2022-08-25

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.

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}.

2 Notions of error

The art of numerics is finding an approximation with a fast algorithm, a form that is easy to analyze, and an error bound. Given a task, we want to engineer an approximation that is good enough, and that composes well with other approximations. To make these goals precise, we need to define types of errors and error propagation, and some associated notation – which is the point of this lecture.

2.1 Absolute and relative error

Suppose $\hat{x}$ is an approximation to x. The *absolute error* is

$$e_{\text{abs}} = |\hat{x} - x|.$$

Absolute error has the same dimensions as x, and can be misleading without some context. An error of one meter per second is dramatic if x is my walking pace; if x is the speed of light, it is a very small error.

The *relative error* is a measure with a more natural sense of scale:

$$e_{\text{rel}} = \frac{|\hat{x} - x|}{|x|}.$$

Relative error is familiar in everyday life: when someone talks about an error of a few percent, or says that a given measurement is good to three significant figures, she is describing a relative error.

We sometimes estimate the relative error in approximating x by $\hat{x}$ using the relative error in approximating $\hat{x}$ by x:

$$\hat{e}_{\text{rel}} = \frac{|\hat{x} - x|}{|\hat{x}|}.$$

As long as $\hat{e}_{\text{rel}} < 1$, a little algebra gives that

$$\frac{\hat{e}_{\text{rel}}}{1 + \hat{e}_{\text{rel}}} \leq e_{\text{rel}} \leq \frac{\hat{e}_{\text{rel}}}{1 - \hat{e}_{\text{rel}}}.$$

If we know $\hat{e}_{\text{rel}}$ is much less than one, then it is a good estimate for e_{rel}. If $\hat{e}_{\text{rel}}$ is not much less than one, we know that $\hat{x}$ is a poor approximation to x. Either way, $\hat{e}_{\text{rel}}$ is often just as useful as e_{rel}, and may be easier to estimate.

Relative error makes no sense for $x = 0$, and may be too pessimistic when the property of x we care about is "small enough." A natural intermediate between absolute and relative errors is the mixed error

$$e_{\text{mixed}} = \frac{|\hat{x} - x|}{|x| + \tau}$$

where τ is some natural scale factor associated with x.

2.2 Errors beyond scalars

Absolute and relative error make sense for vectors as well as scalars. If $\|\cdot\|$ is a vector norm and $\hat{x}$ and x are vectors, then the (normwise) absolute and relative errors are

$$e_{\text{abs}} = \|\hat{x} - x\|, \qquad e_{\text{rel}} = \frac{\|\hat{x} - x\|}{\|x\|}.$$

We might also consider the componentwise absolute or relative errors

$$e_{\text{abs},i} = |\hat{x}_i - x_i| \qquad\qquad e_{\text{rel},i} = \frac{|\hat{x}_i - x_i|}{|x_i|}.$$

The two concepts are related: the maximum componentwise relative error can be computed as a normwise error in a norm defined in terms of the solution vector:

$$\max_i e_{\text{rel},i} = |||\hat{x} - x|||$$

where $|||z||| = \| \operatorname{diag}(x)^{-1} z\|_\infty$. More generally, absolute error makes sense whenever we can measure distances between the truth and the approximation; and relative error makes sense whenever we can additionally measure the size of the truth. However, there are often many possible notions of distance and size; and different ways to measure give different notions of absolute and relative error. In practice, this deserves some care.

2.3 Dimensions and scaling

The first step in analyzing many application problems is *nondimensionalization*: combining constants in the problem to obtain a small number of dimensionless constants. Examples include the aspect ratio of a rectangle, the Reynolds number in fluid mechanics[1], and so forth. There are three big reasons to nondimensionalize:

- Typically, the physics of a problem only really depends on dimensionless constants, of which there may be fewer than the number of dimensional constants. This is important for parameter studies, for example.
- For multi-dimensional problems in which the unknowns have different units, it is hard to judge an approximation error as "small" or "large," even with a (normwise) relative error estimate. But one can usually tell what is large or small in a non-dimensionalized problem.
- Many physical problems have dimensionless parameters much less than one or much greater than one, and we can approximate the physics in

[1] Or any of a dozen other named numbers in fluid mechanics. Fluid mechanics is a field that appreciates the power of dimensional analysis

these limits. Often when dimensionless constants are huge or tiny and asymptotic approximations work well, naive numerical methods work work poorly. Hence, nondimensionalization helps us choose how to analyze our problems — and a purely numerical approach may be silly.

3 Forward and backward error

We often approximate a function f by another function $\hat{f}$. For a particular x, the *forward* (absolute) error is

$$|\hat{f}(x) - f(x)|.$$

In words, forward error is the function *output*. Sometimes, though, we can think of a slightly wrong *input*:

$$\hat{f}(x) = f(\hat{x}).$$

In this case, $|x - \hat{x}|$ is called the *backward* error. An algorithm that always has small backward error is *backward stable*.

A *condition number* a tight constant relating relative output error to relative input error. For example, for the problem of evaluating a sufficiently nice function $f(x)$ where x is the input and $\hat{x} = x + h$ is a perturbed input (relative error $|h|/|x|$), the condition number $\kappa[f(x)]$ is the smallest constant such that

$$\frac{|f(x+h) - f(x)|}{|f(x)|} \leq \kappa[f(x)] \frac{|h|}{|x|} + o(|h|)$$

If f is differentiable, the condition number is

$$\kappa[f(x)] = \lim_{h \neq 0} \frac{|f(x+h) - f(x)|/|f(x)|}{|(x+h) - x|/|x|} = \frac{|f'(x)||x|}{|f(x)|}.$$

If f is Lipschitz in a neighborhood of x (locally Lipschitz), then

$$\kappa[f(x)] = \frac{M_{f(x)}|x|}{|f(x)|}.$$

where M_f is the smallest constant such that $|f(x+h) - f(x)| \leq M_f|h| + o(|h|)$. When the problem has no linear bound on the output error relative to the

input error, we sat the problem has an *infinite* condition number. An example is $x^{1/3}$ at $x = 0$.

A problem with a small condition number is called *well-conditioned*; a problem with a large condition number is *ill-conditioned*. A backward stable algorithm applied to a well-conditioned problem has a small forward error.

4 Perturbing matrix problems

To make the previous discussion concrete, suppose I want $y = Ax$, but because of a small error in A (due to measurement errors or roundoff effects), I instead compute $\hat{y} = (A + E)x$ where E is "small." The expression for the *absolute* error is trivial:

$$\|\hat{y} - y\| = \|Ex\|.$$

But I usually care more about the *relative error*.

$$\frac{\|\hat{y} - y\|}{\|y\|} = \frac{\|Ex\|}{\|y\|}.$$

If we assume that A is invertible and that we are using consistent norms (which we will usually assume), then

$$\|Ex\| = \|EA^{-1}y\| \leq \|E\|\|A^{-1}\|\|y\|,$$

which gives us

$$\frac{\|\hat{y} - y\|}{\|y\|} \leq \|A\|\|A^{-1}\|\frac{\|E\|}{\|A\|} = \kappa(A)\frac{\|E\|}{\|A\|}.$$

That is, the relative error in the output is the relative error in the input multiplied by the condition number $\kappa(A) = \|A\|\|A^{-1}\|$. Technically, this is the condition number for the problem of matrix multiplication (or solving linear systems, as we will see) with respect to a particular (consistent) norm; different problems have different condition numbers. Nonetheless, it is common to call this "the" condition number of A.

For some problems, we are given more control over the structure of the error matrix E. For example, we might suppose that A is symmetric, and ask whether we can get a tighter bound if in addition to assuming a bound on $\|E\|$, we also assume E is symmetric. In this particular case, the answer is "no" — we have the same condition number either way, at least for the 2-norm or Frobenius norm[2]. In other cases, assuming a structure to the

[2]This is left as an exercise for the student

perturbation does indeed allow us to achieve tighter bounds.

As an example of a refined bound, we consider moving from condition numbers based on small norm-wise perturbations to condition numbers based on small *element-wise* perturbations. Suppose E is elementwise small relative to A, i.e. $|E| \leq \epsilon |A|$. Suppose also that we are dealing with a norm such that $\|X\| \leq \| \, |X| \, \|$, as is true of all the norms we have seen so far. Then

$$\frac{\|\hat{y} - y\|}{\|y\|} \leq \|EA^{-1}\| \leq \| \, |A| \, |A^{-1}| \, \| \epsilon.$$

The quantity $\kappa_{\text{rel}}(A) = \| \, |A| \, |A^{-1}| \, \|$ is the *relative condition number*; it is closely related to the *Skeel condition number* which we will see in our discussion of linear systems[3]. Unlike the standard condition number, the relative condition number is invariant under column scaling of A; that is $\kappa_{\text{rel}}(AD) = \kappa_{\text{rel}}(A)$ where D is a nonsingular diagonal matrix.

What if, instead of perturbing A, we perturb x? That is, if $\hat{y} = A\hat{x}$ and $y = Ax$, what is the condition number relating $\|\hat{y} - y\|/\|y\|$ to $\|\hat{x} - x\|/\|x\|$? We note that

$$\|\hat{y} - y\| = \|A(\hat{x} - x)\| \leq \|A\| \|\hat{x} - x\|;$$

and

$$\|x\| = \|A^{-1}y\| \leq \|A^{-1}\| \|y\| \quad \Longrightarrow \quad \|y\| \geq \|A^{-1}\|^{-1} \|x\|.$$

Put together, this implies

$$\frac{\|\hat{y} - y\|}{\|y\|} \leq \|A\| \|A^{-1}\| \frac{\|\hat{x} - x\|}{\|x\|}.$$

The same condition number appears again!

[3]The Skeel condition number involves the two factors in the reverse order.