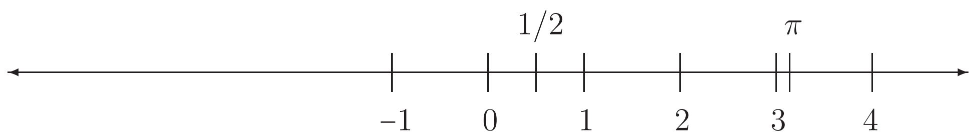
Computer Numbers and Arithmetic

References:

- M. Overton, Numerical Computing with IEEE Floating Point Arithmetic, SIAM, 2004
- IEEE Computer Society, IEEE Standard for Floating-Point Arithmetic, 2008
- Cheney & Kincaid, Numerical Mathematics and Computing, Sections 1.1 and 1.3

Classes of Real Numbers

All real numbers can be represented by a line:



The Real Line

$$\text{real numbers}\begin{cases}\text{rational numbers}\begin{cases}\text{integers}\\ \text{non-integral fractions}\end{cases}\\ \text{irrational numbers}\end{cases}$$

Rational numbers
All of the real numbers which consist of a ratio of two integers.

Irrational numbers
Most real numbers are **not** rational, i.e. there is no way of writing them as the ratio of two integers. These numbers are called **irrational**.
Familiar examples of irrational numbers are: $\sqrt{2}$, π and e.

How to represent numbers?

- The **decimal**, or **base 10**, system requires 10 symbols, $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$.
- The **binary**, or **base 2**, system is convenient for electronic computers: here, every number is represented as a string of **0**'s and **1**'s.

Decimal and binary representation of **integers** is simple, requiring an expansion in nonnegative powers of the base; e.g.

$$(71)_{10} = 7 \times 10 + 1$$

and its **binary equivalent:** $(1000111)_2 =$

$$1 \times 64 + 0 \times 32 + 0 \times 16 + 0 \times 8 + 1 \times 4 + 1 \times 2 + 1 \times 1.$$

Non-integral **fractions** have entries to the **right** of the point. e.g. **finite** representations

$$\frac{11}{2} = (5.5)_{10} = 5 \times 1 + 5 \times \frac{1}{10},$$

$$\frac{11}{2} = (101.1)_2 = 1 \times 4 + 0 \times 2 + 1 \times 1 + 1 \times \frac{1}{2}$$

Infinitely Long Representations

But 1/10, with finite **decimal** expansion $(0.1)_{10}$, has the **binary** representation

$$\begin{aligned} \frac{1}{10} &= (0.0001100110011\ldots)_2 \\ &= \frac{1}{16} + \frac{1}{32} + \frac{0}{64} + \frac{0}{128} + \frac{1}{256} + \frac{1}{512} + \cdots. \end{aligned}$$

This, while **infinite**, is **repeating**.
1/3 has **both** representations infinite *and* repeating:

$$1/3 = (0.333\ldots)_{10} = (0.010101\ldots)_2.$$

If the representation of a **rational** number is *infinite*, it **must** be *repeating*. e.g.

$$1/7 = (0.142857142857\ldots)_{10}.$$

Irrational numbers always have **infinite, non-repeating** expansions. e.g.

$$\begin{aligned} \sqrt{2} &= (1.414213...)_{10}, \\ \pi &= (3.141592\ldots)_{10}, \\ e &= (2.71828182845\ldots)_{10}. \end{aligned}$$

Converting between binary & decimal numbers

- **Binary $\longrightarrow$ decimal**:
  Easy. e.g. $(1001.11)_2$ is the decimal number

  $$1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} = 9.75$$

- **Decimal $\longrightarrow$ binary**:
  Convert the integer and fractional parts separately.
  e.g. if x is a **decimal integer**, we want coefficients $a_0, a_1, \ldots, a_n$, all 0 or 1, so that

  $$a_n \times 2^n + a_{n-1} \times 2^{n-1} + \cdots + a_0 \times 2^0 = x,$$

  has representations $(a_n a_{n-1} \cdots a_0)_2 = (x)_{10}$.

  Clearly dividing x by 2 gives **remainder** a_0, leaving as **quotient**

  $$a_n \times 2^{n-1} + a_{n-1} \times 2^{n-2} + \cdots + a_1 \times 2^0,$$

  and so we can continue to find a_1 then a_2 etc.

  Q: What is a similar approach for decimal fractions?

Computer Representation of Numbers

- **Integers** — three ways:

  1. **sign-and-modulus** — a simple approach.

     Use 1 bit to represent the **sign**, and store the **binary** representation of the magnitude of the integer. e.g. decimal 71 is stored as the bitstring

0	00...01000111

     If the computer **word size** is 32 bits, $2^{31} - 1$ is the **largest** magnitude which will fit.

  2. **2's complement representation (CR)**
     more convenient, & used by most machines.

     (i) The **nonnegative** integers 0 to $2^{31} - 1$ are stored as before, e.g., 71 is stored as the bitstring

     $$\boxed{000\ldots01000111}$$

     (ii) A **negative integer** $-x$, where $1 \le x \le 2^{31}$, is stored as the **positive integer** $2^{32} - x$.

     e.g. -71 is stored as the bitstring

     $$\boxed{111\ldots10111001}.$$

     Converting x to its 2's CR $2^{32} - x$ of $-x$:

     $$2^{32} - x = (2^{32} - 1 - x) + 1,$$

     $$2^{32} - 1 = (111\ldots111)_2.$$

     Chang all zero bits of x to ones, one bits to zero and adding one.

     Q: What is the quickest way of deciding if a number is negative or nonnegative using 2's CR?

     An advantage of 2's CR:
     Form $y + (-x)$, where $0 \le x, y \le 2^{31} - 1$.

     $$\text{2's CR of } y \text{ is } y\,; \quad \text{2's CR of } -x \text{ is } 2^{32} - x$$

     Adding these two **representations** gives

     $$y + (2^{32} - x) = 2^{32} + y - x = 2^{32} - (x - y).$$

     – If $y \ge x$, the LHS will not fit in a 32-bit word, and the **leading bit** can be dropped, giving the **correct result**, $y - x$.
     – If $y < x$, the RHS is **already correct**, since it *represents* $-(x - y)$.

Thus, **no special hardware is needed for integer subtraction**. The addition hardware can be used, once $-x$ has been represented using 2's complement.

3. **1's complement representation:**
   a negative integer $-x$ is stored as $2^{32} - x - 1$.
   This system *was* used, but no longer.

- **Non-integral real numbers.**

  Real numbers are approximately stored using the **binary** representation of the number.

  Two possible methods:
  . **fixed point** and **floating point**.

  Fixed point: the computer word is divided into **three fields**, one for each of:
  • the **sign** of the number
  • the number **before** the point
  • the number **after** the point.

  In a **32-bit word** with field widths of **1,15** and **16**, the number 11/2 would be stored as:

0	000000000000101	1000000000000000

  The fixed point system has a severe limitation on the **size** of the numbers to be stored.

  Q: What are the **smallest** and **largest** numbers in magnitude the above system can store?

  Thus fixed point system is inadequate for most scientific computing.

  Note: The advantage of the fixed point system is that fixed point arithmetic is orders of magnitude faster than floating point arithmetic.

Normalized Exponential Notation

In (normalized) **exponential notation**,
a nonzero real number is written as

$$\pm m \times 10^E, \qquad 1 \le m < 10,$$

- m is called the **significand** or mantissa,
- E is an **integer**, called the **exponent**.

For the **computer** we need **binary**, write $x \neq 0$ as

$$x = \pm m \times 2^E, \quad \text{where } 1 \le m < 2.$$

The binary expansion for m is

$$m = (b_0.b_1b_2b_3\ldots)_2, \quad \text{with } b_0 = 1.$$

IEEE Floating Point Arithmetic

Through the efforts of **W. Kahan** & others, a **binary** floating point standard was developed: IEEE 754-1985. It has now been adopted by almost all computer manufacturers. Another standard, IEEE 854-1987 for radix independent floating point arithmetic, is devoted to both binary (radix-2) and decimal (radix-10) arithmetic. The current version is IEEE 754-2008, including nearly all of the original IEEE 754-1985 and IEEE 854-1987. When we say the **IEEE standard** in this course, we refer to the binary standard.

The standard defines:

- **arithmetic formats:** sets of binary floating-point data, which consist of finite numbers (including signed zeros and subnormal numbers), infinities, and special "not a number" values (NaNs)
- **interchange formats:** encodings (bit strings) that may be used to exchange floating-point data in an efficient and compact form
- **rounding rules:** properties to be satisfied when rounding numbers during arithmetic and conversions
- **operations:** arithmetic and other operations (such as trigonometric functions) on arithmetic formats
- **exception handling:** indications of exceptional conditions (such as division by zero, overflow, etc.)

IEEE Single Format (binary32)

There are **3 binary floating point basic formats** in the IEEE standard: binary32, binary64 and binary128, which are also referred to as **single**, **double**, and **quadruple** formats or precision, respectively. A computer which adopts the IEEE standard must
Single format numbers use **32-bit words**.
A 32-bit **word** is divided into **3 fields**:

- **sign field**: **1** bit (**0** for positive, **1** for negative).
- **exponent field**: **8** bits for E.
- **significand field**: **23** bits for m.

In the **IEEE single format system**, the **23** significand bits are used to store $b_1 b_2 \ldots b_{23}$.
Do not store b_0, since we know $b_0 = 1$. This idea is called **hidden bit normalization**.
The **stored bitstring** $b_1 b_2 \ldots b_{23}$ is now the **fractional part** of the significand, the significand field is also referred to as the **fraction field**.
It may not be possible to store x with such a scheme, because

- either E is outside the permissible range (see later).
- or $b_{24}, b_{25}, \ldots$ are **not all zero**.

Def. A number is called a **(computer) floating point number** if it can be stored **exactly** this way, e.g.,

$$71 = (1.000111)_2 \times 2^6$$

can be represented by

0	$E = 6$	00011100000000000000000

.

If x is not a floating point number, it must be **rounded** before it can be stored on the computer.

Special Numbers

- 0. Zero cannot be **normalized**.

  A pattern of **all 0s** in the fraction field of a normal number represents the significand **1.0**, not **0.0**.

- -0. -0 and 0 are **two different representations for the same value**
- ∞. This allows e.g. $1.0/0.0 \to \infty$, instead of terminating with an **overflow** message.
- $-\infty$. $-\infty$ and ∞ represent **two very different numbers**.
- NaN, or **"Not a Number"**, is an **error pattern**.
- Subnormal numbers (see later)

All **special numbers** are represented by **a special bit pattern** in the **exponent field**.

IEEE Single format

$\pm$	$a_1a_2a_3\ldots a_8$	$b_1b_2b_3\ldots b_{23}$

If exponent $a_1\ldots a_8$ is	Then value is
$(00000000)_2 = (0)_{10}$	$\pm(0.b_1..b_{23})_2 \times 2^{-126}$
$(00000001)_2 = (1)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{-126}$
$(00000010)_2 = (2)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{-125}$
$(00000011)_2 = (3)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{-124}$
$\downarrow$	$\downarrow$
$(01111111)_2 = (127)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{0}$
$(10000000)_2 = (128)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{1}$
$\downarrow$	$\downarrow$
$(11111100)_2 = (252)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{125}$
$(11111101)_2 = (253)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{126}$
$(11111110)_2 = (254)_{10}$	$\pm(1.b_1..b_{23})_2 \times 2^{127}$
$(11111111)_2 = (255)_{10}$	$\pm\infty$ if $b_1,\ldots,b_{23} = 0$; NaN otherwise.

The $\pm$ refers to the sign, **0** for **positive**, **1** for **negative**.

- All lines *except the first and the last* refer to the **normal** numbers, i.e. **not special**.

  The exponent representation $a_1a_2\ldots a_8$ uses **biased representation**: this bitstring is the binary representation of $E + 127$. 127 is the **exponent bias**.
  e.g. $1 = (1.000\ldots0)_2 \times 2^0$ is stored as

0	01111111	00000000000000000000000

  Exponent range for normal numbers is 00000001 to 11111110 (1 to 254), representing **actual exponents**

  $$\boxed{E_{min} = -126 \text{ to } E_{max} = 127}$$

  The **smallest positive normal number** is

  $$(1.000\ldots0)_2 \times 2^{-126}:$$

0	00000001	00000000000000000000000

  approximately 1.2×10^{-38}.

  The **largest positive normal number** is

  $$(1.111\ldots1)_2 \times 2^{127}:$$

0	11111110	11111111111111111111111

  approximately 3.4×10^{38}.

- Last line:

If exponent $a_1 \dots a_8$ is	Then value is
$(11111111)_2 = (255)_{10}$	$\pm\infty$ if $b_1, \dots, b_{23} = 0$; NaN otherwise

This shows an **exponent bitstring of all ones** is a special pattern for $\pm\infty$ or NaN, depending on the value of the fraction.

There are two kinds of NaN: **a quite NaN (qNaN)** if $b_1 = 1$ and **a signaling NaN (sNaN)** if $b_1 = 0$.

- First line

$(00..00)_2 = (0)_{10}$	$\pm(0.b_1..b_{23})_2 \times 2^{-126}$

shows **zero** requires a zero bitstring for the *exponent* field **as well as** for the *fraction*:

0	00000000	00000000000000000000000

Initial unstored bit is 0, not 1, in line 1.

If **exponent** is **zero**, but **fraction** is **nonzero**, the number represented is **subnormal**.

Although 2^{-126} is the smallest positive **normal** number, we can represent **smaller** numbers called **subnormal** numbers.

e.g. $2^{-127} = (0.1)_2 \times 2^{-126}$:

0	00000000	10000000000000000000000

and $2^{-149} = (0.0000\dots01)_2 \times 2^{-126}$:

0	00000000	00000000000000000000001

This is the smallest positive number we can store.

Subnormal numbers cannot be normalized, as this gives exponents which do not fit.

Subnormal numbers are **less accurate**, (less room for nonzero bits in the fraction). e.g. $(1/10) \times 2^{-133} = (0.11001100\dots)_2 \times 2^{-136}$ is

0	00000000	00000000011001100110011001

IEEE Double Format (binary64)

The **double format (binary64)** uses 64 bits. The ideas are the same as those for the single format. The differences are that the field widths are (1, 11 & 52) and exponent bias is $2^{11-1} - 1 = 1023$.

$\pm$	$a_1a_2a_3\ldots a_{11}$	$b_1b_2b_3\ldots b_{52}$

If exponent is $a_1..a_{11}$	Then value is
$(000..0000)_2 = (0)_{10}$	$\pm(0.b_1..b_{52})_2 \times 2^{-1022}$
$(000..0001)_2 = (1)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{-1022}$
$(000..0010)_2 = (2)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{-1021}$
$(000..0011)_2 = (3)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{-1020}$
$\downarrow$	$\downarrow$
$(01..111)_2 = (1023)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{0}$
$(10..000)_2 = (1024)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{1}$
$\downarrow$	$\downarrow$
$(11..100)_2 = (2044)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{1021}$
$(11..101)_2 = (2045)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{1022}$
$(11..110)_2 = (2046)_{10}$	$\pm(1.b_1..b_{52})_2 \times 2^{1023}$
$(11..111)_2 = (2047)_{10}$	$\pm\infty$ if $b_1, \ldots, b_{52} = 0$; NaN otherwise

IEEE Quadruple Format (binary128)

The **quadruple format (binary128)** uses 128 bits, the field widths are (1, 15 & 112) and exponent bias is $2^{14} - 1 = 16383$.

$\pm$	$a_1a_2a_3\ldots a_{15}$	$b_1b_2b_3\ldots b_{112}$

If exponent is $a_1..a_{15}$	Then value is
$(000..0000)_2 = (0)_{10}$	$\pm(0.b_1..b_{112})_2 \times 2^{-16382}$
$(000..0001)_2 = (1)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{-16382}$
$(000..0010)_2 = (2)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{-16381}$
$(000..0011)_2 = (3)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{-16380}$
$\downarrow$	$\downarrow$
$(01..111)_2 = (16383)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{0}$
$(10..000)_2 = (16384)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{1}$
$\downarrow$	$\downarrow$
$(11..100)_2 = (32764)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{16381}$
$(11..101)_2 = (32765)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{16382}$
$(11..110)_2 = (32766)_{10}$	$\pm(1.b_1..b_{112})_2 \times 2^{16383}$
$(11..111)_2 = (32767)_{10}$	$\pm\infty$ if $b_1, \ldots, b_{112} = 0$; NaN otherwise

Precision, Machine Epsilon

Def. Precision:

The number of bits in the significand (including the hidden bit) is called the **precision** of the floating point system, denoted by p.

In the **single format** system, $p = 24$, i.e., the "single precision" is 24. Recall the term "single precision" is also used to mean the single format.

Def. Machine Epsilon:

The gap between the number **1** and the **next larger** floating point number is called the **machine epsilon** of the floating point system, denoted by ϵ.

In the **single format** system, the number after 1 is

$$b_0.b_1 \ldots b_{23} = 1.00000000000000000000001,$$

so $\epsilon = 2^{-23}$.

Q: Gap between two consecutive IEEE single format numbers:

(i) How is 2 represented?

0	10000000	00000000000000000000000

(ii) What is the next smallest IEEE single format number larger than 2?

0	10000000	00000000000000000000001

(iii) What is the gap between 2 and the first IEEE single format number larger than 2?

$$2^{-23} \times 2 = 2^{-22}.$$

General result:

Let $x = m \times 2^E$ be a single format number with $1 \le m < 2$. The **gap** between x and the next single format number is

$$\epsilon \times 2^E.$$

Machine Precision and Epsilon of the 3 Formats

Name	Common name	Precision p	Machine ϵ
binary32	Single format	24	$2^{-23} \approx 1.2 \times 10^{-7}$
binary64	Double format	53	$2^{-52} \approx 2.2 \times 10^{-16}$
binary128	Quadruple format	113	$2^{-112} \approx 1.9 \times 10^{-34}$

Significant Digits

- The single precision $p = 24$ corresponds to **approximately** 7 **significant decimal digits**, since
$$2^{-24} \approx 10^{-7}.$$
Equivalently,
$$\log_{10}(2^{24}) \approx 7.$$
- The double precision $p = 53$ corresponds to **approximately** 16 **significant decimal digits**, since
$$2^{-53} \approx 10^{-16}.$$
- The quadruple precision $p = 114$ corresponds to **approximately** 34 **significant decimal digits**, since
$$2^{-114} \approx 10^{-34}.$$

Here we use the word "approximately" because it is difficult to define "significant digits" precisely.

Rounding

We use **Floating Point Data (FPD)** to include

± 0, subnormal & normal floating point numbers (FPNs), & $\pm\infty$ and NaNs

in a given format, e.g., single. These form a **finite** set.

Let N_{min} denote the minimum positive **normal** FPN and let N_{max} denote the maximum positive **normal** FPN;

A real number x is in the **"normal range"** if
$$N_{min} \leq |x| \leq N_{max}.$$

Q: Let x be a real number and $|x| \leq N_{\max}$. If x is <u>not</u> a FPN, what are two obvious choices for the floating point **approximation** to x?

x_- the closest FPN **less** than x; x_+ the closest FPN **greater** than x.

Using **IEEE single** format, if x is positive with
$$x = (b_0.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots)_2 \times 2^E,$$
$$b_0 = 1 \text{ (normal), or } b_0 = 0,\ E = -126 \text{ (subnormal)}$$
then **discarding** b_{24}, b_{25}, ... gives.
$$x_- = (b_0.b_1b_2\ldots b_{23})_2 \times 2^E,$$
An algorithm for x_+ is more complicated since it may involve some bit "carries".
$$x_+ = \left[(b_0.b_1b_2\ldots b_{23})_2 + (0.00\ldots 1)_2\right] \times 2^E.$$
If x is **negative**, the situation is reversed: x_+ is obtained by dropping bits b_{24}, b_{25}, etc.

Correctly Rounded Arithmetic

The IEEE standard defines the **correctly rounded value of** x: round(x).
If x **is** a floating point number, round$(x) = x$. Otherwise round(x) depends on the **rounding mode** in effect:

- **Round down:** round$(x) = x_-$.
- **Round up:** round$(x) = x_+$.
- **Round towards zero:**
  round(x) is either x_- or x_+,
  whichever is between zero and x.
- **Round to nearest:** round(x) is either x_- or x_+, whichever is nearer to x. In the case of a tie, the one with its **least significant bit equal to zero** is chosen.

  This rounding mode is almost always used.

If x is **positive**, then x_- is between zero and x, so **round down** and **round towards zero** have the same effect. **Round towards zero** simply requires **truncating** the binary expansion, i.e. discarding bits.
"Round" with no qualification usually means **"round to nearest"**.

Absolute Rounding Error

Def. The **absolute rounding error** associated with x:

$$|\text{round}(x) - x|.$$

Its value depends on mode.
For all modes $|\text{round}(x) - x| < |x_+ - x_-|$.
Suppose $N_{\min} \le x \le N_{\max}$,

$$x = (b_0.b_1b_2\dots b_{23}b_{24}b_{25}\dots)_2 \times 2^E, \quad b_0 = 1.$$

$$\text{IEEE } \textbf{single} \quad x_- = (b_0.b_1b_2\dots b_{23})_2 \times 2^E.$$

$$\text{IEEE } \textbf{single} \quad x_+ = x_- + (0.00\dots01)_2 \times 2^E.$$

So for **any** mode

$$|\text{round}(x) - x| < 2^{-23} \times 2^E.$$

In general for **any rounding mode**:

$$|\text{round}(x) - x| < \epsilon \times 2^E. \quad (*)$$

Q: (i) For **round towards zero**, could the absolute rounding error equal $\epsilon \times 2^E$?
(ii) Does $(*)$ hold if $0 < x < N_{\min}$, i.e. $E = -126$ and $b_0 = 0$?

Relative Rounding Error, $x \neq 0$

The **relative rounding error** is defined by $|\delta|$, where

$$\delta \equiv \frac{\text{round}(x)}{x} - 1 = \frac{\text{round}(x) - x}{x}.$$

Assuming x is in the normal range,

$$x = \pm m \times 2^E, \qquad \text{where } m \geq 1,$$

so $|x| \geq 2^E$. Since $|\text{round}(x) - x| < \epsilon \times 2^E$, we have, for **all** rounding modes,

$$|\delta| < \frac{\epsilon \times 2^E}{2^E} = \epsilon. \quad (*)$$

Q: Does $(*)$ necessarily hold if $0 < |x| < N_{\min}$, i.e. $E = -126$ and $b_0 = 0$? Why?
Note for any real x in the **normal range**,

$$\boxed{\text{round}(x) = x(1+\delta), \quad |\delta| < \epsilon.}$$

An Important Idea

From the definition of δ we see

$$\text{round}(x) = x(1+\delta),$$

so the **rounded value** of an arbitrary number x in the **normal range** is **equal to** $x(1+\delta)$, where, regardless of the rounding mode,

$$|\delta| < \epsilon.$$

This is very important, because you can think of the stored value of x as **not exact**, but as **exact within a factor** of $1 + \epsilon$.
IEEE **single format numbers** are good to a factor of about $1 + 10^{-7}$, which means that they **have approximately** 7 **significant decimal digits**.

Special Case of Round to Nearest

For **round to nearest**, the **absolute** rounding error can be **no more than** half **the gap between** x_- **and** x_+. This means in IEEE single, for all $|x| \leq N_{\max}$:

$$|\text{round}(x) - x| \leq 2^{-24} \times 2^E,$$

and in general

$$|\text{round}(x) - x| \leq \frac{1}{2}\epsilon \times 2^E.$$

The previous analysis for **round to nearest** then gives for x in the **normal range**:

$$\text{round}(x) = x(1+\delta),$$

$$|\delta| \leq \frac{\frac{1}{2}\epsilon \times 2^E}{2^E} = \frac{1}{2}\epsilon.$$

Operations on Floating Point Numbers

IEEE standard requires corrected rounded operations:

- correctly rounded basic arithmetic operations ($+$, $-$, $*$, $/$);
- correctly rounded remainder and square root operations;
- correctly rounded format conversions.

Correctly rounded means rounded to fit the destination of the result, using rounding mode in effect.

IEEE Rule of Correctly Rounded Arithmetic

The exact result of an operation may **not** be a floating point number, e.g. the **multiplication** of two **24-bit** significands generally gives a **48-bit** significand.

When the result is **not** a floating point number, the IEEE standard requires that the computed result be the **correctly** rounded value of the **exact** result.

Let x and y be floating point numbers, and $\oplus,\ominus,\otimes,\oslash$ denote the **implementations** of $+,-,*,/$ on the computer. Thus $x \circledast y$ is the computer's **approximation** to $x * y$, where $* = +, -, \times, /$.

The **IEEE rule** for the four basic arithmetic operations is then precisely:

$$x \circledast y = \text{round}(x * y).$$

From our discussion of relative rounding errors, when $x * y$ is in the **normal range**,

$$\boxed{x \circledast y = (x * y)(1 + \delta), \quad |\delta| < \epsilon}$$

for **all** rounding modes. Note that $|\delta| \le \epsilon/2$ for the **round to nearest** mode.

Note. The computed result of a sequence of two or more arithmetic operations may **not** be the correctly rounded value of the exact result.

Square Root and Remainder

In additional to requiring that the basic arithmetic operations be correctly rounded, the IEEE standard also requires that correctly rounded remainder and square root operations be provided. The remainder operation x REM y is valid for finite x and nonzero y and produces $r = x - n * y$, where n is the integer nearest the exact value x/y.

Format Conversions

Numbers are usually input to the computer using some kind of high-level programming language, to be processed by a compiler or an interpreter.

Two different ways that a number such as $1/10$ might be input:

- Input the decimal string 0.1 directly, either in the program itself or in the input to the program. The complier or interpreter then calls a standard input-handling procedure which generates machine instructions to convert the decimal string to a binary format and store the correctly rounded result in memory or register.

- The integers 1 and 10 might be input to the program and the ratio 1/10 generated by a division operation. The input-handling procedure must be called to read the integer strings 1 and 10 and convert them to binary representation. Either integer or floating point format might be used to store these values, depending on the type of the variables used in the program, but the values must be converted to floating point format before the division operation computes the quotient 1/10.

Just as decimal to binary conversion is usually performed to input data to the computer and binary to decimal conversion is usually performed to output results when computation is completed.
The IEEE standard requires support for correctly rounded format conversions:

- Conversion between floating point formats.
- Conversion between floating point and integer formats.
- Rounding a floating point number to an integral value (not an integer format).
- Binary to decimal and decimal to binary conversion.

  There is an important requirement: if a binary single (say) format number is converted to at least 9 decimal digits (17 decimal digits) and then the converted from this decimal representation back to the binary single format (double format), the original number must be recovered.

Exceptional Situations

When a reasonable response to exceptional data is possible, it should be used.
The simplest example is **division by zero**. Two **earlier** standard responses:

- generate the **largest FPN** as the result.

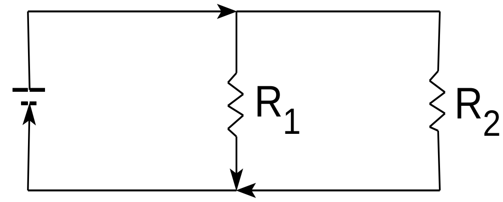
  Rationale: user would notice the large number in the output and conclude something had gone wrong.

  Disaster: e.g. $2/0 - 1/0$ would then have a result of 0, which is **completely meaningless**. In general the user might **not even notice** that any error had taken place.

- generate a **program interrupt**, e.g.
  "fatal error — division by zero".
  The burden was on the programmer to make sure that division by zero would **never** occur.

Example: Consider computing the **total resistance** in an electrical circuit with two resistors (R_1 and R_2 ohms) connected in parallel:
The formula for the **total resistance** is

$$T = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}.$$



What if $R_1 = 0$? If one resistor offers no resistance, **all** the current will flow through that and avoid the other; therefore, the total resistance in the circuit is **zero**. The formula for T also makes perfect sense **mathematically**:

$$T = \frac{1}{\frac{1}{0} + \frac{1}{R_2}} = \frac{1}{\infty + \frac{1}{R_2}} = \frac{1}{\infty} = 0.$$

The IEEE Standard Solution

Why should a **programmer** have to worry about treating division by zero as an exceptional situation here?

In **IEEE floating point arithmetic**, if the initial floating point environment is set properly: **division by zero** does not generate an interrupt but **gives an infinite result**, program execution continuing normally.

In the case of **the parallel resistance formula** this leads to a final **correct result** of $1/\infty = 0$, following the mathematical concepts **exactly**:

$$T = \frac{1}{\frac{1}{0} + \frac{1}{R_2}} = \frac{1}{\infty + \frac{1}{R_2}} = \frac{1}{\infty} = 0.$$

Other uses of ∞

We used some of the following:

$$\begin{aligned} a > 0 \quad &: \quad a/0 \to \infty \\ & \quad\;\; a * \infty \to \infty, \\ a \text{ finite} \quad &: \quad a + \infty \to \infty \\ & \quad\;\; a - \infty \to -\infty \\ & \quad\;\; a/\infty \to 0 \\ & \quad\;\; \infty + \infty \to \infty. \end{aligned}$$

But

$$\infty * 0, \quad 0/0, \quad \infty/\infty, \quad \infty - \infty$$

make no sense. Computing any of these is called an **invalid operation**, and the IEEE standard sets the result to NaN (**Not a Number**). In addition, a real operation with a complex result, e.g., the square root of a negative number, produce NaN. **Almost all** arithmetic operations with at least one NaN operand **also** produce NaN. Whenever a NaN is discovered in the output, the programmer **knows something has gone wrong**. An ∞ in the output may or may not indicate an error, depending on the context.

We stated before that there are two types of NaN: qNaN and sNaN. Their only difference is that sNaN generates interruption while qNaN does not. The application decides if it generates qNaN or sNaN. For instance, GCC C compiler always generates qNaN unless explicitly specified to behave the other way around.

Overflow and Underflow

Overflow is said to occur when

$$N_{max} < |\text{ true result }| < \infty,$$

where N_{max} is the **largest** normal FPN.

Two **pre-IEEE** standard treatments:

(i) Set the result to ($\pm$) N_{max}, or

(ii) Interrupt with an **error message**.

In IEEE arithmetic, the standard response depends on the **rounding mode**:

Suppose that the overflowed value is **positive**. Then

rounding model	result
round up	∞
round down	N_{max}
round towards zero	N_{max}
round to nearest	∞

Round to nearest is the **default** rounding mode and any other choice may lead to very misleading final computational results.

Underflow is said to occur when

$$0 < |\text{ true result }| < N_{min},$$

where N_{min} is the **smallest** normal floating point number.

Historically the response was usually: **replace the result by zero**.

In **IEEE arithmetic**, the result may be a **subnormal** number instead of zero. This allows results **much smaller** than N_{min}. But there may still be a significant loss of accuracy, since subnormal numbers have fewer bits of precision.

IEEE Standard Response to Exceptions

Invalid Opn.	Set result to NaN
Division by 0	Set result to $\pm\infty$
Overflow	Set result to $\pm\infty$ or $\pm N_{max}$
Underflow	Set result to ± 0, $\pm N_{\text{min}}$ or subnormal
Inexact	Set result to correctly rounded value

- The IEEE standard requires that an **exception** must be **signaled** by setting an associated **status flag**,
- The IEEE standard highly recommends that the programmer should have the option of either **trapping the exception** — providing special code to be executed when the exception occurs, or **masking the exception** — the program continues with the response of the table.
- A high level language may not allow trapping.
- It is usually best to rely on these standard responses.

Floating Point in C

In C, the type **float** refers to a **single precision** floating point variable.
e.g. **read** in a floating point number, using the standard input routine `scanf`, and **print** it out again, using `printf`:

```
main ()     /* echo.c:  echo the input */
{
    float x;
    scanf("%f", &x);
    printf("x = %f", x);
}
```

The 2nd argument `&x` to `scanf` is the **address** of `x`. The routine `scanf` needs to know **where** to store the value read.
The 2nd argument `x` to `printf` is the **value** of `x`.
The 1st argument `"%f"` to both routines is a **control string**.

The two standard **format codes** used for specifying floating point numbers in these control strings are:

- `%f`, for **fixed decimal** format;
- `%e`, for **exponential decimal** format.

The two format codes have **identical** effects in `scanf`, which can process input in a **fixed** decimal format (e.g. 0.666) or an **exponential** decimal format (e.g. `6.66e-1`, meaning 6.66×10^{-1}. However, different format codes have different effects when used with the output routine `printf` (see later).
The `scanf` routine calls a decimal to binary conversion routine to convert the input decimal format to internal binary floating point representation, and the `printf` routine calls a binary to decimal conversion routine to convert the binary floating point representation to the output

decimal format. Both conversion routines use the rounding mode that is in effect to correctly round the results.

The following results were for a Sun 4.
With input 0.66666666666666666666 :

```
#cc -o echoinput echo.c
#echoinput
0.66666666666666666666      (typed in)
x= 0.666667                 (printed out)
```

Using Different Output Formats in printf:

```
  Output format       Output
      %f               0.666667
      %e               6.666667e-01
      %8.3f            0.667
      %8.3e            6.667e-01
      %20.15f          0.666666686534882
      %20.15e          6.666666865348816e-01
```

The input is correctly rounded to about 6 or 7 digits of precision, so `%f` and `%e` print, **by default**, 6 digits after the decimal point.
The next two lines print to **less** precision. The 8 refers to the **total** field width, the 3 to the number of digits **after** the point.
In the last two lines about half the digits have **no significance**.
Regardless of the **output format**, the floating point variables are **always** stored in the **IEEE formats**.

Double or Long Float
Double precision variables are declared in C using **double** or **long float**. But changing **float** to **double** above:
{`double x; scanf("%f",&x); printf("%e",x);`}
gives `-6.392091e-236`. **Q:** Why?
`scanf` reads the input and stores it in **single format** in the first half of the **double word** allocated to x, but when x is **printed**, its value is read assuming it is **stored in double format**.
When `scanf` reads a **double** variable **we must use the format** `%lf` (for long float), so that it stores the result in **double** precision format.
`printf` **expects** double precision, and single precision variables are **automatically** converted to double before being passed to it. Since it always receives **long float** arguments, it treats `%e` and `%le` identically;
likewise `%f` and `%lf`, `%g` and `%lg`.

Programs to "test" if x is "zero"

Program 1

```
main() { float x; int n;
  n = 0; x = 1;   /* x = 2^0 */
  while (x != 0){
    n++;
    x = x/2;    /* x = 2^(-n) */
    printf("\n n= %d  x=%e", n,x); }
}
```

Program 2

```
main()
{ float x,y; int n;
  n = 0; y = 2; x = 1;     /* x = 2^0 */
  while (y != 1) {
    n++;
    x = x/2;     /* x = 2^(-n) */
    y = 1 + x;     /* y = 1 + 2^(-n) */
    printf("\n n= %d x= %e y= %e",n,x,y);
  }
}
```

Program 3

```
main()
{ float x,y; int n;
  n = 0;  x = 1;     /* x = 2^0 */
  while (1 + x != 1) {
    n++;
    x = x/2;     /* x = 2^(-n) */
    y = 1 + x;     /* y = 1 + 2^(-n) */
    printf("\n n= %d x= %e y= %e",n,x,y);
  }
}
```

Compile and run the above programs on various machines with different operating systems using different compliers. When do the programs terminate? Do you understand the results?