# Chapter 2

# Systems of Linear Equations - Direct Approach

One of the most important problems in scientific computation is to solve a linear equation

$$Ax = B \tag{2.1}$$

where $A$ is an $n \times n$, square matrix and $x$ and $B$ are $n \times m$ matrices. We are especially interested in the case when $m = 1$ or $m = n$ and $B = I$. Needless to say, if any method is used to solve (2.1), then in general only an approximation $\overline{x}$ to the true solution $x$ is obtained. The accuracy of $\overline{x}$ is usually judged by measuring the norm of $\|\overline{x} - x\|$. In the previous chapter, we have seen there are many ways to define the norm of a vector. It is important to note that all norms in finite dimensional space are equivalent in the following sense:

**Theorem 2.0.1** *Given any two norms $\|\cdot\|$ and $\|\cdot\|'$ for a finite dimensional vector space $V$, there exist constants $0 < m \leq M$ such that for all $v \in V$, we have*

$$m\|v\| \leq \|v\|' \leq M\|v\|.$$

(pf): It suffices to show the theorem if one of the norm, say $\|\cdot\|'$, is the $\|\cdot\|_\infty$. For then there are constants $0 < m_1 \leq M_1$ and $0 < m_2 \leq M_2$ such that $m_1\|v\| \leq \|v\|_\infty \leq M_1\|v\|$ and $m_2\|v\|'\|v\|_\infty \leq M_2\|v\|'$. Then original theorem is proved with $m = \frac{m_1}{M_2}$ and $M = \frac{M_1}{m_2}$.

We now show that given any norm $\|\cdot\|$, there exist constants $0 < m \leq M$ such that $m\|v\| \leq \|v\|_\infty \leq M\|v\|$ for all $v \in V$.

Observe first that the theorem is true for any $m$ and $M$ if $v = 0$. Hence we only need to consider $v \neq 0$. Let $S := \{v \in V \mid \|v\|_\infty = 1)$. Obviously this set is bounded and closed. Since $\|\cdot\|$ is continuous in each of itss component (Why?), it

attains its maximum $v_M$ and minimum $v_m$ in $S$. Thus $0 < \|v_m\| \leq \|v\| \leq \|v_M\|$ for every $v \in S$. Now for every $0 \neq v \in V$, we consider $\frac{v}{\|v\|_\infty} \in S$. Thus

$$0 \leq \|v_m\| \leq \frac{\|v\|}{\|v\|_\infty} \leq \|v_M\|$$

$$\|v\|_\infty \|v_m\| \leq \|v\| \leq \|v\|_\infty \|v_M\|$$

It follows

$$\frac{1}{\|v_M\|}\|v\| \leq \|v\|_\infty \leq \frac{1}{\|v_m\|}\|v\|.$$

The theorem is proved by choosing $m = \frac{1}{\|v_M\|}$ and $M = \frac{1}{\|v_m\|}$.

**Example** For $A \in R^{m\times n}$, it can be proved that

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2,$$

$$\tfrac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty,$$

$$\tfrac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1.$$

**Remarks** (1) Based on the norm equivalent theorem, a sequence of vectors $(v_n)$ converges in one norm if an only if it converges in any other norm.

(2) It is worth noting that $\|Av\| \leq \|A\|\|v\|$ if the matrix norm $\|A\|$ is induced from the vector norm $\|v\|$. (See the definition of an induced norm.)

(3) For any square matrix $A$ and for any induced norm, we always have

$$\|A\| \geq \rho(A)$$

where $\rho(A) := \max_i |\lambda_i|$ = the spectral radius of $A$, and $\lambda_i$: = eigenvalues of $A$. (Prove this fact!)

## 2.1 Linear Systems – General Consideration

Consider the linear equation (2.1). Suppose that some $y$ is found that satisfies the equation

$$(A + E)y = B + F \tag{2.2}$$

where $E$ and $F$ are some "small" perturbations of $A$ and $B$ (One source of such perturbations is due to the use of the floating point numbers in the computer). We expect that $y$ is close to $x$ whenever $E$ and $F$ are small. We now derive some error bounds for $x - y$.

Assume both $A^{-1}$ and $(A+E)^{-1}$ exist. (How do we know that this assumption is reasonable?) It is clear that

$$x = (A + E)^{-1}(B + Ex).$$

From (2.2) it follows that

$$x - y = (A + E)^{-1}(Ex - F). \tag{2.3}$$

Taking any induced norm on both sides of (2.3) yields

$$\|x-y\| \le \|(A+E)^{-1}\|(\|E\|\|x\|+\|F\|). \tag{2.4}$$

Since $\|Ax\| = \|B\| \le \|A\|\|x\|$ implies $\|B\|/\|A\| \le \|x\|$, it follows that

$$\begin{aligned}\frac{\|x-y\|}{\|x\|} &\le \|(A+E)^{-1}\|(\|E\| + \frac{\|F\|\,\|A\|}{\|B\|}) \\ &= \|(A+E)^{-1}\|\|A\|(\frac{\|E\|}{\|A\|}+\frac{\|F\|}{\|B\|}).\end{aligned} \tag{2.5}$$

Obviously, the quantity $\|(A+E)^{-1}\|\|A\|$ needs our special attention.

**Definition 2.1.1** *The number*

$$k(A) := \|A^{-1}\|\|A\| \tag{2.6}$$

*of a square matrix $A$ is called the condition number of $A$ with respect to the norm chosen.*

**Remarks**

1. Since the eigenvlaues of $A^{-1}$ are the reciprocals of those of $A$, together with the fact $\|A\| \ge \rho(A)$, we know
$$k(A) \ge \frac{\max|\lambda_i|}{\min|\lambda_i|}.$$
2. We define $k(A) = +\infty$ when $A$ is singular.
3. The condition number $k(A)$ is used to estimate the conditioning of a matrix. From (2.1.4) it is seen that, if $E = 0$, the relative error of the solution will be large if $k(A)$ is large.
4. A perfectly conditioned matrix $A$ should have condition number $k(A) = 1$. (Why?).

**Theorem 2.1.1** *(Banach Lemma) If $D$ is an $n \times n$ matrix with $\|D\| < 1$, then $(I+D)^{-1}$ exists and satisfies*

$$\|(I+D)^{-1}\| \le \frac{1}{1-\|D\|}. \tag{2.7}$$

(pf): By the triangle inequality, we have

$$\|(I+D)x\| = \|x+Dx\| \ge \|x\| - \|Dx\| \ge (1-\|D\|)\|x\|$$

for every $x$. It follows that $\|(I+D)x\| > 0$ if $x \ne 0$. That is, $(I+D)x = 0$ has only the trivial solution $x = 0$. This shows $I+D$ is nonsingular. Furthermore,

$$\begin{aligned}1 &= \|I\| = \|(I+D)(I+D)^{-1}\| = \|(I+D)^{-1} + D(I+D)^{-1}\| \\ &\ge \|(I+D)^{-1}\| - \|D\|\|(I+D)^{-1}\| = \|(I+D)^{-1}\|(1-\|D\|) > 0.\end{aligned}$$

The assertion of the theorem is proved.

With the Banach lemma, we may estimate $\|(A+E)^{-1}\|$ as follows:

$$\begin{aligned}\|(A+E)^{-1}\| &= \|[A(I+A^{-1}E)]^{-1}\| \le \|A^{-1}\|\|(I+A^{-1}E)^{-1}\| \\ &\le \frac{\|A^{-1}\|}{1-\|A^{-1}E\|} \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|E\|} \end{aligned} \tag{2.8}$$

provided $\|A^{-1}\|\|E\| < 1$. Together with the inequality (2.5), we can now conclude an a prior error estimate for the system (2.1).

**Theorem 2.1.2** *Suppose* $\|E\| < \frac{1}{\|A^{-1}\|}$. *Then*

$$\frac{\|x-y\|}{\|x\|} \le \frac{k(A)}{1-k(A)\frac{\|E\|}{\|A\|}}\left(\frac{\|E\|}{\|A\|}+\frac{\|F\|}{\|B\|}\right). \tag{2.9}$$

**Example** The Hilbert matrix $H_n$ of order $n$ is defined by

$$H_n := \begin{bmatrix} 1, 1/2 & , \ldots, 1/n \\ 1/2, 1/3 & , \ldots, 1/(n+1) \\ \vdots & \\ 1/n, 1/(n+1), & \ldots\ 1/(2n-1) \end{bmatrix}.$$

This matrix can be inverted exactly and its eigenvalues can be computed exactly. However, the Hilbert matrix is very ill-conditioned.

| n | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| $k(H_n)$ | $5.24\times10^2$ | $1.55\times10^4$ | $4.77\times10^5$ | $1.50\times10^7$ | $4.75\times10^8$ | $1.53\times10^{10}$ |

Therefore, such matrices are often used as test problems for testing the efficiency of numerical methods for solving linear systems. (Ref: N. J. Higham, The test matrix toolbox for Matlab, University of Manchester/UMIST, NA Report No. 237, ftp vtx.ma.man.ac.uk).

**Example.** Consider the matrix $A = \begin{bmatrix} 1, 1 \\ 1, 1.0001 \end{bmatrix}$ and the equation $Ax = B$ with $B = [2, 2.0001]^T$ and $[2, 2.0002]^T$, respectively. The solutions are $[1, 1]^T$ and $[0, 2]^T$. This sensitivity of solution to small changes in data is related to the ill-conditioning of the matrix $A$.

We now consider a posterior error estimate. Let $y$ denote an approximate solution for the system (2.1). Define $r := Ay - B$ to be a residue. Then $y - x = A^{-1}r$ and, hence,

$$\frac{\|y-x\|}{\|x\|} = \frac{\|A^{-1}r\|}{\|x\|} \le \frac{\|A^{-1}\|\|r\|}{\|x\|} = \frac{k(A)\|r\|}{\|A\|\|x\|} \le \frac{k(A)\|r\|}{\|B\|}. \tag{2.10}$$

In the special case when $B = I$ so that $x = A^{-1}$, then we have

$$\frac{\|y-A^{-1}\|}{\|A^{-1}\|} \le k(A)\|r\|. \tag{2.11}$$

**Remark.** We expect that if $r$ is small, then $y$ would be close to the true solution $x$. However, it should be noted that even if we have two approximate solutions $y_1$ and $y_2$ such that $\|r_1\| < \|r_2\|$, it is possible that $y_2$ is a more reasonable solution than $y_1$. For example, consider the system

$$\begin{bmatrix} 0.780, & 0.563 \\ 0.913, & 0.659 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix}.$$

The exact solution is $[1, -1]^T$. Consider the two approximate solutions $y_1 = [0.341, -0.087]^T$ and $y_2 = [0.999, -1.001]^T$. It can be checked that $r_1 = [-0.000010, 0.000000]^T$ and $r_2 = [-0.001343, -0.001572]^T$.

**Remark.** The information of the residue can be utilized to improve the accuracy of the approximate solution $y$. This could be done as follows:???

## 2.2 Gaussian Elimination

Consider the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{12}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2. \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}$$

Gauss elimination method consists in using the first equation (assuming $a_{11} \neq 0$) to eliminate all $x_1$ terms from the second equation on, then using the new second equation to eliminate all $x_2$ terms from the third equation on, until in the end, we obtain a new system

$$\begin{aligned} \overline{a}_{11}x_1 + \overline{a}_{12}x_2 + \ldots + \overline{a}_{1n}x_n &= \overline{b}_1 \\ \overline{a}_{22}x_2 + \ldots + \overline{a}_{2n}x_n &= \overline{b}_2 \\ &\vdots \\ \overline{a}_{nn}x_n &= \overline{b}_n \end{aligned}$$

Assuming that all $\overline{a}_{ii} \neq 0$, we can solve (2.2) from the backward substitution

$$x_i = \frac{\overline{b}_i - \sum_{k=i+1}^{n} \overline{a}_{ik}x_k}{\overline{a}_{ii}}, \; i = n, n-1, \ldots, 1.$$

**Remark.** Given a general square matrix $A$, suppose we can express $A$ as $A = LU$ where $L$ is a lower triangular matrix and $U$ is an upper triangular matrix. Then the system $Ax = b$ can be written as $L(Ux) = b$. We can solve the triangular system $Ly = b$ for $y$ first, and then solve $Ux = y$ for $x$. The $LU$-decomposition of $A$ is closely related to the Gauss elimination method.

**Algorithm 2.2.1.** (Gauss Elimination Method) Given $A \in \mathbf{R}^{n\times n}$, the following algorithm computes the factorization $A = LU$ where $L$ is a lower triangular

matrix with 1 along the diagonal and $U$ is an upper triangular matrix. The elements $A_{ij}$ is overwritten by $1_{ij}$ if $i > j$, and by $u_{ij}$, otherwise. If $A$ has a singular leading principal submatrix, then the algorithm may terminate prematurely.

For $k = 1, \ldots, n$
If $a_{kk} = 0$ then quit.
Else

$$w_j := a_{kj} \quad (j = k+1, \ldots, n)$$

$$\text{For} \quad i = k+1, \ldots, n$$

$$\eta := a_{ik}/a_{kk}$$

$$a_{ik} := \eta$$

$$\text{For } j = k+1, \ldots, n$$

$$a_{ij} := a_{ij} - \eta w_j$$

**Comments.** (1) Assume $a_{11}^{(1)} \neq 0$. The row multipliers are given by $m_{i1} := a_{i1}^{(1)}/a_{11}^{(1)}$ for $i = 2, \ldots, n$. Thus from the second row on, we generate new elements $a_{ij}^{(2)} := a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)}$ for $i, j = 2, \ldots, n$. In general, after $k-1$ steps, we should have constructed the matrix

$$A^{(k)} := \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & & & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & & & & a_{2n}^{2} \\ & \ddots & \ddots & & & \vdots \\ 0 & & 0 & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ & & & & \vdots & \\ 0 & & 0 & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{bmatrix}$$

Assuming $a_{kk}^{(k)} \neq 0$, we define the multipliers $m_{ik} := a_{ik}^{(k)}/a_{kk}^{(k)}$ for $i = k+1, \ldots, n$. Then we generate $a_{ij}^{(k+1)} := a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}$ for all $i, j = k+1, \ldots, n$. In do so, the earlier $k$ rows are left unchanged, and zeros are introduced into column $k$ below the diagonal element.

(2) Observe that the operation of multiplying the $i$-th row by a number $p$ and then adding the result to the $j$-th row can be accomplished by premultiplying the matrix by the elementary matrix $E_{ji}(p)$ where

$$E_{ji}(p) := \begin{matrix} \\ \\ \\ \\ \text{j-th row} \longrightarrow \\ \\ \\ \end{matrix} \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & \ddots & & & \\ & & & p & \ddots & \\ & & & & & \\ 0 & & & & & 1 \end{bmatrix}$$

$$\uparrow$$

$$\text{i-th column}$$

Thus the transformation from $A^{(1)}$ to $A^{(2)}$ can be summarized as

$$E_{n1}(-m_{n1})\dots E_{21}(-m_{21})A^{(1)} = A^{(2)}.$$

It can be checked easily that

$$E_1 := E_{n1}(-m_{n1})\dots E_{21}(-m_{21}) = \begin{bmatrix} 1 & & & \\ -m_{21} & 1 & & \\ \vdots & & & \\ \vdots & & & \\ -m_{n1} & & & 1 \end{bmatrix}.$$

In general, from $A^{(k)}$ to $A^{(k+1)}$, assuming $A^{(k)}_{kk} \neq 0$, we construct so called Frobenius matrix

$$E_k := \left[\begin{array}{c|ccc} I_{k-1} & & & \\ \hline & 1 & & \\ & -m_{k+1,k} & 1 & \\ & \vdots & & \ddots \\ & -m_{n,k} & & 1 \end{array}\right]$$

where $m_{ik} := a^{(k)}_{ik}/a^{(k)}_{kk}$ for $i = k+1, \dots, n$. Then

$$E_{n-1}E_{n-2}\dots E_2E_1A^{(1)} = A^{(n)} := U \tag{2.12}$$

is an upper triangular matrix. Note each elementary matrix is nonsingular. In fact, $(E_{ji}(p))^{-1} = E_{ji}(-p)$. Therefore

$$A = A^{(1)} = E_1^{-1}\dots E_{n-1}^{-1}U. \tag{2.13}$$

We note that the matrix $L := E_1^{-1}\dots E_{n-1}^{-1}$ is a lower triangular matrix. In fact, it can be proved that

$$L = \begin{bmatrix} 1 & & & \\ m_{21} & 1 & & \\ & m_{22} & & \\ \vdots & \vdots & & \\ m_{n1} & m_{n2} & & 1 \end{bmatrix}$$

**Remark.** The $LU$-decomposition as described as above can be carried out (without pivoting) if and only if $a^{(k)}_{kk} \neq 0$ for $k = 1, \dots, n-1$.

## 2.3 Mathematical Pivoting

The system $\begin{bmatrix} 0, 1 \\ 2, 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ obviously has solutions for every $[b_1, b_2]^T$. But Algorithm 2.2.1 fails to work because $a_{11} = 0$. This difficulty can easily be remedied by simply interchanging the two equations. Such a process of interchanging rows or columns to bring a nonzero element to the pivoting position is called mathematical pivoting.

Theoretically Algorithm 2.2.1 should work for the system $\begin{bmatrix} \epsilon, 1 \\ 2, 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ where $0 \neq \epsilon << 1$. But in practice we still prefer to do pivoting so as to avoid division by a very small number $\epsilon$.

There are two types of pivoting strategies used in practice:

**Definition 2.3.1** *(1) For $1 \leq k \leq n-1$, let*

$$\max_{k \leq i \leq n} |a_{ik}^{(k)}| = |a_{rk}^{(k)}|.$$

*We then swap the $k$-th row and the $r$-th row of the matrix $A^{(k)}$. In case there are more than one such indices, we choose $r$ to be the smallest index. Such a pivoting procedure is called Partial Pivoting.*

*(2) For $1 \leq k \leq n-1$, let*

$$\max_{k \leq i \leq n_{k \leq j \leq n}} |a_{ij}^{(k)}| = |a_{rs}^{(k)}|.$$

*We then swap the $k$-th row and the $r$-th row, and the $k$-th column and the $s$-th column of the matrix $A^{(k)}$. Such a pivoting procedure is called Complete Pivoting.*

**Example.** $\begin{bmatrix} \epsilon_1, & 2\epsilon_1 + \epsilon_2, & 3 \\ 1, & 2, & 0 \\ 0, & 4, & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ (Original system)

$$\begin{bmatrix} 1, & 2, & 0 \\ \epsilon_1, & 2\epsilon_1 + \epsilon_2, & 3 \\ 0, & 4, & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \text{ (Partial Poviting)}$$

$$\begin{bmatrix} 5, & 4, & 0 \\ 0, & 2, & 1 \\ 3, & 2\epsilon_1 + \epsilon_2, & \epsilon_1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \text{ (Complete Poviting)}$$

**Remark.** When compared to what might happen if no pivoting is used, Gaussian elimination with complete pivoting has been shown to have slower propagation of roundoff errors. The theoretical analysis of error propagation for partial pivoting is not as good as that for complete pivoting. But in almost all practical problems, the error behavior is essentially the same. Obviously, complete

pivoting is more expensive. Thus, partial pivoting is used in most practical algorithms.

The interchanging of the $k$-th and the $r$-th rows of a matrix can be achieved from premultiplying the matrix by the permutation matrix $P_{rk}$ where

$$P_{rk} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & 1 & & & \longleftarrow & k-th\ row \\ & & & 1 & 0 & & & \longleftarrow & r-th\ row \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \end{bmatrix}$$

Thus the Gaussian elimination process with partial pivoting can be represented as

$$E_{n-1}P_{n-1}E_{n-2}\dots P_3E_2P_2E_1P_1A = U \tag{2.14}$$

where each $P_k$ represents a permutation matrix $P_{rk}$ for some $r \geq k$ and each $E_k$ represents a Frobenius matrix. Consider the matrix

$$\begin{aligned} N &:= (E_{n-1}P_{n-1}E_{n-2}\cdots P_3E_2P_2E_1P_1)^{-1} \\ &= P_1E_1^{-1}\cdots P_{n-1}E_{n-1}^{-1} \end{aligned} \tag{2.15}$$

In general, $A = NU$ and $N$ is not expected to be lower triangular. Let

$$P := P_{n-1}\cdots P_1. \tag{2.16}$$

we claim

$$PN = L. \tag{2.17}$$

That is, if we do all the necessary row permutations $P_1, \cdots, P_{n-1}$ at the beginning to get the matrix $PA$ from $A$, then the matrix $PA$ has $LU$-decomposition.

(pf): Since $E_{n-1}P_{n-1}\cdots E_1P_1A = U$, we have $P_{n-1}\dots E_1P_1A = E_{n-1}^{-1}U$. We claim $P_{n-1}E_{n-2} = \hat{E}_{n-2}P_{n-1}$ for some new lower triangular matrix $\hat{E}_{n-2}$. More generally, we claim $P_{k+p}E_k = \hat{E}_kP_{k+p}$ where $\hat{E}_k$ is some new lower triangular matrix for $k = 1, \cdots, n-2, p \geq 1$ and $k+p \leq n-1$. If these claims are true, then we have $E_{n-1}^{-1}U = P_{n-1}E_{n-2}\cdots E_1P_1A = \hat{E}_{n-2}P_{n-1}P_{n-2}\cdots E_1P_1A$. Hence $\hat{E}_{n-2}^{-1}E_{n-1}^{-1}U = P_{n-1}P_{n-2}E_{n-3}\cdots E_1P_1A$. Continuing this process, we eventually get $LU = P_{n-1}\cdots P_1A$ where $L$ is the product of lower triangular matrices.

We now proceed to prove the claim that $P_{k+p}E_k = \hat{E}_kP_{k+p}$. This is done

from the following observations:

$$P_{k+p}E_k = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & 1 & & & \\ & & & 1 & 0 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \end{bmatrix} \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -m_{k+p,k} & 1 & & 0 & \\ & & -m_{r,k} & 0 & & 1 & \\ & & & & & & \ddots \\ & & -m_{n,k} & & & & \end{bmatrix}$$

($k-th$ column, $(k+p)-th$ column, $r-th$ column)

$$= \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -m_{r,k} & 0 & & 1 & \\ & & -m_{k+p,k} & 1 & & 0 & \\ & & & & & & \ddots \\ & & -m_{n,k} & & & & 1 \end{bmatrix}.$$

($k_th$ column, $(k+p)-th$ column, $r-th$ column)

So

$$P_{k+p}E_kP_{k+p} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -m_{r,k} & 1 & & 0 & \\ & & -m_{k+p,k} & 0 & & 1 & \\ & & & & & & \ddots \\ & & -m_{n,k} & & & & 1 \end{bmatrix} = \hat{E}_k.$$

($k_th$ column, $(k+p)-th$ column, $r-th$ column)

**Theorem 2.3.1** *If $A$ is nonsingular, then the Gauss elimination method with partial pivoting to reduce $A$ to upper triangular form can always be carried out.*

## 2.4 Error Analysis for Gaussian Elimination

Suppose the system

$$Ax = b$$

is be solved on the computer by the Gaussian elimination method. We usually will only obtain an approximate solution $y$. We shall regard that $y$ satisfies a perturbed system

$$(A + E)y = b + f$$

for some perturbations $E$ and $f$. The sources of errors are

(1) The roundoff errors in representing $A$ and $b$, say

$$\begin{aligned} A_r &= A + \delta A, \\ b_r &= b + \delta b. \end{aligned} \tag{2.18}$$

(2) The errors occurred in the decomposition of $A_r$, say

$$L_r U_r = P A_r + \delta A_r. \tag{2.19}$$

(3) The floating-point arithmetic errors in solving the triangular systems.

Since we constantly need to deal with the inner product of two vectors, we first analyze the error in this important operation. Suppose we need to calculate

$$< a, b >= \Sigma a_i b_i \tag{2.20}$$

by summing the products $a_i b_i$ and rounding is done after each multiplication and after each addition. Then

$$\begin{aligned} s_1 &= fl(a_1 b_1) = (a_1 b_1)(1 + \delta_1), \\ s_i &= fl(s_{i-1} + fl(a_i b_i)) \\ &= (s_{i-1} + (a_i b_i)(1 + \delta_i))(1 + \eta_i). \end{aligned}$$

Thus

$$\begin{aligned} s_n &= fl(< a, b >) \\ &= a_n b_n (1 + \delta_n)(1 + \eta_n) + a_{n-1} b_{n-1} (1 + \delta_{n-1})(1 + \eta_{n-1})(1 + \eta_n) \\ &\quad + \cdots + a_1 b_1 (1 + \delta_1)(1 + \eta_1) \cdots (1 + \eta_n) \text{ with } \eta_1 = 0. \\ &= \textstyle\sum_{i=1}^{n} a_i b_i (1 + \epsilon_i) \text{ with } 1 + \epsilon_i := (1 + \delta_i) \prod_{j=i}^{n} (1 + \eta_j). \end{aligned}$$

**Lemma 2.4.1.** Assume $|\delta_i| \le u$ for $1 \le i \le n$, $|\eta_i| \le u$ for $2 \le i \le n$, and $\eta_1 = 0$. (Recall $u = \frac{1}{2}\beta^{1-t}$.) Assume $u < \frac{1}{2n}$. Then

$$|\epsilon_1| \le 2nu, |\epsilon_i| \le 2(n - i + 2)u, \quad \text{for } 2 \le i \le n. \tag{2.21}$$

(pf): Since $u < 1$, for all factors in $(1+\delta_i)\prod_{j=1}^{n}(1+\eta_i)$ are positive. Thus

$$\begin{aligned}(1-u)^n \le 1 \quad &+\epsilon_1 \le (1+u)^n,\\ (1-u)^{n-i+2} \quad &\le 1+\epsilon_i \le (1+u)^{n-i+2}, \text{ for } 2 \le i \le n.\end{aligned}$$

It follows that

$$\begin{aligned}|\epsilon_1| \quad &\le (1+u)^n - 1\\ |\epsilon_i| \quad &\le (1+u)^{n-i+2} - 1, \text{ for } 2 \le i \le n.\end{aligned}$$

Consider, for $p \le n$,

$$\begin{aligned}(1+u)^p - 1 &= pu + \frac{p(p-1}{2}u^2 + \cdots + u^p\\ &= pu(1 + \frac{p-1}{2}u + \cdots + \frac{1}{p}u^{p-1})\\ &\le pu(1 + \frac{1}{2} + (\frac{1}{2})^2 + \cdots + (\frac{1}{2})^p) \le 2pu.\end{aligned}$$

**Remark.** According to the above, we may write

$$fl(\Sigma a_i b_i) = \Sigma \tilde{a}_i b_i = \Sigma a_i \tilde{b}_i$$

where $\tilde{a}_i = a_i(1+\epsilon_i), \tilde{b}_i = b_i(1+\epsilon_i)$.

We now consider the errors involved the solving a triangular system

$$Tx = r.$$

Suppose $T$ is a lower triangular matrix. In exact arithmetic, we should have

$$\begin{aligned}x_1 &= \frac{r_1}{t_{11}},\\ x_2 &= \frac{r_2 - t_{21}x_1}{t_{22}},\\ &\vdots\\ x_i &= \frac{r_i - \Sigma_{j=1}^{i-1} t_{ij}x_j}{t_{ii}}, \quad i = 1, \ldots, n.\end{aligned}$$

On a computer, we produce an approximate solution $z$ where

$$\begin{aligned}z_1 &= fl(\frac{r_1}{t_{11}}) = \frac{r_1}{t_{11}}\frac{1}{1+\delta_1},\\ z_i &= fl(\frac{fl(r_i - fl(\sum_{j=1}^{i-1} t_{ij}z_j)}{t_{ii}}) = \frac{fl(r_i - fl(\Sigma_{j=1}^{i-1} t_{ij}z_j)}{t_{ii}}\frac{1}{1+\delta_i}\end{aligned}$$

where $|\delta_i| \leq u = \frac{1}{2}\beta^{1-t}$. But

$$fl(r_i - fl(\Sigma_{j=1}^{i-1} t_{ij} z_j) = (r_i - fl(\Sigma_{j=1}^{i-1} t_{ij} z_j) \frac{1}{1+\eta_i}$$
$$= (r_i - \Sigma_{j=1}^{i-1}(t_{ij} + \delta t_{ij}) z_j) \frac{1}{1+\eta_i}$$

where

$$|\delta t_{ij}| \leq \begin{cases} |t_{ij}| 2(i-j+1)u, & \text{for } 2 \leq j \leq i-1, \\ |t_{i1}| 2(i-1)u, & \text{for } j = 1. \end{cases}$$

Therefore, we find

$$t_{11}(1+\delta_1) z_1 = r_1,$$
$$t_{ii}(1+\delta_i)(1+\eta_i) z_i + \textstyle\sum_{j=1}^{i=1}(t_{ij} + \delta t_{ij}) z_j = r_i.$$

In summary, we find that

**Lemma 2.4.2.** The floating-point solution to the triangular system (2.4.9) satisfies the system

$$(T + \delta T) z = r$$

where

$$(\delta T)_{ij} = \begin{cases} \delta T_{ij}, & \text{if } i \neq j; \\ t_{ii}((1+\delta_i)(1+\eta_i) - 1), & \text{if } i = j. \end{cases}$$

**Remark.** Note that $|(\delta T)_{ij}| \leq 2nu|t_{ij}|$ for all $i$ and $j$. Thus the computed solution can be interpreted as the exact solution of a slightly changed problem, showing that the process of solving a triangular system is stable.

We now apply Lemma 2.4.2 to study the third source of errors in solving the system (2.4). We find that the approximate solution $y$ is given from

$$(L_r + \delta L_r) z = P b_r \tag{2.22}$$
$$(U_r + \delta U_r) y = z \tag{2.23}$$

for some $\delta L_r$ and $\delta U_r$. Putting all things together, we obtain

$$\begin{aligned} (L_r + \delta L_r)(U_r + \delta U_r) y &= P b_r, \\ (PA + P\delta A + \delta A_r + \delta L_r U_r + L_r \delta U_r + \delta L_r \delta U_+ r) y &= Pb + P\delta b, \\ (PA + E) y &= Pb + f \end{aligned}$$

where

$$E: = P\delta A + \delta A_r + \delta L_r U_r + L_r \delta U_r + \delta L_r \delta U_r \tag{2.24}$$
$$f: = P\delta b. \tag{2.25}$$

We now analyze the errors in the $LU$-decomposition. Note that permutations of rows do not introduce errors. We thus assume that rows of $A_r$ are already

arranged so that $A$ has a $LU$-decomposition. At step $k$, we are working on the matrix

$$A^{(k)} := \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & & & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & & & & a_{2n}^{2} \\ & \ddots & \ddots & & & \vdots \\ 0 & & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & & & & \vdots & \\ 0 & & 0 & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}$$

Recall that the multipliers are $m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}},\ k+1 \le i \le n$, and that the exact arithmetic should be

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & ,1 \le i \le k, i \le j \le n; \\ a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}, & k+1 \le i \le n, k+1 \le j \le n; \\ 0 & \text{otherwise.} \end{cases}$$

Now due to the floating-point arithmetic, we obtain

$$\begin{aligned} m_{ik} &= fl(\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}) = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}(1+\eta_{ik}) \\ a_{ij}^{(k+1)} &= fl(a_{ij}^{(k)} - fl(m_{ik}a_{kj}^{(k)})) \\ &= (a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)})(1+\delta_{ij}))\frac{1}{1+\xi_{ij}},\ \text{for } k+1 \le i,\ j \le n \end{aligned}$$

where $|\eta_{ij}|,\ |\delta_{ij}|,\ |\xi_{ij}| \le u$. So

$$A_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} - m_{ik}a_{kj}^{(k)}\delta_{ij} - a_{ij}^{(k+1)}\xi_{ij}.$$

Let $E_k$ represents the Frobenius matrix formed from the floating-point numbers of $m_{ik}$. Then we have

$$A^{(k+1)} = E_k A^{(k)} + \Omega^{(k)}$$

wher $\Omega^{(k)} = (\omega_{ij}^{(k)})$ is given by

$$\omega_{ij}^{(k)} = \begin{cases} -m_{ik}a_{kj}^{(k)}\delta_{ij} - a_{ij}^{(k+1)}\xi_{ij} & ,\ \text{for } k+1 \le i,j \le n, \\ a_{ik}^{(k)}\eta_{ik} & ,\ \text{for } k+1 \le i \le n, \\ 0 & ,\ \text{otherwise} \end{cases}$$

Thus

$$\begin{aligned} U_r &= A^{(n)} = E_{n-1}A^{(n-1)} + \Omega^{(n-1)} \\ &= E_{n-1}(E_{n-2}A^{(n-2)} + \Omega^{(n-2)}) + \Omega^{(n-1)} \\ &= E_{n-1}\dots E_1 A^{(1)} + E_{n-1}\dots E_2\Omega^{(1)} + \dots + \Omega^{(n-1)}. \end{aligned}$$

Note that in general, $E_i^{-1}\Omega^{(j)} = \Omega^{(j)}$, if $i \leq j$. So we obtain

$$E_1^{-1} \ldots E_{n-1}^{-1} U_r = A^{(1)} + \Omega^{(1)} + \ldots + \Omega^{(n-1)}$$

That is,

$$L_r U_r = A_r + \delta A_r$$

where

$$\delta A_r = \Omega^{(1)} + \ldots + \Omega^{(n-1)}$$

Because of pivoting, it is reasonable to assume $|m_{ij}| \leq 1$. Let it be assumed that

$$\max_{i,j,k} |a_{ij}^{(k)}| = \rho(n)\|A_r\|_\infty$$

where $\rho(n) :=$ the growth factor (which will be derived later). Then from (2.4), we see

$$|\omega_{ij}^{(k)}| \leq \begin{array}{ll} 2\rho(n)\|A_r\|_\infty u, & \text{for } k+1 \leq i,j \leq n; \\ \rho(n)\|A_r\|_\infty u, & \text{for } k+1 \leq i \leq n; \\ 0 & \text{, otherwise .} \end{array}$$

So $\|(\delta A_r)_{ij}\| \leq \Sigma_{k=1}^{n-1} |\omega_{ij}^{(k)}|$. In terms of matrix inequality, we denote this as

$$|\delta A_r| \leq \rho(n)\|A_r\|_\infty u \begin{bmatrix} 0 & 0 & & & \\ 1 & 2 & & \ldots & 2 \\ 1 & 3 & 4 & \ldots & 4 \\ 1 & 3 & 5 & & \\ \vdots & \vdots & \vdots & & \\ 1 & 3 & 5 & \ldots & 2n-2 \end{bmatrix}$$

or simply

$$\|\delta A_r\|_\infty \leq (\sum_{j=1}^{n} (2j-1)-1)-1)\rho(n)\|A_r\| u \leq n^2 \rho(n) u \|A_r\|_\infty.$$

Recall we have established that the computed solution $y$ for the system (2.4) actually satisfies the perturbed system (2.4). Now we have all the estimates:

$$\begin{array}{rcl} \|f\|_\infty & = & \|P\delta b\|_\infty = \|\delta b\|_\infty \leq u\|b\|_\infty, \\ \|P\delta A\|_\infty & = & \|\delta A\|_\infty \leq u\|A_r\|_\infty, \\ \|\delta A_r\|_\infty & \leq & n^2\rho(n)u\|A_r\|_\infty, \\ \|L_r\|_\infty & \leq & n, (\text{Since } |m_{ij}| \leq 1), \\ \|U_r\|_\infty & = & \|A^{(n-1)}\|_\infty \leq n\rho(n)\|A_r\|_\infty, \ (\text{By definition of } \rho(n)), \\ \|\delta L_r\|_\infty & \leq & 2nu\|L_r\|_\infty \leq 2n^2 u, \ (\text{Since } |(\delta T)_{ij}| \leq 2nu|t_{ij}|), \\ \|\delta U_r\|_\infty & \leq & 2nu\|U_r\|_\infty \leq 2n^2\rho(n)u\|A_r\|_\infty. \end{array}$$

So the size of the perturbed matrix $E$ is bounded by

$$\begin{aligned} \|E\| &\leq \|A_r\|_\infty(1 + n^2\rho + 2n^3\rho + 2n^3\rho + 4n^4u\rho) \\ &\leq \rho u\|A_r\|_\infty(5n^2 + 4n^3). \text{ (Since } \rho >> 1, \text{ and } n^2u < 1). \quad (2.26) \end{aligned}$$

Now we give an estimate of the growth factor $\rho(n)$ used in (2.4). Recall that $a_{ij}^{(k+1)} = fl(a_{ij}^{(k)} - fl(m_{ik}a_{kj}^{(k)}))$. We claim $\max|a_{ij}^{(k+1)}| \leq 2\max|a_{ij}^{(k)}|$. This follows from

$$\begin{aligned} & |a_{ij}^{(k)} - fl(m_{ik}a_{kj}^{(k)})| \leq |a_{ij}^{(k)}| + |fl(m_{ik}a_{kj}^{(k)}| \\ \leq\ & |a_{ij}^{(k)}| + |a_{kj}^{(k)}| \text{ (Since } |m_{ik}a_{kj}^{(k)}| \leq |a_{kj}^{(k)}|) \\ \leq\ & 2\max|a_{ij}^{(k)}|. \end{aligned}$$

Thus we conclude $\max|a_{ij}^{(k)}| \leq 2^{k-1}\max|a_{ij}^{(1)}|$ and that

$$\max_k \max_{i,j} |a_{ij}^{(k)}| \leq 2^{n-1}\|A_r\|_\infty.$$

We call $\rho(n) = 2^{n-1}$ the growth factor in the Gauss elimination. Note that this number is overestimated.

**Remarks.** The growth factor $\rho(n)$ can be greatly reduced for the special cases that

(1) If $A_r$ is symmetric and positive definite, then $\rho(n) = 1$.

(2) If $A_r$ is an upper Hessenberg matrix, i.e., $a_{ij}^{(1)} = 0$ for $j > i + 1$, then $\rho(n) = n$.

## 2.5 QR Decomposition

The technique to be discussed in this section has extremely important impact in many areas of numerical linear algebra. We begin with the major result:

**Theorem 2.5.1** *Suppose* $A \in \mathbf{R^{m\times n}}$, $\mathbf{m \geq n}$ *and suppose rank* $(A) = n$ *(i.e., suppose* $A$ *has linearly independent columns). Then* $A$ *can be decomposed as*

$$A = QR$$

*where* $Q = \mathbf{R^{m\times m}}$ *is an orthogonal matrix (i.e.,* $Q^TQ = I_m$*), and* $R \in \mathbf{R^{m\times m}}$ *is of the form* $R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$ *where* $R_1 \in \mathbf{R^{n\times n}}$ *is an upper triangular matrix.*

**Remark.** If we rewrite $Q = [Q_1, Q_2]$ with $Q_1 \in \mathbf{R^{m\times n}}, \mathbf{Q_2} \in \mathbf{R^{n\times(m-n)}}$. Then $A = QR = Q_1R_1$. Note also that the columns of $Q_1$ are still mutually orthonormal, i.e., $Q_1^TQ_1 = I_n$.

The proof of Theorem 2.5.1 can be seen directly from the way we construct $Q$ and $R$. There are two approaches for the construction:

**1. $QR$ decomposition via Gram-Schmidt process:**

A set of linearly independent column vectors $\{v_1, \ldots, v_n\} \subset \mathbf{R^m}(\mathbf{m} \geq \mathbf{n})$ spans an $n$-dimensional vector subspace of $\mathbf{R^m}$. We want to construct an orthonormal set of vectors $\{q_1, \ldots, q_n\}$ that spans the same subspace. The well-known Gram-Schmidt process provides a recipe for accomplishing this:

(a) Set $w_1 := v_1$.

(b) Set, in general, for $j = 2, \ldots, n$

$$w_j := v_j - \sum_{i=1}^{j-1} \frac{< w_i, v_j >}{< w_i, w_i >} w_i.$$

It is really to be verified that $< w_i, w_j >= 0$ for $i \neq j$, and that $w_j \neq 0$. If we define $q_j := \frac{w_j}{\|w_j\|_2}$, then $< q_i, q_j >= \delta_{ij}$. The above relationship may be rewritten as

$$\begin{aligned} v_1 &= w_1; \\ v_j &= w_j + \sum_{i=1}^{j-1} \frac{< w_i, v_j >}{< w_i, w_i >} w_i. \end{aligned} \tag{2.27}$$

In matrix form, (2.27) may be recorded as

$$\begin{aligned} & [v_1, \ldots, v_n] \\ = \; & [w_1, \ldots, w_j, \ldots, w_n] \begin{bmatrix} 1 & \frac{<w_1,v_2>}{<w_i,w_1>} & \frac{<w_1,v_3>}{<w_1,w_1>} & & \\ 0 & 1 & \frac{<w_2,v_3.}{<w_2,w_2>} & & \\ 0 & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}. \end{aligned}$$

That is, $V = W\tilde{R}$. Let $D :=$ diag $\{\frac{1}{\|w_1\|}, \ldots, \frac{1}{\|w_n\|}\}$. Then $V = WDD^{-1}\tilde{R} = Q_1 R_1$ where $Q_1 := WD$ and $R_1 := D^{-1}\tilde{R}$.

Obviously, for our consideration, the matrix $V$ is identified as our original $A$.

**2. $QR$ decomposition via Householder transformation:**

Let $u \in \mathbf{R^m}$ be a normalized column vector. The associated Householder matrix is defined to be

$$V := I - 2uu^T.$$

It can be shown that $V$ is an orthogonal matrix (Show this!). Consider the so called Householder transformation

$$Vx = x - 2uu^T x.$$

Note that $Vx$ is the reflection of $x$ with respect to the hyperplane which is normal to the vector $u$;

This idea, therefore, may be applied in the following way: Given $x \in R^m$, find an Householder matrix $V$ such that $Vx = \pm e\sigma$ where $\sigma = \|x\|_2$ (Why?).

There are two choices of $u$:

$$\begin{aligned} u_1 : &= \frac{\frac{x}{\sigma} + e_1}{\|\frac{x}{\sigma} + e_1\|_2}, \\ u_2 : &= \frac{\frac{x}{\sigma} - e_1}{\|\frac{x}{\sigma} - e_1\|_2}. \end{aligned} \tag{2.28}$$

Suppose the second choice is used, then $Vx = [\sigma, 0, \ldots, 0]^T$. Applied to our matrix $A \in R^{m\times n}$ with $m \geq n$, then we may construct an orthogonal matrix $V_1 \in R^{m\times n}$ such that

$$A^{(2)} := V_1 A = \begin{bmatrix} x, & x, & \ldots & , x \\ 0, & x, & & \\ & & B^{(2)} & \\ 0, & x, & & , x \end{bmatrix}.$$

Now consider the $(m-1)\times(n-1)$ lower right submatrix $B^{(2)}$ of $A^{(2)}$. We then can construct an orthogonal matrix $V_2 \in R^{(m-1)\times(m-1)}$ such that

$$V_2' B^{(2)} = \begin{bmatrix} x, & x, & \ldots & , x \\ 0, & x, & & \\ & & & \\ 0, & x, & & , x \end{bmatrix}.$$

Now let $V_2 := \begin{bmatrix} 1, 0 \\ 0, V_2 \end{bmatrix} \in R^{m\times n}$. Then

$$V_2 A^{(2)} = \begin{bmatrix} x, & x, & \dots & , & x \\ 0, & x, & & & \\ & 0, & x, & & ,x \\ & & & B^{(3)} & \\ 0, & 0, & x, & & ,x \end{bmatrix}.$$

Continuing this procedure $n-1$ times, we obtain

$$A^{(n)} := V_{n-1}A^{(n-1)} - V_{n-1}\dots V_1 A = \begin{bmatrix} x, & x, & \dots & ,x \\ 0, & x, & & \\ & 0, & x, & ,x \\ & & & \\ & & & ,x \\ 0, & & & ,0 \\ & & & \\ 0, & \dots & & ,0 \end{bmatrix} = R.$$

Let

$$Q := V^{-1} := (V_{n-1}\dots V_1)^{-1}.$$

Then $Q$ is also an orthogonal matrix. We have proved that $A = QR$ as asserted in Theorem 2.5.1.

**Remark.** It can be proved that the QR decomposition via the Gram-Schmidt process is not numerically stable if the columns of matrix $A$ are nearly linearly dependent. The Householder transformation, in contrast, can avoid the instability by carefully monitoring the choice of $u$ (That is, use $u_1$ if $x_1 > 0$, use $u_2$ if $x_1 < 0$).

**Algorithm 2.5.1.** ($QR$ decomposition via Householder transformation) Given $A \in \mathbf{R^{m\times n}}$ with $n \geq n$, the following algorithm reduces $A$ to upper triangular form by using Householder transformations. The strictly upper triangular portion of $A$ and an non-dimensional and $d$ are overwritten by the resulting $R$. The lower triangular portion of $A$ contains the vectors $u$ used in constructing the Householder matrices.

For $j = 1, \dots, n$
  $\sigma := 0$
    For $i = j, \dots, n$
    $\sigma := \sigma + a_{ij}^2$
  If $\sigma = 0$
    go to singular
  Else
    If $a_{jj} < 0$,
      $s := dj := sqrt(\sigma)$ (This choice is to avoid cancellation!)
    Else

$s := d_j := -sqrt(\sigma)$ (The vector $d$ stores the diagonals!)
$\beta := 1/(sa_{jj} - \sigma)$ (Note that $sa_{jj}$ is always negative!)
$a_{jj} := a_{jj} - s$
For $k = j + 1, \ldots, n$ ( Do not change the lower triangular part!)
  $t := 0$
  For $i = j, \ldots, m$
    $t := t + a_{ij}a_{ik}$ ( $t$ is actually the scalar $u^T x$.)
  $t := \beta t$
  For $i = j, \ldots, m$
    $a_{ik}; = a_{ik} + a_{ij}t$

**Remark.** Analogous to the $LU$ decomposition, the $QR$ decomposition may be used to solve the system $Ax = b$. Indeed, if $A = QR$, then $Rx = Q^T b$. The advantage is that the columns of $Q$ has unit length while, in contrast, the column of $L$ in the $LU$ decomposition may have larger length (due to the a small pivot element). The $QR$ decomposition usually is considered to a more stable method, although the cost is about 4 times higher. (Justify this!)