2019-09-09

1 Matrix calculus

Numerical linear algebra is not just about algebra, but also about *analysis*, the branch of mathematics that deals with real functions and operations such as differentiation and integration (and their generalizations). This is particularly relevant when we deal with error analysis, as we will soon.

1.1 Warm up: derivative of a dot product

Consider the real-valued expression $y^T x$ as a function of the vector variables $x, y \in \mathbb{R}^n$. How would we compute the gradient of $y^T x$ with respect to these variables? The usual method taught in a first calculus class would be to write the expression in terms of each of the components of x and y, and then compute partial derivatives, i.e.

$$\begin{aligned} y^T x &= \sum_{i=1}^{n} x_i y_i \\ \frac{\partial (y^T x)}{\partial x_j} &= y_j \\ \frac{\partial (y^T x)}{\partial y_j} &= x_j. \end{aligned}$$

This notation is fine for dealing with a dot product, in which we are summing over only one variable; but when we deal with more complicated matrix expressions, it quickly becomes painful to deal with coordinates. A neat trick of notation is to work not with derivatives along the coordinate directions, but with derivatives in an *arbitrary* direction $(\delta x, \delta y) \in \mathbb{R}^n \times \mathbb{R}^n$:

$$\left. \frac{d}{ds} \right|_{s=0} (y + s\delta y)^T (x + s\delta x) = \delta y^T x + y^T \delta x.$$

We denote the directional derivative by $\delta(y^T x)$, giving the tidy expression

$$\delta(y^T x) = \delta y^T x + y^T \delta x.$$

This is *variational notation* for reasoning about directional (Gateaux) derivatives. It is often used in mechanics and in PDE theory and functional analysis

(where the vector spaces involved are infinite-dimensional), and I have always felt it deserves to be used more widely.

1.2 Some calculus facts

We will make frequent use of the humble product rule in this class:

$$\delta(AB) = \delta A\, B + A\, \delta B.$$

As is always the case, the order of the terms in the products is important. To differentiate a product of three terms (for example), we would have

$$\delta(ABC) = (\delta A)BC + A(\delta B)C + AB(\delta C).$$

The product rule and implicit differentiation gives us

$$0 = \delta(A^{-1}A) = \delta(A^{-1})A + A^{-1}\delta A.$$

Rearranging slightly, we have

$$\delta(A^{-1}) = -A^{-1}(\delta A)A^{-1},$$

which is again a matrix version of the familiar rule from Calculus I, differing only in that we have to be careful about the order of products. This rule also nicely illustrates the advantage of variational notation; if you are unconvinced, I invite you to write out the elements of the derivative of a matrix inverse using conventional coordinate notation!

The vector 2-norm and the Frobenius norm for matrices are convenient because the (squared) norm is a differentiable function of the entries. For the vector 2-norm, we have

$$\delta(\|x\|^2) = \delta(x^*x) = (\delta x)^*x + x^*(\delta x);$$

observing that $y^*x = (x^*y)^*$ and $z + \bar{z} = 2\Re(z)$, we have

$$\delta(\|x\|^2) = 2\Re(\delta x^* x).$$

Similarly, the Frobenius norm is associated with a dot product (the unsurprisingly-named Frobenius inner product) on all the elements of the matrix, which we can write in matrix form as

$$\langle A, B\rangle_F = \operatorname{tr}(B^*A),$$

and we therefore have

$$\delta(\|A\|_F^2) = \delta \operatorname{tr}(A^*A) = 2\Re \operatorname{tr}(\delta A^* A).$$

1.3 The 2-norm revisited

In the previous lecture, we discussed the matrix 2-norm in terms of the singular value decomposition. What if we did not know about the SVD? By the definition, we would like to maximize $\phi(v)^2 = \|Av\|^2$ subject to $\|v\|^2 = 1$. Flexing our new variational notation, let's work through the first-order condition for a maximum. To enforce the condition, we form an augmented Lagrangian

$$L(v, \mu) = \|Av\|^2 - \mu(\|v\|^2 - 1)$$

and differentiating gives us

$$\delta L = 2\Re(\delta v^*(A^*Av - \mu v)) - \delta\mu(\|v\|^2 - 1).$$

The first-order condition for a maximum or minimum is $\delta L = 0$ for all possible δv and $\delta\mu$; this gives

$$A^*Av = \mu v, \quad \|v\|^2 = 1,$$

which is an eigenvalue problem involving the Gram matrix A^*A. We will see this eigenvalue problem again — and the more general idea of the connection between eigenvalue problems and optimizing quadratic forms — later in the course.

1.4 Norms and Neumann series

We will do a great deal of operator norm manipulation this semester, almost all of which boils down to repeated use of the triangle inequality and the submultiplicative property. For now, we illustrate the point by a simple, useful example: the matrix version of the geometric series.

Suppose F is a square matrix such that $\|F\| < 1$ in some operator norm, and consider the power series

$$\sum_{j=0}^{n} F^j.$$

Note that $\|F^j\| \leq \|F\|^j$ via the submultiplicative property of induced operator norms. By the triangle inequality, the partial sums satisfy

$$(I - F)\sum_{j=0}^{n} F^j = I - F^{n+1}.$$

Hence, we have that

$$\|(I-F)\sum_{j=0}^{n} F^j - I\| \leq \|F\|^{n+1} \to 0 \text{ as } n \to \infty,$$

i.e. $I-F$ is invertible and the inverse is given by the convergent power series (the geometric series or *Neumann series*)

$$(I-F)^{-1} = \sum_{j=0}^{\infty} F^j.$$

By applying submultiplicativity and triangle inequality to the partial sums, we also find that

$$\|(I-F)^{-1}\| \leq \sum_{j=0}^{\infty} \|F\|^j = \frac{1}{1-\|F\|}.$$

Note as a consequence of the above that if $\|A^{-1}E\| < 1$ then

$$\|(A+E)^{-1}\| = \|(I+A^{-1}E)^{-1}A^{-1}\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}E\|}.$$

That is, the Neumann series gives us a sense of how a small perturbation to A can change the norm of A^{-1}.