Vector, Matrix, and Tensor Derivatives

Erik Learned-Miller

The purpose of this document is to help you learn to take derivatives of vectors, matrices, and higher order tensors (arrays with three dimensions or more), and to help you take derivatives *with respect to* vectors, matrices, and higher order tensors.

1 Simplify, simplify, simplify

Much of the confusion in taking derivatives involving arrays stems from trying to do too many things at once. These "things" include taking derivatives of multiple components simultaneously, taking derivatives in the presence of summation notation, and applying the chain rule. By doing all of these things at the same time, we are more likely to make errors, at least until we have a lot of experience.

1.1 Expanding notation into explicit sums and equations for each component

In order to simplify a given calculation, it is often useful to write out the explicit formula for *a single scalar element* of the output in terms of nothing but *scalar variables*. Once one has an explicit formula for a single scalar element of the output in terms of other scalar values, then one can use the calculus that you used as a beginner, which is much easier than trying to do matrix math, summations, and derivatives all at the same time.

Example. Suppose we have a column vector $\vec{y}$ of length C that is calculated by forming the product of a matrix W that is C rows by D columns with a column vector $\vec{x}$ of length D:

$$\vec{y} = \overset{(C,D)}{W}\overset{(D,1)}{\vec{x}}. \in \mathbb{R}^C \qquad \nabla_{\vec{x}} \vec{y} \tag{1}$$

Suppose we are interested in the derivative of $\vec{y}$ with respect to $\vec{x}$. A full characterization of this derivative requires the (partial) derivatives of each component of $\vec{y}$ with respect to each component of $\vec{x}$, which in this case will contain $C \times D$ values since there are C components in $\vec{y}$ and D components of $\vec{x}$.

Let's start by computing one of these, say, the 3rd component of $\vec{y}$ with respect to the 7th component of $\vec{x}$. That is, we want to compute

$$\frac{\partial \vec{y}_3}{\partial \vec{x}_7},$$

$$\vec{y} : \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_c \end{bmatrix} \qquad \vec{x} : \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

$$\underset{(C,1)}{\vec{y}} = \underset{(C,D)(D,1)}{W\vec{x}} \quad \vec{y}: \underset{(C,1)}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_C \end{bmatrix}} = W = \underset{(C,D)}{\begin{bmatrix} W_1 & \cdots & W_{1D} \\ \vdots & & \vdots \\ W_{C1} & \cdots & W_{CD} \end{bmatrix}} \cdot \vec{x}: \underset{(D,1)}{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}} = \begin{bmatrix} W_{11} \\ W_{21} \\ \vdots \\ W_{C1} \end{bmatrix} x_1 + \begin{bmatrix} W_{12} \\ W_{22} \\ \vdots \\ W_{C2} \end{bmatrix} x_2 + \cdots + \begin{bmatrix} W_{1D} \\ W_{2D} \\ \vdots \\ W_{CD} \end{bmatrix} x_D$$

$\frac{\partial \vec{y}_3}{\partial \vec{x}_7}$ which is just the derivative of one scalar with respect to another.

The first thing to do is to write down the formula for computing $\vec{y}_3$ so we can take its derivative. From the definition of matrix-vector multiplication, the value $\vec{y}_3$ is computed by taking the dot product between the 3rd row of W and the vector $\vec{x}$:

$$\underset{(1,1)}{\vec{y}_3} = W_{31} \cdot x_1 + W_{32} x_2 + \cdots W_{3D} x_D$$

$$\vec{y}_3 = \sum_{j=1}^{D} W_{3,j}\, \vec{x}_j. \quad \underset{(1,D)}{\begin{bmatrix} W_{31} & W_{32} & \cdots & W_{3D} \end{bmatrix}} \quad \underset{(D,1)}{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}} \tag{2}$$

At this point, we have reduced the original matrix equation (Equation 1) to a scalar equation. This makes it much easier to compute the desired derivatives.

$(1,D)(D,1) = (1,1)$

1.2 Removing summation notation

While it is certainly possible to compute derivatives directly from Equation 2, people frequently make errors when differentiating expressions that contain summation notation ($\sum$) or product notation ($\prod$). When you're beginning, it is sometimes useful to write out a computation without any summation notation to make sure you're doing everything right. Using "1" as the first index, we have:

$$\vec{y}_3 = W_{3,1}\vec{x}_1 + W_{3,2}\vec{x}_2 + \ldots + W_{3,7}\vec{x}_7 + \ldots + W_{3,D}\vec{x}_D.$$

Of course, I have explicitly included the term that involves $\vec{x}_7$, since that is what we are differenting with respect to. At this point, we can see that the expression for y_3 only depends upon $\vec{x}_7$ through a single term, $W_{3,7}\vec{x}_7$. Since none of the other terms in the summation include $\vec{x}_7$, their derivatives with respect to $\vec{x}_7$ are all 0. Thus, we have

$$\begin{aligned}
\frac{\partial \vec{y}_3}{\partial \vec{x}_7} &= \frac{\partial}{\partial \vec{x}_7}\left[W_{3,1}\vec{x}_1 + W_{3,2}\vec{x}_2 + \ldots + W_{3,7}\vec{x}_7 + \ldots + W_{3,D}\vec{x}_D\right] &(3)\\
&= 0 + 0 + \ldots + \underset{(1,1)}{\frac{\partial}{\partial \vec{x}_7}}\,[\underset{(1,1)}{W_{3,7}}\underset{(1,1)}{\vec{x}_7}] + \ldots + 0 &(4)\\
&= \frac{\partial}{\partial \vec{x}_7}\,[W_{3,7}\vec{x}_7] &(5)\\
&= W_{3,7}. &(6)
\end{aligned}$$

By focusing on one component of $\vec{y}$ and one component of $\vec{x}$, we have made the calculation about as simple as it can be. In the future, when you are confused, it can help to try to reduce a problem to this most basic setting to see where you are going wrong.

1.2.1 Completing the derivative: the Jacobian matrix $\frac{\partial \vec{y}}{\partial \vec{x}}$

Recall that our original goal was to compute the derivatives of each component of $\vec{y}$ with respect to each component of $\vec{x}$, and we noted that there would be $C \times D$ of these. They

$$\frac{\partial \vec{y}}{\partial \vec{x}} \qquad \underset{(C,1)}{\vec{y}} = \underset{(C,D)}{W}\,\underset{(D,1)}{\vec{x}}$$

can be written out as a matrix in the following form:

$$\frac{\partial \vec{y}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial \vec{y}_1}{\partial \vec{x}_1} & \frac{\partial \vec{y}_1}{\partial \vec{x}_2} & \frac{\partial \vec{y}_1}{\partial \vec{x}_3} & \dots & \frac{\partial \vec{y}_1}{\partial \vec{x}_D} \\ \frac{\partial \vec{y}_2}{\partial \vec{x}_1} & \frac{\partial \vec{y}_2}{\partial \vec{x}_2} & \frac{\partial \vec{y}_2}{\partial \vec{x}_3} & \dots & \frac{\partial \vec{y}_2}{\partial \vec{x}_D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \vec{y}_C}{\partial \vec{x}_1} & \frac{\partial \vec{y}_C}{\partial \vec{x}_2} & \frac{\partial \vec{y}_C}{\partial \vec{x}_3} & \dots & \frac{\partial \vec{y}_C}{\partial \vec{x}_D} \end{bmatrix} \longleftarrow \text{Jacobian matrix}$$

In this particular case, this is called the *Jacobian matrix*, but this terminology is not too important for our purposes.

Notice that for the equation

$$\vec{y} = W\vec{x},$$

the partial of $\vec{y}_3$ with respect to $\vec{x}_7$ was simply given by $W_{3,7}$. If you go through the same process for other components, you will find that, for all i and j,

$$\frac{\partial \vec{y}_i}{\partial \vec{x}_j} = W_{i,j}.$$

This means that the matrix of partial derivatives is

$$\begin{bmatrix} \frac{\partial \vec{y}_1}{\partial \vec{x}_1} & \frac{\partial \vec{y}_1}{\partial \vec{x}_2} & \frac{\partial \vec{y}_1}{\partial \vec{x}_3} & \dots & \frac{\partial \vec{y}_1}{\partial \vec{x}_D} \\ \frac{\partial \vec{y}_2}{\partial \vec{x}_1} & \frac{\partial \vec{y}_2}{\partial \vec{x}_2} & \frac{\partial \vec{y}_2}{\partial \vec{x}_3} & \dots & \frac{\partial \vec{y}_2}{\partial \vec{x}_D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \vec{y}_C}{\partial \vec{x}_1} & \frac{\partial \vec{y}_C}{\partial \vec{x}_2} & \frac{\partial \vec{y}_C}{\partial \vec{x}_3} & \dots & \frac{\partial \vec{y}_C}{\partial \vec{x}_D} \end{bmatrix} = \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} & \dots & W_{1,D} \\ W_{2,1} & W_{2,2} & W_{2,3} & \dots & W_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{C,1} & W_{C,2} & W_{C,3} & \dots & W_{C,D}. \end{bmatrix} = W$$

This, of course, is just W itself.

Thus, after all this work, we have concluded that for

$$\vec{y} = W\vec{x},$$

we have

$$\frac{d\vec{y}}{d\vec{x}} = W.$$

2 Row vectors instead of column vectors

It is important in working with different neural networks packages to pay close attention to the arrangement of weight matrices, data matrices, and so on. For example, if a data matrix X contains many different vectors, each of which represents an input, is each data vector a row or column of the data matrix X?

In the example from the first section, we worked with a vector $\vec{x}$ that was a column vector. However, you should also be able to use the same basic ideas when $\vec{x}$ is a row vector.

2.1 Example 2

$$\underset{(1,C)}{\vec{y}} = [\vec{y}_1 \; \vec{y}_2 \cdots \vec{y}_c] \qquad \underset{(1,D)}{\vec{x}} = [\vec{x}_1 \; \vec{x}_2 \cdots \vec{x}_D] \qquad W = \underset{(D,C)}{\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1C} \\ W_{21} & W_{22} & \cdots & W_{2C} \\ \vdots & & & \\ W_{D1} & W_{D2} & \cdots & W_{DC} \end{bmatrix}}$$

Let $\vec{y}$ be a *row vector* with C components computed by taking the product of another row vector $\vec{x}$ with D components and a matrix W that is D rows by C columns.

$$\underset{(1,C)}{\vec{y}} = \underset{(1,D)}{\vec{x}}\underset{(D,C)}{W}.$$

Importantly, despite the fact that $\vec{y}$ and $\vec{x}$ have the same number of components as before, the shape of W is the *transpose* of the shape that we used before for W. In particular, since we are now left-multiplying by $\vec{x}$, whereas before $\vec{x}$ was on the right, W must be transposed for the matrix algebra to work.

$$\vec{y}_1 = \vec{x} \cdot \vec{W}_1 = [\vec{x}_1 \; \vec{x}_2 \cdots \vec{x}_D] \cdot \begin{bmatrix} W_{11} \\ W_{21} \\ \vdots \\ W_{D1} \end{bmatrix}$$

$$\underset{(1,1)}{\vec{y}_3} = \vec{x} \cdot \vec{W}_3 = [\vec{x}_1 \; \vec{x}_2 \cdots \vec{x}_D] \cdot \begin{bmatrix} W_{13} \\ W_{23} \\ \vdots \\ W_{D3} \end{bmatrix}$$

$$= \vec{x}_1 W_{13} + \vec{x}_2 W_{23} + \cdots + \vec{x}_D W_{D3}$$

$$\frac{\partial \vec{y}_3}{\partial x_\eta} = \frac{\partial}{\partial x_\eta}\{\vec{x}_1 W_{13} + \vec{x}_2 W_{23} + \cdots \vec{x}_\eta W_{\eta 3} + \cdots \vec{x}_D W_{D3}\}$$

$$= W_{\eta 3}$$

In this case, you will see, by writing

$$\frac{\partial \vec{y}_i}{\partial x_j} = W_{ji} \qquad W = D\left\{\overbrace{\begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix}}^{C}\right.$$

j = from 1 to D

i = from 1 to C

$$\underset{(1,D)}{\vec{x}} \quad \underset{(1,C)}{\vec{y}}$$

$$\vec{y}_3 = \sum_{j=1}^{D} \vec{x}_j W_{j,3}$$

that

$$\frac{\partial \vec{y}_3}{\partial \vec{x}_7} = W_{7,3}.$$

Notice that the indexing into W is the opposite from what it was in the first example. However, when we assemble the full Jacobian matrix, we can still see that in this case as well,

$$\frac{d\vec{y}}{d\vec{x}} = W. \tag{7}$$

3 Dealing with more than two dimensions

Let's consider another closely related problem, that of computing

$$\frac{d\vec{y}}{dW}.$$

In this case, $\vec{y}$ varies along one coordinate while W varies along two coordinates. Thus, the entire derivative is most naturally contained in a *three*-dimensional array. We avoid the term "three-dimensional matrix" since it is not clear how matrix multiplication and other matrix operations are defined on a three-dimensional array.

Dealing with three-dimensional arrays, it becomes perhaps more trouble than it's worth to try to find a way to display them. Instead, we should simply define our results as formulas which can be used to compute the result on any element of the desired three dimensional array.

Let's again compute a scalar derivative between one component of $\vec{y}$, say $\vec{y}_3$ and one component of W, say $W_{7,8}$. Let's start with the same basic setup in which we write down an equation for $\vec{y}_3$ in terms of other scalar components. Now we would like an equation that expresses $\vec{y}_3$ in terms of scalar values, and shows the role that $W_{7,8}$ plays in its computation.

$$\frac{\partial \vec{y}_3}{\partial W_{7,8}}$$

However, what we see is that $W_{7,8}$ plays *no role* in the computation of $\vec{y}_3$, since

$$\vec{y}_3 = \vec{x}_1 W_{1,3} + \vec{x}_2 W_{2,3} + ... + \vec{x}_D W_{D,3}. \tag{8}$$

In other words,

$$\frac{\partial \vec{y}_3}{\partial W_{7,8}} = 0.$$

However, the partials of $\vec{y}_3$ with respect to elements of the 3rd column of W will certainly be non-zero. For example, the derivative of $\vec{y}_3$ with respect to $W_{2,3}$ is given by

$$\frac{\partial \vec{y}_3}{\partial W_{2,3}} = \frac{\partial}{\partial W_{2,3}} \left\{ \vec{x}_1 W_{13} + \vec{x}_2 W_{23} + \cdots + \vec{x}_D W_{D3} \right\} = \underset{(1,1)}{\vec{x}_2}$$

$$\frac{\partial \vec{y}_3}{\partial W_{2,3}} = \vec{x}_2, \tag{9}$$

as can be easily seen by examining Equation 8.

In general, when the index of the $\vec{y}$ component is equal to the second index of W, the derivative will be non-zero, but will be zero otherwise. We can write:

$$\frac{\partial \vec{y}_j}{\partial W_{i,j}} = \vec{x}_i,$$

but the other elements of the 3-d array will be 0. If we let F represent the 3d array representing the derivative of $\vec{y}$ with respect to W, where

$$F_{i,j,k} = \frac{\partial \vec{y}_i}{\partial W_{j,k}},$$

then

$$F_{i,j,i} = \vec{x}_j, = \frac{\partial \vec{y}_i}{\partial W_{ji}} = \vec{x}_j$$

but all other entries of F are zero.

Finally, if we define a new *two-dimensional* array G as

$$G_{i,j} = F_{i,j,i}$$

we can see that all of the information we need about F can be stored in G, and that the non-trivial portion of F is really two-dimensional, not three-dimensional.

Representing the important part of derivative arrays in a compact way is critical to efficient implementations of neural networks.

4 Multiple data points

It is a good exercise to repeat some of the previous examples, but using multiple examples of $\vec{x}$, stacked together to form a matrix X. Let's assume that each individual $\vec{x}$ is a row vector of length D, and that X is a two-dimensional array with N rows and D columns. W, as in our last example, will be a matrix with D rows and C columns. Y, given by

$$\vec{x} = [x_1 x_2 \cdots x_D]$$

$$\underset{(N,C)}{Y} = \underset{(N,D)}{X}\underset{(D,C)}{W},$$

$$X : \begin{bmatrix} | & | & & | \\ \vec{x}_1^T & \vec{x}_2^T & \cdots & \vec{x}_N \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix} \in \mathbb{R}^{N\times D}$$

$$W \in \mathbb{R}^{D\times C} : \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1C} \\ W_{21} & W_{22} & \cdots & W_{2C} \\ \vdots & & & \vdots \\ W_{D1} & W_{D2} & \cdots & W_{DC} \end{bmatrix}$$

$$Y = XW$$
$$(N,C) \quad (N,D) \quad (D,C)$$

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1C} \\ Y_{21} & Y_{22} & \cdots & Y_{2C} \\ \vdots & \vdots & & \\ Y_{N1} & Y_{N2} & \cdots & Y_{NC} \end{bmatrix} \quad X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1D} \\ X_{21} & X_{22} & \cdots & X_{2D} \\ \vdots & & & \\ X_{N1} & X_{N2} & \cdots & X_{ND} \end{bmatrix} \quad W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1C} \\ W_{21} & W_{22} & \cdots & W_{2C} \\ \vdots & & & \vdots \\ W_{D1} & W_{D2} & \cdots & W_{DC} \end{bmatrix}$$

$$Y_{11} = X_{11}W_{11} + X_{12}W_{21} + \cdots + X_{1D}W_{D1}$$
$$Y_{12} = X_{11}W_{12} + X_{12}W_{22} + \cdots + X_{1D}W_{D2}$$
$$\vdots$$

$Y \in \mathbb{R}^{N \times C}$ will also be a matrix, with N rows and C columns. Thus, each row of Y will give a row vector associated with the corresponding row of the input X.

Sticking to our technique of writing down an expression for a given component of the output, we have

$$Y_{i,j} = \sum_{k=1}^{D} X_{i,k} W_{k,j}.$$

We can see immediately from this equation that among the derivatives

$$\frac{\partial Y_{a,b}}{\partial X_{c,d}}, = \frac{\partial}{\partial X_{c,d}} \{ X_{a1}W_{1b} + X_{a2}W_{2b} + \cdots X_{aD}W_{Db} \}$$

they are all zero unless $a = c$. That is, since each component of Y is computed using only the corresponding row of X, derivatives of components between different rows of Y and X are all zero.

Furthermore, we can see that

$$\frac{\partial Y_{i,j}}{\partial X_{i,k}} = W_{k,j} \tag{10}$$

doesn't depend at all upon which row of Y and X we are comparing.

In fact, the matrix W holds all of these partials as it is–we just have to remember to index into it according to Equation 10 to obtain the specific partial derivative that we want.

If we let $Y_{i,:}$ be the ith row of Y and let $X_{i,:}$ be the ith row of X, then we see that

$$\frac{\partial Y_{i,:}}{\partial X_{i,:}} = W,$$

which is a simple generalization of our previous result from Equation 7.

5 The chain rule in combination with vectors and matrices

Now that we have worked through a couple of basic examples, let's combine these ideas with an example of the chain rule. Again, assuming $\vec{y}$ and $\vec{x}$ are column vectors, let's start with the equation

$$\vec{y} = VW\vec{x},$$

and try to compute the derivative of $\vec{y}$ with respect to $\vec{x}$. We could simply observe that the product of two matrices V and W is simply another matrix, call it U, and therefore

$$\frac{d\vec{y}}{d\vec{x}} = VW = U.$$

However, we want to go through the process of using the chain rule to define intermediate results, so that we can see how the chain rule applies in the context of non-scalar derivatives.

$$\vec{y} = V\underset{\sim}{W\vec{x}}$$

Let us define the intermediate result

$$\vec{m} = W\vec{x}.$$

Then we have that

$$\vec{y} = V\vec{m}.$$

We can then write, using the chain rule, that

$$\frac{d\vec{y}}{d\vec{x}} = \frac{d\vec{y}}{d\vec{m}}\frac{d\vec{m}}{d\vec{x}}.$$

To make sure that we know exactly what this means, let's take the old approach of analyzing one component at a time, starting with a single component of $\vec{y}$ and a single component of $\vec{x}$:

$$\frac{d\vec{y}_i}{d\vec{x}_j} = \frac{d\vec{y}_i}{d\vec{m}}\frac{d\vec{m}}{d\vec{x}_j}.$$

** But how exactly should we interpret the product on the right? The idea with the chain rule is to *multiply* the change in $\vec{y}_i$ with respect to *each scalar* intermediate variable by the change in the scalar intermediate variable with respect to $\vec{x}_j$. In particular, if $\vec{m}$ has M components, then we write

$$\frac{d\vec{y}_i}{d\vec{x}_j} = \sum_{k=1}^{M} \frac{d\vec{y}_i}{d\vec{m}_k}\frac{d\vec{m}_k}{d\vec{x}_j}. = \left\{ \frac{dy_i}{d\vec{m}_1} \cdot \frac{d\vec{m}_1}{d\vec{x}_j} + \frac{d\vec{y}_i}{d\vec{m}_2}\frac{d\vec{m}_2}{dx_j} + \cdots + \frac{d\vec{y}_i}{d\vec{m}_M}\frac{d\vec{m}_M}{d\vec{x}_j} \right\}$$

Recall from our previous results about derivatives of a vector with respect to a vector that

$$\frac{d\vec{y}_i}{d\vec{m}_k}$$

is just $V_{i,k}$ and that

$$\frac{d\vec{m}_k}{d\vec{x}_j}$$

is just $W_{k,j}$. So we can write

$$\frac{d\vec{y}_i}{d\vec{x}_j} = \sum_{k=1}^{M} V_{i,k}W_{k,j},$$

which is just the component expression for VW, our original answer to the problem.

To summarize, we can use the chain rule in the setting of vector and matrix derivatives by

- Clearly stating intermediate results and the variables used to represent them,
- Expressing the chain rule for individual components of the final derivatives,
- Summing appropriately over the intermediate results within the chain rule expression.