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 $\partial \mathbf{y}_j/\partial \mathbf{x}_i$. 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:

- 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.
- For $\mathbf{x}$ of size $n \times 1$ and y of size 1×1, $\partial y/\partial \mathbf{x}$ (also written $\nabla_x y$) is a vector of size $n \times 1$ with the i^{th} entry $\partial y/\partial \mathbf{x}_i$.
- For x of size 1×1 and $\mathbf{y}$ of size $m \times 1$, $\partial \mathbf{y}/\partial x$ is a vector of size $1 \times m$ with the j^{th} entry $\partial \mathbf{y}_j/\partial x$.
- 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 $\partial \mathbf{y}_j/\partial \mathbf{x}_i$.
- For $\mathbf{X}$ of size $n \times m$ and y of size 1×1, $\partial y/\partial \mathbf{X}$ is a matrix of size $n \times m$ with the (i, j) entry $\partial y/\partial \mathbf{X}_{i,j}$.

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 $\mathbf{g}$.

(In the discussion below, the words are relevant to the *preceding* formula.)

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

This is an $n \times m$ matrix of 0s.

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

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.

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T \tag{A.3}$$

The dimension of $\mathbf{A}$ here would be $m \times n$. Each element of the derivative is $\partial(\mathbf{A}\mathbf{x})_j / \partial \mathbf{x}_i$ which is $\partial \sum_k \mathbf{A}_{jk}\mathbf{x}_k / \partial \mathbf{x}_i$. This is only non-zero when $i = k$ and in that case it's $\mathbf{A}_{ji}$. So, this gives us $\mathbf{A}^T$.

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

$$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} = a \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \tag{A.5}$$

The dimension of $\mathbf{u}$ is $m \times 1$.

$$\frac{\partial v\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial v}{\partial \mathbf{x}} \mathbf{a}^T \tag{A.6}$$

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 ij of the answer is $\partial v \mathbf{a}_j / \partial x_i = (\partial v / \partial x_i)\mathbf{a}_j$ which corresponds to element ij of $(\partial v / \partial \mathbf{x})\mathbf{a}^T$.

$$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} = v \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}} \mathbf{u}^T \tag{A.7}$$

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

$$\frac{\partial (\mathbf{u} + \mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \tag{A.9}$$

$$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \tag{A.10}$$

This is the basic chain rule. Assume $\mathbf{u}$ is $d \times 1$. Then $\partial \mathbf{u} / \partial \mathbf{x}$ is $n \times d$ and $\partial \mathbf{g}(\mathbf{u}) / \partial \mathbf{u}$ is $d \times m$, where element ij is $\partial \mathbf{g}(\mathbf{u})_j / \partial \mathbf{u}_i$.

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

Even more chain rule!

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*).

$$\frac{\partial \mathbf{u}^T \mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \mathbf{u} \tag{A.12}$$

$$\frac{\partial \mathbf{u}^T}{\partial x} = \left(\frac{\partial \mathbf{u}}{\partial x}\right)^T \tag{A.13}$$

A.4 Derivation of gradient for linear regression

Applying identities A.4, A.12, A.9, A.3 A.1

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

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:

$$J(\theta) = \frac{1}{n} (\tilde{X}\theta - \tilde{Y})^T (\tilde{X}\theta - \tilde{Y}) \,, \tag{A.14}$$

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

$$\nabla_\theta J = \frac{2}{n} \tilde{X}^T (\tilde{X}\theta - \tilde{Y}) \;\; ? \tag{A.15}$$

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

$$J(\theta) = \frac{1}{n} \left(\tilde{X}_{ab}\theta_b - \tilde{Y}_a\right) \left(\tilde{X}_{ac}\theta_c - \tilde{Y}_a\right) \,. \tag{A.16}$$

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:

$$\begin{aligned}
\frac{dJ}{d\theta_d} &= \frac{1}{n}\left[\tilde{X}_{ab}\delta_{bd}\left(\tilde{X}_{ac}\theta_c - \tilde{Y}_a\right) + \left(\tilde{X}_{ab}\theta_b - \tilde{Y}_a\right)\tilde{X}_{ac}\delta_{cd}\right] && \text{(A.17)}\\
&= \frac{1}{n}\left[\tilde{X}_{ad}\left(\tilde{X}_{ac}\theta_c - \tilde{Y}_a\right) + \left(\tilde{X}_{ab}\theta_b - \tilde{Y}_a\right)\tilde{X}_{ad}\right] && \text{(A.18)}\\
&= \frac{2}{n}\tilde{X}_{ad}\left(\tilde{X}_{ab}\theta_b - \tilde{Y}_a\right), && \text{(A.19)}
\end{aligned}$$

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

$$\begin{aligned}
\frac{dJ}{d\theta_d} &= \frac{2}{n}\tilde{X}^T_{da}\left(\tilde{X}_{ab}\theta_b - \tilde{Y}_a\right) && \text{(A.20)}\\
&= \frac{2}{n}\left[\tilde{X}^T\left(\tilde{X}\theta - \tilde{Y}\right)\right]_d, && \text{(A.21)}
\end{aligned}$$

which is the desired result.