$$\nabla_\theta J = \begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \frac{\partial J}{\partial \theta_1} \\ \vdots \\ \frac{\partial J}{\partial \theta_n} \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$f : \mathbb{R}^{m \times n} \longmapsto \mathbb{R}$$

$$A \in \mathbb{R}^{m \times n}$$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

If $A \in \mathbb{R}^{n \times n}$ (square matrix)

$\mathrm{tr}\, A = \sum_{i=1}^{n} A_{ii}$ (sum of diagonal elements)

Fact:

$\mathrm{tr}\, AB = \mathrm{tr}\, BA$

$\mathrm{tr}\, ABC = \mathrm{tr}\, CAB = \mathrm{tr}\, BCA$ (cyclic permutations)

$f(A) = \mathrm{tr}\, AB \; : \; \nabla_A \mathrm{tr}\, AB = B^T$

$\mathrm{tr}\, A = \mathrm{tr}\, A^t$

$a \in \mathbb{R} : \quad \mathrm{tr}\, a = a$

$\nabla_A \, \mathrm{tr}\, ABA^TC = CAB + C^TAB^T$

$$\nabla_\theta (X\theta - y)^T (X\theta - y)$$

$$= \nabla_\theta \underbrace{\left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta - y^T y\right)}$$

This is a real number.

Recall that $\mathrm{tr}\, a = a$ if $a \in \mathbb{R}$.

So the above becomes

$$= \nabla_\theta \, \mathrm{tr}\left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta - \cancel{y^T y}\right)$$

0. does not depend upon θ

$$= \nabla_\theta \, \mathrm{tr}\, \theta^T X^T X\theta - \nabla_\theta \, \mathrm{tr}\, \underbrace{\theta^T X^T y} - \nabla_\theta \, \mathrm{tr}\, y^T X\theta$$

This a real number.

Transpose of a real number is itself.

$\therefore (\theta^T X^T y)^T = y^T X\theta$

$$= \nabla_\theta \, \mathrm{tr}\, \theta^T X^T X\theta - \nabla_\theta \, \mathrm{tr}\, y^T X\theta - \nabla_\theta \, \mathrm{tr}\, y^T X\theta$$

$$= \underbrace{\nabla_\theta \, \mathrm{tr}\, \theta^T X^T X\theta} - 2\nabla_\theta \, \mathrm{tr}\, y^T X\theta$$

$$\nabla_\theta \operatorname{tr} \theta^T X^T X \theta = \nabla_\theta \operatorname{tr} \theta \theta^T X^T X$$

$$= \nabla_\theta \operatorname{tr} \underbrace{\theta}_{A} \underbrace{I}_{B} \underbrace{\theta^T}_{A^T} \underbrace{X^T X}_{C}$$

$$= \underbrace{X^T X}_{C} \underbrace{\theta}_{A} \underbrace{I}_{B} + \underbrace{(X^T X)^T}_{C^T} \underbrace{\theta}_{A} \underbrace{I^T}_{B^T}$$

$$= X^T X \theta + (X^T X)^T \theta$$

$X^T X$ is a square-symmetric matrix

$$\therefore \; X^T X = (X^T X)^T$$

$$\therefore \; \nabla_\theta \, \theta \theta^T X^T X = 2 X^T X \theta \quad \text{—— } A$$

$$\nabla_\theta \operatorname{tr} \underbrace{y^T X}_{B} \underbrace{\theta}_{A} = X^T y \quad \text{—— } B$$

So $\nabla_\theta \underbrace{(X\theta - y)^T (X\theta - y)} = 2 X^T X \theta - 2 X^T y$

This is a convex function.
We can find its minima by setting its derivative to 0.

$$2 X^T X \theta - 2 X^T y = 0$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T y .$$