The Fixed-Point Arithmetic and Its Limitations

Kedar Mhaswade

04 November 2024

Abstract

This paper discusses personal exploration of the so-called Fixed-Point Arithmetic performed by a computer. The intended audience is an interested high school student.

1 Introduction

Computers have limited resources. We can always find a number that exceeds a computer's "memory capacity." By "computer", we do not always mean a computer based on binary digits, or bits, 0 and 1, although such computers are ubiquitous now. In a binary computer, everything must be represented in 0s and 1s. With some flight of imagination, however, a "decimal computer" in which everything needs to be represented using the ten decimal symbols (digits) can be conceived. The computational part of the human mind can be said to be such a decimal computer.

For this article, we can assume the existence of a "decimal computer." It does not matter in the context of this article if the computer is decimal, binary, or something else. These computers differ only in the number of *distinct pieces of information* a digit represents. For the binary computer it's two, for a decimal computer it's ten. The issues discussed here apply to every such computer.

The finiteness of a computer comes from the fact that data must be represented using a finite number of digits. We have learned in mathematics that starting with a number 0 and a *successor* function: $succ(n) = n + 1$, we can imagine infinitely many integers. In other words, there is no largest integer in mathematics. However, for even the most powerful computer, there *is* a largest integer.

The number of digits available to represent integers (using the familiar place-value system) on a computer decides the largest integer it represents. Thus, a decimal computer with 20 digits for integer representation can represent all the integers in the interval $[0, 10^{20} - 1] = [0, 99999999999999999999]$. That is a wide range, however, "wide" is a relative term. In this sense, a computer is better than a desk calculator only in terms of ranges of numbers represented.

It appears that this is an acceptable practical limit for calculations involving only integers. Once we enter the realm of decimals, however, the complexity

increases. It turns out that we had to go through a long struggle before we could reliably[1] represent on *every* computer a decimal like 0.1 as easily as writing it by hand.

(How will you represent negative integers in our decimal computer?)

2 Representing Decimals

A fraction, or rational number, can be represented as a ratio of two integers. Therefore, we can consider representing any fraction as a ratio of two integers: numerator and denominator. Why do we need "decimals[2]" on computers?

We need computers to have the ability to specify decimals because we are used to them since grade school. It feels like a severe limitation if the computer were to take away the convenience of expressing *one-half* simply as 0.5 by forcing us to specify it as numbers 1 and 2.

It is certainly conceivable that the computer lets us specify a fraction as a decimal but *transparently* treats it as an ordered pair of integers that best represents that fraction. Thus, it could treat 4.6 as $(46, 10)$ and 0.217 as $(217, 1000)$. It would then add the fractions using the familiar algorithm:

$$\frac{46}{10} + \frac{217}{1000} = \frac{100 \times 46 + 1 \times 217}{1000} = \frac{4817}{1000} \text{ represented as } (4817, 1000)$$

and display the result as 4.817.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

to determine the distance between any two points (x_1, y_1) and (x_2, y_2) in $\mathbb{R}^2$. For our example, $(x_1, y_1) = (-1, 16)$ and $(x_2, y_2) = (3, 1)$, so plugging these values into the distance formula (1) tell us the distance between the two points is

$$d = \sqrt{(3 - (-1))^2 + (1 - 16)^2} = \sqrt{4^2 + (-15)^2} = \sqrt{241}.$$

3 Linear Fit

Consider a linear equation $y = mx + b$ through the two points. We will first determine the slope m of the line in Section 3.1, and we will then determine the y-intercept b of the line in Section 3.2.

[1] Such reliability implies 'portability' of a computer program across computers.

[2] In mathematics, we often use the same name to express disparate ideas. A 'decimal' is another name for a fraction. A 'decimal representation' uses ten symbols ($\{0, 1, 2, \ldots, 9\}$) to represent a number.

3.1 Slope

The slope of the line passing through the two points is given by the forumula

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}.$$

Plugging in our two points, we find the slope of the line between them is

$$m = \frac{1 - 16}{3 - (-1)} = -\frac{15}{4}. \tag{2}$$

3.2 Intercept

To find the y-intercept of the line, we start with the point-slope form of the line of slope m through the point (x_0, y_0):

$$y - y_0 = m(x - x_0).$$

We plug in the point $(x_0, y_0) = (-1, 16)$ and the slope we found previously (2) to obtain the equation

$$y - 16 = -\frac{15}{4}(x + 1).$$

Solving for y, we find the slope-intercept form of the line:

$$\begin{aligned} y &= -\frac{15}{4}x - \frac{15}{4} + 16 \\ &= -\frac{15}{4}x + \frac{49}{4}. \end{aligned}$$

Therefore, the y-intercept is $b = 49/4$, and the equation $y = -\frac{15}{4}x + \frac{49}{4}$ describes the line through the two points.

4 Exponential Fit

Let us consider the exponential function $y = Ae^{kx}$. For this function to pass through both points, we must find constants A and k that satisfy both equations $16 = Ae^{-k}$ and $1 = Ae^{3k}$. To solve these two simultaneous equations, we first take the ratio of the two equations, which gives us a single equation involving only k:

$$16 = \frac{Ae^{-k}}{Ae^{3k}} = e^{-4k}.$$

We can take the natural logarithm of this equation to solve for k:

$$-4k = \ln(16) = 4\ln(2),$$

which means $k = -\ln(2)$.

We can then use this value of k, along with either of the two points to solve for A. Let us consider the point $(-1, 16)$:

$$16 = Ae^{(-\ln(2))(-1)} = Ae^{\ln 2} = 2A.$$

Solving for A, we find $A = 8$, and the exponential equation through both points is

$$y = 8e^{-\ln(2)x} = 82^{-x} = 8\left(\frac{1}{2}\right)^{x}.$$

Here are examples of piecewise functions:

$$\chi_{\mathbb{Q}}(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases} \tag{3}$$

$$C_k = \begin{cases} 1 & \text{if } k = 1 \\ 1 & \text{if } k = 2 \\ C_{k-1} + C_{k-2} & \text{otherwise} \end{cases} \tag{4}$$