CS314 Class Note

Week 1 (6/13-6/17)

by Yulai Liu liu264@purdue.edu

Floating/Fixed Point Representation

N.F = Interger.Fraction

……$b_2b_1b_0.b_{-1}b_{-2}b_{-3}$…….

- **Binary to Decimal:**

  Decimal = …… $b_2 \times 2^2 + b_1 \times 2^1 + b_0 \times 2^0 . b_{-1} \times 2^{-1} + b_{-2} \times 2^{-2} + b_{-3} \times 2^{-3}$ .....

  Example:

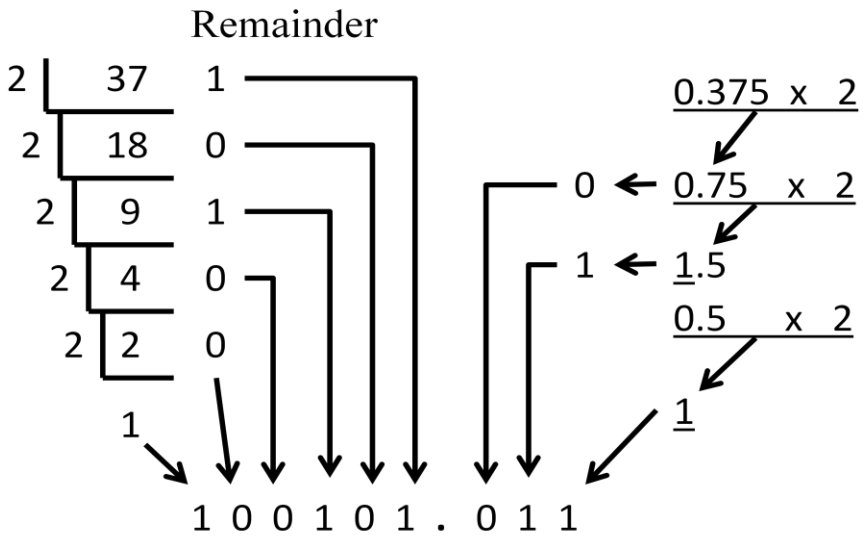
  101.011 -> $1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3}$

  = 4 + 0 + 1 + 0 + 0.25 + 0.125 = 5.375

- **Decimal to Binary:**

  Example:

  $37.375_{(10)}$ -> $100101.011_{(2)}$



- **Fixed Point:**

  The number of bits for the integer part and the fraction part is fixed.

  Pros: Simple and fast arithmetic operations

  Cons: Little Range and bad precision

- **Floating Point:**

  The point **"."** Moves around to keep a desired precision.

  Numbers are stored in scientific form

  Example:

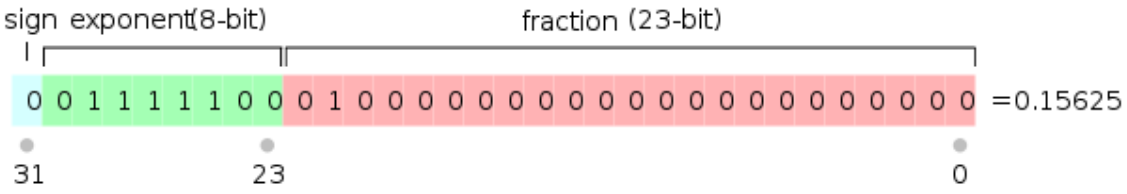
  $101.1101_{(2)} => 1.011101_{(2)} \times 2^2$

  Fraction: 0.011101

  Exponent: 2

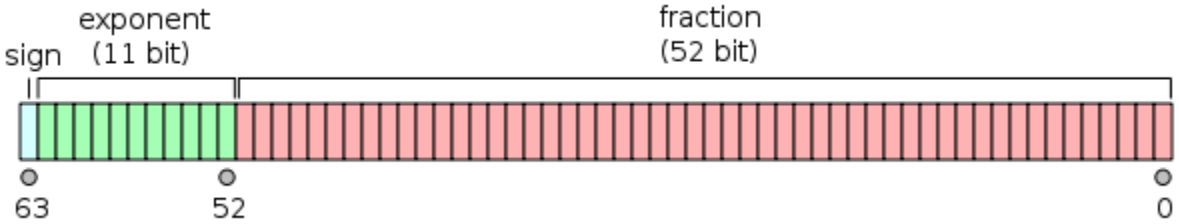
- **2 Floating Point Representation:**

  - Single Precision (Float) – 4 bytes (32 bits)



  Binary = $(-1)^{sign} \times 1.fraction \times 2^{exponent-127}$

  - Double Precision (Double) – 8 bytes (64 bits)



  Binary = $(-1)^{sign} \times 1.fraction \times 2^{exponent-1023}$

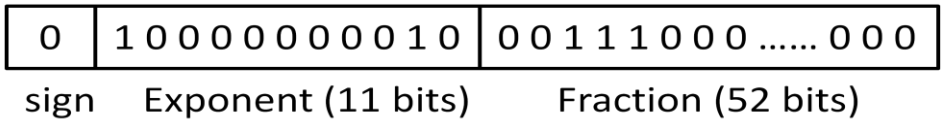
Example:

$9.75_{(10)} \Rightarrow 1001.11_{(2)}$

Normalize $1001.11_{(2)} \Rightarrow 1.00111_{(2)} \times 2^3$

Exponent = $3 + 1023 = 1026 = 2^{10} + 2^1$

Double:



More on **Standard for Binary Floating-Point Arithmetic**

http://en.wikipedia.org/wiki/IEEE_754-1985

Computational Error

Absolute Error = $| p - p^* |$ (p^* = approximation of p)

Relative Error = $\frac{| p - p^* |}{| p |}$ ($p \neq 0$)

Propagation of Errors $p^* = p + e_p$

$q^* = q + e_q$

Sum: $p^* + q^* = \underbrace{p + q}_{\text{Real Value}} + \underbrace{e_p + e_q}_{\text{New Error}}$

The new error equals the sum of the initial errors.

Product: $p^* + q^* = \underbrace{pq}_{\text{Real Value}} + \underbrace{pe_q + qe_p + e_pe_q}_{\text{New Error}}$

The new error amplified by the magnifying of p and q.

Stability:

Stable if small initial errors stay small

Unstable if small initial errors get larger at each step

Solution of Nonlinear Equations

f(x) = Ø, use "iteration" to solve

(It's usually the intersection between y = g(x) and y = x)

Fix Point: A fixed point of g(x) is a real number p such that p = g(p)

Fixed Point Iteration: $P_{n+1} = g(P_n)$

Iteration

P_0

$P_1 = g(P_0)$

$P_2 = g(P_1)$

.

.

.

.

.

Until $|P_k - P_{k-1}| < \varepsilon$

For some k, ε

Fixed Point Theorem

If $|g'(x)| < 1$ for all $x \in [a,b]$ and $g(x) \in [a,b]$,

then iteration $P_n = g(P_{n-1})$ converge to a fixed point $P \in [a,b]$.

If $|g'(x)| > 1$, then not converge.

Example:

$x^2 - 2x + 1 = 0$

$x^2 + 1 = 2x$

$x = \frac{x^2+1}{2}$

$x_0 = 0$

$x_1 = 0.5$

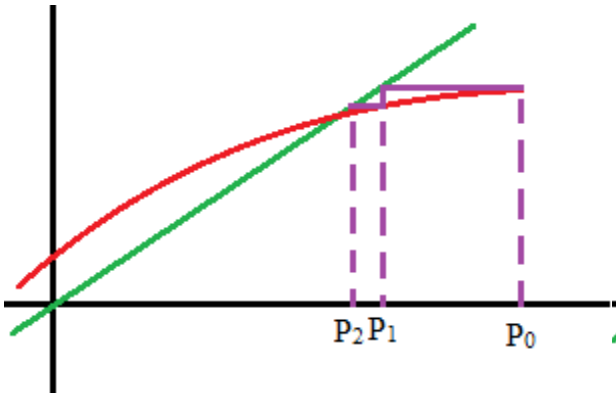
.

.

.

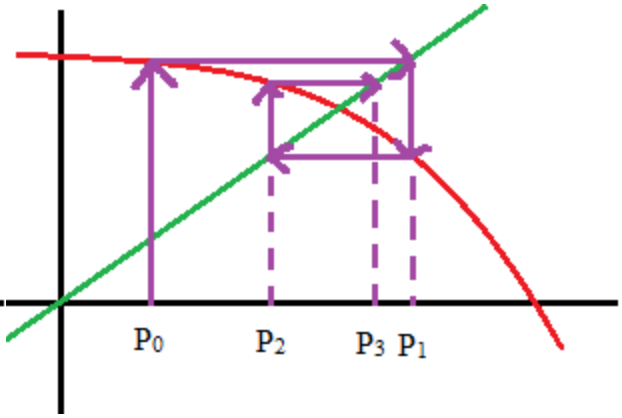
$x_n = 1$

Types of convergence



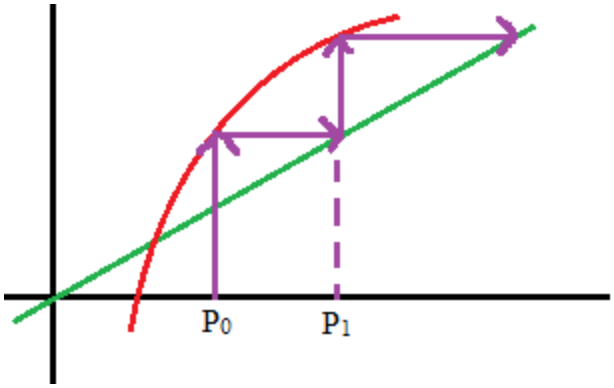
Monotone Convergence

$0 < g'(p) < 1$



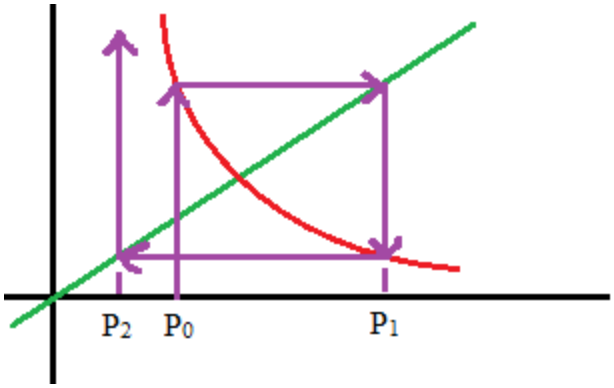
Oscillating Convergence

$-1 < g'(p) < 0$



Monotone Divergence

$1 < g'(p)$ (never converge)



Oscillating Divergence

$g'(p) < -1$ (never converge)

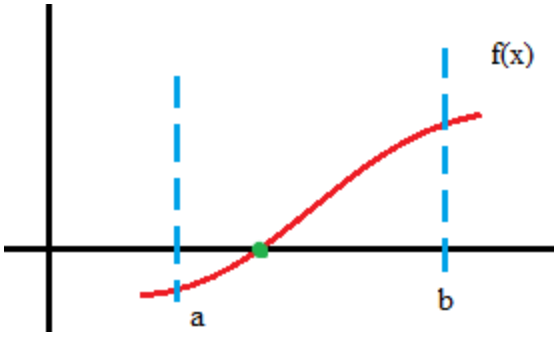
Bisection Method

Find solution of f(x) = 0 (similar to binary search)

f(x) is contiguous, so there is a f(x) = Ø

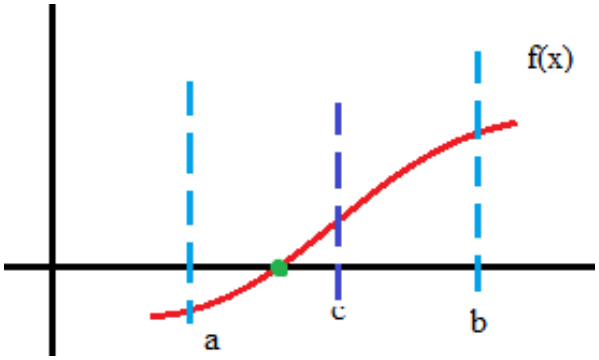
Get middle point c = (a+b) / 2

If a,c opposite sign, range change to [a,c]



If c,b opposite sign, range change to [c,b]

If f(c) = 0, c is the solution



Stop when $| a - b | < \varepsilon$

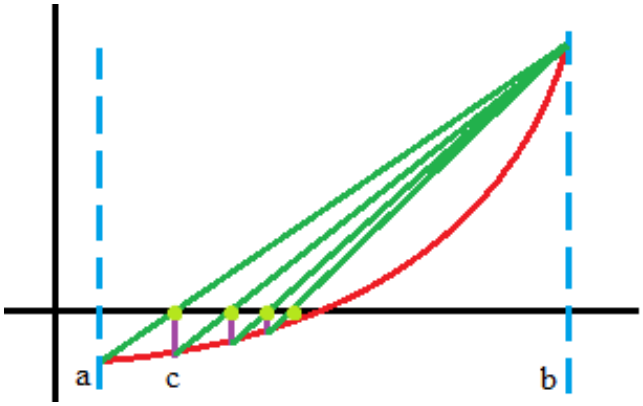
Assume finding solution takes step 'n'

$| a_n - b_n | = | a - b | / 2^n < \varepsilon$

$\log_2 (| a - b | / \varepsilon) < n$

False Position Method

- Created because Bisection Method is too slow
- Need [a,b] with a zero in between



Line between (a,f(a)) , (b,f(b)) with c intercept x axis in between.

If a,c opposite sign, range change to [a,c]

If c,b opposite sign, range change to [c,b]

Assume slop 'm'

m = (f(b) – f(a)) / (b-a)

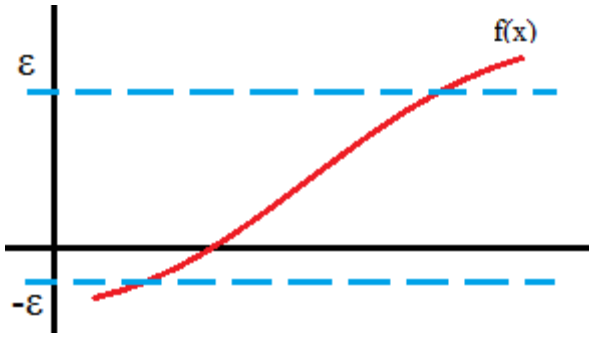
m = (f(b) – 0) / (b-c)

$$\Rightarrow \quad c = b - \frac{(b-a)\,f(b)}{f(b) - f(a)}$$

Check for Convergence

1. Horizontal Convergence

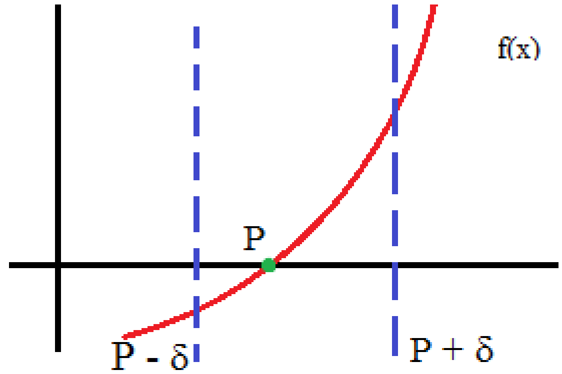
   $| f(x) | < \varepsilon$



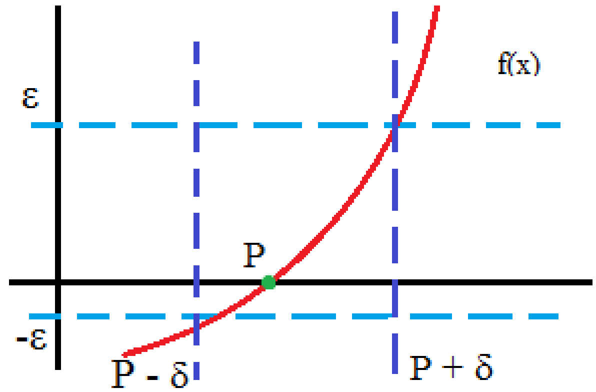
2. Vertical Convergence

   $| x - p | < \delta$ $(f(p) = 0)$

   (don't know where p is)

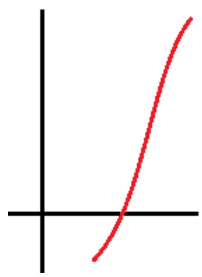


3. Both Vertical and Horizontal Convergence

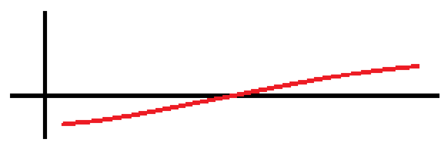


Troublesome Functions

- If graph is steep to a root (P, ø), can get a good precise solution (Well Conditioned)



- If is shallow (Ill Conditioned), root finding may have a few significant digits

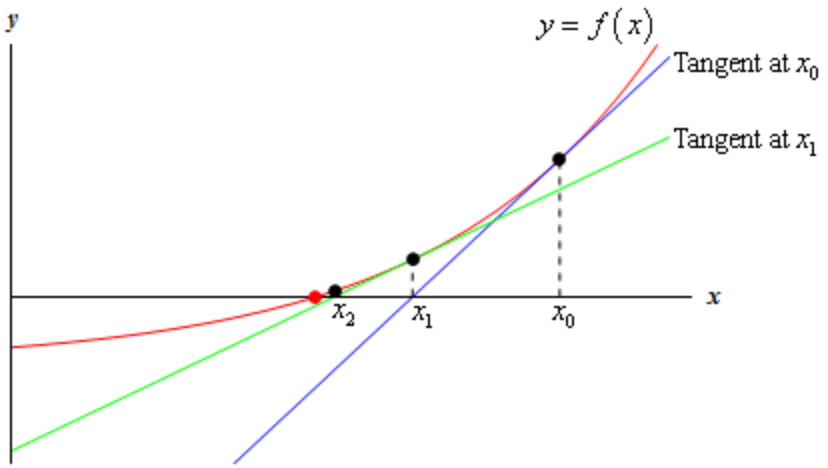


Newton Raphson

If f(x), f'(x), f''(x) are continuous

$m = f'(P_0)$

$m = (f(P_0) - 0) / (P_0 - P_1)$ => $$P_1 = P_0 - \frac{f(P_0)}{f'(P_0)}$$



Graph from *Paul's Online Math Notes*

Taylor Expansion around P_0 is $f(x) = f(P_0) + f'(P_0)(x- P_0) + [f''(P_0)(x- P_0)^2] / 2!$ ……

Newton Raphson method approximates f(x) by using Taylor Expansion to first 2 terms:

$f(x) = f(P_0) + f'(P_0)(x- P_0)$

If $x = P_1$, $f(P_1) \approx f(P_0) + f'(P_0)(P_1- P_0)$

Use 3 terms of Taylor Expansion to find improved Newton Raphson that approximates the function using a parabola (quadratic equation) and a second derivative.

<u>Example:</u>

$\sin(x) = ½$

$f(x) = \sin(x) – ½ = 0$

$P_0 = 0$

k	P_k	$f(P_k)$	$f'(P_k)$	P_{k+1}
1	0	-½	$\cos(0) = 1$	$P_{k+1} = P_k - f(P_k) / f'(P_k) = 0 + ½ = ½$
2	½	$\sin(½) – ½ = -0.02$	$\cos(½) = 0.8775$	$P_2 = 0.522$
……				

It's fast but needs derivative of the equation.

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Derivative can be obtained numerically:

$$\Delta x = \varepsilon$$