CS 4220: Numerical Analysis

Neumann series, sensitivity, conditioning

David Bindel

2026-01-30

Norms revisited

Earlier, we discussed norms, including induced norms: if A maps between two normed vector spaces $\mathcal{V}$ and $\mathcal{W}$, the *induced norm* on A is

$$\|A\|_{\mathcal{V},\mathcal{W}} = \sup_{v \neq 0} \frac{\|Av\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} = \sup_{\|v\|_{\mathcal{V}}=1} \|Av\|_{\mathcal{W}}.$$

When $\mathcal{V}$ is finite-dimensional (as it always is in this class), the unit ball $\{v \in \mathcal{V} : \|v\| = 1\}$ is compact, and $\|Av\|$ is a continuous function of v, so the supremum is actually attained. Induced norms have a number of nice properties, not the least of which are the submultiplicative properties

$$\|Av\| \leq \|A\|\|v\|$$
$$\|AB\| \leq \|A\|\|B\|.$$

The first property ($\|Av\| \leq \|A\|\|v\|$) is clear from the definition of the vector norm. The second property is almost as easy to prove:

$$\|AB\| = \max_{\|v\|=1} \|ABv\| \leq \max_{\|v\|=1} \|A\|\|Bv\| = \|A\|\|B\|.$$

The matrix norms induced when $\mathcal{V}$ and $\mathcal{W}$ are supplied with a 1-norm, 2-norm, or ∞-norm are simply called the matrix 1-norm, 2-norm, and ∞-norm. The matrix 1-norm and ∞-norm are given by

$$\|A\|_1 = \max_j \sum_i |a_{ij}|$$
$$\|A\|_\infty = \max_i \sum_j |a_{ij}|.$$

These norms are nice because they are easy to compute; the two norm is nice for other reasons, but is not easy to compute.

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}.

A matrix calculus aside

A *directional derivative* of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ in the direction u is

$$\frac{\partial f}{\partial u}(x) \equiv \left. \frac{d}{ds} \right|_{s=0} f(x + su) = f'(x)u.$$

A nice notational convention, sometimes called *variational* notation (as in "calculus of variations") is to write

$$\delta f = f'(x)\delta u,$$

where δ should be interpreted as "first order change in." You can also always rewrite these expressions as derivatives with respect to some scalar parameter s (as is done in the definition). In introductory calculus classes, this sometimes is called a total derivative or total differential, though there one usually uses d rather than δ. There is a good reason for using δ in the calculus of variations, though, so that's typically what I do.

Variational notation can tremendously simplify the calculus book-keeping for dealing with multivariate functions. For example, consider the problem of differentiating A^{-1} with respect to every element of A. I would compute this by thinking of the relation between a first-order change to A^{-1} (written $\delta[A^{-1}]$) and a corresponding first-order change to A (written δA). Using the product rule and differentiating the relation $I = A^{-1}A$, we have

$$0 = \delta[A^{-1}A] = \delta[A^{-1}]A + A^{-1}\delta A.$$

Rearranging a bit gives

$$\delta[A^{-1}] = -A^{-1}[\delta A]A^{-1}.$$

One *can* do this computation element by element, but it's harder to do it without the computation becoming horrible.

The 2-norm

The matrix 2-norm is very useful, but it is also not so straightforward to compute. Last time, we showed how to think about computing $\|A\|_2$ via the SVD. We now take a different tack, foreshadowing topics to come in the class. I will likely not talk about this in lecture, but I think it is worth mentioning in the notes.

If A is a real matrix, then we have

$$\begin{aligned}
\|A\|_2^2 &= \left(\max_{\|v\|_2=1} \|Av\|\right)^2 \\
&= \max_{\|v\|_2^2=1} \|Av\|^2 \\
&= \max_{v^Tv=1} v^TA^TAv.
\end{aligned}$$

This is a constrained optimization problem, to which we will apply the method of Lagrange multipliers: that is, we seek critical points for the functional

$$L(v,\mu) = v^TA^TAv - \mu(v^Tv - 1).$$

Differentiate in an arbitrary direction $(\delta v, \delta\mu)$ to find

$$\begin{aligned} 2\delta v^T(A^TAv - \mu v) &= 0, \\ \delta\mu(v^Tv - 1) &= 0. \end{aligned}$$

Therefore, the stationary points satisfy the eigenvalue problem

$$A^TAv = \mu v.$$

The eigenvalues of A^TA are non-negative (why?), so we will call them σ_i^2. The positive values σ_i are exactly the *singular values* of A — the diagonal elements of the matrix Σ in the singular value decomposition from last lecture — and the eigenvectors of A^TA are the right singular vectors (V).

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.

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.

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

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

We might also consider the componentwise absolute or relative errors

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

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 $\|z\|_* = \|\operatorname{diag}(x)^{-1} z\|$. 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.

Forward and backward error and conditioning

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.

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.

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 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.

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

Problems to ponder

1. Show that as long as $\hat{e}_{\text{rel}} < 1$,

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

2. Show that $A + E$ is invertible if A is invertible and $\|E\| < 1/\|A^{-1}\|$ in some operator norm.

3. In this problem, we will walk through an argument about the bound on the relative error in approximating the relative error in solving a perturbed linear system: that is, how well does $\hat{y} = (A + E)^{-1}b$ approximate $y = A^{-1}b$ in a relative error sense? We will assume throughout that $\|E\| < \epsilon$ and $\kappa(A)\epsilon < 1$.

    1. Show that $\hat{y} = (I + A^{-1}E)y$.

    2. Using Neumann series bounds, argue that

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

    3. Conclude that

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