CS724: Topics in Algorithms
Solving Linear Systems
Slide Set 6

Prof. Dan A. Simovici

Consider the following set of linear equalities

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + \ldots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{m1}x_1 + \ldots + a_{mn}x_n &= b_m, \end{aligned}$$

where a_{ij} and b_i belong to a field F. This set constitutes a *system of linear equations*. Solving this system means finding $x_1, \ldots, x_n$ that satisfy all equalities.

The system can be written succinctly in a matrix form as $A\boldsymbol{x} = \boldsymbol{b}$, where

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \cdots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix},$$

and

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

If the set of solutions of a system $A\boldsymbol{x} = \boldsymbol{b}$ is not empty we say that the system is *consistent*. Note that $A\boldsymbol{x} = \boldsymbol{b}$ is consistent if and only if $\boldsymbol{b} \in \text{range}(A)$.

Let $A\boldsymbol{x} = \boldsymbol{b}$ be a linear system in matrix form, where $A \in \mathbb{C}^{m\times n}$. The matrix $[A\ \ \boldsymbol{b}] \in \mathbb{C}^{m\times(n+1)}$ is the *augmented matrix* of the system $A\boldsymbol{x} = \boldsymbol{b}$.

Theorem

Let $A \in \mathbb{C}^{m \times n}$ be a matrix and let $\boldsymbol{b} \in \mathbb{C}^{n \times 1}$. The linear system $A\boldsymbol{x} = \boldsymbol{b}$ is consistent if and only if $rank(A\ \boldsymbol{b}) = rank(A)$.

Proof.

If $A\boldsymbol{x} = \boldsymbol{b}$ is consistent and $\boldsymbol{x}' = (x_1, \ldots, x_n)$ is a solution of this system, then $\boldsymbol{b} = x_1\boldsymbol{c}_1 + \cdots + x_n\boldsymbol{c}_n$, where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ are the columns of A. This implies $\text{rank}([A\ \ \boldsymbol{b}]) = \text{rank}(A)$.
Conversely, if $\text{rank}(A\ \boldsymbol{b}) = \text{rank}(A)$, the vector $\boldsymbol{b}$ is a linear combination of the columns of A, which means that $A\boldsymbol{x} = \boldsymbol{b}$ is a consistent system. □

Definition

An *homogeneous linear system* is a linear system of the form $A\mathbf{x} = \mathbf{0}_m$, where $A \in \mathbb{C}^{m \times n}$, $\mathbf{x} \in \mathbb{C}^{n,1}$ and $\mathbf{0} \in \mathbb{C}^{m \times 1}$.

Clearly, any homogeneous system $A\mathbf{x} = \mathbf{0}_m$ has the solution $\mathbf{x} = \mathbf{0}_n$. This solution is referred to as the *trivial solution*. The set of solutions of such a system is null(A), the null space of the matrix A.

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two solutions of the system $A\boldsymbol{x} = \boldsymbol{b}$. Then $A(\boldsymbol{u} - \boldsymbol{v}) = \boldsymbol{0}_m$, so $\boldsymbol{z} = \boldsymbol{u} - \boldsymbol{v}$ is a solution of the homogeneous system $A\boldsymbol{x} = \boldsymbol{0}_m$, or $\boldsymbol{z} \in \text{null}(A)$. Thus, the set of solutions of $A\boldsymbol{x} = \boldsymbol{b}$ can be obtained as a "translation" of the null space of A by any particular solution of $A\boldsymbol{x} = \boldsymbol{b}$. In other words the set of solution of $A\boldsymbol{x} = \boldsymbol{b}$ is $\{\boldsymbol{x} + \boldsymbol{z} \mid \boldsymbol{z} \in \text{null}(A)\}$. Thus, for $A \in \mathbb{C}^{m \times n}$, the system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution if and only if $\text{null}(A) = \{\boldsymbol{0}_n\}$, that is, if $\text{rank}(A) = n$.

Theorem

Let $A \in \mathbb{C}^{n \times n}$. Then, A is invertible (which is to say that rank$(A) = n$) if and only if the system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for every $\boldsymbol{b} \in \mathbb{C}^n$.

Proof.

If A is invertible, then $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, so the system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution.
Conversely, if the system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for every $\boldsymbol{b} \in \mathbb{C}^n$, let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ be the solution of the systems $A\boldsymbol{x} = \boldsymbol{e}_1, \ldots, A\boldsymbol{x} = \boldsymbol{e}_n$, respectively. Then, we have

$$A(\boldsymbol{c}_1 | \cdots | \boldsymbol{c}_n) = I_n,$$

which shows that A is invertible and $A^{-1} = (\boldsymbol{c}_1 | \cdots | \boldsymbol{c}_n)$. □

Corollary

An homogeneous linear system $A\mathbf{x} = \mathbf{0}$, where $A \in \mathbb{C}^{n\times n}$ has a non-trivial solution if and only if A is a singular matrix.

Definition

A matrix $A \in \mathbb{C}^{n \times n}$ is *diagonally dominant* if
$|a_{ii}| > \sum\{|a_{ik}| \mid 1 \leqslant k \leqslant n \text{ and } k \neq i\}$.

Theorem

A diagonally dominant matrix is non-singular.

Proof

Suppose that $A \in \mathbb{C}^{n \times n}$ is a diagonally dominant matrix that is singular. The homogeneous system $A\mathbf{x} = \mathbf{0}$ has a non-trivial solution $\mathbf{x} \neq \mathbf{0}$. Let x_k be a component of $\mathbf{x}$ that has the largest absolute value. Since $\mathbf{x} \neq \mathbf{0}$, we have $|x_k| > 0$. We can write

$$a_{kk}x_k = -\sum\{a_{kj}x_j \mid 1 \leqslant j \leqslant n \text{ and } j \neq k\},$$

which implies

$$\begin{aligned} |a_{kk}|\,|x_k| &= \left|\sum\{a_{kj}x_j \mid 1 \leqslant j \leqslant n \text{ and } j \neq k\}\right| \\ &\leqslant \sum\{|a_{kj}|\,|x_j| \mid 1 \leqslant j \leqslant n \text{ and } j \neq k\}\Big| \\ &\leqslant |x_k| \sum\{|a_{kj}| \mid 1 \leqslant j \leqslant n \text{ and } j \neq k\}. \end{aligned}$$

Thus, we obtain

$$|a_{kk}| \leqslant \sum\{|a_{kj}| \mid 1 \leqslant j \leqslant n \text{ and } j \neq k\},$$

which contradicts the fact that A is diagonally dominant.

We begin with a class of linear systems that can easily be solved.

Definition

A matrix $C \in \mathbb{C}^{m \times n}$ is in row echelon form if the following conditions are satisfied:

- rows that contain a non-zero elements precede zero rows (that is, rows that contain only zeros);
- if c_{ij} is the first non-zero element of the row i, all elements in the j^{th} column located below c_{ij}, that is, entries of the form c_{kj} with $k > j$ are zero;
- if $i < \ell$, c_{ij_i} is the first non-zero element of the row i, and $c_{\ell j_\ell}$ is the first non-zero element of the row ℓ, then $j_i < j_\ell$.

The first non-zero element of a row i (if it exists) is called the *pivot of the row i*.

Example

Let $C \in \mathbb{R}^{4\times 5}$ be the matrix

$$C = \begin{pmatrix} 1 & 2 & 0 & 2 & 0 \\ 0 & 2 & 3 & 0 & 1 \\ 0 & 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

It is clear that C is in row echelon form; the pivots of the first, second and third rows are $c_{11} = 1$, $c_{22} = 2$, and $c_{34} = -1$.

Theorem

Let $C \in \mathbb{C}^{m\times n}$ be a matrix in row echelon form such that the rows that contain non-zero elements are the first r rows. Then, $\text{rank}(C) = r$.

Proof: Let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_r$ be the non-zero rows of C. Suppose that the row $\boldsymbol{c}_i$ has the first non-zero element in the column j_i for $1 \leqslant i \leqslant r$. By the definition of the echelon form we have $j_1 < j_2 < \cdots < j_r$.

Suppose that $a_1\boldsymbol{c}_1 + \cdots + a_r\boldsymbol{c}_r = \mathbf{0}$. This equality can be written as:

$$\begin{aligned} a_1 c_{1j_1} &= 0, \\ a_1 c_{1j_2} + a_2 c_{2j_2} &= 0, \\ &\vdots \\ a_1 c_{1n} + a_2 c_{2n} + \ldots + a_r c_{rn} &= 0. \end{aligned}$$

Since $c_{1j_1} \neq 0$, we have $a_1 = 0$. Substituting a_1 by 0 in the second equality implies $a_2 = 0$ because $c_{2j_2} \neq 0$, *etc*. Thus, we obtain $a_1 = a_2 = \ldots = a_r = 0$, which proves that the rows $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_r$ are linearly independent. Since this is a maximal set of rows of C that is linearly independent, it follows that $\text{rank}(C) = r$.

Linear systems whose augmented matrices are in row echelon form can be easily solved using a process called *back substitution*. Consider the augmented matrix in row echelon form of a system with m equations and n unknowns:

$$\begin{pmatrix} 0 & \cdots & 0 & a_{1j_1} & \cdots & \cdots & \cdots & a_{1n} & b_1 \\ 0 & \cdots & 0 & 0 & \cdots & a_{2j_2} & \cdots & a_{2n} & b_2 \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \cdots & a_{rj_r} & \cdots & b_r \\ 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 & \cdots & b_{r+1} \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 & \cdots & b_m \end{pmatrix}$$

The system of equations has the form:

$$\begin{aligned} a_{1j_1}x_{j_1} + \cdots\cdots\cdots + a_{1n}x_n &= b_1 \\ a_{2j_2}x_{j_2} + \cdots\cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{rj_r}x_{j_r} + \cdots + a_{rn}x_n &= b_r \end{aligned}$$

The variables $x_{j_1}, x_{j_2}, \ldots, x_{j_r}$ that correspond to the columns where the pivot element occur are referred to as the *basic variables*. The remaining variables are *non-basic*.

Note that we have $r \leqslant \min\{m, n\}$. If $r < m$ and there exists $b_\ell \neq 0$ for $r < \ell \leqslant m$, then the system is inconsistent and no solutions exist. If $r = m$ or $b_\ell = 0$ for $r < \ell \leqslant m \leqslant n$, one can choose the variables that do not correspond to the pivot elements, $\{x_i \mid i \not\in \{j_1, j_2, \ldots, j_r\}$ as parameters and express the basic variables as functions of these parameters.

UMASS BOSTON

The process starts with the last basic variable, x_{j_r} (because every other variable in the equation $a_{rj_r}x_{j_r} + \cdots + a_{rn}x_n = b_r$ is a parameter), and then substitutes this variable in the previous equality. This allows us to express $x_{j_{r-1}}$ as a function of parameters, *etc*. This explains the term *back substitution* previously introduced. If $r = n$, then no parameters exist. To conclude, if $r < m$ the system has a solution if and only if $b_j = 0$ for $j > r$. If $r = m$, the system has a solution. This solution is unique if $r = n$.

Example

Consider the system

$$\begin{array}{ccccccccccc}
x_1 & + & 2x_2 & + & & & & & 2x_4 & = & b_1 \\
 & & 2x_2 & + & 3x_3 & + & & & x_5 & = & b_2 \\
 & & & & & & -x_4 & + & x_5 & = & b_3 \\
 & & & & & & & & 0 & = & b_4
\end{array}$$

The augmented matrix of this system is

$$\begin{pmatrix} 1 & 2 & 0 & 2 & 0 & b_1 \\ 0 & 2 & 3 & 0 & 1 & b_2 \\ 0 & 0 & 0 & -1 & 2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & b_4 \end{pmatrix}.$$

The basic variables are x_1, x_2 and x_4. If $b_4 = 0$ the system is consistent. Under this assumption we can choose x_3 and x_5 as parameters. Let $x_3 = p$ and $x_5 = q$.

Example

The third equation yields $x_4 = q - b_3$. Similarly, the second equation implies $x_2 = 0.5(b_2 - 3p - q)$. Substituting these value in the first equation allows us to write $x_1 = b_1 - b_2 + 2b_3 - 3p - q$.
Further transformations of this system allow us to construct an equivalent linear system whose matrix contain the columns of the matrix I_3.
Subtracting the second row from the first yields:

$$\begin{pmatrix} 1 & 0 & -3 & 2 & -1 & b_1 - b_2 \\ 0 & 2 & 3 & 0 & 1 & b_2 \\ 0 & 0 & 0 & -1 & 2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & b_4 \end{pmatrix}.$$

Example

Then, dividing the second row by 2 will produce:

$$\begin{pmatrix} 1 & 0 & -3 & 2 & -1 & b_1 - b_2 \\ 0 & 1 & 3/2 & 0 & 1/2 & b_2/2 \\ 0 & 0 & 0 & -1 & 2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & b_4 \end{pmatrix}.$$

which creates the second column of I_3. Next, multiply the third row by -1:

$$\begin{pmatrix} 1 & 0 & -3 & 2 & -1 & b_1 - b_2 \\ 0 & 1 & \frac{3}{2} & 0 & \frac{1}{2} & \frac{b_2}{2} \\ 0 & 0 & 0 & 1 & -2 & -b_3 \\ 0 & 0 & 0 & 0 & 0 & b_4 \end{pmatrix}.$$

Finally, multiply the third by -2 and add it to the first row:

$$\begin{pmatrix} 1 & 0 & -3 & 0 & 1 & b_1 - b_2 + 2b_3 \\ 0 & 1 & \frac{3}{2} & 0 & \frac{1}{2} & \frac{b_2}{2} \\ 0 & 0 & 0 & 1 & -2 & -b_3 \\ 0 & 0 & 0 & 0 & 0 & b_4 \end{pmatrix}.$$

Example

Thus, if we choose for the basic variables

$$x_1 = b_1 - b_2 + 2b_3, x_2 = \frac{b_2}{2}, \text{ and } x_4 = -b_3$$

and for the non-basic variables $x_3 = x_5 = 0$ we obtain a solution of the system.

We shall see that the extended echelon form of a system can be achieved by applying certain transformations on the rows of the augmented matrix of the system (which amount to transformations involving the equations of the system). In preparation, a few special invertible matrices are introduced in the next examples.

Example

Consider the matrix

$$T^{(i)\leftrightarrow(j)} = \begin{pmatrix} 1 & & \cdots & & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 0 & \cdots & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \cdots & & 1 \end{pmatrix},$$

where line i contains exactly one 1 in position j and line j contains exactly one 1 in position i. If $T^{(i)\leftrightarrow(j)} \in \mathbb{C}^{p\times p}$ and $A \in \mathbb{C}^{p\times q}$, it is easy to see that the matrix $T^{(i)\leftrightarrow(j)}A$ is obtained from the matrix A by permuting the lines i and j.

Example

For instance, consider the matrix $T^{(2)\leftrightarrow(4)} \in \mathbb{C}^{4\times 4}$ defined by:

$$T^{(2)\leftrightarrow(4)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

and the matrix $A \in \mathbb{C}^{4\times 5}$.

Example

We have:

$$T^{(2)\leftrightarrow(4)}A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \end{pmatrix}.$$

The inverse of $T^{(i)\leftrightarrow(j)}$ is $T^{(i)\leftrightarrow(j)}$ itself.

Example

Let $T^{a(i)} \in \mathbb{C}^{p \times p}$ be the matrix

$$T^{a(i)} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots 0 \\ 0 & 1 & \cdots & 0 & \cdots 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots 0 \\ \cdots & \cdots & \cdots & a & \cdots 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots 0 \\ 0 & 0 & \cdots & 0 & \cdots 1 \end{pmatrix}$$

that has $a \in F - \{0\}$ on the i^{th} diagonal element, 1 on the remaining diagonal elements and 0 everywhere else. The product $T^{a(i)}A$ is obtained from A by multiplying the i^{th} row by a. The inverse of this matrix is $T^{\frac{1}{a}(i)}$.

UMASS BOSTON

Example

Consider the matrix $T^{3(2)} \in \mathbb{C}^{4\times 4}$ given by

$$T^{3(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

If $A \in \mathbb{C}^{4\times 5}$ the matrix $T^{3(2)}A$ is obtained from A by multiplying its second line by 3. We have

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 3a_{21} & 3a_{22} & 3a_{23} & 3a_{24} & 3a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix}.$$

UMASS BOSTON

Example

Let $T^{(i)+a(j)} \in \mathbb{C}^{p\times p}$ be the matrix whose entries are identical to the matrix I_p with the exception of the element located in row i and column j that equals a:

$$T^{(i)+a(j)} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 & \cdots 0 \\ 0 & 1 & \cdots & \cdots & \cdots & 0 & \cdots 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots 0 \\ 0 & 0 & \cdots & a & \cdots & 1 & \cdots 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots 1 \end{pmatrix}.$$

The result of the multiplication $T^{(i)+a(j)}A$ is a matrix that can be obtained from A by adding the j^{th} line of A multiplied by a to the i^{th} line of A. The inverse of the matrix $T^{(i)+a(j)}A$ is $T^{(i)-a(j)}A$.

UMASS BOSTON

Example

We have

$$T^{(4)+2(2)}A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} + 2a_{21} & a_{42} + 2a_{22} & a_{43} + 2a_{23} & a_{44} + 2a_{24} & a_{45} + 2a_{25} \end{pmatrix}.$$

It is easy to see that if one multiplies a matrix A at the right by $T^{(i)\leftrightarrow(j)}$, $T^{a(i)}$, and $T^{(i)+a(j)}$ the effect on A consists of exchanging the columns i and j, multiplying the i^{th} column by a, and adding the j^{th} column multiplied by a to the i^{th} column, respectively.

Definition

Let $\mathbb{F}$ be a field, $A, C \in \mathbb{F}^{m \times n}$ and let $\boldsymbol{b}, \boldsymbol{d} \in \mathbb{F}^{m \times 1}$. Two systems of linear equations $A\mathbf{x} = \boldsymbol{b}$ and $C\mathbf{x} = \boldsymbol{d}$ are *equivalent* if they have the same set of solutions.

If $A\mathbf{x} = \boldsymbol{b}$ is a system of linear equations in matrix form, where $A \in \mathbb{C}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{C}^{m \times 1}$, and $T \in \mathbb{C}^{m \times m}$ is a matrix that has an inverse, then the systems $A\mathbf{x} = \boldsymbol{b}$ and $(TA)\mathbf{x} = (T\boldsymbol{b})$ are equivalent. Indeed, any solution of $A\mathbf{x} = \boldsymbol{b}$ satisfies the system $(TA)\mathbf{x} = (T\boldsymbol{b})$. Conversely, if $(TA)\mathbf{x} = (T\boldsymbol{b})$, by multiplying this equality by T^{-1} to the left, we get $(T^{-1}T)A\mathbf{x} = (T^{-1}T)\boldsymbol{b}$, that is, $A\mathbf{x} = \boldsymbol{b}$.

The matrices $T^{(i)\leftrightarrow(j)}$, $T^{a(i)}$, and $T^{(i)+a(j)}$ play a special role in an algorithm that transforms a linear system $A\boldsymbol{x} = \boldsymbol{b}$ into an equivalent system in row echelon form. These transformations are known as *elementary transformation matrices*.

Algorithm for the Row Echelon Form of a Matrix

Data:An matrix $A \in \mathbb{F}^{p \times q}$.
Result: A row echelon form of A.

$r = 1$; $c = 1$;
while($r \leqslant p$ and $c \leqslant q$) do {
  while ($A(*, c) = \mathbf{0}$) {c=c+1}
  $j = r$;
  while ($A(j, c) = 0$) {j = j+1}
  if ($j \neq r$) {exchange line r with line j}
multiply line r by $\frac{1}{A(r,c)}$
ForEach ($k = r + 1$ to p)
  {add line r multiplied by $-A(k, c)$ to line k}
$r = r + 1$; $c = c + 1$;
}

Example

Consider the linear system

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 4 \\ x_1 + 2x_2 + x_3 &= 3 \\ x_1 + 3x_2 + x_3 &= 1. \end{aligned}$$

The augmented matrix of this system is

$$[A|\boldsymbol{b}] = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 1 & 3 \\ 1 & 3 & 1 & 1 \end{pmatrix}.$$

By subtracting the first row from the second and the third we obtain the matrix

$$T^{(3)-1(1)} T^{(2)-1(1)} [A|\boldsymbol{b}] = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & -2 & -1 \\ 0 & 1 & -2 & -3 \end{pmatrix}.$$

Example

Next, the second and third row are exchanged yielding the matrix

$$T^{(2)\leftrightarrow(3)}T^{(3)-1(1)}T^{(2)-1(1)}[A|\boldsymbol{b}] = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & -2 & -1 \end{pmatrix}.$$

To obtain an 1 in the pivot of the third row we multiply the third row by $-\frac{1}{2}$:

$$T^{-0.5(3)}T^{(2)\leftrightarrow(3)}T^{(3)-1(1)}T^{(2)-1(1)}[A|\boldsymbol{b}] = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & 0.5 \end{pmatrix},$$

which is the row echelon form of the matrix $[A|\boldsymbol{b}]$.

Example

To achieve the row echelon form we needed to multiply the matrix $[A|\boldsymbol{b}]$ by the matrix

$$T = T^{-0.5(3)} T^{(2)\leftrightarrow(3)} T^{(3)-1(1)} T^{(2)-1(1)}.$$

The solutions of the system can now be obtained by back substitution from the linear system

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 4, \\ x_2 - 2x_3 &= -3, \\ x_3 &= 0.5. \end{aligned}$$

The last equation yields $x_3 = 0.5$. Substituting x_3 in the second equation implies $x_2 = -2$; finally, from the first equality we have $x_1 = 6.5$.

Theorem

Let $T^{a(i)}$, $T^{(p)\leftrightarrow(q)}$, and $T^{(i)+a(j)}$ be the matrices in $\mathbb{R}^{m\times m}$ that correspond to the row transformations applied to matrices in $\mathbb{R}^{m\times n}$, where $i \neq j$ and $p \neq q$. We have:

$$T^{a(i)}T^{(p)\leftrightarrow(q)} = \begin{cases} T^{(p)\leftrightarrow(q)}T^{a(i)} & \text{if } i \not\in \{p,q\}, \\ T^{(p)\leftrightarrow(q)}T^{a(q)} & \text{if } i = p, \\ T^{(p)\leftrightarrow(q)}T^{a(p)} & \text{if } i = q, \end{cases},$$

$$T^{(i)+a(j)}T^{(p)\leftrightarrow(q)} = \begin{cases} T^{(p)\leftrightarrow(q)}T^{(i)+a(j)} & \text{if } \{i,j\} \cap \{p,q\} = \emptyset, \\ T^{(q)+a(j)}T^{(p)\leftrightarrow(q)} & \text{if } i = p \text{ and } j \neq q, \\ T^{(i)+a(p)}T^{(p)\leftrightarrow(q)} & \text{if } i \neq p \text{ and } j = q, \\ T^{(q)+a(p)}T^{(p)\leftrightarrow(q)} & \text{if } i = p \text{ and } j = q. \end{cases}$$

UMASS BOSTON

The matrices that describe elementary transformations are of two types: lower triangular matrices of the form $T^{a(i)}$ or $T^{(i)+a(j)}$ or permutations matrices of the form $T^{(p)\leftrightarrow(q)}$.

If all pivots encountered in the construction of the row echelon form of the matrix A are not-zero then there is no need to have any permutation matrix $T^{(p)\leftrightarrow(q)}$ among the matrices that multiply A at the left. Thus, there is a lower matrix T and an upper triangular matrix U such that $TA = U$.

The matrix T is a product of invertible matrices and therefore it is invertible. Since the inverse $L = T^{-1}$ of a lower triangular matrix is lower triangular, it follows that $A = LU$; in other words, A can be decomposed into a product of a lower triangular and an upper triangular matrix. This decomposition is known as an *LU-decomposition* of A.

Example

Let $A \in \mathbb{R}^{3\times 3}$ be the matrix

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

Initially, we add the first row multiplied by -2 to the second row, and the same first row, multiplied by -1 to the third row. This amounts to

$$T^{(3),-(1)}T^{(2),-2(1)}A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

Example

Next, we add the second row to the third to produce the matrix

$$T^{(3)+(2)}T^{(3),-(1)}T^{(2),-2(1)}A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix},$$

which is an upper triangular matrix. We can conclude that $\text{rank}(A) = 2$ and we can write:

$$A = (T^{(2),-2(1)})^{-1}(T^{(3),-(1)})^{-1}(T^{(3)+(2)})^{-1} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

Example

Thus, the lower triangular matrix we are seeking is

$$L = (T^{(2)-2(1)})^{-1}(T^{(3),-(1)})^{-1}(T^{(3)+(2)})^{-1} = T^{(2)+2(1)}T^{(3)+(1)}T^{(3)-(2)}$$

$$= \begin{pmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 1 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & -1 & 1\end{pmatrix} = \begin{pmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 1 & -1 & 1\end{pmatrix},$$

which shows that A can be written as:

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

where the first matrix is lower triangular and the second is upper triangular.

Suppose that during the construction of the matrix U some of the elementary transformation matrices are permutations matrices of the form $T^{(p)\leftrightarrow(q)}$.

Matrices of the form $T^{(p)\leftrightarrow(q)}$ can be shifted to the right. Therefore, instead on the previous factorization of the matrix A we have a lower triangular matrix T and a permutation matrix (which results as a product of all permutation matrices of the form $T^{(p)\leftrightarrow(q)}$ used in the algorithm such that $TPA = U$. In this case we obtain an LU-factorization of PA instead of A.

Theorem

Let $A \in \mathbb{R}^{m \times n}$ be a matrix. If rank$(A) = k$, then a largest non-singular square submatrix B of A is a $k \times k$-matrix.

The function `inv` computes the inverse of an invertible square matrix.

Example

Let `A` be the invertible matrix

```
A =

     1     2
     2     3
```

Its inverse is given by;

```
>> inv(A)

ans =

   -3.0000    2.0000
    2.0000   -1.0000
```

Example

On other hand, if `inv` is applied to a singular matrix

```
A =
     1     2
     2     4
```

an error message is posted:

```
Warning: Matrix is singular to working precision.
```

If A is non-singular, the function `inv` can be used to solve the system $A\mathbf{x} = \mathbf{b}$ by writing `x = inv(A)*b`, although a better method is described below.

Example

For

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 13 \\ 23 \end{pmatrix}$$

the solution of the system $A\mathbf{x} = \mathbf{b}$ is

```
x = inv(A)*b


x =
    7.0000
    3.0000
```

UMASS BOSTON

This is not the best way for solving a system of linear equations. In certain circumstances, this method produces errors and has a poor time performance.

A better approach is for solving a linear system $A\boldsymbol{x} = \boldsymbol{b}$ is to use the backslash operator `x = A \ b` or `x = mldivide(A,b)`. The term `mldivide` is related to the position of the matrix A at the left of $\boldsymbol{x}$.

Example

Define A and $\boldsymbol{b}$ as

```
>> A = [5 11 2; 10 6 -4; -2 9 7]
A =
     5    11     2
    10     6    -4
    -2     9     7
>> b=[53;26;48]
b =
    53
    26
    48
```

Then either `x=A\b` or `x=mldivide(A,b)` produces

```
x =
    1.0000
    4.0000
    2.0000
```

BOSTON

The system $\boldsymbol{x}A = \boldsymbol{c}$, where $A \in \mathbb{R}^{n \times m}$ and $\boldsymbol{c}' \in \mathbb{R}^m$ can be solved using either `x = A / b` or `x = mrdivide(A,b)`.
It is easy to see that these operations are related by:

$$A \backslash b = (A'/b')'.$$

The function `rref` produces the reduced row echelon form of a matrix A, when called as `R = rref(A)`.

A variant of this function, `[R,r] = rref(A)` also yields a vector `r` so that `r` indicates the non-zero pivots, `length(r)` is the rank of A, and `A(:,r)` is a basis for the range of A. Roundoff errors may cause this algorithm to produce a rank for A that is different from the actual rank.

A pivot tolerance `tol` used by the algorithm to determine negligible columns can be specified using `rref(A,tol)`.

Example

Starting from the matrix

```
A =
     1     2     3     4     5     6
     7     8     9    10    11    12
     1     3     5     7     9    11
```

the function call `[R,r]=rref(A)` returns

```
R =
     1     0    -1    -2    -3    -4
     0     1     2     3     4     5
     0     0     0     0     0     0


r =
     1     2
```

showing that the rank of A is 2.

Let $A\boldsymbol{u} = \boldsymbol{b}$ be a linear system, where $A \in \mathbb{C}^{n\times n}$ is a non-singular matrix and $\boldsymbol{b} \in \mathbb{R}^n$.
We examine the sensitivity of the solution of this system to small variations of $\boldsymbol{b}$. So, together with the original system, we work with a system of the form $A\boldsymbol{v} = \boldsymbol{b} + \boldsymbol{h}$, where $\boldsymbol{h} \in \mathbb{R}^n$ is the *perturbation* of $\boldsymbol{b}$. Note that $A(\boldsymbol{v} - \boldsymbol{u}) = \boldsymbol{h}$, so $\boldsymbol{v} - \boldsymbol{u} = A^{-1}\boldsymbol{h}$. Using a vector norm $\| \cdot \|$ and its corresponding matrix norm $|||\cdot|||$ we have

$$\| \boldsymbol{v} - \boldsymbol{u} \| = \| A^{-1}\boldsymbol{h} \| \leqslant |||A^{-1}||| \, \| \boldsymbol{h} \| .$$

Since

$$\| \boldsymbol{v} - \boldsymbol{u} \| = \| A^{-1}\boldsymbol{h} \| \leqslant |||A^{-1}||| \, \| \boldsymbol{h} \| .$$

and $\| \boldsymbol{b} \| = \| A\boldsymbol{u} \| \leqslant |||A||| \, \| \boldsymbol{u} \|$, it follows that

$$\begin{aligned} \frac{\| \boldsymbol{v} - \boldsymbol{u} \|}{\| \boldsymbol{u} \|} &\leqslant \frac{|||A^{-1}||| \, \| \boldsymbol{h} \|}{\frac{\|\boldsymbol{b}\|}{|||A|||}} \\ &= \frac{|||A||| |||A^{-1}||| \, \| \boldsymbol{h} \|}{\| \boldsymbol{b} \|}. \end{aligned} \quad (1)$$

Thus, the relative variation of the solution, $\frac{\|\boldsymbol{v}-\boldsymbol{u}\|}{\|\boldsymbol{u}\|}$ is upper bounded by the number $\frac{|||A||| |||A^{-1}||| \|\boldsymbol{h}\|}{\|\boldsymbol{b}\|}$.

Definition

Let $A \in \mathbb{C}^{n \times n}$ be a non-singular matrix. The *condition number of A relative to the matrix norm* $\|\!|\cdot|\!\|$ is the number $\text{cond}(A) = \|\!|A|\!\|\|\!|A^{-1}|\!\|$.

The Equality

$$\frac{\parallel \boldsymbol{v} - \boldsymbol{u} \parallel}{\parallel \boldsymbol{u} \parallel} = \frac{\|A\| \|A^{-1}\| \parallel \boldsymbol{h} \parallel}{\parallel \boldsymbol{b} \parallel}.$$

implies that if the condition number is large, then small variations in $\boldsymbol{b}$ may generate large variations in the solution of the system $A\boldsymbol{u} = \boldsymbol{b}$, especially when $\boldsymbol{b}$ is close to $\mathbf{0}$. When this is the case, we say that the system $A\boldsymbol{u} = \boldsymbol{b}$ is *ill-conditioned*. Otherwise, the system $A\boldsymbol{u} = \boldsymbol{b}$ is *well-conditioned*.

Theorem

Let $A \in \mathbb{C}^{n \times n}$ be a non-singular matrix. The following statements hold for every matrix norm induced by a vector norm:

- $cond(A) = cond(A^{-1})$;
- $cond(cA) = |c| cond(A)$;
- $cond(A) \geqslant 1$.

Proof.

We prove here only Part (iii). Since $AA^{-1} = I$, by the properties of a matrix norm induced by a vector norm we have:

$$\mathrm{cond}(A) = \|A\|\|A^{-1}\| \geqslant \|AA^{-1}\| = \|I_n\| = 1.$$

□

Let A, B be two non-singular matrices in $C^{n\times n}$ such that $B = aA$, where $a \in \mathbb{C}$.
We have $B^{-1} = a^{-1}A^{-1}$, $\|B\| = |a|\|B\|$ and $\|B^{-1}\| = |a|^{-1}\|A^{-1}\|$ so $\text{cond}(B) = \text{cond}(A)$.
On another hand, $\det(B) = a^n \det(A)$. Thus, if n is large enough and $a < 1$, then $\det(B)$ can be quite close to 0, while the condition number of B may be quite large. This shows that the determinant and the condition number are relatively independent.

Example

Let $A \in \mathbb{C}^{2\times 2}$ be the matrix

$$A = \begin{pmatrix} a & a+\alpha \\ a+\alpha & a+2\alpha \end{pmatrix},$$

where $a > 0$ and $\alpha < 0$. We have

$$A^{-1} = \begin{pmatrix} -\frac{a+2\alpha}{\alpha^2} & \frac{a+\alpha}{\alpha^2} \\ \frac{a+\alpha}{\alpha^2} & -\frac{a}{\alpha^2} \end{pmatrix},$$

so $\|A\|_1 = a$ and $\|A^{-1}\|_1 = \frac{a}{\alpha^2}$. Thus, $\text{cond}(A) = \left(\frac{a}{\alpha}\right)^2$ and, if $|\alpha|$ is small a system of the for $A\boldsymbol{u} = \boldsymbol{b}$ may be ill-conditioned.

Theorem

Let $A \in \mathbb{C}^{n\times n}$ be an invertible matrix and let $\| \cdot \|$ be a norm on $\mathbb{C}^n$. We have:

$$|||A^{-1}||| = \frac{1}{\min\{\| A\mathbf{x} \| \mid \| \mathbf{x} \| = 1\}},$$

where $||| \cdot |||$ is the matrix norm generated by $\| \cdot \|$.

Proof

We claim that

$$\{\| A^{-1}\boldsymbol{t} \| \mid \| \boldsymbol{t} \| = 1\} = \left\{ \frac{1}{\| A\boldsymbol{x} \|} \mid \| \boldsymbol{x} \| = 1 \right\}.$$

Let $a = \| A^{-1}\boldsymbol{t} \|$ for some $\boldsymbol{t} \in \mathbb{C}^n$ such that $\| \boldsymbol{t} \| = 1$. Define $\boldsymbol{x}$ as

$$\boldsymbol{x} = \frac{1}{\| A^{-1}\boldsymbol{t} \|} A^{-1}\boldsymbol{t}.$$

Clearly, we have $\| \boldsymbol{x} \| = 1$. In addition,

$$\| A\boldsymbol{x} \| = \frac{\| \boldsymbol{t} \|}{\| A^{-1}\boldsymbol{t} \|} = \frac{1}{\| A^{-1}\boldsymbol{t} \|} = \frac{1}{a},$$

so $a \in \left\{ \frac{1}{\|A\boldsymbol{x}\|} \mid \| \boldsymbol{x} \| = 1 \right\}$. Thus,

$$\{\| A^{-1}\boldsymbol{t} \| \mid \| \boldsymbol{t} \| = 1\} \subseteq \left\{ \frac{1}{\| A\boldsymbol{x} \|} \mid \| \boldsymbol{x} \| = 1 \right\}.$$

Proof (cont'd)

Let $b = \frac{1}{\|A\mathbf{x}\|}$ for some $\mathbf{x}$ such that $\| \mathbf{x} \| = 1$.
Define $\mathbf{y} = \frac{1}{\|A\mathbf{x}\|} A\mathbf{x}$. We have $\| \mathbf{y} \| = 1$. Also,

$$A^{-1}\mathbf{y} = \frac{1}{\| A\mathbf{x} \|}\mathbf{x},$$

so $b \in \left\{ \| A^{-1}\mathbf{z} \| \mid \| \mathbf{z} \| = 1 \right\}$. Thus

$$\{ \| A^{-1}\mathbf{z} \| \mid \| \mathbf{z} \| = 1 \} \supseteq \left\{ \frac{1}{\| A\mathbf{x} \|} \mid \| \mathbf{x} \| = 1 \right\},$$

hence we have the equality.

Ill-conditioned linear systems $A\mathbf{u} = \mathbf{b}$ may occur when large differences in scale exists among the columns of A, or among the rows of A.

Theorem

Let $A = (\mathbf{a}_1 \ \cdots \ \mathbf{a}_n)$ be an invertible matrix in $\mathbb{C}^{n\times n}$, where $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are the columns of A. Then,

$$cond(A) \geqslant \max\left\{\frac{\parallel \mathbf{a}_i \parallel}{\parallel \mathbf{a}_j \parallel} \mid 1 \leqslant i, j \leqslant n\right\}.$$

Proof

Since $\text{cond}(A) = \|A\|\|A^{-1}\|$, we have

$$\text{cond}(A) = \frac{\max\{\| A\mathbf{x} \| \mid \| \mathbf{x} \| = 1\}}{\min\{\| A\mathbf{x} \| \mid \| \mathbf{x} \| = 1\}}$$

Note that $A\mathbf{e}_k = \mathbf{a}_k$, where $\mathbf{a}_k$ is the k^{th} column of A and that $\| \mathbf{e}_k \| = 1$. Therefore,

$$\begin{aligned} \max\{\| A\mathbf{x} \| \mid \| \mathbf{x} \| = 1\} &\geqslant \| \mathbf{a}_i \|, \\ \min\{\| A\mathbf{x} \| \mid \| \mathbf{x} \| = 1\} &\leqslant \| \mathbf{a}_j \|, \end{aligned}$$

which implies

$$\text{cond}(A) \geqslant \frac{\| \mathbf{a}_i \|}{\| \mathbf{a}_j \|}$$

for all $1 \leqslant i, j \leqslant n$. This yields the inequality of the theorem.

Example

Let

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

where $\alpha \in \mathbb{R}$ and $\alpha > 0$. The matrix A is invertible and

$$A^{-1} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\alpha} & \frac{1}{\alpha} \end{pmatrix}.$$

It is easy to see that the condition number of A relative to the Frobenius norm is

$$\text{cond}(A) = \frac{2 + \alpha^2}{\alpha}.$$

Thus, if α is sufficiently close to 0, the condition number can reach arbitrarily large values.

UMASS BOSTON

In general, we use as matrix norms, norms of the form $\|\cdot\|_p$. The corresponding condition number of a matrix A is denoted by $\text{cond}_p(A)$.

Example

Let $A = \text{diag}(a_1, \ldots, a_n)$ be a diagonal matrix. Then, $\|A\|_2 = \max_{1 \leqslant i \leqslant n} |a_i|$. Since $A^{-1} = \text{diag}\left(\frac{1}{a_1}, \ldots, \frac{1}{a_n}\right)$, it follows that $\|A^{-1}\|_2 = \frac{1}{\min_{1 \leqslant i \leqslant n} |a_i|}$, so $\text{cond}_2(A) = \frac{\max_{1 \leqslant i \leqslant n} |a_i|}{\min_{1 \leqslant i \leqslant n} |a_i|}$.

The condition number of a matrix A is computed using the function `cond(A,p)` which returns the p-norm condition of matrix A. When used with a single parameter, as in `cond(A)`, the 2-norm condition number of A is returned.

Example

Let A be the matrix

```
>> A=[10.1 6.2; 5.1 3.1]
A =
   10.1000    6.2000
    5.1000    3.1000
```

The condition number `cond(A)` is 567.966, which is quite large indicating significant sensitivity to inverse calculations. The inverse of A is

```
>> inv(A)
ans =
  -10.0000   20.0000
   16.4516  -32.5806
```

Example

If we make a small change in A yielding the matrix

```
>> B=[10.2 6.3;5.1 3.1]
B =
   10.2000    6.3000
    5.1000    3.1000
```

the inverse of B changes completely:

```
>> inv(B)
ans =
   -6.0784   12.3529
   10.0000  -20.0000
```

Values of the condition number close to 1 indicate a well-conditioned matrix, and the opposite is true for large values of the condition number.

Example

Consider the linear systems:

$$\begin{aligned} 10.1x_1 + 6.2x_2 &= 12 \\ 5.1x_1 + 3.1x_2 &= 6 \end{aligned} \quad \text{and} \quad \begin{aligned} 10.2x_1 + 6.3x_2 &= 12 \\ 5.1x_1 + 3.1x_2 &= 6 \end{aligned}$$

that correspond to $A\boldsymbol{x} = \boldsymbol{b}$ and $B\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{b} = \begin{pmatrix} 12 \\ 6 \end{pmatrix}$. In view of the resemblance of A and B one would expect their solutions to be close.

Example

However, this is not the case. The solution of $A\boldsymbol{x} = \boldsymbol{b}$ is

```
>> x=inv(A)*b
x =
         0
    1.9355
```

while the solution of $B\boldsymbol{x} = \boldsymbol{b}$ is

```
>> x=inv(B)*b
x =
    1.1765
         0
```