Chapter 3

Numerical linear algebra

3.1 Motivation

In this chapter we will consider the two following problems:

① Solve linear systems $Ax = b$, where $x, b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$.

② Find $x \in \mathbb{R}^n$ that minimizes

$$\sum_{i=1}^{m} (Ax - b)_i^2,$$

where $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$. When $m > n$ there are more equations than unknowns, so that in general, $Ax = b$ cannot be solved.

Example 3.1 (Stokes flow in a cavity) Three equations,

$$\begin{aligned}\frac{\partial p}{\partial x} &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\\ \frac{\partial p}{\partial y} &= \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\\ \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} &= 0,\end{aligned}$$

for the functions $u(x, y)$, $v(x, y)$, and $p(x, y)$; $(x, y) \in [0, 1]^2$. The boundary conditions are

$$\begin{gathered} u(0, y) = u(1, y) = u(x, 0) = 0, \qquad u(x, 1) = 1\\ v(0, y) = v(1, y) = v(x, 0) = v(x, 1) = 0.\end{gathered}$$

Solve with a staggered grid. A linear system in $n^2 + 2n(n-1)$ unknowns. (And by the way, it is singular).

Example 3.2 (Curve fitting) We are given a set of m points (a_i, b_i) in the plane, and we want to find the best cubic polynomial through these points. I.e, we are looking for the coefficients x_1, x_2, x_3, x_4, such that the polynomial

$$p(y) = \sum_{j=1}^{4} x_j y^{j-1}$$

minimizes

$$\sum_{i=1}^{m} \left[p(y_i) - b_i\right]^2,$$

where the vector $p(y_i)$ is of the form Ax, and

$$A = \begin{pmatrix} 1 & y_1 & y_1^2 & y_1^3 \\ 1 & y_2 & y_2^2 & y_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_m & y_m^2 & y_m^3 \end{pmatrix}$$

3.2 Vector and matrix norms

Definition 3.1 (Norm) Let X be a (real or complex) linear space. It is ***normed*** *if there exists a function $\|\cdot\| : X \mapsto \mathbb{R}$ (the* ***norm****) with the following properties:*

① $\|x\| \geq 0$ *with* $\|x\| = 0$ *iff* $x = 0$.

② $\|\alpha x\| \leq |\alpha| \|x\|$.

③ $\|x + y\| \leq \|x\| + \|y\|$.

Example 3.3 The most common vector norms are the ***p-norms*** defined (on $\mathbb{C}^n$) by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

which are norms for $1 \le p < \infty$. Another common norm is the ***infinity-norm***,

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|.$$

It can be shown that $\|\cdot\|_\infty = \lim_{p\to\infty} \|\cdot\|_p$.

✎ *Exercise 3.1* Show that the p-norms do indeed satisfy the properties of a norm.

Lemma 3.1 (Hölder inequality) Let $p, q > 1$ with $1/p + 1/q = 1$. Then,

$$|\sum_{k=1}^n x_k y_k| \le \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} \left(\sum_{k=1}^n |x_k|^q\right)^{1/q}.$$

Proof: From ***Young's inequality***

$$|ab| \le \frac{|a|^p}{p} + \frac{|b|^q}{q},$$

follows

$$\frac{|\sum_{k=1}^n x_k y_k|}{\|x\|_p \|y\|_q} \le \sum_{k=1}^n \frac{|x_k|}{\|x\|_p}\frac{|y_k|}{\|y\|_q} \le \sum_{k=1}^n \frac{1}{p}\frac{|x_k|^p}{\|x\|_p^p} + \sum_{k=1}^n \frac{1}{q}\frac{|y_k|^q}{\|y\|_q^q} \le \frac{1}{p} + \frac{1}{q} = 1.$$

■

Lemma 3.2 (Minkowski inequality) Let $p, q > 1$ with $1/p + 1/q = 1$, then

$$\left(\sum_{k=1}^n |x_k + y_k|^p\right)^{1/p} \le \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} + \left(\sum_{k=1}^n |y_k|^p\right)^{1/p}.$$

Proof: We write

$$|x_k + y_k|^p \le |x_k||x_k + y_k|^{p-1} + |y_k||x_k + y_k|^{p-1}.$$

Using Hölder's inequality for the first term,

$$\sum_{k=1}^n |x_k||x_k + y_k|^{p-1} \le \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} \left(\sum_{k=1}^n |x_k + y_k|^{q(p-1)}\right)^{1/q}.$$

Note that $q(p-1)=p$. Similarly, for the second term

$$\sum_{k=1}^n |y_k||x_k+y_k|^{p-1} \le \left(\sum_{k=1}^n |y_k|^p\right)^{1/p} \left(\sum_{k=1}^n |x_k+y_k|^p\right)^{1/q},$$

Summing up,

$$\sum_{k=1}^n |x_k+y_k|^p \le \left(\sum_{k=1}^n |x_k+y_k|^p\right)^{1/q} \left(\|x\|_p + \|y\|_p\right).$$

Dividing by the factor on the right-hand side, and using the fact that $1-1/q = 1/p$ we get the required result. ■

Definition 3.2 (Inner product space) Let X be a (complex) vector space. The function $(\cdot,\cdot) : X \times X \mapsto \mathbb{C}$ is called an ***inner product*** *if:*

① $(x,y) = \overline{(y,x)}$.
② $(x,y+z) = (x,y) + (x,z)$ *(bilinearity).*
③ $(\alpha x, y) = \alpha(x,y)$.
④ $(x,x) \ge 0$ *with* $(x,x) = 0$ *iff* $x = 0$.

Example 3.4 For $X = \mathbb{C}^n$ the form

$$(x,y) = \sum_{i=1}^n x_i \overline{y}_i$$

is an inner product.

Lemma 3.3 (Cauchy-Schwarz inequality) The following inequality holds in an inner product space.

$$|(x,y)|^2 \le (x,x)(y,y).$$

Proof: We have,

$$0 \le (x-\alpha y, x-\alpha y) = (x,x) - \alpha(y,x) - \overline{\alpha}(x,y) + |\alpha|^2(y,y).$$

Suppose that $(y,x) = r\,\exp(\imath\theta)$, then take $\alpha = t\,\exp(-\imath\theta)$. For every t,

$$(x,x) - 2rt + t^2(y,y) \ge 0.$$

Since we have a quadratic inequality valid for all t we must have

$$r^2 - (x,x)(y,y) \le 0,$$

which completes the proof. ■

Comments:

① The Cauchy-Schwarz inequality is a special case of Hölder's inequality.

② A third method of proof is from the inequality

$$0 \le ((y,y)x - (x,y)y, (y,y)x - (x,y)y) = (y,y)\left[(x,x)(y,y) - |(x,y)|^2\right].$$

Lemma 3.4 In an inner product space $\sqrt{(x,x)}$ is a norm.

Proof: Let $\|x\| = \sqrt{(x,x)}$. The positivity and the homogeneity are immediate. The triangle inequality follows from the Cauchy-Schwarz inequality

$$\begin{aligned}\|x+y\|^2 &= (x+y,x+y) = \|x\|^2 + \|y\|^2 + (x,y) + (y,x) \\ &\le \|x\|^2 + \|y\|^2 + 2|(x,y)| \le \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| = (\|x\| + \|y\|)^2.\end{aligned}$$

■

Definition 3.3 An Hermitian matrix A is called **positive definite** *(s.p.d) if*

$$x^\dagger A x > 0$$

for all $x \neq 0$.

Definition 3.4 (Convergence of sequences) Let (x_n) be a sequence in a normed linear space X. It is said to converge to a limit x if $\|x_n - x\| \to 0$.

In $\mathbb{R}^n$ convergence in norm always implies convergence of each of the component.

Lemma 3.5 The norm $\|\cdot\|$ is a continuous mapping from X to $\mathbb{R}$.

Proof: This is an immediate consequence of the triangle inequality, for

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|,$$

hence

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Take now $y = x_n$ and the limit $n \to \infty$. ■

Definition 3.5 *Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on X. They are called equivalent if there exist constants $c_1, c_2 > 0$ such that*

$$c_1\|x\| \leq \|x\|' \leq c_2\|x\|$$

for all $x \in X$.

Theorem 3.1 *All norms over a finite dimensional vector space are equivalent.*

Proof: Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms. It is sufficient to show the existence of a constant $c > 0$ such that

$$\|x\|' \leq c\|x\|$$

for all x. In fact, it is sufficient to restrict this on the unit ball of the norm $\|\cdot\|$. Thus, we need to show that for all x on the unit ball of $\|\cdot\|$, the norm $\|\cdot\|'$ is bounded. This follows from the fact that the norm is a continuous function and that the unit ball of a finite-dimensional vector space is compact. ■

Lemma 3.6 *In $\mathbb{R}^n$ the following inequalities hold:*

$$\begin{aligned} \|x\|_2 \leq \|x\|_1 &\leq \sqrt{n}\|x\|_2 \\ \|x\|_\infty \leq \|x\|_2 &\leq \sqrt{n}\|x\|_\infty \\ \|x\|_\infty \leq \|x\|_1 &\leq n\,\|x\|_\infty. \end{aligned}$$

✎ *Exercise 3.2* Prove the following inequalities for vector norms:

$$\begin{aligned} \|x\|_2 \leq \|x\|_1 &\leq \sqrt{n}\|x\|_2 \\ \|x\|_\infty \leq \|x\|_2 &\leq \sqrt{n}\|x\|_\infty \\ \|x\|_\infty \leq \|x\|_1 &\leq n\,\|x\|_\infty. \end{aligned}$$

Definition 3.6 (Subordinate matrix norm) Let $\|\cdot\|$ be a norm of $X = \mathbb{R}^n$. For every $A : X \mapsto X$ (an operator on the space) we define the following function $\|\cdot\| : \mathscr{B}(X, X) \mapsto \mathbb{R}$,

$$\|A\| = \sup_{0 \neq x \in X} \frac{\|Ax\|}{\|x\|}. \tag{3.1}$$

Comments:

① By the homogeneity of the norm we have

$$\|A\| = \sup_{0 \neq x \in X} \left\| A \frac{x}{\|x\|} \right\| = \sup_{\|x\|=1} \|Ax\|.$$

② Since the norm is continuous and the unit ball is compact then,

$$\|A\| = \max_{\|x\|=1} \|Ax\|,$$

and the latter is always finite.

③ By definition, for all A and x,

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

Theorem 3.2 Eq. (3.1) *defines a norm on the space of matrices $\mathbb{R}^n \mapsto \mathbb{R}^n$, which we call the matrix norm* ***subordinate*** *to the vector norm $\|\cdot\|$.*

Proof: The positivity and the homogeneity are immediate. It remains to show the triangle inequality:

$$\begin{aligned} \|A + B\| &= \sup_{\|x\|=1} \|(A + B)x\| \\ &\leq \sup_{\|x\|=1} (\|Ax\| + \|Bx\|) \\ &\leq \sup_{\|x\|=1} \|Ax\| + \sup_{\|x\|=1} \|Bx\|. \end{aligned}$$

■

Lemma 3.7 For every two matrices A, B and subordinate norm $\|\cdot\|$,

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

In particular,

$$\|A^k\| \leq \|A\|^k.$$

Proof: Obvious. ■

✎ *Exercise 3.3* Show that for every invertible matrix A and norm $\|\cdot\|$,

$$\|A\|\|A^{-1}\| \geq 1.$$

Example 3.5 (infinity-norm) Consider the infinity norm on vectors. The matrix norm subordinate to the infinity norm is

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \max_i \left| \sum_j a_{i,j} x_j \right| = \max_i \sum_j |a_{i,j}|.$$

✎ *Exercise 3.4* Prove that the matrix norm subordinate to the vector norm $\|\cdot\|_1$ is

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|.$$

Example 3.6 (2-norm) Consider now the matrix 2-norm subordinate to the vector 2-form

$$\|x\|_2 = \sqrt{(x,x)}.$$

By definition,

$$\|A\|_2^2 = \sup_{\|x\|_2=1} (Ax, Ax) = \sup_{\|x\|_2=1} (A^\dagger Ax, x).$$

The matrix $A^\dagger A$ is Hermitian, hence it can be diagonalized $A^\dagger A = Q^\dagger \Lambda Q$, where Q is unitary. Then

$$\|A\|_2^2 = \sup_{\|x\|_2=1} (Q^\dagger \Lambda Q x, x) = \sup_{\|x\|_2=1} (\Lambda Q x, Qx) = \sup_{\|y\|_2=1} (\Lambda y, y),$$

where we have used the fact that $y = Q^{-1}x$ has unit norm. This gives,

$$\|A\|_2^2 = \sup_{\|y\|_2=1} \sum_{i=1}^n \lambda_i |y_i|^2,$$

which is maximized by taking y_i to choose the maximal eigenvalue. Thus,

$$\|A\|_2 = \sqrt{\operatorname{spr} A^\dagger A},$$

where we have used the fact that all the eigenvalue of an Hermitian matrix of the form $A^\dagger A$ are real and positive.

✎ *Exercise 3.5* ① Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and S be an n-by-n nonsingular matrix. Define $\|x\|' = \|Sx\|$, and prove that $\|\cdot\|'$ is a norm on $\mathbb{R}^n$.

② Let $\|\cdot\|$ be the matrix norm subordinate to the above vector norm. Define $\|A\|' = \|SAS^{-1}\|$, and prove that $\|\cdot\|'$ is the matrix norm subordinate to the corresponding vector norm.

✎ *Exercise 3.6* True or false: if $\|\cdot\|$ is a matrix norm subordinate to a vector norm, so is $\|\cdot\|' = \frac{1}{2}\|\cdot\|$ (the question is not just whether $\|\cdot\|'$ satisfies the definition of a norm; the question is whether there exists a vector norm, for which $\|\cdot\|'$ is the subordinate matrix norm!).

Neumann series Let A be an n-by-n matrix and consider the infinite series

$$\sum_{k=0}^{\infty} A^k,$$

where $A^0 = I$. Like for numerical series, this series is said to converge to a limit B, if the sequence of partial sums

$$B_n = \sum_{k=0}^{\infty} A^k$$

converges to B (in norm). Since all norms on finite dimensional spaces are equivalent, convergence does not depend on the choice of norm. Thus, we may consider any arbitrary norm $\|\cdot\|$.

Recall the root test for the convergence of numerical series. Since it only relies on the completeness of the real numbers, it can be generalized as is for arbitrary complete normed spaces. Thus, if the limit

$$L = \lim_{n\to\infty} \|A^n\|^{1/n}$$

exists, then $L < 1$ implies the (absolute) convergence of the above series, and $L > 1$ implies that the series does not converge.

Proposition 3.1 *If the series converges absolutely then*

$$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}$$

(and the right hand side exists). It is called the ***Neumann series*** *of* $(I - A)^{-1}$.

Proof: We may perform a term-by-term multiplication

$$(I - A) \sum_{k=0}^{\infty} A^k = \sum_{k=0}^{\infty} (A^k - A^{k+1}) = I - \lim_{k\to\infty} A^k,$$

but the limit must vanish (in norm) if the series converges. ■

We still need to establish the conditions under which the Neumann series converges. First, we show that the limit L always exists:

Proposition 3.2 *The limit* $\lim_{n\to\infty} \|A^n\|^{1/n}$ *exists and is independent of the choice of norms. The limit is called the* ***spectral radius*** *of* A *and is denoted by* $\operatorname{spr}(A)$.

Proof: Let $a_n = \log \|A^n\|$. Clearly,

$$a_{n+m} = \log \|A^{n+m}\| \le \log \|A^n\| \|A^m\| = a_n + a_m,$$

i.e., the sequence (a_n) is ***sub-additive***. Since the logarithm is a continuous function on the positive reals, we need to show that the limit

$$\lim_{n\to\infty} \log \|A^n\|^{1/n} = \lim_{n\to\infty} \frac{a_n}{n}$$

exists. This follows directly from the sub-additivity.

Indeed, set m. Then, any integer n can be written as $m = mq + r$, with $0 \le r < m$. We have,

$$\frac{a_n}{n} = \frac{a_{mq+r}}{n} \le \frac{q}{n} a_m + \frac{r}{n} a_r.$$

Taking $n \to \infty$, the right hand side converges to a_m/m, hence,

$$\limsup \frac{a_n}{n} \le \frac{a_m}{m}.$$

Taking then $m \to \infty$ we have

$$\limsup \frac{a_n}{n} \le \liminf \frac{a_m}{m}$$

which proves the existence of the limit. The independence on the choice of norm results from the equivalence of norms, as

$$c^{1/n}\|A^n\|^{1/n} \le (\|A^n\|')^{1/n} \le C^{1/n}\|A^n\|^{1/n}.$$

■

Corollary 3.1 *The Neumann series* $\sum_k A^k$ *converges if* $\operatorname{spr} A < 1$ *and diverges if* $\operatorname{spr} A > 1$.

Thus, the spectral radius of a matrix is always defined, and is a property that does not depend on the choice of norm. We now relate the spectral radius with the eigenvalues of A. First, a lemma:

Lemma 3.8 *Let* S *be an invertible matrix.Then,* $\operatorname{spr} S^{-1}AS = \operatorname{spr} A$.

Proof: This is an immediate consequence of the fact that $\|S^{-1} \cdot S\|$ is a matrix norm and the independence of the spectral radius on the choice of norm. ■

Proposition 3.3 *Let* $\Sigma(A)$ *be the set of eigenvalues of* A *(the* ***spectrum****). Then,*

$$\operatorname{spr} A = \max_{\lambda\in\Sigma(A)} |\lambda|.$$

Proof: By the previous lemma it is sufficient to consider A in Jordan canonical form. Furthermore, since all power of A remain block diagonal, and we are free to choose, say, the infinity norm, we can consider the spectral radius of a single Jordan block; the spectral radius of A is the maximum over the spectral radii of its Jordan blocks.

Let then A be an m-by-m Jordan block with eigenvalue λ, i.e.,

$$A = \lambda I + D,$$

where D has ones above its main diagonal, i.e., it is nil-potent with $D^m = 0$. Raising this sum to the n-th power ($n > m$) we get

$$A^n = \lambda^n I + n\,\lambda^{n-1} D + \binom{n}{2}\lambda^{n-2}D^2 + \cdots \binom{n}{m-1}\lambda^{n-m+1}D^{m-1}.$$

Taking the infinity norm we have

$$|\lambda|^n \leq \|A^n\| \leq m \binom{n}{m-1} |\lambda|^{n-m+1} \max\left(|\lambda|^{m-1}, 1\right).$$

Taking the n-th root and going to the limit we obtain that $\operatorname{spr} A = |\lambda|$. ■

Proposition 3.4 For every matrix A,

$$\operatorname{spr} A \leq \inf_{\|\cdot\|} \|A\|,$$

where the infimum is over all choices of subordinate matrix norms.

Proof: For every eigenvalue λ with (normalized) eigenvector u, and every subordinate matrix norm $\|\cdot\|$,

$$\|A\| \geq \|Au\| = |\lambda|\|u\| = |\lambda|.$$

It remains to take the maximum over all $\lambda \in \Sigma(A)$ and the infimum over all norms. ■

We will now prove that this inequality is in fact an identity. For that we need the following lemma:

Lemma 3.9 Every matrix A can be "almost" diagonalized in the following sense: for every $\epsilon > 0$ there exists a non-singular matrix S such that

$$A = S^{-1}(\Lambda + T)S,$$

where Λ is diagonal with its element coinciding with the eigenvalues of A, and T is strictly upper triangular with $\|T\|_\infty < \epsilon$.

Proof: There exists a trasformation into the Jordan canonical form:

$$A = P^{-1}(\Lambda + D)P,$$

where D is nil-potent with ones above its main diagonal. Let now

$$E = \begin{pmatrix} \epsilon & 0 & \cdots & 0 \\ 0 & \epsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \epsilon^n \end{pmatrix}.$$

and set $E^{-1}P = S$. Then

$$A = S^{-1}E^{-1}(\Lambda + D)S = S^{-1}(\Lambda + E^{-1}DE)S,$$

where $T = EDE^{-1}$ is given by

$$T_{i,j} = \sum_{k,l} E^{-1}_{i,k} D_{k,l} E_{l,j} = \epsilon^{j-i} D_{i,j}.$$

But since the only non-zero elements are $D_{i,i+1} = 1$, we have $T^{i,i+1} = \epsilon$, and $\|T\|_\infty = \epsilon$. ■

Theorem 3.3 *For every matrix A,*

$$\operatorname{spr} A = \inf_{\|\cdot\|} \|A\|.$$

Proof: We have already proved the less-or-equal relation. It remains to show that for every $\epsilon > 0$ there exists a subordinate matrix norm $\|\cdot\|$ such that

$$\|A\| \le \operatorname{spr} A + \epsilon.$$

This follows from the fact that every matrix is similar to an almost diagonal matrix, and that the spectral radius is invariant under similarity transformations. Thus, for every ϵ we take S as in the lemma above, and set $\|\cdot\| = \|S^{-1} \cdot S\|_\infty$, hence

$$\|A\| = \|\Lambda + T\|_\infty \le \|\Lambda\|_\infty + \|T\|_\infty = \operatorname{spr} A + \epsilon.$$

■

✎ *Exercise 3.7* A matrix is called ***normal*** if it has a complete set of orthogonal eigenvectors. Show that for normal matrices,

$$\|A\|_2 = \operatorname{spr} A.$$

✎ *Exercise 3.8* Show that $\operatorname{spr} A < 1$ if and only if

$$\lim_{k\to\infty} A^k x = 0, \qquad \forall x.$$

✎ *Exercise 3.9* True or false: the spectral radius $\operatorname{spr} A$ is a matrix norm.

✎ *Exercise 3.10* Is the inequality $\operatorname{spr} AB \le \operatorname{spr} A \operatorname{spr} B$ true for all pairs of n-by-n matrices? What about if A and B were upper-triangular? Hint: try to take $B = A^T$ and

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}.$$

✎ *Exercise 3.11* Can you use the Neumann series to approximate the inverse of a matrix A? Under what conditions will this method converge?

✎ *Exercise 3.12 (Computer exercise)* Construct a "random" 6-by-6 matrix A. Then plot the 1,2, and infinity norms of $\|A^n\|^{1/n}$ as function of n with the maximum n large enough so that the three curves are sufficiently close to the expected limit.

Normal operators

Definition 3.7 A matrix A is called **normal** *if it commutes with its adjoint, $A^\dagger A = AA^\dagger$.*

Lemma 3.10 A is a normal operator if and only if

$$\|Ax\|_2 = \|A^\dagger x\|_2$$

for every $x \in \mathbb{R}^n$.

Proof: Suppose first that A is normal, then for all $x \in \mathbb{R}^n$,

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^\dagger Ax) = (x, AA^\dagger x) = (A^\dagger x, A^\dagger x) = \|A^\dagger x\|_2^2.$$

Conversely, let $\|Ax\|_2 = \|A^\dagger x\|_2$. Then,

$$(x, AA^\dagger x) = (A^\dagger x, A^\dagger x) = (Ax, Ax) = (x, A^\dagger Ax),$$

from which follows that

$$(x, (AA^\dagger - A^\dagger A)x) = 0, \qquad \forall x \in \mathbb{R}^n.$$

Since $AA^\dagger - A^\dagger A$ is symmetric then it must be zero (e.g., because all its eigenvalues are zero, and it cannot have any nilpotent part). ■

Lemma 3.11 *For every matrix* A,

$$\|A^\dagger A\|_2 = \|A\|_2^2.$$

Proof: Recall that the 2-norm of A is given by

$$\|A\|_2^2 = \operatorname{spr} A^\dagger A.$$

On the other hand, since $A^\dagger A$ is Hermitian, its $2-$norm coincides with its largest eigenvalue. ■

Theorem 3.4 *If* A *is a normal operator then*

$$\|A^n\|_2 = \|A\|_2^n,$$

and in particular $\operatorname{spr} A = \|A\|_2$.

Proof: Suppose first that A was Hermitian. Then, by the previous lemma

$$\|A^2\|_2 = \|A^\dagger A\|_2 = \|A\|_2^2.$$

Since A^2 is also Hermitian we then have $\|A^4\|_2 = \|A\|_2^4$, and so on for every $n = 2^m$. Suppose then that A is normal (but no necessarily Hermitian), then for every $n = 2^m$,

$$\|A^n\|_2^2 = \|(A^\dagger)^n A^n\|_2 = \|(A^\dagger A)^n\|_2 = \|(A^\dagger A)\|_2^n = \|A\|_2^{2n},$$

hence $\|A_n\|_2 = \|A\|_2^n$. It remains to treat the case of general n. Write then $n = 2^m - r$, $r \geq 0$. We then have

$$\|A\|_2^{n+r} = \|A^{n+r}\|_2 \leq \|A^n\|_2 \|A\|_2^r,$$

hence $\|A\|_2^n \leq \|A^n\|_2$. The reverse inequality is of course trivial, which proves the theorem. ■

3.3 Perturbation theory and condition number

Consider the linear system

$$Ax = b,$$

and a “nearby” linear system

$$(A+\delta A)\hat{x} = (b+\delta b).$$

The question is under what conditions the smallness of δA, δb guarantees the smallness of $\delta x = \hat{x} - x$. If δx is small the problem is well-conditioned, and it is ill-conditioned otherwise.

Subtracting the two equations we have

$$A(\hat{x}-x) + \delta A\,\hat{x} = \delta b,$$

or,

$$\delta x = A^{-1}\left(-\delta A\,\hat{x} + \delta b\right).$$

Taking norms we obtain an inequality

$$\|\delta x\| \le \|A^{-1}\|\left(\|\delta A\|\,\|\hat{x}\| + \|\delta b\|\right),$$

which we further rearrange as follows,

$$\frac{\|\delta x\|}{\|\hat{x}\|} \le \|A^{-1}\|\|A\|\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\|\hat{x}\|}\right).$$

We have thus expressed the relative change in the output as the product of the relative change in the input (we’ll look more carefully at the second term later) and the number

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

which is the (relative) ***condition number***. When $\kappa(A)$ is large a small perturbation in the input can produce a large perturbation in the output.

In practice, $\hat{x}$ will be the computed solution. Then, provided we have estimates on the “errors” δA, and δb, we can estimate the relative error $\|\delta x\|/\|\hat{x}\|$. From a theoretical point of view, however, it seems “cleaner” to obtain an error bound which in independent of δx (via $\hat{x}$). This can be achieved as follows. First from

$$(A+\delta A)(x+\delta x) = (b+\delta b) \quad \Rightarrow \quad (A+\delta A)\delta x = (-\delta A\,x + \delta b)$$

we extract

$$\begin{aligned}\delta x &= (A+\delta A)^{-1}(-\delta A\,x+\delta b)\\ &= [A(I+A^{-1}\delta A)]^{-1}(-\delta A\,x+\delta b)\\ &= (I+A^{-1}\delta A)^{-1}A^{-1}(-\delta A\,x+\delta b).\end{aligned}$$

Taking now norm and applying the standard inequalities we get

$$\frac{\|\delta x\|}{\|x\|} \le \|(I + A^{-1}\delta A)^{-1}\| \|A^{-1}\| \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right).$$

Now, if $\operatorname{spr} A^{-1}\delta A < 1$, we can use the Neumann series to get the following estimate,

$$\|(I+A^{-1}\delta A)^{-1}\| = \| \sum_{n=0}^{\infty} (-A^{-1}\delta A)^n \| \le \sum_{n=0}^{\infty} \|A^{-1}\|^n \|\delta A\|^n = \frac{1}{1 - \|A^{-1}\| \|\delta A\|}.$$

Combining with the above,

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|} \left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|} \right) \\ &= \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \, \|x\|} \right) \\ &\le \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right), \end{aligned}$$

where we have used the fact that $\|A\| \|x\| \ge \|Ax\| = \|b\|$. In this (cleaner) formulation the condition number is

$$\frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}},$$

which is close to $\kappa(A)$ provided that δA is sufficiently small, and more precisely, that $\kappa(A) \frac{\|\delta A\|}{\|A\|} = \|A^{-1}\| \|\delta A\| < 1$.

We conclude this section by establishing another meaning to the condition number. It is *the reciprocal on the distance to the nearest ill-posed problem.* A large condition number means that the problem is close *in a geometrical sense* to a singular problem.

Theorem 3.5 *Let A be non-singular, then*

$$\frac{1}{\kappa(A)} = \min \left\{ \frac{\|\delta A\|_2}{\|A\|_2} : A + \delta A \textit{ is singular} \right\},$$

where $\kappa(A)$ is expressed in terms of 2-norm (Euclidean).

Proof: Since $\kappa(A) = \|A\|_2 \, \|A^{-1}\|_2$, we need to show that

$$\frac{1}{\|A^{-1}\|_2} = \min\left\{\|\delta A\|_2 : A + \delta A \text{ is singular}\right\}.$$

If $\|\delta A\|_2 < \frac{1}{\|A^{-1}\|_2}$, then $\|A^{-1}\|_2\|\delta A\|_2 < 1$, which implies the convergence of the Neumann series

$$\sum_{n=0}^{\infty}(-A^{-1}\delta A)^n = (1 + A^{-1}\delta A)^{-1} = A^{-1}(A + \delta A)^{-1},$$

i.e.,

$$\|\delta A\|_2 < \frac{1}{\|A^{-1}\|_2} \qquad \Rightarrow \qquad A + \delta A \text{ is not singular,}$$

or,

$$\min\left\{\|\delta A\|_2 : A + \delta A \text{ is singular}\right\} \geq \frac{1}{\|A^{-1}\|_2}.$$

To show that this is an equality it is sufficient to construct a δA of norm $\frac{1}{\|A^{-1}\|_2}$ so that $A + \delta A$ is singular. By definition, there exists an $x \in \mathbb{R}^n$ on the unit sphere for which $\|A^{-1}x\|_2 = \|A^{-1}\|_2$. Let then $y = \frac{A^{-1}x}{\|A^{-1}x\|_2}$, be another unit vector and construct

$$\delta A = -\frac{xy^T}{\|A^{-1}\|_2}.$$

First note that

$$\|\delta A\|_2 = \frac{1}{\|A^{-1}\|_2} \max_{\|z\|_2=1} \|xy^T z\|_2 = \frac{1}{\|A^{-1}\|_2} \max_{\|z\|_2=1} |y^T z| = \frac{1}{\|A^{-1}\|_2},$$

where we have used the fact that $\|x\|_2 = 1$, and the fact that $|y^T z|$ is maximized for $z = y^T$. Finally, $A + \delta A$ is singular because

$$(A + \delta A)y = \left(A - \frac{xy^T}{\|A^{-1}\|_2}\right) y = Ay - \frac{x}{\|A^{-1}\|_2} = 0.$$

■

Comment: Note how the theorem relies on the use of the Euclidean norm.

✎ *Exercise 3.13* The ***spectrum*** $\Sigma(A)$ of a matrix A is the set of its eigenvalues. The ϵ***-pseudospectrum*** of A, which we denote by $\Sigma_\epsilon(A)$, is defined as the set of complex numbers z, for which there exists a matrix δA such that $\|\delta A\|_2 \le \epsilon$ and z is an eigenvalue of $A + \delta A$. In mathematical notation,

$$\Sigma_\epsilon(A) = \{z \in \mathbb{C} : \; \exists\, \delta A, \; \|\delta A\|_2 \le \epsilon, \; z \in \Sigma(A + \delta A)\} .$$

Show that

$$\Sigma_\epsilon(A) = \left\{z \in \mathbb{C} : \; \|(zI - A)^{-1}\|_2 \ge 1/\epsilon\right\} .$$

✎ *Exercise 3.14* Let $Ax = b$ and $(A + \delta A)\hat{x} = (b + \delta b)$. We showed in class that $\delta x = \hat{x} - x$ satisfies the inequality,

$$\|\delta x\|_2 \le \|A^{-1}\|_2 \left(\|\delta A\|_2 \|\hat{x}\|_2 + \|\delta b\|_2\right) .$$

Show that this is not just an upper bound: that for sufficiently small $\|\delta A\|_2$ there exist non-zero δA, δb such that the above in an equality. (**Hint**: follow the lines of the proof that links the reciprocal of the condition number to the distance to the nearest ill-posed problem.)

3.4 Direct methods for linear systems

Algorithms for solving the linear system $Ax = b$ are divided into two sorts: ***direct methods*** give, in the absence of roundoff errors, an exact solution after a finite number of steps (of floating point operations); all direct methods are variations of ***Gaussian elimination***. In contrast, ***iterative methods*** compute a sequence of iterates (x_n), until x_n is sufficiently close to satisfying the equation. Iterative methods may be much more efficient in certain cases, notably when the matrix A is ***sparse***.

3.4.1 Matrix factorization

The basic direct method algorithm uses ***matrix factorization***—the representation of a matrix A as a product of "simpler" matrices. Suppose that A

was ***lower-triangular***:

$$\begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Then the system can easily be solved using ***forward-substitution***:

```
Algorithm 3.4.1: FORWARD-SUBSTITUTION($A, b$)

 for $i = 1$ to $n$
   do $x_i = \left(b_i - \sum_{k=1}^{i-1} a_{ik}x_k\right) / a_{ii}$
```

Similarly, if A was ***upper-diagonal***,

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Then the system can easily be solved using ***backward-substitution***:

```
Algorithm 3.4.2: BACKWARD-SUBSTITUTION($A, b$)

 for $i = n$ downto 1
   do $x_i = \left(b_i - \sum_{k=i+1}^{n} a_{ik}x_k\right) / a_{ii}$
```

Finally, if A is a ***permutation matrix***, i.e., an identity matrix with its rows permuted, then the system $Ax = b$ only requires the permutation of the rows of b.

Matrix factorization consists of expressing any non-singular matrix A as a product $A = PLU$, where P is a permutation matrix, L is non-singular lower-triangular and U is non-singular upper-triangular. Then, the system $Ax = b$ is solved as follows:

$$\begin{aligned} LUx &= P^{-1}b = P^T b && \text{permute the entries of } b \\ Ux &= L^{-1}(P^T b) && \text{forward substitution} \\ x &= U^{-1}(L^{-1}P^T b) && \text{backward substitution.} \end{aligned}$$

This is the general idea. We now review these steps is a systematic manner.

Lemma 3.12 *Let P, P_1, P_2 be n-by-n permutation matrices and A be an n-by-n matrix. Then,*

① *PA is the same as A with its rows permuted and AP is the same as A with its column permuted.*
② *$P^{-1} = P^T$.*
③ $\det P = \pm 1$.
④ *$P_1 P_2$ is also a permutation matrix.*

Proof: Let $\pi : [1, n] \mapsto [1, n]$ be a permutation function (ono-to-one and onto). Then, the entries of the matrix P are of the form $P_{ij} = \delta_{\pi^{-1}(i),j}$. Now,

$$(PA)_{i,j} = \sum_{k=1}^{n} \delta_{\pi^{-1}(i),k} a_{kj} = a_{\pi^{-1}(i),j}$$

$$(AK)_{i,j} = \sum_{k=1}^{n} A_{ik} \delta_{\pi^{-1}(k),j} a_{kj} = a_{i,\pi(j)},$$

which proves the first assertion. Next,

$$(P^T P)_{i,j} = \sum_{k=1}^{n} \delta_{\pi^{-1}(i),k} \delta_{k,\pi^{-1}(j)} = \sum_{k=1}^{n} \delta_{i,\pi(k)} \delta_{\pi(k),j} = \delta_{i,j},$$

which proves the second assertion. The determinant of a permutation matrix is ±1 because when two rows of a matrix are interchanged the determinant changes sign. Finally, if P_1 and P_2 are permutation matrices with maps π_1 and π_2, then

$$\begin{aligned}(P_1 P_2)_{i,j} &= \sum_{k=1}^{n} \delta_{\pi_1^{-1}(i),k} \delta_{\pi_2^{-1}(k),j} = \sum_{k=1}^{n} \delta_{\pi_1^{-1}(i),k} \delta_{k,\pi_2(j)} \\ &= \delta_{\pi_1^{-1}(i),\pi_2(j)} = \delta_{\pi_2^{-1}(\pi_1^{-1}(i)),j}.\end{aligned}$$

■

Definition 3.8 *The m-th* ***principal sub-matrix*** *of an n-by-n matrix A is the square matrix with entries a_{ij}, $1 \le i, j \le m$.*

Definition 3.9 *A lower triangular matrix L is called* **unit lower triangular** *if its diagonal entries are* 1.

Theorem 3.6 *A matrix A has a unique decomposition $A = LU$ with L unit lower triangular and U non-singular upper triangular if and only if all its principal sub-matrices are non-singular.*

Proof: Suppose first that $A = LU$ with the above properties. Then, for every $1 \le m \le n$,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} \\ & U_{22} \end{pmatrix},$$

where A_{11} is the m-th principal sub-matrix, L_{11} and L_{22} are unit lower triangular and U_{11} and U_{22} are upper triangular. Now,

$$A_{11} = L_{11} U_{11}$$

is non-singular because $\det A_{11} = \det L_{11} \det U_{11} = \prod_{i=1}^m u_{ii} \neq 0$, where the last step is a consequence of U being triangular and non-singular.

Conversely, suppose that all the principal sub-matrices of A are non-singular. We will show the existence of L, U by induction on n. For $n = 1$, $a = 1 \cdot a$. Suppose that the decomposition holds all $(n-1)$-by-$(n-1)$ matrices. Let A' be of the form

$$A' = \begin{pmatrix} A & b \\ c^T & d \end{pmatrix}$$

where b, c are column vectors of length $(n-1)$ and d is a scalar. By assumption, $A = LU$. Thus, we need to find vectors $l, u \in \mathbb{R}^{n-1}$ and a scalar γ such that

$$\begin{pmatrix} A & b \\ c^T & d \end{pmatrix} = \begin{pmatrix} L & \\ l^T & 1 \end{pmatrix} \begin{pmatrix} U & u \\ & \gamma \end{pmatrix}.$$

Expanding we have

$$\begin{aligned} b &= Lu \\ c^T &= l^T U \\ d &= l^T u + \gamma. \end{aligned}$$

The first and second equation for u, l can be solved because by assumption L and U are invertible. Finally, γ is extracted from the third equation. It must be non-zero otherwise A' would be singular. ■

A matrix A may be regular and yet the LU decomposition may fail. This is where permutations are necessary.

Theorem 3.7 Let A be a non-singular n-by-n matrix. Then there exist permutation matrices P_1, P_2, a unit lower triangular matrix L and an upper triangular matrix L, such that

$$P_1 A P_2 = LU.$$

Either P_1 or P_2 can be taken to be a unit matrix.

Proof: The proof is by induction. The case $n = 1$ is trivial. Assume this is true for dimension $n-1$. Let then A be a non-singular matrix. Thus, every row and every column has a non-zero element, and we can find permutation matrices P_1', P_2' such that $a_{11} = (P_1' A P_2')_{11} \neq 0$ (only one of them is necessary).

Now, we solve the block problem

$$P_1' A P_2' = \begin{pmatrix} a_{11} & A_{12}^T \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{A}_{22} \end{pmatrix},$$

where A_{22}, I and $\tilde{A}_{22}$ are $(n-1)$-by-$(n-1)$ matrices, and A_{12}, A_{21} L_{21}, U_{12} and are $(n-1)$-vectors; u_{11} is a scalar. Expanding, we get

$$u_{11} = a_{11}, \qquad A_{12} = U_{12}, \qquad A_{21} = L_{21} u_{11}, \qquad A_{22} = L_{21} U_{12}^T + \tilde{A}_{22}.$$

Since $\det A \neq 0$ and multiplication by a permutation matrix can at most change the sign of the determinant, we have

$$0 \neq \det P_1' A P_2' = 1 \cdot u_{11} \cdot \det \tilde{A}_{22},$$

from which we deduce that $\tilde{A}_{22}$ is non-singular. Applying the induction, there exist permutation matrices $\tilde{P}_1$, $\tilde{P}_2$ and triangular matrices $\tilde{L}_{22}$, $\tilde{U}_{22}$ such that

$$\tilde{P}_1 \tilde{A}_{22} \tilde{P}_2 = \tilde{L}_{22} \tilde{U}_{22}.$$

Substituting we get

$$\begin{aligned}
P_1'AP_2' &= \begin{pmatrix} 1 & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{P}_1^T \tilde{L}_{22} \tilde{U}_{22} \tilde{P}_2^T \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_1^T \tilde{L}_{22} \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{U}_{22} \tilde{P}_2^T \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ L_{21} & \tilde{P}_1^T \tilde{L}_{22} \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \\ 0 & \tilde{U}_{22} \tilde{P}_2^T \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_1^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \tilde{P}_1 L_{21} & \tilde{L}_{22} \end{pmatrix} \begin{pmatrix} u_{11} & U_{12}^T \tilde{P}_2 \\ 0 & \tilde{U}_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{P}_2^T \end{pmatrix}
\end{aligned}$$

The two outer matrices are permutation matrices whereas the two middle matrices satisfy the required conditions. This completes the proof. ■

A practical choice of the permutation matrix, known as Gaussian elimination with partial pivoting (GEPP) is given is the following corollary:

Corollary 3.2 It is possible to choose $P_2' = I$ and P_1' so that a_{11} is the largest entry in absolute value in its column.

The PLU with partial pivoting algorithm is implemented as follows:

Algorithm 3.4.3: LU FACTORIZATION(A)

```
for $i = 1$ to $n - 1$
  /* permute only with rows under $i$ */
  permute the rows of $A, L$ such that $a_{ii} \neq 0$
  /* calculate $L_{21}$ */
  for $j = i + 1$ to $n$
    do $l_{ji} = a_{ji}/a_{ii}$
  /* calculate $U_{12}$ */
  for $j = i$ to $n$
    do $u_{ij} = a_{ij}$
  /* change $A_{22}$ into $\tilde{A}_{22}$ */
  for $j = i + 1$ to $n$
    do for $k = i + 1$ to $n$
    do $a_{jk} = a_{jk} - l_{ji}u_{ik}$
```

Comments:

① It can be checked that once l_{ij} and u_{ij} are computed, the corresponding entries of A are not used anymore. This means that U, L can overwrite A. (No need to keep the diagonal terms of L.)

② Since the algorithm involves row permutation, the output must also provide the permutation matrix, which can be represented by a vector.

③ In practice, there is no need to actually permute the entries of the matrix. This can be done "logically" only.

Operation count The number of operations needed for LU factorization can be deduced directly from the algorithm:

$$\sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} + \sum_{j=i+1}^{n}\sum_{k=i+1}^{n} 2\right) = \sum_{i=1}^{n-1}\left[(n-i)+2(n-i)^2\right] = \frac{2}{3}n^3 + O(n^2).$$

Since the forward and backward substitution require $O(n^2)$ operations, the number of operations needed to solve the system $Ax = b$ is roughly $\frac{2}{3}n^3$.

✎ *Exercise 3.15* Show that every matrix of the form

$$\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix}$$

$a, b, \neq 0$, has an LU decomposition. Show that even if the diagonal elements of L are 1 the decomposition is not unique.

✎ *Exercise 3.16* Show that if $A = LU$ is symmetric then the columns of L are proportional to the rows of U.

✎ *Exercise 3.17* Show that every symmetric positive-definite matrix has an LU-decomposition.

✎ *Exercise 3.18* Suppose you want to solve the equation $AX = B$, where A is n-by-n and X, B are n-by-m. One algorithm would factorize $A = PLU$ and then solve the system column after column using forward and backward substitution. The other algorithm would compute A^{-1} using Gaussian elimination and then perform matrix multiplication to get $X = A^{-1}B$. Count the number of operations in each algorithm and determine which is more efficient.

✎ *Exercise 3.19* Determine the LU factorization of the matrix

$$\begin{pmatrix} 6 & 10 & 0 \\ 12 & 26 & 4 \\ 0 & 9 & 12 \end{pmatrix}.$$

✎ *Exercise 3.20 (Computer exercise)* Construct in Matlab an n-by-n matrix A (its entries are not important, but make sure it is non-singular), and verify how long its takes to perform the operation `B=inv(A);`. Repeat the procedure for $n = 10, 100, 1000, 2000$.

3.4.2 Error analysis

The two-step approach for obtaining error bounds is as follows:

① Analyze the accumulation of roundoff errors to show that the *algorithm* for solving $Ax = b$ generates the exact solution $\hat{x}$ of the nearby problem $(A+\delta A)\hat{x} = (b+\delta b)$, where δA, δb (the ***backward errors***) are small.

② Having obtained estimates for the backward errors, apply perturbation theory to bound the error $\hat{x} - x$.

Note that perturbation theory assumes that δA, δb are given. In fact, these perturbations are just "backward error estimates" of the roundoff errors present in the computation.

We start with backward error estimates, in the course of which we will get a better understanding of the role of ***pivoting*** (row permutation). As a demonstration, consider the matrix

$$A = \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}$$

with an arithmetic device accurate to three decimal digits. Note first that

$$\kappa(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 2 \times 2,$$

so that the result is quite insensitive to perturbations in the input. Consider now an LU decomposition, taking into account roundoff errors:

$$\begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \ell_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}.$$

Then,

$$\begin{aligned}
u_{11} &= \mathrm{fl}(0.0001/1) = 0.0001 \\
\ell_{21} &= \mathrm{fl}(1/u_{11}) = 10000 \\
u_{12} &= 1 \\
u_{22} &= \mathrm{fl}(1 - \ell_{21}u_{12}) = \mathrm{fl}(1 - 10000 \cdot 1) = -10000.
\end{aligned}$$

However,

$$\begin{pmatrix} 1 & 0 \\ 10000 & 1 \end{pmatrix} \begin{pmatrix} 0.0001 & 1 \\ 0 & -10000 \end{pmatrix} = \begin{pmatrix} 0.0001 & 1 \\ 1 & 0 \end{pmatrix}.$$

Thus, the a_{22} entry has been completely forgotten! In our terminology, the method is not backward stable because

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} = \frac{\|A - LU\|_\infty}{\|A\|_\infty} = \frac{1}{2}.$$

The relative backward error is large, and combined with the estimated condition number, the relative error in x could be as large as 2.

Had we used GEPP, the order of the rows would have been reversed,

$$\begin{pmatrix} 1 & 1 \\ 0.0001 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \ell_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix},$$

yielding

$$\begin{aligned}
u_{11} &= \mathrm{fl}(1/1) = 1 \\
\ell_{21} &= \mathrm{fl}(0.0001/u_{11}) = 0.0001 \\
u_{12} &= \mathrm{fl}(1/1) = 1 \\
u_{22} &= \mathrm{fl}(1 - \ell_{21}u_{12}) = \mathrm{fl}(1 - 0.0001 \cdot 1) = 1,
\end{aligned}$$

which combined back gives

$$\begin{pmatrix} 1 & 0 \\ 0.0001 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0.0001 & 1.0001 \end{pmatrix},$$

and

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} = \frac{\|A - LU\|_\infty}{\|A\|_\infty} = \frac{0.0001}{2}.$$

3.5 Iterative methods

3.5.1 Iterative refinement

Let's start with a complementation of direct methods. Suppose we want to solve the system $Ax = b$, i.e., we want to find the vector $x = A^{-1}b$, but due to roundoff errors (and possible other sources of errors), we obtain instead a vector

$$x_0 = \tilde{A}^{-1}b.$$

Clearly, we can substitute the computed solution back into the linear system, and find out that the ***residual***,

$$b - Ax_0 \overset{\text{def}}{=} r_0$$

differs from zero. Let $e_0 = x_0 - x$ be the ***error***. Subtracting $b - Ax = 0$ from the residual equation, we obtain

$$Ae_0 = r_0.$$

That is, *the error satisfies a linear equation with the same matrix A and the residual vector on its right hand side.*

Thus, we will solve the equation for e_0, but again we can only do it approximately. The next approximation we get for the solution is

$$x_1 = x_0 + \tilde{A}^{-1}r_0 = x_0 + \tilde{A}^{-1}(b - Ax_0).$$

Once more, we define the residual,

$$r_1 = b - Ax_1,$$

and notice that the error satisfies once again a linear system, $Ae_1 = r_1$, thus the next correction is $x_2 = x_1 + \tilde{A}^{-1}(b - Ax_1)$, and inductively, we get

$$x_{n+1} = x_n + \tilde{A}^{-1}(b - Ax_n). \tag{3.2}$$

The algorithm for iterative refinement is given by

```
Algorithm 3.5.1: ITERATIVE REFINEMENT$(A, b, \epsilon)$

$x = 0$
for $i = 1$ to $n$
  do $r = b - Ax$
     if $\|r\| < \epsilon$
       then break
     Solve $Ae = r$
     $x = x + e$
return $(x)$
```

Of course, if the solver is exact, the refinement procedure ends after one cycle.

Theorem 3.8 If $\tilde{A}^{-1}$ is sufficiently close to A^{-1} in the sense that $\mathrm{spr}(I - A\tilde{A}^{-1}) < 1$*, then the iterative refinement procedure converges to the solution x of the system $Ax = b$. (Note that equivalently, we need $\|I - A\tilde{A}^{-1}\|$ in any subordinate matrix norm.)*

Proof: We start by showing that

$$x_n = \tilde{A}^{-1} \sum_{k=0}^{n} (I - A\tilde{A}^{-1})^k b.$$

We do it inductively. For $n = 0$ we have $x_0 = \tilde{A}^{-1} b$. Suppose this was correct for $n - 1$, then

$$\begin{aligned} x_n &= x_{n-1} + \tilde{A}^{-1}(b - Ax_{n-1}) \\ &= \tilde{A}^{-1} \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k b + \tilde{A}^{-1} b - \tilde{A}^{-1} A \tilde{A}^{-1} \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k b \\ &= \tilde{A}^{-1} \left[\sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k + I - A\tilde{A}^{-1} \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k \right] b \\ &= \tilde{A}^{-1} \left[I + (I - A\tilde{A}^{-1}) \sum_{k=0}^{n-1} (I - A\tilde{A}^{-1})^k \right] b \\ &= \tilde{A}^{-1} \sum_{k=0}^{n} (I - A\tilde{A}^{-1})^k b. \end{aligned}$$

We have a Neumann series which converges if and only if $\operatorname{spr}(I - A\tilde{A}^{-1}) < 1$, giving in the limit

$$\lim_{n\to\infty} x_n = \tilde{A}^{-1}(A\tilde{A}^{-1})^{-1}b = A^{-1}b = x.$$

■

3.5.2 Analysis of iterative methods

Example 3.7 (Jacobi iterations) Consider the following example

$$\begin{aligned} 7x_1 - 6x_2 &= 3 \\ -8x_1 + 9x_2 &= -4, \end{aligned}$$

whose solution is $x = (1/5, -4/15)$. We may try to solve this system by the following iterative procedure:

$$\begin{aligned} x_1^{(n+1)} &= \frac{3 + 6\,x_2^{(n)}}{7} \\ x_2^{(n+1)} &= \frac{-4 + 8\,x_1^{(n)}}{9}. \end{aligned}$$

From a matrix point of view this is equivalent to taking the system

$$\begin{pmatrix} 7 & -6 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \end{pmatrix},$$

and splitting it as follows,

$$\begin{pmatrix} 7 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n+1)} = -\begin{pmatrix} 0 & -6 \\ -8 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n)} + \begin{pmatrix} 3 \\ -4 \end{pmatrix}.$$

This iterative methods, based on a splitting of the matrix A into its diagonal part and its off-diagonal part is called ***Jacobi's method***.

The following table gives a number of iterates:

n	$x_1^{(n)}$	$x_2^{(n)}$
1	0.4286	−0.4444
10	0.1487	−0.1982
20	0.1868	−0.2491
40	0.1991	−0.2655
80	0.2000	−0.2667

Example 3.8 (Gauss-Seidel iterations) Consider now the same system, but with a slightly different iterative method:

$$
\begin{aligned}
x_1^{(n+1)} &= \frac{3 + 6\,x_2^{(n)}}{7} \\
x_2^{(n+1)} &= \frac{-4 + 8\,x_1^{(n+1)}}{9}.
\end{aligned}
$$

The idea here is to use the entries which have already been computed in the present iteration. In matrix notation we have

$$
\begin{pmatrix} 7 & 0 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n+1)} = - \begin{pmatrix} 0 & -6 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{(n)} + \begin{pmatrix} 3 \\ -4 \end{pmatrix}.
$$

This iterative method, based on a splitting of the matrix A into its lower-triangular part and the remainder is called the ***Gauss-Seidel method***.

The following table gives a number of iterates:

n	$x_1^{(n)}$	$x_2^{(n)}$
1	0.4286	−0.0635
10	0.2198	−0.2491
20	0.2013	−0.2655
40	0.2000	−0.2667
80	0.2000	−0.2667

✎ *Exercise 3.21* Write an algorithm (i.e., a list of intructions in some pseudo-code) that calculates the solution to the linear system, $Ax = b$, by Gauss-Seidel's iterative procedure. The algorithm receives as input the matrix A and the vector b, and returns the solution x. Try to make the algorithm efficient.

✎ *Exercise 3.22 (Computer exercise)* Solve the system

$$
\begin{pmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}
$$

using both the Jacobi and the Gauss-Seidel iterations. Plot a graph of the norm of the errors as function of the number of iterations. Use the same graph for both methods for comparison.

We are now ready for a general analysis of iterative methods. Suppose we want to solve the system $Ax = b$. For any non-singular matrix Q we can equivalently write $Qx = (Q - A)x + b$, which leads to the iterative method

$$Qx_{n+1} = (Q - A)x_n + b.$$

Definition 3.10 An iterative method is said to be convergent if it converges for any initial vector x_0.

The goal is to choose a ***splitting matrix*** Q such that (1) Q is easy to invert, and (2) the iterations converge fast.

Theorem 3.9 Let A be a non-singular matrix, and Q be such that $\text{spr}(I - Q^{-1}A) < 1$*. Then the iterative method is convergent.*

Proof: We have

$$x_{n+1} = (I - Q^{-1}A)x_n + Q^{-1}b.$$

It is easy to see by induction that

$$x_n = (I - Q^{-1}A)^n x_0 + \sum_{k=0}^{n-1} (I - Q^{-1}A)^k Q^{-1}b,$$

and as we've already seen, the Neumann series converges iff $\text{spr}(I - Q^{-1}A) < 1$. If it converges, the first term also converges to zero (the initial condition is forgotten). The limit is

$$\lim_{n\to\infty} x_n = (Q^{-1}A)^{-1}Q^{-1}b = A^{-1}b = x.$$

■

Definition 3.11 A matrix A is called **diagonally dominant** *if for any row i,*

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|.$$

Proposition 3.5 *If A is diagonally dominant then Jacobi's method converges.*

Proof: For Jacobi's method the matrix Q comprises the diagonal of A, therefore, $Q^{-1}A$ consists of the rows of A divided by the diagonal term, and

$$(I - Q^{-1}A)_{ij} = \begin{cases} 0 & i = j \\ -\frac{a_{ij}}{a_{ii}} & i \neq j \end{cases}.$$

Because A is diagonally dominant,

$$\|I - Q^{-1}A\|_\infty = \max_i \sum_j |(I - Q^{-1}A)_{ij}| = \max_i \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1.$$

■

✎ *Exercise 3.23* Show that the Jacobi iteration converges for 2-by-2 symmetric positive-definite systems.

Hint Suppose that the matrix to be inverted is

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

First, express the positive-definiteness of A as a condition on a, b, c. Then, proceed to write the matrix $(I - Q^{-1}A)$, where Q is the splitting matrix corresponding to the Jacobi iterative procedure. It remains to find a norm in which $\|I - Q^{-1}A\| < 1$ or compute the spectral radius.

✎ *Exercise 3.24* Will Jacobi's iterative method converge for

$$\begin{pmatrix} 10 & 2 & 3 \\ 4 & 50 & 6 \\ 7 & 8 & 90 \end{pmatrix}.$$

✎ *Exercise 3.25* Explain why at least one eigenvalue of the Gauss-Seidel iterative matrix must be zero.

✎ *Exercise 3.26* Show that if A is strictly diagonally dominant then the Gauss-Seidel iteration converges.

✎ *Exercise 3.27* What is the explicit form of the iteration matrix $G = (I - Q^{-1}A)$ in the Gauss-Seidel method when

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

3.6 Acceleration methods

3.6.1 The extrapolation method

Consider a general iterative method for linear systems

$$x_{n+1} = Gx_n + c.$$

For the system $Ax = b$ we had $G = (I - Q^{-1}A)$ and $c = Q^{-1}b$, but for now this does not matter. We know that the iteration will converge if $\operatorname{spr} G < 1$.

Consider now the one-parameter family of methods,

$$\begin{aligned} x_{n+1} &= \gamma(Gx_n + c) + (1-\gamma)x_n \\ &= [\gamma G + (1-\gamma)I]x_n + \gamma c \overset{\text{def}}{=} G_\gamma x_n + \gamma c, \end{aligned}$$

$\gamma \in \mathbb{R}$. Can we choose γ such to optimize the rate of convergence, i.e., such to minimize the spectral radius of G_γ? Note that (1) if the method converges then it converges to the desired solution, and (2) $\gamma = 1$ reduces to the original procedure.

Recall that (1) the spectral radius is the largest eigenvalue (in absolute value), and that (2) if $\lambda \in \Sigma(A)$ and $p(\lambda) \in \Sigma(p(A))$ for any polynomial p. Suppose that we even don't really know the eigenvalues of the original matrix G, but we only know that they are real (true for symmetric or Hermitian matrices) and within the segment $[a, b]$. Then, the spectrum of G_γ lies within

$$\Sigma(G_\gamma) \subseteq \{\gamma z + (1-\gamma) : z \in [a, b]\}.$$

This means that

$$\operatorname{spr} G_\gamma \leq \max_{a\leq\lambda\leq b} |\gamma\lambda + (1-\gamma)|.$$

The expression on the right-hand side is the quantity we want to minimize,

$$\gamma^* = \arg\min_{\gamma\in\mathbb{R}} \max_{a\leq z\leq b} |\gamma z + (1-\gamma)|.$$

Problems of this type are call ***min-max problems***. They are very common in optimization.

Theorem 3.10 *If* $1 \notin [a,b]$, *then*

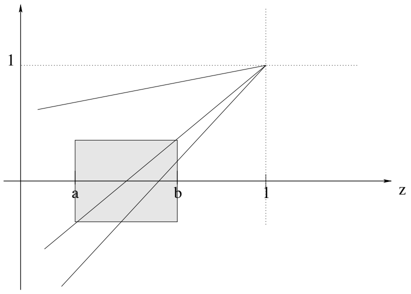
$$\gamma^* = \frac{2}{2-a-b},$$

and

$$\operatorname{spr} G^*_\gamma \leq 1 - |\gamma^*| d,$$

where $d = \operatorname{dist}(1,[a,b])$.

Proof: Since $1 \notin [a,b]$, then we either have $b < 1$ or $a > 1$. Let's focus on the first case; the second case is treated the same way. The solution to this problem is best viewed graphically:



From the figure we see that the optimal γ is when the absolute values of the two extreme cases coincide, i.e., when

$$\gamma(a-1) + 1 = -\gamma(b-1) - 1,$$

from which we readily obtain $2 = (2-a-b)\gamma^*$. Substituting the value of γ^* into

$$\max_{a\leq z\leq b} |\gamma z + (1-\gamma)|,$$

whose maximum is attained at either $z = a, b$, we get

$$\operatorname{spr} G_{\gamma^*} \leq \gamma^*(b-1) + 1 = 1 - |\gamma^*| d,$$

since γ^* is positive and $d = 1 - b$. ■

Example 3.9 The method of extrapolation can be of use even if the original method does not converge, i.e., even if $\operatorname{spr} G > 1$. Consider for example the following iterative method for solving the linear systems $Ax = b$,

$$x_{n+1} = (I - A)x_n + b.$$

It is known as Richardson's method. If we know that A has real eigenvalues ranging between λ_{min} and λ_{max}, then in the above notation

$$a = 1 - \lambda_{\text{max}} \qquad \text{and} \qquad b = 1 - \lambda_{\text{min}}.$$

If $1 \notin [a, b]$, i.e, all the eigenvalues of A have the same sign, then This means that the optimal extrapolation method is

$$x_{n+1} = [\gamma^*(I - A) + (1 - \gamma^*)I]\, x_n + \gamma^* b,$$

where

$$\gamma^* = \frac{2}{\lambda_{\text{max}} + \lambda_{\text{min}}}.$$

Suppose that $\lambda_{\text{min}} > 0$, then the spectral radius of the resulting iteration matrix is bounded by

$$\operatorname{spr} G_{\gamma^*} \leq 1 - \frac{2\lambda_{\text{min}}}{\lambda_{\text{max}} + \lambda_{\text{min}}} = \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}} + \lambda_{\text{min}}}.$$

It is easy to see that the bounds remains unchanged if $\lambda_{\text{max}} < 0$.

3.6.2 Chebyshev acceleration

Chebyshev's acceleration method takes the idea even further. Suppose we have an iterative method,

$$x_{n+1} = Gx_n + c,$$

and that we have used it to generate the sequence $x_0, x_1, \dots, x_n$. Can we use this existing sequence to get even closer to the solution? Specifically, consider a linear combination,

$$u_n = \sum_{k=0}^{n} a_{n,k} x_k.$$

We want to optimize this expression, with respect to the coefficients $a_{n,k}$ such that u_n is as close as possible to the fixed point $x = Gx + c$. Assume that for all n,

$$\sum_{k=0}^{n} a_{n,k} = 1.$$

Then,

$$u_n - x = \sum_{k=0}^{n} a_{n,k} x_k - x = \sum_{k=0}^{n} a_{n,k}(x_k - x).$$

Now, since $(x_k - x) = (Gx_{k-1} + c) - (Gx + c) = G(x_{k-1} - x)$, repeated application of this recursion gives

$$u_n - x = \sum_{k=0}^{n} a_{n,k} G^k (x_0 - x) \overset{\text{def}}{=} p_n(G)(x_0 - x),$$

where $p_n(z) = \sum_{k=0}^{n} a_{n,k} z^k$. Optimality will be achieved if we take the coefficients $a_{n,k}$ such to minimize the norm of $p_n(G)$, or instead, its spectral radius. Note that

$$\operatorname{spr} p_n(G) = \max_{z \in \Sigma(p_n(G))} |z| = \max_{z \in \Sigma(G)} |p_n(z)|.$$

Suppose all we knew was that the eigenvalues of G lie in a set S. Then, our goal is to find a polynomial of degree n, satisfying $p_n(1) = 1$, which minimizes

$$\max_{z \in S} |p_n(z)|.$$

That is, we are facing another min-max problem,

$$p_n^* = \arg \min_{p_n} \max_{z \in S} |p_n(z)|.$$

This can be quite a challenging problem. We will solve it again for the case where the spectrum of G is real, and confined to the set $S = [a, b]$.

Definition 3.12 (Chebyshev polynomials) The Chebyshev polynomials, $T_k(x)$, $k = 0, 1, \ldots,$ are a family of polynomials defined recursively by

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2x\, T_n(x) - T_{n-1}(x). \end{aligned}$$

Applying the iterative relation we have

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1. \end{aligned}$$

Note that for $y \in [-1, 1]$, we can express y as $\cos x$, in which case

$$\begin{aligned} T_2(y) &= T_2(\cos x) = 2\cos^2 x - 1 = \cos 2x = \cos(2\cos^{-1} y) \\ T_3(y) &= T_3(\cos x) = 4\cos^3 x - 3\cos x = \cos 3x = \cos(3\cos^{-1} y), \end{aligned}$$

and so on. This suggests the following relation:

Lemma 3.13 For $x \in [-1, 1]$ the Chebyshev polynomials have the following explicit representation:

$$T_n(x) = \cos(n\cos^{-1} x).$$

Proof: We have the following relations,

$$\begin{aligned} \cos[(n+1)\theta] &= \cos\theta\cos n\theta - \sin\theta\sin n\theta \\ \cos[(n-1)\theta] &= \cos\theta\cos n\theta + \sin\theta\sin n\theta, \end{aligned}$$

which upon addition gives

$$\cos[(n+1)\theta] = 2\ \cos\theta\cos n\theta - \cos[(n-1)\theta].$$

Set now $x = \cos\theta$, we get

$$\cos[(n+1)\cos^{-1} x] = 2\, x\cos[n\cos^{-1} x] - \cos[(n-1)\cos^{-1} x],$$

i.e., the functions $\cos[n\cos^{-1} x]$ satisfy the same recursion relations as the Chebyshev polynomials. It only remains to verify that they are identical for $n = 0, 1$. ■

Properties of the Chebyshev polynomials

① $T_n(x)$ is a polynomial of degree n.

② $|T_n(x)| \le 1$ for $x \in [-1, 1]$.

③ For $j = 0, 1, \ldots, n$,

$$T_n\left(\cos\frac{j\pi}{n}\right) = \cos(j\pi) = (-1)^j.$$

These are the extrema of $T_n(x)$.

④ For $j = 1, 2, \ldots, n$,

$$T_n\left(\cos\frac{(j-\frac{1}{2})\pi}{n}\right) = \cos\left((j-\frac{1}{2})\pi\right) = 0.$$

That is, the n-th Chebyshev polynomial has n real-valued roots and *all reside within the segment* $[-1, 1]$.

Proposition 3.6 *Let $p_n(z)$ be a polynomial of degree n with $p(\tilde{z}) = 1$, $\tilde{z} \notin [-1, 1]$. Then*

$$\max_{-1\le z\le 1} |p_n(z)| \ge \frac{1}{|T_n(\tilde{z})|}.$$

Equality is satisfied for $p_n(z) = T_n(z)/T_n(\tilde{z})$.

This proposition states that given that p_n equals one at a point z_n, there is a limit on how small it can be in the interval $[-1, 1]$. The Chebyshev polynomials are optimal, within the class of polynomials of the same degree, in that they can fit within a strip of minimal width.

Proof: Consider the $n + 1$ points $z_i = \cos(i\pi/n) \in [-1, 1]$, $i = 0, 1, \ldots, n$. Recall that these are the extrema of the Chebyshev polynomials, $T_n(z_i) = (-1)^i$.

We now proceed by contradiction, and assume that

$$\max_{-1\le z\le 1} |p_n(z)| < \frac{1}{|T_n(\tilde{z})|}.$$

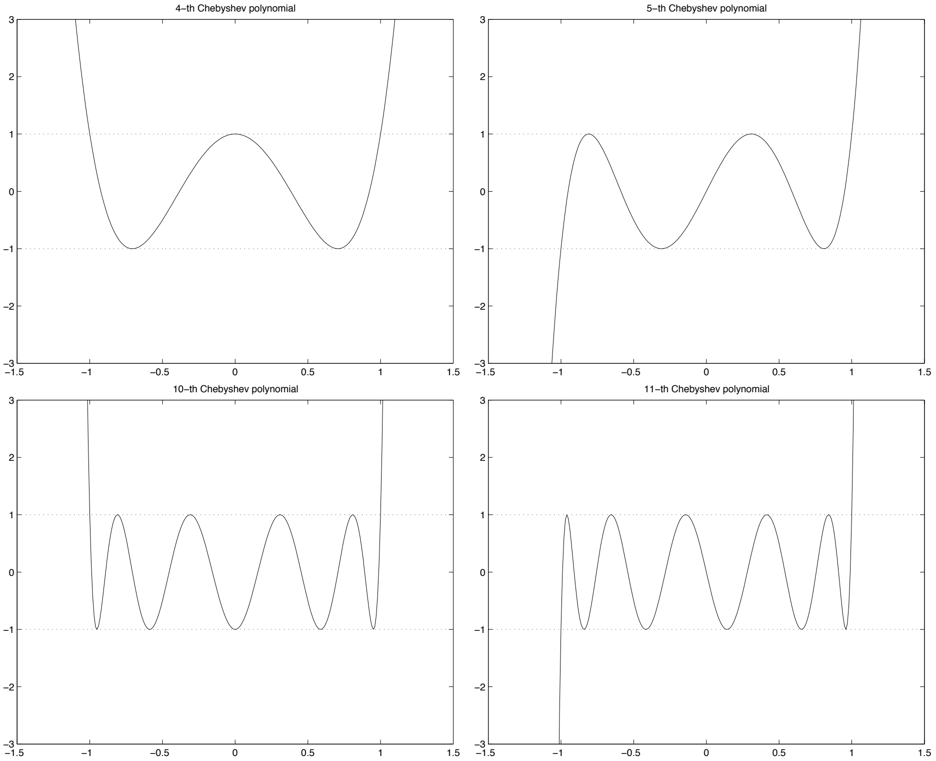


Figure 3.1: The functions $T_4(x)$, $T_5(x)$, $T_{10}(x)$, and $T_{11}(x)$.

If this holds, then a-forteriori,

$$|p_n(z_i)| - \frac{1}{|T_n(\tilde{z})|} < 0, \qquad i = 0, 1, \dots, n.$$

This can be re-arranged as follows

$$\operatorname{sgn}[T_n(\tilde{z})](-1)^i p_n(z_i) - \frac{(-1)^i T_n(z_i)}{\operatorname{sgn}[T_n(\tilde{z})]\, T_n(\tilde{z})} < 0,$$

or,

$$\operatorname{sgn}[T_n(\tilde{z})](-1)^i \left[p_n(z_i) - \frac{T_n(z_i)}{T_n(\tilde{z})} \right] < 0.$$

Consider now the function

$$f(z) = p_n(z) - \frac{T_n(z)}{T_n(\tilde{z})}.$$

It is a polynomial of degree at most n; its sign alternates at the z_i, implying the presence of n roots on the interval $[-1, 1]$; it has a root at $z = \tilde{z}$. This is impossible, contradicting the assumption. ■

Proposition 3.7 *Let $p_n(z)$ be a polynomial of degree n, $p_n(1) = 1$, and let a, b be real numbers such that $1 \notin [a, b]$. Then,*

$$\max_{a \le z \le b} |p_n(z)| \ge \frac{1}{|T_n(w(1))|},$$

where

$$w(z) = \frac{2z - b - a}{b - a}.$$

Equality is obtained for $p_n(z) = T_n(w(z))/T_n(w(1))$.

Note that a polynomial of degree n composed with a linear function is still a polynomial of degree n,

Proof: Take the case $a < b < 1$. Then,

$$w(1) = \frac{2 - b - a}{b - a} = 1 + 2\frac{1 - b}{b - a} \overset{\text{def}}{=} \tilde{w} > 1.$$

The converse relation is

$$z(w) = \frac{1}{2}[(b-a)w + a + b],$$

and $z(\tilde{w}) = 1$.

Let p_n we a polynomial of degree n satisfying $p_n(1) = 1$, and define $q_n(w) = p_n(z(w))$. We have $q_n(\tilde{w}) = p_n(1) = 1$, hence, by the previous proposition,

$$\max_{-1 \le w \le 1} |q_n(w)| \ge \frac{1}{|T_n(\tilde{w})|},$$

Substituting the definition of q_n, this is equivalent to

$$\max_{-1 \le w \le 1} |p_n(z(w))| = \max_{a \le z \le b} |p_n(z)| \ge \frac{1}{|T_n(\tilde{w})|}.$$

■

We have thus shown that among all polynomials of degree n satisfying $p_n(1) = 1$, the one that minimizes its maximum norm in the interval $[a, b]$ is

$$p_n(z) = \frac{T_n(w(z))}{T_n(w(1))}, \qquad \text{with} \qquad w(z) = \frac{2z - b - a}{b - a}.$$

What does this have to do with acceleration methods? Recall that we assume the existence of an iterative procedure,

$$x_{n+1} = Gx_n + c,$$

where $\operatorname{spr} G \in [a, b]$, and we want to improve it by taking instead

$$u_n = \sum_{k=0}^{n} a_{n,k} x_k,$$

where $\sum_{k=0}^{n} a_{n,k} = 1$. We've seen that this amounts to an iterative method with iteration matrix $p_n(G)$, where p_n is the polynomial with coefficients $a_{n,k}$. Thus, what we want is to find the polynomial that minimizes

$$\max_{a \le z \le b} |p_n(z)|,$$

and now we know which it is. This will ensure that

$$\text{error(n)} \le \frac{\text{error}(0)}{|T_n(w(1))|},$$

and the right hand side decays exponentially fast in n. We are still facing a practical problem of implementation. This will be dealt with now.

Lemma 3.14 The family of polynomials $p_n(z) = \frac{T_n(w(z))}{T_n(w(1))}$ *can be constructed recursively as follows:*

$$\begin{aligned}
p_0(z) &= 1 \\
p_1(z) &= \frac{2z-b-a}{2-b-a} \\
p_n(z) &= \sigma_n p_1(z) p_{n-1}(z) + (1-\sigma_n) p_{n-2},
\end{aligned}$$

where the constants σ_n *are defined by*

$$\sigma_1 = 2 \qquad \sigma_n = \left(1 - \frac{\sigma_{n-1}}{2[w(1)]^2}\right)^{-1}.$$

Proof: By the recursive property of the Chebyshev polynomials,

$$T_n(w(z)) = 2w(z)\, T_{n-1}(w(z)) - T_{n-2}(w(z)).$$

Dividing by $T_n(w(1))$, and converting T_k's into p_k's:

$$p_n(z) = \frac{2w(1)\, T_{n-1}(w(1))}{T_n(w(1))} p(z) p_{n-1}(w(z)) - \frac{T_{n-2}(w(1))}{T_n(w(1))} T_{n-2}(w(z)).$$

It remains to show that

$$\rho_n \overset{\text{def}}{=} \frac{2w(1)\, T_{n-1}(w(1))}{T_n(w(1))} = \sigma_n \qquad \text{and} \qquad -\frac{T_{n-2}(w(1))}{T_n(w(1))} = 1 - \sigma_n.$$

That their sum is indeed one follows from the Chebyshev recursion relation. It is also obvious that $\rho_1 = 2$. Finally,

$$\begin{aligned}
\rho_{n-1} &= \frac{2w(1)\, T_{n-2}(w(1))}{T_{n-1}(w(1))} \\
&= \frac{2w(1)\, \frac{T_{n-2}(w(1))}{T_n(w(1))} T_n(w(1))}{\frac{2w(1) T_{n-1}(w(1))}{T_n(w(1))} \frac{T_n(w(1))}{2w(1)}} \\
&= -[2w(1)]^2 \frac{1-\rho_k}{\rho_k}.
\end{aligned}$$

It only remains to invert this relation. ■

Theorem 3.11 *The sequence (u_n) of Chebyshev's acceleration's method can be constructed as follows:*

$$\begin{aligned} u_1 &= \gamma\,(Gx_0 + c) + (1-\gamma)x_0 \\ u_n &= \sigma_k\left[\gamma\,(Gx_{n-1} + c) + (1-\gamma)x_{n-1}\right] + (1-\sigma_n)u_{n-2}, \end{aligned}$$

where $\gamma = 2/(2-b-a)$ and the σ_n are as above.

Comments:

① The (u_n) are constructed directly without generating the (x_n).

② The first step is extrapolation, and the next ones are "weighted extrapolations".

③ The Chebyshev polynomials are not apparent (they are hiding...).

Proof: Start with $n = 1$,

$$u_1 = a_{1,1}x_1 + a_{1,0}x_0 = a_{1,1}(Gx_0 + c) + a_{1,0}x_0.$$

The coefficients $a_{1,0}$ and $a_{1,1}$ are the coefficients of the polynomial $p_1(z)$. By Lemma 3.14,

$$a_{1,1} = \frac{2}{2-b-a} = \gamma \qquad a_{1,0} = -\frac{a+b}{2-b-a} = 1-\gamma.$$

Now to the n-th iterate. Recall that

$$u_n = \sum_{k=0}^n a_{n,k}x_k = x + \sum_{k=0}^n a_{n,k}(x_k - x) = x + p_n(G)(x_0 - x).$$

By Lemma 3.14,

$$p_n(G) = \sigma_n p_1(G)p_{n-1}(G) + (1-\sigma_n)p_{n-2}(G),$$

and $p_1(G) = \gamma G + (1-\gamma)I$. Applying this on $x_0 - x$ we get

$$\begin{aligned} u_n - x &= \sigma_n\left[\gamma G + (1-\gamma)I\right](u_{n-1} - x) + (1-\sigma_n)(u_{n-2} - x) \\ &= \sigma_n\left[\gamma G u_{n-1} + (1-\gamma)u_{n-1}\right] - \sigma_n\left[\gamma Gx + (1-\gamma)x\right] \\ &\quad + (1-\sigma_n)u_{n-2} - (1-\sigma_n)x. \end{aligned}$$

It remains to gather the terms multiplying x. Since $x = Gx + c$ is a fixed point,

$$-\sigma_n\left[\gamma Gx + (1-\gamma)x\right] - (1-\sigma_n)x = \sigma_n\gamma c - x.$$

Substituting into the above we get the desired result. ■

✎ *Exercise 3.28 (Computer exercise)* The goal is to solve the system of equations:

$$\begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -4 \\ 0 \\ 4 \\ -4 \end{pmatrix}.$$

① Write explicitly the Jacobi iterative procedure,

$$x^{k+1} = Gx^k + c.$$

② What is is range of eigenvalues of the matrix G?

③ Is the Jacobi iterative procedure convergent?

④ Write an algorithm for the Chebyshev acceleration method based on Jacobi iterations.

⑤ Implement both procedures and compare their performance.

3.7 The singular value decomposition (SVD)

Relevant, among other things, to the mean-square minimization: find $x \in \mathbb{R}^n$ that minimizes $\|Ax - b\|_2$, where $A \in \mathbb{R}^{m\times n}$, and $\in \mathbb{R}^m$ (more equations than unknowns). It has many other uses.

Since we are going to consider vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, and operators between these two spaces, we will use the notation $\|\cdot\|_m$ and $\|\cdot\|_n$ for the corresponding vector 2-norms. Similarly, we will use $\|\cdot\|_{m\times n}$, etc., for the operator 2-norms. We will also use I_m, I_n to denote the identity operators in the two spaces.

Recall that the norm of an m-by-n matrix (it will always be assumed that $m \geq n$) is defined by

$$\|A\|_{m\times n} = \sup_{\|x\|_n=1} \|Ax\|_m = \sup_{(x,x)_n=1} \sqrt{(Ax, Ax)_m}.$$

A matrix Q is called ***orthogonal*** if its columns form an orthonormal set. If the matrix is n-by-n, then its columns form a basis in $\mathbb{R}^n$, and $Q^TQ = I_n$. Since Q is invertible, it immediately follows that $Q^T = Q^{-1}$, hence $QQ^T = I_n$ as well. If Q is an m-by-n orthogonal matrix, then $Q^TQ = I_n$, but the m-by-m matrix QQ^T is not an identity.

Lemma 3.15 *Let $x \in \mathbb{R}^n$, and Q be an orthogonal m-by-n matrix, $m \geq n$, then $\|Qx\|_m = \|x\|_n^2$.*

Proof: This is immediate by

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

■

Lemma 3.16 *Let A be an n-by-n matrix, V be an orthogonal n-by-n matrix, and U by an orthogonal m-by-n matrix. Then,*

$$\|UAV^T\|_{m\times n} = \|A\|_{n\times n}.$$

Proof: By definition,

$$\begin{aligned}\|UAV^T\|_{m\times n}^2 &= \sup_{(x,x)_n=1} (UAV^T x, UAV^T x)_m \\ &= \sup_{(x,x)_n=1} (AV^T x, AV^T x)_n \\ &= \sup_{(y,y)_n=1} (Ay, Ay)_n \\ &= \|A\|_{n\times n}^2,\end{aligned}$$

where we have used the previous lemma in the passage from the first to the second line, and the fact that and x on the unit sphere can be expressed as Vy, with y on the unit sphere. ■

Theorem 3.12 (SVD decomposition) *Let A be an m-by-n matrix, $m \geq n$. Then, A can be decomposed as*

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

where U is an m-by-n orthogonal matrix, V is an n-by-n orthogonal matrix, and Σ is an n-by-n diagonal matrix with entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

The columns of U, u_i, are called the ***left singular vectors***, the columns of V, v_i, are called the ***right singular vectors***, and the σ_i are called the ***singular values***. This theorem states that in some sense "every matrix is diagonal". Indeed, for every right singular vector v_i,

$$Av_i = U\Sigma V^T v_i = U\Sigma e_i = \sigma_i U e_i = \sigma_i u_i.$$

Thus, it is always possible to find an orthogonal basis $\{v_i\}$ in $\mathbb{R}^n$, and an orthogonal set $\{u_i\}$ in $\mathbb{R}^m$, such that any $x = \sum_{i=1}^n a_i v_i$ is mapped into $Ax = \sum_{i=1}^n \sigma_i a_i u_i$.

Proof: The proof goes by induction, assuming this can be done for an $(m-1)$-by-$(n-1)$ matrix. The basis of induction is a column vector, which can always be represented as a normalized column vector, times its norm, times one.

Let then A be given, and set v to be a vector on the unit sphere, $\|v\|_n = 1$, such that $\|Av\|_m = \|A\|_{m\times n}$ (such a vector necessarily exists). Set then $u = Av/\|Av\|_m$, which is a unit vector in $\mathbb{R}^m$. We have one vector $u \in \mathbb{R}^m$, which we complete (by Gram-Schmidt orthonormalization) into an orthogonal basis $U = (u, \tilde{U}) \in \mathbb{R}^{m\times m}$, $U^T U = UU^T = I_m$. Similarly, we complete $v \in \mathbb{R}^n$ into an orthonormal basis $V = (v, \tilde{V}) \in \mathbb{R}^{n\times n}$. Consider the m-by-n matrix

$$U^T AV = \begin{pmatrix} u^T \\ \tilde{U}^T \end{pmatrix} A \begin{pmatrix} v & \tilde{V} \end{pmatrix} = \begin{pmatrix} u^T Av & u^T A\tilde{V} \\ \tilde{U}^T Av & \tilde{U}^T A\tilde{V} \end{pmatrix}.$$

Note that $u \in \mathbb{R}^m$, $\tilde{U} \in \mathbb{R}^{m\times(m-1)}$, $v \in \mathbb{R}^n$ and $\tilde{V} \in \mathbb{R}^{n\times(n-1)}$. Hence, $u^T Av \in \mathbb{R}$, $u^T A\tilde{V} \in \mathbb{R}^{1\times(n-1)}$, $\tilde{U}^T Av \in \mathbb{R}^{(m-1)\times 1}$, and $\tilde{U}^T A\tilde{V} \in \mathbb{R}^{(m-1)\times(n-1)}$.

Now,

$$u^T Av = \|Av\|_m u^T u = \|A\|_{m\times n} \overset{\text{def}}{=} \sigma,$$

and

$$\tilde{U}^T Av = \|Av\|_m \tilde{U}^T u = 0,$$

due to the orthogonality of u and each of the rows of $\tilde{U}$. Thus,

$$U^T AV = \begin{pmatrix} \sigma & w^T \\ 0 & A_1 \end{pmatrix},$$

where $w^T = u^T A\tilde{V}$ and $A_1 = \tilde{U}^T A\tilde{V}$. We are going to prove that $w = 0$ as well. On the one hand we have

$$\left\| U^T AV \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_m^2 = \left\| \begin{pmatrix} \sigma^2 + w^T w \\ A_1 w \end{pmatrix} \right\|_m^2 \geq (\sigma^2 + w^T w)^2.$$

On the other hand

$$\left\| U^T AV \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_m^2 \leq \left\| U^T AV \right\|_{m\times n}^2 \left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_m^2 = \|A\|_{m\times n}^2 (\sigma^2 + w^T w),$$

where we have used the above lemma for $\left\|U^T AV\right\|_{m\times n}^2 = \|A\|_{m\times n}^2$. Since $\|A\|_{m\times n}^2 = \sigma^2$, it follows from these two inequalities that

$$(\sigma^2 + w^T w)^2 \leq \sigma^2(\sigma^2 + w^T w),$$

i.e., $w = 0$ as claimed.

Thus,

$$U^T AV = \begin{pmatrix} \sigma & 0 \\ 0 & A_1 \end{pmatrix},$$

At this stage, we use the inductive hypothesis for matrices of size $(m-1) \times (n-1)$, and write $A_1 = U_1 \Sigma_1 V_1^T$, which gives,

$$U^T AV = \begin{pmatrix} \sigma & 0 \\ 0 & U_1\Sigma_1 V_1^T \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V_1 \end{pmatrix}^T,$$

hence

$$A = \left[U \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix}\right] \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix} \left[V \begin{pmatrix} 1 & 0 \\ 0 & V_1 \end{pmatrix}\right]^T.$$

It remains to show that σ is larger or equal to all the diagonal entries of Σ, but this follows at once from the fact that

$$\sigma = \|A\|_{m\times n} = \left\| \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix} \right\|_{n\times n} = \left| \max_i \begin{pmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{pmatrix}_{ii} \right|.$$

This concludes the proof. ■

Having proved the existence of such a decomposition, we turn to prove a number of algebraic properties of SVD.

Theorem 3.13 *Let $A = U\Sigma V^T$ be an SVD of the m-by-n matrix A. Then,*

① *If A is square symmetric with eigenvalues λ_i, and orthogonal diagonalizing transformation $U = (u_1, \ldots, u_n)$, i.e., $A = U\Lambda U^T$, then an SVD of A is with $\sigma_i = |\lambda_i|$, the same U and V with columns $v_i = \operatorname{sgn}(\lambda_i)u_i$.*

② *The eigenvalues of the n-by-n (symmetric) matrix $A^T A$ are σ_i^2, and the corresponding eigenvalues are the right singular vectors v_i.*

③ *The eigenvalues of the m-by-m (symmetric) matrix AA^T are σ_i^2 and $m-n$ zeros. The corresponding eigenvectors are the left singular vectors supplemented with a set of $m-n$ orthogonal vectors.*

④ *If A has full rank (its columns are independent), then the vector $x \in \mathbb{R}^n$ that minimizes $\|Ax - b\|_m$ is $x = V\Sigma^{-1}U^T b$.*

⑤ $\|A\|_{m\times n} = \sigma_1$. *If, furthermore, A is square and non-singular then $\|A^{-1}\|_{n\times n} = 1/\sigma_n$, hence the condition number is σ_1/σ_n.*

⑥ *Suppose that $\sigma_1 \geq \sigma_n \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma - n = 0$. Then the rank of A is r, and*

$$\begin{aligned} \operatorname{null} A &= \operatorname{span}(v_{r+1}, \dots, v_n) \\ \operatorname{range} A &= \operatorname{span}(u_1, \dots, u_r). \end{aligned}$$

⑦ *Write $V = (v_1, \dots, v_n)$ and $U = (u_1, \dots, u_n)$. Then,*

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T,$$

*i.e., it is a sum of rank-*1 *matrices. The matrix of rank $k < n$ that is closest to A is*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T,$$

and $\|A - A_k\|_2 = \sigma_{k+1}$. A_k can also be written as

$$A_k = U\Sigma_k V^T,$$

where $\Sigma_k = \operatorname{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0)$.

Proof:

① This is obvious.

② We have

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

where we have used the fact that $U^T U = I_m$. This is an eigen-decomposition of $A^T A$.

③ Take an m-by-$(m-n)$ matrix $\tilde{U}$ such that $(U, \tilde{U})$ is orthogonal (use Gram-Schmidt). Then,

$$AA^T = (U, \tilde{U})\Sigma V^T V \Sigma^T (U, \tilde{U})^T = (U, \tilde{U})\Sigma\Sigma^T (U, \tilde{U})^T,$$