1.2 Round-off Errors and Computer Arithmetic

1. **Quote.** "Approximating mere numbers, the task of floating point arithmetic, is indeed a rather small topic and maybe even a tedious one. The deeper business of numerical analysis is approximating unknowns not knowns. Rapid convergence of approximations is the aim."

   (Nick Trefethen, Oxford University (1955-))

2. **Why study rounding errors?**

3. **The Main Issue -** Real Numbers versus storage of a finite number of digits.

4. **Notation.**

   (a) A number consists of an integer part and a fractional part.

   (b) The base of a number is the number of unique digits used to represent numbers in a positional numeral system.

   (c) If it is not clear from context, the base will be indicated with a subscript following the number. For example $(24)_{10}$ is 24 in base 10.

5. **Example.** Meaning and representations of $(234.1)_{10}$ and $(1010.01)_2$

6. **Remarks**

1) Computers use the binary (base 2) system.

2) The octal (base 8) and hexidecimal (base 16) following naturally from this.

7. **Machine Number Properties.**

(a) Each machine number has an integer **base**, $\beta > 1$.

(b) Each machine number has an integer **precision**, $k \geq 1$.

(c) Each machine number has an integer exponent, n, which falls within a given **range**, $n_{min}, n_{min+1}, ..., n_{max-1}, n_{max}$.

8. **Floating Point Numbers.**

9. The floating point number system is the set of numbers

$$x = (\pm 0.d_1 d_2 d_3 d_4 ... d_k)\beta^n$$

where

(a) $d_1 \in \{1, 2, 3, ..., \beta - 1\}$

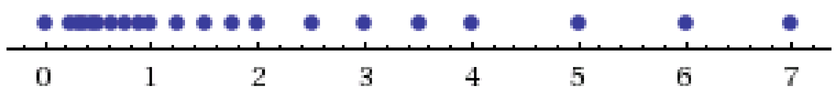
(b) $d_2, d_3, ..., d_k \in \{0, 1, 2, ..., \beta - 1\}$

(c) $n_{min} \leq n \leq n_{max}$

(d) The number $d_1 d_2 d_3 d_4 ... d_k$ is called the mantissa

10. **Example.** What numbers can be represented with $\beta = 2$, $k = 3$, $n_{min} = -1$, and $n_{max} = 3$?

Number line:



11. **Example.** When representing $(1.75)_{10}$ in base 2 what are the mantissa and exponent?

12. **The IEEE Standard (754) for double precision floating point numbers.**
    (a) base is 2
    (b) Exponent is 11 bits in offset binary
    (c) Mantissa is 52 bits (which equates to a k=53 because we don't need to store d_1 since we know it is a one).

13. **Storage Example.**

14. **Example.** Finding the smallest and largest non-zero numbers that can be represented.

15. Use the command **realmin** and **realmax** in Matlab and confirm that you get the following: 2.2251e-308 and 1.7977e+308

16. **Overflow and Underflow.** Refer to demo code Overflow.m which is in Canvas on the lecture notes page.

    Overflow occurs when numbers get bigger than realmax and results in the program terminating or the answer being set to infinity.

    Underflow occurs when numbers get smaller than realmin and results in the answer being set to zero.

17. **Example** Calculating the determinant of an N by N matrix.

18. **Rounding and Chopping** How do we represent a given number, x, in our floating point number system?

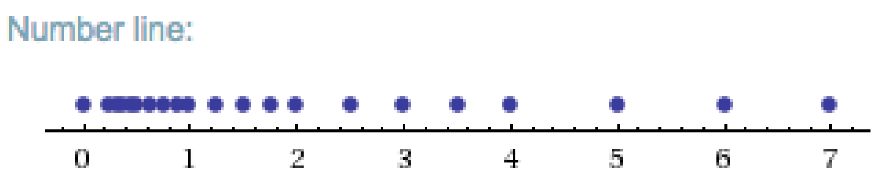
19. **Roundoff Error.** Let p^* be an approximation of a number p. Then,

The **Absolute Error** is $|p - p^*|$

The **Relative Error** is $\dfrac{|p-p^*|}{|p|}$, given $p \neq 0$

p^* approximates p to k signficant digits in base β if $\dfrac{|p-p^*|}{|p|} \leq \frac{1}{2}\beta^{1-k}$

20. **Example** What is the maximum relative error for IEEE double precision floating point? What does Matlab say?



21. **Floating Point Arithmetic**

    Computers must be able to perform the basic arithmetic operations of addition, subtraction, mutliplication and division.

22. **Example** Compute the relative error when adding $\frac{2}{7}$ and $\frac{1}{3}$ using 5 digit chopping arithmetic.

23. **3 important issues that arise when doing arithmetic with finite precision.**
    (a) Cancellation Error - this happens when you subtract nearly equal numbers
    (b) Amplification of Round-off error (this is when your round-off error is multiplied by a factor greater than 1)
    (c) Accumulation of Round-off error (this happens when round-off errors are added together often as the result of performing many operations)

24. **Example** Cancellation Error

25. **Example** Amplification of Round-off Error

26. **Example** Accumulation of Round-off Error