1.3 Floating Point Form

Floating point numbers are used by computers to approximate real numbers. On the surface, the question is a simple one. There are an infinite number of real numbers, but a computer is a finite machine so it can only represent a finite number of real numbers. That is, not all real numbers can be stored exactly. Therein lies the problem.

If the computer can only store an approximation of a real number, then it is essential that there is a discussion of the error involved.

In this section we will address each of these issues.

Floating Point Numbers

Each of the following numbers is equal to 123.4567 in base ten:

$$12345.67 \cdot 10^{-2}, \qquad 1.234567 \cdot 10^{2}, \qquad \text{and} \qquad 0.01234567 \cdot 10^{4}.$$

- In the first case, multiplying by 10^{-2} moves the decimal point two places to the left, so $12345.67 \cdot 10^{-2} = 123.4567$.
- In the second case, multplying by 10^{2} moves the decimal point two places to the right, so $1.234567 \cdot 10^{2} = 123.4567$.
- In the third case, multiplying by 10^{4} moves the decimal point four places to the right, so $0.01234567 \cdot 10^{4} = 123.4567$.

Computers use a form of scientific notation to store approximations of real numbers in n-digit floating point form.

n-digit Floating Point Form. An n-digit floating point has the form

$$\pm d_1.d_2d_3 \ldots d_n \cdot b^m.$$

Note that the sign occurs first (plus or minus), followed by an n-digit number $d_1.d_2d_3 \ldots d_n$ called the **matissa**. The number b is called the **base** and m is called its **exponent**. Each digit d_i of the mantissa is an integer such that $0 \leq d_i < b$, for $i = 2, 3, \ldots, n$. The first digit must satisfy $0 < d_1 < b$ unless the floating point number is zero.

For example, in base ten, the number $2.3854 \cdot 10^{-13}$ is in 5-digit floating point format, but the numbers $238.54 \cdot 10^{-12}$ and $0.0023854 \cdot 10^{12}$ are not.

- In the first case, $238.54 \cdot 10^{-12}$ has more than one digit to the left of the decimal point. We can place this number in 5-digit floating point form by repositioning the decimal point and adjusting the exponent. That is,

$$238.54 \cdot 10^{-12} = 2.3854 \cdot 10^{-10}.$$

- In the second case, the first digit to the left of the decimal point in the number $0.0023854 \cdot 10^{12}$ is zero, but $0.0023854 \cdot 10^{12}$ is not zero. Again, we can place this number in floating point form by repositioning the decimal point and adjusting the exponent. That is,

$$0.0023854 \cdot 10^{12} = 2.3854 \cdot 10^{9}.$$

In the examples that follow, let's assume that we are working on a base ten machine that stores numbers in 5-digit floating point format.

▶ **Example 1.** *Change the number* 888.341983765 *into 5-digit base ten floating point format.*

First, reposition the decimal point so that there is exactly one nonzero digit to the left of the decimal point.

$$888.341983765 = 8.88341983765 \cdot 10^{2}$$

The machine we are working on can only handle 5-digit floating point form. It can't store all the digits of the mantissa above. Therefore, we must determine the closest 5-digit floating point number available and use that as an approximation for our number. Note that our number lies between the two 5-digit floating point numbers

$$8.8834 \cdot 10^{2} < 8.88341983765 \cdot 10^{2} < 8.8835 \cdot 10^{2},$$

but it is closer to $8.8834 \cdot 10^{2}$. Hence, in 5-digit floating point form,

$$888.341983765 \approx 8.8834 \cdot 10^{2}.$$

Note that we *rounded* towards zero in this example. Because the digit following the 4 in $8.88341983765 \cdot 10^{2}$ is a 1, which is less than 5, we truncate the number at $8.8834 \cdot 10^{2}$.

Let's look at another example.

▶ **Example 2.** *Change the number* 0.00075493671278 *into 5-digit base ten floating point form.*

First, reposition the decimal point so that there is exactly one nonzero digit to the left of the decimal point.

$$0.00075493671278 = 7.5493671278 \cdot 10^{-4}$$

Again, our machine can only handle 5-digit mantissas. Our number lies between the following two 5-digit floating point numbers

$$7.5493 \cdot 10^{-4} < 7.5493671278 \cdot 10^{-4} < 7.5494 \cdot 10^{-4},$$

but it is closer to the number $7.5494 \cdot 10^{-4}$. Hence, in 5-digit floating point form,

$$0.00075493671278 = 7.5494 \cdot 10^{-4}.$$

Note that we *rounded away* from zero in this example. Because the digit following the 3 in $7.5493671278 \cdot 10^{-4}$ is a 6, which is 5 or greater, we add 1 to the previous place before truncating to get $7.5494 \cdot 10^{-4}$.

Let's look at another example.

▶ **Example 3.** *Suppose that the 5-digit base ten floating point representation of a real number x is $x^* = 2.3086 \cdot 10^{-4}$. Find the range of possible values for the real number x.*

In **Examples 1** and **2**, we saw that the computer will sometime rounds towards zero and other times round away from zero, depending on the value of the sixth digit in 5-digit floating point format. In this example, we're given the 5-digit base ten floating point form of the number, namely

$$x^* = 2.3086 \cdot 10^{-4}.$$

- The very smallest that x could be is $x = 2.30855 \cdot 10^{-4}$. Any smaller, such as $x = 2.30854999\ldots \cdot 10^{-4}$, and x would have been rounded towards zero to $x^* = 2.3085 \cdot 10^{-4}$.

- The very largest that x could be is $x = 2.30864999\ldots \cdot 10^{-4}$. Any larger, such as $x = 2.30865 \cdot 10^{-4}$, and x would have been rounded away from zero to $x^* = 2.3087 \cdot 10^{-4}$.

  Therefore, x could be any number in the range

$$2.30855 \cdot 10^{-4} < x < 2.30864999\ldots \cdot 10^{-4}.$$

Binary Floating Point Form

In binary (base two), things work pretty much the same. Note that the number $1.0011 \cdot 2^{-3}$ is in 5-digit base two floating point form, but the numbers $1101.1 \cdot 2^{-4}$ and $0.00011001 \cdot 2^5$ are not.

- In the first case, $1101.1 \cdot 2^{-4}$ has more than one digit to the left of the decimal point. We can place this number in floating point form by repositioning the decimal point and adjusting the exponent.

$$1101.1 \cdot 2^{-4} = 1.1011 \cdot 2^{-1}$$

- In the second case, the first digit to the left of the decimal point in the number $0.00011001 \cdot 2^5$ is not zero, but we can again reposition the decimal point and adjust the exponent.

$$0.00011001 \cdot 2^5 = 1.1001 \cdot 2^1$$

Error

In this section we discuss the error made when storing a real number in n-digit floating point form on a computer.

We will discuss two important types of error: (1) *absolute* error, and (2) *relative* error.

In the discussion that follows, we will let x represent the real number and x^* represent the n-digit floating point approximation of x.

Absolute and Relative Error. Let x^* be the n-digit floating point representation of the real number x. Then the absolute and relative error in approximating x with x^* is given by the formulae

$$\text{Absolute Error} = |x^* - x|$$

and

$$\text{Relative Error} = \frac{|x^* - x|}{|x|}.$$

Let's look at an example.

▶ **Example 4.** *Calculate both the absolute and relative error when the real number $x = 938\,756$ is stored in 3-digit base ten floating point form.*

First, reposition the decimal point so that there is one nonzero digit to the left of the decimal point.

$$x = 938\,756 = 9.38756 \cdot 10^5$$

We can only use 3 digits in the mantissa. The next digit to the right of 8 is a 7, which is greater than 5, so we round up (away from zero) to

$$9.38756 \cdot 10^5 = 9.39 \cdot 10^5.$$

The result $x^* = 9.39 \cdot 10^5$ is in 3-digit base ten floating point form. We calculate the absolute error with the following computation.

$$|x^* - x| = \left|9.39 \cdot 10^5 - 938\,756\right| = 244$$

That seems to be an very large error! But on second glance, note what the relative error reveals.

$$\frac{|x^* - x|}{|x|} = \frac{\left|9.39 \cdot 10^5 - 938\,756\right|}{|938\,756|} \approx 2.6 \cdot 10^{-4}$$

A calculator was used to determine the approximation. Note that the number $x = 9.38756 \cdot 10^5$ and its approximation $x^* = 9.39 \cdot 10^5$ agree in about 3 places and the exponent in the relative error $2.6 \cdot 10^{-4}$ is -4.

We'll see that the relative error is more useful. Let's look at another example.

▶ **Example 5.** *Calculate both the absolute and relative error when the real number 0.000005823417658 is stored in 5-digit base ten floating point form.*

Reposition the decimal point so that there is one nonzero digit to the left of the decimal point.

$$x = 0.000005823417658 = 5.823417658 \cdot 10^{-6}$$

The mantissa is allowed 5 digits. Note that the next digit after the 4 is a 1, which is less than 5, so we round down (towards zero) by truncating.

$$5.823417658 \cdot 10^{-6} = 5.8234 \cdot 10^{-6}$$

The result $x^* = 5.8234 \cdot 10^{-6}$ is in 5-digit base ten floating point form. The absolute error is

$$|x^* - x| = \left|5.8234 \cdot 10^{-6} - 0.000005823417658\right| = 1.7658 \cdot 10^{-11},$$

which at first glance, appears very small indeed. But again, how small is the error relative to the numbers involved? The relative error reveals the answer.

$$\frac{|x^* - x|}{|x|} = \frac{\left|5.8234 \cdot 10^{-6} - 0.000005823417658\right|}{|0.000005823417658|} \approx 3.0 \cdot 10^{-6} \tag{1.1}$$

A calculator was used to find an approximation for the relative error. Note that this error is much larger than the absolute error.

Also, note that $x = 5.823417658 \cdot 10^{-6}$ and $x^* = 5.8234 \cdot 10^{-6}$ agree in approximately 5 digits and the exponent on the relative error $3.0 \cdot 10^{-6}$ is -6.

Indeed, there is a technical definition for the number of *significant digits*.

Significant Digits. The number x^* is said to approximate x to n significant digits if n is the largest nonnegative integer for which

$$\frac{|x^* - x|}{|x|} < 5 \cdot 10^{-n}.$$

Thus, for example, in **Example 4**, we approximated $x = 938\,756$ with $x^* = 9.39 \cdot 10^5$ and found that the relative error was

$$\frac{|x^* - x|}{|x|} \approx 2.6 \cdot 10^{-4},$$

so the relative error is less than $5 \cdot 10^{-4}$. Thus, by the definition, we say that $x^* = 9.39 \cdot 10^5$ approximates $x = 938\,756$ to 4 significant digits. However, note that only the first two leading digits are the same.

In **Example 5**, we approximated $x = 5.823417658 \cdot 10^{-6}$ with $x^* = 5.8234 \cdot 10^{-6}$ and found that the relative error was

$$\frac{|x^* - x|}{|x|} \approx 3.0 \cdot 10^{-6},$$

so the relative error is less than $5 \cdot 10^{-6}$. Thus, by the definition, we say that $x^* = 5.8234 \cdot 10^{-6}$ approximates $x = 5.823417658 \cdot 10^{-6}$ to 6 significant digits. Note, however, that only the first 5 leading digits are the same.

It is important to realize that the notion of significant digits and the the number of digits of agreement between a number and its floating point form are related, but not exactly the same. For example, in 5-digit floating point form, approximating $x = 7.899966666 \cdot 10^3$ with with its 5-digit floating pont form $x^* = 7.9000 \cdot 10^3$ provides a relative error

$$\frac{|x^* - x|}{|x|} \approx 4.2 \cdot 10^{-6},$$

which is less than $5 \cdot 10^{-6}$. Thus, x^* approximates x to 6 significant digits. However, the numbers $x = 7.89996666 \cdot 10^3$ and $x^* = 7.9000 \cdot 10^3$ have only the first leading digit in common. Still, in the sense of the relative error, it's not difficult to imagine the closeness of the digits in $x^* = 7.9000 \cdot 10^{-6}$ to the first 6 digits of $7.89996666 \cdot 10^{-6}$.

n-Digit Floating Point Form and Significant Digits. What is most important to understand is the fact that there is a definite relationship between the the number of digits used to store the mantissa, the relative error, and the number of significant digits.

Propogation of Error

Whenever we store an n-digit floating point form of a real number, we are making an error. This error has a special name.

Roundoff Error. The error incurred when we store a real number in n-digit flooting point form is called **roundoff error**.

With today's modern computers, we can store numbers so that the initial roundoff error is fairly insignificant. The difficulty lies in the fact that computers can literally do billions of computations very quickly, so it is not uncommon to see the original roundoff error propogate through a series of calculations, diverging quickly so as to make the final outcome meaningless.

In the next section we will study some ways to keep this propogation of error under control.

1.3 Exercises

In **Exercises 1-8**, place the given number in 4-digit base ten floating point form. In each case, calculate the absolute and relative error made.

1. 1 885 934

2. 12 345 612

3. 0.0001234567

4. 0.0085188342

5. 888.456123

6. 1 765.33458

7. 0.0002312316

8. 0.00000556781245

In **Exercises 9-12**, a 4-digit base ten floating point approximation x^* of a real number x is given. Determine a range of possible values for x.

9. $2.446 \cdot 10^{-12}$

10. $4.453 \cdot 10^{8}$

11. $5.684 \cdot 10^{5}$

12. $1.104 \cdot 10^{-6}$

In **Exercises 13-16**, Place the given number into n digit floating point format for the given value of n, calculate the relativer error, then use the result to determine the the number of significant digits in the approximation.

13. 1 789.23456, $n = 5$

14. $0.008456174 \cdot 10^{-6}$, $n = 3$

15. $0.0000456712345 \cdot 10^{-11}$, $n = 6$

16. $18.9123456 \cdot 10^{6}$, $n = 4$

1.3 Answers

1. $1.886 \cdot 10^{6}$

3. $1.235 \cdot 10^{-4}$

5. $8.885 \cdot 10^{2}$

7. $2.312 \cdot 10^{-4}$

9. Range from $2.4455 \cdot 10^{-12}$ to $2.4464999\ldots \cdot 10^{-12}$.

11. Range from $5.6835 \cdot 10^{5}$ to $5.6844999\ldots \cdot 10^{5}$.

13. $1.7892 \cdot 10^{3}$, relative error is approximately $1.9 \cdot 10^{-5}$, 5 significant digits.

15. $4.56712 \cdot 10^{-16}$, relative error approximately $7.6 \cdot 10^{-7}$, 6 significant digits.