# 线性方程组

## Linear Equations

Baobin Li

Email: libb@ucas.ac.cn

School of Computer and Control

University of Chinese Academy of Sciences

- A fundamental problem that surfaces in all mathematical sciences is that of analyzing and solving m algebraic equations in n unknowns.
- The earliest recorded analysis of simultaneous equations is found in the ancient Chinese book Chiu-chang Suan-shu (Nine Chapters on Arithmetic), estimated to have been written some time around 200 B.C.
- In the beginning of Chapter VIII, there appears a problem of the following form.

*Three sheafs of a good crop, two sheafs of a mediocre crop, and one sheaf of a bad crop are sold for 39 dou. Two sheafs of good, three mediocre, and one bad are sold for 34 dou; and one good, two mediocre, and three bad are sold for 26 dou. What is the price received for each sheaf of a good crop, each sheaf of a mediocre crop, and each sheaf of a bad crop?*

- This problem would be formulated as three equations in three unknowns by writing

$$\begin{aligned} 3x + 2y + z &= 39, \\ 2x + 3y + z &= 34, \\ x + 2y + 3z &= 26, \end{aligned}$$

  where $x, y$, and $z$ represent the price for one sheaf of a good, mediocre, and bad crop, respectively.
- The Chinese saw right to the heart of the matter.
- Place coefficients of this system in a square array on a "counting board" and then manipulated the lines of the array according to prescribed rules of thumb.
- In Europe, the technique became known as Gaussian elimination in honor of the German mathematician Carl Gauss, whose extensive use of it popularized the method.

# Gaussian Elimination and Matrices

- Because this elimination technique is fundamental, we begin the study of our subject by learning how to apply this method in order to compute solutions for linear equations.
- A system of $m$ linear algebraic equations in $n$ unknowns is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

  where the $x_i$'s are the unkowns and the $a_{ij}$'s and the $b_i$'s are known constants.
- The $a_{ij}$'s are called the coefficients of the system, and the set of $b_i$'s is referred to as the right-hand side of the system.
- For any such system, there are exactly three possibilities for the set of solutions.

## Three Possibilities

- **UNIQUE SOLUTION:** There is one and only one set of values for the $x_i$'s that satisfies all equations simultaneously.
- **NO SOLUTION:** There is no set of values for the $x_i$'s that satisfies all equations simultaneously—the solution set is empty.
- **INFINITELY MANY SOLUTIONS:** There are infinitely many different sets of values for the $x_i$'s that satisfy all equations simultaneously. It is not difficult to prove that if a system has more than one solution, then it has infinitely many solutions. For example, it is impossible for a system to have exactly two different solutions.

- Part of the job in dealing with a linear system is to decide which one of these three possibilities is true.
- The other part of the task is to compute the solution if it is unique or to describe the set of all solutions if there are many solutions.
- Gaussian elimination is a tool that can accomplish all of these goals.

- Gaussian elimination is a methodical process of systematically transforming one system into another simpler, but equivalent, system by successively eliminating unknowns and eventually arriving at a system that is easily solvable.
- The elimination process relies on three simple operations by which to transform one system to another equivalent system.
- Let $E_k$ denote the $k^{th}$ equation

$$E_k : \quad a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n = b_k$$

and write the system as

$$S = \left\{ \begin{array}{l} E_1 \\ E_2 \\ \vdots \\ E_m \end{array} \right\}.$$

- For a linear system $S$, each of the following three elementary operations results in an equivalent system $S'$.

  (1) Interchange the $i^{th}$ and $j^{th}$ equations. That is, if

$$S = \left\{ \begin{array}{c} E_1 \\ \vdots \\ E_i \\ \vdots \\ E_j \\ \vdots \\ E_m \end{array} \right\}, then \quad S' = \left\{ \begin{array}{c} E_1 \\ \vdots \\ E_j \\ \vdots \\ E_i \\ \vdots \\ E_m \end{array} \right\}.$$

  (2) Replace the $i^{th}$ equation by a nonzero multiple of itself. That is

$$S' = \left\{ \begin{array}{c} E_1 \\ \vdots \\ \alpha E_i \\ \vdots \\ E_m \end{array} \right\}, \quad where \quad \alpha \neq 0.$$

(3) Replace the $j^{th}$ equation by a combination of itself plus a multiple of the $i^{th}$ equation. that is

$$S' = \left\{ \begin{array}{c} E_1 \\ \vdots \\ E_i \\ \vdots \\ E_j + \alpha E_i \\ \vdots \\ E_m \end{array} \right\}.$$

- The most common problem encountered in practice is the one in which there are $n$ equations as well as $n$ unknowns—called a square system.
- Since Gaussian elimination is straightforward for this case, we begin here and later discuss the other possibilities.
- What follows is a detailed description of Gaussian elimination as applied to the following simple (but typical) square system:

$$\begin{aligned} 2x &+ y &+ z &= 1, \\ 6x &+ 2y &+ z &= -1, \\ -2x &+ 2y &+ z &= 7. \end{aligned}$$

- At each step, the strategy is to focus on one position, called the **pivot position**, and to eliminate all terms below this position using the three elementary operations.
- The coefficient in the pivot position is called a pivotal element (or simply a pivot), while the equation in which the pivot lies is referred to as the pivotal equation.

- Only nonzero numbers are allowed to be pivots.
- If a coefficient in a pivot position is ever 0, then the pivotal equation is interchanged with an equation below the pivotal equation to produce a nonzero pivot.
- Unless it is 0, the first coefficient of the first equation is taken as the first pivot.
- For example, the circled ② in the system below is the pivot for the first step:

$$\begin{aligned} \textcircled{2}x &+ y + z = 1, \\ 6x &+ 2y + z = -1, \\ -2x &+ 2y + z = 7. \end{aligned}$$

**Step 1.** Eliminate all terms below the first pivot.

$$\begin{aligned} \textcircled{2}x &+ y + z = 1, & \\ - y &- 2z = -4, & (E_2 - 3E_1), \\ 3y &+ 2z = 8 & (E_3 + E_1). \end{aligned}$$

**Step 2.**Select a new pivot.

- For the time being, select a new pivot by moving down and to the right.
- If this coefficient is not 0, then it is the next pivot.
- Otherwise, interchange with an equation below this position so as to bring a nonzero number into this pivotal position.

**Step 3.**Eliminate all terms below the second pivot.

$$\begin{aligned} 2x &+ y &+ z &= 1, \\ &\textcircled{-1}y &- 2z &= -4, \\ & &- 4z &= -4 \quad (E_3 + 3E_2). \end{aligned}$$

- In general, at each step you move down and to the right to select the next pivot, then eliminate all terms below the pivot until you can no longer proceed.
- At this point, we say that the system has been **triangularized**.

- A triangular system is easily solved by a simple method known as back substitution.
- The last equation is solved for the value of the last unknown and then substituted back into the penultimate equation, which is in turn solved for the penultimate unknown, etc., until each unknown has been determined.
- It should be clear that there is no reason to write down the symbols such as "x," "y," "z, " and "= " at each step since we are only manipulating the coefficients.
- If such symbols are discarded, then a system of linear equations reduces to a rectangular array of numbers in which each horizontal line represents one equation.

$$\left(\begin{array}{rrr|r} 2 & 1 & 1 & 1 \\ 6 & 2 & 1 & -1 \\ -2 & 2 & 1 & 7 \end{array}\right)$$

- The array of coefficients—the numbers on the left-hand side of the vertical line—is called the **coefficient matrix** for the system.
- The entire array—the coefficient matrix augmented by the numbers from the right-hand side of the system—is called the **augmented matrix** associated with the system.
- If the coefficient matrix is denoted by $\mathbf{A}$ and the right-hand side is denoted by $\mathbf{b}$ , then the augmented matrix associated with the system is denoted by $[\mathbf{A}|\mathbf{b}]$.
- Formally, a scalar is either a real number or a complex number, and a matrix is a rectangular array of scalars. For example,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

- A submatrix of a given matrix $\mathbf{A}$ is an array obtained by deleting any combination of rows and columns from $\mathbf{A}$.
- Matrix $\mathbf{A}$ is said to have shape or size $m \times n$ whenever $\mathbf{A}$ has exactly m rows and n columns.
- To emphasize that matrix $\mathbf{A}$ has shape $m \times n$, subscripts are sometimes placed on $\mathbf{A}$ as $\mathbf{A}_{m\times n}$. Whenever $m = n$ (i.e., when $\mathbf{A}$ has the same number of rows as columns), $\mathbf{A}$ is called a square matrix. Otherwise, $\mathbf{A}$ is said to be rectangular.
- Matrices consisting of a single row or a single column are often called **row vectors** or **column vectors**, respectively.
- The symbol $\mathbf{A}_{i*}$ is used to denote the $i^{th}$ row, while $\mathbf{A}_{*j}$ denotes the $j^{th}$ column of matrix $\mathbf{A}$.
- For a linear system of equations, Gaussian elimination can be executed on the associated augmented matrix $[\mathbf{A}|\mathbf{b}]$ by performing elementary operations to the rows of $[\mathbf{A}|\mathbf{b}]$.

- For an $m \times n$ matrix $\mathbf{M}$, the three types of elementary row operations on $\mathbf{M}$ are as follows.

$$M = \begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{j*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}, \quad M' = \begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \mathbf{M}_{j*} \\ \vdots \\ \mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}, \quad \begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \alpha\mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}, \quad \begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{j*} + \alpha\mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}.$$

- Type I: Interchange rows $i$ and $j$.
- Type II: Replace row $i$ by a nonzero multiple of itself.
- Type III: Replace row $j$ by a combination of itself plus a multiple of row $i$.

- In general, if an $n \times n$ system has been triangularized to the form

$$\left(\begin{array}{cccc|c} t_{11} & t_{12} & \cdots & t_{1n} & c_1 \\ 0 & t_{22} & \cdots & t_{2n} & c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & t_{nn} & c_n \end{array}\right)$$

in which each $t_{ii} \neq 0$ (i.e., there are no zero pivots), then the general algorithm for back substitution is as follows.

**Algorithm for Back Substitution**

Determine the $x_i$'s by first setting $x_n = c_n/t_{nn}$ and then recursively computing

$$x_i = \frac{1}{t_{ii}}\left(c_i - t_{i,i+1}x_{i+1} - t_{i,i+2}x_{i+2} - \cdots - t_{in}x_n\right)$$

for $i = n-1,\, n-2, \ldots, 2,\, 1.$

- One way to gauge the efficiency of an algorithm is to count the number of arithmetical operations required.

## Gaussian Elimination Operation Counts

Gaussian elimination with back substitution applied to an $n \times n$ system requires

$$\frac{n^3}{3} + n^2 - \frac{n}{3} \quad \text{multiplications/divisions}$$

and

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} \quad \text{additions/subtractions.}$$

As $n$ grows, the $n^3/3$ term dominates each of these expressions. Therefore, the important thing to remember is that Gaussian elimination with back substitution on an $n \times n$ system requires about $n^3/3$ multiplications/divisions and about the same number of additions/subtractions.

**Problem:** Solve the following system using Gaussian elimination with back substitution:

$$\begin{aligned} v - w &= 3, \\ -2u + 4v - w &= 1, \\ -2u + 5v - 4w &= -2. \end{aligned}$$

**Solution:** The associated augmented matrix is

$$\left(\begin{array}{ccc|c} 0 & 1 & -1 & 3 \\ -2 & 4 & -1 & 1 \\ -2 & 5 & -4 & -2 \end{array}\right).$$

Since the first pivotal position contains 0, interchange rows one and two before eliminating below the first pivot:

$$\left(\begin{array}{ccc|c} \textcircled{0} & 1 & -1 & 3 \\ -2 & 4 & -1 & 1 \\ -2 & 5 & -4 & -2 \end{array}\right) \xrightarrow{\textit{Interchange } R_1 \textit{ and } R_2} \left(\begin{array}{ccc|c} \textcircled{-2} & 4 & -1 & 1 \\ 0 & 1 & -1 & 3 \\ -2 & 5 & -4 & -2 \end{array}\right) \begin{array}{l} \\ \\ R_3 - R_1 \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} -2 & 4 & -1 & 1 \\ 0 & \textcircled{1} & -1 & 3 \\ 0 & 1 & -3 & -3 \end{array}\right) \begin{array}{l} \\ \\ R_3 - R_2 \end{array} \longrightarrow \left(\begin{array}{ccc|c} -2 & 4 & -1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & -2 & -6 \end{array}\right).$$

Back substitution yields

$$\begin{aligned} w &= \frac{-6}{-2} = 3, \\ v &= 3 + w = 3 + 3 = 6, \\ u &= \frac{1}{-2}(1 - 4v + w) = \frac{1}{-2}(1 - 24 + 3) = 10. \end{aligned}$$

# Gauss-Jordan Method

- The purpose of this section is to introduce a variation of Gaussian elimination that is known as the **Gauss-Jordan method**.
- The two features that distinguish the Gauss-Jordan method from standard Gaussian elimination are as follows.
  1. At each step, the pivot element is forced to be 1.
  2. At each step, all terms above the pivot as well as all terms below the pivot are eliminated.
- In other words, if

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right)$$

is the augmented matrix associated with a linear system.

- Then elementary row operations are used to reduce this matrix to

$$\left(\begin{array}{cccc|c} 1 & 0 & \cdots & 0 & s_1 \\ 0 & 1 & \cdots & 0 & s_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & s_n \end{array}\right)$$

- The solution then appears in the last column (i.e., $x_i = s_i$ ) so that this procedure circumvents the need to perform back substitution.

**Problem:** Apply the Gauss–Jordan method to solve the following system:

$$\begin{aligned} 2x_1 + 2x_2 + 6x_3 &= 4, \\ 2x_1 + x_2 + 7x_3 &= 6, \\ -2x_1 - 6x_2 - 7x_3 &= -1. \end{aligned}$$

$$\left(\begin{array}{ccc|c} \textcircled{2} & 2 & 6 & 4 \\ 2 & 1 & 7 & 6 \\ -2 & -6 & -7 & -1 \end{array}\right) R_1/2 \longrightarrow \left(\begin{array}{ccc|c} \textcircled{1} & 1 & 3 & 2 \\ 2 & 1 & 7 & 6 \\ -2 & -6 & -7 & -1 \end{array}\right) \begin{array}{l} \\ R_2 - 2R_1 \\ R_3 + 2R_1 \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} \textcircled{1} & 1 & 3 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -4 & -1 & 3 \end{array}\right) (-R_2) \longrightarrow \left(\begin{array}{ccc|c} 1 & 1 & 3 & 2 \\ 0 & \textcircled{1} & -1 & -2 \\ 0 & -4 & -1 & 3 \end{array}\right) \begin{array}{l} R_1 - R_2 \\ \\ R_3 + 4R_2 \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 4 & 4 \\ 0 & \textcircled{1} & -1 & -2 \\ 0 & 0 & -5 & -5 \end{array}\right) -R_3/5 \longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 4 & 4 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & \textcircled{1} & 1 \end{array}\right) \begin{array}{l} R_1 - 4R_3 \\ R_2 + R_3 \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & \textcircled{1} & 1 \end{array}\right).$$

Therefore, the solution is $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$.

## Gauss–Jordan Operation Counts

For an $n \times n$ system, the Gauss–Jordan procedure requires

$$\frac{n^3}{2} + \frac{n^2}{2} \quad \text{multiplications/divisions}$$

and

$$\frac{n^3}{2} - \frac{n}{2} \quad \text{additions/subtractions.}$$

In other words, the Gauss–Jordan method requires about $n^3/2$ multiplications/divisions and about the same number of additions/subtractions.

# Making Gaussian Elimination Work

- It's time to turn into a practical algorithm for realistic applications.
- Practical applications usually demand the use of a computer.
- Computers will produce a more predictable kind of error: *roundoff error*.
- Numerical computation in digital computers is performed by approximating the infinite set of real numbers with a finite set of numbers as described below.

**Floating-Point Numbers**

A $t$-digit, base-$\beta$ ***floating-point number*** has the form

$$f = \pm .d_1 d_2 \cdots d_t \times \beta^{\epsilon} \quad \text{with} \quad d_1 \neq 0,$$

where the base $\beta$, the exponent $\epsilon$, and the digits $0 \leq d_i \leq \beta - 1$ are integers. For internal machine representation, $\beta = 2$ (binary representation) is standard, but for pencil-and-paper examples it's more convenient to use $\beta = 10$. The value of $t$, called the ***precision,*** and the exponent $\epsilon$ can vary with the choice of hardware and software.

- Floating-point numbers are just adaptations of the familiar concept of scientific notation where $\beta = 10$.
- For any fixed set of values for $t, \beta$, and $\epsilon$ , the corresponding set $\mathcal{F}$ of floating-point numbers is necessarily a finite set, so some real numbers can't be found in $\mathcal{F}$.
- There is more than one way of approximating real numbers with floating-point numbers. The common rounding convention is adopted.
- Given a real number $x$, the floating-point approximation $fl(x)$ is defined to be the nearest element in $\mathcal{F}$ to $x$, and in case of a tie we round away from 0.
- This means that for $t$-digit precision with $\beta = 10$, we need to look at digit $d_{t+1}$ in $x = .d_1 d_2 \cdots d_t d_{t+1} \cdots \times 10^{\epsilon}$ and then set

$$fl(x) = \begin{cases} .d_1 d_2 \cdots d_t \times 10^{\epsilon} & if \quad d_{t+1} < 5, \\ ([.d_1 d_2 \cdots d_t] + 10^{-t}) \times 10^{\epsilon} & if \quad d_{t+1} \geq 5. \end{cases}$$

- For example, in 2-digit, base-10 floating-point arithmetic, $fl(3/80) = fl(0.0375) = fl(0.375 \times 10^{-1}) = 0.38 \times 10^{-1} = 0.038$.

- Several familiar rules of real arithmetic do not hold for floating-point arithmetic—associativity is one outstanding example.
- By considering $\eta = 1/3$, $\xi = 3$ with $t$-digit base-10 arithmetic,

$$fl(\eta + \xi) \neq fl(\eta) + fl(\xi) \quad and \quad fl(\eta\xi) \neq fl(\eta)fl(\xi).$$

- This makes the analysis of floating-point computation difficult.
- To understand how to execute Gaussian elimination using floating-point arithmetic, let's compare the use of exact arithmetic with the use of 3-digit base-10 arithmetic to solve the following system:

$$47x + 28y = 19,$$

$$89x + 53y = 36.$$

  - Using Gaussian elimination with exact arithmetic, we multiply the first equation by the multiplier m $= 89/47$ and subtract the result from the second equation to produce exact solution $x = 1$ and $y = -1$.
  - Using 3-digit arithmetic, the multiplier is $fl(m) = 1.89$. Apply 3-digit back substitution to obtain the 3-digit floating-point solution $y = 1$ and $x = -0.191$.

- The vast discrepancy between the exact solution (1,.1) and the 3-digit solution (-0.191, 1) illustrates some of the problems we can expect to encounter while trying to solve linear systems with floating-point arithmetic.
- Sometimes using a higher precision may help, but this is not always possible because on all machines there are natural limits that make extended precision arithmetic impractical past a certain point.
- Even if it is possible to increase the precision, it may not buy you very much because there are many cases for which an increase in precision does not produce a comparable decrease in the accumulated roundoff error.
- Although the effects of rounding can almost never be eliminated, there are some simple techniques that can help to minimize these machine induced errors.

## Partial Pivoting

At each step, search the positions on and below the pivotal position for the coefficient of *maximum magnitude*. If necessary perform the appropriate row interchange to bring this maximal coefficient into the pivotal position. Illustrated below is the third step in a typical case:

$$\left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{S} & * & * & * \\ 0 & 0 & S & * & * & * \\ 0 & 0 & S & * & * & * \end{array}\right).$$

Search the positions in the third column marked "$S$" for the coefficient of maximal magnitude and, if necessary, interchange rows to bring this coefficient into the circled pivotal position. Simply stated, the strategy is to maximize the magnitude of the pivot at each step by using only row interchanges.

- On the surface, it is probably not apparent why partial pivoting should make a difference.

The following example not only shows that partial pivoting can indeed make a great deal of difference, but it also indicates what makes this strategy effective.

It is easy to verify that the exact solution to the system

$$\begin{aligned} -10^{-4}x + y &= 1, \\ x + y &= 2, \end{aligned}$$

is given by

$$x = \frac{1}{1.0001} \quad \text{and} \quad y = \frac{1.0002}{1.0001}.$$

If 3-digit arithmetic *without* partial pivoting is used, then the result is

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) R_2 + 10^4 R_1 \longrightarrow \left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 0 & 10^4 & 10^4 \end{array}\right)$$

because

$$fl(1 + 10^4) = fl(.10001 \times 10^5) = .100 \times 10^5 = 10^4$$

and

$$fl(2 + 10^4) = fl(.10002 \times 10^5) = .100 \times 10^5 = 10^4.$$

Back substitution now produces

$$x = 0 \quad \text{and} \quad y = 1.$$

Although the computed solution for $y$ is close to the exact solution for $y$, the computed solution for $x$ is not very close to the exact solution for $x$—the computed solution for $x$ is certainly not accurate to three significant figures as you might hope. If 3-digit arithmetic *with* partial pivoting is used, then the result is

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \longrightarrow \left(\begin{array}{cc|c} 1 & 1 & 2 \\ -10^{-4} & 1 & 1 \end{array}\right) R_2 + 10^{-4}R_1$$
$$\longrightarrow \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array}\right)$$

because

$$fl(1 + 10^{-4}) = fl(.10001 \times 10^1) = .100 \times 10^1 = 1$$

and

$$fl(1 + 2 \times 10^{-4}) = fl(.10002 \times 10^1) = .100 \times 10^1 = 1.$$

This time, back substitution produces the computed solution

$$x = 1 \quad \text{and} \quad y = 1,$$

which is as close to the exact solution as one can reasonably expect—the computed solution agrees with the exact solution to three significant digits.

- Why did partial pivoting make a difference?
- In summary, the large multiplier prevents some smaller numbers from being fully accounted for, thereby resulting in the exact solution of another system that is very different from the original system.
- When partial pivoting is used, no multiplier ever exceeds 1 in magnitude.

$$\left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{p} & * & * & * \\ 0 & 0 & q & * & * & * \\ 0 & 0 & r & * & * & * \end{array}\right) \begin{array}{l} \\ \\ \\ R_4 - (q/p)R_3 \\ R_5 - (r/p)R_3 \end{array} \longrightarrow \left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{p} & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * & * \end{array}\right).$$

  - The pivot is $p$, while $q/p$ and $r/p$ are the multipliers.
  - If partial pivoting has been employed, then $|p| \geq |q|$ and $|p| \geq |r|$ so that

$$\left|\frac{q}{p}\right| \leq 1 \quad and \quad \left|\frac{r}{p}\right| \leq 1.$$

- Other example:

The exact solution to the system

$$\begin{aligned} -10x + 10^5 y &= 10^5, \\ x + \quad y &= 2, \end{aligned}$$

is given by

$$x = \frac{1}{1.0001} \quad \text{and} \quad y = \frac{1.0002}{1.0001}.$$

Suppose that 3-digit arithmetic with partial pivoting is used. Since $|-10| > 1$, no interchange is called for and we obtain

$$\left(\begin{array}{cc|c} -10 & 10^5 & 10^5 \\ 1 & 1 & 2 \end{array}\right) R_2 + 10^{-1} R_1 \longrightarrow \left(\begin{array}{cc|c} -10 & 10^5 & 10^5 \\ 0 & 10^4 & 10^4 \end{array}\right)$$

because

$$fl(1 + 10^4) = fl(.10001 \times 10^5) = .100 \times 10^5 = 10^4$$

and

$$fl(2 + 10^4) = fl(.10002 \times 10^5) = .100 \times 10^5 = 10^4.$$

Back substitution yields

$$x = 0 \quad \text{and} \quad y = 1,$$

which must be considered to be very bad—the computed 3-digit solution for $y$ is not too bad, but the computed 3-digit solution for $x$ is terrible!

# Complete Pivoting

If $[\mathbf{A}|\mathbf{b}]$ is the augmented matrix at the $k^{th}$ step of Gaussian elimination, then search the pivotal position together with every position in $\mathbf{A}$ that is below or to the right of the pivotal position for the coefficient of maximum magnitude. If necessary, perform the appropriate row and column interchanges to bring the coefficient of maximum magnitude into the pivotal position. Shown below is the third step in a typical situation:

$$\left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{\scriptstyle S} & S & S & * \\ 0 & 0 & S & S & S & * \\ 0 & 0 & S & S & S & * \end{array}\right)$$

Search the positions marked "$S$" for the coefficient of maximal magnitude. If necessary, interchange rows and columns to bring this maximal coefficient into the circled pivotal position.

**Problem:** Use 3-digit arithmetic together with complete pivoting to solve the following system:

$$\begin{aligned} x - \quad y &= -2, \\ -9x + 10y &= 12. \end{aligned}$$

**Solution:** Since 10 is the coefficient of maximal magnitude that lies in the search pattern, interchange the first and second rows and then interchange the first and second columns:

$$\left(\begin{array}{cc|c} 1 & -1 & -2 \\ -9 & 10 & 12 \end{array}\right) \longrightarrow \left(\begin{array}{cc|c} -9 & 10 & 12 \\ 1 & -1 & -2 \end{array}\right)$$
$$\longrightarrow \left(\begin{array}{cc|c} 10 & -9 & 12 \\ -1 & 1 & -2 \end{array}\right) \longrightarrow \left(\begin{array}{cc|c} 10 & -9 & 12 \\ 0 & .1 & -.8 \end{array}\right).$$

The effect of the column interchange is to rename the unknowns to $\hat{x}$ and $\hat{y}$, where $\hat{x} = y$ and $\hat{y} = x$. Back substitution yields $\hat{y} = -8$ and $\hat{x} = -6$ so that

$$x = \hat{y} = -8 \quad \text{and} \quad y = \hat{x} = -6.$$

# Ill-Conditioned Systems

- Gaussian elimination with partial pivoting on a system is perhaps the most fundamental algorithm in the practical use of linear algebra.
- However, it is not a universal algorithm nor can it be used blindly.
- There are some systems that are so inordinately sensitive to small perturbations that no numerical technique can be used with confidence.
- Consider the system

$$.835x + .667y = .168,$$

$$.333x + .266y = .067,$$

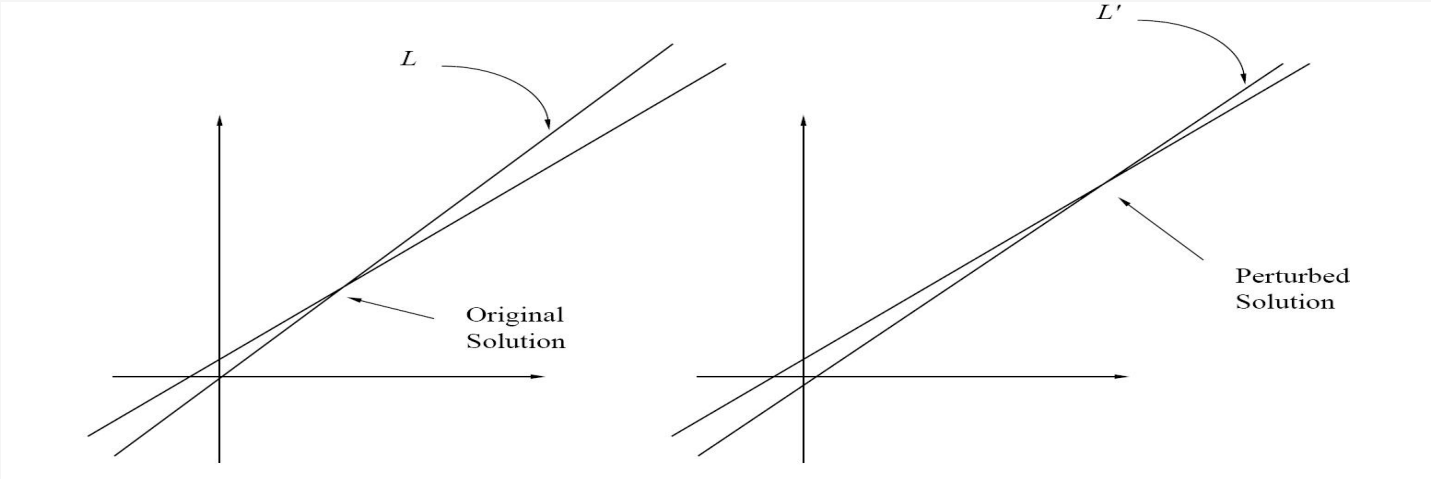
  for which the exact solution is $x = 1$ and $y = -1$.
- If $b_2 = .067$ is only slightly perturbed to become $\hat{b}_2 = .066$, then the exact solution changes dramatically to become $\hat{x} = -666$ and $\hat{y} = 834$.

- This is an example of a system whose solution is extremely sensitive to a small perturbation.
- This sensitivity is intrinsic to the system itself and is not a result of any numerical procedure.
- Therefore, you cannot expect some "numerical trick" to remove the sensitivity.
- If the exact solution is sensitive to small perturbations, then any computed solution cannot be less so, regardless of the algorithm used.

**Ill-Conditioned Linear Systems**

A system of linear equations is said to be ***ill-conditioned*** when some small perturbation in the system can produce relatively large changes in the exact solution. Otherwise, the system is said to be ***well-conditioned.***

- It is easy to visualize what causes a $2 \times 2$ system to be ill- conditioned.
- Geometrically, two equations in two unknowns represent two straight lines, and the point of intersection is the solution for the system.
- An ill-conditioned system represents two straight lines that are almost parallel.
- If two straight lines are almost parallel and if one of the lines is tilted only slightly, then the point of intersection is drastically altered.




- In general, ill-conditioned systems are those that represent almost parallel lines, almost parallel planes, and generalizations of these notions.

- In dealing with an ill-conditioned system, the engineer or scientist is often confronted with a much more basic problem than that of simply trying to solve the system.
- Even if a minor miracle could be performed so that the exact solution could be extracted, the scientist or engineer might still have a nonsensical solution that could lead to totally incorrect conclusions.
- The problem stems from the fact that the coefficients are often empirically obtained and are therefore known only within certain tolerances.
- For an ill-conditioned system, a small uncertainty in any of the coefficients can mean an extremely large uncertainty may exist in the solution.
- Rather than trying to extract accurate solutions from ill-conditioned systems, engineers and scientists are usually better off investing their time and resources in trying to redesign the associated experiments or their data collection methods so as to avoid producing ill-conditioned systems.

- There is one other discomforting aspect of ill-conditioned systems—checking the answer.
- Substituting a computed solution back into the left-hand side of the original system of equations to see how close it comes to satisfying the system.
- Suppose that you compute a solution $x_c$ and substitute it back to find that all the residuals are relatively small. Does this guarantee that $x_c$ is close to the exact solution?
- Surprisingly, the answer is a resounding "no!" whenever the system is ill-conditioned.
- This raises the question, "*How can I check a computed solution for accuracy?*"
- Fortunately, if the system is well-conditioned, then the residuals do indeed provide a more effective measure of accuracy.

- But this means that you must be able to answer some additional questions.
  - For example, how can one tell beforehand if a given system is ill-conditioned?
  - How can one measure the extent of ill-conditioning in a linear system?
- One technique to determine the extent of ill-conditioning might be to experiment by slightly perturbing selected coefficients and observing how the solution changes.
  - If a radical change in the solution is observed for a small perturbation to some set of coefficients, then you have uncovered an ill-conditioned situation.
  - If a given perturbation does not produce a large change in the solution, then nothing can be concluded-perhaps you perturbed the wrong set of coefficients.
- By performing several such experiments using different sets of coefficients, a feel (but not a guarantee) for the extent of ill-conditioning can be obtained.

# Exercises

1. Explain why a linear system can never have exactly two different solutions. Extend your argument to explain the fact that if a system has more than one solution, then it must have infinitely many different solutions.
2. Suppose the matrix $\mathbf{B}$ is obtained by performing a sequence of row operations on matrix $\mathbf{A}$. Explain why $\mathbf{A}$ can be obtained by performing row operations on $\mathbf{B}$.
3. By solving a $3 \times 3$ system by Gaussian elimination, find the coefficients in the equation of the parabola $y = \alpha + \beta x + \gamma x^2$ that passes through the points $(1, 1)$, $(2, 2)$ and $(3, 0)$.
4. Verify that the operation counts given in the text for Gaussian elimination with back substitution are correct for a general $3 \times 3$ system. If you up to the challenge, try to verify these counts for a general $n \times n$ system.

5. Verify that the operation counts given in the text for Gauss-Jordan method are correct for a general $3 \times 3$ system. If you up to the challenge, try to verify these counts for a general $n \times n$ system.
6. Consider the following system:

$$\begin{aligned} 10^{-3}x - y &= 1, \\ x + y &= 0. \end{aligned}$$

   (a) Use 3-digit arithmetic with no pivoting to solve this system.
   (b) Now use partial pivoting and 3-digit arithmetic to solve the original system.