Some Vector Calculus Identities

Sohom Paul

February 9, 2021

Fact 1. $\nabla_x(a^\intercal x) = a$

Proof. Here, we have some intimidating notation (gradients and inner products!) hiding a really simple computation. Let's start with a concrete example. Suppose that $x = \langle u, v\rangle$ and $a = \langle 3, 4\rangle$. Then $f(u, v) = 3u + 4v$, so $f_u = 3$ and $f_v = 4$ (where subscripts denote partial derivatives), so $\nabla f = \langle 3, 4\rangle = a$. In general, we have

$$\begin{aligned}
a^\intercal x &= a_1x_1 + a_2x_2 + \ldots a_nx_n \\
\frac{\partial}{\partial x_i} a^\intercal x &= \frac{\partial}{\partial x_i}\left(a_1x_1 + a_2x_2 + \ldots + a_ix_i + \ldots a_nx_n\right) \\
&= 0 + 0 + \ldots + a_i + \ldots + 0 \\
&= a_i
\end{aligned}$$

Because the ith component of $\nabla_x(a^\intercal x)$ matches the ith component of a, the two vectors are equal, which is what we set out to prove. □

Fact 2. $\nabla_x(x^\intercal Ax) = (A + A^\intercal)x$

Proof. Now, this compuation is a little bit trickier than the previous one. We notice that

$$\begin{aligned}
x^\intercal Ax &= \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
&= \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{pmatrix} \\
&= x_1a_{11}x_1 + x_1a_{12}x_2 \ldots + x_2a_{21}x_1 + x_2a_{22}x_2 + \ldots \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} x_ia_{ij}x_j
\end{aligned}$$

Now, to take the partial derivative with respect to x_k, there are 4 cases:

$$\frac{\partial}{\partial x_k}(x_i a_{ij} x_j) = \begin{cases} a_{kj}x_j & i = k, j \neq k \\ a_{ik}x_i & i \neq k, j = k \\ 2a_{kk}x_k & i = j = k \\ 0 & \text{otherwise} \end{cases}$$

It follows

$$\begin{aligned}
\frac{\partial}{\partial x_k}(x^\intercal A x) &= \frac{\partial}{\partial x_k}(x_i a_{ij} x_j) \\
&= 2a_{kk} + \sum_{\substack{1 \leq j \leq n \\ j \neq k}} a_{kj}x_j + \sum_{\substack{1 \leq i \leq n \\ i \neq k}} a_{ik}x_i \\
&= 2a_{kk} + \sum_{\substack{1 \leq j \leq n \\ j \neq k}} (a_{kj} + a_{jk})x_j \\
&= \sum_{1 \leq j \leq n} (a_{kj} + a_{jk})x_j \\
&= \begin{pmatrix} a_{k1} + a_{1k} & a_{k2} + a_{2k} & \dots & a_{kn} + a_{nk} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
\nabla_x(x^\intercal A x) &= \begin{pmatrix} a_{11} + a_{11} & a_{12} + a_{21} & \dots & a_{1n} + a_{n1} \\ a_{21} + a_{12} & a_{22} + a_{22} & \dots & a_{2n} + a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + a_{1n} & a_{n2} + a_{2n} & \dots & a_{nn} + a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
&= (A + A^\intercal)x
\end{aligned}$$

which is what we set out to prove. □

Theorem 1. $\arg\max_w \|Xw - y\|_2 = (X^\intercal X)^{-1}(X^\intercal y)$

Proof. You've likely seen a proof of this formula in your linear algebra class using some properties of orthogonal projections. I want to outline here an alternate proof using vector calculus that is easier to generalize to nonlinear models.

The key insight to recall the fact from elementary calculus that the derivative of a smooth function is 0 when it achieves a local minimum or maximum. The same holds for a function of multiple variables, except we set the gradient of the function to 0 (which is equivalent to setting all the partials to 0). Thus, minimizing the squared L^2 loss, we get

$$\begin{aligned}
f(w) &= \|Xw - y\|_2^2 \\
&= (Xw - y)^\intercal (Xw - y) \\
&= (w^\intercal X^\intercal - y^\intercal)(Xw - y) \\
&= w^\intercal X^\intercal X w - y^\intercal X w - w^\intercal X^\intercal y + y^\intercal y \\
&= w^\intercal X^\intercal X w - 2(X^\intercal y)^\intercal w + \|y\|_2^2 \\
\nabla_w f(w) &= (X^\intercal X + (X^\intercal X)^\intercal)w - 2X^\intercal y \\
&= 2(X^\intercal X)w - 2X^\intercal y
\end{aligned}$$

where we used Facts 1 and 2 when taking the gradient. Setting the last line to 0 immediately gives the optimal coefficient vector as $w^* = (X^\intercal X)^{-1} X^\intercal y$. (It can be checked that $X^\intercal X$ is invertible if X has linearly independent columns, which always holds in practice.) As this is the only point where the gradient is 0, we know that this is the global optimum. □

Note that to a machine learning practitioner, our matrix X has rows that correspond to *data points* and columns that correspond to *data features*. Our model is that the output is linear in the different features, with vector w telling us how much weight we assign to each. The vector y tells us the observed *labels*. Our goal is to fit some weight vector that minimizes the L^2 loss. As you will see in your homework, the "linear" in linear regression only refers to the fact that your model is a linear combination of features; by appropriately choosing data features, you can fit polynomials or other functions using linear regression. Neat!