And now for sgd what would I need, for gradient descent. Just in case you have forgotten what gds, what do we need? Nothing, you have done, it will give you gradient of this, which is a sum of the derivatives of the two quantities of which you know one right. You know this already, and what is the other guy;  $\alpha w$  ok.

(Refer Slide Time: 01:59)



• For  $l_2$  regularization we have,

$$\tilde{L}(w) = L(w) + \frac{\alpha}{2} \|w\|^2$$

• For SGD (or its variants), we are interested in

$$\nabla \tilde{L}(w) = \nabla L(w) + \alpha w$$

• Update rule:

$$w_{t+1} = w_t - \eta \nabla L(w_t) - \eta \alpha w_t .$$

• Requires a very small modification to the code

• Let us see the geometric interpretation of this

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So, you see this  $l_2$  regularization right one reason why it is preferred is now imagine you have already written code for gradient descent. All you need to do is change it at one place add this to your update rule; that is all you need and you can think of the vector form of this, where you have a vector of parameters, you can think of the matrix form of this variable vector matrix of parameters. All you need to do is add one term to your update rule, so it can be done with very minimalistic change and this would be your update rule. Now let us see geometric interpretation of this.

(Refer Slide Time: 02:36)

• Assume  $w^*$  is the optimal solution for  $L(w)$  [not  $\tilde{L}(w)$ ] i.e. the solution in the absence of regularization ( $w^*$  optimal  $\rightarrow \nabla L(w^*) = 0$ )

$w^*$   $w = w^* + h$   $h = w - w^*$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

Now, from here onwards some of you will start getting a bit uncomfortable with some of the math, because of these assumptions that it only works for squared eggs in a vacuum right. So, you will see those kind of things, I will not tell you upfront what is the assumption I am making, because that will just spoil the analysis, you will just not enjoy it as much as you would ignorance is bliss right. So, if you do not know what the assumptions are, you will probably enjoy it more.

But for some of you will pick it up, just keep it to yourself, at the end I will tell you what are the assumptions I had made ok. There are some tricky assumptions that I want to make, but just live with it and just try to enjoy it while those assumptions last right ok. So, now, let us assume that  $w^*$  is the optimal solution for  $L(w)$ , what is  $L(w)$ ? The train error, not our regularized error, just the train error ok.

And, so if  $w^*$  is the optimal solution what can you take tell about the derivative with respect to  $w^*$  or derivative at  $w^*$  sorry, it is going to be 0 from basic calculus right. So, which I say minimize  $x$  square, the minima is where derivative of  $x$  squared with respect to  $x$  is equal to 0 right, everyone knows this ok.

So, now consider one point which is ok. So, what I actually want to consider is that, let me just see how to see this. So, let us see my  $w^*$  and I want to consider some point in the neighborhood of  $w^*$  ok; that is what I want do. So, one way of saying it is that  $h$  is equal to  $w - w^*$ , is that fine ok. So, that is what I am going to use in the next few steps.

(Refer Slide Time: 04:16)

- Assume  $w^*$  is the optimal solution for  $L(w)$  [not  $\tilde{L}(w)$ ] i.e. the solution in the absence of regularization ( $w^*$  optimal  $\rightarrow \nabla L(w^*) = 0$ )
- Consider  $h = w - w^*$ . Using Taylor series approximation (upto 2<sup>nd</sup> order)

$$L(w^* + h) = L(w^*) + (h)^T \nabla L(w^*) + \frac{1}{2} (h)^T H(h)$$

$$L(w) = L(w^*) + (w - w^*)^T \nabla L(w^*) + \frac{1}{2} (w - w^*)^T H(w - w^*)$$

$$= L(w^*) + \frac{1}{2} (w - w^*)^T H(w - w^*) \quad (\because \nabla L(w^*) = 0)$$

$$\nabla L(w) = \nabla L(w^*) + H(w - w^*)$$

Handwritten annotations in red include:
 

- A circle around  $\tilde{L}(w)$  in the first bullet point.
- A circle around  $(h)^T H(h)$  in the first equation.
- A circle around  $\nabla L(w)$  in the final equation.
- Handwritten  $w^T H u$  next to the first equation.
- Handwritten  $h^T H h$  next to the second equation.

Footer: Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So, suppose I have such an  $h$  which is equal to  $w - w^*$ ; that means, I can move from  $w^*$  to some point in its neighborhood by using  $h$ . And what does Taylor series tell us? This is what Taylor series tells us right, that the value of the function at this neighborhood point is equal to this. All of you know Taylor series well now, it is that fine I do not need to really go over this right everyone is with.

This is approximation up to the second term second order derivative. Now what was  $h$  actually  $w - w^*$ . So, I will just substitute that and this is what I get, is that fine. What is this quantity?  $1 - 0 - \infty - \infty - 0$  right we just did that ok. So, that term will disappear, what am I left with? This quantity and I have forgotten what is next ok.

Now again I am interested in the derivative of this ok. So, what will happen if I take the derivative what would I get? I am interested in computing  $\text{grad } L w$ , what will the R H S be, how many of you fine with this? Remember this is a quadratic form right. So, this is of the form  $x$  square that is, I mean that is roughly how I remember it is not correct, because of the form  $x$  square. So, when you take the derivative one of the  $x$  is will disappear and this quantity will remain ok. So, everyone gets this ok.

So, now what do I have is, I have the formula for the gradient with respect to  $L w$  and it is in terms of the gradient with respect to or rather the gradient at  $L w^*$ ; that is what I have achieved. So, far, but what am I actually interested in, the regularized loss, I am, what I am still dealing with, is the non regularized loss. This is just the empirical estimate of the training error that is not what I am interested in, I am interested in the regularized loss.

How many of you lost at this point? Oh  $h$  is the second order derivative oh. So, these are brackets just for clarity, but I see it is making it more unclear yeah, actually we should have used  $u$  and then call it  $u^T H u$ . So, it is the brackets here are not indicating function this is just  $h^T H$ . Now let us say it I realize how bad it is. So, last step what are we taking gradients with respect to is  $w$  right, is it fine.


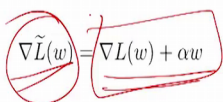
So, we have a. So, I mean do not get too confused right. So, up till this point we have a formula for  $L w$  right, and I am just interested in the derivative of that ok. And all I have achieved by this is that, I have ok. In fact, I have one more step right.

(Refer Slide Time: 06:50)

- Assume  $w^*$  is the optimal solution for  $L(w)$  [not  $\tilde{L}(w)$ ] i.e. the solution in the absence of regularization ( $w^*$  optimal  $\rightarrow \nabla L(w^*) = 0$ )
- Consider  $h = w - w^*$ . Using Taylor series approximation (upto 2<sup>nd</sup> order)

$$L(w^* + h) = L(w^*) + (h)^T \nabla L(w^*) + \frac{1}{2} (h)^T H(h)$$
$$L(w) = L(w^*) + (w - w^*)^T \nabla L(w^*) + \frac{1}{2} (w - w^*)^T H(w - w^*)$$
$$= L(w^*) + \frac{1}{2} (w - w^*)^T H(w - w^*) \quad (\because \nabla L(w^*) = 0)$$
$$\nabla L(w) = \nabla L(w^*) + H(w - w^*)$$
$$= H(w - w^*)$$

- Now,



Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

What is this quantity zero ok. So, we now know that the derivative of the loss function with respect to  $w$  can be written as this quantity. Is it ok, and I have just derived it step by step, there is nothing great about it, anyone is can. Why I am doing this is not clear that will become clear hopefully, but what I am doing is clear right, is that fine can I move i.

Now what we are actually interested in is this quantity, because this is the true loss that we are going to deal with right and we just saw in the previous slide that this quantity which is on the L H S is equal to this thing on the R H S, this is what we saw on the previous slide, can I just go back to the previous slide, because the derivative of this was just  $\alpha w$ . Now let us start with this. So, on the next slide, let me just see if there is anything else that I need to see here ok.

So, far everyone is clear what I have derived so far why is not clear, but what is clear, what is being derived so far. So, I have said that the derivative of the loss function or the regular is loss function can be written as this quantity ok, is that fine, where  $w^*$  is the optimal solution for with respect to the un regularized loss function ok. And now I have what I am interested in this solution with respect to the regularized loss function ok.

(Refer Slide Time: 08:08)

• Let  $\tilde{w}$  be the optimal solution for  $\tilde{L}(w)$  [i.e regularized loss]

$$\because \nabla \tilde{L}(\tilde{w}) = 0$$
$$H(\tilde{w} - w^*) + \alpha \tilde{w} = 0$$
$$\therefore (H + \alpha \mathbb{I})\tilde{w} = Hw^*$$

*Handwritten red text:*  $H\tilde{w} - Hw^* + \alpha\tilde{w} = 0$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

Now, let  $w$  tilde be that solution for the regularized loss function. So; that means, the derivative of the loss, the regularized loss function at  $w$  tilde is going to be 0, nothing great about this, but I just told you on the previous slide that I can write this quantity as this quantity that is what we derived on the previous slide ok. Just take my word that is what we derived on the previous slide ok, let just, no confidence in me ok that is fine. Now can you. Are you if I write it as this, just rearranging some terms oh sorry.

So, I am just grouping all the  $w$  tilde some terms and this is, a matrix is needed here right, because I need to, I can only add two matrices. So, what I am just doing is, putting the elements across the diagonal. Everyone understands this, everyone gets this step.

(Refer Slide Time: 09:19)

• Let  $\tilde{w}$  be the optimal solution for  $\tilde{L}(w)$  [i.e regularized loss]

$$\because \nabla \tilde{L}(\tilde{w}) = 0$$
$$H(\tilde{w} - w^*) + \alpha \tilde{w} = 0$$
$$\therefore (H + \alpha \mathbb{I}) \tilde{w} = H w^*$$
$$\therefore \tilde{w} = (H + \alpha \mathbb{I})^{-1} H w^*$$

• Notice that if  $\alpha \rightarrow 0$  then  $\tilde{w} \rightarrow w^*$  [no regularization]

• But we are interested in the case when  $\alpha \neq 0$

• Let us analyse the case when  $\alpha \neq 0$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So, now I have a formula for  $w$  tilde in terms of  $w$  star ok. I am going to go a bit further and be a bit bold and compute the inverse also. So, now, I have an exact formula for  $w$  tilde in terms of  $w$  star. So, what is this, actually what is this relation that I am trying to establish? Suppose I know the solution with respect to the unregularized loss, and now I have added regularization what happens to the new solution.

So, I am telling you the new solution would be smaller weights and so on that is what L2 regularization tells you, now you are just trying to make an interpretation for that. So, I have given you a closed form solution that  $w$  tilde is actually equal to this quantity that you see on the right hand side ok. Why you are doing this is still not clear but right now I just focus on the what part of it this is just some mathematical steps that I am doing, anyone who is not comfortable with this.

Now notice what would happen if  $\alpha$  tends to 0 what would be  $w$  tilde be  $w$  star, what do you mean by  $\alpha$  equal to 0, no regularization right. So, that is just one corner case that I want to do, but that is not what we care about anything what that is stupid to do all this and tell you that if you do not use regularization you will get the same solution, but that is not what I am going to tell you right. We are interested in the case when  $\alpha$  is not equal to 0 ok. So, let us look at that case.



(Refer Slide Time: 10:35)

• If H is symmetric Positive Semi Definite  
 $H = Q\Lambda Q^T$  [Q is orthogonal,  $QQ^T = Q^TQ = \mathbb{I}$ ]

$$\begin{aligned}\tilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\ &= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^* \\ &= \underline{(Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1}} Q\Lambda Q^T w^* && a, b + \alpha z, b \\ &= \underline{[Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1}} Q\Lambda Q^T w^* \\ & \quad \underline{(ABC)^{-1}}\end{aligned}$$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

Now, I am going to assume that H is a symmetric positive semi definite matrix squared  
egg in a vacuum ok. So, if that is the case then I can write H as this, I have just done the  
dash of H, eigenvalue decomposition all right ok, and I know that since it is a squared  
symmetric matrix the eigenvalues are going to be eigenvalues are going to be orthogonal  
yes, eigenvalue vectors are going to be orthogonal and that is why I can write this that  
cute suppose as the inverse of Q ok.

Now, let us start with whatever we had on the previous slide and substitute what, what I  
am going to substitute? Instead of H I am going to use Q lambda Q transpose ok. So, I  
am doing that. So, is that ok, I will just go over the steps and let me know at any point if  
you have a problem. What I have done is, I have replaced this I by this and its valid,  
because Q Q transpose is just equal to I, I have just taken q and q transpose as common  
right. So, this is a c b plus some a z b. So, I have taken a and b out right, is that fine ok.

Now, what is the next thing I am going to do? This is of the form a b c inverse. So, I am  
going to write it as, and the inverses are neat right ok.

(Refer Slide Time: 12:08)

• If  $H$  is symmetric Positive Semi Definite  
 $H = Q\Lambda Q^T$  [ $Q$  is orthogonal,  $QQ^T = Q^TQ = \mathbb{I}$ ]

$$\begin{aligned}\tilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\ &= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^* \\ &= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1}Q\Lambda Q^T w^* \\ &= [Q(\Lambda + \alpha\mathbb{I})Q^T]^{-1}Q\Lambda Q^T w^* \\ &= Q^{T^{-1}}(\Lambda + \alpha\mathbb{I})^{-1}Q^{-1}Q\Lambda Q^T w^* \\ &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \quad (\because Q^{T^{-1}} = Q) \\ \tilde{w} &= QDQ^T w^*\end{aligned}$$

where  $D = (\Lambda + \alpha\mathbb{I})^{-1}\Lambda$  is a diagonal matrix which we will see in more detail soon

35/84

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

This is fine, what will happen to this quantity  $\Lambda$ , what is this quantity  $Q$  and this is what I am left with, but there is still something more I can do I guess let us see ok. So, I can write this entire thing as a diagonal matrix. How many of you see that it is a diagonal matrix, because  $\Lambda$  is a diagonal matrix,  $\mathbb{I}$  of course, is a diagonal matrix,  $\Lambda$  is multiplied by a scalar which is also going to be a diagonal matrix and the whole thing is again multiplied by some diagonal matrix ok. What is the inverse of a diagonal matrix? The reciprocal of the diagonal elements.

So,  $\Lambda$  is fine. So, I have a very neat formula for what  $w$  tilde looks like in terms of  $w$  star ok. Again why am I doing all this and God knows, but here  $D$  is equal to this quantity ok.

(Refer Slide Time: 13:00).

$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} =$

$$\tilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$
$$= Q D Q^T w^*$$
$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1 + \alpha} & & & \\ & \frac{1}{\lambda_2 + \alpha} & & \\ & & \dots & \\ & & & \frac{1}{\lambda_n + \alpha} \end{bmatrix}$$
$$D = (\Lambda + \alpha \mathbb{I})^{-1} \Lambda$$

- So what is happening here?
- $w^*$  first gets rotated by  $Q^T$  to give  $Q^T w^*$
- However if  $\alpha = 0$  then  $Q$  rotates  $Q^T w^*$  back to give  $w^*$
- If  $\alpha \neq 0$  then let us see what  $D$  looks like

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So what exactly is happening here, in terms of linear algebra or in terms of geometric interpretations ok. So, let me just see if I have to do something first ok. So, what is happening to  $w^*$  is getting

Student: (Refer Time: 13:15).

Rotated, remember what happens when a matrix where hits a vector, it gets rotated and scaled also. And then what is this diagonal matrix going to do, scale it, element by scaling actually, everyone gets this operation and then I am again rotating it by  $Q$  ok. Again the same stupid question if  $\alpha$  is equal to 0 what would happen?  $Q$  transpose what rotated by something and then  $Q$  would rotate it back way; that means, you will end up getting the same solution ok. If  $\alpha$  is equal to 0 we understand.

Now if  $\alpha$  is not equal to 0. First let us see what does this matrix look like ok. So, what is this matrix actually, it is a diagonal matrix, what are the diagonal elements, the. What is the first element in the diagonal? 1 by, everyone agrees with this, what is the second element fine, and what is the other matrix that I have?  $\Lambda$ . So,  $D$  is equal to the product of these two things right. So, what is  $D$  going to be? What is the first element of this matrix is going to be, how many if you say  $\lambda_1$  by  $\lambda_1$   $\lambda_1 + \alpha$ , this much is clear, everyone gets this ok.

So, this is a diagonal matrix of the form a b c, let us consider a 3 by 3 matrix ok. Now I am going to multiply it by another matrix x y is z which is also a diagonal matrix right, because this is also it. So, this matrix I have already told you what it looks like, the other matrix is also a diagonal matrix. Now what is this product actually a x, b y, c z and everything else has 0. Now everyone gets it. now can you say what would this product look like if you can actually make out, it would be a diagonal matrix and what would the diagonal elements be? Is it fine with everyone now ok, is it ok, fine.

(Refer Slide Time: 15:18)

$$\tilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$

$$= Q D Q^T w^*$$

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1 + \alpha} & & & \\ & \frac{1}{\lambda_2 + \alpha} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n + \alpha} \end{bmatrix}$$

$$D = (\Lambda + \alpha \mathbb{I})^{-1} \Lambda$$

$$(\Lambda + \alpha \mathbb{I})^{-1} \Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1 + \alpha} & & & \\ & \frac{\lambda_2}{\lambda_2 + \alpha} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n + \alpha} \end{bmatrix}$$

- So what is happening here?
- $w^*$  first gets rotated by  $Q^T$  to give  $Q^T w^*$
- However if  $\alpha = 0$  then  $Q$  rotates  $Q^T w^*$  back to give  $w^*$
- If  $\alpha \neq 0$  then let us see what  $D$  looks like
- So what is happening now?

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So, now what is happening? So, first this rotation is happening that no one is denying, after rotating what is happening, this is a, this product is actually a vector; that is fine ok. What are we doing to every element of the vector, scaling it, scaling it by what quantity these quantities that every element is getting scaled by the corresponding entry in the diagonal, in this diagonal right.

So, the first entry is getting scaled by this, the second entry is getting scaled by this and so on ok. I just want you to take some 30 seconds and try to figure out where I am headed from here.

(Refer Slide Time: 16:03)

$$\tilde{w} = Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^*$$
$$= Q D Q^T w^*$$

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1 + \alpha} & & & \\ & \frac{1}{\lambda_2 + \alpha} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n + \alpha} \end{bmatrix}$$

$$D = (\Lambda + \alpha \mathbb{I})^{-1} \Lambda$$

$$(\Lambda + \alpha \mathbb{I})^{-1} \Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1 + \alpha} & & & \\ & \frac{\lambda_2}{\lambda_2 + \alpha} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n + \alpha} \end{bmatrix}$$

- Each element  $i$  of  $Q^T w^*$  gets scaled by  $\frac{\lambda_i}{\lambda_i + \alpha}$  before it is rotated back by  $Q$
- if  $\lambda_i \gg \alpha$  then  $\frac{\lambda_i}{\lambda_i + \alpha} = 1$
- if  $\lambda_i \ll \alpha$  then  $\frac{\lambda_i}{\lambda_i + \alpha} = 0$
- Thus only significant directions (larger eigen values) will be retained.

Effective parameters =  $\sum_i^n \frac{\lambda_i}{\lambda_i + \alpha} < n$

Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

Let us see if I can, yeah maybe look at this sentence and see. First of all everyone agrees with this sentence right. Is there anyone who does not agree with the sentence? I am just trying you to figure out the implication of the sentence, you get it, some people are nodding their heads just in, because if you scale it right, then there is no guarantee that what the vector has changed ok, what happens in the following case; that means, that dimension will be left as it is ok, but if the eigen, if this condition holds what would happen that dimension is almost getting multiplied by a 0 right.

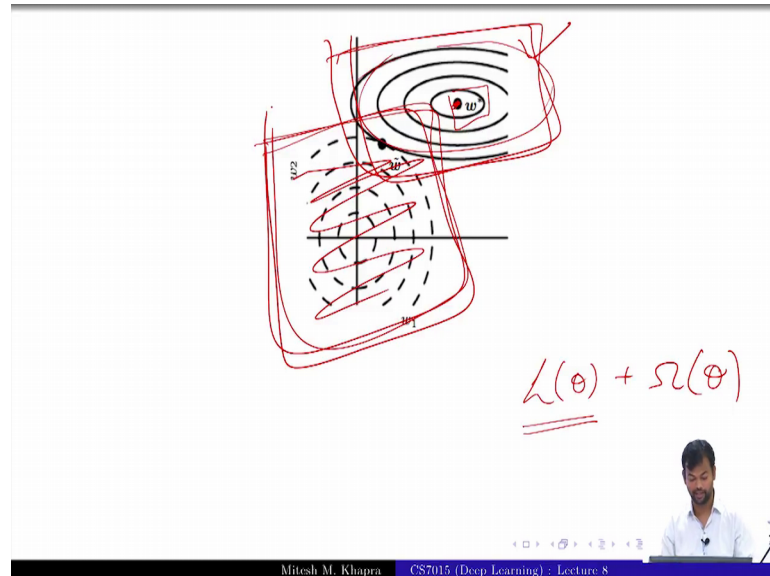
So, see these two extremes, when the eigen value is very large you will end up staying where you were, so those dimensions will not be affected. If the eigen value is very small then you are almost getting scaled down to 0. So, now, what will happen is actually, only the significant directions larger eigen values will be retained. So, what is the effective number of parameters in your model now?

See remember that this  $w$  vector is a vector of all the parameters, what am I telling you that some of these are going to disappear, when which condition holds, the third can, the third bullet holds, some of these are going to disappear. That means, the effective number of parameters, which remain in your model is going to be less right and you see that it is going to be given by this quantity right.

So, that is sometimes known as the effective number of parameters in a neural network. If the effective number of parameters in your neural network is decreasing; that means,

what you are doing, making the model less complex right. So, that is what we have achieved, you see that ok.

(Refer Slide Time: 17:50)



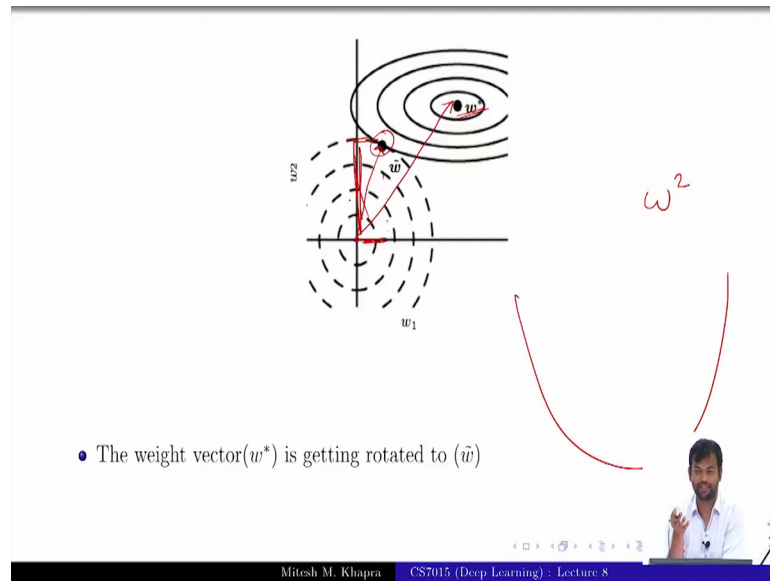
Now, let me end with a pictorial interpretation of this. You see two figures here and there is only one figure, but you see two different things here. Can you tell me what this is and what this is, that is the first question I want to ask you. The hint is that in this lecture we care about, the other hint is what was  $w^*$ , the solution for the.

Student: (Refer Time: 18:23).

Unregulated loss which means which loss  $L$  theta, you need any more hints. Sorry, this box is the contours of  $L$  theta, this box contours of  $\omega$  theta. So, this thing just ignore this part of the figure for now ok. This I have marked as  $w^*$ ,  $w^*$  was the solution when I only had the un regularized loss ok. There is the solution when I had the un regularized loss ok.

So, remember the contour maps that we had seen. So, this is the minimum of that particular function. So, this is the contour map for  $L$  theta; that is clear. Now what probably is not clear is, why is this the contour map of  $\omega$  theta, let me just go ahead actually.

(Refer Slide Time: 19:20)



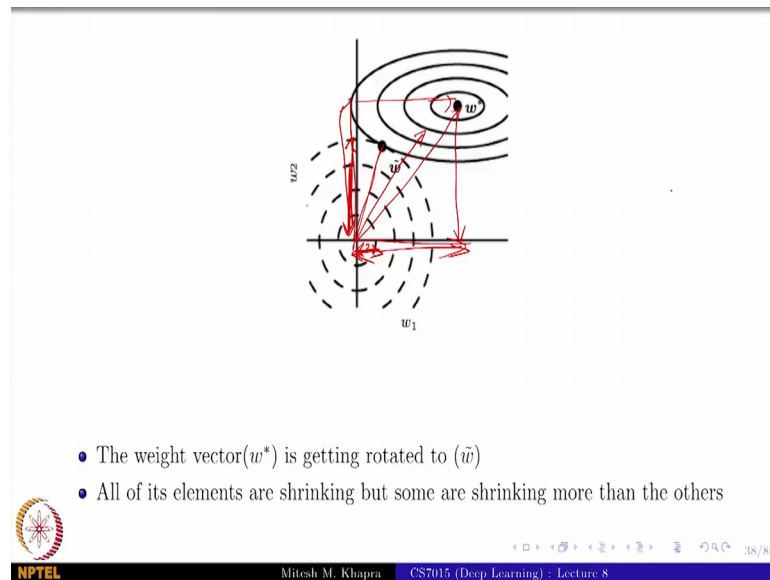
Please do not read this, this is the prestige ok. So, do not read that. So, this is the contour map of omega theta right, because omega theta in the, what is the minima for the omega theta, it is a function of the form  $w$  square, what is the minima? 0 and what does that function look like and what is this point, 0 the origin right. So, that is why this is the contour for omega theta ok.

Now, what is happening? This was the solution when you had without regularization and now this is  $w$  tilde which is a solution with regularization. So, can you make some commentary on this, with respect to not just general commentary, with respect to the things that we saw in the derivation. We talked about rotation, scaling, dimension specific scaling, so what is happening? This was my original solution vector; this was my original solution vector when I did not have the regularization term, now what has happened? The rotation has happened and we saw that there is a rotation operation happening, more importantly what has happened? Scaling has happened.

More importantly what has happened dimension specific scaling is happening right. One dimension has not, this dimension has scaled down this dimension has not scaled down enough; that is exactly what we wanted right. We wanted the less important weights to go down and the more important weights to stay there. We did not want a uniform scaling down; we wanted a dimension specific scaling down.

So, the weight vector has been rotated yes, each dimension after rotation has been scaled, some dimensions have been scaled down more, the other dimensions have been scaled down less, how many of you can make this interpretation from the figure, now that I have told you this interpretation.

(Refer Slide Time: 21:33)



Now, still if you do not how mean if you can still have a doubt with this, you still have a doubt what is doubt fine so, ok. So, this was the original solution vector right. The map told us that what actually happens is when you add this omega theta the solution vector gets rotated ok, at the same time there is also some scaling down and that scaling down is for dimension.

How many dimensions do you have here? Two dimensions right. So, this is one dimension, this is the other dimension. Now in the original case both these weights actually seemed almost equal right. I mean if you look at the  $w_1$  coordinate and the  $w_2$  coordinate they were same. Now after this regularization what has happened is, what are the new coordinates for  $w_1$  and  $w_2$ , this is the coordinate for  $w_1$  right, this is the value of  $w_1$  and this is the value for  $w_2$ .

Both of them are admittedly smaller than the original values for  $w_1$  and  $w_2$  in the absence of regularization or both of them equally smaller. No they are being scaled differently, one rate has been scaled down more, the other weight has been scaled down lesser right and that is exactly what the math was telling us that they get scaled in



proportion to those  $\lambda_1$  by  $\lambda_1 + \alpha$  and that is exactly what we see in the figure is that fine.

How many if you get this interpretation now is that ok. So, all of its elements are shrink oh. You have a question. So, this final resultant right it is. So, what would have happened is that there would have been first rotation then scaling down and then again rotation. So, what you are seeing here is the final rotation right. So, it is not, it should have been showed in three steps by just shown the final step, is that ok.

So, its question was that we first had a rotation, then had a scaling and then again a rotation, but I even as explained in the figure I spoke only about one rotation. So, I basically clubbed both the rotations, and so what you see finally is rotations scaling down and again rotation, is that fine.