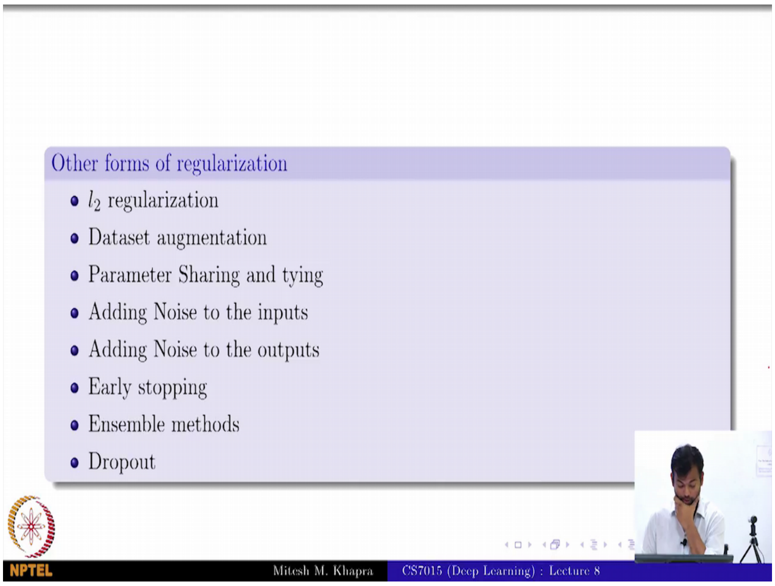
Deep Learning
Prof. Mithesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 8.9
Lecture – 08
Early Stopping

I will do, will do early stopping where again we will get into some of these eigenvector analysis. So, let us see that.

(Refer Slide Time: 00:18)

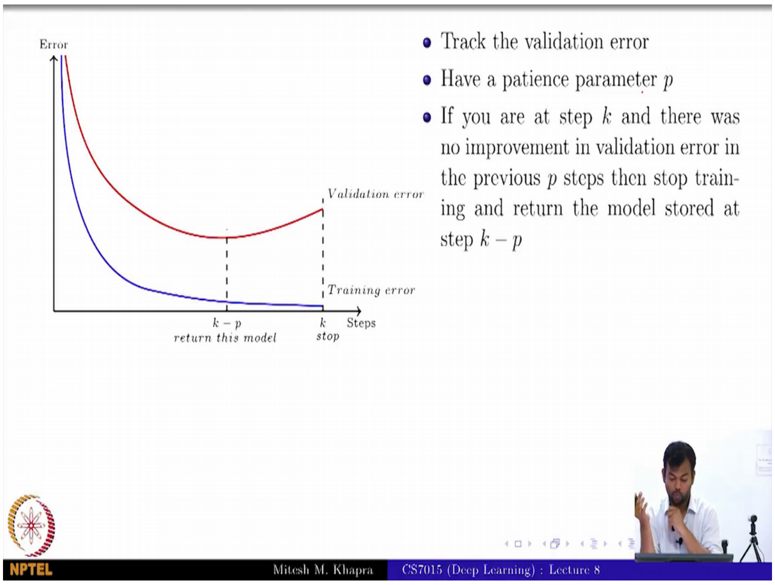


Other forms of regularization

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

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

(Refer Slide Time: 00:21)



- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$

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

So, the idea been early stopping is actually very simple in principle what needs to be done. So we know that, this that this trend exists between the training error and the tester right. So in practice, what you will do is you will continue to optimize the training error; the empirical training error which is the sum of the errors on the m training points.

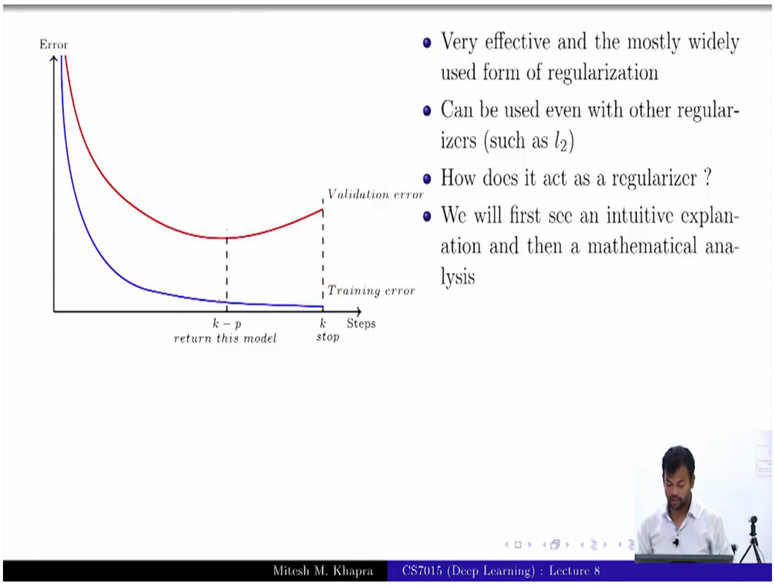
You will also continuously keep track of the validation error; that means, the same quantity you will compute over the n validation or test points. Everyone get this, you can do this and you are actually doing this in your back propagation assignment. Keeping track of the training error as well as the validation error and you keep plotting them. I will keep running for various epochs and keep something known as a patients parameter p.

So, if you are at the 20th epoch and if your patients parameter p is equal to phi; And just do a check whether, in the phi last phi epochs has my validation error. Ever gone down or it has been staying the same or has it been increasing ok. Now I will give you a condition that it was either staying the same or it was actually increasing is this good or bad.

What does it tell you while your training error was of course, decreasing may the more you train your training error will keep going down? So, what does this tell it is just over fitting you are fitting the training error you are just making it 0 or as close to 0 as possible, but that is not helping your validation error. So, the validation error is either worst case increasing or remaining the same right.

So, this is a very commonly used trick which is known as early stopping you keep this passions patients parameter. And you make sure that if you have cross this patients right and the patients here is that I was waiting for the validation error to go down, but it is not going down for some p epochs; So, no point in continuing training anymore I will just stop it does not make sense.

(Refer Slide Time: 02:21)



- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as l_2)
- How does it act as a regularizer ?
- We will first see an intuitive explanation and then a mathematical analysis

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

So, and this can also be used in conjunction with other regularizers right. So, in the quiz also we had this question sorry for bringing up the quiz, but we also at this question where you have the sparsity regularization and I was asking whether I can add the l 2 regularization along with it right. So, these regulations can be added or used in conjunction it is not that you can only use one of them.

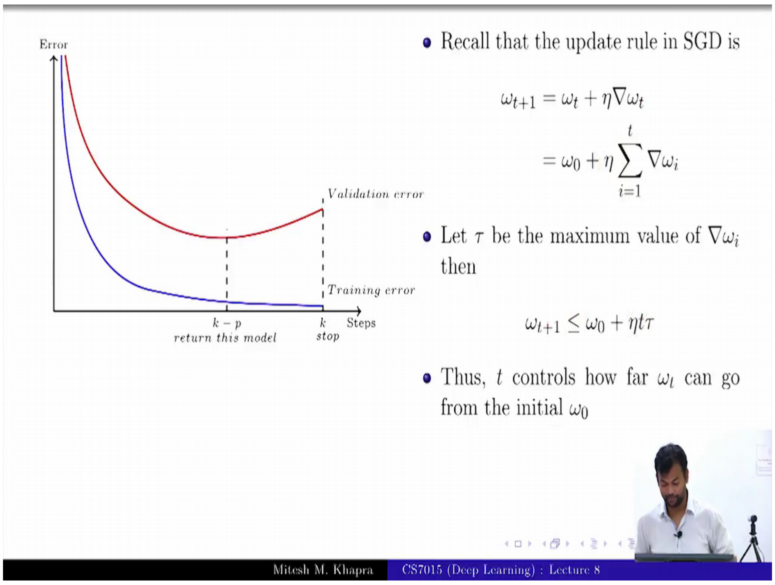
So, early stopping is a way of regularizing, but you could also use it in conjunction with l 2 regularization or any other regularization technique that you do not want right so, but how does this act as a regularizer from the picture? It is probably clear and is the same as the explanation. I was trying to give to his question right, that you are preventing yourself from entering in these regions and trying to enter into more favorable stop at more favorable regions right.

But can you think of slightly more in terms of, what happens in gradient? And what would happen if you stopped it early and so on. Can you try it to connect it to the update rule of gradient descent, what happens as you keep doing it for more and more epoch?

No gradient descent has nothing to do with validation error or backtracking error gradient descent only works on the training data let us think in those terms.

Gradient star diminishing to 0; so, what happens how does gradient descent progress? Where do you start? I started a random point at every epoch which is a collection of high iterations right, or you go or many training points. What happens to this, I start moving ok. I keep moving now, if I fix the number of epochs or do not allow it to change any more after a number of epochs. What am I doing? I am restricting the boundary around the weight right. I am not allowing it to grow beyond a certain boundary do you get that ok.

(Refer Slide Time: 04:13)



- Recall that the update rule in SGD is

$$\omega_{t+1} = \omega_t + \eta\nabla\omega_t$$
$$= \omega_0 + \eta\sum_{i=1}^{t}\nabla\omega_i$$

- Let τ be the maximum value of $\nabla\omega_i$ then

$$\omega_{t+1} \leq \omega_0 + \eta t\tau$$

- Thus, t controls how far ω_t can go from the initial ω_0

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

Let us see that. So, we will first see an intuitive explanation and then, go to a more mathematical analysis are update. So, the update rule for gradient descent is, I always make this mistake, this has to be minus oh the t s I have disappeared is there; so, sorry other to have disappear.

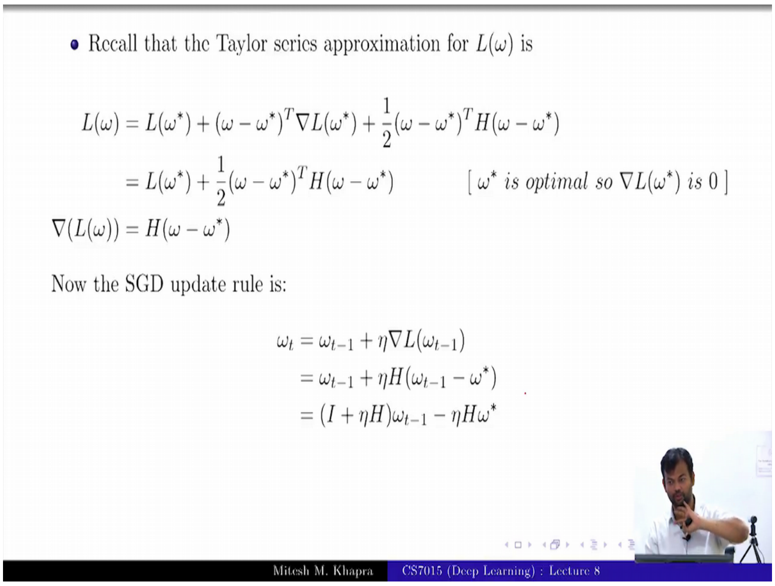
So now, what would actually happen at the t H step is we have w not 3 plus or minus does not matter. It just tells you that, how much it is going to change? This is what is happening actually at the t H step right. You have just subtracted all the previous derivatives that you had so far right. From where you started off now, you are looking at t steps. So, at every point you are computing a certain gradient, but had a certain magnitude.

Now, let me say that across all these steps, the maximum gradient that you had. I will just call it by tau right. So; that means, in this summation there are t terms, I am saying the maximum of those was tau that was the maximum rate gradient that I got at any one point ok, you get that.

Now, what I am going to do after this? I am going to replace this by something. This summation is always going to be less than or equal to this right. Because, I am assuming that each of my steps is less than tau, there are t such steps. So, I could have at matched moved t into tau right, but I would have moved less than that, because tau was the maximum gradient that I had ok.

So, this is going to be less than equal to is that do you get the change from the equality to less than equal to ok. So now, what am I restricting actually in early stopping, what is being restricted? There are only so many symbols there I just take 1 t tau is of course, not in your hands w naught is not in your hands w. So, t is the 1 right. So, I am only allowing that many updates so; that means, from w naught you can only moves that much this looks. You see that analogy that, this is something similar to you not allowing the weights to really grow a lot right is that, fine.

(Refer Slide Time: 06:27)



- Recall that the Taylor series approximation for $L(\omega)$ is

$$\begin{aligned} L(\omega) &= L(\omega^*) + (\omega-\omega^*)^T \nabla L(\omega^*) + \frac{1}{2}(\omega-\omega^*)^T H(\omega-\omega^*) \\ &= L(\omega^*) + \frac{1}{2}(\omega-\omega^*)^T H(\omega-\omega^*) \qquad [\, \omega^* \textit{ is optimal so } \nabla L(\omega^*) \textit{ is } 0 \,] \end{aligned}$$

$$\nabla(L(\omega)) = H(\omega-\omega^*)$$

Now the SGD update rule is:

$$\begin{aligned} \omega_t &= \omega_{t-1} + \eta \nabla L(\omega_{t-1}) \\ &= \omega_{t-1} + \eta H(\omega_{t-1} - \omega^*) \\ &= (I + \eta H)\omega_{t-1} - \eta H \omega^* \end{aligned}$$

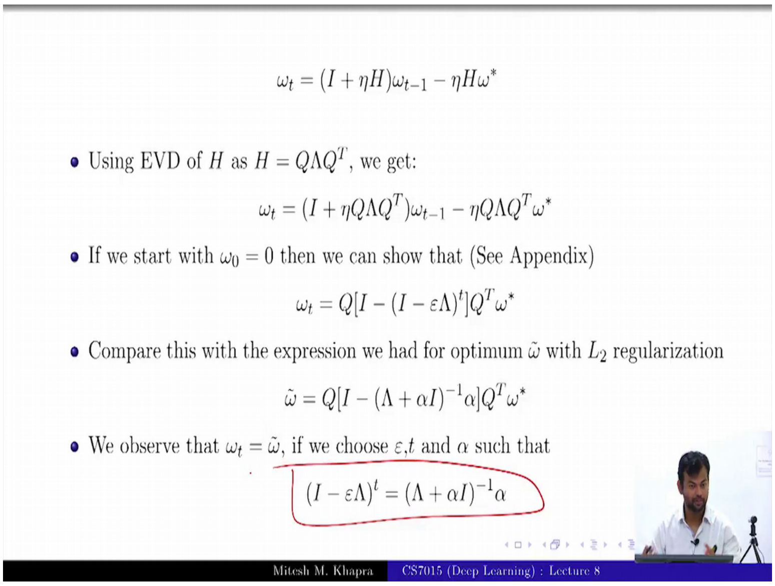
Mitesh M. Khapra CS7015 (Deep Learning) : Lecture 8

So now, but will not end here you will. Of course, do some more stuff on this right ok. So, we now see a mathematical analysis of this. So, recall that a Taylor series approximation for l w is the following. The same thing which I wrote a few slides back or many slides back everyone remembers this right. And now again I am going to do the same thing that, if I know the optimal w star then the gradient at that point is going to be 0.

So, this term disappears ok, and now if I take the derivative, this is what will remain. This is exactly what we did earlier also right. So, we will have derivative of this and derivative of this. So, the derivative of this quantity is just this and the derivative of this is 0. Because, that is exactly what we started off with right that w star is the optimal solution, everyone is fine with this right.

Now, SGD date rule is the following ok, which I can write as this. I just replaced this by this ok, I am just rearranging some terms is that ok. How many if you are fine with this? How many pages to tired to even care about this? I am just raising my hand ok.

(Refer Slide Time: 07:32)



$$\omega_t = (I + \eta H)\omega_{t-1} - \eta H\omega^*$$

- Using EVD of H as $H = Q\Lambda Q^T$, we get:

$$\omega_t = (I + \eta Q\Lambda Q^T)\omega_{t-1} - \eta Q\Lambda Q^T\omega^*$$

- If we start with $\omega_0 = 0$ then we can show that (See Appendix)

$$\omega_t = Q[I - (I - \varepsilon\Lambda)^t]Q^T\omega^*$$

- Compare this with the expression we had for optimum $\tilde{\omega}$ with L_2 regularization

$$\tilde{\omega} = Q[I - (\Lambda + \alpha I)^{-1}\alpha]Q^T\omega^*$$

- We observe that $\omega_t = \tilde{\omega}$, if we choose ε,t and α such that

$$\boxed{(I - \varepsilon\Lambda)^t = (\Lambda + \alpha I)^{-1}\alpha}$$

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

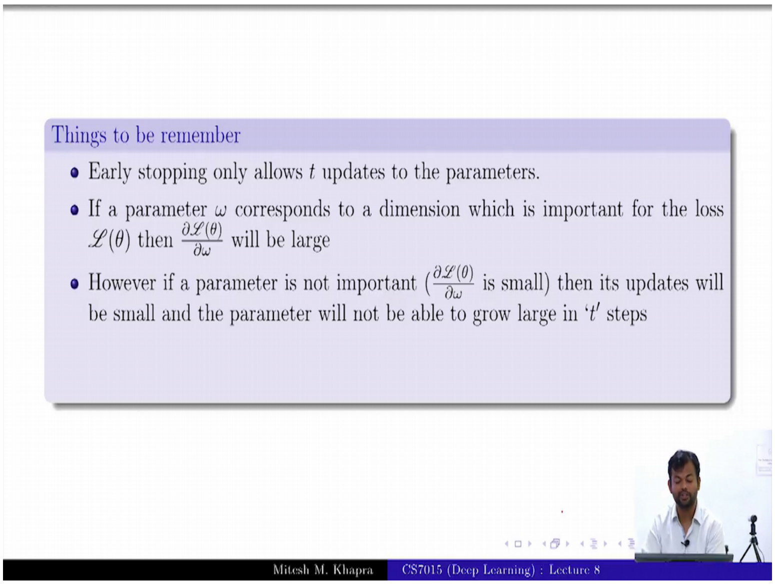
So, this is what w t would be this is again some simple steps leading to some conclusion. The conclusion is, what matters the steps are very easy you can go back and look at them right. So again, I will use the EVD the same trick that I did earlier and it will give me this instead of H ok. Again I will just do some rearrangements and actually, I can show that if I start with w naught equal to 0, then w 2 is actually given by this quantity ok. And there is a proof of this in the appendix you can go and look at it.

Now, what does this look similar to rotation diagonal rotation exactly similar to the analysis that we did for l 2 regularization right. And in fact, if you can you can show that, if we compare this expression with the 1 we had for l 2 regularization. And this is the expression that we had for l 2 regularization right. Rotation some scaling and then again rotation right; Then we can show that, early stopping is actually equivalent to l 2 regularization, if the following condition is satisfied.

This does not mean much because, god knows how you will satisfy this condition? Right, but all it is saying is that there is some equivalence. At under certain conditions and that is what is the intuition was also telling us that, it is somehow preventing the weights from going large and it is doing this; In this very convoluted way where this condition holds for it to be equivalent to l 2 regularization.

As I said for you and me is going to be very hard to create this condition right. How do I make sure that, something like this is true right, but that does not matter what matters is that, there is some equivalence between them?

(Refer Slide Time: 09:34)



Things to be remember

- Early stopping only allows t updates to the parameters.
- If a parameter ω corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ will be large
- However if a parameter is not important ($\frac{\partial \mathscr{L}(\theta)}{\partial \omega}$ is small) then its updates will be small and the parameter will not be able to grow large in 't' steps

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

So, when you are doing early stopping. It is not just a heuristic or a blind thing that you are doing. You know that, it is somehow related to l 2 regularization. Hence, that you are doing it and hence it also works in practice is it fine, we will that work for all of you right; So, the things to remember is that early stopping only allows t updates to the parameters. Ok, this is the important thing rights. So, now, if a parameter w corresponds to a dimension which is important for the loss. Then what would this quantity be the partial derivative of the loss with respect to that parameter it is going to be, if there is a parameter.

For example, let us take the America an example right. That whatever weight you gives to whether, the actor was American or not. If that is very important because, if that feature is on you are lost completely changes and so on right. If you do not learn the weight correctly that feature is very sensitive.

So, for important features the loss would be very sensitive to the changes. In the weights of these features, is that intuition correct right; that means, this gradient would be large and if a parameter corresponds to a feature which is not important. What would this derivative be small now, what is the net effect of this you have some parameter which are important. So, the derivatives are large some parameters which are not important.

So, the derivatives are going to be small and you are going to only allow t updates. So, what is going to happen? The parameters which are important; we will end up getting effectively more updates right; because, each of these magnitudes was higher and you did t of those. The parameters which are not important we will end up getting effectively lesser movement right.

Because, each of these gradients were small and you did only t of those right. So, you again see this that, it is a weird way of ensuring that your important parameters get more updates than your non important parameters right. So, it is very important to see these connections between these different regularization methods, all of you are fine with this fine.