

# CS229 Lecture Notes 1 Derivations

Carson Tang

September 5, 2016

## 1 Update Rule for Linear Regression

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \left( \frac{1}{2} \sum_{i=0}^m (\theta^T x^{(i)} - y^{(i)})^2 \right) \\&= \frac{1}{2} \frac{\partial}{\partial \theta_j} \left[ (\theta^T x^{(0)} - y^{(0)})^2 + (\theta^T x^{(1)} - y^{(1)})^2 + \dots + (\theta^T x^{(m)} - y^{(m)})^2 \right] \\&= \frac{1}{2} \underbrace{\left[ \frac{\partial}{\partial \theta_j} (\theta^T x^{(0)} - y^{(0)})^2 + \frac{\partial}{\partial \theta_j} (\theta^T x^{(1)} - y^{(1)})^2 + \dots + \frac{\partial}{\partial \theta_j} (\theta^T x^{(m)} - y^{(m)})^2 \right]}_{\text{via the sum rule}} \\&= \frac{1}{2} \left[ 2 (\theta^T x^{(0)} - y^{(0)}) \cdot \frac{\partial}{\partial \theta_j} (\theta^T x^{(0)} - y^{(0)}) + 2 (\theta^T x^{(1)} - y^{(1)}) \cdot \frac{\partial}{\partial \theta_j} (\theta^T x^{(1)} - y^{(1)}) \right. \\&\quad \left. + \dots + 2 (\theta^T x^{(m)} - y^{(m)}) \cdot \frac{\partial}{\partial \theta_j} (\theta^T x^{(m)} - y^{(m)}) \right] \\&= \frac{1}{2} \left[ 2 (\theta^T x^{(0)} - y^{(0)}) \cdot x_j^{(0)} + 2 (\theta^T x^{(1)} - y^{(1)}) \cdot x_j^{(1)} + \dots + 2 (\theta^T x^{(m)} - y^{(m)}) \cdot x_j^{(m)} \right] \\&= \frac{1}{2} \underbrace{\sum_{i=0}^m 2 (\theta^T x^{(i)} - y^{(i)}) \cdot x_j^{(i)}}_{\text{extract the 2 via the distributive property of multiplication}} \\&= \sum_{i=0}^m (\theta^T x^{(i)} - y^{(i)}) \cdot x_j^{(i)}\end{aligned}$$

$$\mathbf{2} \quad tr(AB) = tr(BA)$$

$$tr(A) = \sum_{i=1}^n A_{ii} \tag{1}$$

$$tr(AB) = tr(C) \tag{2}$$

$$AB_{ii} = \sum_{k=1}^n A_{ik}B_{ki} \tag{3}$$

$$BA_{ii} = \sum_{k=1}^n B_{ik}A_{ki} \tag{4}$$

$$C_{ij} = \sum_{k=1}^n A_{ik}B_{kj} \tag{5}$$

$$tr(C) = \sum_{i=1}^n C_{ii} \tag{6}$$

$$= \sum_{i=1}^n AB_{ii} \tag{7}$$

$$= \sum_{i=1}^n \sum_{k=1}^n A_{ik}B_{ki} = \sum_{i=1}^n \sum_{k=1}^n B_{ki}A_{ik} = \sum_{k=1}^n \sum_{i=1}^n B_{ki}A_{ik} \tag{8}$$

$$= \sum_{k=1}^n BA_{kk} \tag{9}$$

$$= tr(BA) = tr(AB) \tag{10}$$

(1) is the definition of *trace*. Every element  $AB_{ij}$  in the matrix  $AB$  is the dot product of  $A$ 's  $i$ th row and  $B$ 's  $j$ th column. You can picture in your head iterating through  $A$ 's  $i$ th row, element by element, each one getting multiplied by its corresponding  $B$ 's  $j$ th column's element. To put it in a different way, imagine  $A$ 's row fixed and  $B$ 's column fixed. Now turn  $B$  over into a row. Next, pair up  $A$  and  $B$ 's elements and multiply them, and finally add them. The switches in (8) made because of the commutative property of addition.  $i$  and  $k$  are both iterating up to  $n$ , so these two elements are no different from each other. We're simply rearranging the elements. This might be harder to see, so let's show that for  $n = 2$ , this is true.

$$\begin{aligned}
& \sum_{i=1}^2 \sum_{k=1}^2 A_{ik} B_{ki} \\
&= \sum_{i=1}^2 (A_{i1} B_{1i} + A_{i2} B_{2i}) \\
&= (A_{11} B_{11} + A_{12} B_{21}) + (A_{21} B_{12} + A_{22} B_{22}) \\
&= (B_{11} A_{11} + B_{21} A_{12}) + (B_{12} A_{21} + B_{22} A_{22}) \\
&= (B_{11} A_{11} + B_{12} A_{21}) + (B_{21} A_{12} + B_{22} A_{22}) \\
&= \sum_{i=1}^2 (B_{i1} A_{1i} + B_{i2} A_{2i}) \\
&= \sum_{i=1}^2 \sum_{k=1}^2 B_{ik} A_{ki}
\end{aligned}$$

As you can see, by writing out the double summation for a  $2 \times 2$  matrix, you are rearranging the grouping the products of the elements of  $A$  and  $B$ . Originally, they were ordered by columns of  $B$ . Then, they were ordered by the rows of  $B$ . Yet another way to think about this would be to put on your computer science hat, and see that it does not matter that you iterate through the inside summation first. The inside summation makes you add up the dot product of the elements of the columns of  $A$  with the rows of  $B$ . The outside summation makes you add up the dot product of the elements of columns of  $B$  with the rows of  $A$ . You can also read a crystal clear explanation **here**

$$\mathbf{3} \quad \text{tr}(ABC) = \text{tr}(ACB) = \dots = \text{tr}(CBA)$$

$$\begin{aligned}
\text{tr}(ABC) &= \text{tr}((AB)C) = \text{tr}(DC) = \text{tr}(CD) \\
&= \text{tr}(CAB) = \text{tr}(EB) = \text{tr}(BE) = \text{tr}(BAC) = \text{and so on}
\end{aligned}$$

$$4 \quad \text{tr}(A) = \text{tr}(A^T)$$

$$\begin{aligned} \text{tr}(A^T) &= \sum_{i=1}^N (A^T)_{ii} \\ &= \sum_{i=1}^N A_{ii} \\ &= \text{tr}(A) \end{aligned}$$

$\text{tr}(A^T) = \text{tr}(A)$  is pretty straightforward because the elements along on a matrix's diagonal are equal to the transpose matrix's.

$$5 \quad \text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$$

$$\begin{aligned} C &= A + B \\ \text{tr}(A) + \text{tr}(B) &= \sum_{i=1}^N A_{ii} + \sum_{i=1}^N B_{ii} \\ &= \sum_{i=1}^N (A_{ii} + B_{ii}) \\ &= \sum_{i=1}^N C_{ii} \\ &= \text{tr}(C) \\ &= \text{tr}(A + B) \end{aligned}$$

The sum of the traces (diagonal sums) is equal to the sum of the diagonals of the sum.

$$\mathbf{6} \quad tr(aA) = atr(A)$$

$$\begin{aligned} C_{ii} &= aA_{ii} \\ atr(A) &= a \sum_{i=1}^n A_{ii} \\ &= \sum_{i=1}^n aA_{ii} \quad (\text{distributive property of multiplication}) \\ &= \sum_{i=1}^n C_{ii} \\ &= tr(C) = tr(aA) \end{aligned}$$

$$7 \quad \nabla_A \text{tr}(AB) = B^T$$

Remember, *trace* is function on **square matrices**.

$$f(AB) = \text{tr}(AB) \quad (11)$$

$$\nabla_A \text{tr}(AB) = \nabla_A f(AB) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{n1}} & \cdots & \frac{\partial f}{\partial A_{nn}} \end{bmatrix} \quad (12)$$

$$AB_{ii} = \sum_{k=1}^n A_{ik} B_{ki} \quad (13)$$

$$\text{tr}(AB) = \sum_{i=1}^n \sum_{k=1}^n A_{ik} B_{ki} \quad (14)$$

$$= A_{11}B_{11} + A_{12}B_{21} + \dots + A_{1n}B_{n1} \quad (15)$$

$$+ A_{21}B_{12} + A_{22}B_{22} + \dots + A_{2n}B_{n2} \quad (16)$$

$$\vdots \quad (17)$$

$$+ A_{n1}B_{1n} + A_{n2}B_{2n} + \dots + A_{nn}B_{nn} \quad (18)$$

$$\frac{\partial f}{\partial A_{ij}} = B_{ji} \quad (19)$$

$$\nabla_A \text{tr}(AB) = \begin{bmatrix} B_{11} & \cdots & B_{n1} \\ \vdots & \ddots & \vdots \\ B_{1n} & \cdots & B_{nn} \end{bmatrix} \quad (20)$$

$$= B^T \quad (21)$$

$$(22)$$

$\frac{\partial f}{\partial A_{ij}} = B_{ji}$  from above can be seen if you write out the double summation and take a few partial derivatives, as shown below

$$\frac{\partial f}{\partial A_{11}} = \frac{\partial(\sum_{i=1}^n \sum_{k=1}^n A_{ik} B_{ki})}{\partial A_{11}}$$

$$= B_{11}$$

$$\frac{\partial f}{\partial A_{12}} = \frac{\partial(\sum_{i=1}^n \sum_{k=1}^n A_{ik} B_{ki})}{\partial A_{12}}$$

$$= B_{21}$$

$$8 \quad \nabla_{A^T} f(A) = (\nabla_A f(A))^T$$

$$\nabla_{A^T} f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{1n}} & \cdots & \frac{\partial f}{\partial A_{nn}} \end{bmatrix} = (\nabla_A f(A))^T$$

$$9 \quad \nabla_A \text{tr} ABA^T C = CAB + C^T AB^T$$

$$\nabla_A \text{tr} ABA^T C = \nabla_A \text{tr} A^T ABC$$

$$D = BCA^T$$

$$\nabla_A \text{tr} ABCA^T = \nabla_A \text{tr} AD$$

$$= D^T = (BCA^T)^T$$

$$E = BC$$

$$(EA^T)^T = (A^T)^T E^T$$

$$= AE^T$$

$$= A(BC)^T$$

$$C^T AB^T =$$

$$\text{tr} BA^T C = \text{tr} A^T BC$$