Tayler Series Approximation

$L(\boldsymbol{\theta})$ around $\boldsymbol{\theta} = \boldsymbol{\theta}'$ can be approximated below

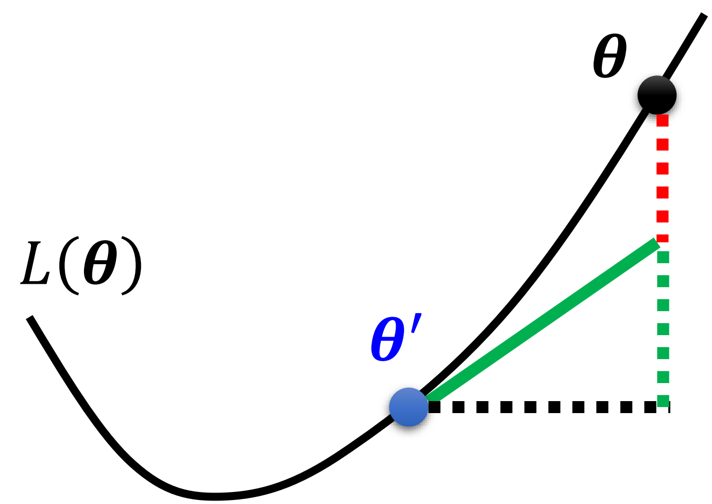
$$L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}') + \boxed{(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{g}} + \boxed{\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T H (\boldsymbol{\theta} - \boldsymbol{\theta}')}$$

Gradient $\boldsymbol{g}$ is a ***vector***

$$\boldsymbol{g} = \nabla L(\boldsymbol{\theta}') \qquad \boldsymbol{g}_i = \frac{\partial L(\boldsymbol{\theta}')}{\partial \boldsymbol{\theta}_i}$$

Hessian H is a ***matrix***

$$H_{ij} = \frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} L(\boldsymbol{\theta}')$$



Source of image:http://www.offconvex.org/2016/03/22/saddlepoints/

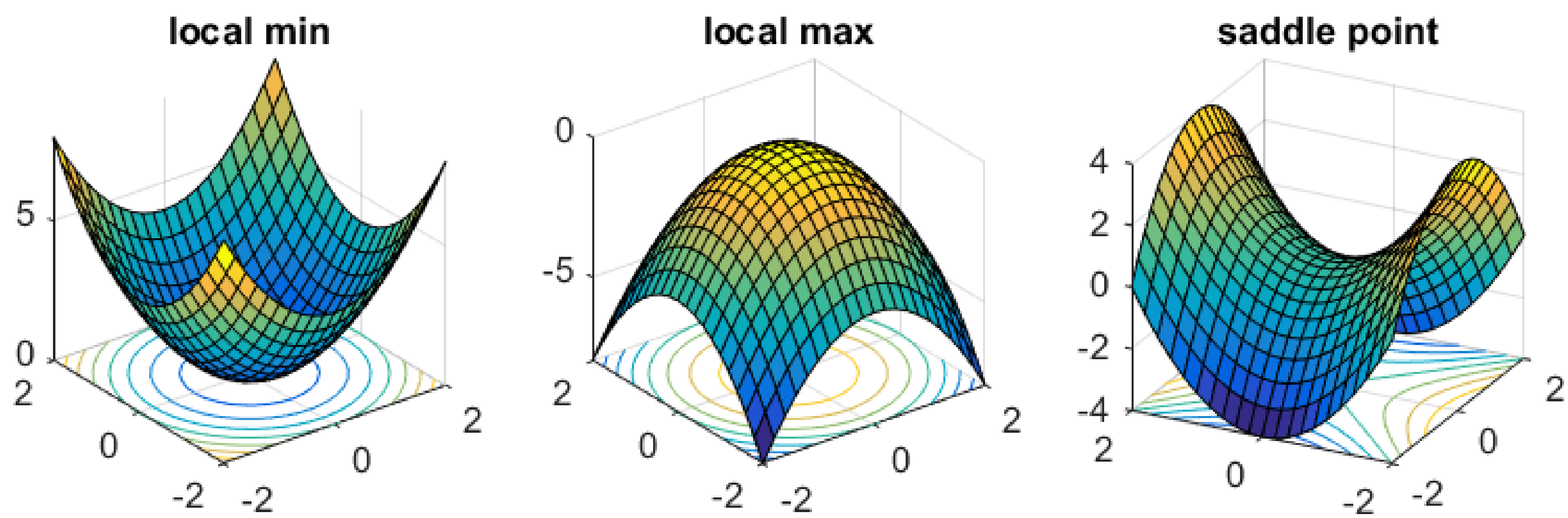
Hessian

$L(\boldsymbol{\theta})$ around $\boldsymbol{\theta} = \boldsymbol{\theta}'$ can be approximated below

$$L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}') + \boxed{\cancel{(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{g}}} + \boxed{\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T H (\boldsymbol{\theta} - \boldsymbol{\theta}')}$$

At critical point

telling the properties of critical points



Hessian

At critical point:

$$\boldsymbol{v}^T H \boldsymbol{v}$$

$$L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}') + \boxed{\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T H (\boldsymbol{\theta} - \boldsymbol{\theta}')}$$

For all $\boldsymbol{v}$

$\boldsymbol{v}^T H \boldsymbol{v} > 0$ ➡ Around $\boldsymbol{\theta}'$: $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}')$ ➡ **Local minima**

= H is positive definite = All eigen values are positive. ⬆

For all $\boldsymbol{v}$

$\boldsymbol{v}^T H \boldsymbol{v} < 0$ ➡ Around $\boldsymbol{\theta}'$: $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}')$ ➡ **Local maxima**

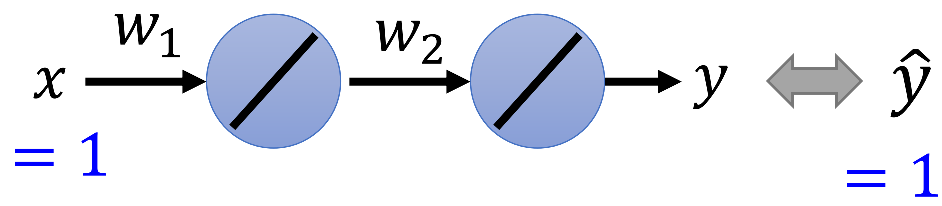
= H is negative definite = All eigen values are negative. ⬆

Sometimes $\boldsymbol{v}^T H \boldsymbol{v} > 0$, sometimes $\boldsymbol{v}^T H \boldsymbol{v} < 0$ ➡ **Saddle point**

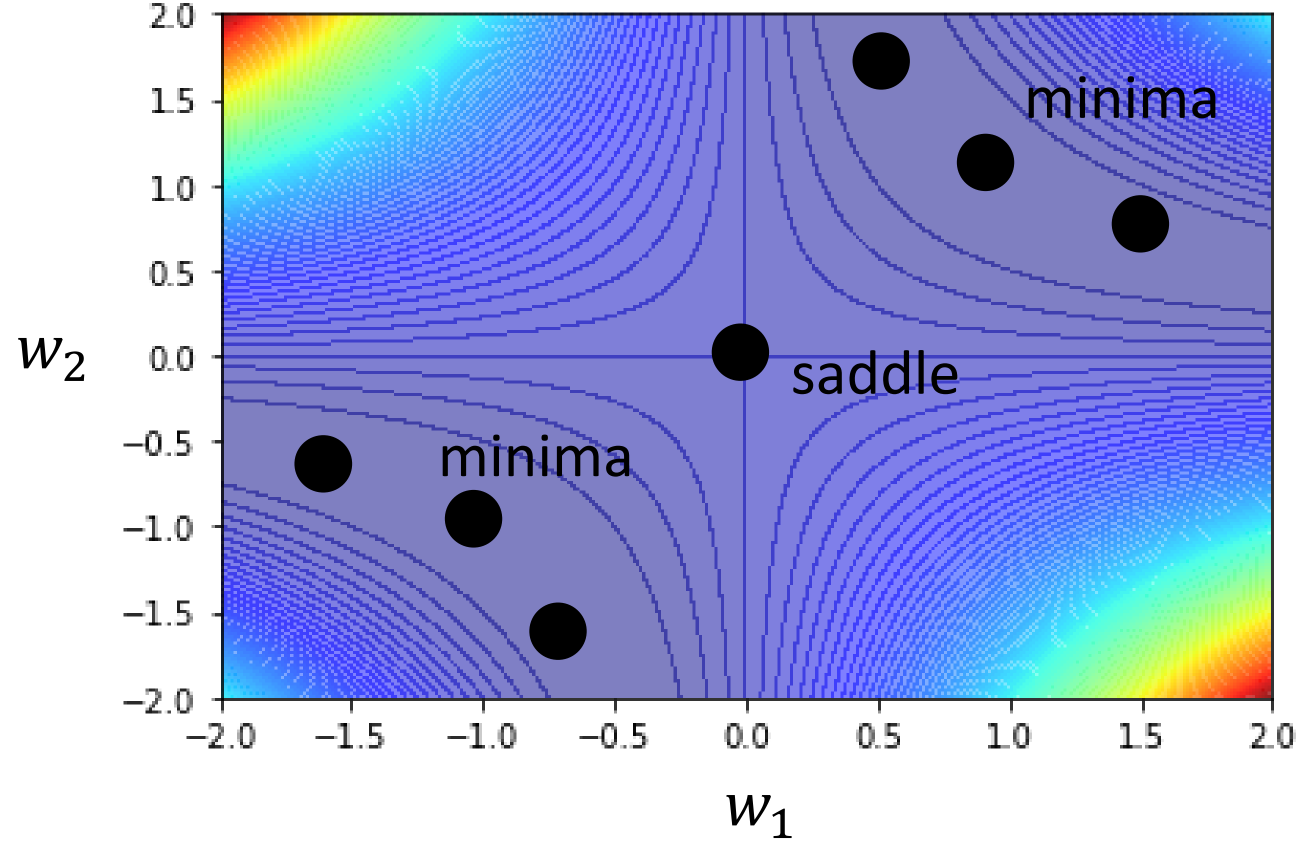
Some eigen values are positive, and some are negative. ⬆

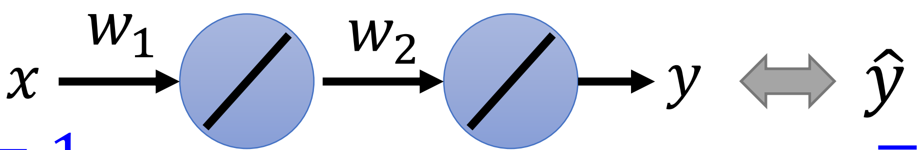
Example

$$y = w_1 w_2 x$$

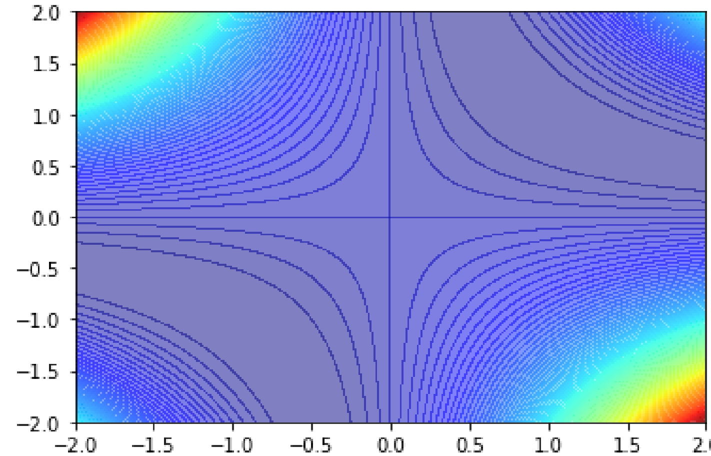


Error Surface





$= 1$ $\qquad$ $= 1$



$$L = (\hat{y} - w_1 w_2 x)^2 \;= (1 - w_1 w_2)^2$$

$$\boxed{\frac{\partial L}{\partial w_1}} = 2(1 - w_1 w_2)(-w_2) \quad = 0$$

$$\boxed{\frac{\partial L}{\partial w_2}} = 2(1 - w_1 w_2)(-w_1) \quad = 0$$

g

Critical point: $w_1 = 0, w_2 = 0$

$$H = \begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix} \quad \lambda_1 = 2, \lambda_2 = -2$$

Saddle point

H

$$\boxed{\frac{\partial^2 L}{\partial {w_1}^2}} = 2(-w_2)(-w_2) \quad = 0$$

$$\boxed{\frac{\partial^2 L}{\partial w_1 \partial w_2}} = -2 + 4 w_1 w_2 \quad = -2$$

$$\boxed{\frac{\partial^2 L}{\partial w_2 \partial w_1}} = -2 + 4 w_1 w_2 \quad = -2$$

$$\boxed{\frac{\partial^2 L}{\partial {w_2}^2}} = 2(-w_1)(-w_1) \quad = 0$$

Don't afraid of saddle point?

At critical point: $L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}') + \overset{\boldsymbol{v}^T H \boldsymbol{v}}{\boxed{\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T H (\boldsymbol{\theta} - \boldsymbol{\theta}')}}$

Sometimes $\boldsymbol{v}^T H \boldsymbol{v} > 0$, sometimes $\boldsymbol{v}^T H \boldsymbol{v} < 0$ ➡ Saddle point

H may tell us parameter update direction!

$\boldsymbol{u}$ is an eigen vector of H

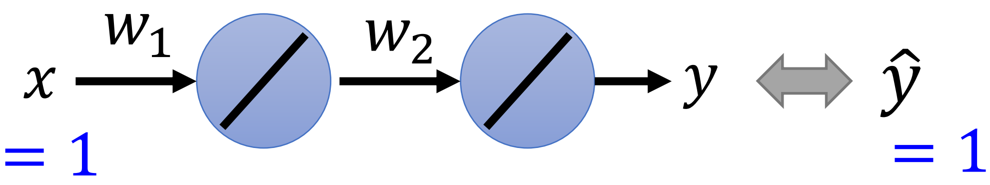
λ is the eigen value of $\boldsymbol{u}$

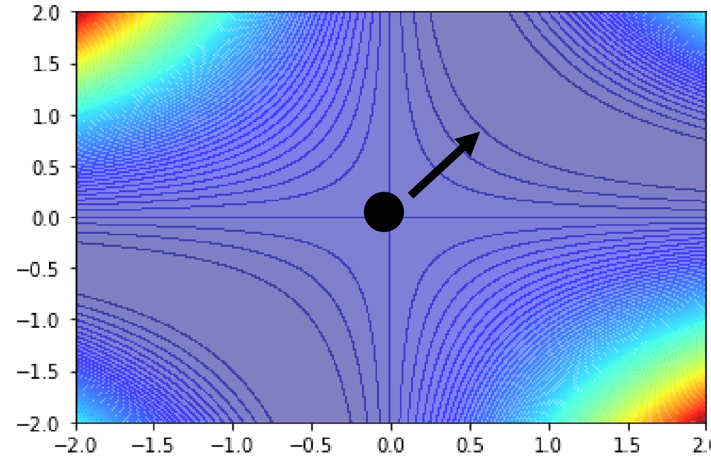
$\lambda < 0$

➡ $\underset{<0}{\boldsymbol{u}^T H \boldsymbol{u}} = \boldsymbol{u}^T(\lambda \boldsymbol{u}) = \underset{<0}{\lambda \|\boldsymbol{u}\|^2}$

$$L(\boldsymbol{\theta}) \approx L(\boldsymbol{\theta}') + \frac{1}{2}\overset{\boldsymbol{u}}{(\boldsymbol{\theta} - \boldsymbol{\theta}')}^T H \overset{\boldsymbol{u}}{(\boldsymbol{\theta} - \boldsymbol{\theta}')} \Rightarrow L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}')$$

$\boldsymbol{\theta} - \boldsymbol{\theta}' = \boldsymbol{u}$ $\quad$ $\boldsymbol{\theta} = \boldsymbol{\theta}' + \boldsymbol{u}$ $\quad$ Decrease L





$$L = (\hat{y} - w_1 w_2 x)^2 \;= (1 - w_1 w_2)^2$$

$$\frac{\partial L}{\partial w_1} = 2(1 - w_1 w_2)(-w_2)$$

$$\frac{\partial L}{\partial w_2} = 2(1 - w_1 w_2)(-w_1)$$

Critical point: $w_1 = 0, w_2 = 0$

$$H = \begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix} \quad \lambda_1 = 2, \lambda_2 = -2$$

Saddle point

$\lambda_2 = -2$ Has eigenvector $\boldsymbol{u} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Update the parameter along the direction of $\boldsymbol{u}$

You can escape the saddle point and decrease the loss.

(this method is seldom used in practice)