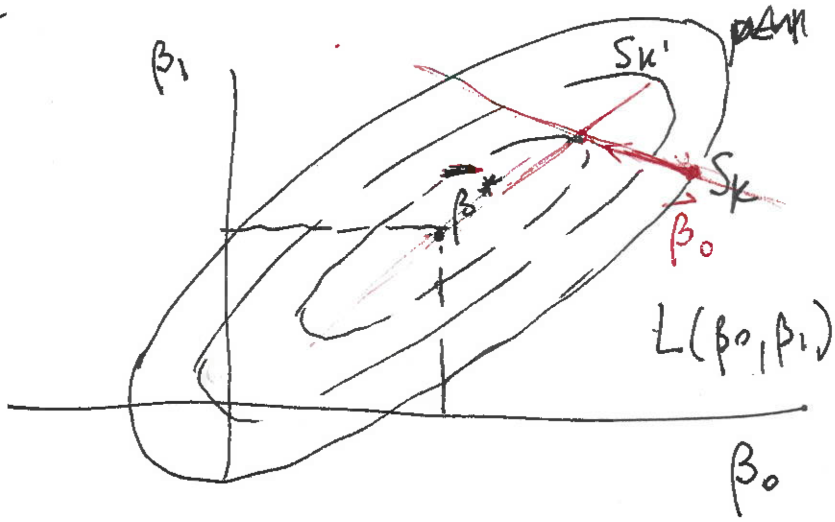
$p=1$



$\vec\beta=(\beta_0,\beta_1)$

$k'<k$

$$S_k=\{\vec\beta \mid L(\vec\beta)=k\}$$

$$L(\vec\beta)=\frac{1}{N}\left[(\vec y - X\vec\beta)^T(\vec y - X\vec\beta)\right] \quad (*)$$

a couple of steps :

$$L(\vec\beta)=L(\vec\beta^*)+\frac{1}{2}(\vec\beta-\vec\beta^*)\underbrace{H}_{\frac{2}{N}(X^TX)}(\vec\beta-\vec\beta^*)$$

$$(*)\ L(\vec\beta)=\frac{1}{N}\left[(\vec y - X\vec\beta^* + X\vec\beta^* - X\vec\beta)^T(\vec y - X\vec\beta^* + X\vec\beta^* - X\vec\beta)\right]=$$

$$=\frac{1}{N}\Big[(\vec y - X\vec\beta^*)^T(\vec y - X\vec\beta^*) + \cancel{(\vec y - X\vec\beta^*)^T X(\vec\beta^*-\vec\beta)} +$$

$$+\cancel{[X(\vec\beta^*-\vec\beta)]^T(\vec y - X\vec\beta^*)} + (\vec\beta^*-\vec\beta)^T X^T X(\vec\beta^*-\vec\beta)\Big]$$

Normal equations: $X^T\vec y - X^TX\vec\beta^* = 0 \Rightarrow X^T[\vec y - X\vec\beta]=0$

see next page

Convexity $\Rightarrow$ local minimum is always ~~equivalent to~~ a global minimum

---

This loss function can be minimised with gradient methods.

$\nabla_{\beta} L$ marks the direction of maximum change of L

Algorithm: Iteratively follow the direction of $-\nabla_{\vec{\beta}} L$

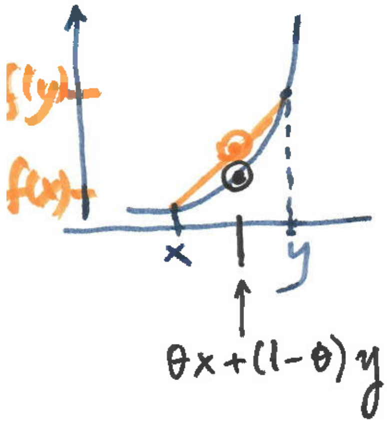
1st order

$$\vec{\beta}_{k+1} = \vec{\beta}_k - \boxed{\eta_k} \vec{\nabla}_{\beta} L(\vec{\beta}_k)$$

- Line search
- Back tracking
- Conjugate gradient

$\leftarrow$ Gradient methods.

Convexity:



$f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$\forall\ \vec{x}, \vec{y} \in \mathbb{R}^n$ and $\theta \in (0,1)$

$$\theta f(\vec{x}) + (1-\theta) f(\vec{y}) \geq f(\theta \vec{x} + (1-\theta)\vec{y})$$

If f is convex, a local minimum is always a global minimum

Let $\vec{x}_{local}, \vec{x}_{global} \in \mathbb{R}^n$, $f(\vec{x}_{global}) < f(\vec{x}_{local})$

By contradiction:

- If $\vec{x}_{local}$ is a local minimum then:

  $f(\vec{x}_{local}) \leq f(\vec{x})$, $\|\vec{x} - \vec{x}_{local}\| < \gamma$

- If f is convex,

  $\exists\ \theta$, $\theta\, \vec{x}_{local} + (1-\theta)\, \vec{x}_{global} = \vec{x}_\theta$

  $\|\vec{x}_\theta - \vec{x}_{local}\| < \gamma$

But then: $\Rightarrow$

$$\begin{aligned} f(\vec{x}_{local}) &\leq f(\vec{x}_\theta) \\ &\leq \theta f(\vec{x}_{local}) + (1-\theta) f(\vec{x}_{gl} \\ &< \theta f(\vec{x}_{local}) + (1-\theta) f(\vec{x}_{local}) \\ &= f(\vec{x}_{local}) \end{aligned}$$

$$\cancel{f(\vec{x}_{local}) < f(\vec{x}_{local})}$$

Another way is using
Newton's method:

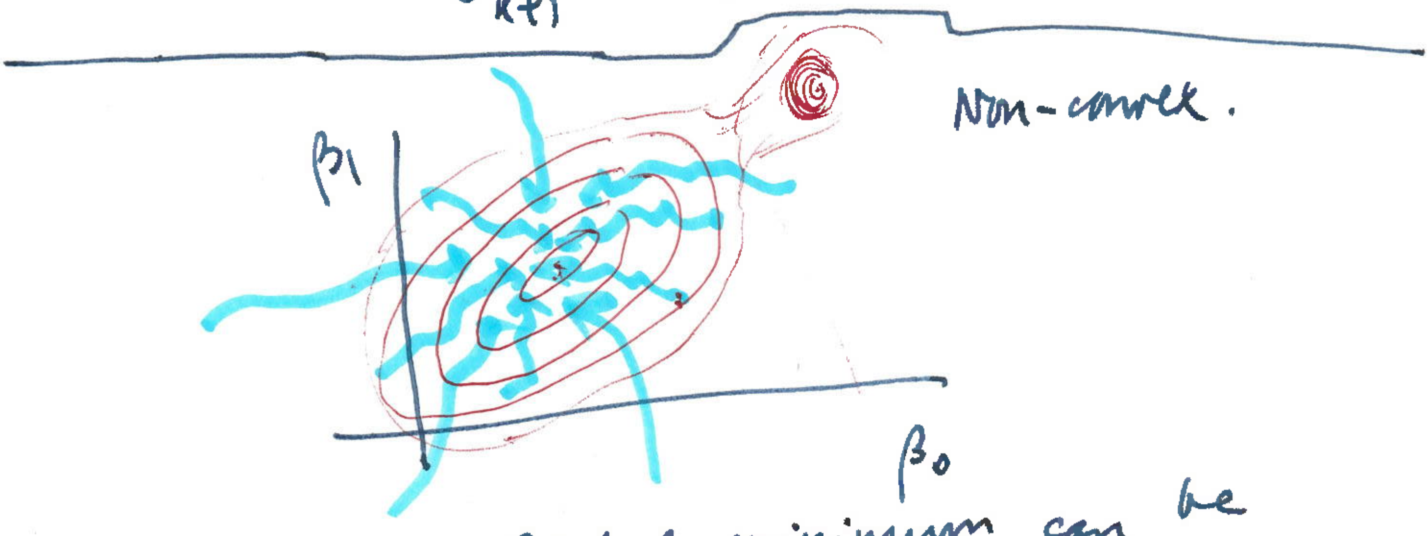
2nd order

$$L \simeq L(\theta_k) + \nabla L\big|_{\theta_k}(\vec{\theta}-\vec{\theta}_k) + \frac{1}{2}(\vec{\theta}-\vec{\theta}_k)^T H(\vec{\theta}_k)\cdot(\vec{\theta}-\vec{\theta}_k)$$

$$\nabla L \simeq \nabla L\big|_{\vec{\theta}_k} + H(\vec{\theta}_k)(\vec{\theta}-\vec{\theta}_k)$$

$$\nabla L(\vec{\theta}_k) + H(\vec{\theta}_k)(\vec{\theta}_{k+1}-\vec{\theta}_k) = 0$$

$$\vec{\theta}_{k+1} = \vec{\theta}_k - H(\vec{\theta}_k)^{-1}\nabla L(\vec{\theta}_k)$$



Non-convex.

Global minimum can be difficult to find

The same formulation applies for

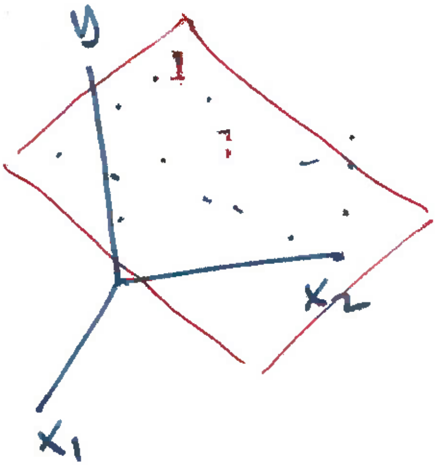
$p > 1$.

$$\left\{ x_1^{(i)} \; x_2^{(i)} \cdots x_p^{(i)} \; , \; y^{(i)} \right\}_{i=1}^{N}$$

$$\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)} = f(\vec{x}^{(i)}) \Longrightarrow \hat{y}^{(i)}$$

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \in \mathbb{R}^{(p+1)}$$

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_p^{(1)} \\ 1 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(N)} & \cdots & x_p^{(N)} \end{bmatrix}_{N\times(p+1)} \qquad \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}_{N\times 1}$$



$X\vec{\beta} = \vec{y}$? No solution

$$\vec{e} = \vec{y} - X\vec{\beta} \qquad \underbrace{\vec{\beta}^*_{LS} = (X^T X)^{-1} X^T \vec{y}}$$

linear regression model ⟵ Least squares solution

$$\min_{\vec{\beta}} \| \vec{y} - X\vec{\beta} \|^2$$

Bias vs variance

Hastie, ESL, Chapter 3

$$\mathbb{E}\left[\vec{\beta}^*\right] = \mathbb{E}\left[(X^T X)^{-1} X^T \vec{y}\right] = \vec{\beta} + \mathbb{E}\left[(X^T X)^{-1} X^T \vec{\varepsilon}\right] = \vec{\beta}$$

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$