# Matrix Derivative

March 8, 2021

## 1 Basic Calculus

$$dy = \frac{dy}{dx}dx \tag{1}$$
$$dy = \frac{dy}{d\mathbf{x}}d\mathbf{x} \tag{2}$$
$$dy = \mathrm{tr}\left(\frac{dy}{d\mathbf{X}}d\mathbf{X}\right) \tag{3}$$
$$d(\mathbf{XY}) = (d\mathbf{X})\mathbf{Y} + \mathbf{X}(d\mathbf{Y}) \tag{4}$$
$$d(\mathbf{X}\otimes\mathbf{Y}) = (d\mathbf{X})\otimes\mathbf{Y} + \mathbf{X}\otimes(d\mathbf{Y}) \tag{5}$$

$$d(\mathbf{X}\circ\mathbf{Y}) = (d\mathbf{X})\circ\mathbf{Y} + \mathbf{X}\circ(d\mathbf{Y}) \tag{6}$$
$$d(\mathbf{X}^\top) = (d\mathbf{X})^\top \tag{7}$$
$$d(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1} \tag{8}$$
$$d(\mathrm{tr}(\mathbf{X})) = \mathrm{tr}(d\mathbf{X}) \tag{9}$$
$$d(|\mathbf{X}|) = \mathrm{tr}(adj(\mathbf{X})d\mathbf{X}) \tag{10}$$
$$= |\mathbf{X}|\mathrm{tr}(\mathbf{X}^{-1}d\mathbf{X}) \tag{11}$$

## 2 Examples

Let's work on some examples

$$\begin{aligned} d(\mathbf{x}^\top\mathbf{x}) &= d(\mathbf{x}^\top)\mathbf{x} + \mathbf{x}^\top d\mathbf{x} && \text{from eq. (4)}\\ &= (d\mathbf{x})^\top\mathbf{x} + \mathbf{x}^\top d\mathbf{x} && \text{from eq. (7)}\\ &= \mathbf{x}^\top d\mathbf{x} + \mathbf{x}^\top d\mathbf{x} = 2\mathbf{x}^\top d\mathbf{x} \end{aligned}$$

therefore, $\frac{d(\mathbf{x}^\top\mathbf{x})}{d\mathbf{x}} = 2\mathbf{x}^\top$.

$$\begin{aligned} d||\mathbf{Wx}+\mathbf{b}||2^2 &= d\left(\mathbf{y}^\top\mathbf{y}\right)\Big|\mathbf{y} = \mathbf{Wx}+\mathbf{b}\\ &= 2\mathbf{y}^\top d\left(\mathbf{y}\right)\Big|\mathbf{y} = \mathbf{Wx}+\mathbf{b}\\ &= 2(\mathbf{Wx}+\mathbf{b})^\top(d\mathbf{W})\mathbf{x}\\ &= \mathrm{tr}(2(\mathbf{Wx}+\mathbf{b})^\top(d\mathbf{W})\mathbf{x})\\ &= \mathrm{tr}(2\mathbf{x}(\mathbf{Wx}+\mathbf{b})^\top d\mathbf{W}) \end{aligned}$$

3rd line is because we are differentiating w.r.t. $\mathbf{W}$, 4th line is because trace of singleton is itself. And the last one is from

$$\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CAB}) \tag{12}$$

therefore, we get $\frac{d||\mathbf{Wx}+\mathbf{b}||_2^2}{d\mathbf{W}} = 2\mathbf{x}(\mathbf{Wx}+\mathbf{b})^\top$

$$\begin{aligned} d\ln|\mathbf{X}| &= |\mathbf{X}|^{-1}d|\mathbf{X}|\\ &= |\mathbf{X}|^{-1}|\mathbf{X}|\mathrm{tr}(\mathbf{X}^{-1}d\mathbf{X}) && \text{from eq. (10)}\\ &= \mathrm{tr}(\mathbf{X}^{-1}d\mathbf{X}) \end{aligned}$$

this is not the norm or absolute value but the determinant of a matrix ($ad - bc$). from eq. (3), $d\ln|\mathbf{X}| = \mathbf{X}^{-1}d\mathbf{X}$, we get $\frac{d\ln|\mathbf{X}|}{d\mathbf{X}} = \mathbf{X}^{-T}$, i.e., transpose of inverse matrix

$$\begin{aligned} d(\mathrm{tr}(\mathbf{AXB})) &= \mathrm{tr}\left(d(\mathbf{AXB})\right) && \text{from eq. (9)}\\ &= \mathrm{tr}(\mathbf{A}(d\mathbf{X})\mathbf{B})\\ &= \mathrm{tr}(\mathbf{BA}d\mathbf{X}) \end{aligned}$$

finally from eq. (3), we get $\frac{d\mathrm{tr}(\mathbf{AXB})}{d\mathbf{X}} = \mathbf{BA}$.

$$\begin{aligned} d\mathbf{X}^{-1}\mathbf{X} &= 0\\ d(\mathbf{X}^{-1})\mathbf{X} + \mathbf{X}^{-1}d\mathbf{X} &= \mathbf{0} \end{aligned}$$

therefore, $d(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1}$

Let's consider a two layer neural network, $\mathcal{L}(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}))$, $\mathcal{L}$ is a loss function such as Softmax, Cross Entropy and MSE, $\sigma$ is an element-wise activation function such as Sigmoid and ReLU. It is more obvious if shape is provided.

### 2.1 derivative w.r.t. $\mathbf{W}_2$

let $\mathbf{y} = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$

$$\begin{aligned} d\mathcal{L}(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})) &= \frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}} d\mathbf{y} \\ &= \frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}} (d\mathbf{W}_2)\sigma(\mathbf{W}_1\mathbf{x}) \\ &= tr\left(\frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}} (d\mathbf{W}_2)\sigma(\mathbf{W}_1\mathbf{x})\right) \\ &= tr\left(\sigma(\mathbf{W}_1\mathbf{x})\frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}} d\mathbf{W}_2\right) \end{aligned}$$

from eq. (3), we get

$$\frac{\partial\mathcal{L}(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}))}{\partial\mathbf{W}_2} = \sigma(\mathbf{W}1\mathbf{x})\frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}}$$

### 2.2 derivative w.r.t. $\mathbf{W}_1$

let $\mathbf{y} = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$

$$d\mathcal{L}(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})) = \frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}}\mathbf{W}_2 d(\sigma(\mathbf{W}_1\mathbf{x}))$$

$$= \frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}}\mathbf{W}_2[\sigma'(\mathbf{W}_1\mathbf{x}) \circ d(\mathbf{W}_1)\mathbf{x}]$$

since $\mathbf{x}^\top(\mathbf{y} \circ \mathbf{z}) = (\mathbf{x} \circ \mathbf{y})^\top\mathbf{z}$

$$= \left[\left(\frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}}\mathbf{W}_2\right)^\top \circ \sigma'(\mathbf{W}_1\mathbf{x})\right]^\top (d\mathbf{W}_1)\mathbf{x}$$

$$= \mathrm{tr}\left(\mathbf{x}\left[\left(\frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}}\mathbf{W}_2\right) \circ \sigma'(\mathbf{W}_1\mathbf{x})^\top\right] d\mathbf{W}_1\right)$$

finally from eq. (3), we get

$$\frac{\partial\mathcal{L}(\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x}))}{\partial\mathbf{W}_1} = \mathbf{x}[(\frac{d\mathcal{L}(\mathbf{y})}{d\mathbf{y}}\mathbf{W}_2) \circ \sigma'(\mathbf{W}1\mathbf{x})^\top]$$

In mathematics, the *convolution* Boeing and Waddell (2017) of two functions $f$ and $g$ is defined as:

## References

Boeing, G. and Waddell, P. (2017). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research*, **37**(4), 457–476.