

to look at digit d_{t+1} in $x = .d_1d_2 \cdots d_t d_{t+1} \cdots \times 10^e$ (making sure $d_1 \neq 0$) and then set

$$fl(x) = \begin{cases} .d_1d_2 \cdots d_t \times 10^e & \text{if } d_{t+1} < 5, \\ ([.d_1d_2 \cdots d_t] + 10^{-t}) \times 10^e & \text{if } d_{t+1} \geq 5. \end{cases}$$

For example, in 2-digit, base-10 floating-point arithmetic,

$$fl(3/80) = fl(.0375) = fl(.375 \times 10^{-1}) = .38 \times 10^{-1} = .038.$$

By considering $\eta = 21/2$ and $\xi = 11/2$ with 2-digit base-10 arithmetic, it's also easy to see that

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

Furthermore, several familiar rules of real arithmetic do not hold for floating-point arithmetic—associativity is one outstanding example. This, among other reasons, makes the analysis of floating-point computation difficult. It also means that you must be careful when working the examples and exercises in this text because although most calculators and computers can be instructed to display varying numbers of digits, most have a fixed internal precision with which all calculations are made before numbers are displayed, and this internal precision cannot be altered. The internal precision of your calculator is greater than the precision called for by the examples and exercises in this book, so each time you make a t -digit calculation with a calculator you should manually round the result to t significant digits and then manually reenter the rounded number in your calculator before proceeding to the next calculation. In other words, *don't "chain" operations in your calculator or computer.*

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:

$$\begin{aligned} 47x + 28y &= 19, \\ 89x + 53y &= 36. \end{aligned}$$

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

$$\left(\begin{array}{cc|c} 47 & 28 & 19 \\ 0 & -1/47 & 1/47 \end{array} \right).$$

Back substitution yields the *exact solution*

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

Using 3-digit arithmetic, the multiplier is

$$fl(m) = fl\left(\frac{89}{47}\right) = .189 \times 10^1 = 1.89.$$