Lesson 5: Fixed-Point Numbers

Recall that our favorite number systems have a very special inclusion chain:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

Each of these are strict subset relations, denoted by the symbol $\subset$. So far, we've only studied how to use MATLAB's native data classes to encode both natural numbers and integers using $m-$bit binary words. We've focused on the case that $m = 8, 16, 32$, and 64 since those word sizes come standard in MATLAB.

More specifically, Lesson 3 focused on how we can encode nonnegative integers using MATLAB's `uint` data classes while Lesson 4 was an introduction to encoding signed integers via MATLAB's `int` data classes. Both the encoding schemes used to represent these numbers are examples of a larger class of data representations. We will refer to this more general encoding scheme known as *fixed-point* number systems in which we always have a constant number of binary digits to the left and right of the radix (binary or decimal) point.

Fixed radix point in decimal and binary integers

Suppose that $x \in \mathbb{Z}$ is a nonnegative (unsigned) integer with $x \geq 0$.Recall from our work in Lesson 3 that when we write x using a decimal representation, we produce a string of $(n+1)-$ decimal digits in the form

$$x = (\, d_n d_{n-1} ... d_2 d_1 d_0 \,)_{10}$$

where $n \geq 0$ and $d_i \in \{0, 1, 2, ..., 9\}$ for all values $i \in \{0, 1, 2, ..., n\}$.

EXAMPLE 5.1

Let's recall how we interpreted the number $x = (215)_{10}$ using a decimal representation. This positive integer is written with $n = 3$ decimal digits. If we set

$$d_0 = 5, \qquad d_1 = 1, \qquad d_2 = 2,$$

then we can rewrite our integer in the following form

$$\begin{aligned} x &= d_2 d_1 d_0, \\ &= d_2 \cdot 10^2 + d_1 \cdot 10^1 + d_0 \cdot 10^0, \\ &= \sum_{i=0}^{n} d_i \cdot 10^i. \end{aligned}$$

This is a decimal representation, since each position d_i is scaled by the power 10^i.

Let's also recall from Lesson 3 that for nonnegative $x \in \mathbb{Z}$ with $x \geq 0$, we can write x using a binary representation to produce a string of $(m+1)-$ binary digits in the form

$$x = (\, b_m b_{m-1} ... b_2 b_1 b_0 \,)_2$$

where $m \geq 0$ and $b_i \in \{0, 1\}$ for all values $i \in \{0, 1, 2, ..., m\}$.

EXAMPLE 5.2

Now, let's revisit our study of binary representations of unsigned integers by considering the number $x = (1101\,0111)_2$. Let's interpret this number using a binary representation. We notice that this positive integer is written with $m = 8$ binary digits. If we set

$$b_0 = 1, \quad b_1 = 1, \quad b_2 = 1, \quad b_3 = 0, \quad b_4 = 1, \quad b_5 = 0, \quad b_6 = 1, \quad b_7 = 1,$$

then we can rewrite our integer in the following form

$$\begin{aligned} x &= b_7 b_6 b_5 b_4 b_3 b_2 b_1 b_0, \\ &= b_7 \cdot 2^7 + b_6 \cdot 2^6 + b_5 \cdot 2^5 + b_4 \cdot 2^4 + b_3 \cdot 2^3 + b_2 \cdot 2^2 + b_1 \cdot 2^1 + b_0 \cdot 2^0, \\ &= \sum_{i=0}^{m} b_i \cdot 2^i. \end{aligned}$$

This is a binary representation since each position b_i is scaled by the power 2^i.

In both Examples 5.1 and 5.2, we notice that the location of the radix point was implied but not explicitly written. Specifically, when we write $x = 215$, we actually mean that the radix point (also known as the decimal point for our decimal representation) is supposed to be immediately to the right of the least significant decimal digit. To be more accurate, we should write

$$x = 215.0$$

where the radix indicates the exact position of the decimal point. Note, we write the radix and zero in gray to represent a digit that is implied but not actually written. Similarly, we can include our radix point (i.e. binary point) to the right of our least significant bit when we write

$$x = 1101\,0111 = 11010111.0$$

In both cases, the radix point is in a fixed position. In fact, every time we use our standard decimal or binary representations

$$x = d_n d_{n-1} ... d_2 d_1 d_0 = b_m b_{m-1} ... b_2 b_1 b_0$$

we have implicitly fixed the radix immediately to the right of the least significant digit. This is our first example of a *fixed-point representation*, since the location of the radix is fixed. Let's extend our view of fixed-point representations and study how we can represent finite fractions using fixed-point number systems.

Decimal fractions

While we have spent a lot of time analyzing decimal and binary encoding of $x \in \mathbb{Z}$, we have not yet considered how to using binary and decimal number systems to represent $x \in \mathbb{Q}$. In other words, we haven't yet discussed how to handle fractions in our number systems. To start this discussion, let's focus on decimal representation of fractions.

For any $x \in \mathbb{Q}$, we can write

$$x = \frac{p}{q}$$

where $p, q \in \mathbb{Z}$ and $q \neq 0$. This *fractional representation* of a rational number uses two integers, namely the numerator p and nonzero denominator q separated by a horizontal line known as the *vinculum*. Let's take a look at an example of this representation in action.

EXAMPLE 5.3

The number

$$x = \frac{2718}{100}$$

has numerator $p = 2718$ and denominator $q = 100$. A corresponding $4-$digit decimal representation of this number is given by

$$\begin{aligned} x &= 27.18, \\ &= d_1\, d_0 \,.\, d_{-1}\, d_{-2}, \\ &= d_1 \cdot 10^1 + d_0 \cdot 10^0 + d_{-1} \cdot 10^{-1} + d_{-2} \cdot 10^{-2}, \\ &= d_1 \cdot 10^1 + d_0 \cdot 10^0 + \frac{d_{-1}}{10^1} + \frac{d_{-2}}{10^2}, \\ &= \sum_{k=-2}^{1} d_k \cdot 10^k. \end{aligned}$$

where $d_{-2} = 8$, $d_{-1} = 1$, $d_0 = 7$, and $d_1 = 2$.

Notice that the decimal expansions from Example 5.3 is exact: no approximations are involved in this work. Moreover, the decimal representation given in this example finite since we can write an exact decimal expansion of the fraction $\frac{2718}{100}$ using a set of 4 decimal digits $\{d_{-2}, d_{-1}, d_0, d_1\}$.

Let's look back at our finite decimal expansion

$$x = 27.18 = d_1\, d_0 \,.\, d_{-1}\, d_{-2}$$

and notice two unique features of the way we write this decimal representations. In particular, this number can be written in two parts

$$\underbrace{d_1\, d_0}_{\text{I}} \,.\, \underbrace{d_{-1}\, d_{-2}}_{\text{II}} = \underbrace{\left(d_1 \cdot 10^1 + d_0 \cdot 10^0 \right)}_{\text{part I}} + \underbrace{\left(d_{-1} \cdot 10^{-1} + d_{-2} \cdot 10^{-2} \right)}_{\text{part II}}$$

We use the radix point (also known as the decimal point for a base 10 representation) to delimit the two parts. We then associate the digits to the left of the radix point, namely $d_1 d_0$, with nonnegative powers of ten. We call the string of digits to the left of the radix point the *integer part* of our decimal expansion. This matches the work we did to express integers using a decimal number system.

The second part of our number on the right side of our radix point includes the digits $d_{-1} d_{-2}$. These digits are paired with negative powers of ten, with

$$0 \, . \, d_{-1} \, d_{-2} = d_{-1} \cdot 10^{-1} + d_{-2} \cdot 10^{-2} = \frac{d_{-1}}{10^1} + \frac{d_{-2}}{10^2}.$$

This string of digits to the right of the radix point is called the *fractional part* of our decimal expansion.

With these observations in mind, we can now generalize our work. To this end, let $i, f \in \mathbb{Z}$ be nonnegative integers. We say that say that a rational number $x \in \mathbb{Q}$ has a *finite decimal expansion* if we can represent our number, using an exact equality, in the form

$$x = \underbrace{d_i \, d_{i-1} \, ... \, d_2 \, d_1 \, d_0}_{\text{integer part}} \, . \, \underbrace{d_{-1} \, d_{-2} \, ... \, d_{-f}}_{\text{fractional part}}$$

where $d_k \in \{0, 1, 2, ..., 9\}$ for all $k \in \{i, i-1, ..., 2, 1, 0, -1, -2, ..., -f\}$. The total number of decimal digits used to represent this finite decimal expansion is

$$n = (i + 1) + f$$

where the integer part of our number $d_i \, d_{i-1} \, ... \, d_2 \, d_1 \, d_0$ is written using $(i + 1)$ decimal digits while the fractional part of our number $0 \, . d_{-1} \, d_{-2} \, ... \, d_{-f}$ is expressed with exactly f decimal digits. Let's take a look at some other examples of rational numbers that yield a finite decimal expansion.

EXAMPLE 5.4

Consider the rational number $x \in \mathbb{Q}$ given by

$$x = \frac{12824}{10}.$$

Here, our numerator is $p = 12824$ and our denominator is $q = 10$. This rational number can be represented exactly using a finite decimal expansion with $n = 5$ decimal digits. More specifically, we can write

$$\begin{aligned} x &= 1282.4, \\ &= d_3 \, d_2 \, d_1 \, d_0 \, . \, d_{-1}, \\ &= d_3 \cdot 10^3 + d_2 \cdot 10^2 + d_1 \cdot 10^1 + d_0 \cdot 10^0 + d_{-1} \cdot 10^{-1}, \\ &= \sum_{k=-1}^{3} d_k \cdot 10^k. \end{aligned}$$

where $d_{-1} = 4$, $d_0 = 2$, $d_1 = 8$, $d_2 = 2$, and $d_3 = 1$. The number $i = 3$ since the integer part of our number, given by $d_3 \, d_2 \, d_1 \, d_0 = 1282$, is expressed using $(i+1) = 4$ decimal digits. The fractional part of our number $0 \, . d_{-1} = 0 \, .4$ uses $f = 1$ digit. We can quickly confirm the statement $n = 5 = 4 + 1 = (i + 1) + f$.

EXAMPLE 5.5

Let's consider a rational number $x \in \mathbb{Q}$ with fractional representation given by

$$x = \frac{314159}{100000}.$$

In this case, our numerator is $p = 314159$ and our denominator is $q = 100000$. We write the associated $n = 6$ digit finite decimal expansion of this number as

$$\begin{aligned}
x &= 3.14159, \\
&= d_0 \, . \, d_{-1} \, d_{-2} \, d_{-3} \, d_{-4} \, d_{-5}, \\
&= d_0 \cdot 10^0 + d_{-1} \cdot 10^{-1} + d_{-2} \cdot 10^{-2} + d_{-3} \cdot 10^{-3} + d_{-4} \cdot 10^{-4} + d_{-5} \cdot 10^{-5}, \\
&= d_0 \cdot 10^0 + \frac{d_{-1}}{10^1} + \frac{d_{-2}}{10^2} + \frac{d_{-3}}{10^3} + \frac{d_{-4}}{10^4} + \frac{d_{-5}}{10^5}, \\
&= \sum_{k=-5}^{0} d_k \cdot 10^k.
\end{aligned}$$

where $d_{-5} = 9$, $d_{-4} = 5$, $d_{-3} = 1$, $d_{-2} = 4$, $d_{-1} = 1$, and $d_0 = 3$. In this case, we have $i = 0$ since we require exactly $(i + 1) = 1$ digits to write the integer part of our number $d_0 = 3$. We also see that, because the fractional part of this number $0.d_{-1}d_{-2}d_{-3}d_{-4}d_{-5} = 0.14159$ is expressed with five digits, we have $f = 5$. Once again, we confirm the fact that $n = 6 = (0 + 1) + 5 = (i + 1) + f$ as we expect.

Not all rational numbers $x \in \mathbb{Q}$ can be expressed exactly using a finite decimal approximation. For example, let's take a look at the decimal representations of all of the following rational numbers

$$\frac{244}{3} = 81.3333... = 81.\overline{3}$$

$$\frac{8972}{7} = 1283.1428571428571... = 1283.\overline{142857}$$

$$\frac{1}{6} = 0.166666... = 0.1\overline{6}$$

In all of these cases, we are unable to write an exact, finite decimal expansion of these $x \in \mathbb{Q}$. To express these numbers exactly requires an infinite decimal representation in the form

$$x = d_i \, d_{i-1} \ldots d_2 \, d_1 \, d_0 \, . \, d_{-1} \, d_{-2} \ldots d_{-f} \ldots = \sum_{k=-\infty}^{i} d_k \cdot 10^k$$

As we will see, there are two types of infinite decimal expansions. The examples above are known as *infinite recurring decimal expansions* since the fractional parts of these decimal representations eventually result in an endlessly repeating sequence of decimal digits. As we establish below, any number x that yields an infinite recurring decimal expansion must be an element of $\mathbb{Q}$ and yields a fractional representation in the form

$$x = \frac{p}{q}$$

for some integers $p, q \in \mathbb{Z}$ where denominator $q \neq 0$. There are two algorithms we can use to convert between an infinite recurring decimal representation and the corresponding fractional representation of this number. Below we illustrate the first of these algorithms based on a convenient algebraic technique.

EXAMPLE 5.6

Suppose we are given the following infinite recurring decimal expansion

$$x = 98.7654321321321... = 98.7654\overline{321}$$

We can use the fact that our fractional part of our number has repeating digits to do something creative. More specifically, we can multiply by the appropriate powers of 10 and then subtract to eliminate the infinite recurring decimals. First, we note that

$$10^4 \cdot x = 10\,000 \cdot x = 987\,654.321321321...$$

We shifted the decimal point four digits to the right so that the fractional part of this new product is purely repeating. We can use this trick again to produce a second number with the exact same repeating sequence of digits. More specifically, we notice that if we shift the decimal point another three units to the right, we get the exact same infinite repeating sequence of digits in the fractional part of this second product.

$$10^7 \cdot x = 10\,000\,000 \cdot x = 987\,654\,321.321321...$$

Using our command of arithmetic, we determine that

$$10^7 \cdot x - 10^4 \cdot x = 987\,654\,321 - 987\,654$$

$$\implies \quad (10^7 - 10^4) \cdot x = 986\,666\,667$$

$$\implies \quad x = \frac{986\,666\,667}{10^7 - 10^4} = \frac{986\,666\,667}{9\,990\,000}$$

$$\implies \quad x = \frac{328\,888\,889}{3\,330\,000}$$

We have now produced an equivalent fractional representation for our original infinite repeating decimal expansion.

Let's consider a second, equivalent method to convert our infinite recurring decimal expansion into an equivalent fractional representation based on the famous geometric series formulation that is often taught as part of an introduction to Taylor Series polynomials in many calculus classes.

EXAMPLE 5.7

Suppose we are given the following infinite recurring decimal expansion

$$x = 98.7654321321321... = 98.7654\overline{321}$$

Let's first note that we can write this number using an infinite sum. To do so, we consider the following relation

$$98.7654321321321... = 98.7654 + 0.0000321 + 0.0000000321 + 0.0000000000321 + \cdots$$

$$= 98.7654 + \frac{321}{10^7} + \frac{321}{10^{10}} + \frac{321}{10^{13}} + \cdots$$

$$= 98.7654 + \sum_{i=0}^{\infty} \frac{321}{10^{\,3i+7}}$$

At this point we can recall our famous geometric series formula that states

$$\sum_{k=1}^{n} r^k = \frac{1 - r^{n+1}}{1 - r}$$

To this end, let's focus on doing some algebra to make our infinite sum ammendable to the geometric series formula we see above. The first technique we use is to turn our infinite sum into the limit of a finite sum, as follows

$$\sum_{i=0}^{\infty} \frac{321}{10^{\,3k+7}} = \lim_{n\to\infty} \left(\sum_{i=0}^{n} \frac{321}{10^{\,3i+7}} \right)$$

We can now use a change of index variables and our dexterity in manipulating finite sums to force our finite sum for this problem into the form we desire in order to apply our geometric series formula. To this end, we notice that

$$\sum_{i=0}^{n} \frac{321}{10^{\,3i+7}} = \frac{321}{10^7} \cdot \sum_{i=0}^{n} \frac{1}{10^{\,3i}} = \frac{321}{10^7} \cdot \sum_{i=0}^{n} \left(\frac{1}{10^{\,3}} \right)^i$$

If we set $r = 10^{-3}$ and $k = i + 1$, we see that we can write

$$98.7654321321321... = \lim_{n\to\infty} \left(\frac{321}{10^7} \cdot \sum_{k=1}^{n} r^k \right) = \lim_{n\to\infty} \frac{321}{10^7} \cdot \frac{1 - r^{n+1}}{1 - r}$$

Since $r < 1$, we know that $\lim_{n\to\infty} r^{n+1} = 0$. With this we conclude that

$$98.7654321321321... = 98.7654 + \frac{321}{10^7} \cdot \frac{1}{1 - r} = 98.7654 + \frac{321}{10^7} \cdot \frac{1}{1 - \frac{1}{10^3}}$$

We now can quickly produce an equivalent fractional representation for our original infinite repeating decimal expansion.

Fractional Representation for Infinite Recurring Decimals

Every infinite decimal expansion yields a corresponding fractional expansion. In other words, if x is a number with an infinite repeating decimal expansion, then $x \in \mathbb{Q}$.

Proof. Suppose that we are given an number x with infinite decimal expansion

$$x = d_i d_{i-1} ... d_2 d_1 d_0 \, . \, d_{-1} d_{-2} ... d_{-n} \overline{d_{-(n+1)} ... d_{-(n+r)}}$$

where a sequence of $r \in \mathbb{N}$ repeating digits. Without loss of generality, suppose that $x > 0$. Notice that the integer part of this number has $(i+1)-$decimal digits while the fractional part of this number contains another n nonrepeating decimal digits. Let's begin our work here by simplifying our expression by setting $a_0 = d_i d_{i-1} ... d_2 d_1 d_0$ and $a_k = d_{-k}$ for all $k \in \mathbb{N}$. Then, we can rewrite our number as

$$x = a_0 . a_1 a_2 ... a_m \overline{a_{n+1} ... a_{(n+r)}}$$

We can use multiplication by the proper power of 10 to shift our decimal point over to the right by n places and produce a new number with a purely repeating fraction part, with

$$10^n x = 10^n a_0 \ + \ a_1 a_2 ... a_n \, . \, \overline{a_{n+1} ... a_{n+r}}$$

If we shift the decimal point another r places to the right, as follows

$$10^{n+r} x = 10^{n+r} a_0 \ + \ a_1 a_2 ... a_n a_{n+1} ... a_{n+r} \, . \, \overline{a_{n+1} ... a_{n+r}}$$

then we get yet another number with an identical fractional part. Then, we find that the difference between these two numbers is given by

$$10^{n+r} x - 10^n x = (10^{n+r} a_0 \ + \ a_1 a_2 ... a_n a_{n+1} ... a_{n+r}) - (10^n a_0 \ + \ a_1 a_2 ... a_n) = y$$

where $y \in \mathbb{N}$ is an integer with no fractional part. Then, we can write

$$x = \frac{y}{10^{n+r} - 10^n} = \frac{y}{10^n \left(10^r - 1\right)}$$

We have just produced a fractional representation of x as the quotient of two integers and thus we conclude that $x \in \mathbb{Q}$. $\square$

Now that we have a few example of rational numbers that yield a finite decimal expansion, we might ask ourselves an important question: given the fractional representation of any $x \in \mathbb{Q}$, how might we be able to immediately ascertain whether or not the corresponding decimal representation is finite?

Finite Decimal Representation Theorem

Let $x \in \mathbb{Q}$ be a positive rational number. The decimal expansion of x will be finite if and only if the fractional representation of x can be expressed with a denominator in the form

$$2^j \cdot 5^k$$

for some nonnegative integers powers $j, k \in \{0, 1, 2, 3, ...\}$.

Proof. Let $x \in \mathbb{Q}$ yield a finite decimal representation in the form

$$\begin{aligned}
x &= a_0 \,.\, a_1\, a_2 \,...\, a_{n-1}\, a_n \\
&= \frac{a_0}{10^0} + \frac{a_1}{10^1} + \frac{a_2}{10^2} + \cdots + \frac{a_{n-1}}{10^{n-1}} + \frac{a_n}{10^n} \\
&= \frac{a_0 \cdot 10^n}{10^n} + \frac{a_1 \cdot 10^{n-1}}{10^n} + \frac{a_2 \cdot 10^{n-2}}{10^n} + \cdots + \frac{a_{n-1} \cdot 10^1}{10^n} + \frac{a_n}{10^n} \\
&= \frac{1}{10^n}\left(a_0 \cdot 10^n + a_1 \cdot 10^{n-1} + a_2 \cdot 10^{n-2} + \cdots + a_{n-1} \cdot 10^1 + a_n\right) \\
&= \frac{\left(\sum_{k=0}^{n} a_k \cdot 10^{n-k}\right)}{10^n} = \frac{p}{q}
\end{aligned}$$

where $p = \sum_{k=0}^{n} a_k \cdot 10^{n-k}$ and $q = 10^n$. This denominator $q = 10^n = 2^n \cdot 5^n$ as was to be shown.

□

We just demonstrated that if have a finite decimal representation of an $x \in \mathbb{Q}$, we must be able to produce a fractional representation of x with a denominator of 10^n. Notice that we have not necessarily reduced this fractional representation to lowest terms. In fact, we may be able to reduce this fraction by eliminating shared factors in the numerator and denominator. However, the fact that we yielded a denominator in the form $2^j \cdot 5^k$ is enough in this instance.

Let's now prove the converse. In other words, let's show that if we start with a fraction that has a denominator equal to $2^j \cdot 5^k$ for some nonnegative integers powers $j, k \in \{0, 1, 2, 3, ...\}$, then we must be able to write a finite decimal expansion for this fraction.

Proof. Assume $x \in \mathbb{Q}$ has a fractional representation $x = \frac{t}{q}$ where $t \in \mathbb{Z}$ and $q = 2^j \cdot 5^k$ for some nonnegative integers powers $j, k \in \{0, 1, 2, 3, ...\}$. Then we can write

$$x = \frac{t}{2^j \cdot 5^k} = \frac{t}{2^j \cdot 5^k} \cdot \frac{2^k}{2^k} \cdot \frac{5^j}{5^j} = \frac{2^k \cdot 5^j \cdot t}{2^{j+k} \cdot 5^{j+k}} = \frac{2^k \cdot 5^j \cdot t}{10^{j+k}}.$$

Using this transformation, we see we can write

$$x = \frac{p}{10^n}$$

where $n = (j + k)$ and $p = 2^k \cdot 5^j \cdot t$. Moreover, since $p \in \mathbb{Z}$, we can use a decimal representation of this integer to write

$$\begin{aligned} p &= 2^k \cdot 5^j \cdot t \\ &= d_n \cdot 10^n + d_{n-1} \cdot 10^{n-1} + \cdots + d_2 \cdot 10^2 + d_1 \cdot 10^1 + d_0 \cdot 10^0 \\ &= \sum_{i=0}^{n} d_i \cdot 10^i \end{aligned}$$

We rewrite combine our fractional representation of x using this equivalent expression of our numerator to confirm that

$$\begin{aligned} x &= \frac{d_n \cdot 10^n + d_{n-1} \cdot 10^{n-1} + \cdots + d_2 \cdot 10^2 + d_1 \cdot 10^1 + d_0 \cdot 10^0}{10^n} \\ &= \frac{d_n}{10^0} + \frac{d_{n-1}}{10^1} + \cdots + \frac{d_2}{10^{n-2}} + \frac{d_1}{10^{n-1}} + \frac{d_0}{10^n} \\ &= a_0 \,.\, a_1\, a_2 \,...\, a_{n-1}\, a_n \end{aligned}$$

where $a_i = d_{n-i}$ for all $i \in \{0, 1, 2, ..., n-1, n\}$. This is the exact finite decimal expansion that we wanted to produce. □

The theorems and proofs offered above provide us with a useful way to think about whether or not a given rational numbers $x \in \mathbb{Q}$ yields a finite decimal expansion. We now know that the only way a number $x \in \mathbb{Q}$ has a finite decimal expansion is if we can reduce x to a fractional representation whose denominator contains only powers of 2 and 5. We also know that the decimal expansion of a rational number is either finite or endlessly repeating. Before we study the more general space of numbers that require infinite, nonrepeating decimal expansions, let's transfer our hard-earned intuition about decimal fractions into binary representations.

Binary fractions

To begin, we recall that binary representations follow a similar pattern as in the decimal case except that, in binary, we work with radix 2. We can then define a *finite binary expansion* using some slight changes to our general structure suggested above. To this end, let $i, f \in \mathbb{Z}$ be nonnegative integers. We say that say that a rational number $x \in \mathbb{Q}$ has a finite binary representation if we can express our number exactly in the form

$$x = \underbrace{b_i \, b_{i-1} \ldots b_2 \, b_1 \, b_0}_{\text{integer part}} \; . \; \underbrace{b_{-1} \, b_{-2} \ldots b_{-f}}_{\text{fractional part}}$$

where $b_k \in \{0, 1\}$ for all $k \in \{i, i-1, ..., 2, 1, 0, -1, -2, ..., -f\}$. This is an $m-$bit binary representation of a rational number, where is

$$m = (i + 1) + f.$$

Once again, we note that the $(i+1)-$bit integer part of x is given by $b_i \, b_{i-1} \ldots b_2 \, b_1 \, b_0$. The $f-$bit fractional part of x is $0 \, . b_{-1} \, b_{-2} \ldots b_{-f}$. Let's take a look at some other examples of rational numbers that yield a finite binary expansion.

EXAMPLE 5.8

Let's begin our work studying finite binary expansions with a relatively simple example. In particular, suppose we want to represent

$$x = \frac{3}{2}$$

using a finite binary expansion. We start by writing the numerator of x in terms of powers of 2. One way to do this is to consider

$$\begin{aligned} 2 \cdot x = 3 &= 2 + 1 \\ &= 2^1 + 2^0 \\ &= (11)_2 \end{aligned}$$

This is an unsigned binary integer representation of the numerator 3. Now, we divide our number $2 \cdot x$ by $2^1 = 2$ to find our corresponding binary expansion

$$\begin{aligned} \frac{2 \cdot x}{2} &= \frac{2^1 + 2^0}{2^1} \\ &= 2^0 + 2^{-1} \\ &= (\, 1.1 \,)_2 \\ &= b_0 \, . \, b_{-1}. \end{aligned}$$

This is a $m-$bit binary expansion of our rational number $x \in \mathbb{Q}$, where

$$m = (i + 1) + f = 2$$

In this example, we have $i = 0$ and $f = 1$ since the integer part has $i + 1 = 0 + 1$ binary digits while the fractional part has $f = 1$ bits.

EXAMPLE 5.9

For our next example, let's represent the rational number

$$x = \frac{19}{8}$$

with a finite binary expansion. Again we begin by eliminating the denominator and writing the numerator of x an unsigned binary integer:

$$\begin{aligned} 8 \cdot x = 19 &= 16 + 2 + 1 \\ &= 2^4 + 2^1 + 2^0 \\ &= (10011)_2 \end{aligned}$$

We now divide $8 \cdot x$ by 2^3 to find our desired finite binary expansion

$$\begin{aligned} \frac{8 \cdot x}{8} &= \frac{2^4 + 2^1 + 2^0}{2^3} \\ &= 2^1 + 2^{-2} + 2^{-3} \\ &= (\, 10.011 \,)_2 \\ &= b_1\, b_0\, .\, b_{-1}\, b_{-2}\, b_{-3}. \end{aligned}$$

This is a $m-$bit binary expansion of our rational number $x \in \mathbb{Q}$, where

$$m = (i + 1) + f = 5$$

In this example, we have $i = 1$ and $f = 3$ since the integer part has $2 = i+1 = 1+1$ bits while the fractional part has $f = 3$ bits.

In our example above, can write equivalent division problems in base 10 or base 2:

$$(2.375)_{10} = \left(\frac{19}{8}\right)_{10} = \left(\frac{10011}{1000}\right)_2 = (10.011)_2$$

When we divide our unsigned binary integer 10011 by $2^3 = 1000$ has the effect of shifting the binary point in the numerator three places to the left to yield the finite binary expansion 10.011.

EXAMPLE 5.10

Let's consider the following fractional representation of the rational number

$$x = \frac{749}{64}$$

Let's begin the process of writing the binary representation of this number by writing an expansion of x in terms of powers of 2. One way to do this is to consider

$$\begin{aligned} 64 \cdot x = 749 &= 512 + 128 + 64 + 32 + 8 + 4 + 1 \\ &= 2^9 + 2^7 + 2^6 + 2^5 + 2^3 + 2^2 + 2^0 \\ &= (10111011001)_2 \end{aligned}$$

This gives us an unsigned binary integer representation of the numerator. Now, we can divide our number $64 \cdot x$ by $2^6 = 64$ to find our corresponding binary representation

$$\begin{aligned} \frac{64 \cdot x}{64} &= \frac{2^9 + 2^7 + 2^6 + 2^5 + 2^3 + 2^2 + 2^0}{2^6} \\ &= 2^3 + 2^1 + 2^0 + 2^{-1} + 2^{-3} + 2^{-4} + 2^{-6} \\ &= (\,1011.101101\,)_2 \\ &= b_3\, b_2\, b_1\, b_0\, .\, b_{-1}\, b_{-2}\, b_{-3}\, b_{-4}\, b_{-5}\, b_{-6}. \end{aligned}$$

This is a $m-$bit binary expansion, with

$$m = (i + 1) + f = 10.$$

Here we have $i = 3$ and $f = 6$ since the integer part has $4 = i+1 = 3+1$ bits while the fractional part has $f = 6$ bits.

In Example 5.10 above, we can write the fractional representation problem in base 10 or base 2, yielding the equivalence

$$(11.703125)_{10} = \left(\frac{749}{64}\right)_{10} = \left(\frac{10111011001}{1000000}\right)_2 = (1011.101101)_2$$

Dividing our unsigned binary integer 10111011001 by the binary number $2^6 = 1000000$, this has the effect of shifting the binary point in the numerator six places to the left. This is the binary analog of the effect of dividing by decimal integers by powers of 10 that we observe in decimal arithmetic. Of course this makes sense since in binary $(2)_{10} = (10)_2$. In other words, to shift the binary point right or left by p positions in a given number $x \in \mathbb{Q}$, we multiply or divide x by the number 2^p.

Just as in the case of finite decimal representations, not all rational numbers $x \in \mathbb{Q}$ can be expressed exactly using a finite binary expansion. For example, let's take a look at the decimal representations of all of the following rational numbers

$$(2.4)_{10} = \left(\frac{12}{5}\right)_{10} = (10.0110\,0110\,0110...)_2 = \left(10.\overline{0110}\right)_2$$

$$(0.9)_{10} = \left(\frac{9}{10}\right)_{10} = (0.1\,1100\,1100\,1100...)_2 = \left(0.1\overline{1100}\right)_2$$

$$\left(45.7\overline{3}\right)_{10} = \left(\frac{686}{15}\right)_{10} = (101101.1011\,1011\,1011...)_2 = \left(101101.\overline{1011}\right)_2$$

For each of these $x \in \mathbb{Q}$, we are unable to write an exact, finite binary expansion. Instead, in order to express these numbers exactly requires an *infinite binary expansion* in the form

$$x = b_i\, b_{i-1} \ldots b_2\, b_1\, b_0\, .\, b_{-1}\, b_{-2} \ldots b_{-f} \ldots = \sum_{k=-\infty}^{i} b_k \cdot 2^k$$

This realization has very neat analogies to our comparisons between finite and infinite recurring decimal expansions. More specifically, we might wonder what type of rational numbers $x \in \mathbb{Q}$ yield an exact, finite binary expansion.

As we will see, there are two types of infinite binary expansions. The examples above are known as *infinite recurring binary expansions* since the fractional parts of these binary representations eventually result in an endlessly repeating sequence of binary digits. Just as we did when our radix is 10, we will show that any number x that has an infinite recurring binary expansion must be an element of $\mathbb{Q}$ and yields a fractional representation in the form

$$x = \frac{p}{q}$$

for some integers $p, q \in \mathbb{Z}$ where denominator $q \neq 0$. For the radix 2 case, we establish this fact using a base 2 analog of the same techniques we used to establish this idea in the base 10 case. We begin with an example to study the major idea we will use to prove this fact.

EXAMPLE 5.11

Suppose we have the following infinite recurring binary expansion

$$x = 101.1\overline{0110}$$

Since the fractional part of our number has repeating binary digits, we can multiply by the appropriate powers of 2 and then subtract to eliminate the infinite recurring bits. To this end, we notice that

$$\left(2^5 \cdot x\right)_{10} = \left(100\,000 \cdot x\right)_2 = \left(10110110.\overline{0110}\right)_2$$

We shifted the decimal point five bits to the right so that the fractional part of this new product is purely repeating. We use this technique once more to produce a second product with the exact same repeating sequence of digits, with

$$\left(2^9 \cdot x\right)_{10} = \left(1\,000\,000\,000 \cdot x\right)_2 = \left(101101100110.\overline{0110}\right)_2$$

Using our command of arithmetic, we determine that

$$\left(2^9 \cdot x - 2^5 \cdot x\right)_{10} = \left(1011\,0110\,0110 - 1011\,0110\right)_2$$

$$\implies \quad \left((2^9 - 2^5) \cdot x\right)_{10} = \left(2918 - 182\right)_{10}$$

$$\implies \quad x = \frac{2736}{2^9 - 2^5} = \frac{2736}{480}$$

$$\implies \quad x = \left(\frac{57}{10}\right)_{10} = \left(5.7\right)_{10}$$

We have now produced an equivalent fractional representation for our original infinite repeating decimal expansion.

Fractional Representation for Infinite Recurring Bits

Every infinite binary expansion yields a corresponding fractional expansion. In other words, if x is a number with an infinite repeating binary expansion, then $x \in \mathbb{Q}$.

Proof. Suppose that we are given an number x with infinite decimal expansion

$$x = b_i b_{i-1} ... b_2 b_1 b_0 \, . \, b_{-1} b_{-2} ... b_{-n} \overline{b_{-(n+1)} ... b_{-(n+r)}}$$

where a sequence of $r \in \mathbb{N}$ repeating bits. Without loss of generality, suppose that $x > 0$. Notice that the integer part of this number has $(i+1)-$decimal digits while the fractional part of this number contains another n nonrepeating binary digits. Let's begin our work here by simplifying our expression by setting $a_0 = b_i b_{i-1} ... b_2 b_1 b_0$ and $a_k = b_{-k}$ for all $k \in \mathbb{N}$. Then, we can rewrite our number as

$$x = a_0 . a_1 a_2 ... a_m \overline{a_{n+1} ... a_{(n+r)}}$$

We can use multiplication by the proper power of 2 to shift our decimal point over to the right by n places and produce a new number with a purely repeating fraction part, with

$$2^n x = 2^n a_0 \ + \ a_1 a_2 ... a_n \, . \, \overline{a_{n+1} ... a_{n+r}}$$

If we shift the decimal point another r places to the right, as follows

$$2^{n+r} x = 2^{n+r} a_0 \ + \ a_1 a_2 ... a_n a_{n+1} ... a_{n+r} \, . \, \overline{a_{n+1} ... a_{n+r}}$$

then we get yet another number with an identical fractional part. Then, we find that the difference between these two numbers is given by

$$2^{n+r} x - 2^n x = (2^{n+r} a_0 \ + \ a_1 a_2 ... a_n a_{n+1} ... a_{n+r}) - (2^n a_0 \ + \ a_1 a_2 ... a_n) = y$$

where $y \in \mathbb{N}$ is an integer with no fractional part. Then, we can write

$$x = \frac{y}{2^{n+r} - 2^n} = \frac{y}{2^n \left(2^r - 1\right)}$$

We have just produced a fractional representation of x as the quotient of two integers and thus we conclude that $x \in \mathbb{Q}$. □

Finite Binary Representation Theorem

Let $x \in \mathbb{Q}$ be a positive rational number. The binary expansion of x will be finite if and only if the fractional representation of x can be expressed with a denominator in the form

$$2^j$$

for some nonnegative integer powers $j \in \{0, 1, 2, 3, ...\}$.

The proof of this proposition is left to the reader. Please look back on our work for the base 10 analog of this theorem for some good ideas on how to start.

Now that we have categorized the type of rational numbers that yield both infinite recurring and finite decimal or binary expansions, let's consider how we might convert between decimal and binary expansions for $x \in \mathbb{Q}$.

Converting $x \in \mathbb{Q}$ from Decimal to Binary
Converting $x \in \mathbb{Q}$ from Binary to Decimal

number of digits used in a decimal expansion = the total number of decimal digits used to express a value. This may include *leading zeros* at the beginning of a number or *trailing zeros* at the end of a number. This may also include zeros at the beginning of (the fractional part of) the number as these zeros only help to indicate the location of the radix point.

minimum number of digits used in a decimal expansion = the minimum number of decimal digits used to express a value. This will not include include *leading zeros* at the beginning of a number or *trailing zeros* at the end of a number. This may also include zeros at the beginning of (the fractional part of) the number as these zeros only help to indicate the location of the radix point.