

Chapter 3

Linear Systems

We present algorithms for solving systems of linear equations whose coefficient matrix is nonsingular, and we discuss the accuracy of these algorithms.

3.1 The Meaning of $Ax = b$

First we examine when a linear system has a solution.

Fact 3.1 (Two Views of a Linear System) Let $A \in \mathbb{C}^{m \times n}$, and $b \in \mathbb{C}^{m \times 1}$.

1. The linear system $Ax = b$ has a solution if and only if there is a vector x that solves the m equations

$$r_1x = b_1 \quad \dots \quad r_mx = b_m,$$

where

$$A = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

2. The linear system $Ax = b$ has a solution if and only if b is a linear combination of columns of A ,

$$b = a_1x_1 + \dots + a_nx_n,$$

where

$$A = (a_1 \quad \dots \quad a_n), \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

When the matrix is nonsingular, the linear system has a solution for any right-hand side, and the solution can be represented in terms of the inverse of A .

Corollary 3.2 (Existence and Uniqueness). *If $A \in \mathbb{C}^{n \times n}$ is nonsingular then $Ax = b$ has the unique solution $x = A^{-1}b$ for every $b \in \mathbb{C}^n$.*

Before we discuss algorithms for solving linear systems we need to take into account, as discussed in Chapter 2, that matrix and right-hand side may be contaminated by uncertainties. This means, instead of solving $Ax = b$, we solve a perturbed system $(A + E)z = b + f$. We want to determine how sensitive the solution is to the perturbations f and E .

Even if we don't know the perturbations E and f , we can estimate them from the approximate solution z . To this end define the residual $r = Az - b$. We can view z as the solution to a system with perturbed right-hand side, $Az = b + r$. If $z \neq 0$, then we can also view z as the solution to a system with perturbed matrix,

$$(A + E)z = b, \quad \text{where } E = -\frac{rz^*}{\|z\|_2^2},$$

see Exercise 1.

Exercises

- (i) Determine the solution to $Ax = b$ when A is unitary (orthogonal).
- (ii) Determine the solution to $Ax = b$ when A is involutory.
- (iii) Let A consists of several columns of a unitary matrix, and b be such that the linear system $Ax = b$ has a solution. Determine a solution to $Ax = b$.
- (iv) Let A be idempotent. When does the linear system $Ax = b$ have a solution for every b ?
- (v) Let A be a triangular matrix. When does the linear system $Ax = b$ have a solution for *any* right-hand side b ?
- (vi) Let $A = uv^*$ be an outer product, where u and v are column vectors. For which b does the linear system $Ax = b$ have a solution?
- (vii) Determine a solution to the linear system $\begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$ when A is nonsingular. Is the solution unique?

1. Matrix Perturbations from Residuals.

This problem shows how to construct a matrix perturbation from the residual. Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, and $z \in \mathbb{C}^n$ a nonzero approximation to x . Show that $(A + E_0)z = b$, where $E_0 = (b - Az)z^\dagger$ and $z^\dagger = (z^*z)^{-1}z^*$; and that $(A + E)z = b$, where $E = E_0 + G(I - zz^\dagger)$ and $G \in \mathbb{C}^{n \times n}$ is any matrix.

2. In Problem 1 above show that, among all matrices F that satisfy $(A + F)z = b$, the matrix E_0 is one with smallest two-norm, i.e. $\|E_0\|_2 \leq \|F\|_2$.

3.2 Conditioning of Linear Systems

We derive normwise bounds for the conditioning of linear systems. The following two examples demonstrate that it is not obvious how to estimate the accuracy of an approximate solution z for a linear system $Ax = b$. In particular, they illustrate

that the residual $r = Az - b$ may give misleading information about how close z is to x .

Example 3.3 We illustrate that a totally wrong approximate solution can have a small residual norm.

Consider the linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 2 + \epsilon \end{pmatrix}, \quad 0 < \epsilon \ll 1, \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

whose solution x is approximated by $z = (2 \ 0)^T$. The residual

$$r = Az - b = \begin{pmatrix} 0 \\ -\epsilon \end{pmatrix}$$

has small norm, $\|r\|_p = \epsilon$, because ϵ is small. This appears to suggest that z does a good job of solving the linear system. However, comparing z to the exact solution,

$$z - x = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

shows that z is a bad approximation to x . Therefore, a small residual norm does not imply that z is close to x . ■

The same thing can happen even for triangular matrices, as the next example shows.

Example 3.4 For the linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 10^8 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 + 10^8 \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

consider the approximate solution

$$z = \begin{pmatrix} 0 \\ 1 + 10^{-8} \end{pmatrix}, \quad r = Az - b = \begin{pmatrix} 0 \\ 10^{-8} \end{pmatrix}.$$

As in the previous example, the residual has small norm, i.e. $\|r\|_p = 10^{-8}$, but z is totally inaccurate,

$$z - x = \begin{pmatrix} -1 \\ 10^{-8} \end{pmatrix}.$$

Again, the residual norm is deceptive. It is small even though z is a bad approximation. ■

The bound below explains why inaccurate approximations can have residuals with small norm.

Fact 3.5 (Residual Bound) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$ and $b \neq 0$. If $r = Az - b$ then

$$\frac{\|z - x\|_p}{\|x\|_p} \leq \kappa_p(A) \frac{\|r\|_p}{\|A\|_p \|x\|_p},$$

Proof. If $b \neq 0$ and A is nonsingular, then $x \neq 0$, see Fact 1.10. The desired bound follows immediately from the perturbation bound for matrix multiplication: Apply Fact 2.22 to $U = \tilde{U} = A^{-1}$, $V = b$, $\tilde{V} = b + r$, $\epsilon_U = 0$ and $\epsilon_V = \|r\|_p / \|b\|_p$ to obtain

$$\frac{\|z - x\|_p}{\|x\|_p} \leq \frac{\|A^{-1}\|_p \|b\|_p}{\|A^{-1}b\|_p} \frac{\|r\|_p}{\|b\|_p} = \|A\|_p \|A^{-1}\|_p \frac{\|r\|_p}{\|A\|_p \|x\|_p}.$$

□

The quantity $\kappa_p(A)$ is the normwise relative condition number of A with respect to inversion, see Definition 2.27. The bound in Fact 3.5 implies that the linear system $Ax = b$ is well-conditioned if $\kappa_p(A)$ is small. In particular, if $\kappa_p(A)$ is small and the relative residual norm $\frac{\|r\|_p}{\|A\|_p \|x\|_p}$ is also small then the approximate solution z has a small error (in the normwise relative sense). However, if $\kappa_p(A)$ is large, then the linear system is ill-conditioned. We return to Examples 3.3 and 3.4 to illustrate the bound in Fact 3.5.

Example. The linear system $Ax = b$ in Example 3.3 is

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 2 + \epsilon \end{pmatrix}, \quad 0 < \epsilon \ll 1, \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and an approximate solution $z = (2 \ 0)^T$ with residual

$$r = Az - b = \begin{pmatrix} 0 \\ -\epsilon \end{pmatrix}.$$

The relative error in the infinity norm is $\|z - x\|_\infty / \|x\|_\infty = 1$, indicating that z has no accuracy whatsoever. To see what the bound in Fact 3.5 predicts, we determine the inverse

$$A^{-1} = \frac{1}{\epsilon} \begin{pmatrix} 1 + \epsilon & -1 \\ -1 & 1 \end{pmatrix},$$

the matrix norms

$$\|A\|_\infty = 2 + \epsilon, \quad \|A^{-1}\|_\infty = \frac{2 + \epsilon}{\epsilon}, \quad \kappa_\infty(A) = \frac{(2 + \epsilon)^2}{\epsilon},$$

as well as the ingredients for the relative residual norm

$$\|r\|_\infty = \epsilon, \quad \|x\|_\infty = 1, \quad \frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = \frac{\epsilon}{2 + \epsilon}.$$

Since $\kappa_\infty(A) \approx 4/\epsilon$, the system $Ax = b$ is ill-conditioned. The bound in Fact 3.5 equals

$$\frac{\|z - x\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A) \frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = 2 + \epsilon,$$

and so correctly predicts the total inaccuracy of z . The small relative residual norm of about $\epsilon/2$ here is deceptive because the linear system is ill-conditioned. \blacksquare

Even triangular systems are not immune from ill-conditioning.

Example 3.6 The linear system $Ax = b$ in Example 3.4 is

$$A = \begin{pmatrix} 1 & 10^8 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 + 10^8 \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and an approximate solution $z = (0 \quad 1 + 10^{-8})^T$ with residual

$$r = Az - b = \begin{pmatrix} 0 \\ 10^{-8} \end{pmatrix}.$$

The normwise relative error in the infinity norm is $\|z - x\|_\infty / \|x\|_\infty = 1$ and indicates that z has no accuracy. From

$$A^{-1} = \begin{pmatrix} 1 & -10^8 \\ 0 & 1 \end{pmatrix}$$

we determine the condition number for $Ax = b$ as $\kappa_\infty(A) = (1 + 10^8)^2 \approx 10^{16}$. Note that conditioning of triangular systems cannot be detected by merely looking at the diagonal elements; the diagonal elements of A are equal to 1 and far from zero but nevertheless A is ill-conditioned with respect to inversion.

The relative residual norm is

$$\frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = \frac{10^{-8}}{1 + 10^8} \approx 10^{-16}.$$

As a consequence the bound in Fact 3.5 equals

$$\frac{\|z - x\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A) \frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = (1 + 10^8)10^{-8} \approx 1,$$

and correctly predicts that z has no accuracy at all. \blacksquare

The residual bound below does not require knowledge of the exact solution. The bound is analogous to the one in Fact 3.5, but bounds the relative error with regard to the perturbed solution.

Fact 3.7 (Computable Residual Bound) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular and $Ax = b$. If $z \neq 0$ and $r = Az - b$ then

$$\frac{\|z - x\|_p}{\|z\|_p} \leq \kappa_p(A) \frac{\|r\|_p}{\|A\|_p \|z\|_p},$$

We will now derive bounds that separate the perturbations in the matrix from those in the right-hand side. We first present a bound with regard to the relative error in the perturbed solution because it is easier to derive.

Fact 3.8 (Matrix and Righthand Side Perturbation) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, and $Ax = b$. If $(A + E)z = b + f$ with $z \neq 0$ then

$$\frac{\|z - x\|_p}{\|z\|_p} \leq \kappa_p(A) (\epsilon_A + \epsilon_f),$$

where

$$\epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \quad \epsilon_f = \frac{\|f\|_p}{\|A\|_p \|z\|_p}.$$

Proof. In the bound in Fact 3.7, the residual r accounts for both perturbations, because if $(A + E)z = b + f$ then $r = Az - b = f - Ez$. Replacing $\|r\|_p \leq \|E\|_p \|z\|_p + \|f\|_p$ in Fact 3.7 gives the desired bound. \square

Below is an analogous bound for the error with regard to the exact solution. In contrast to Fact 3.8, the bound below requires the perturbed matrix to be nonsingular.

Fact 3.9 (Matrix and Righthand Side Perturbation) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, and $Ax = b$ with $b \neq 0$. If $(A + E)z = b + f$ with $\|A^{-1}\|_p \|E\|_p \leq 1/2$ then

$$\frac{\|z - x\|_p}{\|x\|_p} \leq 2\kappa_p(A) (\epsilon_A + \epsilon_f)$$

where

$$\epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \quad \epsilon_f = \frac{\|f\|_p}{\|A\|_p \|x\|_p}.$$

Proof. We could derive the desired bound from the perturbation bound for matrix multiplication in Fact 2.22 and matrix inversion in Fact 2.25. However, the resulting bound would not be tight because it does not exploit any relation between matrix and righthand side. This is why we start from scratch.

Subtracting $(A + E)x = b + Ex$ from $(A + E)z = b + f$ gives $(A + E)(z - x) = f - Ex$. Corollary 2.24 implies that $A + E$ is nonsingular. Hence we can write $z - x = (A + E)^{-1}(-Ex + f)$. Taking norms and applying Corollary 2.24 yields

$$\|z - x\|_p \leq 2\|A^{-1}\|_p (\|E\|_p \|x\|_p + \|f\|_p) = 2\kappa_p(A) (\epsilon_A + \epsilon_f) \|x\|_p.$$

\square

We can simplify the bound in Fact 3.9 and obtain a weaker version.

Corollary 3.10. Let $Ax = b$ with $A \in \mathbb{C}^{n \times n}$ nonsingular and $b \neq 0$. If $(A + E)z = b + f$ with $\|A^{-1}\|_p \|E\|_p < 1/2$ then

$$\frac{\|z - x\|_p}{\|x\|_p} \leq 2\kappa_p(A) (\epsilon_A + \epsilon_b), \quad \text{where } \epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \quad \epsilon_b = \frac{\|f\|_p}{\|b\|_p}.$$

Proof. In Fact 3.9 bound $\|b\|_p \leq \|A\|_p \|x\|_p$. \square

Effect of the Right-Hand Side. So far we have focused almost exclusively on the effect that the matrix has on the conditioning of the linear system, and we have ignored the right-hand side. The advantage of this approach is that the resulting perturbation bounds hold for all right-hand sides. However the bounds can be too pessimistic for some right-hand sides, as the following example demonstrates.

Example 3.11 We illustrate that a favourable right-hand side can improve the conditioning of a linear system. Let's change the right-hand side in Example 3.6 and consider the linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 10^8 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 - 10^8 \\ 1 \end{pmatrix},$$

and approximate solution

$$z = \begin{pmatrix} -10^8 - 9 \\ 1 + 10^{-7} \end{pmatrix}, \quad r = Az - b = \begin{pmatrix} 0 \\ 10^{-7} \end{pmatrix}.$$

Although $\kappa_\infty(A) \approx 10^{16}$ implies that A is ill-conditioned with respect to inversion, the relative error in z is surprisingly small,

$$\frac{\|z - x\|_\infty}{\|x\|_\infty} = \frac{10}{1 - 10^8} \approx 10^{-7}.$$

The bound in Fact 3.5 recognizes this, too. From

$$\kappa_\infty(A) = (1 + 10^8)^2, \quad \frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = \frac{10^{-7}}{(10^8 - 1)(10^8 + 1)}$$

we obtain

$$\frac{\|z - x\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A) \frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = \frac{10^8 + 1}{10^8 - 1} 10^{-7} \approx 10^{-7}.$$

So, what is happening here? Observe that the relative residual norm is extremely small, $\frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} \approx 10^{-23}$, and that the norms of the matrix and right-hand side are large compared to the norm of the right-hand side, i.e. $\|A\|_\infty \|x\|_\infty \approx 10^{16} \gg \|b\|_\infty = 1$. We can represent this situation by writing the bound in Fact 3.5 as

$$\frac{\|z - x\|_\infty}{\|x\|_\infty} \leq \frac{\|A^{-1}\|_\infty \|b\|_\infty}{\|A^{-1}b\|_\infty} \frac{\|r\|_\infty}{\|b\|_\infty},$$

Because $\|A^{-1}\|_\infty \|b\|_\infty / \|A^{-1}b\|_\infty \approx 1$, the matrix multiplication of A^{-1} with b is well-conditioned with regard to changes in b . Hence the linear system $Ax = b$ is well-conditioned for this very particular right-hand side b . \blacksquare

Exercises

(i) Absolute residual bounds.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, and $r = Az - b$ for some $z \in \mathbb{C}^n$. Show

$$\|r\|_p / \|A\|_p \leq \|z - x\|_p \leq \|A^{-1}\|_p \|r\|_p.$$

(ii) Lower bounds for normwise relative error.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and $r = Az - b$ for some $z \in \mathbb{C}^n$.

Show

$$\frac{\|r\|_p}{\|A\|_p \|x\|_p} \leq \frac{\|z - x\|_p}{\|x\|_p}, \quad \frac{1}{\kappa_p(A)} \frac{\|r\|_p}{\|b\|_p} \leq \frac{\|z - x\|_p}{\|x\|_p}.$$

(iii) Relation between relative residual norms.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and $r = Az - b$ for some $z \in \mathbb{C}^n$.

Show

$$\frac{\|r\|_p}{\|A\|_p \|x\|_p} \leq \frac{\|r\|_p}{\|b\|_p} \leq \kappa_p(A) \frac{\|r\|_p}{\|A\|_p \|x\|_p}.$$

(iv) If a linear system is well-conditioned, and the relative residual norm is small, then the approximation has about the same norm as the solution.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, and $b \neq 0$. Prove: If

$$\rho\kappa < 1, \quad \text{where } \kappa = \kappa_p(A), \quad \rho = \frac{\|b - Az\|_p}{\|b\|_p}$$

then

$$1 - \kappa\rho \leq \frac{\|z\|_p}{\|x\|_p} \leq 1 + \kappa\rho.$$

(v) For this special righthand side, the linear system is well-conditioned with regard to changes in the righthand side.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, and $Az = b + f$. Show: If $\|A^{-1}\|_p = \|A^{-1}b\|_p / \|b\|_p$ then

$$\frac{\|z - x\|_p}{\|x\|_p} \leq \frac{\|f\|_p}{\|b\|_p}.$$

1. Let $A \in \mathbb{C}^{n \times n}$ be the bidiagonal matrix

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

a) Show that

$$\kappa_\infty(A) = \begin{cases} \frac{|\alpha|+1}{|\alpha|-1} (|\alpha|^n - 1) & \text{if } |\alpha| \neq 1 \\ 2n & \text{if } |\alpha| = 1. \end{cases}$$

Hint: See Exercise 4 in §1.13.

- b) Suppose we want to compute an approximation to the solution of $Ax = e_n$ when $\alpha = 2$ and $n = 100$. How small, approximately, must the residual norm be, so that the normwise relative error bound is less than .1?
2. Componentwise Condition Numbers.
Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $b \neq 0$, and $Ax = b$. Prove: If $x_j \neq 0$ then

$$\frac{|z_j - x_j|}{|x_j|} \leq \kappa_j \frac{\|b - Az\|_p}{\|b\|_p}, \quad \text{where } \kappa_j = \frac{\|x\|_p}{|x_j|} \|e_j^* A^{-1}\|_p \|A\|_p.$$

We can interpret κ_j as the condition number for x_j . Which components of x would you expect to be sensitive to perturbations?

3. Condition estimation.
Let A be nonsingular. Show how to determine a lower bound for $\kappa_p(A)$ with one linear system solution involving A .

3.3 Solution of Triangular Systems

Linear systems with triangular matrices are easy to solve. In the algorithm below we use the symbol “ \equiv ” to represent an assignment of a value.

Algorithm 3.1. Upper Triangular System Solution.

Input: Nonsingular, upper triangular matrix $A \in \mathbb{C}^{n \times n}$, vector $b \in \mathbb{C}^n$
Output: $x = A^{-1}b$

1. If $n = 1$ then $x \equiv b/A$.
2. If $n > 1$ partition

$$A = \begin{matrix} & n-1 & 1 \\ n-1 & \hat{A} & a \\ 1 & 0 & a_{nn} \end{matrix}, \quad x = \begin{matrix} n-1 \\ 1 \end{matrix} \begin{pmatrix} \hat{x} \\ x_n \end{pmatrix}, \quad b = \begin{matrix} n-1 \\ 1 \end{matrix} \begin{pmatrix} \hat{b} \\ b_n \end{pmatrix}$$

- (i) Set $x_n \equiv b_n/a_{nn}$.
- (ii) Repeat the process on the smaller system $\hat{A}\hat{x} = \hat{b} - x_n a$.

The process of solving an upper triangular system is also called *backsubstitution*, and the process of solving a lower triangular system is called *forward elimination*.

Exercises

- (i) Describe an algorithm to solve a non-singular lower triangular system.
- (ii) Solution of Block Upper Triangular Systems.

Even if A is not triangular, it may have a coarser triangular structure of which one take advantage. For instance, let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

where A_{11} and A_{22} are nonsingular. Show how to solve $Ax = b$ by solving two smaller systems.

(iii) Conditioning of Triangular Systems.

This problem illustrates that a nonsingular triangular matrix is ill-conditioned, if a diagonal element is small in magnitude compared to the other non-zero matrix elements.

Let $A \in \mathbb{C}^{n \times n}$ be upper triangular and nonsingular. Show:

$$\kappa_{\infty}(A) \geq \frac{\|A\|_{\infty}}{\min_{1 \leq j \leq n} |a_{jj}|}.$$

3.4 Stability of Direct Methods

We do **not** solve general nonsingular systems $Ax = b$ by first forming A^{-1} and then multiplying by b (likewise, you would not compute $2/4$ by first forming $1/4$ and then multiplying by 2). It is too expensive and numerically less accurate, see Exercise 4 below.

A more efficient approach factors A into a product of simpler matrices, and then solves a sequence of simpler linear systems. Examples of such factorizations include:

- LU factorization: $A = LU$ (if it exists), where L is lower triangular, and U is upper triangular.
- Cholesky factorization: $A = LL^*$ (if it exists), where L is lower triangular.
- QR factorization: $A = QR$, where Q is unitary and R is upper triangular. If A is real then Q is real orthogonal.

Methods that solve linear systems by first factoring a matrix are called *direct methods*. In general, a direct method factors $A = S_1 S_2$ (where “ S ” stands for “simpler matrix”) and then computes the solution $x = A^{-1}b = S_2^{-1} S_1^{-1} b$ by solving two linear systems.

Algorithm 3.2. Direct Method.

Input: Nonsingular matrix $A \in \mathbb{C}^{n \times n}$, vector $b \in \mathbb{C}^n$

Output: Solution of $Ax = b$

1. Factor $A = S_1 S_2$.
2. Solve the system $S_1 y = b$.
3. Solve the system $S_2 x = y$.

Each step of the above algorithm is itself a computational problem that may be sensitive to perturbations. We need to make sure that the algorithm does not introduce additional sensitivity by containing unnecessary ill-conditioned steps. For a direct method, this means that the factors S_1 and S_2 should be well-conditioned with respect to inversion. The example below illustrates that this cannot be taken for granted. That is, even if A is well-conditioned with respect to inversion, S_1 or S_2 can be ill-conditioned.

Example 3.12 The linear system $Ax = b$ with

$$A = \begin{pmatrix} \epsilon & 1 \\ 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 + \epsilon \\ 1 \end{pmatrix}, \quad 0 < \epsilon \leq 1/2,$$

has the solution $x = (1 \quad 1)^T$. The linear system is well-conditioned because

$$A^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -\epsilon \end{pmatrix}, \quad \kappa_\infty(A) = (1 + \epsilon)^2 \leq 9/4.$$

We can factor $A = S_1 S_2$ where

$$S_1 = \begin{pmatrix} 1 & 0 \\ \frac{1}{\epsilon} & 1 \end{pmatrix}, \quad S_2 = \begin{pmatrix} \epsilon & 1 \\ 0 & -\frac{1}{\epsilon} \end{pmatrix}$$

and then solve the triangular systems $S_1 y = b$ and $S_2 x = y$. Suppose we compute the factorization and the first linear system solution exactly, i.e.

$$A = S_1 S_2, \quad S_1 y = b, \quad y = \begin{pmatrix} 1 + \epsilon \\ -\frac{1}{\epsilon} \end{pmatrix},$$

and that we make errors only in the solution of the second system, i.e.

$$S_2 z = y + r_2 = \begin{pmatrix} 1 \\ -\frac{1}{\epsilon} \end{pmatrix}, \quad r_2 = \begin{pmatrix} -\epsilon \\ 0 \end{pmatrix}.$$

Then the computed solution satisfies

$$z = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \frac{\|z - x\|_\infty}{\|x\|_\infty} = 1.$$

The relative error is large because the leading component of z is completely wrong – although A is very well-conditioned. What happened? The triangular matrices S_1 and S_2 contain elements that are much larger in magnitude than the elements of A ,

$$\|A\|_\infty = 1 + \epsilon, \quad \|S_1\|_\infty = \frac{1 + \epsilon}{\epsilon}, \quad \|S_2\|_\infty = \frac{1}{\epsilon},$$

and the same is true for the inverses

$$\|A^{-1}\|_\infty = 1 + \epsilon, \quad \|S_1^{-1}\|_\infty = \|S_2^{-1}\|_\infty = \frac{1 + \epsilon}{\epsilon}.$$

The condition numbers for S_1 and S_2 are

$$\kappa_\infty(S_1) = \left(\frac{1+\epsilon}{\epsilon}\right)^2 \approx \frac{1}{\epsilon^2}, \quad \kappa_\infty(S_2) = \frac{1+\epsilon}{\epsilon^2} \approx \frac{1}{\epsilon^2}.$$

As a consequence, S_1 and S_2 are ill-conditioned with respect to inversion. Although the original linear system $Ax = b$ is well-conditioned, the algorithm contains steps that are ill-conditioned, namely the solution of the linear systems $S_1y = b$ and $S_2x = y$. ■

We want to avoid methods, like the one above, that factor a well-conditioned matrix into two ill-conditioned matrices. Such methods are called *numerically unstable*.

Definition 3.13. *An algorithm is (very informally) numerically stable in exact arithmetic, if each step in the algorithm is not much worse conditioned than the original problem.*

If an algorithm contains steps that are much worse conditioned than the original problem, the algorithm is called numerically unstable.

The above definition talks about “stability in exact arithmetic” because in this book we do not take into account errors caused by floating arithmetic operations (analyses that estimate such errors can be rather tedious). However, if a problem is numerically unstable in exact arithmetic then it is also numerically unstable in finite precision arithmetic, so that a distinction is not necessary in this case.

Below we analyze how the conditioning of the factors S_1 and S_2 affects the stability of Algorithm 3.2. The bounds are expressed in terms of relative residual norms from the linear systems.

Fact 3.14 (Stability in Exact Arithmetic of Direct Methods) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and

$$\begin{aligned} A + E &= S_1 S_2, & \epsilon_A &= \frac{\|E\|_p}{\|A\|_p} \\ S_1 y &= b + r_1, & \epsilon_1 &= \frac{\|r_1\|_p}{\|b\|_p} \\ S_2 z &= y + r_2, & \epsilon_2 &= \frac{\|r_2\|_p}{\|y\|_p}. \end{aligned}$$

If $\|A^{-1}\|_p \|E\|_p \leq 1/2$ then

$$\frac{\|z - x\|_p}{\|x\|_p} \leq \underbrace{2\kappa_p(A)}_{\text{condition}} (\epsilon_A + \epsilon_1 + \epsilon),$$

where

$$\epsilon = \underbrace{\frac{\|S_2^{-1}\|_p \|S_1^{-1}\|_p}{\|(A + E)^{-1}\|_p}}_{\text{stability}} \epsilon_2 (1 + \epsilon_1).$$

Proof. Expanding the right-hand side gives

$$(A + E)z = S_1 S_2 z = S_1(y + r_2) = S_1 y + S_1 r_2 = b + r_1 + S_1 r_2.$$

The obvious approach would be to apply Fact 3.9 to the perturbed linear system $(A + E)z = b + r_1 + S_1 r_2$. However the resulting bound would be too pessimistic, because we did not exploit the relation between matrix and righthand side. Instead, we can exploit this relation by subtracting $(A + E)x = b + Ex$ to obtain

$$(A + E)(z - x) = -Ex + r_1 + S_1 r_2.$$

Corollary 2.24 implies that $A + E$ is nonsingular, so that

$$z - x = (A + E)^{-1}(-Ex + r_1) + S_2^{-1}r_2.$$

Taking norms gives

$$\|z - x\|_p \leq \|(A + E)^{-1}\|_p (\|E\|_p \|x\|_p + \|r_1\|_p) + \|S_2^{-1}\|_p \|r_2\|_p.$$

Substituting $\|r_1\|_p = \epsilon_1 \|b\|_p \leq \epsilon_1 \|A\|_p \|x\|_p$ gives

$$\|z - x\|_p \leq \|(A + E)^{-1}\|_p \|A\|_p (\epsilon_A + \epsilon_1) \|x\|_p + \|S_2^{-1}\|_p \|r_2\|_p.$$

It remains to bound $\|r_2\|_p$. From $\|r_2\|_p = \epsilon_2 \|y\|_p$ and $y = S_1^{-1}(b + r_1)$ follows

$$\|r_2\|_p = \epsilon_2 \|y\|_p \leq \|S_1^{-1}\|_p (\|b\|_p + \|r_1\|_p).$$

Bounding $\|r_1\|_p$ as above yields

$$\|r_2\|_p \leq \|S_1^{-1}\|_p \|A\|_p \|x\|_p \epsilon_2 (1 + \epsilon_1).$$

We substitute this bound for $\|r_2\|_p$ into the above bound for $\|z - x\|_p$,

$$\|z - x\|_p \leq \|A\|_p \|x\|_p (\|(A + E)^{-1}\|_p (\epsilon_A + \epsilon_1) + \|S_2^{-1}\|_p \|S_1^{-1}\|_p \epsilon_2 (1 + \epsilon_1)).$$

Factoring out $\|(A + E)^{-1}\|_p$ and applying Corollary 2.24 gives the desired bound. \square

Remark 3.15.

- *The numerical stability in exact arithmetic of a direct method can be represented by the condition number for multiplying the two matrices S_2^{-1} and S_1^{-1} , see Fact 2.22, since*

$$\frac{\|S_2^{-1}\|_p \|S_1^{-1}\|_p}{\|(A + E)^{-1}\|_p} = \frac{\|S_2^{-1}\|_p \|S_1^{-1}\|_p}{\|S_2^{-1} S_1^{-1}\|_p}.$$

- *If $\|S_2^{-1}\|_p \|S_1^{-1}\|_p \approx \|(A + E)^{-1}\|_p$ then the matrix multiplication $S_2^{-1} S_1^{-1}$ is well-conditioned. In this case the bound in Fact 3.14 is approximately $2\kappa_p(A)(\epsilon_A + \epsilon_1 + \epsilon_2(1 + \epsilon_1))$, and Algorithm 3.2 is numerically stable in exact arithmetic.*

- If $\|S_2^{-1}\|_p \|S_1^{-1}\|_p \gg \|(A + E)^{-1}\|_p$ then Algorithm 3.2 is unstable.

Example 3.16 Returning to Example 3.12 we see that

$$\kappa_\infty(A) = (1 + \epsilon)^2, \quad \frac{\|S_1^{-1}\|_\infty \|S_2^{-1}\|_\infty}{\|A^{-1}\|_\infty} = \frac{1 + \epsilon}{\epsilon^2}, \quad \frac{\|r_2\|_\infty}{\|y\|_\infty} = \epsilon^2.$$

Hence the bound in Fact 3.14 equals $2(1 + \epsilon)^3$ and correctly indicates the inaccuracy of z . ■

The following bound is similar to the one in Fact 3.14, but it bounds the relative error with regard to the computed solution.

Fact 3.17 (A Second Stability Bound) Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, and

$$\begin{aligned} A + E &= S_1 S_2, & \epsilon_A &= \frac{\|E\|_p}{\|A\|_p} \\ S_1 y &= b + r_1, & \epsilon_1 &= \frac{\|r_1\|_p}{\|S_1\|_p \|y\|_p} \\ S_2 z &= y + r_2, & \epsilon_2 &= \frac{\|r_2\|_p}{\|S_2\|_p \|z\|_p} \end{aligned}$$

where $y \neq 0$ and $z \neq 0$. Then

$$\frac{\|z - x\|_p}{\|z\|_p} \leq \underbrace{\kappa_p(A)}_{\text{condition}} (\epsilon_A + \epsilon),$$

where

$$\epsilon = \underbrace{\frac{\|S_1\|_p \|S_2\|_p}{\|A\|_p}}_{\text{stability}} (\epsilon_2 + \epsilon_1 (1 + \epsilon_2)).$$

Proof. As in the proof of Fact 3.14 we start by expanding the right-hand side,

$$(A + E)z = S_1 S_2 z = S_1(y + r_2) = S_1 y + S_1 r_2 = b + r_1 + S_1 r_2.$$

The residual is $r = Az - b = -Ez + S_1 y + S_1 r_2 = b + r_1 + S_1 r_2$. Take norms and substitute the expressions for $\|r_1\|_p$ and $\|r_2\|_p$ to obtain

$$\|r\|_p \leq \|E\|_p \|z\|_p + \epsilon_1 \|S_1\|_p \|y\|_p + \epsilon_2 \|S_1\|_p \|S_2\|_p \|z\|_p.$$

To bound $\|y\|_p$ write $y = S_2 z - r_2$, take norms and replace $\|r_2\|_p = \epsilon_2 \|S_2\|_p \|z\|_p$ to get

$$\|y\|_p \leq \|S_2\|_p \|y\|_p + \|r_2\|_p = \|S_2\|_p \|z\|_p (1 + \epsilon_2).$$

Substituting this into the bound for $\|r\|_p$ gives

$$\|r\|_p \leq \|z\|_p (\|E\|_p + \|S_1\|_p \|S_2\|_p \epsilon_1 (1 + \epsilon_2) + \|S_1\|_p \|S_2\|_p \epsilon_2) = \|A\|_p \|z\|_p (\epsilon_A + \epsilon).$$

The relative error bound now follows from Fact 3.7. \square

In Fact 3.17, the numerical stability is represented by the factor $\|S_1\|_p\|S_2\|_p/\|A\|_p$. If $\|S_1\|_p\|S_2\|_p \gg \|A\|_p$ then Algorithm 3.2 is unstable.

Exercises

- The following bound is slightly tighter than the one in Fact 3.14. Under the conditions of Fact 3.14 show that

$$\frac{\|z - x\|_p}{\|x\|_p} \leq 2\kappa_p(A) [\epsilon_A + \rho_p(A, b) \epsilon]$$

where

$$\rho_p(A, b) = \frac{\|b\|_p}{\|A\|_p\|x\|_p}, \quad \epsilon = \frac{\|S_2^{-1}\|_p\|S_1^{-1}\|_p}{\|(A + E)^{-1}\|_p} \epsilon_2(1 + \epsilon_1) + \epsilon_1.$$

- The following bound suggests that Algorithm 3.2 is unstable if the first factor is ill-conditioned with respect to inversion. Under the conditions of Fact 3.14 show that

$$\frac{\|z - x\|_p}{\|x\|_p} \leq 2\kappa_p(A) [\epsilon_A + \epsilon_1 + \kappa_p(S_1) \epsilon_2(1 + \epsilon_1)].$$

- The following bound suggests that Algorithm 3.2 is unstable if the second factor is ill-conditioned with respect to inversion. Let $Ax = b$ where A is nonsingular. Also let

$$A = S_1S_2, \quad S_1y = b, \quad S_2z = y + r_2, \quad \text{where } \epsilon_2 = \frac{\|r_2\|_p}{\|S_2\|_p\|z\|_p}$$

and $z \neq 0$. Show that

$$\frac{\|z - x\|_p}{\|z\|_p} \leq \kappa_p(S_2) \epsilon_2.$$

- How Not to Solve Linear Systems.

One could solve a linear system $Ax = b$ by forming A^{-1} , and then multiplying A^{-1} with b . The bound below suggests that this approach is likely to be numerically less accurate than a direct solver.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, and $Ax = b$ with $b \neq 0$. Let $A + E \in \mathbb{C}^{n \times n}$ with $\|A^{-1}\|_p\|E\|_p \leq 1/2$. Compute $Z = (A + E)^{-1}$ and $z = Z(b + f)$. Show that

$$\frac{\|z - x\|_p}{\|x\|_p} \leq \kappa_p(A) \left(2 \frac{\|A^{-1}\|_p\|b\|_p}{\|A^{-1}b\|_p} \epsilon_A + \epsilon_f \right),$$

where

$$\epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \quad \epsilon_f = \frac{\|f\|_p}{\|A\|_p\|x\|_p},$$

and compare this to the bound in Fact 3.9.

Hint: Use the perturbation bounds for matrix multiplication and matrix inversion in Facts 2.22 and 2.25.

3.5 LU Factorization

The LU factorization of a matrix is the basis for Gaussian elimination.

Definition 3.18. Let $A \in \mathbb{C}^{n \times n}$. A factorization $A = LU$, where L is unit lower triangular and U is upper triangular matrix is called a LU factorization of A .

The LU factorization of a nonsingular matrix, if it exists, is unique, see the Exercise 5 in §1.13. Unfortunately there are matrices that do not have a LU factorizations, as the example below illustrates.

Example 3.19 The nonsingular matrix

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

cannot be factored into $A = LU$, where L is lower triangular and U is upper triangular. Suppose to the contrary that it could. Then

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l & 1 \end{pmatrix} \begin{pmatrix} u_1 & u_2 \\ 0 & u_3 \end{pmatrix}.$$

The first column of the equality implies $u_1 = 0$, and $lu_1 = 1$ so $u_1 \neq 0$, a contradiction. ■

Example 3.12 illustrates that a matrix A that is well-conditioned with respect to inversion can have LU factors that are ill-conditioned with respect to inversion. Algorithm 3.3 below shows how to permute the rows of a nonsingular matrix so that the permuted matrix has a LU factorization. Permuting the rows of A is called *partial pivoting* – as opposed to *complete pivoting* where both rows and columns are permuted. In order to prevent the factors from being too ill-conditioned, Algorithm 3.3 chooses a permutation matrix so that the elements of L are bounded.

Algorithm 3.3. LU Factorization with Partial Pivoting.

Input: Nonsingular matrix $A \in \mathbb{C}^{n \times n}$

Output: Permutation matrix P , unit lower triangular matrix L , upper triangular matrix U such that $PA = LU$

1. If $n = 1$ then $P \equiv 1$, $L \equiv 1$ and $U \equiv A$.
2. If $n > 1$ then choose a permutation matrix P_n such that

$$P_n A = \begin{matrix} & & 1 & \dots & n-1 \\ & & \alpha & & a \\ & & d & & A_{n-1} \end{matrix},$$

where α has the largest magnitude among all elements in the leading column, i.e. $|\alpha| \geq \|d\|_\infty$, and factor

$$P_n A = \begin{pmatrix} 1 & 0 \\ l & I_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & S \end{pmatrix},$$

where $l \equiv d\alpha^{-1}$ and $S \equiv A_{n-1} - la$.

3. Compute $P_{n-1}S = L_{n-1}U_{n-1}$, where P_{n-1} is a permutation matrix, L_{n-1} is unit lower triangular, and U_{n-1} is upper triangular.
4. Then

$$P \equiv \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1} \end{pmatrix} P_n, \quad L \equiv \begin{pmatrix} 1 & 0 \\ P_{n-1}l & L_{n-1} \end{pmatrix}, \quad U \equiv \begin{pmatrix} \alpha & a \\ 0 & U_{n-1} \end{pmatrix}.$$

Remark 3.20.

- Each iteration of step 2 in Algorithm 3.3 determines one column of L and one row of U .
- Partial pivoting ensures that the magnitude of the multipliers is bounded by one, i.e. $\|l\|_\infty \leq 1$ in Step 2 of Algorithm 3.3. Therefore all elements of L have magnitude less than or equal to one.
- The scalar α is called a pivot, and the matrix $S = A_{n-1} - d\alpha^{-1}a$ is a Schur complement. We already encountered Schur complements in Fact 1.14, as part of the inverse of a partitioned matrix. In this particular Schur complement S the matrix $d\alpha^{-1}a$ is an outer product.
- The multipliers can be easily recovered from L , because they are elements of L . Step 4 of Algorithm 3.3 shows that the first column of L contains the multipliers $P_{n-1}l$ that zero out elements in the first column. Similarly, column i of L contains the multipliers that zero out elements in column i . However, the multipliers cannot be easily recovered from L^{-1} .
- Step 4 of Algorithm 3.3 follows from $S = P_{n-1}^T L_{n-1} U_{n-1}$, extracting the permutation matrix,

$$P_n A = \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ P_{n-1}l & I_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & L_{n-1}U_{n-1} \end{pmatrix}$$

and separating lower and upper triangular parts

$$\begin{pmatrix} 1 & 0 \\ P_{n-1}l & I_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & L_{n-1}U_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ P_{n-1}l & L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & U_{n-1} \end{pmatrix}.$$

- In the vector $P_{n-1}l$, the permutation P_{n-1} reorders the multipliers l , but does not change their values. To combine all permutations into a single permutation matrix P , we have to pull all permutation matrices in front of the lower triangular matrix. This, in turn, requires reordering the multipliers in earlier steps.

Fact 3.21 (LU Factorization with Partial Pivoting) Every nonsingular matrix A has a factorization $PA = LU$, where P is a permutation matrix, L is unit lower triangular, and U is nonsingular upper triangular.

Proof. Perform an induction proof based on Algorithm 3.3. \square

A factorization $PA = LU$ is in general not unique because there are many choices for the permutation matrix.

With a factorization $PA = LU$, the rows of the linear system $Ax = b$ are rearranged, and the system to be solved is $PAx = Pb$. The process of solving this linear system is called *Gaussian elimination with partial pivoting*.

Algorithm 3.4. Gaussian Elimination with Partial Pivoting.

Input: Nonsingular matrix $A \in \mathbb{C}^{n \times n}$, vector $b \in \mathbb{C}^n$

Output: Solution of $Ax = b$

1. Factor $PA = LU$ with Algorithm 3.3.
2. Solve the system $Ly = Pb$.
3. Solve the system $Ux = y$.

The next bound implies that Gaussian elimination with partial pivoting is stable in exact arithmetic, if the elements of U are not much larger in magnitude than those of A .

Corollary 3.22 (Stability in Exact Arithmetic of Gaussian Elimination with Partial Pivoting). *If $A \in \mathbb{C}^{n \times n}$ is nonsingular, $Ax = b$, and*

$$\begin{aligned} P(A + E) &= LU, & \epsilon_A &= \frac{\|E\|_\infty}{\|A\|_\infty} \\ Ly &= Pb + r_L, & \epsilon_L &= \frac{\|r_L\|_\infty}{\|L\|_\infty \|y\|_\infty} \\ Uz &= y + r_U, & \epsilon_U &= \frac{\|r_U\|_\infty}{\|U\|_\infty \|z\|_\infty}. \end{aligned}$$

where $y \neq 0$ and $z \neq 0$ then

$$\frac{\|z - x\|_\infty}{\|z\|_\infty} \leq \kappa_\infty(A) (\epsilon_A + \epsilon), \quad \text{where } \epsilon = n \frac{\|U\|_\infty}{\|A\|_\infty} (\epsilon_U + \epsilon_L(1 + \epsilon_U)).$$

Proof. Apply Fact 3.17 to $A + E = S_1 S_2$, where $S_1 = P^T L$ and $S_2 = U$. Permutation matrices do not change p-norms, see the Exercise (iv) in §2.6, so that $\|P^T L\|_\infty = \|L\|_\infty$. Because the multipliers are the elements of L , and $|l_{ij}| \leq 1$ with partial pivoting, we get $\|L\|_\infty \leq n$. \square

The ratio $\|U\|_\infty / \|A\|_\infty$ represents the element growth during Gaussian elimination. In practice, $\|U\|_\infty / \|A\|_\infty$ tends to be small, but there are $n \times n$ matrices for which $\|U\|_\infty / \|A\|_\infty = 2^{n-1}/n$ is possible, see Exercise 2. If $\|U\|_\infty \gg \|A\|_\infty$ then Gaussian elimination is unstable.

Exercises

- (i) Determine the LU factorization of a non-singular lower triangular matrix A . Express the elements of L and U in terms of the elements of A .
- (ii) Determine a factorization $A = LU$ when A is upper triangular.
- (iii) For

$$A = \begin{pmatrix} 0 & 0 \\ A_1 & 0 \end{pmatrix},$$

with A_1 nonsingular, determine a factorization $PA = LU$ where L is unit lower triangular and U is upper triangular.

- (iv) LDU factorization.

One can make a LU factorization more symmetric by requiring that both triangular matrices have ones on the diagonal, and factoring $A = LD\tilde{U}$, where L is unit lower triangular, D is diagonal, and \tilde{U} is unit upper triangular.

Given a LU factorization $A = LU$, express the diagonal elements d_{ii} of D and the elements \tilde{u}_{ij} in terms of elements of U .

- (v) Block LU factorization.

Suppose we can partition the invertible matrix A as

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

where A_{11} is invertible. Verify that A has the block factorization $A = LU$ where

$$L = \begin{pmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{pmatrix}, \quad U = \begin{pmatrix} A_{11} & A_{12} \\ 0 & S \end{pmatrix},$$

and $S \equiv A_{22} - A_{21}A_{11}^{-1}A_{12}$ is a *Schur complement*. Note that L is unit lower triangular. However U is only block upper triangular, because A_{11} and S are in general not triangular. Hence a block LU factorization is not the same as a LU factorization.

Determine a block LDU factorization $A = LDU$, where L is unit lower triangular, U is unit upper triangular, and D is block diagonal.

- (vi) The matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 3 & 4 \\ 1 & 2 & 1 & 2 \\ 3 & 4 & 3 & 4 \end{pmatrix}$$

does not have a LU factorization. However, it does have a block LU factorization $A = LU$ with

$$A_{11} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Determine L and U .

- (vii) UL Factorization.

Analogous to Algorithm 3.3, present an algorithm that factors any square matrix A into $PA = UL$, where P is a permutation matrix, U is unit upper triangular, and L is lower triangular.

1. Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, and P a permutation matrix such that

$$PA = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

with A_{11} nonsingular. Show: If all elements of $A_{21}A_{11}^{-1}$ are less than one in magnitude then

$$\kappa_{\infty}(A_{22} - A_{21}A_{11}^{-1}A_{12}) \leq n^2\kappa_{\infty}(A).$$

2. Compute the LU factorization of the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & & & 1 \\ -1 & 1 & & 1 \\ -1 & -1 & 1 & 1 \\ \vdots & & \ddots & \vdots \\ -1 & \dots & \dots & -1 & 1 \end{pmatrix}.$$

Show that pivoting is not necessary. Determine the one norms of A and U .

3. Let $A \in \mathbb{C}^{n \times n}$ and $A + uv^*$ be nonsingular, where $u, v \in \mathbb{C}^n$. Show how to solve $(A + uv^*)x = b$ using two linear system solves with A , two inner products, one scalar vector multiplication, and one vector addition.
4. This problem shows that if Gaussian elimination with partial pivoting encounters a small pivot, then A must be ill-conditioned.

Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, and $PA = LU$, where P is a permutation matrix, L is unit triangular with elements $|l_{ij}| \leq 1$ and U is upper triangular with elements u_{ij} . Show that $\kappa_{\infty}(A) \geq \|A\|_{\infty} / \min_j |u_{jj}|$.

5. The following matrices G are generalizations of the lower triangular matrices in the LU factorization. The purpose of G is to transform all elements of a column vector to zero, except for the k th element.

Let $G = I_n - ge_k^T$, where $g \in \mathbb{C}^n$ and $1 \leq k \leq n$. Which conditions do the elements of g have to satisfy so that G is invertible? Determine G^{-1} when it exists.

Given an index k and a vector $x \in \mathbb{C}^n$, which conditions do the elements of x have to satisfy so that $Gx = e_k$? Determine the vector g when it exists.

3.6 Cholesky Factorization

It would seem natural that a Hermitian matrix should have a factorization that reflects the symmetry of the matrix. For a $n \times n$ Hermitian matrix, we need to store only $n(n+1)/2$ elements, and it would be efficient if the same were true of the factorization. Unfortunately this is not possible in general. For instance, the matrix

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

is nonsingular and Hermitian. But it cannot be factored into a lower times upper triangular matrix, as illustrated in Example 3.19. Fortunately, a certain class of matrices, so-called *Hermitian positive definite* matrices do admit a symmetric factorization.

Definition 3.23. A Hermitian matrix $A \in \mathbb{C}^{n \times n}$ is positive definite if $x^*Ax > 0$ for all $x \in \mathbb{C}^n$ with $x \neq 0$.

A Hermitian matrix $A \in \mathbb{C}^{n \times n}$ is positive semi-definite if $x^*Ax \geq 0$ for all $x \in \mathbb{C}^n$.

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if $x^T Ax > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$, and positive semi-definite if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$.

A positive semi-definite matrix A can have $x^*Ax = 0$ for $x \neq 0$.

Example. The 2×2 Hermitian matrix

$$A = \begin{pmatrix} 1 & \beta \\ \bar{\beta} & 1 \end{pmatrix}$$

is positive definite if $|\beta| < 1$, and positive semi-definite if $|\beta|^2 = 1$. \square

We derive several properties of Hermitian positive definite matrices. We start by showing that all Hermitian positive definite matrices are nonsingular.

Fact 3.24 If $A \in \mathbb{C}^{n \times n}$ is Hermitian positive definite then A is nonsingular.

Proof. Suppose to the contrary that A were singular. Then $Ax = 0$ for some $x \neq 0$, implying $x^*Ax = 0$ for some $x \neq 0$, which contradicts the positive definiteness of A , i.e. $x^*Ax > 0$ for all $x \neq 0$. \square

Hermitian positive definite matrices have positive diagonal elements.

Fact 3.25 If $A \in \mathbb{C}^{n \times n}$ is Hermitian positive definite then its diagonal elements are positive.

Proof. Since A is positive definite, we have $x^*Ax > 0$ for any $x \neq 0$, and in particular $0 < e_j^* A e_j = a_{jj}$, $1 \leq j \leq n$. \square

Below is a transformation that preserves Hermitian positive definiteness.

Fact 3.26 If $A \in \mathbb{C}^{n \times n}$ is Hermitian positive definite, and $B \in \mathbb{C}^{n \times n}$ is nonsingular then B^*AB is also Hermitian positive definite.

Proof. The matrix B^*AB is Hermitian because A is Hermitian. Since B is nonsingular, $y = Bx \neq 0$ if and only if $x \neq 0$. Hence

$$x^* B^* A B x = (Bx)^* A (Bx) = y^* A y > 0$$

for any vector $y \neq 0$, so that B^*AB is positive definite. \square

At last we show that principal submatrices and Schur complements inherit Hermitian positive definiteness.

Fact 3.27 If $A \in \mathbb{C}^{n \times n}$ is Hermitian positive definite then its leading principal submatrices and Schur complements are also Hermitian positive definite.

Proof. Let B be a $k \times k$ principal submatrix of A , for some $1 \leq k \leq n - 1$. The submatrix B is Hermitian because it is a principal submatrix of a Hermitian matrix. To keep the notation simple, we permute the rows and columns of A so that the submatrix B occupies the leading rows and columns. That is, let P be a permutation matrix, and partition

$$\hat{A} = P^T A P = \begin{pmatrix} B & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix}.$$

Fact 3.26 implies that \hat{A} is also Hermitian positive definite. Thus $x^* \hat{A} x > 0$ for any vector $x \neq 0$. In particular, let $x = \begin{pmatrix} y \\ 0 \end{pmatrix}$ for $y \in \mathbb{C}^k$. Then for any $y \neq 0$ we have

$$0 < x^* \hat{A} x = (y^* \ 0) \begin{pmatrix} B & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix} \begin{pmatrix} y \\ 0 \end{pmatrix} = y^* B y.$$

This means $y^* B y > 0$ for $y \neq 0$, so that B is positive definite. Since the submatrix B is a principal submatrix of a Hermitian matrix, B is also Hermitian. Therefore any principal submatrix B of A is Hermitian positive definite.

Now we prove Hermitian positive definiteness for Schur complements. Fact 3.24 implies that B is nonsingular. Hence we can set

$$L = \begin{pmatrix} I_k & 0 \\ -A_{12}^* B^{-1} & I_{n-k} \end{pmatrix}$$

so that

$$L \hat{A} L^* = \begin{pmatrix} B & 0 \\ 0 & S \end{pmatrix}, \quad \text{where } S = A_{22} - A_{12}^* B^{-1} A_{12}.$$

Since L is unit lower triangular, it is nonsingular. From Fact 3.26 follows then that $L \hat{A} L^*$ is Hermitian positive definite. Earlier in this proof we showed that principal submatrices of Hermitian positive definite matrices are Hermitian positive definite, thus the Schur complement S must be Hermitian positive definite. \square

Now we have all the tools we need to factor Hermitian positive definite matrices. The following algorithm produces a symmetric factorization $A = LL^*$ for a Hermitian positive definite matrix A . The algorithm exploits the fact that the diagonal elements of A are positive and the Schur complements are Hermitian positive definite.

Definition 3.28. Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite. A factorization $A = LL^*$, where L is (lower or upper) triangular with positive diagonal elements is called a Cholesky factorization of A .

Below we compute a lower-upper Cholesky factorization $A = LL^*$ where L is a lower triangular matrix.

Algorithm 3.5. Cholesky Factorization.

Input: Hermitian positive definite matrix $A \in \mathbb{C}^{n \times n}$

Output: Lower triangular matrix L with positive diagonal elements such that $A = LL^*$

1. If $n = 1$ then $L \equiv \sqrt{A}$.
2. If $n > 1$ partition and factor

$$A = \begin{array}{c|c} 1 & n-1 \\ \hline \alpha & a^* \\ a & A_{n-1} \end{array} = \begin{pmatrix} \alpha^{1/2} & 0 \\ a\alpha^{-1/2} & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \alpha^{1/2} & \alpha^{-1/2}a^* \\ 0 & I_{n-1} \end{pmatrix},$$

where $S \equiv A_{n-1} - a\alpha^{-1}a^*$.

3. Compute $S = L_{n-1}L_{n-1}^*$, where L_{n-1} is lower triangular with positive diagonal elements.
4. Then

$$L \equiv \begin{pmatrix} \alpha^{1/2} & 0 \\ a\alpha^{-1/2} & L_{n-1} \end{pmatrix}.$$

A Cholesky factorization of a positive matrix is unique.

Fact 3.29 (Uniqueness of Cholesky factorization) Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite. If $A = LL^*$ where L is lower triangular with positive diagonal then L is unique. Similarly, if $A = LL^*$ where L is upper triangular with positive diagonal elements then L is unique.

Proof. This can be shown in the same way as the uniqueness of the LU factorization. \square

The following result shows that one can use a Cholesky factorization to determine whether a Hermitian matrix is positive definite.

Fact 3.30 Let $A \in \mathbb{C}^{n \times n}$ be Hermitian. A is positive definite if and only if $A = LL^*$ where L is triangular with positive diagonal elements.

Proof. Algorithm 3.5 shows that if A is positive definite then $A = LL^*$. Now assume that $A = LL^*$. Since L is triangular with positive diagonal elements, it is nonsingular. Therefore $Lx \neq 0$ for $x \neq 0$, and $x^*Ax = \|L^*x\|_2^2 > 0$. \square

The next bound shows that a Cholesky solver is numerically stable in exact arithmetic.

Corollary 3.31 (Stability of Cholesky Solver). *Let $A \in \mathbb{C}^{n \times n}$ and $A + E$ be Hermitian positive definite matrices, $Ax = b$, $b \neq 0$, and*

$$\begin{aligned} A + E &= LL^*, & \epsilon_A &= \frac{\|E\|_2}{\|A\|_2} \\ Ly &= b + r_1, & \epsilon_1 &= \frac{\|r_1\|_2}{\|b\|_2} \\ L^*z &= y + r_2, & \epsilon_2 &= \frac{\|r_2\|_2}{\|y\|_2}. \end{aligned}$$

If $\|A^{-1}\|_2\|E\|_2 \leq 1/2$ then

$$\frac{\|z - x\|_2}{\|x\|_2} \leq 2\kappa_2(A) (\epsilon_A + \epsilon_1 + \epsilon_2(1 + \epsilon_1)).$$

Proof. Apply Fact 3.14 to $A + E$, where $S_1 = L$ and $S_2 = L^*$. The stability factor is $\|L^{-*}\|_2\|L^{-1}\|_2/\|(A + E)^{-1}\|_2 = 1$ because Fact 2.19 implies

$$\|(A + E)^{-1}\|_2 = \|L^{-*}L^{-1}\|_2 = \|L^{-1}\|_2^2 = \|L^{-*}\|_2\|L^{-1}\|_2.$$

□

Exercises

- (i) The magnitude of an offdiagonal element of a Hermitian positive definite matrix is bounded by the geometric mean of the corresponding diagonal elements. Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite. Show: $|a_{ij}| < \sqrt{a_{ii}a_{jj}}$ for $i \neq j$. Hint: Use the positive definiteness of the Schur complement.
- (ii) The magnitude of an offdiagonal element of a Hermitian positive definite matrix is bounded by the arithmetic mean of the corresponding diagonal elements. Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite. Show: $|a_{ij}| \leq (a_{ii} + a_{jj})/2$ for $i \neq j$. Hint: Use the relation between arithmetic and geometric mean.
- (iii) The largest element in magnitude of a Hermitian positive definite matrix is on the diagonal. Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite. Show: $\max_{1 \leq i, j \leq n} |a_{ij}| = \max_{1 \leq i \leq n} a_{ii}$.
- (iv) Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite. Show: A^{-1} is also positive definite.
- (v) Modify Algorithm 3.5 so it computes a factorization $A = LDL^*$ for a Hermitian positive definite matrix A , where D is diagonal and L is unit lower triangular.

- (vi) Upper-Lower Cholesky Factorization. Modify Algorithm 3.5 so it computes a factorization $A = L^*L$ for a Hermitian positive definite matrix A , where L is lower triangular with positive diagonal elements.
- (vii) Block Cholesky factorization. Partition the Hermitian positive definite matrix A as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Analogous to the block LU factorization in Exercise (v) of Section 3.5 determine a factorization $A = LL^*$, where L is block lower triangular. That is, L is of the form

$$L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix},$$

where L_{11} and L_{22} are in general not lower triangular.

- (viii) Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

be Hermitian positive definite. Show:

$$\|A_{22} - A_{21}A_{11}^{-1}A_{12}\|_2 \leq \|A\|_2,$$

and

$$\kappa_2(A_{22} - A_{21}A_{11}^{-1}A_{12}) \leq \kappa_2(A).$$

- (ix) Prove: $A = MM^*$ for some nonsingular matrix M if and only if A is Hermitian positive definite.
- (x) Generalized Cholesky Factorization. Let $M \in \mathbb{C}^{n \times n}$ be Hermitian positive-definite. Prove: If $M = M_1^*M_1 = M_2^*M_2$ for square matrices M_1 and M_2 then there exists a unitary matrix Q such that $M_2 = QM_1$.
- (xi) Let $M = A + \iota B$ be Hermitian positive definite, where $\iota^2 = -1$, and A and B are real square matrices. Show that the matrix

$$C = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

is real symmetric positive definite.

3.7 QR Factorization

The QR factorization is a matrix factorization where one of the factors is unitary, and the other one is triangular. We derive the existence of a QR factorization from the Cholesky factorization.

Fact 3.32 Every nonsingular matrix $A \in \mathbb{C}^{n \times n}$ has a unique factorization $A = QR$, where Q is unitary and R is upper triangular with positive diagonal elements.

Proof. Since A is nonsingular, $Ax \neq 0$ for $x \neq 0$, and $x^*A^*Ax = \|Ax\|_2^2 > 0$, which implies that $M = A^*A$ is Hermitian positive definite. Let $M = LL^*$ be a Cholesky factorization of M , where L is lower triangular with positive diagonal elements. Then $M = A^*A = LL^*$. Multiplying by A^{-*} on the left gives $A = QR$, where $Q = A^{-*}L$, and where $R = L^*$ is upper triangular with positive diagonal elements. Exercise (ix) in Section 3.6 shows that Q is unitary.

The uniqueness of the QR factorization follows from the uniqueness of the Cholesky factorization, as well as from Exercise 6 in Section 1.13. \square

The bound below shows that a QR solver is numerically stable in exact arithmetic.

Corollary 3.33 (Stability of QR Solver). *Let $A \in \mathbb{C}^{n \times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and*

$$\begin{aligned} A + E &= QR, & \epsilon_A &= \frac{\|E\|_2}{\|A\|_2} \\ Qy &= b + r_1, & \epsilon_1 &= \frac{\|r_1\|_2}{\|b\|_2} \\ Rz &= y + r_2, & \epsilon_2 &= \frac{\|r_2\|_2}{\|y\|_2}. \end{aligned}$$

If $\|A^{-1}\|_2\|E\|_2 \leq 1/2$ then

$$\frac{\|z - x\|_2}{\|x\|_2} \leq 2\kappa_2(A) (\epsilon_A + \epsilon_1 + \epsilon_2(1 + \epsilon_1)).$$

Proof. Apply Fact 3.14 to $A + E$, where $S_1 = Q$ and $S_2 = R$. The stability factor is $\|R^{-1}\|_2\|Q^*\|_2/\|(A + E)^{-1}\|_2 = 1$ because Exercise (v) in Section 2.6 implies $\|Q^*\|_2 = 1$ and $\|(A + E)^{-1}\|_2 = \|R^{-1}\|_2$. \square

There are many ways to compute a QR factorization. Here we present an algorithm that is based on *Givens rotations*, see Definition 1.17. Givens rotations are unitary, see Example 1.16, and they are often used to introduce zeros into matrices. Let's start by using a Givens rotation to introduce a single zero into a vector.

Example. Let $x, y \in \mathbb{C}$.

$$\begin{pmatrix} c & s \\ -\bar{s} & \bar{c} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}, \quad \text{where } d = \sqrt{|x|^2 + |y|^2}.$$

If $x = y = 0$ then $c = 1$ and $s = 0$; otherwise $c = \bar{x}/d$ and $s = \bar{y}/d$. That is, if both components of the vector are zero then there is nothing to do and the unitary matrix is the identity. Note that $d \geq 0$, and $|c|^2 + |s|^2 = 1$. \square

When introducing zeros into a longer vector, we embed each Givens rotation in an identity matrix.

Example. Suppose we want to zero out elements 2, 3 and 4 in a 4×1 vector with a unitary matrix. We can apply three Givens rotations in the following order.

1. Apply a Givens rotation to rows 3 and 4 to zero out element 4,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -\bar{s}_4 & \bar{c}_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ y_3 \\ 0 \end{pmatrix}$$

where $y_3 = \sqrt{|x_3|^2 + |x_4|^2} \geq 0$. If $x_4 = x_3 = 0$ then $c_4 = 1$ and $s_4 = 0$, otherwise $c_4 = \bar{x}_3/y_3$ and $s_4 = \bar{x}_4/y_3$.

2. Apply a Givens rotation to rows 2 and 3 to zero out element 3,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -\bar{s}_3 & \bar{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_3 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_2 \\ 0 \\ 0 \end{pmatrix}$$

where $y_2 = \sqrt{|x_2|^2 + |y_3|^2} \geq 0$. If $y_3 = x_2 = 0$ then $c_3 = 1$ and $s_3 = 0$, otherwise $c_3 = \bar{x}_2/y_2$ and $s_3 = \bar{y}_3/y_2$.

3. Apply a Givens rotation to rows 1 and 2 to zero out element 2,

$$\begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\bar{s}_2 & \bar{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} y_1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

where $y_1 = \sqrt{|x_1|^2 + |y_2|^2} \geq 0$. If $y_2 = x_1 = 0$ then $c_2 = 1$ and $s_2 = 0$, otherwise $c_2 = \bar{x}_1/y_1$ and $s_2 = \bar{y}_2/y_1$.

Therefore $Qx = y_1 e_1$, where $y_1 = \|Qx\|_2$ and

$$Q = \begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\bar{s}_2 & \bar{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -\bar{s}_3 & \bar{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -\bar{s}_4 & \bar{c}_4 \end{pmatrix}.$$

□

There are many possible orders in which to apply Givens rotations, and Givens rotations don't have to operate on adjacent rows either. The example below illustrates this.

Example. Here is another way to zero out elements 2, 3 and 4 in a 4×1 vector. We can apply three Givens rotations that all involve the leading row.

1. Apply a Givens rotation to rows 1 and 4 to zero out element 4,

$$\begin{pmatrix} c_4 & 0 & 0 & s_4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\bar{s}_4 & 0 & 0 & \bar{c}_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} y_1 \\ x_2 \\ x_3 \\ 0 \end{pmatrix}$$

where $y_1 = \sqrt{|x_1|^2 + |x_4|^2} \geq 0$. If $x_4 = x_1 = 0$ then $c_4 = 1$ and $s_4 = 0$, otherwise $c_4 = \bar{x}_1/y_1$ and $s_4 = \bar{x}_4/y_1$.

2. Apply a Givens rotation to rows 1 and 3 to zero out element 3,

$$\begin{pmatrix} c_3 & 0 & s_3 & 0 \\ 0 & 1 & 0 & 0 \\ -\bar{s}_3 & 0 & \bar{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ x_2 \\ x_3 \\ 0 \end{pmatrix} = \begin{pmatrix} z_1 \\ x_2 \\ 0 \\ 0 \end{pmatrix}$$

where $z_1 = \sqrt{|y_1|^2 + |x_3|^2} \geq 0$. If $x_3 = y_1 = 0$ then $c_3 = 1$ and $s_3 = 0$, otherwise $c_3 = \bar{y}_1/z_1$ and $s_3 = \bar{x}_3/z_1$.

3. Apply a Givens rotation to rows 1 and 2 to zero out element 2,

$$\begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\bar{s}_2 & \bar{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

where $u_1 = \sqrt{|z_1|^2 + |x_2|^2} \geq 0$. If $x_2 = z_1 = 0$ then $c_2 = 1$ and $s_2 = 0$, otherwise $c_2 = \bar{z}_1/u_1$ and $s_2 = \bar{x}_2/u_1$.

Therefore $Qx = u_1 e_1$, where $u_1 = \|Qx\|_2$ and

$$Q = \begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\bar{s}_2 & \bar{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_3 & 0 & s_3 & 0 \\ 0 & 1 & 0 & 0 \\ -\bar{s}_3 & 0 & \bar{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_4 & 0 & 0 & s_4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\bar{s}_4 & 0 & 0 & \bar{c}_4 \end{pmatrix}$$

□

The preceding examples demonstrate that if a Givens rotation operates on rows i and j , then the c and s elements occupy positions (i, i) , (i, j) , (j, i) and (j, j) .

At last here is a sketch of how one can reduce a square matrix to upper triangular form by means of Givens rotations

Example. We introduce zeros one column at a time, from left to right, and within a column from bottom to top. The Givens rotations operate on adjacent rows. Elements that can be nonzero are represented by *. Elements that were affected by

the i th Givens rotation have the label i . We start by introducing zeros into column 1,

$$\begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{1} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{2} \begin{pmatrix} * & * & * & * \\ 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{3} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 \\ 0 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

Now we introduce zeros into column 2, and then into column 3,

$$\begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 \\ 0 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{4} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 \\ 0 & 4 & 4 & 4 \\ 0 & 0 & 4 & 4 \end{pmatrix} \xrightarrow{5} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 5 & 5 & 5 \\ 0 & 0 & 5 & 5 \\ 0 & 0 & 4 & 4 \end{pmatrix} \xrightarrow{6} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 5 & 5 & 5 \\ 0 & 0 & 6 & 6 \\ 0 & 0 & 0 & 6 \end{pmatrix}$$

□

Below is the general algorithm.

Algorithm 3.6. QR Factorization for Nonsingular Matrices.

Input: Nonsingular matrix $A \in \mathbb{C}^{n \times n}$

Output: Unitary matrix $Q \in \mathbb{C}^{n \times n}$ and upper triangular matrix $R \in \mathbb{C}^{n \times n}$ with positive diagonal elements such that $A = QR$

1. If $n = 1$ then $Q \equiv A/|A|$ and $R \equiv |A|$
2. If $n > 1$ zero out elements $n, n-1, \dots, 2$ in column 1 of A as follows.

(i) Set $(b_{n1} \ b_{n2} \ \dots \ b_{nn}) = (a_{n1} \ a_{n2} \ \dots \ a_{nn})$.

(ii) For $i = n, n-1, \dots, 2$

Zero out element $(i, 1)$ by applying a rotation to rows i and $i-1$

$$\begin{pmatrix} c_i & s_i \\ -\bar{s}_i & \bar{c}_i \end{pmatrix} \begin{pmatrix} a_{i-1,1} & a_{i-1,2} & \dots & a_{i-1,n} \\ b_{i1} & b_{i2} & \dots & b_{in} \end{pmatrix} = \begin{pmatrix} b_{i-1,1} & b_{i-1,2} & \dots & b_{i-1,n} \\ 0 & \hat{a}_{i2} & \dots & \hat{a}_{in} \end{pmatrix}$$

where $b_{i-1,1} \equiv \sqrt{|b_{i1}|^2 + |a_{i-1,1}|^2}$. If $b_{i1} = a_{i-1,1} = 0$ then $c_i \equiv 1$ and $s_i \equiv 0$; otherwise $c_i \equiv \bar{a}_{i-1,1}/b_{i-1,1}$ and $s_i \equiv \bar{b}_{i1}/b_{i-1,1}$.

(iii) Multiply all $n-1$ rotations

$$Q_n^* \equiv \begin{pmatrix} c_2 & s_2 & 0 \\ -\bar{s}_2 & \bar{c}_2 & 0 \\ 0 & 0 & I_{n-2} \end{pmatrix} \dots \begin{pmatrix} I_{n-2} & 0 & 0 \\ 0 & c_n & s_n \\ 0 & -\bar{s}_n & \bar{c}_n \end{pmatrix}$$

(iv) Partition the transformed matrix

$$Q_n^* A = \begin{pmatrix} r_{11} & r^* \\ 0 & \hat{A} \end{pmatrix}, \quad \text{where } \hat{A} \equiv \begin{pmatrix} \hat{a}_{22} & \dots & \hat{a}_{2n} \\ \vdots & & \vdots \\ \hat{a}_{n2} & \dots & \hat{a}_{nn} \end{pmatrix},$$

$r^* \equiv (b_{12} \ \dots \ b_{1n})$, and $r_{11} \equiv b_{11} > 0$.

3. Compute $\hat{A} = Q_{n-1}R_{n-1}$, where Q_{n-1} is unitary and R_{n-1} is upper triangular with positive diagonal elements.
4. Then

$$Q \equiv Q_n \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix}, \quad R \equiv \begin{pmatrix} r_{11} & r^* \\ 0 & R_{n-1} \end{pmatrix}.$$

Exercises

- (i) Determine the QR factorization of a real upper triangular matrix.
- (ii) QR Factorization of Outer Product.
Let $x, y \in \mathbb{C}^n$, and apply Algorithm 3.6 to xy^* . How many Givens rotations do you have to apply at the most? What does the upper triangular matrix R look like?
- (iii) Let $A \in \mathbb{C}^{n \times n}$ be a tridiagonal matrix, that is, only elements a_{ii} , $a_{i+1,i}$ and $a_{i,i+1}$ can be nonzero; all other elements are zero. We want to compute a QR factorization $A = QR$ with $n - 1$ Givens rotations. In which order do the elements have to be zeroed out, on which rows do the rotations act, and which elements of R can be nonzero?
- (iv) QL Factorization.
Show: Every nonsingular matrix $A \in \mathbb{C}^{n \times n}$ has a unique factorization $A = QL$, where Q is unitary, and L is lower triangular with positive diagonal elements.
- (v) Computation of QL Factorization.
Suppose we want to compute the QL factorization of a nonsingular matrix $A \in \mathbb{C}^{n \times n}$ with Givens rotations. In which order do the elements have to be zeroed out, and on which rows do the rotations act?
- (vi) The elements in a Givens rotation

$$G = \begin{pmatrix} c & s \\ -\bar{s} & \bar{c} \end{pmatrix}$$

are named to invoke an association with sine and cosine, because $|c|^2 + |s|^2 = 1$. One can also express the elements in terms of tangents or cotangents. Let

$$G \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}, \quad \text{where } d = \sqrt{|x|^2 + |y|^2}.$$

Show the following: If $|y| > |x|$ then

$$\tau = \frac{x}{y}, \quad s = \frac{\bar{y}}{|y|} \frac{1}{\sqrt{1 + |\tau|^2}}, \quad c = \bar{\tau}s,$$

and if $|x| > |y|$ then

$$\tau = \frac{y}{x}, \quad c = \frac{\bar{x}}{|x|} \frac{1}{\sqrt{1 + |\tau|^2}}, \quad s = \bar{\tau}c.$$

(vii) Householder Reflections.

Here is another way to introduce zeros into a vector without changing its two norm. Let $x \in \mathbb{C}^n$ and $x_1 \neq 0$. Define $Q = I - 2vv^*/v^*v$, where $v = x + \alpha\|x\|_2 e_1$ and $\alpha = x_1/|x_1|$. Show that Q is unitary and that $Qx = -\alpha\|x\|_2 e_1$. The matrix Q is called a *Householder reflection*.

(viii) Householder Reflections for Real Vectors.

Let $x, y \in \mathbb{R}^n$ with $\|x\|_2 = \|y\|_2$. Show how to choose a vector v in the Householder reflection so that $Qx = y$.

3.8 QR Factorization of Tall and Skinny Matrices

We look at rectangular matrices $A \in \mathbb{C}^{m \times n}$ with at least as many rows as columns, i.e. $m \geq n$. If A is involved in a linear system $Ax = b$, then we must have $b \in \mathbb{C}^m$ and $x \in \mathbb{C}^n$. Such linear systems do not always have a solution; and if they do happen to have a solution then the solution may not be unique.

Example. If

$$A = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

then the linear system $Ax = b$ has a solution only for those b all of whose elements are the same, i.e. $\beta = b_1 = b_2 = b_3$. In this case the solution is $x = \beta$. \square

Fortunately there is one righthand side for which a linear system $Ax = b$ always has a solution, namely $b = 0$. That is, $Ax = 0$ always has the solution $x = 0$. However $x = 0$ may not be the only solution.

Example. If

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

then $Ax = 0$ has infinitely many solutions $x = (x_1 \ x_2)^T$ with $x_1 = x_2$. \square

We distinguish matrices A where $x = 0$ is the unique solution for $Ax = 0$.

Definition 3.34. Let $A \in \mathbb{C}^{m \times n}$. The columns of A are linearly independent if $Ax = 0$ implies $x = 0$. If $Ax = 0$ has infinitely many solutions, then the columns of A are linearly dependent.

Example.

- The columns of a nonsingular matrix A are linearly independent.
- If A is nonsingular then the matrix $\begin{pmatrix} A \\ 0 \end{pmatrix}$ has linearly independent columns.

- Let $x \in \mathbb{C}^n$. If $x \neq 0$ then x consists of a single, linearly independent column. If $x = 0$ then x is linearly dependent.
- If $A \in \mathbb{C}^{m \times n}$ with $A^*A = I_n$ then A has linearly independent columns. This is because multiplying $Ax = 0$ on the left by A^* implies $x = 0$.
- If the linear system $Ax = b$ has a solution x , then the matrix $B = \begin{pmatrix} A & b \end{pmatrix}$ has linearly dependent columns. That is because $B \begin{pmatrix} x \\ -1 \end{pmatrix} = 0$.

□

How can we tell whether a tall and skinny matrix has linearly independent columns? We can use a QR factorization.

Algorithm 3.7. QR Factorization for Tall and Skinny Matrices.

Input: Matrix $A \in \mathbb{C}^{m \times n}$ with $m \geq n$

Output: Unitary matrix $Q \in \mathbb{C}^{m \times m}$ and upper triangular matrix $R \in \mathbb{C}^{n \times n}$ with nonnegative diagonal elements such that $A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$

1. If $n = 1$ then Q is a unitary matrix that zeros out elements $2, \dots, m$ of A , and $R \equiv \|A\|_2$.
2. If $n > 1$ then, as in Algorithm 3.6, determine a unitary matrix $Q_m \in \mathbb{C}^{m \times m}$ to zero out elements $2, \dots, m$ in column 1 of A , so that

$$Q_m^* A = \begin{pmatrix} r_{11} & r^* \\ 0 & \hat{A} \end{pmatrix},$$

where $r_{11} \geq 0$ and $\hat{A} \in \mathbb{C}^{(m-1) \times (n-1)}$.

3. Compute $\hat{A} = Q_{m-1} \begin{pmatrix} R_{n-1} \\ 0 \end{pmatrix}$, where $Q_{m-1} \in \mathbb{C}^{(m-1) \times (m-1)}$ is unitary, and $R_{n-1} \in \mathbb{C}^{(n-1) \times (n-1)}$ is upper triangular with nonnegative diagonal elements.
4. Then

$$Q \equiv Q_m \begin{pmatrix} 1 & 0 \\ 0 & Q_{m-1} \end{pmatrix}, \quad R \equiv \begin{pmatrix} r_{11} & r^* \\ 0 & R_{n-1} \end{pmatrix}.$$

Fact 3.35 Let $A \in \mathbb{C}^{m \times n}$ with $m \geq n$, and $A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$ where $Q \in \mathbb{C}^{m \times m}$ is unitary, and $R \in \mathbb{C}^{n \times n}$ is upper triangular. Then A has linearly independent columns if and only if R has nonzero diagonal elements.

Proof. Since Q is nonsingular, $Ax = Q \begin{pmatrix} R \\ 0 \end{pmatrix} 0 \Rightarrow x = 0$ if and only if $Rx = 0 \Rightarrow x = 0$. This is the case if and only if R has is nonsingular and has nonzero diagonal elements. □

One can make a QR factorization more economical by reducing the storage and omitting part of the unitary matrix.

Fact 3.36 (Thin QR Factorization) If $A \in \mathbb{C}^{m \times n}$ with $m \geq n$ then there exists a matrix $Q_1 \in \mathbb{C}^{m \times n}$ with $Q_1^* Q_1 = I_n$, and an upper triangular matrix $R \in \mathbb{C}^{n \times n}$ with nonnegative diagonal elements so that $A = Q_1 R$.

Proof. Let $A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$ be a QR factorization as in Fact 3.35. Partition $Q = (Q_1 \ Q_2)$, where Q_1 has n columns. Then $A = Q_1 R$. \square

Definition 3.37. If $A \in \mathbb{C}^{m \times n}$ and $A^* A = I_n$ then the columns of A are orthonormal.

For a square matrix the thin QR decomposition is identical to the full QR decomposition.

Example 3.38 The columns of a unitary or an orthogonal matrix $A \in \mathbb{C}^{n \times n}$ are orthonormal because $A^* A = I_n$, and so are the rows because $AA^* = I_n$. This means, a square matrix with orthonormal columns must be a unitary matrix. A real square matrix with orthonormal columns is an orthogonal matrix. \blacksquare

Exercises

- (i) Let $A \in \mathbb{C}^{m \times n}$, $m \geq n$, with thin QR factorization $A = QR$. Show: $\|A\|_2 = \|R\|_2$.
- (ii) Uniqueness of Thin QR Factorization.
Let $A \in \mathbb{C}^{m \times n}$ have linearly independent columns. Show: If $A = QR$, where $Q \in \mathbb{C}^{m \times n}$ satisfies $Q^* Q = I_n$ and R is upper triangular with positive diagonal elements, then Q and R are unique.
- (iii) Generalization of Fact 3.35.
Let $A \in \mathbb{C}^{m \times n}$, $m \geq n$, and $A = B \begin{pmatrix} C \\ 0 \end{pmatrix}$, where $B \in \mathbb{C}^{m \times n}$ has linearly independent columns, and $C \in \mathbb{C}^{n \times n}$. Show: A has linearly independent columns if and only if C is nonsingular.
- (iv) Let $A \in \mathbb{C}^{m \times n}$ where $m > n$. Show: There exists a matrix $Z \in \mathbb{C}^{m \times (m-n)}$ such that $Z^* A = 0$.
- (v) Let $A \in \mathbb{C}^{m \times n}$, $m \geq n$, have a thin QR factorization $A = QR$. Express the k th column of A as a linear combination of columns of Q and elements of R . How many columns of Q are involved?
- (vi) Let $A \in \mathbb{C}^{m \times n}$, $m \geq n$, have a thin QR factorization $A = QR$. Determine a QR factorization of $A - Qe_1 e_1^* R$ from the QR factorization of A .
- (vii) Let $A = (a_1 \ \dots \ a_n)$ have linearly independent columns a_j , $1 \leq j \leq n$. Let $A = QR$ be a thin QR factorization where $Q = (q_1 \ \dots \ q_n)$ and R is

upper triangular with positive diagonal elements. Express the elements of R in terms of the columns a_j of A and the columns q_j of Q .

(viii) Let A be a matrix with linearly independent columns. Show how to compute the lower-upper Cholesky factorization of A^*A without forming the product A^*A .

(ix) Bessel's Inequality.

Let $V \in \mathbb{C}^{m \times n}$ with $V = (v_1 \ \dots \ v_n)$ have orthonormal columns, and let $x \in \mathbb{C}^m$. Show

$$\sum_{j=1}^n |v_j^* x|^2 \leq x^* x.$$

1. QR factorization with column pivoting.

This problem presents a method to compute QR factorizations of arbitrary matrices. Let $A \in \mathbb{C}^{m \times n}$ with $\text{rank}(A) = r$. Then there exists a permutation matrix P so that

$$AP = Q \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix},$$

where R_1 is an upper triangular nonsingular matrix.

(a) Show how to modify Algorithm 3.7 so that it computes such a factorization. In the first step, choose a permutation matrix P_n that brings the column with largest two-norm to the front, i.e.

$$\|AP_n e_1\|_2 = \max_{1 \leq j \leq n} \|AP_n e_j\|_2.$$

(b) Show that the diagonal elements of R_1 have decreasing magnitudes, i.e. $(R_1)_{11} \geq (R_1)_{22} \geq \dots \geq (R_1)_{rr}$.