# Floating Point Numbers

# (From http://pages.stat.wisc.edu/~bwu62/lec\_fp.html)

## Goals

1. Develp basic understanding of floating point numbesr
2. Develop some basic notions of errors, and their consequences

## References

Optional texts:

1. ["What Every Computer Scientist Should Know About Floating-Point Arithmetic"](https://www.itu.dk/~sestoft/bachelor/IEEE754_article.pdf) , David Goldberg, 1991
2. ["Anatomy of a floating point number"](https://www.johndcook.com/blog/2009/04/06/anatomy-of-a-floating-point-number/) , John D Cook, 2009
3. ["Accuracy and Stability of Numerical Algorithms"](http://ftp.demec.ufpr.br/CFD/bibliografia/Higham_2002_Accuracy%20and%20Stability%20of%20Numerical%20Algorithms.pdf) , Hingham, 2002

# Outline

1. Positional Numeral System
2. Floating point format
3. Errors
4. Operations, fundametal axiom

## Positional Numeral System

Ingredients   
1. A base $\beta\in\br{2,3,4...}$   
2. A sequence of digits, $d\_0,d\_1,...,\in\br{0,1,...,\beta-1}$   
3. Exponent $e\in\Z$

representation: $d\_0 d\_1 d\_2 ... \times \beta^e$

#### Homework

1. Can all nonnegative real numbers be represented in such a manner for an arbitrary base $\beta\in\br{2,3,...}$ ?
2. Suppose $e=-1$, what are the range of numbers that can be represented for an arbitrary base $\beta$ ?
3. Characterize the numbers that have a unique representation in a base $\beta$ (iff condition+proof)

## Floating Point Format

**Definition.** A floating point is one that can be represented in a base $\beta$ with a fixed digits $p$ (precision), and with exponent $e\in\br{e\_{min},e\_{max}}$

**Example.** $\beta=10, p=3, e\_{min}=-1, e\_{max}=1$. Then, $0.1$ can be represented as

d0 d1 d2 e

0 0 1 1

0 1 0 0

1 0 0 -1

**Definition.** A normalized floating point number has $d\_0\neq0$

**Example.** The total number of values that can be prepresented in normalized floating point format with $\beta,p,e\_{min},e\_{max}$?

**Answer.** $\beta^{p-1}(\beta-1)(e\_{max}-e\_{min}+1)$ ?

#### Homework

1. Take decimal number, base, precision, return closest approximation of normalized floating point with vector of digits, exponent
2. given base, precision, $e\_{min},e\_{max}$, list all normalized floating point representable

## Floating Point Approximations

A IEEEE 16-bit (Half-precision standard) has   
- 1 bit for sign   
- 5 bits exponent 00000 and 11111 reserved for 0 and $\infty$, 30 exp   
- 11 significant (only 10 stored since leading bit must be 1 since normalized)

**Question.** What are smallest and largest positive numbers that can be represented?

**Answer.**   
- Exponents: $2^5 = 32$ exponents. Minus $0$ and $\infty=30$ possible exponents, $-14,..,15$.   
- Smallest non-normalized: $(0+0(2^{-1})...0(2^{-9})+1(2^{-10}))\times2^{-14}=2^{-24}\approx5.96\times10^{-8}$   
- Smallest normalized: $(1+0(2^{-1})...0(2^{-9})+0(2^{-10}))\*2^{-14}=2^{-14}\approx6.1\times10^{-5}$   
- Largest finite: $(1+1(2^{-1})...1(2^{-9})+1(2^{-10}))\times2^{15}=65504$

**Note.** Julia has implementation allowing non-normalized numbers, and somehow allows 11-bits of digits.

A IEEEE 64-bit (Double-precision standard) has   
- 1 bit for sign   
- 11 bits exponent 0...0 and 1...1 reserved for 0 and $\infty$   
- 11 significant (only 10 stored since leading bit must be 1 since normalized)

#### Homework

* (Look up 64-bit standard) What is smallest normalized and non-normalized positive value? Largest noninfinite value?
* Is there a general formula for determining the largest positive value for a given base $\beta$, precision $p$, and $e\_{max}$?
* Verify the smallest non-normalized positive number that can be represented? Longest noninfinite?

## Errors

* Units in the Last Place (ULP) eps
* Absolute+Relative Error

$\fl:\R\_{20}\to S$   
- Absolute Error: $|\fl(z)-z|$   
- Relative Error: $\tfrac12|\fl(z)-z|$

**Lemma.** If $z$ has exponent $e$, the max absolute error is $\tfrac12\beta^{e-p+1}$

**Proof.** : **homework**

**Lemma.** If $z$ has exponent $e$, the relative error is bounded above by $\tfrac12\beta^{1-p}$

**Proof.** if $z$ has exponent $e$, $\beta^e\in\Z$,   
$\tfrac12|\fl(z)-z|\leq\frac{\beta^{e-p+1}}{2\beta^e}=\tfrac12\beta^{1-p}~~~$ $\leftarrow$(eps)

$\varepsilon=\tfrac12\beta^{1-p}$ called **machine epsilon**

#### Homework

What happens if we consider negative numbers?

## Operations

#### Fundamental Axiom

For any $\operatorname{op}$ of the 4 arithmetic operations $(+-\times\div)$ we have the following bound   
$$\fl(x \op y)=(x\op y)(1+\delta)$$

where $|\delta|\leq U$ where $U$ is commonly $2\varepsilon$

Matrix Storage $A\in\R^{m\times n}$   
$|\fl(A)-A|\leq U|A|$

Homework question:   
1. Show that $||A||$ is equal to maximum of the $\ell'$ norms of the columns of $A$   
2. Show that $||A||\_\infty$ is equal to maximum of the $\ell'$ norms of the rows of $A$   
3. $||\fl(A)-A||\_p\leq U||A||\_p$

#### Dot Product

Let $X,y\in\R^n$, $\fl(x'y)-x'y$

$x'y=\sum x\_iy\_i$ pseudo-code:

fl(x'y)

length(x) == length(y) || error("wrong dim")

s = 0

for i in 1..n

s += x[i]\*y[i]

return s

**Lemma** Let $x,y\in\R^n$ $nU<0.01$   
$|\fl(x'y)-x'y|\leq 1.01 nU|x|'|y|$