10 NUMERICAL METHODS

Carl Gustav Jacob
Jacobi
1 8 0 4 – 1 8 5 1

$\mathcal{C}$arl Gustav Jacob Jacobi was the second son of a successful banker in Potsdam, Germany. After completing his secondary schooling in Potsdam in 1821, he entered the University of Berlin. In 1825, having been granted a doctorate in mathematics, Jacobi served as a lecturer at the University of Berlin. Then he accepted a position in mathematics at the University of Königsberg.

Jacobi's mathematical writings encompassed a wide variety of topics, including elliptic functions, functions of a complex variable, functional determinants (called Jacobians), differential equations, and Abelian functions. Jacobi was the first to apply elliptic functions to the theory of numbers, and he was able to prove a longstanding conjecture by Fermat that every positive integer can be written as the sum of four perfect squares. (For instance, $10 = 1^2 + 1^2 + 2^2 + 2^2$.) He also contributed to several branches of mathematical physics, including dynamics, celestial mechanics, and fluid dynamics.

In spite of his contributions to applied mathematics, Jacobi did not believe that mathematical research needed to be justified by its applicability. He stated that the sole end of science and mathematics is "the honor of the human mind" and that "a question about numbers is worth as much as a question about the system of the world."

Jacobi was such an incessant worker that in 1842 his health failed and he retired to Berlin. By the time of his death in 1851, he had become one of the most famous mathematicians in Europe.

10.1 GAUSSIAN ELIMINATION WITH PARTIAL PIVOTING

In Chapter 1 two methods for solving a system of n linear equations in n variables were discussed. When either of these methods (Gaussian elimination and Gauss-Jordan elimination) is used with a digital computer, the computer introduces a problem that has not yet been discussed—**rounding error.**

Digital computers store real numbers in **floating point form,**

$$\pm M \times 10^k,$$

where k is an integer and the **mantissa** M satisfies the inequality $0.1 \leq M < 1$. For instance, the floating point forms of some real numbers are as follows.

Real Number	Floating Point Form
527	0.527×10^3
-3.81623	-0.381623×10^1
0.00045	0.45×10^{-3}

The number of decimal places that can be stored in the mantissa depends on the computer. If n places are stored, then it is said that the computer stores **n significant digits.** Additional digits are either truncated or rounded off. When a number is **truncated** to n significant digits, all digits after the first n significant digits are simply omitted. For instance, truncated to two significant digits, the number 0.1251 becomes 0.12.

When a number is **rounded** to n significant digits, the last retained digit is increased by one if the discarded portion is greater than half a digit, and the last retained digit is not changed if the discarded portion is less than half a digit. For instance, rounded to two significant digits, 0.1251 becomes 0.13 and 0.1249 becomes 0.12. For the special case in which the discarded portion is precisely half a digit, round so that the last retained digit is even. So, rounded to two significant digits, 0.125 becomes 0.12 and 0.135 becomes 0.14.

Whenever the computer truncates or rounds, a rounding error that can affect subsequent calculations is introduced. The result after rounding or truncating is called the **stored value.**

EXAMPLE 1   *Finding the Stored Value of Number*

Determine the stored value of each of the following real numbers in a computer that rounds to three significant digits.

(a) 54.7   (b) 0.1134   (c) -8.2256

(d) 0.08335   (e) 0.08345

Solution

Number	Floating Point Form	Stored Value
(a) 54.7	0.547×10^2	0.547×10^2
(b) 0.1134	0.1134×10^0	0.113×10^0
(c) -8.2256	-0.82256×10^1	-0.823×10^1
(d) 0.08335	0.8335×10^{-1}	0.834×10^{-1}
(e) 0.08345	0.8345×10^{-1}	0.834×10^{-1}

Note in parts (d) and (e) that when the discarded portion of a decimal is precisely half a digit, the number is rounded so that the stored value ends in an even digit.

REMARK: Most computers store numbers in binary form (base two) rather than decimal form (base ten). Because rounding occurs in both systems, however, this discussion will be restricted to the more familiar base ten.

Rounding error tends to propagate as the number of arithmetic operations increases. This phenomenon is illustrated in the following example.

EXAMPLE 2   *Propagation of Rounding Error*

Evaluate the determinant of the matrix

$$A = \begin{bmatrix} 0.12 & 0.23 \\ 0.12 & 0.12 \end{bmatrix},$$

when rounding each intermediate calculation to two significant digits. Then find the exact solution and compare the two results.

Solution   Rounding each intermediate calculation to two significant digits, produces the following.

$$|A| = (0.12)(0.12) - (0.12)(0.23)$$

$$= 0.0144 - 0.0276$$

$$\approx 0.014 - 0.028 \qquad \text{**Round to two significant digits**}$$

$$= -0.014$$

However, the exact solution is

$$|A| = 0.0144 - 0.0276$$

$$= -0.0132.$$

So, to two significant digits, the correct solution is -0.013. Note that the rounded solution is not correct to two significant digits, even though each arithmetic operation was performed with two significant digits of accuracy. This is what is meant when it is said that arithmetic operations tend to propagate rounding error.

In Example 2, rounding at the intermediate steps introduced a rounding error of

$$-0.0132 - (-0.014) = 0.0008. \qquad \text{**Rounding error**}$$

Although this error may seem slight, it represents a **percentage error** of

$$\frac{0.0008}{0.0132} \approx 0.061 = 6.1\%. \qquad \text{**Percentage error**}$$

In most practical applications, a percentage error of this magnitude would be intolerable. Keep in mind that this particular percentage error arose with only a few arithmetic steps. When the number of arithmetic steps increases, the likelihood of a large percentage error also increases.

Gaussian Elimination with Partial Pivoting

For large systems of linear equations, Gaussian elimination can involve hundreds of arithmetic computations, each of which can produce rounding error. The following straightforward example illustrates the potential magnitude of the problem.

EXAMPLE 3   *Gaussian Elimination and Rounding Error*

Use Gaussian elimination to solve the following system.

$$
\begin{aligned}
0.143x_1 + 0.357x_2 + 2.01x_3 &= -5.173 \\
-1.31x_1 + 0.911x_2 + 1.99x_3 &= -5.458 \\
11.2x_1 - 4.30x_2 - 0.605x_3 &= 4.415
\end{aligned}
$$

After *each* intermediate calculation, round the result to three significant digits.

Solution   Applying Gaussian elimination to the augmented matrix for this system produces the following.

$$\begin{bmatrix} 0.143 & 0.357 & 2.01 & -5.17 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 11.2 & -4.30 & -0.605 & 4.42 \end{bmatrix}$$

$$\begin{bmatrix} 1.00 & 2.50 & 14.1 & -36.2 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 11.2 & -4.30 & -0.605 & 4.42 \end{bmatrix}$$ ⟵ Dividing the first row by 0.143 produces a new first row.

$$\begin{bmatrix} 1.00 & 2.50 & 14.1 & -36.2 \\ 0.00 & 4.19 & 20.5 & -52.9 \\ 11.2 & -4.30 & -0.605 & 4.42 \end{bmatrix}$$ ⟵ Adding 1.31 times the first row to the second row produces a new second row.

$$\begin{bmatrix} 1.00 & 2.50 & 14.1 & -36.2 \\ 0.00 & 4.19 & 20.5 & -52.9 \\ 0.00 & -32.3 & -159. & 409. \end{bmatrix}$$ ⟵ Adding -11.2 times the first row to the third row produces a new third row.

$$\begin{bmatrix} 1.00 & 2.50 & 14.1 & -36.2 \\ 0.00 & 1.00 & 4.89 & -12.6 \\ 0.00 & -32.3 & -159. & 409. \end{bmatrix}$$ ⟵ Dividing the second row by 4.19 produces a new second row.

$$\begin{bmatrix} 1.00 & 2.50 & 14.1 & -36.2 \\ 0.00 & 1.00 & 4.89 & -12.6 \\ 0.00 & 0.00 & -1.00 & 2.00 \end{bmatrix}$$ ⟵ Adding 32.3 times the second row to the third row produces a new third row.

$$\begin{bmatrix} 1.00 & 2.50 & 14.1 & -36.2 \\ 0.00 & 1.00 & 4.89 & -12.6 \\ 0.00 & 0.00 & 1.00 & -2.00 \end{bmatrix}$$ ⟵ Multiplying the third row by -1 produces a new third row.

So $x_3 = -2.00$, and using back-substitution, you can obtain $x_2 = -2.82$ and $x_1 = -0.950$. Try checking this "solution" in the original system of equations to see that it is not correct. (The correct solution is $x_1 = 1$, $x_2 = 2$, and $x_3 = -3$.)

What went wrong with the Gaussian elimination procedure used in Example 3? Clearly, rounding error propagated to such an extent that the final "solution" became hopelessly inaccurate.

Part of the problem is that the original augmented matrix contains entries that differ in orders of magnitude. For instance, the first column of the matrix

$$\begin{bmatrix} 0.143 & 0.357 & 2.01 & -5.17 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 11.2 & -4.30 & -0.605 & 4.42 \end{bmatrix}$$

has entries that increase roughly by powers of ten as one moves down the column. In subsequent elementary row operations, the first row was multiplied by 1.31 and -11.2 and

the second row was multiplied by 32.3. When floating point arithmetic is used, such large row multipliers tend to propagate rounding error. This type of error propagation can be lessened by appropriate row interchanges that produce smaller multipliers. One method for restricting the size of the multipliers is called **Gaussian elimination with partial pivoting.**

Gaussian Elimination with Partial Pivoting

1. Find the entry in the left column with the largest absolute value. This entry is called the **pivot.**
2. Perform a row interchange, if necessary, so that the pivot is in the first row.
3. Divide the first row by the pivot. (This step is unnecessary if the pivot is 1.)
4. Use elementary row operations to reduce the remaining entries in the first column to zero.

The completion of these four steps is called a **pass.** After performing the first pass, ignore the first row and first column and repeat the four steps on the remaining submatrix. Continue this process until the matrix is in row-echelon form.

Example 4 shows what happens when this partial pivoting technique is used on the system of linear equations given in Example 3.

EXAMPLE 4   *Gaussian Elimination with Partial Pivoting*

Use Gaussian elimination with partial pivoting to solve the system of linear equations given in Example 3. After *each* intermediate calculation, round the result to three significant digits.

Solution   As in Example 3, the augmented matrix for this system is

$$\left[\begin{array}{ccc|c} 0.143 & 0.357 & 2.01 & -5.17 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 11.2 & -4.30 & -0.605 & 4.42 \end{array}\right].$$

$\uparrow$
Pivot

In the left column 11.2 is the pivot because it is the entry that has the largest absolute value. So, interchange the first and third rows and apply elementary row operations as follows.

$$\left[\begin{array}{ccc|c} 11.2 & -4.30 & -0.605 & 4.42 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 0.143 & 0.357 & 2.01 & -5.17 \end{array}\right]$$ ⟵ **Interchange the first and third rows.**
⟵

$$\left[\begin{array}{ccc|c} 1.00 & -0.384 & -0.0540 & 0.395 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 0.143 & 0.357 & 2.01 & -5.17 \end{array}\right]$$ ⟵ **Dividing the first row by 11.2 produces a new first row.**

$$\begin{bmatrix} 1.00 & -0.384 & -0.0540 & 0.395 \\ 0.00 & 0.408 & 1.92 & -4.94 \\ 0.143 & 0.357 & 2.01 & -5.17 \end{bmatrix}$$ ⟵ **Adding 1.31 times the first row to the second row produces a new second row.**

$$\begin{bmatrix} 1.00 & -0.384 & -0.0540 & 0.395 \\ 0.00 & 0.408 & 1.92 & -4.94 \\ 0.00 & 0.412 & 2.02 & -5.23 \end{bmatrix}$$ ⟵ **Adding −0.143 times the first row to the third row produces a new third row.**

This completes the first pass. For the second pass consider the submatrix formed by deleting the first row and first column. In this matrix the pivot is 0.412, which means that the second and third rows should be interchanged. Then proceed with Gaussian elimination as follows.

Pivot

$$\begin{bmatrix} 1.00 & -0.384 & -0.0540 & 0.395 \\ 0.00 & 0.412 & 2.02 & -5.23 \\ 0.00 & 0.408 & 1.92 & -4.94 \end{bmatrix}$$ ⟵ **Interchange the second and third rows.**

$$\begin{bmatrix} 1.00 & -0.384 & -0.0540 & 0.395 \\ 0.00 & 1.00 & 4.90 & -12.7 \\ 0.00 & 0.408 & 1.92 & -4.94 \end{bmatrix}$$ ⟵ **Dividing the second row by 0.412 produces a new second row.**

$$\begin{bmatrix} 1.00 & -0.384 & -0.0540 & 0.395 \\ 0.00 & 1.00 & 4.90 & -12.7 \\ 0.00 & 0.00 & -0.0800 & 0.240 \end{bmatrix}$$ ⟵ **Adding −0.408 times the second row to the third row produces a new third row.**

This completes the second pass, and you can complete the entire procedure by dividing the third row by −0.0800 as follows.

$$\begin{bmatrix} 1.00 & -0.384 & -0.0540 & 0.395 \\ 0.00 & 1.00 & 4.90 & -12.7 \\ 0.00 & 0.00 & 1.00 & -3.00 \end{bmatrix}$$ ⟵ **Dividing the third row by −0.0800 produces a new third row.**

So $x_3 = -3.00$, and back-substitution produces $x_2 = 2.00$ and $x_1 = 1.00$, which agrees with the exact solution of $x_1 = 1$, $x_2 = 2$, and $x_3 = -3$ when rounded to three significant digits.

REMARK: Note that the row multipliers used in Example 4 are 1.31, −0.143, and −0.408, as contrasted with the multipliers of 1.31, 11.2, and 32.3 encountered in Example 3.

The term *partial* in partial pivoting refers to the fact that in each pivot search only entries in the left column of the matrix or submatrix are considered. This search can be extended to include every entry in the coefficient matrix or submatrix; the resulting technique is called **Gaussian elimination with complete pivoting.** Unfortunately, neither complete pivoting nor partial pivoting solves all problems of rounding error. Some systems of linear

equations, called **ill-conditioned** systems, are extremely sensitive to numerical errors. For such systems, pivoting is not much help. A common type of system of linear equations that tends to be ill-conditioned is one for which the determinant of the coefficient matrix is nearly zero. The next example illustrates this problem.

EXAMPLE 5   *An Ill-Conditioned System of Linear Equations*

Use Gaussian elimination to solve the following system of linear equations.

$$x + \quad y = \quad 0$$

$$x + \frac{401}{400}y = 20$$

Round each intermediate calculation to four significant digits.

Solution   Using Gaussian elimination with rational arithmetic, you can find the exact solution to be $y = 8000$ and $x = -8000$. But rounding $401/400 = 1.0025$ to four significant digits introduces a large rounding error, as follows.

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1.002 & 20 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0.002 & 20 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1.00 & 10{,}000 \end{bmatrix}$$

So $y = 10{,}000$ and back-substitution produces

$$x = -y$$

$$= -10{,}000.$$

This "solution" represents a percentage error of 25% for both the x-value and the y-value. Note that this error was caused by a rounding error of only 0.0005 (when you rounded 1.0025 to 1.002).

SECTION 10.1 ☐ EXERCISES

In Exercises 1–8, express the real number in floating point form.

1. 4281   **2.** 321.61   **3.** -2.62 **4.** -21.001
5. -0.00121 **6.** 0.00026   **7.** $\frac{1}{8}$ **8.** $16\frac{1}{2}$

In Exercises 9–16, determine the stored value of the real number in a computer that rounds to (a) three significant digits and (b) four significant digits.

9. 331   **10.** 21.4   **11.** -92.646 **12.** 216.964
13. $\frac{7}{16}$ **14.** $\frac{5}{32}$ **15.** $\frac{1}{7}$ **16.** $\frac{1}{6}$

In Exercises 17 and 18, evaluate the determinant of the matrix, rounding each intermediate calculation to three significant digits. Then compare the rounded value with the exact solution.

17. $\begin{bmatrix} 1.24 & 56.00 \\ 66.00 & 1.02 \end{bmatrix}$ **18.** $\begin{bmatrix} 2.12 & 4.22 \\ 1.07 & 2.12 \end{bmatrix}$

In Exercises 19 and 20, use Gaussian elimination to solve the system of linear equations. After each intermediate calculation, round the result to three significant digits. Then compare this solution with the exact solution.

19. $1.21x + 16.7y = 28.8$ **20.** $14.4x - 17.1y = 31.5$
$\quad\ 4.66x + 64.4y = 111.0$ $\quad\ 81.6x - 97.4y = 179.0$

⊞ In Exercises 21–24, use Gaussian elimination without partial pivoting to solve the system of linear equations, rounding to three significant digits after each intermediate calculation. Then use partial pivoting to solve the same system, again rounding to three significant digits after each intermediate calculation. Finally, compare both solutions with the given exact solution.

21. $\quad\ x + 1.04y = 2.04$ **22.** $\quad 0.51x + 92.6y = 97.7$
$\quad\ 6x + 6.20y = 12.20$ $\quad\ 99.00x - 449.0y = 541.0$
$\quad$ (Exact: $x = 1, y = 1$) $\quad$ (Exact: $x = 10, y = 1$)

23. $\quad\ x + 4.01y + 0.00445z = 0.00$
$\quad -x - 4.00y + 0.00600z = 0.21$
$\quad\ 2x - 4.05y + 0.05000z = -0.385$
$\quad$ (Exact: $x = -0.49, y = 0.1, z = 20$)

24. $\quad 0.007x + 61.20y + 0.093z = 61.3$
$\quad\ 4.810x - 5.92y + 1.110z = 0.0$
$\quad 81.400x + 1.12y + 1.180z = 83.7$
$\quad$ (Exact: $x = 1, y = 1, z = 1$)

In Exercises 25 and 26, use Gaussian elimination to solve the ill-conditioned system of linear equations, rounding each intermediate calculation to three significant digits. Then compare this solution with the given exact solution.

25. $x + \quad y = 2$ **26.** $\quad x - \frac{800}{801}y = 10$
$\quad x + \frac{600}{601}y = 20$ $\quad -x + \quad y = 50$
$\quad$ (Exact: $x = 10,820,$ $\quad$ (Exact: $x = 48,010,$
$\quad\ y = -10,818$) $\quad\ y = 48,060$)

27. Consider the ill-conditioned systems

$\quad x + \quad\quad y = 2$ and $x + \quad\quad y = 2$
$\quad x + 1.0001y = 2$ $x + 1.0001y = 2.0001.$

Calculate the solution to each system. Notice that although the systems are almost the same, their solutions differ greatly.

28. Repeat Exercise 27 for the systems

$\quad\quad x - y = 0$ and $\quad\quad x - y = 0$
$\quad -1.001x + y = 0.001$ $\quad -1.001x + y = 0.$

29. The **Hilbert matrix** of size $n \times n$ is the $n \times n$ symmetric matrix $H_n = [a_{ij}]$, where $a_{ij} = 1/(i + j - 1)$. As n increases, the Hilbert matrix becomes more and more ill-conditioned. Use Gaussian elimination to solve the following system of linear equations, rounding to two significant digits after each intermediate calculation. Compare this solution with the exact solution ($x_1 = 3, x_2 = -24,$ and $x_3 = 30$).

$\quad x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = 1$
$\quad \frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 = 1$
$\quad \frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = 1$

30. Repeat Exercise 29 for $H_4\mathbf{x} = \mathbf{b}$, where $\mathbf{b} = (1, 1, 1, 1)^T$, rounding to four significant digits. Compare this solution with the exact solution ($x_1 = -4, x_2 = 60, x_3 = -180,$ and $x_4 = 140$).

⊞ **31.** The inverse of the $n \times n$ Hilbert matrix H_n has integer entries. Use your computer of graphing calculator to calculate the inverses of the Hilbert matrices H_n for $n = 4, 5, 6,$ and 7. For what values of n do the inverses appear to be accurate?