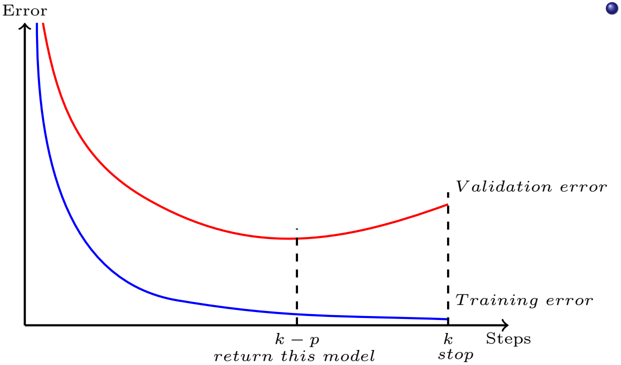
Module 8.9 : Early stopping

Other forms of regularization

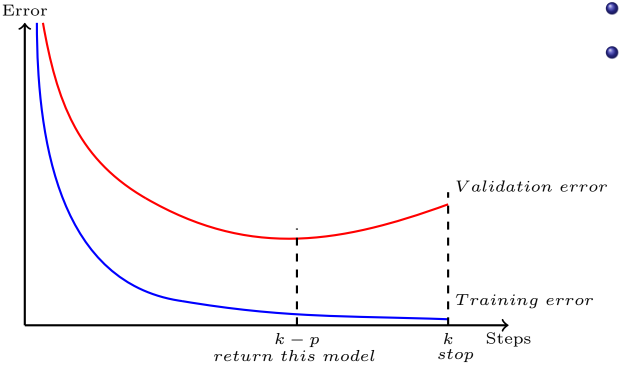
- l_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout

Other forms of regularization

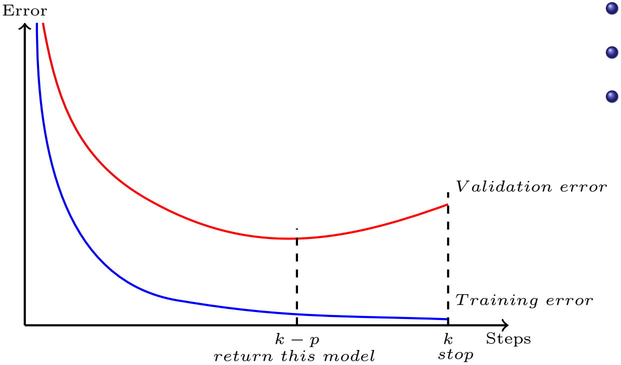
- l_2 regularization
- Dataset augmentation
- Parameter Sharing and tying
- Adding Noise to the inputs
- Adding Noise to the outputs
- Early stopping
- Ensemble methods
- Dropout



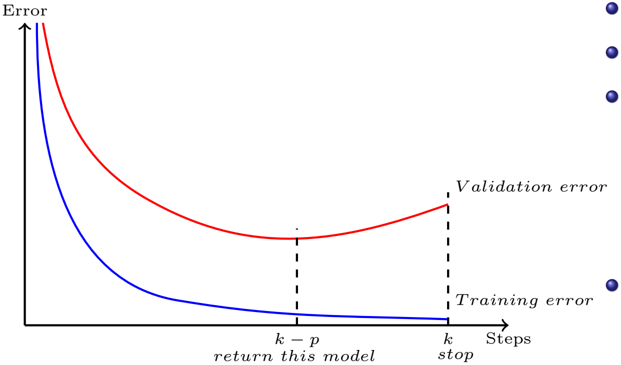
- Track the validation error



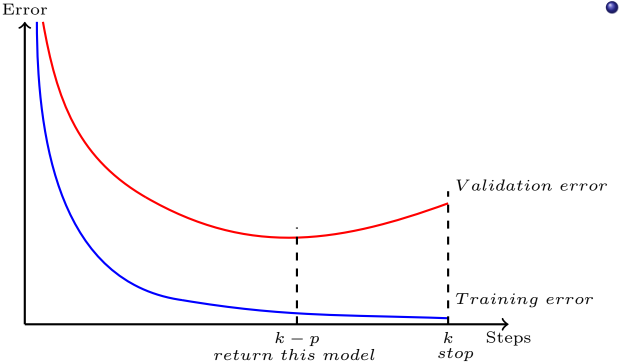
- Track the validation error
- Have a patience parameter p



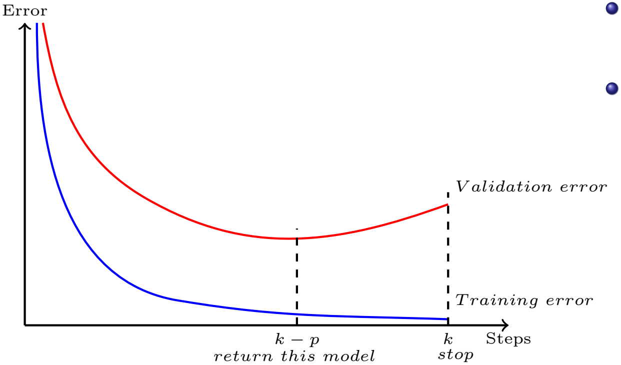
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$



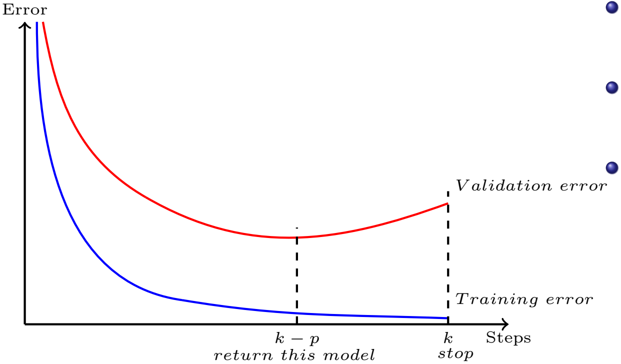
- Track the validation error
- Have a patience parameter p
- If you are at step k and there was no improvement in validation error in the previous p steps then stop training and return the model stored at step $k - p$
- Basically, stop the training early before it drives the training error to 0 and blows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



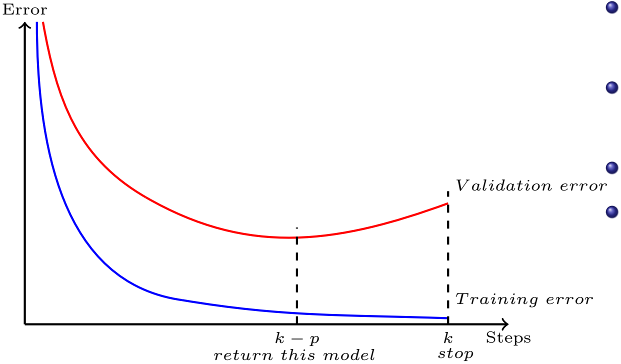
- Very effective and the mostly widely used form of regularization



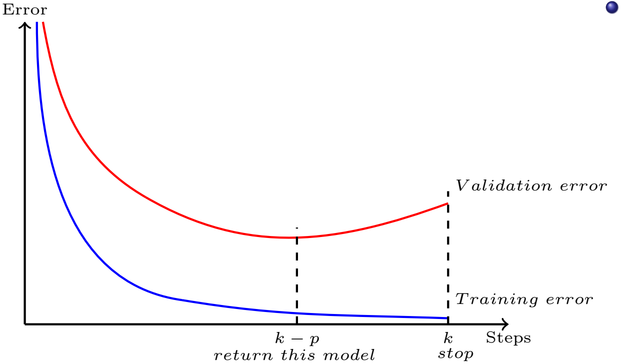
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as l_2)



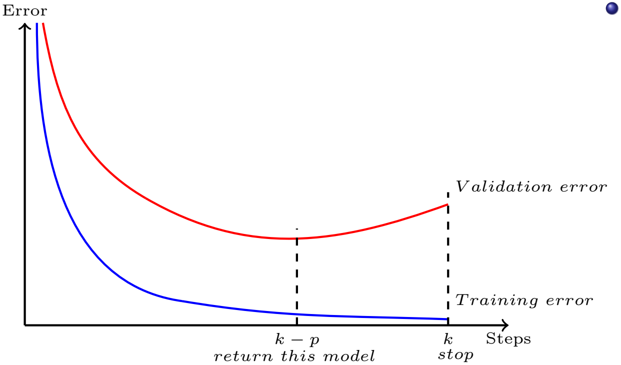
- Very effective and the mostly widely used form of regularization
- Can be used even with other regularizers (such as l_2)
- How does it act as a regularizer ?



- 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

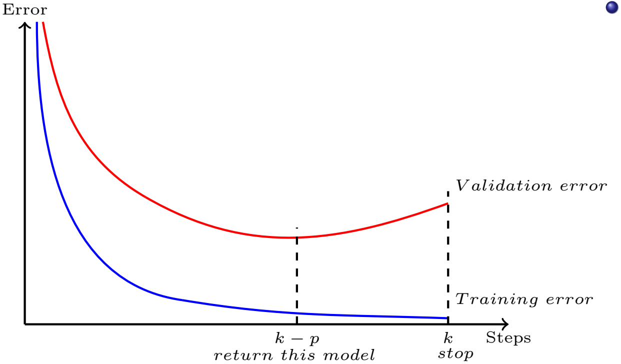


- Recall that the update rule in SGD is



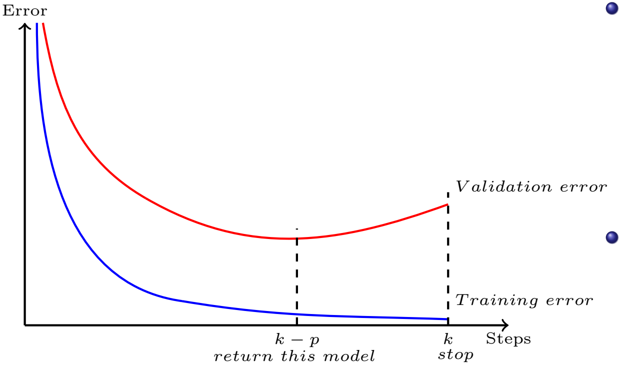
- Recall that the update rule in SGD is

$$w_{t+1} = w_t - \eta \nabla w_t$$



- Recall that the update rule in SGD is

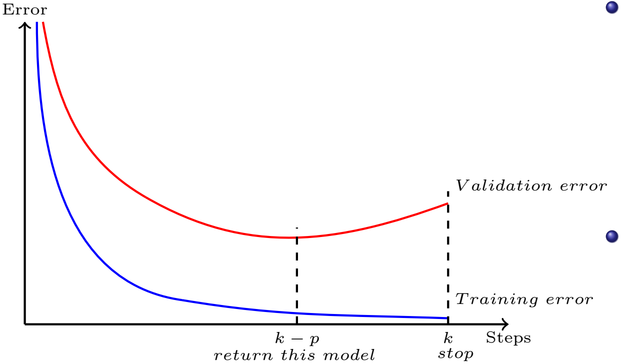
$$w_{t+1} = w_t - \eta \nabla w_t$$
$$= w_0 - \eta \sum_{i=1}^{t} \nabla w_i$$



- Recall that the update rule in SGD is

$$w_{t+1} = w_t - \eta \nabla w_t$$
$$= w_0 - \eta \sum_{i=1}^{t} \nabla w_i$$

- Let τ be the maximum value of ∇w_i then

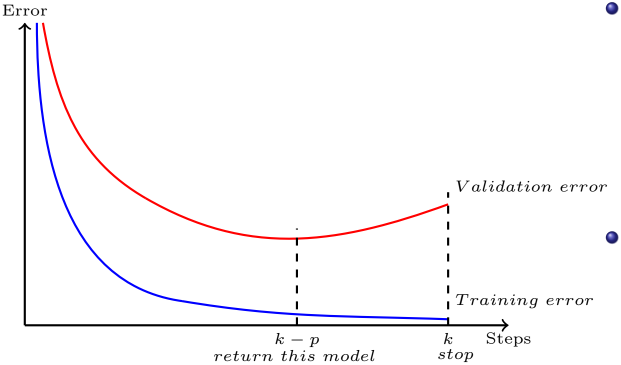


- Recall that the update rule in SGD is

$$w_{t+1} = w_t - \eta \nabla w_t$$
$$= w_0 - \eta \sum_{i=1}^{t} \nabla w_i$$

- Let τ be the maximum value of ∇w_i then

$$|w_{t+1} - w_0| \leq \eta t |\tau|$$



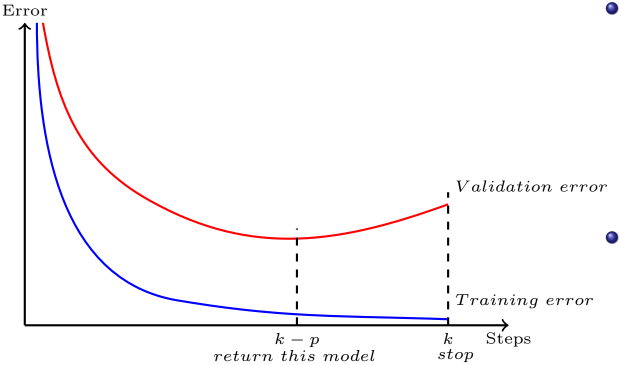
- Recall that the update rule in SGD is

$$w_{t+1} = w_t - \eta \nabla w_t$$
$$= w_0 - \eta \sum_{i=1}^{t} \nabla w_i$$

- Let τ be the maximum value of ∇w_i then

$$|w_{t+1} - w_0| \leq \eta t |\tau|$$

- Thus, t controls how far w_t can go from the initial w_0



- Recall that the update rule in SGD is

$$w_{t+1} = w_t - \eta \nabla w_t$$
$$= w_0 - \eta \sum_{i=1}^{t} \nabla w_i$$

- Let τ be the maximum value of ∇w_i then

$$|w_{t+1} - w_0| \leq \eta t |\tau|$$

- Thus, t controls how far w_t can go from the initial w_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 $\mathscr{L}(w)$ is

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

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

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

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

$$\nabla(\mathscr{L}(w)) = H(w - w^*)$$

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

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

$$\nabla(\mathscr{L}(w)) = H(w - w^*)$$

Now the SGD update rule is:

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

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

$$\nabla(\mathscr{L}(w)) = H(w - w^*)$$

Now the SGD update rule is:

$$w_t = w_{t-1} - \eta \nabla \mathscr{L}(w_{t-1})$$

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

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

$$\nabla(\mathscr{L}(w)) = H(w - w^*)$$

Now the SGD update rule is:

$$w_t = w_{t-1} - \eta \nabla \mathscr{L}(w_{t-1})$$

$$= w_{t-1} - \eta H (w_{t-1} - w^*)$$

- Recall that the Taylor series approximation for $\mathscr{L}(w)$ is

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

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

$$\nabla(\mathscr{L}(w)) = H(w - w^*)$$

Now the SGD update rule is:

$$w_t = w_{t-1} - \eta \nabla \mathscr{L}(w_{t-1})$$
$$= w_{t-1} - \eta H (w_{t-1} - w^*)$$
$$= (I - \eta H) w_{t-1} + \eta H w^*$$

$$w_t = (I - \eta H) w_{t-1} + \eta H w^*$$

$$w_t = (I - \eta H)w_{t-1} + \eta H w^*$$

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

$$w_t = (I - \eta Q\Lambda Q^T)w_{t-1} + \eta Q\Lambda Q^T w^*$$

$$w_t = (I - \eta H)w_{t-1} + \eta H w^*$$

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

$$w_t = (I - \eta Q\Lambda Q^T)w_{t-1} + \eta Q\Lambda Q^T w^*$$

- If we start with $w_0 = 0$ then we can show that (See Appendix)

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

$$w_t = (I - \eta H)w_{t-1} + \eta H w^*$$

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

$$w_t = (I - \eta Q\Lambda Q^T)w_{t-1} + \eta Q\Lambda Q^T w^*$$

- If we start with $w_0 = 0$ then we can show that (See Appendix)

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

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

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

$$w_t = (I - \eta H)w_{t-1} + \eta H w^*$$

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

$$w_t = (I - \eta Q\Lambda Q^T)w_{t-1} + \eta Q\Lambda Q^T w^*$$

- If we start with $w_0 = 0$ then we can show that (See Appendix)

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

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

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

- We observe that $w_t = \tilde{w}$, 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 w corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial w}$ will be large

Things to be remember

- Early stopping only allows t updates to the parameters.
- If a parameter w corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial w}$ will be large

Things to be remember

- Early stopping only allows t updates to the parameters.
- If a parameter w corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial w}$ will be large
- However if a parameter is not important ($\frac{\partial \mathscr{L}(\theta)}{\partial w}$ 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 w corresponds to a dimension which is important for the loss $\mathscr{L}(\theta)$ then $\frac{\partial \mathscr{L}(\theta)}{\partial w}$ will be large
- However if a parameter is not important ($\frac{\partial \mathscr{L}(\theta)}{\partial w}$ 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).