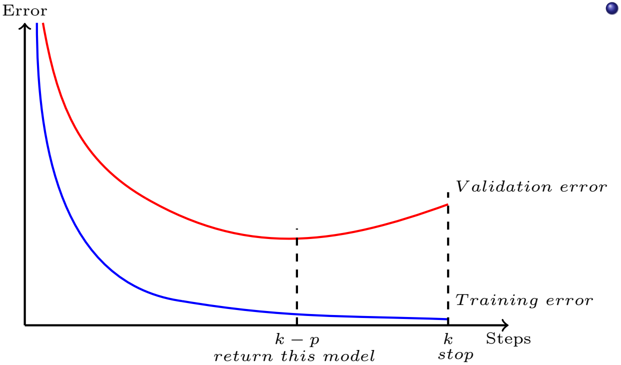
Module 8.9 : Early stopping

Other forms of regularization

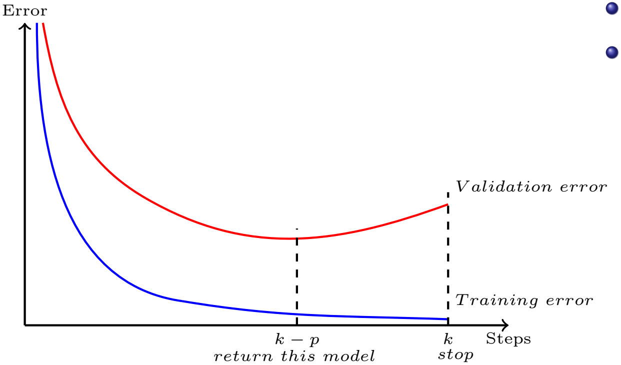
- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

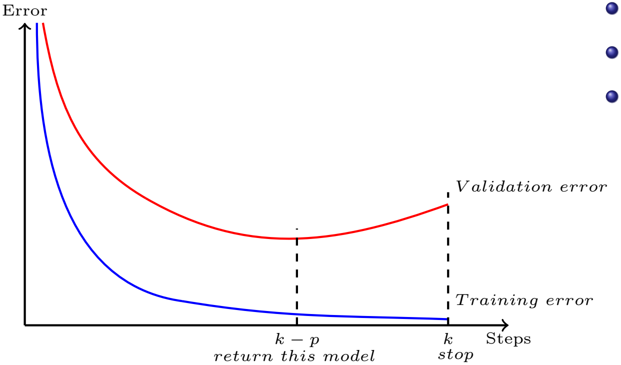
- L_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



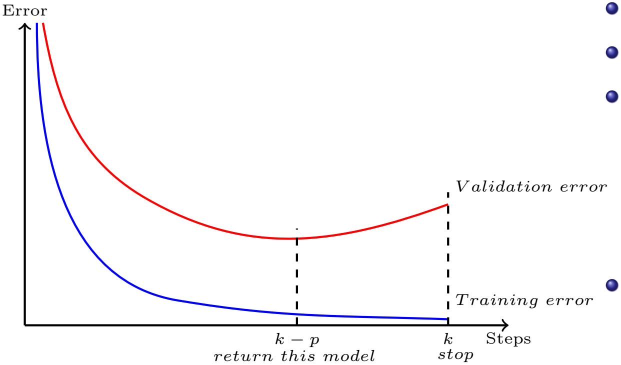
- Track the validation error



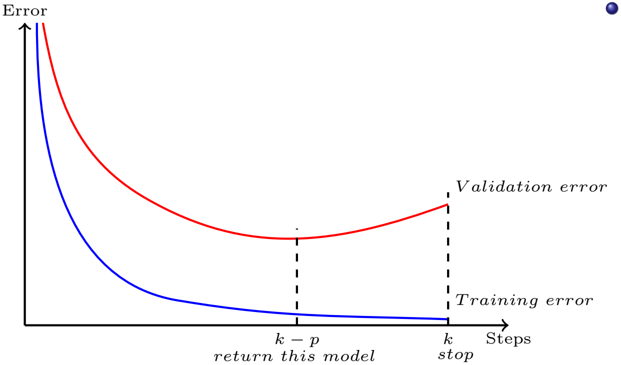
- Track the validation error
- Have a patience parameter p



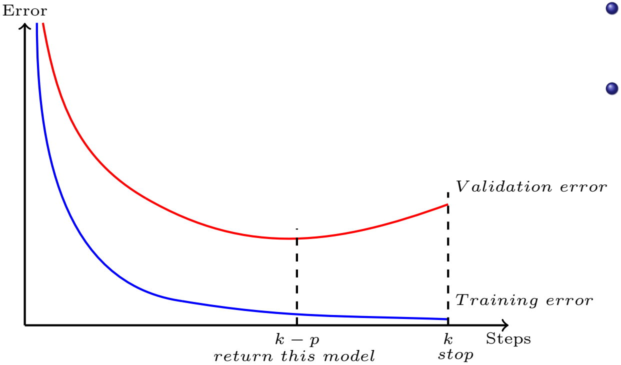
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k-p$



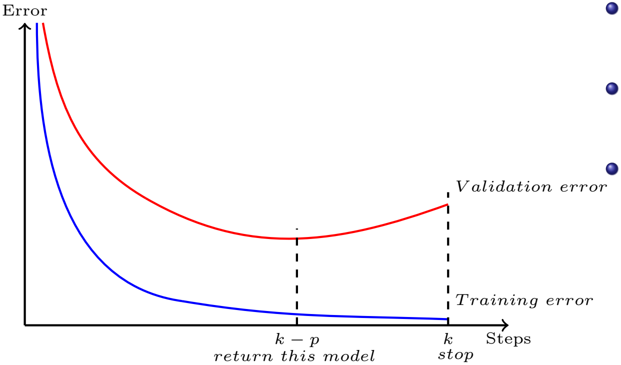
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$
- Basically, stop the training early before it drives the training error to 0 and blows up the validation error



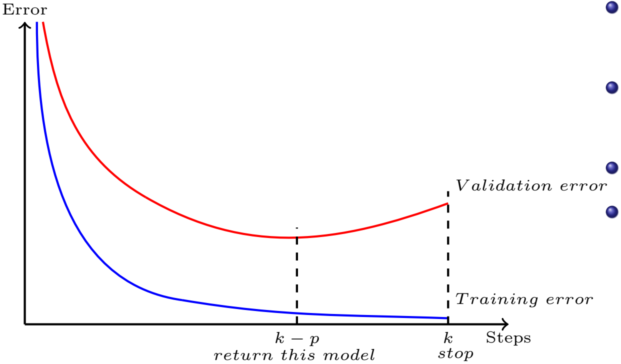
- Very effective and the mostly widely used form of regularization



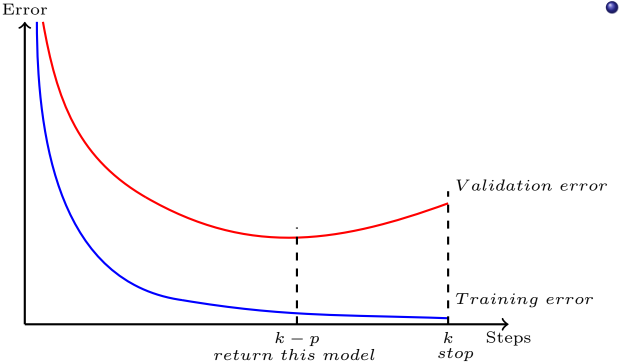
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as L_2)



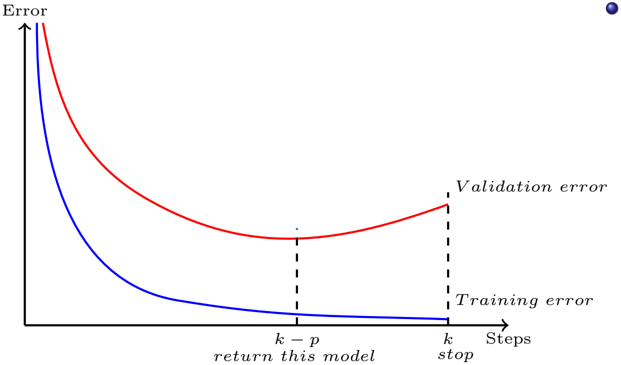
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as L_2)
- How does it act as a regularizer ?



- 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

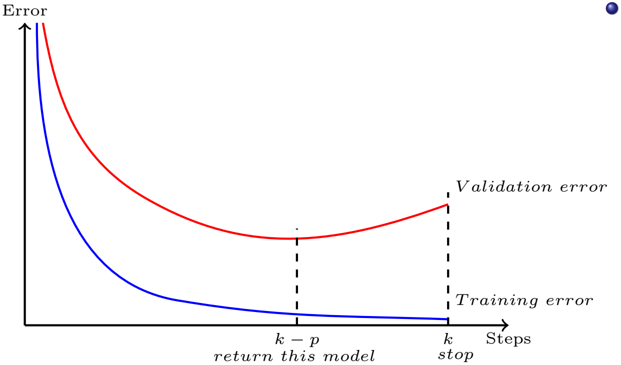


- Recall that the update rule in SGD is :-



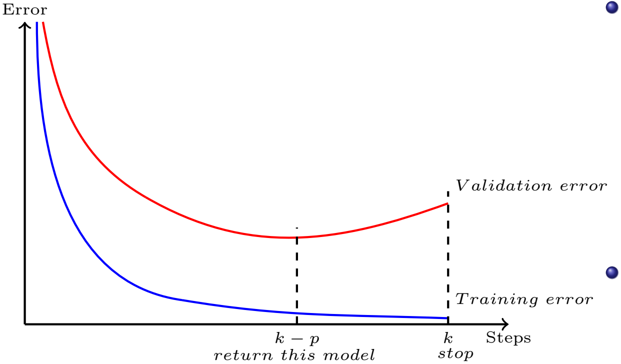
- Recall that the update rule in SGD is :-

$$\omega_{t+1} = \omega_t + \eta\nabla\omega_t$$



- Recall that the update rule in SGD is :-

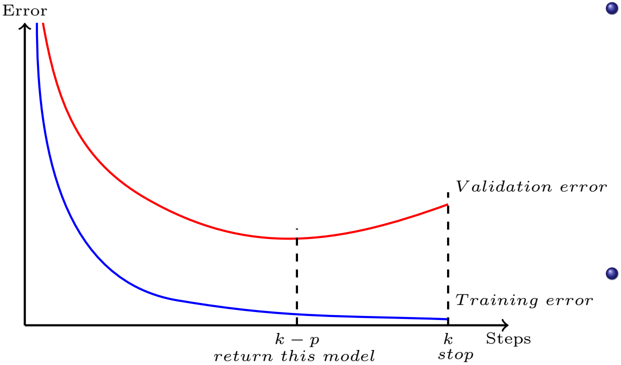
$$\begin{aligned}\omega_{t+1} &= \omega_t + \eta \nabla \omega_t \\ &= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i\end{aligned}$$



- 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

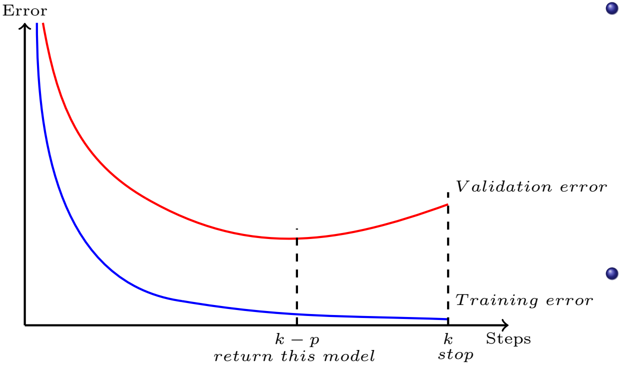


- Recall that the update rule in SGD is :-

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

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

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



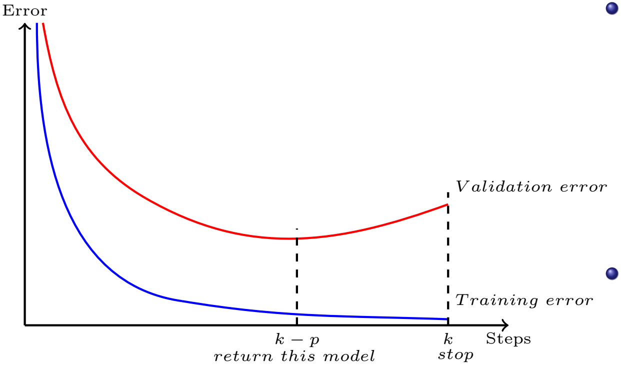
- Recall that the update rule in SGD is :-

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

- 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



- 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
- In other words it controls the space of exploration

We will now see a mathematical analysis of this

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

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

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

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

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

$$= J(\omega^*) + \frac{1}{2}(\omega - \omega^*)^T H(\omega - \omega^*) \qquad [\ \omega^* \textit{ is optimal so } \nabla J(\omega^*) \textit{ is } 0\ ]$$

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

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

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

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

Now the SGD update rule is:

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

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

Now the SGD update rule is:

$$\omega_t = \omega_{t-1} + \eta \nabla J(\omega_{t-1})$$

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

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

Now the SGD update rule is:

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

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

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

Now the SGD update rule is:

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

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

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

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

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

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

- 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

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

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

- 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

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

- Let us see the derivation

- To prove: The below two equations are equivalent

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

- To prove: The below two equations are equivalent

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

- Proof by induction:

- To prove: The below two equations are equivalent

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

- Proof by induction:
- Base case: $t = 1$ and ω_0=0:

- To prove: The below two equations are equivalent

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

- Proof by induction:
- Base case: $t = 1$ and ω_0=0:
- ω_1 according to the first equation:

$$\begin{aligned}\omega_1 &= (I - \eta Q \Lambda Q^T)\omega_0 + \eta Q \Lambda Q^T \omega^* \\ &= \eta Q \Lambda Q^T \omega^*\end{aligned}$$

- To prove: The below two equations are equivalent

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

- Proof by induction:
- Base case: $t = 1$ and ω_0=0:
- ω_1 according to the first equation:

$$\begin{aligned}\omega_1 &= (I - \eta Q\Lambda Q^T)\omega_0 + \eta Q\Lambda Q^T \omega^* \\ &= \eta Q\Lambda Q^T \omega^*\end{aligned}$$

- ω_1 according to the second equation:

$$\begin{aligned}\omega_1 &= Q(I - (I - \eta\Lambda)^1)Q^T w^* \\ &= \eta Q\Lambda Q^T w^*\end{aligned}$$

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Proof that this will hold for $(t+1)^{th}$ step

- Induction step: Let the two equations be equivalent for t^{th} step

$$\begin{aligned}\therefore \omega_t &= (I + \eta Q\Lambda Q^T)\omega_{t-1} - \eta Q\Lambda Q^T\omega^* \\ &= Q[I - (I - \varepsilon\Lambda)^t]Q^T\omega^*\end{aligned}$$

- Proof that this will hold for $(t+1)^{th}$ step

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

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Proof that this will hold for $(t+1)^{th}$ step

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

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Proof that this will hold for $(t+1)^{th}$ step

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

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Proof that this will hold for $(t+1)^{th}$ step

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

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

$$= (I - \eta Q\Lambda Q^T)Q(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^*$$

(Opening this bracket)

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Proof that this will hold for $(t+1)^{th}$ step

$$\omega_{t+1} = (I - \eta Q \Lambda Q^T)\omega_t + \eta Q \Lambda Q^T \omega^*$$
$$(\text{using } \omega_t = Q[I - (I - \varepsilon \Lambda)^t] Q^T \omega^*)$$
$$= (I - \eta Q \Lambda Q^T) Q (I - (I - \eta \Lambda)^t) Q^T \omega^* + \eta Q \Lambda Q^T \omega^*$$
$$(\text{Opening this bracket})$$
$$= {\color{red}I} Q (I - (I - \eta \Lambda)^t) Q^T \omega^* - {\color{red}\eta Q \Lambda Q^T} Q (I - (I - \eta \Lambda)^t) Q^T \omega^* + \eta Q \Lambda Q^T \omega^*$$

- Induction step: Let the two equations be equivalent for t^{th} step

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

- Proof that this will hold for $(t+1)^{th}$ step

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

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

$$= (I - \eta Q\Lambda Q^T)Q(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^*$$

(Opening this bracket)

$$= IQ(I - (I - \eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^*$$
$$= Q(I - (I - \eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^*$$

- Continuing

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

- Continuing

$$\begin{aligned}\omega_{t+1} &= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\ &= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* (\because Q^TQ = I)\end{aligned}$$

- Continuing

$$
\begin{aligned}
\omega_{t+1} &= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* (\because Q^TQ = I) \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q\left[(I-(I-\eta\Lambda)^t) - \eta\Lambda(I-(I-\eta\Lambda)^t) + \eta\Lambda\right]Q^T\omega^*
\end{aligned}
$$

- Continuing

$$
\begin{aligned}
\omega_{t+1} &= Q(I - (I - \eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q(I - (I - \eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* (\because Q^T Q = I) \\
&= Q(I - (I - \eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I - (I - \eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q\big[(I - (I - \eta\Lambda)^t) - \eta\Lambda(I - (I - \eta\Lambda)^t) + \eta\Lambda\big]Q^T\omega^* \\
&= Q\big[I - (I - \eta\Lambda)^t + \eta\Lambda(I - \eta\Lambda)^t\big]Q^T\omega^*
\end{aligned}
$$

- Continuing

$$
\begin{aligned}
\omega_{t+1} &= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* (\because Q^TQ = I) \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q\Big[(I-(I-\eta\Lambda)^t) - \eta\Lambda(I-(I-\eta\Lambda)^t) + \eta\Lambda\Big]Q^T\omega^* \\
&= Q\Big[I-(I-\eta\Lambda)^t + \eta\Lambda(I-\eta\Lambda)^t\Big]Q^T\omega^* \\
&= Q\Big[I-(I-\eta\Lambda)^t(I-\eta\Lambda)\Big]Q^T\omega^*
\end{aligned}
$$

- Continuing

$$
\begin{aligned}
\omega_{t+1} &= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* (\because Q^TQ = I) \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q\Big[(I-(I-\eta\Lambda)^t) - \eta\Lambda(I-(I-\eta\Lambda)^t) + \eta\Lambda\Big]Q^T\omega^* \\
&= Q\Big[I-(I-\eta\Lambda)^t + \eta\Lambda(I-\eta\Lambda)^t\Big]Q^T\omega^* \\
&= Q\Big[I-(I-\eta\Lambda)^t(I-\eta\Lambda)\Big]Q^T\omega^* \\
&= Q(I-(I-\eta\Lambda)^{t+1})Q^T\omega^*
\end{aligned}
$$

- Continuing

$$
\begin{aligned}
\omega_{t+1} &= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda Q^T Q(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* (\because Q^TQ = I) \\
&= Q(I-(I-\eta\Lambda)^t)Q^T\omega^* - \eta Q\Lambda(I-(I-\eta\Lambda)^t)Q^T\omega^* + \eta Q\Lambda Q^T\omega^* \\
&= Q\big[(I-(I-\eta\Lambda)^t) - \eta\Lambda(I-(I-\eta\Lambda)^t) + \eta\Lambda\big]Q^T\omega^* \\
&= Q\big[I-(I-\eta\Lambda)^t + \eta\Lambda(I-\eta\Lambda)^t\big]Q^T\omega^* \\
&= Q\big[I-(I-\eta\Lambda)^t(I-\eta\Lambda)\big]Q^T\omega^* \\
&= Q(I-(I-\eta\Lambda)^{t+1})Q^T\omega^*
\end{aligned}
$$

Hence, proved!

- Coming back...

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

- Coming back...

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

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

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

- Coming back...

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

- 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

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

Things to be remember

- Early stopping only allows t updates to the parameters.

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

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

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

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
- Early stopping will thus effectively shrink the parameters corresponding to less important directions (same as weight decay).