(IEEE) Floating-point representation

MATH 6610 Lecture 05

September 11, 2020

(Lecture 13 in Trefethen & Bau)

Finite representation of numbers

Numbers in decimal are represented as

$$34.1503,$$

Finite representation of numbers

Numbers in decimal are represented as

$$34.1503,$$

This actually means the more complicated expression

$$3 \times 10^{1} + 4 \times 10^{0} + 1 \times 10^{-1} + 5 \times 10^{-2} + 0 \times 10^{-3} + 3 \times 10^{-4}$$

Finite representation of numbers

Numbers in decimal are represented as

$$34.1503,$$

This actually means the more complicated expression

$$3 \times 10^{1} \; + \; 4 \times 10^{0} \; + \; 1 \times 10^{-1} \; + \; 5 \times 10^{-2} \; + \; 0 \times 10^{-3} \; + \; 3 \times 10^{-4}$$

The portion 34 is the "integer" part, and 1503 is the "fractional" part.

They are separated by the radix (point).

The base 10 is to be understood (implied) by context.

Finite representation of numbers

Numbers in decimal are represented as

$$34.1503,$$

This actually means the more complicated expression

$$3 \times 10^1 \;+\; 4 \times 10^0 \;+\; 1 \times 10^{-1} \;+\; 5 \times 10^{-2} \;+\; 0 \times 10^{-3} \;+\; 3 \times 10^{-4}$$

The portion 34 is the "integer" part, and 1503 is the "fractional" part.

They are separated by the radix (point).

The base 10 is to be understood (implied) by context.

Without context, the base b can be other positive integers:

$$b = 6 \quad \Rightarrow \quad 3 \times 6^1 \;+\; 4 \times 6^0 \;+\; 1 \times 6^{-1} \;+\; 5 \times 6^{-2} \;+\; 0 \times 6^{-3} \;+\; 3 \times 6^{-4}$$

Computer representations

Circuits typically detect the presence (1) or absence (0) of an electrical signal.

For this reason, base $b = 2$ is the standard computer format, e.g.,:

$$100010.0010011001111$$

(b=2 ⇒ "binary")

is approximately equal to the decimal 34.1503.

Computer representations

Circuits typically detect the presence (1) or absence (0) of an electrical signal.

For this reason, base $b = 2$ is the standard computer format, e.g.,:

$$\textcolor{blue}{100010}\textcolor{red}{.}0010011001111$$

is approximately equal to the decimal 34.1503.

Computers store this *binary* representation of these numbers, and each digit is a "bit".

8 bits = 1 "byte" (Above: 19 bits)

Computer representations

Circuits typically detect the presence (1) or absence (0) of an electrical signal.

For this reason, base $b = 2$ is the standard computer format, e.g.,:

$$100010.0010011001111$$

is approximately equal to the decimal 34.1503.

Computers store this *binary* representation of these numbers, and each digit is a "bit".

8 bits = 1 "byte"

Binary representations must store the integer part, the fractional part, and typically also a sign.

Fixed point vs floating point

Fixed-point representations have a fixed radix point, with the size of the fractional part predetermined:

$$100010.\quad \underbrace{0010011001111}_{\text{fixed number of entries}}$$

Then the fractional precision of this representation is fixed for any number.

This *truncation* of finite representations is one of the main challenges to address in numerical computations.

Above: 6 bits for integer part
13 bits for fractional part.

Fixed precision: can't represent numbers smaller than 2^{-13}
(or larger than 2^{7})

Fixed point vs floating point

Fixed-point representations have a fixed radix point, with the size of the fractional part predetermined:

$$100010.\quad \underbrace{0010011001111}_{\text{fixed number of entries}}$$

Then the fractional precision of this representation is fixed for any number.

This *truncation* of finite representations is one of the main challenges to address in numerical computations.

Fixed-point representation has a restriced range of precision (pre)defined by the size of the fractional part.

Fixed point vs floating point

Fixed-point representations have a fixed radix point, with the size of the fractional part predetermined:

$$100010.\ \underbrace{0010011001111}_{\text{fixed number of entries}}$$

Then the fractional precision of this representation is fixed for any number.

This *truncation* of finite representations is one of the main challenges to address in numerical computations.

Fixed-point representation has a restriced range of precision (pre)defined by the size of the fractional part.

Floating-point representations allow the radix to float:

"radix shift"

$$1.\underbrace{0001000100110011111}_{\text{fixed number of entries}} \quad + \quad \text{exponent} \quad 5$$

e.g. exponent $\cancel{6}$ $\Rightarrow$

The exponent encodes which exponent the radix is aligned with. (Above: 5)

Floating-point representations

encodes integer + fractional part

Generally speaking, floating-point representations store:

$$\color{blue}{\text{significand}} + \color{red}{\text{exponent}} + \text{sign}$$

The significand combines the integer and fractional parts of the number.

The exponent encodes the location of the radix.

Floating-point representations allows for a (much) larger operating range than fixed-point representations.

exponent

In decimal: $\pm I \cdot 10^{E}$

significand

Floating-point representations

Generally speaking, floating-point representations store:

$$\text{significand} + \text{exponent} + \text{sign}$$

The significand combines the integer and fractional parts of the number.

The exponent encodes the location of the radix.

Floating-point representations allows for a (much) larger operating range than fixed-point representations.

The most popular representation is the IEEE 754 standard, defining various formats:

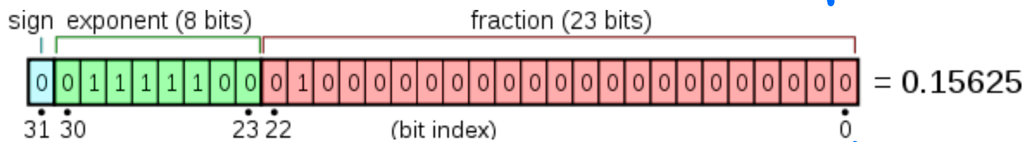
- Binary 16 # of bits used for storage.
- Binary 32 ("single precision")
- Binary 64 ("double precision") ☆ (most of the time)
- ⋮

Most high-level scientific computing languages use double precision as default.

Floating-point details

32 bits (single precision)

These standards define bit allocation:



Source: `https://en.wikipedia.org/wiki/File:Float_example.svg`

significand

exponent: encoded as a signed integer.

significand: encoded as an unsigned integer

DP: exponent: 11 bits

significand: 52 bits

In DP: what's the largest # I can represent?

$I \cdot 2^E$, Exponent E: 11 bits

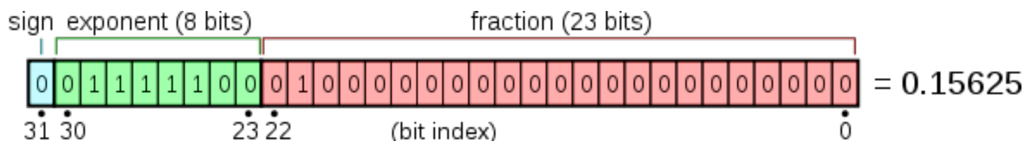
Encodes integers between -2^{10} and 2^{10}.

Ignoring significand: 2^E can be as large as $2^{2^{10}} = 2^{1024}$

Similarly: smallest number (in magnitude) is $2^{-2^{10}} = 2^{-1024}$.

Floating-point details

These standards define bit allocation:



Source: `https://en.wikipedia.org/wiki/File:Float_example.svg` The number of bits allocated to the exponent indicates the rounding precision of the format.

Machine precision or machine epsilon is the maximum relative rounding error due to finite representation,

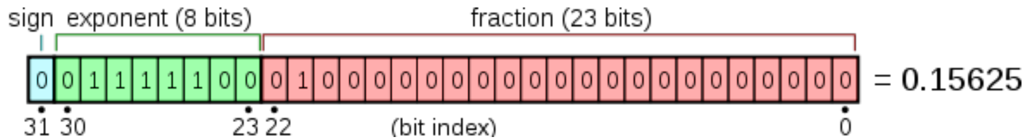
$$\left| \frac{x - \mathrm{fl}(x)}{x} \right| .$$

floating-point representation of x.

Roughly speaking, machine epsilon is also the largest ϵ such that $1 + \epsilon$ is rounded to 1.

Floating-point details

These standards define bit allocation:



Source: `https://en.wikipedia.org/wiki/File:Float_example.svg` The number of bits allocated to the exponent indicates the rounding precision of the format.

Machine precision or machine epsilon is the maximum relative rounding error due to finite representation,

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right|.$$

Roughly speaking, machine epsilon is also the largest ϵ such that $1 + \epsilon$ is rounded to 1.

base^{-E}, $-E$: smallest integer representable by significand.

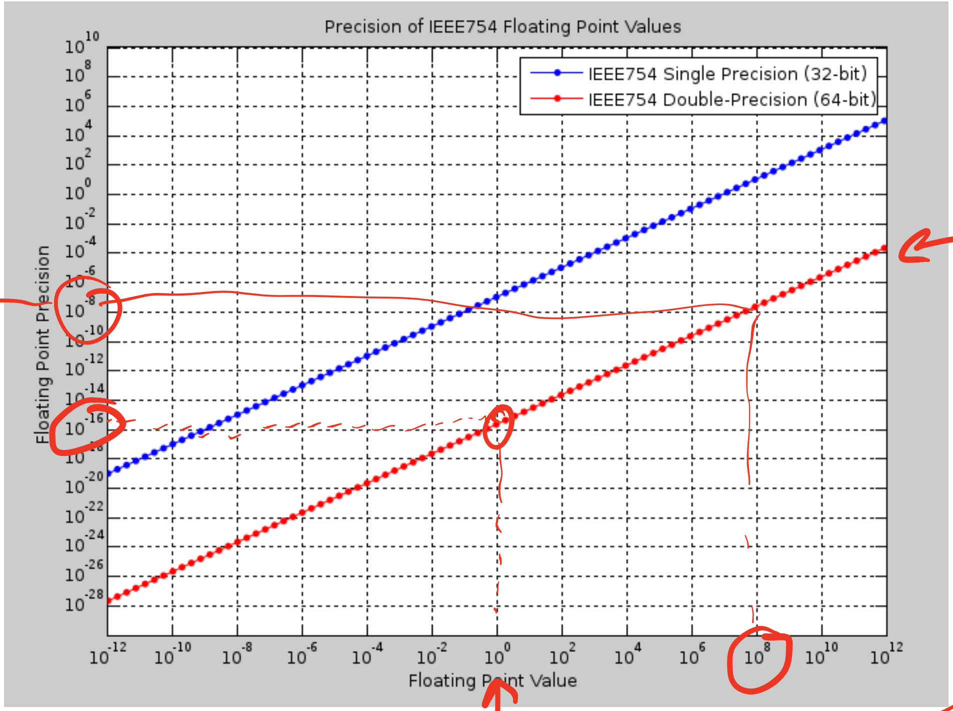
The roundoff or truncation error associated with a floating-point format is essentially equal to $\text{base}^{-\#(\text{exponent digits})}$.

Machine precision

Floating-point representation ensures that there is not really an absolute error committed by computer representations, there is only a relative error.

$\approx 10^{-8}$

$\approx 10^{-16}$



1

Source: `https://en.wikipedia.org/wiki/File:IEEE754.svg`

Roundoff and computing

The rounding truncation imparted by finite representations requires attention to how algorithms are implemented. E.g.,

- Implementation of the formula $\underbrace{\sqrt{1+x^4}-1}$ for small positive x.
- Evaluation of e^x for $x < 0$.
- Evaluation of $\frac{f(x+h)-f(x)}{h}$ for small h.

$f(x) := \sqrt{1+x}$

Taylor Series around $x=0$

$f(x) \approx 1 + \frac{x}{2}$

$\Rightarrow \sqrt{1+x^4} \approx 1+\frac{x^4}{2}$ for x^4 close to 0.

$\underbrace{\sqrt{1+x^4}}-1 \approx \frac{x^4}{2}$ for x^4 close to 0.