MATH 466 Numerical Linear Algebra with Applications
MATH 462 Engineering Numerical Analysis

Jun Liu

Department of Mathematics and Statistics

Southern Illinois University Edwardsville

To be covered Sections:

1. Sec 0.1 Evaluating a Polynomial
2. Sec 0.2 Binary Numbers
3. Sec 0.3 Floating Point Representation of Real Numbers
4. Sec 0.4 Loss of Significance

Key ideas:

- **Round-off error** are everywhere, but you may not SEE it!
  IEEE 754 (64 bits) has a precision: $eps = 2^{-52} \approx 2.22 \times 10^{-16}$.
- Mathematically Equivalent $\neq$ Numerically the SAME.
- Small errors can be a disaster if magnified MANY times!
- Different algorithms lead to VERY different results in:
  efficiency (CPU times) and **accuracy** (approximation errors).

Best way to evaluate $P(1/2)$?

$$P(x) = 2x^4 + 3x^3 - 3x^2 + 5x - 1.$$

- Direct approach: 10 *, 4 +/- =14 operations.
- Store/reuse powers of 1/2: 7 *, 4 +/- =11 operations.
- Nested (Horner's) Multiplication: 4 *, 4 +/- =8 operations.

$$P(x) = -1 + x * (5 + x * (-3 + x * (3 + x * 2))).$$

A polynomial of degree d needs d * and d +/- operations.

Sec 0.1 Horn's method

A general polynomial of degree 4:

$$P_4(x) = c_1 + x(c_2 + x(c_3 + x(c_4 + xc_5)))$$

its shifted version for given base points $b_i, i = 1, 2, 3, 4$:

$$Q_4(x) = c_1 + (x - b_1)(c_2 + (x - b_2)(c_2 + (x - b_3)(c_4 + (x - b_4)c_5)))$$

```
function y=nest(d,c,x,b)%Program 0.1 (nest.m)
if(nargin<4) %input arguments less than 4
  b=zeros(d,1); %set shift vector to zero
end
y=c(d+1);
for i=d:-1:1 %decreasing by 1
y = y.*(x-b(i))+c(i);% compatible to a vector x
end
```

Run in Command Window: outputs make sense?

$>>$ nest(4,[-1 5 -3 3 2],1/2,[0 0 0 0])

$>>$ nest(4,[-1 5 -3 3 2],[-2 -1 0 1 2])

1. $P(1.00001) = ? Q(1.00001) = ?$ Which one is more accurate?

$$P(x) = 1 + x + x^2 + \cdots + x^{50} = \frac{x^{51} - 1}{x - 1} = Q(x)$$

2. $P(1.00001) = ?$ $Q(1.00001) = ?$ What is the error?

$$P(x) = 1 - x + x^2 - \cdots + x^{98} - x^{99} = \frac{1 - x^{100}}{1 + x} = Q(x)$$

Try different values of x to find the maximum possible error?

- Decimal number (base 10):
  $(466)_{10} = 4 * 10^2 + 6 * 10^1 + 6 * 10^0$
- Binary number (base 2):
  $(100)_2 = 1 * 2^2 + 0 * 2^1 + 0 * 2^0$
- Decimal to Binary: $(53.7)_{10} = (110101.1\overline{0110})_2$, how?

  Integer part: $(53)_{10} = (110101.)_2$
  (Divde by 2 repeatedly, record remainder to the left)

  Fractional part: $(0.7)_{10} = (.1\ 0110\ 0110\ 0110...)_2 = (.1\overline{0110})_2$
  (Multiply by 2 repeatedly, record the integer parts to the right).
  Here overbar line $\overline{0110}$ denotes infinitely repeated pattern.

- Binary to Decimal: match position with power
  Integer part: $(10101)_2 = 2^4 + 2^2 + 2^0 = (21)_{10}$
  Fractional part: $(.1011)_2 = 2^{-1} + 2^{-3} + 2^{-4} = (\frac{11}{16})_{10}$
- Repeating case: $x = (.\overline{0110})_2$ satisfies
  $(2^4 x - x) = 1011.\overline{0110} - .\overline{0110} = (1011)_2 = (11)_{10} \to x = 11/15.$
- How about the case $x = (.10\overline{101})_2$?
  $y = 2^2 x = 10.\overline{101} = (2)_{10} + (.\overline{101})_2 = (2 + 5/7)_{10} \to x = 19/28.$