Contents

1. Linear Algebra: theory and conditioning

References:

- D. Bindel's and J. Goodman's book "Principles of Scientific Computing", Chapter 4.
- J. Demmel, "Applied Numerical Linear Algebra", Section 1.7 (vector and matrix norms).

Linear algebra is one of the most important tools of modern computational science. In recent years, the importance of numerical linear algebra has increased due to the need to solve large-scale problems arising in data science. For example, numerous personal recommendations that you encounter in such services as Netflix, Amazon, etc, are obtained for you by solving certain large scale optimization problems by algorithms heavy on the use of linear algebra. New methods for solving large-scale linear algebra problems have been developed in recent years. These include, e.g., the butterfly algorithm for fast Fourier transform, fast direct algorithms for solving structured linear systems, etc.

The operations of linear algebra include but not limited to:

- solving linear systems of algebraic equations;
- finding subspaces;
- matrix factorization (PLU, QR, SVD, CUR, etc);
- solving least squares problems;
- computing eigenvectors and eigenvalues.

In this section, we will go over the aspect of linear algebra that you should know as a user of linear algebra software: basic concepts, basic theory, and conditioning. The last item is extremely important as you should be aware of what can go wrong when you are using some standard linear algebra operations.

There are publicly available linear algebra libraries that you are strongly encouraged to use: `clapack` (C/C++), `lapack` (Fortran). Matlab contains excellent linear algebra commands for both dense and sparse matrices.

Standard linear algebra algorithms are *backward stable*. This means that the output of any standard linear algebra algorithm is as accurate as the *condition number* for the problem allows. Recall that

- an algorithm is backward stable if its output is the exact answer for a slightly perturbed input;
- the condition number for the problem is the strict upper bound for the ratio of the relative error in the output to the relative error in the input that caused it.

This means, in particular, that the error produced by a backward stable algorithm can be large if the condition number of the problem being solved is large.

We start with reviewing basic concepts of linear algebra.

1.1. **Vector spaces.** Typically we are happy with the results of any numerical algorithm if the produces error is small. If the error is multi- or infinite dimensional, in order to say that it is small, we need some reasonable way to convert it to a single nonnegative number and compare it with some threshold. That's why we need the concept of norm.

A norm is defined as a function on vector spaces. Let's recall what it is.

Definition 1. *A vector space V is a set closed with respect to the operations of addition "$+$": $V \times V \to V$, and multiplication by a scalar "$\cdot\alpha$": $V \to V$ The operations satisfy the following properties.*

(1) $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$,
(2) $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$,
(3) $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$,
(4) $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$,
(5) there is $0 \in V$ s.t. $\mathbf{a} + \mathbf{0} = \mathbf{a}$ for any $\mathbf{a} \in V$,
(6) for any $\mathbf{a} \in V$ there is $(-\mathbf{a}) \in V$ s.t. $\mathbf{a} + (-\mathbf{a}) = 0$,
(7) $\alpha(\beta\mathbf{a}) = (\alpha\beta)\mathbf{a}$,
(8) $1\mathbf{a} = \mathbf{a}$ for any $\mathbf{a} \in V$.

Exercise Prove that for any $\mathbf{a} \in V$ $0\mathbf{a} = \mathbf{0}$ where $0 \in \mathbb{R}$ while $\mathbf{0} \in V$.

Below we remind some basic concepts. Please read Sections 4.2.1 and 4.2.2 in Bindel and Goodman for more details.

- A *subspace* W of a vector space V is a subset of V that is a vector space itself with respect to the same operations as in V, i.e., W is closed under addition and scalar multiplication: for any $w_1, w_2 \in W$ and $\alpha \in \mathbb{R}$ or $\mathbb{C}$, $w_1 + W_2 \in W$ and $\alpha w_1 \in W$. Therefore, to check is W is a subspace, it suffices to check if it closed under addition

and scalar multiplication. The properties of the operations are inherited for those in V.

- The span of vectors v_1, ..., v_n in V is the set of their all possible linear combinations.
- We say that vectors v_1, ..., v_n are *linearly independent* if any their zero linear combination implies that all of its coefficients are zero.
- A *basis* of V is a subset of vectors $\{b_i\}_{i\in\mathcal{I}}$ such that:
  (1) any $v \in V$ can be represented as
  $$v = \sum_{i\in\mathcal{I}} \alpha_i b_i,$$
  (2) and the $\{b_i\}_{i\in\mathcal{I}}$ is minimal in the sense such that for any $m \in \mathcal{I}$ one can find $v \in V$ such that
  $$v - \sum_{i\in\mathcal{I}\setminus m} \alpha_i b_i \neq 0$$
  for any set of values of α_i, $i \in \mathcal{I}\setminus\{m\}$.

  Recall a theorem in linear algebra saying that if there is a basis in V $\{b_i\}_{i=1}^n$, then any other basis in V also has n vectors.
- If the number of vectors in a basis of V is finite, this number is called the *dimension* of V. Otherwise, the vector space is infinitely dimensional.
- A *linear transformation* or a linear map for a vector space V to a vector space W is a map $L : V \to W$ such that for any $v_1, v_2 \in V$ and any $\alpha \in \mathbb{R}$ or $\mathbb{C}$
  $$L(v_1 + v_2) = L(v_1) + L(v_2) \quad \text{and} \quad L(\alpha v_1) = \alpha L(v_1).$$
  Let $\mathcal{B} = \{b_i\}$ be a basis in V and $\mathcal{E} = \{e_i\}$ be a basis in W. Then by linearity we have:
  $$L(v) = L\left(\sum_j v_j b_j\right) = \sum_j v_j L(b_j) = \sum_j v_j \sum_i a_{ij} e_i \quad \text{where} \quad L(b_j) = \sum_i a_{ij} e_i.$$
  Therefore, we can define the matrix of the linear transformation
  $$A =_{\mathcal{E}} [L]_{\mathcal{B}} = (a_{ij}).$$
  Its columns are the images of the basis vectors in V written in the basis in W.
- A matrix product AB is defined if and only if the number of columns in A is equal the number of rows in B. The matrix product AB corresponds to a composition of linear transformations with matrices A and B. Matrix multiplication is associative but not commutative.
- For a matrix $A = (a_{ij})$ the *transpose* is defined by $A^\top := (a_{ji})$. If A has complex entries, than its *adjoint* is defined as its transpose with complex conjugation: $A^* := (\bar{a}_{ji})$.

Now let us list some examples illustrating these concepts.

Example (1) $\mathbb{R}^n$ is an n-dimensional vector space. Its standard basis is $\{e_i\}$ where e_i is a vector with entry 1 at the ith place and the rest of entries being zeros. Its

subset of vectors satisfying $\sum_{i=1}^n a_i = 0$ is an $n-1$-dimensional subspace, while the subset of vectors satisfying $\sum_{i=1}^n a_i = 1$ is not a subspace as it is not closed under addition and scalar multiplication.

(2) The set of polynomials of degree less or equal than n denoted by $\mathcal{P}_n$ is an $(n+1)$-dimensional vector space. One basis in it is the set

$$\mathcal{X} := \{1, x, \dots, x^n\}.$$

(3) An example of linear transformation from $\mathcal{P}_n$ to $\mathcal{P}_{n-1}$ is the differentiation:

$$\frac{d}{dx} : \mathcal{P}_n \to \mathcal{P}_{n-1}.$$

Its matrix in the basis $\mathcal{X}$ is

$$D_{\mathcal{X}} := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2 & \dots & \\ & & & \ddots & \\ 0 & \dots & & 0 & n \end{bmatrix}.$$

If we pick another basis, for example, Chebyshev's basis, the differentiation matrix will be different.

(4) Example of an infinite-dimensional space is the space of all polynomials, the space of all continuous functions on an interval $[a, b]$, the space of all continuous functions on $[a, b]$ satisfying the homogeneous boundary conditions $f(a) = f(b) = 0$, etc.

1.2. **Vector norms.**

Definition 2. *Norm is a function defined on a vector space V:*

$$\mathcal{N} : V \longrightarrow \overline{\mathbb{R}}_+ \equiv [0, +\infty]$$

such that

(1) $\|\mathbf{a}\| \geq 0$, $\|\mathbf{a}\| = 0$ iff $\mathbf{a} = \mathbf{0}$,
(2) $\|\alpha\mathbf{a}\| = |\alpha|\|\mathbf{a}\|$,
(3) $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$.

Example The space of continuous functions on the interval $[a, b]$ with the maximum norm

$$V = C([a, b]), \ \|f\| = \sup_{[a,b]} |f(x)|.$$

If the interval is finite, $\|f\| = \max_{[a,b]} |f(x)|$.

Example The space of continuous functions on the interval $[a, b]$ with the maximum norm

$$V = L_p([a, b]), \ \|f\| = \left(\int_a^b |f(x)|^p dx \right)^{1/p}.$$

Example The space $V = l_p$ of all sequences $\{a_k\}_{k=1}^{\infty}$ such that

$$\|\{a\}\| := \left(\sum_{k=1}^{\infty} |a_k|^p\right)^{1/p} < \infty.$$

In particular, l_1 is the space of all absolutely convergent sequences as

$$|a| := \sum_{k=1}^{\infty} |a_k| < \infty.$$

Example The space $V = l_\infty$ of all sequences $\{a_k\}_{k=1}^{\infty}$ such that

$$\|\{a\}\| := \sup_k |a_k| < \infty.$$

In other words, l_∞ is the space of all bounded sequences.

The concept of orthogonality is generalized to vector spaces via the notion of the inner product.

Definition 3. *An inner product is a function* $(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ *or* $\mathbb{C}$ *satisfying*

(1) $(\mathbf{a}, \mathbf{a}) \geq 0$, $(\mathbf{a}, \mathbf{a}) = 0$ iff $\mathbf{a} = \mathbf{0}$,
(2) $(\mathbf{a}, \mathbf{b}) = \overline{(\mathbf{b}, \mathbf{a})}$,
(3) $(\mathbf{a}, \mathbf{b} + \mathbf{c}) = (\mathbf{a}, \mathbf{b}) + (\mathbf{a}, \mathbf{c})$,
(4) $(\alpha\mathbf{a}, \mathbf{b}) = \alpha(\mathbf{a}, \mathbf{b})$.

The norm induced by an inner product is given by $\|f\| = \sqrt{(f, f)}$. The norms that are associated with inner products are especially important.

Example (1)

$$f, g \in L_2([a, b]),\ (f, g) = \int_a^b f(x)\overline{g(x)}dx.$$

(2) Chebyshev inner product.

$$f, g \in C([-1, 1]),\ (f, g) = \int_a^b \frac{f(x)g(x)}{\sqrt{1 - x^2}}dx.$$

Suppose we are looking at the error $e(x) = f(x) - p(x)$ where f is a given function and p is its approximation. The Chebyshev norm puts more weight to the ends of the interval, i.e., the error near the ends of the interval contributes more to the norm than the error near its midpoint.

(3) Hermite inner product.

$$f, g \in C([-\infty, \infty]),\ (f, g) = \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2}dx.$$

Suppose we are looking at the error $e(x) = f(x) - p(x)$ where f is a given function and p is its approximation. Only the error around the origin will contribute significantly to the norm.

1.3. **Matrix norm.**

Definition 4. *The norm of a matrix associated with the vector norm $\|\cdot\|$ is defined as*

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

The geometric sense of the matrix norm is the maximal elongation of a unit vector as a result of the corresponding linear transformation.

Example Let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

(1) Find $\|A\|_1$ associated with the vector 1-norm in $\mathbb{R}^2$: $\|x\|_1 = |x_1| + |x_2|$.

Solution

$$Ax = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_2 \end{bmatrix}.$$

On one hand,

$$\|A\|_1 = \max_{|x_1|+|x_2|=1} (|x_1 + x_2| + |x_2|) \leq \max_{|x_1|+|x_2|=1} (|x_1| + |x_2| + |x_2|) = 1 + \max_{|x_1|+|x_2|=1} |x_2| = 2.$$

On the other hand, $\|Ax\|_1 = 2$ if $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Hence

$$\|A\|_1 = 2.$$

(2) Find $\|A\|_\infty$ associated with the vector max-norm in $\mathbb{R}^2$: $\|x\|_\infty = \max\{|x_1|, |x_2|\}$.

Solution

On one hand,

$$\|A\|_\infty = \max_{\{|x_1|,|x_2|\}=1} \max\{(|x_1 + x_2|, |x_2|\} \leq \max_{\{|x_1|,|x_2|\}=1} \max\{(|x_1| + |x_2|, |x_2|\} \leq 2.$$

On the other hand, $\|Ax\|_\infty = 2$ if $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Hence

$$\|A\|_\infty = 2.$$

(3) Find $\|A\|_2$ associated with the 2-vector norm in $\mathbb{R}^2$: $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$.

Solution

On one hand,

$$\|A\|_2 = \max_{\sqrt{x_1^2+x_2^2}=1} \sqrt{|x_1 + x_2|^2 + |x_2|^2}.$$

Since $\sqrt{x_1^2 + x_2^2} = 1$, x_1 and x_2 are cosine and sine of some angle $t \in [-\pi, \pi]$. Therefore, we can write

$$\|A\|_2 = \max_{t\in[-\pi,\pi]} \sqrt{(\cos t + \sin t)^2 + \sin^2 t} = \max_{t\in[-\pi,\pi]} \sqrt{1 + \sin 2t + \tfrac{1}{2}(1 - \cos 2t)}.$$

Here we used the trigonometric identities

$$2\sin t\cos t = \sin 2t \quad \text{and} \quad \sin^2 t = \tfrac{1}{2}(1-\cos 2t).$$

To find the maximum we differentiate the expression under the square root and set it to zero:

$$\frac{d}{dt}\left(1+\sin 2t+\tfrac{1}{2}(1-\cos 2t)\right) = 2\cos 2t + \sin 2t = 0.$$

Hence $\tan 2t = -2$. This corresponds either to $\cos 2t = -1/\sqrt{5}$ and $\sin 2t = 2/\sqrt{5}$, or to $\cos 2t = 1/\sqrt{5}$ and $\sin 2t = -2/\sqrt{5}$. The first pair gives maximum while the second pair gives minimum of the expression under the square root. Hence

$$\|A\|_2 = \sqrt{\tfrac{3}{2}+\tfrac{2}{\sqrt{5}}+\tfrac{1}{2\sqrt{5}}} = \sqrt{\frac{3\sqrt{5}+5}{2\sqrt{5}}} = \sqrt{\frac{3+\sqrt{5}}{2}} = \frac{1+\sqrt{5}}{2}.$$

Exercise Let $A=(a_{ij})$ be an $m\times n$ matrix, $m\geq n$. Show that then:

(1) For the l_1-norm,

$$\|A\|_1 = \max_j \sum_i |a_{ij}|,$$

i.e., the maximal column sum of absolute values.

(2) For the max-norm or l_∞-norm

$$\|A\|_{\max} = \max_i \sum_j |a_{ij}|,$$

i.e., the maximal row sum of absolute values

1.4. **Eigenvalues and eigenvectors.** Finding eigenvalues and eigenvectors is often very useful in many different contexts. For example, the general analytic solution to a linear system of ODEs $\dot{x} = Ax$ is often written in terms of eigenvalues and eigenvectors of A. The 2-norm of A is expressed in terms of eigenvalues of $A^\top A$.

1.4.1. *Diagonalizable matrices.* Recall that an $n\times n$ matrix A is called *diagonalizable* if it has n linearly independent eigenvectors. In this case, A can we written as

$$A = R\Lambda R^{-1} \equiv R\Lambda L = \begin{bmatrix} r_1 & r_2 & \cdots & r_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} l_1 & \rightarrow \\ l_2 & \rightarrow \\ \vdots & \\ l_n & \rightarrow \end{bmatrix}. \tag{2}$$

The columns of R are the right eigenvectors of A. They satisfy:

$$Ar_j = \lambda_j r_j.$$

The rows of $L := R^{-1}$ are the left eigenvectors of A satisfying

$$l_j A = \lambda_j l_j.$$

Even if A is real, eigenvectors and eigenvalues do not need to be real. they are complex in the general case.

1.4.2. *Symmetric matrices.* In the special case where A is real and symmetric, there always exists an orthonormal basis of real eigenvectors, the eigenvalues are real, and the eigenvectors corresponding to distinct eigenvalues are orthogonal. Let us show this. First note that if λ is an eigenvalue, and r is the corresponding unit right eigenvector, then $r^* := \bar{r}^\top$ is the left eigenvector for $\bar{\lambda}$. Indeed, since A is real and symmetric, we have:

$$Ar = \lambda r, \text{ hence } (Ar)^* = (\lambda r)^*, \text{ i.e. } r^*A = \bar{\lambda} r^*,$$

which shows that r^* is the left eigenvector for $\bar{\lambda}$. Now we consider the number

$$r^*Ar = r^*Ar = \lambda r^* r = \lambda \|r\| = \lambda.$$

On the other , applying A to r^*, we get

$$r^*Ar = r^*Ar = \bar{\lambda} r^* r = \bar{\lambda} \|r\| = \bar{\lambda}.$$

Therefore, $\lambda = \bar{\lambda}$, i.e., λ is real. Now we show that eigenvectors corresponding to distinct eigenvalues are orthogonal. Let $Ar_1 = \lambda_1 r_1$ and $Ar_2 = \lambda_2 r_2$ with $\lambda_1 \neq \lambda_2$. Then

$$r_1^* A r_2 = \lambda_1 r_1^* r_2 = \lambda_2 r_1^* r_2.$$

Since $\lambda_1 \neq \lambda_2$, $r_1^* r_2$ must be zero. Hence r_1 and r_2 are orthogonal. Note that we always can pick real eigenvectors for real eigenvalues of a real symmetric matrix.

Exercise Let $A = (a_{ij})$ be an $m \times n$ matrix, $m \geq n$. Show that then for the vector l_2-norm,

$$\|A\|_2 = \sqrt{\rho(A^\top A)}.$$

Solution Recall that the vector 2-norm is given by $\|x\|_2 = \sqrt{x^\top x}$. Using this we get

$$\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2 = \max_{x^\top x = 1} \sqrt{x^\top A^\top A x} = \max_{x^\top x = 1} \sqrt{x^\top A^2 x}.$$

Since $A^\top A$ is symmetric, its eigen-decomposition is given by

$$A^\top A = U \Lambda U^\top,$$

where U is an orthogonal matrix (i.e., $U^\top U = U U^\top = I$, or $U^\top = U^{-1}$) whose columns are the eigenvectors of A, and Λ is a diagonal matrix whose diagonal entries are the corresponding eigenvalues. Using this we continue:

$$\|A\|_2 = \max_{x^\top x = 1} \sqrt{x^\top U \Lambda U^\top x} = \max_{x^\top x = 1} \sqrt{(U^\top x)^\top \Lambda (U^\top x)}.$$

Now we note that

$$\|x\|_2 = \|U^\top x\|_2$$

because

$$\|x\|_2^2 = x^\top x = x^\top U U^\top x = (U^\top x)^\top (U^\top x) = \|U^\top x\|^2.$$

Let us denote $U^\top x$ by y. Then

$$\|A\|_2 = \max_{y^\top y = 1} \sqrt{y^\top \Lambda y} = \max_{y^\top y = 1} \sqrt{y_1^2 \lambda_1 + y_2^2 \lambda_2 + \ldots + y_n^2 \lambda_n} = \max_{j = 1, \ldots, n} \sqrt{|\lambda_n|} \equiv \sqrt{\rho(A^\top A)}.$$

Remark If A is a real symmetric matrix, then the eigenvalues of $A^\top A$ are squares of the eigenvalues of A. Hence the 2-norm of A is the spectral radius of A:

$$\|A\|_2 = \max_i |\lambda_i| = \rho(A),$$

1.4.3. *Defective matrices and the Jordan form.* If matrix is not diagonalizable, it is called *defective*. An example of such a matrix is

$$A = \begin{bmatrix} 1 & 10 \\ 0 & 1 \end{bmatrix}. \tag{3}$$

This matrix has eigenvalue 1 of algebraic multiplicity 1 and just one eigenvector $[1, 0]^\top$. In linear algebra, the Jordan form is often considered for such matrices:

$$A = VJV^{-1} \tag{4}$$

where J is a block-diagonal matrix with blocks of the form

$$J_j := \begin{bmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_j & 1 \\ & & & & \lambda_j \end{bmatrix}.$$

There is a unique eigenvector v_j corresponding to each block. The columns of V form the Jordan basis.

Exercise Find the Jordan form and the Jordan basis for the matrix in (3).

In numerical linear algebra, the Jordan form is rarely computed. The reason is that it is unstable with respect to small perturbations of A. For example, consider a 16×16 matrix A

$$A := \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}. \tag{5}$$

It is already in the Jordan form consisting of a single block, and its unique eigenvalue of algebraic multiplicity 16 is zero. Indeed,

$$\det(\lambda I - A) = \lambda^{16} = 0.$$

Now consider a perturbation of A such that the zero at its bottom left corner is replaced with 10^{-16}:

$$A + \delta A := \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ 10^{-16} & & & & 0 \end{bmatrix}. \tag{6}$$

The eigenvalues of $A + \delta A$ are the roots of

$$\det(\lambda I - A) = \lambda^{16} - 10^{-16} = 0.$$

There are 16 distinct complex eigenvalues located at the corners of the 16-gon in the complex plane:

$$\lambda_k = 0.1e^{i2\pi k/16}, \quad k = 0, 1, \ldots 15.$$

Hence, the Jordan form of A will be $\mathrm{diag}\{\lambda_0, \ldots, \lambda_{15}\}$ which is not close to (6). Thus, we see that a perturbation of the size of the machine epsilon has a dramatic effect on the Jordan form and on the magnitudes of the eigenvalues of A.

1.4.4. *The Schur form.* For reasons indicated in Section 1.4.3 the Jordan form of a matrix is rarely computed. Another eigenvalue revealing form is much more preferable: the Schur form defined by:

$$A = QTQ^\top$$

where T is upper-triangular,

$$T = \begin{bmatrix} \lambda_1 & t_{12} & t_{13} & \ldots & t_{1n} \\ & \lambda_2 & t_{23} & \ldots & t_{2n} \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & t_{n-1,n} \\ & & & & \lambda_n \end{bmatrix}.$$

and Q is orthogonal (or unitary if it is complex), i.e., its columns form an orthonormal basis, or $Q^*Q = I$. Often it is more preferable to deal with the so-called real Schur form in which complex pairs of eigenvalues form 2×2 blocks along the diagonal of T. Then both Q and T are real. The Matlab command to compute the Schur form is

```
A = rand(10);
[Q,T] = schur(A);
```

If A is real, this command computes the real Schur form. If you would like the complex Schur form, type

```
[Q,T] = schur(A,'complex');
```

Exercise Let $u + iv$ be a complex eigenvector of a real matrix A, and $\mu + i\nu$ be the corresponding eigenvalue. Show that

$$A[u, v] = [u, v] \begin{bmatrix} \mu & \nu \\ -\nu & \mu \end{bmatrix}, \tag{7}$$

i.e., the vectors u and v span a 2-dimensional invariant subspace of A.

1.5. **Normal equations.** Consider the overdetermined system of linear equations

$$Ax = b, \quad A_{m\times n}, \quad m \geq n.$$

Such problems arise, for example, when we have want to find a line that best fits measured data points (x_i, y_i), $i = 1, \ldots, m$, that ideally lie on a straight line $ax + b = y$. Thus, we set up the following system:

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}. \tag{8}$$

The system (8) typically does not have a solution unless the points happen to lie on the same line. However, we always can find a line $ax + b$ that fits the data best in the least squares sense, the so called least squares solution.

Definition 5. *We say that x^* is the least squares solution of $Ax = b$, A is $m \times n$, $m \geq n$, if*

$$x^* = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|.$$

Now we will show that x^* is given by the formula

$$x^* = (A^\top A)^{-1} A^\top b. \tag{9}$$

Note that x^* is the solution of the so called normal equation that is obtained from $Ax = b$ by multiplication by $A^\top$ from the left. If the matrix A has full rank, i.e., $\operatorname{rank}(A) = n$, the matrix $A^\top A$ is symmetric positive definite. Write $x = x^* + e$ and consider $\|Ax - b\|^2$. We want to show that it is minimal if and only if $e = 0$, i.e., $x = x^*$.

$$\begin{aligned}
&\|Ax - b\|^2 = (Ax - b)^\top (Ax - b) = (Ax^* + Ae - b)^\top (Ax^* + Ae - b) = \\
&\|Ae\|^2 + \|Ax^* - b\|^2 + 2(Ae)^\top (Ax^* - b) = \\
&\|Ae\|^2 + \|Ax^* - b\|^2 + 2e^\top (A^\top Ax^* - A^\top b) = \\
&\|Ae\|^2 + \|Ax^* - b\|^2 \geq \|Ax^* - b\|^2
\end{aligned}$$

The equality occurs if and only if $e = 0$, i.e., the norm $\|Ax - b\|$ is minimal if and only if $x = x^*$ given by Eq. (9). A good way to compute the QR decomposition is by using the Householder reflections. We will discuss this later.

The geometric sense of the least squares solution is the following: the residual

$$r := Ax - b$$

is orthogonal to the space spanned by the columns of the matrix A, i.e., r dotted with any column of A is zero, or

$$A^\top r = 0.$$

1.6. **The QR decomposition and Gram-Schmidt Algorithm.**

Theorem 1. *Let A be $m \times n$, $m \geq n$. Suppose that A has full column rank. Then there exist a unique $m \times n$ orthogonal matrix Q, i.e., $Q^\top Q = I_{n\times n}$, and a unique $n \times n$ upper-triangular matrix R with positive diagonals $r_{ii} > 0$ such that $A = QR$.*

Proof. The proof of this theorem is given by the Gram-Schmidt orthogonalization process.

Algorithm 1: Gram-Schmidt orthogonalization

Input : matrix $A = [a_1\ a_2\ \dots\ a_n]$, $m \times n$, $\mathrm{rank}(A) = n$.
Output: orthogonal matrix Q $m \times n$, $Q^\top Q = I_{n\times n}$, and upper-triangular $n \times n$ matrix R with $r_{ii} > 0$.

for $i = 1, \dots, N$ **do**
  $q_i = a_i$;
  for $j = 1, \dots, i-1$ **do**
    $\begin{cases} r_{ji} = q_j^\top a_i & \text{CGS} \\ r_{ji} = q_j^\top q_i & \text{MGS} \end{cases}$;
    $q_i = q_i - r_{ji} q_j$;
  end
  $r_{ii} = \|q_i\|$;
  $q_i = q_i / r_{ii}$;
end

Here CGS and MGS stand for the Classic Gram-Schmidt and the Modified Gram-Schmidt respectively. □

Unfortunately the classic Gram-Schmidt algorithm is numerically unstable when the columns of A are nearly linearly dependent. The modified Gram-Schmidt is better but still can result in Q that is far from orthogonal (i.e., $\|Q^\top Q - I\|$ is much larger than the machine ϵ) when A is ill-conditioned. There are numerically stable ways to compute the QR-decomposition, i.e., by using the Householder reflections or Givens' rotations. We will consider the Householder reflections in homework exercises.

Exercise Show that the least squares solution of $Ax = b$ is given by

$$x^* = R^{-1} Q^\top b,$$

where $A = QR$ is the QR decomposition of A.

In Matlab, the least squares solution of $Ax = b$ is found by $A\backslash b$. The QR decomposition per se can be obtained by [Q,R]=`qr`(A).

1.7. **Singular Value Decomposition (SVD).** The Singular Value Decomposition is a very useful decomposition. It has numerous practical applications. Examples are image compression and determination of effective dimensionality of a data set.

Theorem 2. *Let A be an arbitrary $m \times n$ matrix with $m \geq n$. Then we can write*

$$A = U \Sigma V^\top,$$

where

$$U \text{ is } m \times n \text{ and } U^\top U = I_{n\times n},$$
$$\Sigma = \operatorname{diag}\{\sigma_1, \ldots, \sigma_n\},\ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0,$$
$$\text{and } V \text{ is } n \times n \text{ and } V^\top V = I_{n\times n}.$$

The columns of U, u_1, ..., u_n, are called left singular vectors. The columns of V, v_1, ..., v_n are called right singular vectors. The numbers σ_1, ..., σ_n are called singular values. If $m < n$, the SVD is defined for $A^\top$.

The geometric sense of this theorem is the following. Let us view the matrix A as a map from $\mathbb{R}^n$ into $\mathbb{R}^m$:

$$A : \mathbb{R}^n \to \mathbb{R}^m,\ x \mapsto Ax.$$

Then one can find orthogonal bases in $\mathbb{R}^n$, v_1, ... ,v_n, and in $\mathbb{R}^m$, u_1, ..., u_m and numbers σ_1, ..., σ_n, such that

$$v_j \mapsto \sigma_j u_j,\ j = 1, \ldots, n.$$

Then for any $x \in \mathbb{R}^n$ we have:

$$\text{if } x = \sum_{j=1}^{n} x_j v_j \text{ then } Ax = \sum_{j=1}^{n} x_j \sigma_j u_j.$$

Proof. We use induction in m and n. We assume that the SVD exists for $(m-1)\times(n-1)$ matrices and prove it for $m \times n$. We assume $A \neq 0$; otherwise we take $\Sigma = 0$ and U and V are arbitrary orthogonal matrices.

The basic step occurs when $n = 1$ (since $m > n$). We write

$$A = U\Sigma V^\top \text{ with } U = \frac{A}{\|A\|}, \quad \Sigma = \|A\|,\ V = 1,$$

where $\|\cdot\|$ is the 2-norm.

For the induction step, choose v so that

$$\|v\| = 1 \text{ and } \|A\| = \|Av\| > 0.$$

Let

$$u = \frac{Av}{\|Av\|},$$

which is a unit vector. Choose $\tilde{U}$ and $\tilde{V}$ so that $U = [u, \tilde{U}]$ and $V = [v, \tilde{V}]$ are $m \times n$ and $n \times n$ orthogonal matrices respectively. Now write

$$U^\top A V = \begin{bmatrix} u^\top \\ \tilde{U}^\top \end{bmatrix} \cdot A \cdot [v\ \tilde{V}] = \begin{bmatrix} u^\top Av & u^\top A\tilde{V} \\ \tilde{U}^\top Av & \tilde{U}^\top A\tilde{V} \end{bmatrix}.$$

Then

$$u^\top Av = \frac{(Av)^\top (Av)}{\|Av\|} = \|Av\| := \sigma$$

and

$$\tilde{U}^\top Av = \tilde{U}^\top u \|Av\| = 0.$$

We claim that $u^\top A\tilde{V} = 0$ too because otherwise

$$\sigma = \|A\| = \|U^\top AV\| \geq \|[1, 0, \ldots, 0]U^\top AV\| = \|[\sigma,\ u^\top A\tilde{V}]\| > \sigma,$$

a contradiction. Therefore,

$$U^\top AV = \begin{bmatrix} \sigma & 0 \\ 0 & \tilde{U}^\top AV \end{bmatrix} = \begin{bmatrix} u^\top Av & 0 \\ 0 & \tilde{A} \end{bmatrix}.$$

Now we apply the induction hypothesis that

$$\tilde{A} = U_1\Sigma_1 V_1^\top.$$

Hence,

$$U^\top AV = \begin{bmatrix} \sigma & 0 \\ 0 & U_1\Sigma_1 V_1^\top \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix} \begin{bmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix}^\top$$

or

$$A = \left(U \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix}\right) \begin{bmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{bmatrix} \left(V \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix}\right)^\top,$$

which is our desired decomposition.

□

The SVD has a large number of important algebraic and geometric properties, the most important of which are summarized in the following theorem.

Theorem 3. *Let $A = U\Sigma V^\top$ be the SVD of the $m \times n$ matrix A, $m \geq n$.*

(1) *Suppose A is symmetric and $A = U\Lambda U^\top$ be an eigendecomposition of A Then the SVD of A is $U\Sigma V^\top$ where $\sigma_i = |\lambda_i|$ and $v_i = u_i \mathrm{sign}(\lambda_i)$, where $\mathrm{sign}(0) = 1$.*

(2) *The eigenvalues of the symmetric matrix $A^\top A$ are σ_i^2. The right singular vectors v_i are the corresponding orthonormal eigenvectors.*

(3) *The eigenvectors of the symmetric matrix $AA^\top$ are σ_i^2 and $m - n$ zeroes. The left singular vectors u_i are the corresponding orthonormal eigenvectors for the eigenvalues σ_i^2. One can take any $m - n$ orthogonal vectors as eigenvectors for the eigenvalue 0.*

(4) *If A has full rank, the solution of*

$$\min_x \|Ax - b\| \quad \text{is} \quad x = V\Sigma^{-1}U^\top b.$$

(5)

$$\|A\|_2 = \sigma_1.$$

If A is square and nonsingular , then

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}.$$

(6) *Suppose*

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0.$$

Then

$$\mathrm{rank}(A) = r,$$

$$\text{null}(A) = \{x \in \mathbb{R}^n \ : \ Ax = 0 \in \mathbb{R}^m\} = \text{span}(v_{r+1}, \ldots, v_n),$$
$$\text{range}(A) = \text{span}(u_1, \ldots, u_r).$$

(7)
$$A = U\Sigma V^\top = \sum_{i=1}^{n} \sigma_i u_i v_i^\top,$$

i.e., A is a sum of rank 1 matrices. Then a matrix of rank $k < n$ closest to A is

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top, \quad \text{and} \quad \|A - A_k\| = \sigma_{k+1}.$$

Example This example illustrates the low rank approximation of a large matrix. The original image is shown in Fig. 1(a). The rank 3, 10, and 20 approximations are shown in Figs. 1 (b), (c), and (d) respectively. The sequence of matlab commands to create an approximation of rank m for a given image is the following.

```
>> clear all
>> im=imread('IMG_1413.jpg');
>> [m n k]=size(im)
m =        1600
n =        1200
k =     3
>> mimi=zeros(m,n);
>> mimi=sum(im,3);
>> fig=figure;
>> imagesc(mimi)
>> colormap gray
>> set(gca,'DataAspectRatio',[1 1 1])
>> [U S V]=svd(mimi);
>> size(U)
ans =        1600        1600
>> size(V)
ans =        1200        1200
>> size(S)
ans =        1600        1200
>> fig=figure;
>> m=10;
>> rm=U(:,1:m)*S(1:m,1:m)*V(:,1:m)';
>> colormap gray
>> set(gca,'DataAspectRatio',[1 1 1])
```

2. Condition number

We start with making the definition of the condition number more precise. Let $f(x)$ be a generally vector-valued function that we need to evaluate. The condition number $\kappa(f; x)$

FIGURE 1. Low rank approximations of image. (a): original; (b): rank 3; (c) rank 10; (d) rank 20.

is the ratio of the relative error in f caused by the relative error in x provided that the change in x is small. Hence, we define κ as

$$\kappa(f;x) := \lim_{\epsilon\to 0} \max_{\|\Delta x\|=\epsilon} \frac{\|f(x+\delta x)-f(x)\|/\|f(x)\|}{\|\Delta x\|/\|x\|}. \tag{10}$$

2.1. **Condition numbers for differentiable functions.** Let us calculate the condition numbers for differentiable functions. Let $f(x)$ be a differentiable function $f:\mathbb{R}^n\to\mathbb{R}$. Then

$$f(x+\Delta x) = f(x) + \nabla f(x+\theta\Delta x)^\top \Delta x, \quad \text{where} \quad \theta\in(0,1).$$

Then

$$\kappa(f;x) = \lim_{\epsilon\to 0} \max_{\|\Delta x\|=\epsilon} \frac{\|x\||\nabla f(x+\theta\Delta x)^\top \Delta x|}{|f(x)|\|\Delta x\|}.$$

The maximum over $\Delta x \in \mathbb{R}^n$ such that $\|\Delta x\| = \epsilon$ is achieved at

$$\Delta x = \frac{\nabla f(x + \theta \Delta x)}{\|\nabla f(x + \theta \Delta x)\|} \epsilon.$$

Therefore,

$$\kappa(f; x) = \frac{\|\nabla f(x)\| \|x\|}{|f(x)|}.$$

Now let $f(x)$ be a differentiable vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$. Then

$$f(x + \Delta x) = f(x) + J(x + \theta \Delta x)\Delta x, \quad \text{where} \quad \theta \in (0, 1),$$

and J is the jacobian matrix of f with entries:

$$J_{ij}(x) := \frac{\partial f_i}{\partial x_j}.$$

Then

$$\kappa(f; x) = \lim_{\epsilon \to 0} \max_{\|\Delta x\| = \epsilon} \frac{\|x\| |J(x + \theta \Delta x)\Delta x|}{\|f(x)\| \|\Delta x\|}.$$

The maximum over $\Delta x \in \mathbb{R}^n$ such that $\|\Delta x\| = \epsilon$ is achieved if Δx is parallel to the first right singular vector v_1 of $J(x + \Delta x) = U \Sigma V^\top$. Therefore,

$$\kappa(f; x) = \frac{\|J\| \|x\|}{\|f(x)\|}.$$

2.2. **Condition number for matrix-vector multiplication.** A particular case is when $f(x)$ is a linear function, i.e., $f(x) = Ax$ where A is an $m \times n$ matrix. Then the Jacobian matrix of f is constant and is equal to A. Hence, the condition number for matrix-vector multiplication is

$$\kappa(A; x) = \frac{\|A\| \|x\|}{\|Ax\|} = \|A\| \frac{\|x\|}{\|Ax\|}. \tag{11}$$

Identity (11) shows that the condition number will be large if

$$\frac{\|Ax\|}{\|x\|} \ll \|A\|,$$

i.e., if there is a vector y that is elongated by A by much larger factor than x. Let us illustrate this phenomenon on a simple example from D. Bindel's and J. Goodman's book "Principles of Scientific Computing", Chapter 4, page 89. Let

$$A = \begin{bmatrix} 1000 & 0 \\ 0 & 10 \end{bmatrix}, \quad \text{and} \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then

$$Ax = \begin{bmatrix} 0 \\ 10 \end{bmatrix}.$$

Suppose x is perturbed by

$$\Delta x = \begin{bmatrix} \epsilon \\ 0 \end{bmatrix}. \quad \text{Then} \quad A(x + \Delta x) - Ax = A\Delta x = \begin{bmatrix} 1000\epsilon \\ 0 \end{bmatrix}.$$

The error in x is amplified by the factor of 1000 that is 100 times larger than the elongation of x. It is easy to check that for this example, $\kappa(A; x) = 100$.

2.3. **Condition number for solving linear system.** On the other hand, let us consider the problem of solving a linear system $Ax = b$, i.e., $f(x) = A^{-1}b$. We find:

$$\kappa(A^{-1}; b) = \|A^{-1}\| \frac{\|b\|}{\|A^{-1}b\|} = \|A^{-1}\| \frac{\|Ax\|}{\|x\|}. \tag{12}$$

The condition number for the linear system $Ax = b$ is large if some vector is stretched by A much less than the solution x (recall that $\|A^{-1}\| = 1/\sigma_n$, where σ is the smallest singular value of A).

What we often call the condition number of a matrix A defined as

$$\kappa(A) = \|A\|\|A^{-1}\|$$

is the worst-case scenario condition number for either of the problems: matrix-vector multiplication and solving of linear system.

2.4. **Condition numbers for eigenvalue problem.** Read Section 4.3.3 from D. Bindel's and J. Goodman's book "Principles of Scientific Computing". The method of virtual perturbations is described in the same book in Section 4.2.6.