As we know, $A^TA = \sum_i A_i^2$

$$\frac{1}{2}(X\theta - \bar{y})^T(X\theta - \bar{y}) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$= J(\theta)$$

$X \in R^{m \times n}$; $\theta \in R^{n \times 1}$

$y \in R^{m \times 1}$

To minimize J, derivative wrt 'θ'

$\nabla_{A^T} \text{tr}\, ABA^TC = B^TA^TCT + BA^TC$ [1×1 → scalar]

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - \bar{y})^T_{1\times1}(X\theta - \bar{y})_{1\times1}$$

$$= \frac{1}{2}\nabla_\theta(\theta^TX^TX\theta - \theta^TX^T\bar{y} - \bar{y}^TX\theta + \bar{y}^T\bar{y}) \rightarrow 1$$

$$= \frac{1}{2}\nabla_\theta \text{tr}(\theta^TX^TX\theta - \theta^TX^T\bar{y} - \bar{y}^TX\theta + \bar{y}^T\bar{y}) \rightarrow 2$$

$$= \frac{1}{2}\nabla_\theta(\text{tr}\,\theta^TX^TX\theta - 2\text{tr}\,\bar{y}^TX\theta) \rightarrow 3$$

$$= \frac{1}{2}(X^TX\theta + X^TX\theta - 2X^T\bar{y}) \rightarrow 4$$

$$= X^TX\theta - X^T\bar{y}$$

Understandy

- $X \in R^{m \times n}$, $\theta \in R^{n \times 1}$, $y \in R^{m \times 1}$
- The term $(X\theta - \bar{y})^T(X\theta - \bar{y})$ is a scalar (1×1 matrix)
- we multiplied $\frac{1}{2}$ bcoz of convenience:

↳ expanding the $(X\theta - \bar{y})^T(X\theta - \bar{y})$

$$= \theta^TX^TX\theta - \theta^TX^T\bar{y} - \bar{y}^TX\theta + \bar{y}^T\bar{y}$$

$\rightarrow \theta^T X^T X \theta - \theta^T X^T \bar{y} - \bar{y}^T X \theta + \bar{y}^T \bar{y}$ [1×1] → all are scalars.

[1×1] [1×1] [1×1] (const wrt θ)

As we know that all are scalars.

So, for a = scalar

$\boxed{tr\, a = a.}$

$\left[\theta^T \rightarrow (1\times n),\ x^T \rightarrow (n\times m),\ x \rightarrow (m\times n),\ \theta \rightarrow (n\times 1),\ (1\times m)(m\times n)(n\times 1) \rightarrow (1\times 1)\right]$

$$= \frac{1}{2} \nabla_\theta tr(\theta^T X^T X \theta - \theta^T X^T \bar{y} - \bar{y}^T X \theta + \bar{y}^T \bar{y})$$

Now, $\theta^T X^T \bar{y}$ is equal to $\bar{y}^T X \theta$ because both are scalar

$[\theta^T \in R^{n\times 1}, X^T \in R^{n\times m}, y \in R^{m\times 1}] \rightarrow X^T y \in R^{n\times 1} \rightarrow \theta^T X^T y \in R^{1\times 1}$

& $\bar{y}^T \bar{y}$ is independent of 'θ' → so it will vanish.

$$= \frac{1}{2} \nabla_\theta tr(\theta^T X^T X \theta - 2\theta(\bar{y}^T X \theta))$$

$$= \frac{1}{2} \nabla_\theta (tr(\theta^T X^T X \theta) - 2\, tr(\bar{y}^T X \theta))$$

↳ Matrix derivative. $[\nabla_\theta tr(\theta^T A \theta) = A\theta + A^T\theta]$

$[\nabla_\theta tr(B^T \theta) = B]$ Here $A = X^T X$

↳ $= X^T X \theta + X^T X \theta - 2X^T \bar{y}$ In first $\nabla_\theta tr(\theta^T \underbrace{X^T X}_{A} \theta)$

$= X^T X \theta - X^T \bar{y}$ $A = X^T X \in R^{n\times n}$

↳ $tr(\theta^T A \theta) = (A + A^T)\theta$

Since, A is symmetric ($A = X^T X$ ✓, $A^T = A$)

[A matrix is symmetric only iff $A^T = A$

[Here, $A = X^T X \rightarrow A^T = (X^T X)^T = X^T X$

So, property $\rightarrow \nabla_\theta \, tr(\theta^T A \theta) = (A + A^T)\theta$; Here

$= 2A\theta$ [As $A = A^T$]

$= 2x^T x\theta$.

Second term, $(\vec{y}^T x\theta)$

$y^T x \in R^{1 \times n}$

$\theta \in R^{n \times 1} \rightarrow y^T x\theta \in R^{1 \times 1}$ — scalar

$tr(\vec{y}^T x\theta) = y^T x\theta$

$\hookrightarrow \nabla_\theta \, tr(A\theta) = A^T$ (If A is const wrt θ)

$A = y^T x$

$\hookrightarrow \nabla_\theta tr = x^T \vec{y}$

Final gradient

$$\nabla_\theta J(\theta) = \frac{1}{2}(2x^T x\theta - 2x^T y) = \boxed{x^T x\theta - x^T \vec{y}}$$

To minimize J, we set $= 0$

$x^T x\theta - x^T \vec{y} = 0$

$x^T x\theta = x^T \vec{y}$

$$\boxed{\theta = (x^T x)^{-1}(x^T \vec{y})}$$