- Vectorized Gradients:
  - Jacobian Matrix:

suppose a function $f: \mathbb{R}^n \to \mathbb{R}^m$

$$f(x) = [f_1(x_1, \cdots, x_n), f_2(x_1, \cdots, x_n), \cdots, f_m(x_1, \cdots, x_n)]$$

its Jacobian is

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

That is, $\left(\frac{\partial f}{\partial x}\right)_{ij} = \frac{\partial f_i}{\partial x_j}$

  - e.g. with chain rule:

$$f(x) = [f_1(x), f_2(x)] \quad , \quad g(y) = [g_1(y_1, y_2), g_2(y_1, y_2)]$$

And $g(x) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$

$$\Rightarrow \frac{\partial g}{\partial x} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial g_1}{\partial f_1} & \frac{\partial g_1}{\partial f_2} \\ \frac{\partial g_2}{\partial f_1} & \frac{\partial g_2}{\partial f_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \end{bmatrix}$$

- Useful Identities

(1) $W \in \mathbb{R}^{n \times m}$, $x \in \mathbb{R}^m$, consider $z = W \cdot x \in \mathbb{R}^n$

$$z_i = \sum_{k=1}^{m} W_{ik} x_k$$

So an entry $\left(\frac{\partial z}{\partial x}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{m} W_{ik} x_k = \sum_{k=1}^{m} W_{ik} \frac{\partial}{\partial x_j} x_k$

$$= W_{ij}$$

$$\Rightarrow \frac{\partial z}{\partial x} = W$$

(2) $W \in \mathbb{R}^{n\times m}$, $x \in \mathbb{R}^{1\times n}$, consider $z = xW \in \mathbb{R}^{1\times m}$

Basically, this maps from $\mathbb{R}^n$ to $\mathbb{R}^m$.

So we expect Jacobian to be $\mathbb{R}^{m\times n}$

$$z_i = \sum_{k=1}^{n} x_k W_{ki}$$

So $$\left(\frac{\partial z}{\partial x}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} x_k W_{ki} = W_{ji}$$

$$\Rightarrow \quad \frac{\partial z}{\partial x} = W^T$$

(3) consider $z = x \in \mathbb{R}^n$

$$\frac{\partial z}{\partial x} = \mathbb{1}$$

(4) f is an elementwise function applied on $x \in \mathbb{R}^n$

consider $z = f(x)$

$$\left(\frac{\partial z}{\partial x}\right)_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j} f(x_i) = \begin{cases} f'(x_i) & \text{if } i=j \\ 0 & \text{if otherwise} \end{cases}$$

$$\Rightarrow \quad \frac{\partial z}{\partial x} = \operatorname{diag}(f'(x))$$

(5) Matrix times column vector with respect to the matrix

($z = Wx$, $\delta = \frac{\partial J}{\partial z}$, what's $\frac{\partial J}{\partial W} = \frac{\partial J}{\partial z}\frac{\partial z}{\partial W} = \delta \frac{\partial z}{\partial W}$?)

$J \in \mathbb{R}$, $x \in \mathbb{R}^n$, $W \in \mathbb{R}^{m\times n}$, $z \in \mathbb{R}^m$

$$\frac{\partial J}{\partial W} = \delta^T x^T$$

(6) $z = xW$, $\delta = \frac{\partial J}{\partial z}$, what's $\frac{\partial J}{\partial W} = \delta \frac{\partial z}{\partial W}$?

$J \in \mathbb{R}$, $x \in \mathbb{R}^{1\times n}$, $W \in \mathbb{R}^{n\times m}$, $z \in \mathbb{R}^{1\times m}$

$$\frac{\partial J}{\partial W} = x^T \delta$$

(7) $\hat{y} = \text{softmax}(\theta)$, $J = CE(y, \hat{y})$, what's $\frac{\partial J}{\partial \theta}$?)

$y, \hat{y} \in \mathbb{R}^n$

$$\frac{\partial J}{\partial \theta} = (\hat{y} - y)^T$$

(8) $\frac{\partial \|x\|^2}{\partial x}$?, $x \in \mathbb{R}^n$

function of $\mathbb{R}^n \mapsto \mathbb{R}$ so Jacobian: $\mathbb{R}^{1\times n}$

$$\left(\frac{\partial \|x\|^2}{\partial x}\right)_{ij} = \frac{\partial \sum_{k=1}^{n} x_k^2}{\partial x_j} = 2x_j$$

$$\Rightarrow \frac{\partial \|x\|^2}{\partial x} = 2x^T$$

- Return to linear regression:

$\hat{Y} = X\theta$ where $\hat{Y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times(p+1)}$, $\theta \in \mathbb{R}^{p+1}$

Loss: $R(\theta) = \frac{1}{n}\|Y - \hat{Y}\|_2^2 = \frac{1}{n}\|Y - X\theta\|_2^2$

By identity (8)

$$\frac{\partial R}{\partial \theta} = \frac{\partial R}{\partial (Y - X\theta)} \frac{\partial (Y - X\theta)}{\partial \theta}$$

$$= 2(Y - X\theta)^T \cdot \frac{\partial\, -X\theta}{\partial\theta} \qquad \text{by identity (1)}$$

$$= 2(Y - X\theta)^T \cdot (-X)$$

let $\frac{\partial R}{\partial\theta} = 0$, we get

$$(Y^T - \theta^T X^T) X = Y^T X - \theta^T X^T X = 0$$

Transpose each side, we get

$$X^T Y - X^T X \theta = 0$$

$$\Rightarrow \quad \theta = (X^T X)^{-1} X^T Y \qquad \text{assuming } X^T X \text{ is invertible}$$