2022-01-31

1 Matrices and mappings

A matrix represents a mapping between two vector spaces. That is, if $L : \mathcal{V} \to \mathcal{W}$ is a linear map, then the associated matrix A with respect to bases V and W satisfies $A = W^{-1}LV$. The same linear mapping corresponds to different matrices depending on the choices of basis. But matrices can reresent several other types of mappings as well. Over the course of this class, we will see several interpretations of matrices:

- **Linear maps**. A map $L : \mathcal{V} \to \mathcal{W}$ is linear if $L(x+y) = Lx + Ly$ and $L(\alpha x) = \alpha Lx$. The corresponding matrix is $A = W^{-1}LV$.

- **Linear operators**. A linear map from a space to itself ($L : \mathcal{V} \to \mathcal{V}$) is a linear operator. The corresponding (square) matrix is $A = V^{-1}LV$.

- **Bilinear forms**. A map $a : \mathcal{V} \times \mathcal{W} \to \mathbb{R}$ (or $\mathbb{C}$ for complex spaces) is bilinear if it is linear in both slots: $a(\alpha u + v, w) = \alpha a(u, w) + a(v, w)$ and $a(v, \alpha u + w) = \alpha a(v, u) + a(v, w)$. The corresponding matrix has elements $A_{ij} = a(v_i, w_j)$; if $v = Vc$ and $w = Wd$ then $a(v, w) = d^T Ac$.

  We call a bilinear form on $\mathcal{V} \times \mathcal{V}$ *symmetric* if $a(v, w) = a(w, v)$; in this case, the corresponding matrix A is also symmetric ($A = A^T$). A symmetric form and the corresponding matrix are called *positive semi-definite* if $a(v, v) \geq 0$ for all v. The form and matrix are *positive definite* if $a(v, v) > 0$ for any $v \neq 0$.

  A *skew-symmetric* matrix ($A = -A^T$) corresponds to a skew-symmetric or anti-symmetric bilinear form, i.e. $a(v, w) = -a(w, v)$.

- **Sesquilinear forms**. A map $a : \mathcal{V} \times \mathcal{W} \to \mathbb{C}$ (where $\mathcal{V}$ and $\mathcal{W}$ are complex vector spaces) is sesquilinear if it is linear in the first slot and the conjugate is linear in the second slot: $a(\alpha u + v, w) = \alpha a(u, w) + a(v, w)$ and $a(v, \alpha u + w) = \bar{\alpha} a(v, u) + a(v, w)$. The matrix has elements $A_{ij} = a(v_i, w_j)$; if $v = Vc$ and $w = Wd$ then $a(v, w) = d^* Ac$.

  We call a sesquilinear form on $\mathcal{V} \times \mathcal{V}$ *Hermitian* if $a(v, w) = a(w, v)$; in this case, the corresponding matrix A is also Hermitian ($A = A^*$). A

Hermitian form and the corresponding matrix are called *positive semi-definite* if $a(v,v) \geq 0$ for all v. The form and matrix are *positive definite* if $a(v,v) > 0$ for any $v \neq 0$.

A *skew-Hermitian* matrix ($A = -A^*$) corresponds to a skew-Hermitian or anti-Hermitian bilinear form, i.e. $a(v,w) = -a(w,v)$.

- **Quadratic forms**. A quadratic form $\phi : \mathcal{V} \to \mathbb{R}$ (or $\mathbb{C}$) is a homogeneous quadratic function on $\mathcal{V}$, i.e. $\phi(\alpha v) = |\alpha|^2 \phi(v)$ for which the map $b(v,w) = \phi(v+w) - \phi(v) - \phi(w)$ is bilinear. Any quadratic form on a finite-dimensional space can be represented as $c^* A c$ where c is the coefficient vector for some Hermitian matrix A. The formula for the elements of A given ϕ is left as an exercise.

We care about linear maps and linear operators almost everywhere, and most students come out of a first linear algebra class with some notion that these are important. But apart from very standard examples (inner products and norms), many students have only a vague notion of what a bilinear form, sesquilinear form, or quadratic form might be. Bilinear forms and sesquilinear forms show up when we discuss large-scale solvers based on projection methods. Quadratic forms are important in optimization, physics (where they often represent energy), and statistics (e.g. for understanding variance and covariance).

1.1 Matrix norms

The space of matrices forms a vector space; and, as with other vector spaces, it makes sense to talk about norms. In particular, we frequently want norms that are *consistent* with vector norms on the range and domain spaces; that is, for any w and v, we want

$$w = Av \implies \|w\| \leq \|A\|\|v\|.$$

One "obvious" consistent norm is the *Frobenius norm*,

$$\|A\|_F^2 = \sum_{i,j} a_{ij}^2.$$

Even more useful are *induced norms* (or *operator norms*)

$$\|A\| = \sup_{v \neq 0} \frac{\|Av\|}{\|v\|} = \sup_{\|v\|=1} \|Av\|.$$

The induced norms corresponding to the vector 1-norm and ∞-norm are

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \quad \text{(max column sum)}$$
$$\|A\|_\infty = \max_i \sum_j |a_{ij}| \quad \text{(max row sum)}$$

The norm induced by the vector Euclidean norm (variously called the matrix 2-norm or the spectral norm) is more complicated.

The Frobenius norm and the matrix 2-norm are both *orthogonally invariant* (or *unitarily invariant* in a complex vector space. That is, if Q is a square matrix with $Q^* = Q^{-1}$ (an orthogonal or unitary matrix) of the appropriate dimensions

$$\|QA\|_F = \|A\|_F, \qquad \|AQ\|_F = \|A\|_F,$$
$$\|QA\|_2 = \|A\|_2, \qquad \|AQ\|_2 = \|A\|_2.$$

This property will turn out to be frequently useful throughout the course.

1.2 Decompositions and canonical forms

Matrix decompositions (also known as *matrix factorizations*) are central to numerical linear algebra. We will get to know six such factorizations well:

- $PA = LU$ (a.k.a. Gaussian elimination). Here L is unit lower triangular (triangular with 1 along the main diagonal), U is upper triangular, and P is a permutation matrix.
- $A = LL^*$ (a.k.a. Cholesky factorization). Here A is Hermitian and positive definite, and L is a lower triangular matrix.
- $A = QR$ (a.k.a. QR decomposition). Here Q has orthonormal columns and R is upper triangular. If we think of the columns of A as a basis, QR decomposition corresponds to the Gram-Schmidt orthogonalization process you have likely seen in the past (though we rarely compute with Gram-Schmidt).
- $A = U\Sigma V^*$ (a.k.a. the singular value decomposition or SVD). Here U and V have orthonormal columns and Σ is diagonal with non-negative entries.

- $A = Q\Lambda Q^*$ (a.k.a. symmetric eigendecomposition). Here A is Hermitian (symmetric in the real case), Q is orthogonal or unitary, and Λ is a diagonal matrix with real numbers on the diagonal.
- $A = QTQ^*$ (a.k.a. Schur form). Here A is a square matrix, Q is orthogonal or unitary, and T is upper triangular (or nearly so).

The last three of these decompositions correspond to *canonical forms* for abstract operators. That is, we can view these decompositions as finding bases in which the matrix representation of some operator or form is particularly simple. More particularly:

- **SVD**: For any linear mapping $L : \mathcal{V} \to \mathcal{W}$, there are orthonormal bases for the two spaces such that the corresponding matrix is diagonal
- **Symmetric eigendecomposition**: For any Hermitian sesquilinear map on an inner product space, there is an orthonormal basis for the space such that the matrix representation is diagonal.
- **Schur form**: For any linear operator $L : \mathcal{V} \to \mathcal{V}$, there is an orthonormal basis for the space such that the matrix representation is upper triangular. Equivalently, if $\{u_1, \ldots, u_n\}$ is the basis in question, then $\mathrm{sp}(\{u_j\}_{j=1}^k)$ is an *invariant subspace* for each $1 \leq k \leq n$.

The Schur form turns out to be better for numerical work than the Jordan canonical form that you should have seen in an earlier class. We will discuss this in more detail when we discuss eigenvalue problems.

1.3 The SVD and the 2-norm

The singular value decomposition is useful for a variety of reasons; we close off the lecture by showing one such use.

Suppose $A = U\Sigma V^*$ is the singular value decomposition of some matrix. Using orthogonal invariance (unitary invariance) of the 2-norm, we have

$$\|A\|_2 = \|U^*AV\|_2 = \|\Sigma_2\|,$$

i.e.

$$\|A\|_2 = \max_{\|v\|^2=1} \frac{\sum_j \sigma_j |v_j|^2}{\sum |v_j|^2}.$$

That is, the spectral norm is the largest weighted average of the singular values, which is the same as just the largest singular value.

The small singular values also have a meaning. If A is a square, invertible matrix then

$$\|A^{-1}\|_2 = \|V\Sigma^{-1}U^*\|_2 = \|\Sigma_{-1}\|_2,$$

i.e. $\|A^{-1}|_2$ is the inverse of the smallest singular value of A.

The smallest singular value of a nonsingular matrix A can also be interpreted as the "distance to singularity": if σ_n is the smallest singular value of A, then there is a matrix E such that $\|E\|_2 = \sigma_n$ and $A + E$ is singular; and there is no such matrix with smaller norm.

These facts about the singular value decomposition are worth pondering, as they will be particularly useful in the next lecture when we ponder sensitivity and conditioning.

2 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.

2.1 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}.

3 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.

3.1 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.

3.2 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

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

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.

3.3 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.

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.

5 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.

6 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 $\kappa(A)\epsilon < 1$.
   (a) Show that $\hat{y} = (I + A^{-1}E)y$.

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

(b) Using Neumann series bounds, argue that

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

(c) Conclude that

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