An examplary computation for derivatives with matrix multiplications

Michael Moeller

November 13, 2019

1 Exemplary "matrix derivative"

To exemplify the situation we discussed in the lecture, consider the function

$$E(A,B,C) = \frac{1}{2}\|AB - C\|_F^2 \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times o}$, $A \in \mathbb{R}^{m\times o}$, and the squared Frobenius norm $\|\cdot\|_F^2$ is defined by squaring all entries of a matrix and summing them up, i.e.,

$$E(A,B,C) = \frac{1}{2}\|AB - C\|_F^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{o}((AB-C)_{ik})^2. \tag{2}$$

Let us consider how derivatives with respect to A, B and C look like.

1.1 Derivative w.r.t. C

The easiest derivative is with respect to C. We consider the partial derivative of E w.r.t. C_{st} and find

$$\frac{\partial}{\partial C_{st}} E(A,B,C) = -(AB-C)_{st}$$

because only one the summands in (2) contains a C_{st}. The rest is just a simple application of the chain rule in 1-d.

Since $\frac{\partial}{\partial C_{st}} E(A,B,C) = (C - AB)_{st}$, one could summarize such a result as

$$\nabla_C E(A,B,C) = C - AB, \tag{3}$$

but of course this is not based on any formal definition of the gradient we made in the lecture.

1.2 Derivative w.r.t. A

Slightly more difficult is the derivativative w.r.t. A. Again we consider the partial derivative w.r.t. A_{st}, but have to write out the matrix-matrix product:

$$E(A,B,C) = \frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{o}((AB-C)_{ik})^2 \tag{4}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{o}(\sum_{j} A_{ij}B_{jk} - C_{ik})^2 \tag{5}$$

and find

$$\frac{\partial}{\partial A_{st}}E(A,B,C) = \sum_{k=1}^{o}(\sum_{j} A_{sj}B_{jk} - C_{sk})B_{tk}, \tag{6}$$

which we could rewrite as

$$\frac{\partial}{\partial A_{st}}E(A,B,C) = \sum_{k=1}^{o}(\sum_{j} A_{sj}B_{jk} - C_{sk})(B^T)_{kt}, \tag{7}$$

$$= \sum_{k=1}^{o}((AB-C)_{sk})(B^T)_{kt}, \tag{8}$$

$$= ((AB-C)B^T)_{st}. \tag{9}$$

This means $\frac{\partial}{\partial A_{st}}E(A,B,C) = ((AB-C)B^T)_{st}$ and one again could summarize

$$\nabla_A E(A,B,C) = (AB-C)B^T. \tag{10}$$

Moreover, viewing $AB-C$ as the outer derivative of the loss function, we can identify "right-multiplication with B^T" as the (inner) derivative of the function $A \mapsto AB$.

1.3 Derivative w.r.t. B

Similar to the computation above, one can show that $\frac{\partial}{\partial B_{st}}E(A,B,C) = (A^T(AB-C))_{st}$, such that we could summarize

$$\nabla_B E(A,B,C) = A^T(AB-C), \tag{11}$$

and view "left-multiplication with A^T" as the derivative of the function $B \mapsto AB$.

For the more math-interested: The above concepts can be nicely formalized with what is called a *Frechet-Derivative*. In our case it would say that the derivative of a function $f_B : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times o}$, $f_B(A) = AB$, must be a linear function, also from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times o}$. In our example, it is "left multiplication with B". The adjoint of this function maps from $\mathbb{R}^{m\times o}$ to $\mathbb{R}^{m\times n}$ and is what we called the gradient. In our example it is "right-multiplication with B^T.