CME 108/MATH 114
Introduction to Scientific Computing
Summer 2019

Gradients involving matrices

In your walk about life, you may often encounter (as in HW4), functions which involving terms of the form $c^T x$, Ax and $x^T Ax$, with $c, x \in \mathbb{R}^n$, and an $n \times n$ matrix A. You may need to compute the corresponding differential (as in HW4).

In this note we discuss the rather satisfying formulas which result from such computations. We develop intuition for the formulas by appealing to the familiar $n = 1$ case.

Proposition 1. *Given $c \in \mathbb{R}^n$, the differential of $f(x) = c^T x$ is given by*

$$\nabla f(x) = c^T.$$

Proof. To motivate the formula, consider the case $n = 1$. In this case, $f(x) = cx$ and

$$\nabla f(x) = \frac{d}{dx} f(x) = \frac{d}{dx}[cx] = c,$$

since c, x are both scalars. The general formula then follows, recalling that the differential is a **row** vector.

Let's prove the general formula. This sort of computation is best done entry by entry. In this case,

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix},$$

and each partial derivative is evaluated at the point x. Therefore we need to compute $\frac{\partial f}{\partial x_k}$, for each $k = 1, 2, \ldots, n$. Since

$$f(x) = c^T x = \sum_{j=1}^{n} c_j x_j,$$

the derivative of a sum is the sum of the derivatives, and since each c_j is a constant (the differential operator $\frac{\partial}{\partial x_k}$ is linear),

$$\frac{\partial f}{\partial x_k}(x) = \frac{\partial}{\partial x_k}\left(\sum_{j=1}^{n} c_j x_j \right) = \sum_{j=1}^{n} c_j \frac{\partial}{\partial x_k} x_j = \sum_{j=1}^{n} c_j \delta_{jk} = c_k.$$

The second-to-last equality follows from the fact that

$$\frac{\partial}{\partial x_k} x_j = \begin{cases} 1, & \text{if } k = j, \\ 0, & \text{otherwise,} \end{cases}$$

since clearly the function x_j depends on x_j and not on any of the other $x_i's$. This fact is indeed key, and it will be used throughout this note. So please take a moment and ensure you've fully digested it.

The last equality shows a "collapsing" property of the Kronecker delta: it gobbles one index in a summation.

Combining our results gives the desired formula:

$$\begin{aligned} \nabla f(x) &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \\ &= c^T. \end{aligned}$$

□

Proposition 2. *Now consider the vector-valued function $f(x) = Ax$, for $x \in \mathbb{R}^n$ and an $n \times n$ matrix A. In this case, the differential is the Jacobian and it is given by*

$$J_f(x) = A.$$

Proof. Again, we appeal to the simple $n = 1$ case for some intuition. When $n = 1$, $f(x) = ax$ for some scalar a and therefore $J_f(x) = \frac{d}{dx} f = a$.

As above, one strategy here is to compute the entries of J individually. In this case,

$$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

by definition, with $f_i(x) = (Ax)_i$ denoting the ith row of f. Therefore we need to compute $\frac{\partial f_i}{\partial x_j}$. Again, since the derivative is linear, and since

$$f_i(x) = (Ax)_i = \sum_{k=1}^{n} a_{ik} x_k,$$

we have

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\sum_{k=1}^{n} a_{ik}x_k\right) = \sum_{k=1}^{n} a_{ik}\frac{\partial}{\partial x_j}x_k = \sum_{k=1}^{n} a_{ik}\delta_{jk} = a_{ij}.$$

Hence,

$$\begin{aligned} J_f(x) &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \\ &= A, \end{aligned}$$

as claimed.

It turns out, however, that in this case there is a better strategy for computing J_f, based on noting that

$$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \vdots \\ \nabla f_n(x) \end{bmatrix},$$

and that

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix} = Ax = \begin{bmatrix} \tilde{a}_1^T x \\ \tilde{a}_2^T x \\ \vdots \\ \tilde{a}_n^T x \end{bmatrix},$$

and using Proposition 1! (Here $\tilde{a}_i^T$ denotes the ith row of A.)

Proposition 1 implies $\nabla f_i(x) = \tilde{a}_i^T$, so

$$J_f(x) = \begin{bmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \vdots \\ \nabla f_n(x) \end{bmatrix} = \begin{bmatrix} \tilde{a}_1^T \\ \tilde{a}_2^T \\ \vdots \\ \tilde{a}_n^T \end{bmatrix} = A.$$

□

Proposition 3. *Finally, consider the scalar-valued function $f(x) = x^T Ax$ for $x \in \mathbb{R}^n$ and a **symmetric** $n \times n$ matrix A. In this case, the corresponding differential is given by*

$$\nabla f(x) = 2x^T A.$$

Remark. *The formula $\nabla f(x) = 2Ax$, which interprets the gradient as a column vector, is more standard. It boils down to a rather technical difference between the "gradient" and the "differential." Beware!*

Proof. As above, we extract intuition from the familiar $n = 1$ case. In this case $f(x) = xax = ax^2$ for some scalar a and $\nabla f(x) = \frac{d}{dx} f(x) = 2xa$. The formula follows by recalling that the differential is a **row**.

In this case,

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix},$$

so we must compute $\frac{\partial f}{\partial x_k}$, as in Proposition 1.

For starters, let's not assume A is symmetric. Since

$$f(x) = x^T Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j,$$

linearity and the product rule imply that

$$\begin{aligned}
\frac{\partial f}{\partial x_k} &= \frac{\partial}{\partial x_k} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \frac{\partial}{\partial x_k} (x_i x_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left[x_j \frac{\partial}{\partial x_k} x_i + x_i \frac{\partial}{\partial x_k} x_j \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} [x_j \delta_{ik} + x_i \delta_{jk}] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_j \delta_{ik} + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i \delta_{jk} \\
&= \sum_{j=1} a_{kj} x_j + \sum_{i=1} a_{ik} x_i \\
&= (x^T A^T)_k + (x^T A)_k \\
&= (x^T (A^T + A))_k
\end{aligned}$$

If A is symmetric, the result follows since then $A^T = A$. □

Corollary 3.1. *If* $f(x) = x^T x$, *then* $\nabla f(x) = 2x^T$.

Proof. This follows directly from Proposition 3 applied with $A = I$. □