CS 189 - Some matrix calculus

Noah Golmant

January 25, 2017

This material is very similar to a Piazza post made last semester by some other TAs. I found it very helpful when I took the course.

Consider an inner product space X equipped with inner product $\langle \cdot, \cdot \rangle$. Let $\|x\| = \sqrt{\langle x, x \rangle}$ denote the norm induced by the inner product. We have some functional $f : X \to \mathbb{R}$.

Let us define the *gradient* of f at a point $x \in X$. We denote the gradient, $\nabla f(x)$, as the unique element in X such that for all $\Delta \in X$,

$$f(x + \Delta) = f(x) + \langle \nabla f(x), \Delta \rangle + o(\|\Delta\|_X) \tag{1}$$

The little o means that the magnitude of the residual term grows slower than $\|\Delta\|$.

In this class we will really only consider two important inner product spaces: $\mathbb{R}^n$ and $\mathbb{R}^{m \times n}$. You may be familiar with how to find the gradient of f in $\mathbb{R}^n$, but using (1) we can also more easily find the gradient of some function with respect to a matrix.

Why does it make it easier? Well, one method to find the gradient is to take the partial derivatives of f with respect to each element of the matrix X, as we have done before with vectors. However, this can prove very tedious. It can also be hard to combine the resulting entries into a single matrix expression.

The trick with this method is that we can plug $X + \Delta$ into our equation, expand out, and rearrange the terms until we have isolated the $\nabla f(X)$ term as the left argument of the inner product on the RHS of (1).

Let's use the standard inner product on $\mathbb{R}^{m \times n}$. We define $\langle A, B \rangle_{\mathbb{R}^{m \times n}} = \text{Tr}(A^T B)$. The trace has some particularly nice properties that you can read about here.

As an example, consider the function $f(Y) = \text{Tr}(AYBY^TYC)$. Y is an $m \times n$ matrix and the other matrices are size-conforming. Let's find the gradient.

$$\begin{aligned}
f(Y + \Delta) &= \text{Tr}(A(Y + \Delta)B(Y + \Delta)^T(Y + \Delta)C) \\
&= \text{Tr}(AYBY^TYC + 2AYBY^T\Delta C + AYB\Delta^T\Delta C + A\Delta BY^TYC + 2A\Delta BY^T\Delta C + A\Delta B\Delta^T\Delta C) \\
&= f(Y) + \text{Tr}(2AYBY^T\Delta C + A\Delta BY^TYC) + o(\|\Delta\|) \\
&= f(Y) + \text{Tr}((2CAYBY^T + BY^TYCA)\Delta) + o(\|\Delta\|) \\
&= f(Y) + \langle 2YB^TY^TA^TC^T + A^TC^TY^TYB^T, \Delta \rangle + o(\|\Delta\|)
\end{aligned}$$

So by (1), we can see that the gradient is:

$$\nabla f(Y) = 2YB^TY^TA^TC^T + A^TC^TY^TYB^T$$

We managed to find the gradient of an ugly expression without too much pain! Just as a note, it's important to be careful with transposes here.