APPENDIX A

Matrix derivative common cases

What are some conventions for derivatives of matrices and vectors? It will always work to explicitly write all indices and treat everything as scalars, but we introduce here some shortcuts that are often faster to use and helpful for understanding.

There are at least two consistent but different systems for describing shapes and rules for doing matrix derivatives. In the end, they all are correct, but it is important to be consistent.

We will use what is often called the 'Hessian' or denominator layout, in which we say that for

$\mathbf{x}$ of size $n \times 1$ and $\mathbf{y}$ of size $m \times 1$, $\partial \mathbf{y}/\partial \mathbf{x}$ is a matrix of size $n \times m$ with the (i, j) entry $\partial y_j/\partial x_i$. This denominator layout convention has been adopted by the field of machine learning to ensure that the shape of the gradient is the same as the shape of the shape of the respective derivative. This is somewhat controversial at large, but alas, we shall continue with denominator layout.

The discussion below closely follows the Wikipedia on matrix derivatives.

A.1 The shapes of things

Here are important special cases of the rule above:

- Scalar-by-scalar: For x of size 1×1 and y of size 1×1, $\partial y/\partial x$ is the (scalar) partial derivative of y with respect to x.
- Scalar-by-vector: For $\mathbf{x}$ of size $n \times 1$ and y of size 1×1, $\partial y/\partial \mathbf{x}$ (also written $\nabla_x y$, the gradient of y with respect to $\mathbf{x}$) is a column vector of size $n \times 1$ with the i^{th} entry $\partial y/\partial x_i$:

$$\partial y/\partial \mathbf{x} = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_n \end{bmatrix}.$$

- Vector-by-scalar: For x of size 1×1 and $\mathbf{y}$ of size $m \times 1$, $\partial \mathbf{y}/\partial x$ is a row vector of size $1 \times m$ with the j^{th} entry $\partial \mathbf{y}_j/\partial x$:

$$\partial \mathbf{y}/\partial x = \begin{bmatrix} \partial y_1/\partial x & \partial y_2/\partial x & \cdots & \partial y_m/\partial x \end{bmatrix}.$$

- Vector-by-vector: For $\mathbf{x}$ of size $n \times 1$ and $\mathbf{y}$ of size $m \times 1$, $\partial \mathbf{y}/\partial \mathbf{x}$ is a matrix of size $n \times m$ with the (i, j) entry $\partial y_j/\partial x_i$:

$$\partial \mathbf{y}/\partial \mathbf{x} = \begin{bmatrix} \partial y_1/\partial x_1 & \partial y_2/\partial x_1 & \cdots & \partial y_m/\partial x_1 \\ \partial y_1/\partial x_2 & \partial y_2/\partial x_2 & \cdots & \partial y_m/\partial x_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial y_1/\partial x_n & \partial y_2/\partial x_n & \cdots & \partial y_m/\partial x_n \end{bmatrix}.$$

- Scalar-by-matrix: For $\mathbf{X}$ of size $n \times m$ and y of size 1×1, $\partial y/\partial \mathbf{X}$ (also written $\nabla_{\mathbf{X}} y$, the gradient of y with respect to $\mathbf{X}$) is a matrix of size $n \times m$ with the (i, j) entry $\partial y/\partial X_{i,j}$:

$$\partial y/\partial \mathbf{X} = \begin{bmatrix} \partial y/\partial X_{1,1} & \cdots & \partial y/\partial X_{1,m} \\ \vdots & \ddots & \vdots \\ \partial y/\partial X_{n,1} & \cdots & \partial y/\partial X_{n,m} \end{bmatrix}.$$

You may notice that in this list, we have not included matrix-by-matrix, matrix-by-vector, or vector-by-matrix derivatives. This is because, generally, they cannot be expressed nicely in matrix form and require higher order objects (e.g., tensors) to represent their derivatives. These cases are beyond the scope of this course.

Additionally, notice that for all cases, you can explicitly compute each element of the derivative object using (scalar) partial derivatives. You may find it useful to work through some of these by hand as you are reviewing matrix derivatives.

A.2 Some vector-by-vector identities

Here are some examples of $\partial \mathbf{y}/\partial \mathbf{x}$. In each case, assume $\mathbf{x}$ is $n \times 1$, $\mathbf{y}$ is $m \times 1$, a is a scalar constant, $\mathbf{a}$ is a vector that does not depend on $\mathbf{x}$ and $\mathbf{A}$ is a matrix that does not depend on $\mathbf{x}$, u and v are scalars that do depend on $\mathbf{x}$, and $\mathbf{u}$ and $\mathbf{v}$ are vectors that do depend on $\mathbf{x}$. We also have vector-valued functions $\mathbf{f}$ and $\mathbf{g}$.

A.2.1 Some fundamental cases

First, we will cover a couple of fundamental cases: suppose that $\mathbf{a}$ is an $m \times 1$ vector which is not a function of $\mathbf{x}$, an $n \times 1$ vector. Then,

$$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0}, \tag{A.1}$$

is an $n \times m$ matrix of 0s. This is similar to the scalar case of differentiating a constant. Next, we can consider the case of differentiating a vector with respect to itself:

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I} \tag{A.2}$$

This is the $n \times n$ identity matrix, with 1's along the diagonal and 0's elsewhere. It makes sense, because $\partial \mathbf{x}_j/\mathbf{x}_i$ is 1 for $i = j$ and 0 otherwise. This identity is also similar to the scalar case.

A.2.2 Derivatives involving a constant matrix

Let the dimensions of $\mathbf{A}$ be $m \times n$. Then the object $\mathbf{Ax}$ is an $m \times 1$ vector. We can then compute the derivative of $\mathbf{Ax}$ with respect to $\mathbf{x}$ as:

$$\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}} = \begin{bmatrix} \partial(\mathbf{Ax})_1/\partial x_1 & \partial(\mathbf{Ax})_2/\partial x_1 & \cdots & \partial(\mathbf{Ax})_m/\partial x_1 \\ \partial(\mathbf{Ax})_1/\partial x_2 & \partial(\mathbf{Ax})_2/\partial x_2 & \cdots & \partial(\mathbf{Ax})_m/\partial x_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial(\mathbf{Ax})_1/\partial x_n & \partial(\mathbf{Ax})_2/\partial x_n & \cdots & \partial(\mathbf{Ax})_m/\partial x_n \end{bmatrix} \tag{A.3}$$

Note that any element of the column vector $\mathbf{Ax}$ can be written as, for $j = 1, \ldots, m$:

$$(\mathbf{Ax})_j = \sum_{k=1}^{n} A_{j,k} x_k.$$

Thus, computing the (i, j) entry of $\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}}$ requires computing the partial derivative $\partial(\mathbf{Ax})_j/\partial x_i$:

$$\partial(\mathbf{Ax})_j/\partial x_i = \partial \left(\sum_{k=1}^{n} A_{j,k} x_k \right) /\partial x_i = A_{j,i}$$

Therefore, the (i, j) entry of $\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}}$ is the (j, i) entry of $\mathbf{A}$:

$$\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}} = \mathbf{A}^T \tag{A.4}$$

Similarly, for objects $\mathbf{x}, \mathbf{A}$ of the same shape, on can obtain,

$$\frac{\partial \mathbf{x}^T \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A} \tag{A.5}$$

A.2.3 Linearity of derivatives

Suppose that $\mathbf{u}, \mathbf{v}$ are both vectors of size $m \times 1$. Then,

$$\frac{\partial(\mathbf{u} + \mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \tag{A.6}$$

Suppose that a is a scalar constant and $\mathbf{u}$ is an $m \times 1$ vector that is a function of $\mathbf{x}$. Then,

$$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} = a \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \tag{A.7}$$

One can extend the previous identity to vector- and matrix-valued constants. Suppose that $\mathbf{a}$ is a vector with shape $m \times 1$ and v is a scalar which depends on $\mathbf{x}$. Then,

$$\frac{\partial v\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial v}{\partial \mathbf{x}} \mathbf{a}^T \tag{A.8}$$

First, checking dimensions, $\partial v/\partial \mathbf{x}$ is $n \times 1$ and $\mathbf{a}$ is $m \times 1$ so $\mathbf{a}^T$ is $1 \times m$ and our answer is $n \times m$ as it should be. Now, checking a value, element (i, j) of the answer is $\partial v \mathbf{a}_j/\partial x_i = (\partial v/\partial x_i)\mathbf{a}_j$ which corresponds to element (i, j) of $(\partial v/\partial \mathbf{x})\mathbf{a}^T$.

Similarly, suppose that $\mathbf{A}$ is a matrix which does not depend on $\mathbf{x}$ and $\mathbf{u}$ is a column vector which does depend on $\mathbf{x}$. Then,

$$\frac{\partial \mathbf{Au}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{A}^T \tag{A.9}$$

A.2.4 Product rule (vector-valued numerator)

Suppose that v is a scalar which depends on $\mathbf{x}$, while $\mathbf{u}$ is a column vector of shape $m \times 1$ and $\mathbf{x}$ is a column vector of shape $n \times 1$. Then,

$$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} = v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}\mathbf{u}^T \tag{A.10}$$

One can see this relationship by expanding the derivative as follows:

$$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} = \begin{bmatrix} \partial(vu_1)/\partial x_1 & \partial(vu_2)/\partial x_1 & \cdots & \partial(vu_m)/\partial x_1 \\ \partial(vu_1)/\partial x_2 & \partial(vu_2)/\partial x_2 & \cdots & \partial(vu_m)/\partial x_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial(vu_1)/\partial x_n & \partial(vu_2)/\partial x_n & \cdots & \partial(vu_m)/\partial x_n \end{bmatrix}.$$

Then, one can use the product rule for scalar-valued functions,

$$\partial(vu_j)/\partial x_i = v(\partial u_j/\partial x_i) + (\partial v/\partial x_i)u_j,$$

to obtain the desired result.

A.2.5 Chain rule

Suppose that $\mathbf{g}$ is a vector-valued function with output vector of shape $m \times 1$, and the argument to $\mathbf{g}$ is a column vector $\mathbf{u}$ of shape $d \times 1$ which depends on $\mathbf{x}$. Then, one can obtain the chain rule as,

$$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \tag{A.11}$$

Following "the shapes of things," $\partial \mathbf{u}/\partial \mathbf{x}$ is $n \times d$ and $\partial \mathbf{g}(\mathbf{u})/\partial \mathbf{u}$ is $d \times m$, where element (i, j) is $\partial \mathbf{g}(\mathbf{u})_j/\partial \mathbf{u}_i$. The same chain rule applies for further compositions of functions:

$$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}} \tag{A.12}$$

A.3 Some other identities

You can get many scalar-by-vector and vector-by-scalar cases as special cases of the rules above, making one of the relevant vectors just be 1 x 1. Here are some other ones that are handy. For more, see the Wikipedia article on Matrix derivatives (for consistency, only use the ones in *denominator layout*).

$$\frac{\partial \mathbf{u}^T\mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{u} \tag{A.13}$$

$$\frac{\partial \mathbf{u}^T}{\partial x} = \left(\frac{\partial \mathbf{u}}{\partial x}\right)^T \tag{A.14}$$

A.4 Derivation of gradient for linear regression

Applying identities A.5, A.13, A.6, A.4 A.1

$$\begin{aligned}\frac{\partial(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}})^T(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}})/n}{\partial\theta} &= \frac{2}{n}\frac{\partial(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}})}{\partial\theta}(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}}) \\ &= \frac{2}{n}\Big(\frac{\partial\tilde{\mathbf{X}}\theta}{\partial\theta}-\frac{\partial\tilde{\mathbf{Y}}}{\partial\theta}\Big)(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}}) \\ &= \frac{2}{n}\big(\tilde{\mathbf{X}}^T-\mathbf{0}\big)(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}}) \\ &= \frac{2}{n}\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}}\theta-\tilde{\mathbf{Y}})\end{aligned}$$

A.5 Matrix derivatives using Einstein summation

You do not have to read or learn this! But you might find it interesting or helpful.

Consider the objective function for linear regression, written out as products of matrices:

$$J(\theta) = \frac{1}{n}(\tilde{X}\theta-\tilde{Y})^T(\tilde{X}\theta-\tilde{Y}), \tag{A.15}$$

where $\tilde{X} = X^T$ is $n \times d$, $\tilde{Y} = Y^T$ is $n \times 1$, and θ is $d \times 1$. How does one show, with no shortcuts, that

$$\nabla_\theta J = \frac{2}{n}\tilde{X}^T(\tilde{X}\theta-\tilde{Y}) \ ? \tag{A.16}$$

One neat way, which is very explicit, is to simply write all the matrices as variables with row and column indices, e.g., $\tilde{X}_{ab}$ is the row a, column b entry of the matrix $\tilde{X}$. Furthermore, let us use the convention that in any product, all indices which appear more than once get summed over; this is a popular convention in theoretical physics, and lets us suppress all the summation symbols which would otherwise clutter the following expressions. For example, $\tilde{X}_{ab}\theta_b$ would be the implicit summation notation giving the element at the a^{th} row of the matrix-vector product $\tilde{X}\theta$.

Using implicit summation notation with explicit indices, we can rewrite $J(\theta)$ as

$$J(\theta) = \frac{1}{n}\left(\tilde{X}_{ab}\theta_b-\tilde{Y}_a\right)\left(\tilde{X}_{ac}\theta_c-\tilde{Y}_a\right). \tag{A.17}$$

Note that we no longer need the transpose on the first term, because all that transpose accomplished was to take a dot product between the vector given by the left term, and the vector given by the right term. With implicit summation, this is accomplished by the two terms sharing the repeated index a.

Taking the derivative of J with respect to the d^{th} element of θ thus gives, using the chain rule for (ordinary scalar) multiplication:

$$\begin{aligned}\frac{dJ}{d\theta_d} &= \frac{1}{n}\left[\tilde{X}_{ab}\delta_{bd}\left(\tilde{X}_{ac}\theta_c-\tilde{Y}_a\right)+\left(\tilde{X}_{ab}\theta_b-\tilde{Y}_a\right)\tilde{X}_{ac}\delta_{cd}\right] && \text{(A.18)} \\ &= \frac{1}{n}\left[\tilde{X}_{ad}\left(\tilde{X}_{ac}\theta_c-\tilde{Y}_a\right)+\left(\tilde{X}_{ab}\theta_b-\tilde{Y}_a\right)\tilde{X}_{ad}\right] && \text{(A.19)} \\ &= \frac{2}{n}\tilde{X}_{ad}\left(\tilde{X}_{ab}\theta_b-\tilde{Y}_a\right), && \text{(A.20)}\end{aligned}$$

where the second line follows from the first, with the definition that $\delta_{bd} = 1$ only when $b = d$ (and similarly for δ_{cd}). And the third line follows from the second by recognizing that the two terms in the second line are identical. Now note that in this implicit summation notation, the a, b element of the matrix product of A and B is $(AB)_{ac} = A_{ab}B_{bc}$. That is, ordinary matrix multiplication sums over indices which are adjacent to each other, because a row of A times a column of B becomes a scalar number. So the term in the above equation with $\tilde{X}_{ad}\tilde{X}_{ab}$ is not a matrix product of $\tilde{X}$ with $\tilde{X}$. However, taking the transpose $\tilde{X}^T$ switches row and column indices, so $\tilde{X}_{ad} = \tilde{X}^T_{da}$. And $\tilde{X}^T_{da}\tilde{X}_{ab}$ *is* a matrix product of $\tilde{X}^T$ with $\tilde{X}$! Thus, we have that

$$\frac{dJ}{d\theta_d} = \frac{2}{n}\tilde{X}^T_{da}\left(\tilde{X}_{ab}\theta_b - \tilde{Y}_a\right) \tag{A.21}$$

$$= \frac{2}{n}\left[\tilde{X}^T\left(\tilde{X}\theta - \tilde{Y}\right)\right]_d , \tag{A.22}$$

which is the desired result.