Module 8.4 : L_2 regularization

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

- For L_2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

- For L_2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

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

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- For L_2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

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

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Update rule:

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

- For L_2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

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

$$\nabla\widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Update rule:

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

- Requires a very small modification to the code

- For L_2 regularization we have,

$$\widetilde{J}(w) = J(w) + \frac{\alpha}{2}\|w\|^2$$

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

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Update rule:

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

- Requires a very small modification to the code
- Let us see the geometric interpretation of this

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

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

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

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$
\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ )
\end{aligned}
$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$\begin{aligned} J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\ &= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \end{aligned}$$

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

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \\
\nabla J(w) &= \nabla J(w^*) + H(w - w^*) \\
&= H(w - w^*)
\end{aligned}$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \\
\nabla J(w) &= \nabla J(w^*) + H(w - w^*) \\
&= H(w - w^*)
\end{aligned}$$

- Now,

$$\nabla \widetilde{J}(w) = \nabla J(w) + \alpha w$$

- Assume w^* is the optimal solution for $J(w)$ [not $\widetilde{J}(w)$] i.e. the solution in the absence of regularization (w^* optimal $\rightarrow \nabla J(w^*) = 0$)
- Using Taylor series approximation (upto 2^{nd} order)

$$
\begin{aligned}
J(w) &= J(w^*) + (w - w^*)^T \nabla J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \\
&= J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*) \qquad (\because \nabla J(w^*) = 0\ ) \\
\nabla J(w) &= \nabla J(w^*) + H(w - w^*) \\
&= H(w - w^*)
\end{aligned}
$$

- Now,

$$
\begin{aligned}
\nabla \widetilde{J}(w) &= \nabla J(w) + \alpha w \\
&= H(w - w^*) + \alpha w
\end{aligned}
$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha \widetilde{w} = 0$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha \widetilde{w} = 0$$

$$\therefore (H + \alpha \mathbb{I})\widetilde{w} = H w^*$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha \widetilde{w} = 0$$
$$\therefore (H + \alpha \mathbb{I})\widetilde{w} = Hw^*$$
$$\therefore \widetilde{w} = (H + \alpha \mathbb{I})^{-1} Hw^*$$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha\widetilde{w} = 0$$
$$\therefore (H + \alpha\mathbb{I})\widetilde{w} = Hw^*$$
$$\therefore \widetilde{w} = (H + \alpha\mathbb{I})^{-1}Hw^*$$

- Notice that if $\alpha \to 0$ then $\widetilde{w} \to w^*$ [no regularization]

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha\widetilde{w} = 0$$
$$\therefore (H + \alpha\mathbb{I})\widetilde{w} = Hw^*$$
$$\therefore \widetilde{w} = (H + \alpha\mathbb{I})^{-1}Hw^*$$

- Notice that if $\alpha \to 0$ then $\widetilde{w} \to w^*$ [no regularization]
- But we are interested in the case when $\alpha \neq 0$

- Let $\widetilde{w}$ be the optimal solution for $\widetilde{J}(w)$ [i.e regularized loss]

$$\because \nabla \widetilde{J}(\widetilde{w}) = 0$$

$$H(\widetilde{w} - w^*) + \alpha\widetilde{w} = 0$$
$$\therefore (H + \alpha\mathbb{I})\widetilde{w} = Hw^*$$
$$\therefore \widetilde{w} = (H + \alpha\mathbb{I})^{-1} Hw^*$$

- Notice that if $\alpha \to 0$ then $\widetilde{w} \to w^*$ [no regularization]
- But we are interested in the case when $\alpha \neq 0$
- Let us analyse the case when $\alpha \neq 0$

- If H is symmetric Positive Semi Definite

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\widetilde{w} = (H + \alpha\mathbb{I})^{-1} H w^*$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}\widetilde{w} &= (H + \alpha\mathbb{I})^{-1}Hw^* \\ &= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1}Q\Lambda Q^T w^*\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

$$\begin{aligned}
\widetilde{w} &= (H + \alpha\mathbb{I})^{-1} H w^* \\
&= (Q\Lambda Q^T + \alpha\mathbb{I})^{-1} Q\Lambda Q^T w^* \\
&= (Q\Lambda Q^T + \alpha Q\mathbb{I}Q^T)^{-1} Q\Lambda Q^T w^*
\end{aligned}$$

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

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

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

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

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

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

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

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

- If H is symmetric Positive Semi Definite

$$H = Q\Lambda Q^T \qquad [Q \text{ is orthogonal, } QQ^T = Q^TQ = \mathbb{I}]$$

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

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

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

- So what is happening here?

$$\begin{aligned}\widetilde{w} &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \\ &= QDQ^T w^*\end{aligned}$$

- So what is happening here?
- w^* first gets rotated by Q^T to give $Q^T w^*$

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

- 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^*

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

- 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

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

$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

- 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

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

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

- 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

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

- 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

$$\begin{aligned}\widetilde{w} &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \\ &= QDQ^T w^*\end{aligned}$$

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

- 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

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

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

- 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

$$
\begin{aligned}
\widetilde{w} &= Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^* \\
&= Q D Q^T w^*
\end{aligned}
$$

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

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

- 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

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

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

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

$$(\Lambda + \alpha\mathbb{I})^{-1}\Lambda = \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

- 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

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

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

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

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

- 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

$$
\begin{aligned}
\widetilde{w} &= Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^* \\
&= Q D Q^T w^*
\end{aligned}
$$

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

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

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

- 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

$$
\begin{aligned}
\widetilde{w} &= Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^* \\
&= Q D Q^T w^*
\end{aligned}
$$

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

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

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

- 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

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

- 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

$$
\begin{aligned}
\widetilde{w} &= Q(\Lambda + \alpha \mathbb{I})^{-1} \Lambda Q^T w^* \\
&= Q D Q^T w^*
\end{aligned}
$$

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

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

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

- 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

$$\begin{aligned}\widetilde{w} &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \\ &= QDQ^T w^*\end{aligned}$$

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

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

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

- 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?

$$
\begin{aligned}
\widetilde{w} &= Q(\Lambda + \alpha\mathbb{I})^{-1}\Lambda Q^T w^* \\
&= QDQ^T w^*
\end{aligned}
$$

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

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

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

- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q

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

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

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

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

- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q
- if $\lambda_i >> \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 1$

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

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

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

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

- Each element i of $Q^T w^*$ gets scaled by $\frac{\lambda_i}{\lambda_i+\alpha}$ before it is rotated back by Q
- if $\lambda_i >> \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 1$
- if $\lambda_i << \alpha$ then $\frac{\lambda_i}{\lambda_i+\alpha} = 0$

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

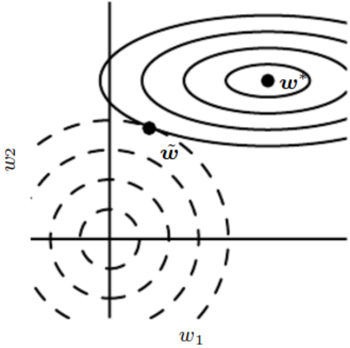
$$(\Lambda + \alpha \mathbb{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\alpha} & & & \\ & \frac{1}{\lambda_2+\alpha} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_n+\alpha} \end{bmatrix}$$

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

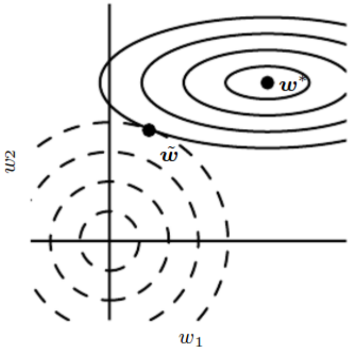
$$(\Lambda + \alpha \mathbb{I})^{-1} \Lambda = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+\alpha} & & & \\ & \frac{\lambda_2}{\lambda_2+\alpha} & & \\ & & \ddots & \\ & & & \frac{\lambda_n}{\lambda_n+\alpha} \end{bmatrix}$$

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

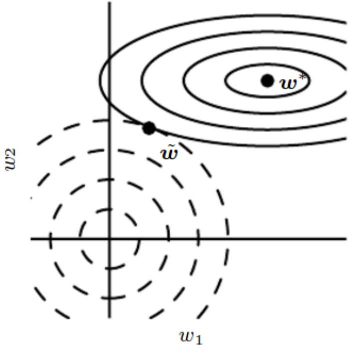
$$\text{Effective parameters} = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \alpha} < n$$



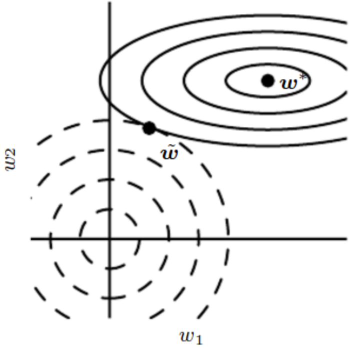
w^*
$\tilde{w}$
w_2
w_1



- The weight vector(w^*) is getting rotated to ($\tilde{w}$)



- The weight vector(w^*) is getting rotated to ($\tilde{w}$)
- All of its elements are shrinking but some are shrinking more than the others



- The weight vector(w^*) is getting rotated to ($\tilde{w}$)
- All of its elements are shrinking but some are shrinking more than the others
- This ensures that only important features are given high weights