Week 3: Wednesday, Feb 8

Spaces and bases

I have two favorite vector spaces[1]: $\mathbb{R}^n$ and the space $\mathcal{P}_d$ of polynomials of degree at most d. For $\mathbb{R}^n$, we have a *canonical basis*:

$$\mathbb{R}^n = \operatorname{span}\{e_1, e_2, \ldots, e_n\},$$

where e_k is the kth column of the identity matrix. This basis is frequently convenient both for analysis and for computation. For $\mathcal{P}_d$, an obvious-seeming choice of basis is the *power basis*:

$$\mathcal{P}_d = \operatorname{span}\{1, x, x^2, \ldots, x^d\}.$$

But this obvious-looking choice turns out to often be terrible for computation. Why? The short version is that powers of x aren't all that strongly linearly dependent, but we need to develop some more concepts before that short description will make much sense.

The *range space* of a matrix or a linear map A is just the set of vectors y that can be written in the form $y = Ax$. If A is *full (column) rank*, then the columns of A are linearly independent, and they form a basis for the range space. Otherwise, A is *rank-deficient*, and there is a non-trivial *null space* consisting of vectors x such that $Ax = 0$.

Rank deficiency is a delicate property[2]. For example, consider the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

This matrix is rank deficient, but the matrix

$$\hat{A} = \begin{bmatrix} 1+\delta & 1 \\ 1 & 1 \end{bmatrix}.$$

is not rank deficient for any $\delta \neq 0$. Technically, the columns of $\hat{A}$ form a basis for $\mathbb{R}^2$, but we should be disturbed by the fact that $\hat{A}$ is so close to a singular matrix. We will return to this point in some detail next week.

[1] This is a fib, but not by too much.

[2] Technically, we should probably say that rank deficiency is *non-generic* rather than "delicate."

Norm!

In order to talk sensibly about a matrix being "close to" singular or a basis being "close to" linear dependence, we need the right language.

First, we need the concept of a *norm*, which is a measure of the length of a vector. A norm is a function from a vector space into the real numbers with three properties

1. Positive definiteness: $\|x\| > 0$ when $x \neq 0$ and $\|0\| = 0$.
2. Homogeneity: $\|\alpha x\| = |\alpha|\|x\|$.
3. Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.

One of the most popular norms is the Euclidean norm (or 2-norm):

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{x^T x}.$$

We will also use the 1-norm and the ∞-norm (*a.k.a.* the max norm or the Manhattan norm):

$$\|x\|_1 = \sum_i |x_i|.$$
$$\|x\|_\infty = \max_i |x_i|$$

Second, we need a way to relate the norm of an input to the norm of an output. We do this with matrix norms. Matrices of a given size form a vector space, so in one way a matrix norm is just another type of vector norm. However, the most useful matrix norms are *consistent* with vector norms on their domain and range spaces, i.e. for all vectors x in the domain,

$$\|Ax\| \leq \|A\|\|x\|.$$

Given norms for vector spaces, a commonly-used consistent norm is the *induced* norm (operator norm):

$$\|A\| \equiv \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$

The matrix 1-norm and the matrix ∞-norm (the norms induced by the vector 1-norm and vector ∞-norm) are:

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

If we think of a vector as a special case of an n-by-1 matrix, the vector 1-norm matches the matrix 1-norm, and likewise with the ∞-norm. This is how I remember which one is the max row sum and which is the max column sum!

The matrix 2-norm is very useful, but it is actually much harder to compute than the 1-norm or the ∞-norm. There is a related matrix norm, the Frobenius norm, which is much easier to compute:

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

The Frobenius norm is consistent, but it is not an operator norm[3]

MATLAB allows us to compute all the vector and matrix norms describe above with the `norm` command. For example, `norm(A, 'fro')` computes the Frobenius norm of a matrix A, while `norm(x,1)` computes the 1-norm of a vector x. The default norm, which we get if we just write `norm(A)` or `norm(x)`, is the Euclidean vector norm (*a.k.a.* the 2-norm) and the corresponding operator norm.

The ideas of vector norms and operator norms make sense on spaces other than $\mathbb{R}^n$, too. For example, one choice of norms for $\mathcal{P}_d$ is

$$\|p\|_{L^2([-1,1])} = \sqrt{\int_{-1}^{1} p(x)^2 \, dx}.$$

You will note that this looks an awful lot like the standard Euclidean norm; we also have analogues of the 1-norm and the ∞-norm in this case. The norms for spaces of functions (like $\mathcal{P}_d$) are actually a more interesting topic than the norms of $\mathbb{R}^n$, but an extended discussion is (lamentably) beyond the scope of what I can reasonably fit into this course.

[3] The first half of this sentence is basically Cauchy-Schwarz; the second half of the sentence can be seen by looking at $\|I\|_F$. If you don't understand this footnote, no worries.

Inner products

Norms are the tools we need to measure lengths and distances. *Inner products* are the tools we need to measure angles. In general, an inner product satisfies three axioms:

- *Positive definiteness*: $\langle u, u \rangle \geq 0$, with equality iff $u = 0$.
- *Symmetry*: $\langle u, v \rangle = \overline{\langle v, u \rangle}$
- *Linearity*: $\langle \alpha u, v \rangle = \alpha \langle u, v \rangle$ and $\langle u_1 + u_2, v \rangle = \langle u_1, v \rangle + \langle u_2, v \rangle$.

For every inner product, we have an associated norm: $\|u\| = \sqrt{\langle u, u \rangle}$. An important identity relating the inner product to the norm is the *Cauchy-Schwartz* inequality:

$$\langle u, v \rangle \leq \|u\| \|v\|.$$

Equality holds only if u and v are parallel. Vectors u and v are *orthogonal* if $\langle u, v \rangle = 0$. In general, the angle α between nonzero vectors u and v is *defined* by the relation

$$\cos(\alpha) = \frac{\langle u, v \rangle}{\|u\| \|v\|}.$$

If x and y are in $\mathbb{R}^n$, the standard inner product is:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i.$$

We say vectors $u_1, u_2, \ldots, u_k$ are *orthonormal* if they mutually orthogonal and have unit Euclidean length, i.e.

$$\langle u_i, u_j \rangle = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & \text{otherwise.} \end{cases}$$

Somewhat oddly, though, we define an *orthogonal matrix* to be a square matrix whose columns are orthonormal (i.e. a matrix Q such that $Q^T Q = I$). When we say a matrix is orthogonal, we usually really mean "orthogonal with respect to the standard inner product on $\mathbb{R}^n$"; if the matrix is orthogonal with respect to some other inner product, we say so explicitly.

One very useful property of orthogonal matrices is that they *preserve Euclidean length*. That is, if Q is orthogonal, then

$$\|Qx\|^2 = (Qx)^T (Qx) = x^T Q^T Q x = x^T x = \|x\|^2.$$

From time to time, I may talk about "unitary operations"; if I do, I generally mean linear maps that have this property of preserving Euclidean length[4]

Of course, other spaces can also have useful inner products. For example, a standard choice of inner products for $\mathcal{P}_d$ is

$$\langle p, q \rangle_{L^2([-1,1])} = \int_{-1}^{1} p(x)q(x)\,dx.$$

The power basis $\{1, x, x^2, \ldots, x^d\}$ is decidedly *not* orthonormal with respect to this inner product. On the other hand the *Legendre polynomials*, which play a critical role in the theory of Gaussian integration, do form an orthogonal basis for $\mathcal{P}_d$ with respect to this inner product.

Symmetric matrices and quadratic forms

The multi-dimensional version of Taylor's theorem says that we can write any sufficiently nice function from $\mathbb{R}^n \rightarrow \mathbb{R}$ as

$$f(x_0 + z) = f(x_0) + \sum_i \frac{\partial f}{\partial x_i} z_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j} z_i z_j + O(\|z\|^3).$$

We sometimes write this more concisely as

$$f(x_0 + z) = f(x_0) + \nabla f(x_0)^T z + \frac{1}{2} z^T H_f(x_0) z + O(\|z\|^3),$$

where the *Hessian matrix* $H_f(x_0)$ has entries which are second partials of f at x_0. Still assuming that f is nice, we have that

$$(H_f(x_0))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} = (H_f(x_0))_{ji};$$

that is, the Hessian matrix is *symmetric.*

A *quadratic form* on $\mathbb{R}^n$ is function of the form

$$\phi(x) = x^T A x.$$

[4] I'll expect you to know what an orthogonal matrix is going forward, but if I ever say "unitary operation" and you forget what I mean, just ask me.

We typically assume A is symmetric, since only the symmetric part of the matrix matters.[5] Quadratic forms show up frequently throughout applied mathematics, partly because second-order Taylor expansions show up frequently. Symmetric matrices also show up more-or-less constantly; and when they do, there is often a quadratic form lurking behind the scenes.

A symmetric matrix A is *positive definite* if the corresponding quadratic form $\phi(x) = x^T A x$ is positive definite — that is, $\phi(x) \geq 0$ for all x, with equality only at $x = 0$. You've likely seen the notion of positive definiteness before in multivariable calculus: if a function f has a critical point at x_0 and $H_f(x_0)$ is positive definite, then x_0 is a local minimum. You've also seen the notion of positive definiteness earlier in these notes, since the quadratic form associated with an inner product ($\|u\|^2 = \langle u, u \rangle$) must be positive definite. Matrices that are symmetric and positive definite occur so frequently in numerical linear algebra that we often just call them SPD matrices[6].

Quadratic forms are characterized by the fact that they are quadratic; that is, $\phi(\alpha x) = \alpha^2 \phi(x)$. It is sometimes convenient to get rid of the effects of scaling vectors, and so we define the *Rayleigh quotient*:

$$\rho_A(x) = \frac{x^T A x}{x^T x}.$$

It is interesting to differentiate $\rho_A(x)$ to try to find critical points:

$$\begin{aligned} \frac{d}{dt}\rho_A(x + tw) &= \frac{w^T A x + x^T A w}{x^T x} - \frac{(x^T A x)(w^T x + x^T w)}{(x^T x)^2} \\ &= \frac{2w^T}{x^T A x}\left(Ax - \rho_A(x) x\right). \end{aligned}$$

At a critical point, where all the directional derivatives are zero, we have

$$Ax = \rho_A(x) x,$$

i.e. x is an *eigenvector* and $\rho_A(x)$ is an *eigenvalue*. This connection between eigenvalues of symmetric matrices and ratios of quadratic forms is immensely powerful. For example, we can use it to characterize the operator two-norm

$$\|A\|_2^2 = \max_{x \neq 0} \frac{\|Ax\|^2}{\|x\|^2} = \max_{x \neq 0} \frac{x^T A^T A x}{x^T x} = \lambda_{\max}(A^T A)$$

[5]The symmetric part of a general matrix A is $(A + A^T)/2$.

[6]Abbreviations are our way of stalling RSI. Why do you think CS has so many TLAs?

The other eigenvalues of $A^T A$ (the squared *singular values*) are also sometimes handy, and we'll talk about them later.

We can also look at the eigenvalues of a symmetric matrix A to determine whether the corresponding quadratic form is positive definite (all eigenvalues of A positive), negative definite (all eigenvalues of A negative), or indefinite.

Problems to ponder

1. We said earlier that
$$\|A\| \equiv \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|.$$
Why is the equality true?

2. What are the range and null space of $\frac{d}{dx}$ viewed as a linear operator acting on $\mathcal{P}_d$? In terms of the power basis, how might you write $\frac{d}{dx}$ as a matrix?

3. Using the inner product $\langle \cdot, \cdot \rangle_{L^2([-1,1])}$, what is the angle between the monomials x^j and x^k?

4. The Cauchy-Schwartz inequality says
$$\langle u, v \rangle \leq \|u\|\|v\|.$$
The easiest way I know to prove Cauchy-Schwartz is to write
$$\phi(t) = \langle u + tv, u + tv \rangle \geq 0,$$
then use the properties of inner products to write $\phi(t)$ as a quadratic function in t with coefficients given in terms of $\|u\|^2$, $\|v\|^2$, and $\langle u, v \rangle$. Do this expansion, and write the discriminant of the resulting quadratic. This discriminant must be non-positive in order for $\phi(t)$ to be non-negative for all values of t; using this fact, show that Cauchy-Schwartz must hold.

5. Given matrices $X, Y \in \mathbb{R}^{m \times n}$, we define the *Frobenius inner product* to be
$$\langle X, Y \rangle = \operatorname{tr}(X^T Y),$$
where $\operatorname{tr}(A)$ is the sum of the diagonal elements of A. Argue that this is an inner product, and that the associated norm is the Frobenius norm.

6. Show that when we have a norm induced by an inner product,

$$(\|u+v\|^2 - \|u-v\|^2)/4 = \langle u, v \rangle$$

7. Show that the operation $p(x) \mapsto p(-x)$ is unitary for $\mathcal{P}_d$ with the inner product $L^2([-1,1])$.

8. Show that if A is an SPD matrix, then

$$\langle x, y \rangle_A = x^T A y$$

   is a valid inner product (sometimes called an *energy inner product*).

9. Assuming A is symmetric, define

$$\psi(x) = \left(\frac{1}{2} x^T A x - x^T b \right).$$

   Give an expression for the directional derivatives

$$\frac{d}{dt} \psi(x + tu).$$

   What equation must be satisfied at a critical point (i.e. a point where all the directional derivatives are zero)?