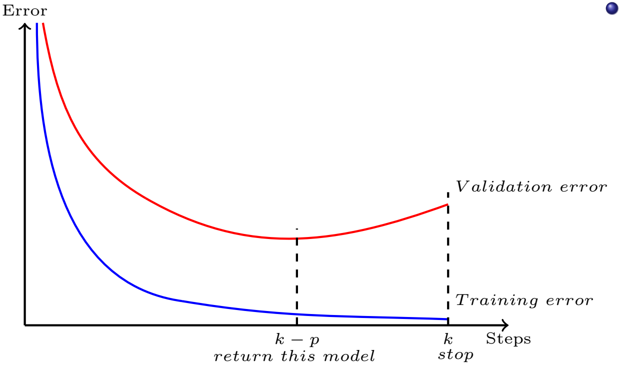
Module 8.9 : Early stopping

Other forms of regularization

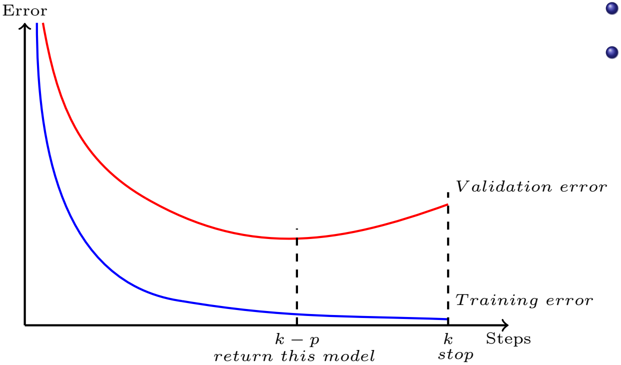
- l_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

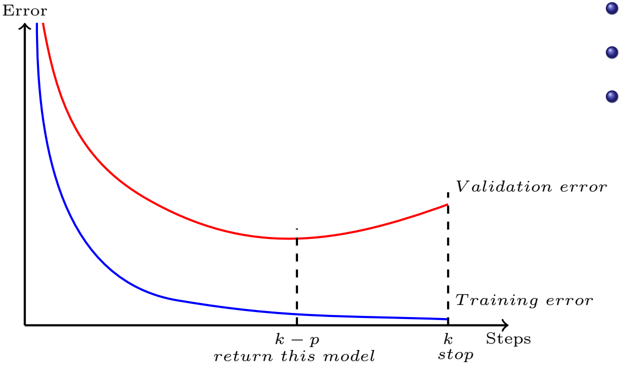
- l_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



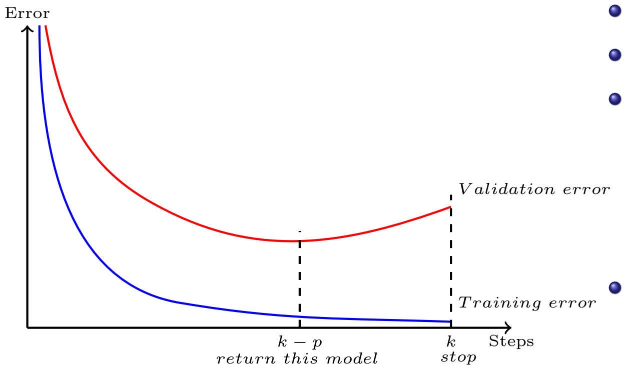
- Track the validation error



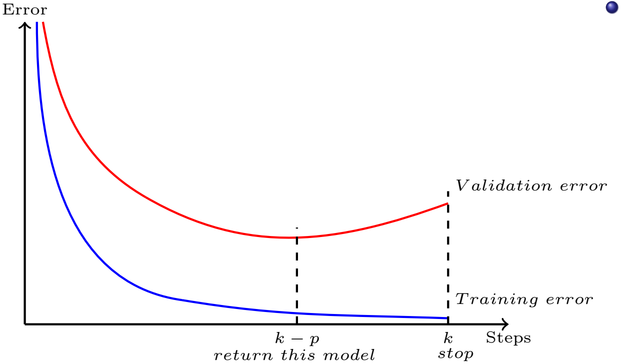
- Track the validation error
- Have a patience parameter p



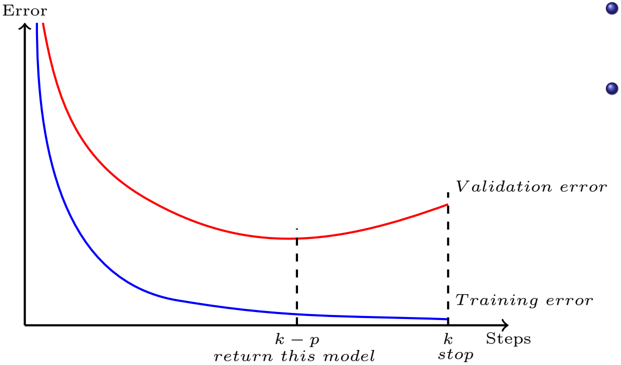
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$



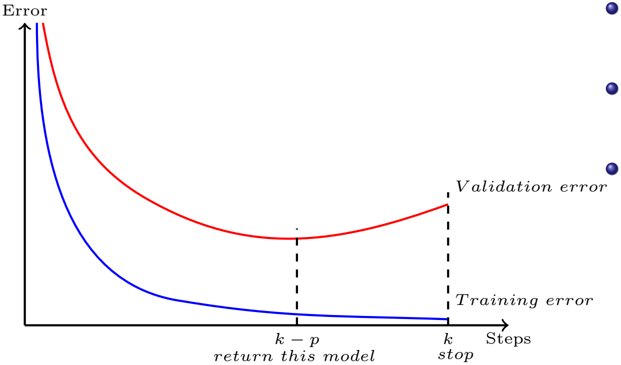
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$
- Basically, stop the training early before it drives the training error to 0 and blows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



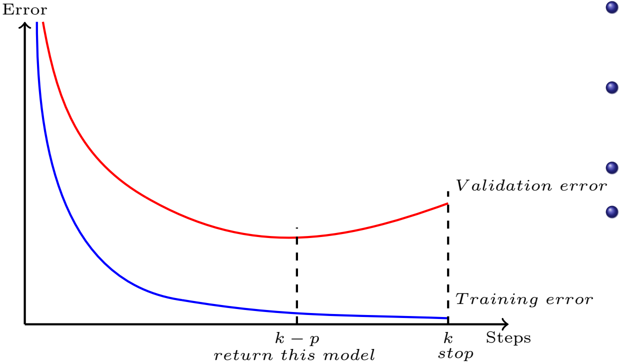
- Very effective and the mostly widely used form of regularization



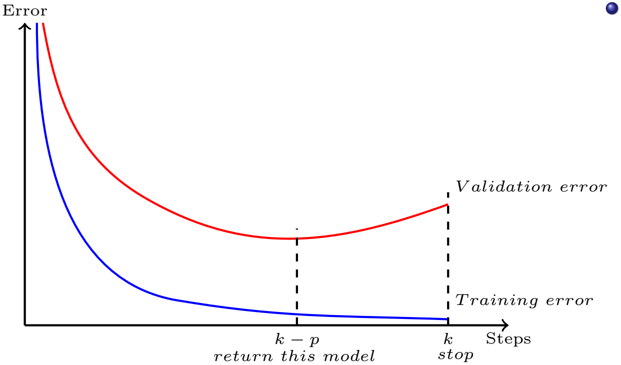
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as l_2)



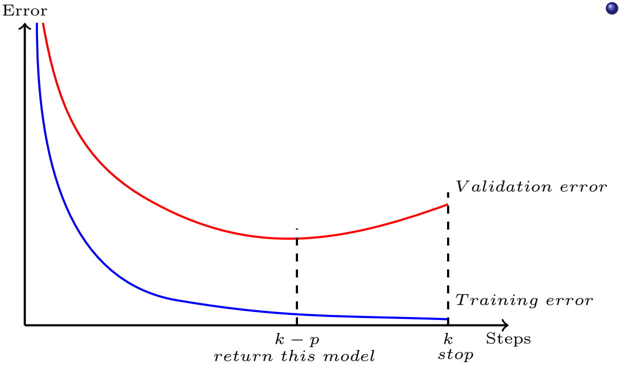
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as l_2)
- How does it act as a regularizer ?



- 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

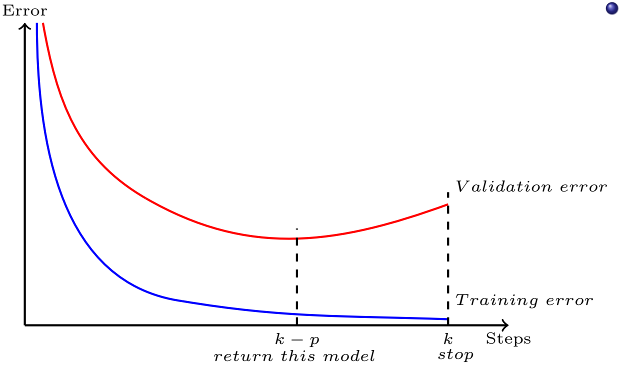


- Recall that the update rule in SGD is



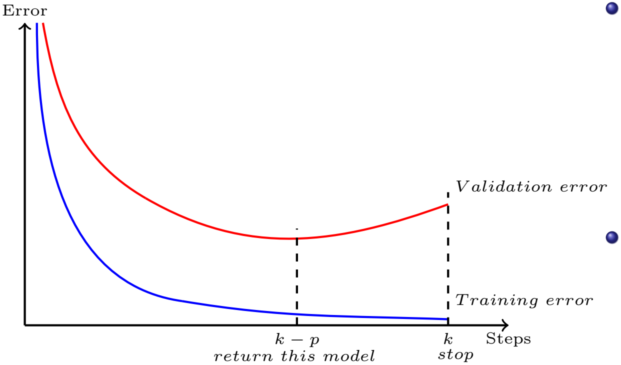
- Recall that the update rule in SGD is

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



- Recall that the update rule in SGD is

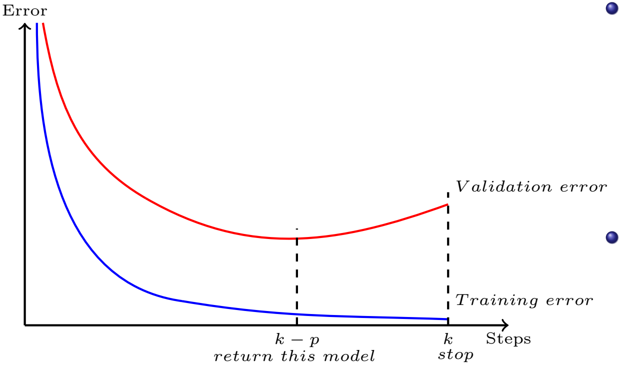
$$\omega_{t+1} = \omega_t + \eta \nabla \omega_t$$
$$= \omega_0 + \eta \sum_{i=1}^{t} \nabla \omega_i$$



- 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

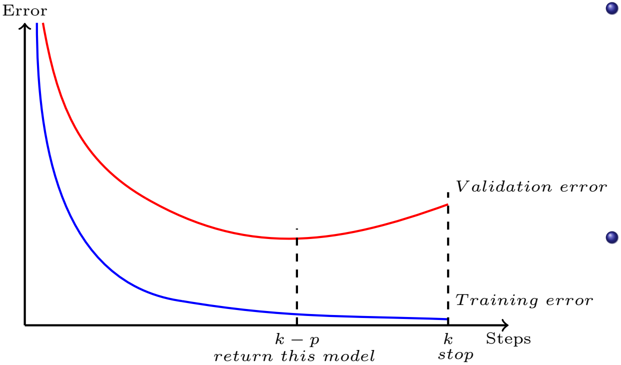


- Recall that the update rule in SGD is

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

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

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



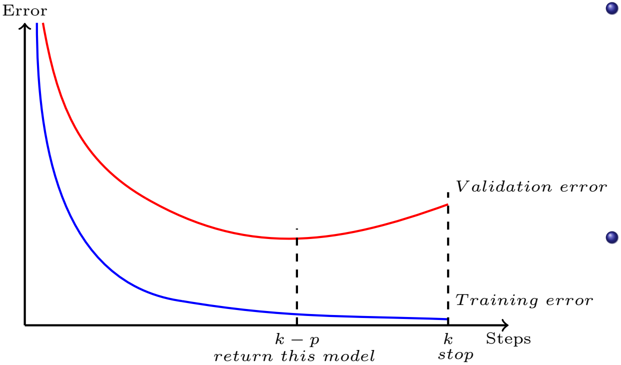
- Recall that the update rule in SGD is

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

- 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 $L(\omega)$ is

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

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

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

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

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

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

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

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

Now the SGD update rule is:

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

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

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

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

Now the SGD update rule is:

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

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

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

Now the SGD update rule is:

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

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

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

Now the SGD update rule is:

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

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

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

- 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 (See Appendix)

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

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

- 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 (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

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

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

- 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

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

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).