

Numerical Algorithms

* What

* Contents

- Solution of equation of single variable
- Linear System of eq.
- Matrix decomposition
- Optimization

Real Numbers \mathbb{R}

→ Analytical

→ Graphical

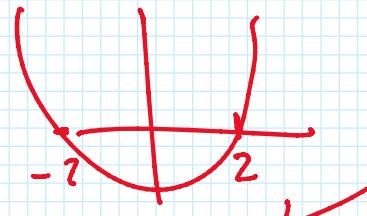
→ Numerical
initial guess

↓ Fast

Approximate
solution

$$y = \boxed{x^2 - 4} = 0 \\ x^2 = 4 \\ x = \pm \sqrt{4} \leftarrow \pm 2$$

1.9999 --



①, ②, $\rightarrow \infty x$

$$0.4 + 0.8 = 1.2 ?$$

$$1.2 - 0.8 = 0.4 ?$$

$$1.2 - 0.4 = 0.8 ?$$

$$1.2 - 0.4 - 0.8 = 0 ?$$

}

$$0.1 + 0.1 = 0.2 ?$$

$$0.1 + 0.1 + 0.1 = 0.3 ?$$

}

Decimal System

$$\begin{array}{r}
 \textcircled{1} \textcircled{5} \textcircled{3} \\
 = \boxed{100} + \boxed{50} + \boxed{3} \\
 = 100 + 50 + 3 \\
 = 153
 \end{array}$$

7531

Binary System

<u>many</u>	<u>system</u>	0.1
7 2	6 2	5 2
4 2	3 2	2 2
1 2		0
0 0 0 0 0 2 1 0		

$$\begin{array}{r} 1 \quad 1 \times 2 \\ 1 \times 2 \quad + \\ \hline \end{array}$$

$$= 2+4 = \boxed{6}$$

Quiz

- $(0000\ 0000)_2 = 0$ $(\underline{29})_{10} = (?)_2$
- $(0000\ 0001)_2 = 1$ $(\underline{231})_{10} = (?)_2$
- $(0000\ 1111)_2 = \underline{15}$
- $(\underline{1000}\ 0000)_2 = 128$
- $(1111\ \underline{1111})_2 = \boxed{255}$ $29 = \boxed{16} + \boxed{?}$

$$\begin{array}{r} \underline{29} \\ - \end{array} \quad \begin{array}{r} 16 \\ 8 \\ 4 \\ 1 \end{array} \quad \begin{array}{r} + 3 \\ + 5 \\ - 1 \end{array}$$

00011101

$\xrightarrow{10^{-1}}$

Decimal

3.14

$$3 \times 1 + 1 \times 0.1 + 4 \times 0.01 = 3.14$$

$$\begin{array}{r} 3 \\ \frac{1}{10} \\ \hline 2 \end{array}$$

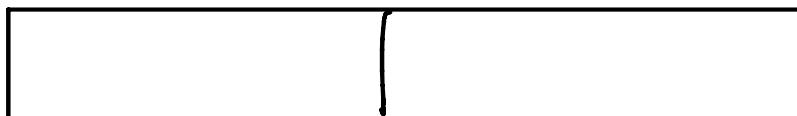
$$\begin{array}{r} 1 \\ \frac{1}{10} \\ \hline 0 \end{array}$$

$$\begin{array}{r} 4 \\ \frac{1}{100} \\ \hline 0 \end{array}$$

$$\begin{array}{r} 1 \\ \frac{1}{1000} \\ \hline 0 \end{array}$$

$$\dots$$

$$\begin{array}{r} -1 -2 -3 -4 \\ 10^3 10^2 10^1 10^0 10^{-1} 10^{-2} 10^{-3} 10^{-4} \dots \end{array}$$

Quiz

$$\rightarrow (0.5625)_{10} = (\quad)_2$$

$$\bullet (0.1)_2 =$$

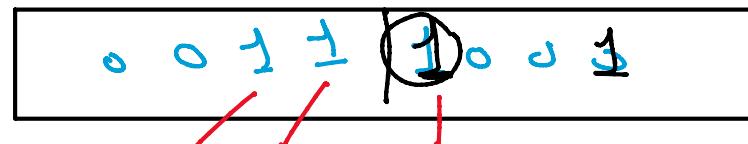
$$\bullet (0.011)_2 = \rightarrow 0.(\underline{\underline{0.1}})_{10} = (\quad)_2$$

$$\bullet (0.101)_2 =$$

$$0.5625 = \frac{0.5}{112} + \underline{0.0625} (\cdot 16)$$

Binary

$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	\dots
-2	-4	-8	-16	\dots



$$2 + 1 + \frac{1}{2} \\ = \boxed{3.5}$$

Decimal Integer \rightarrow Binary

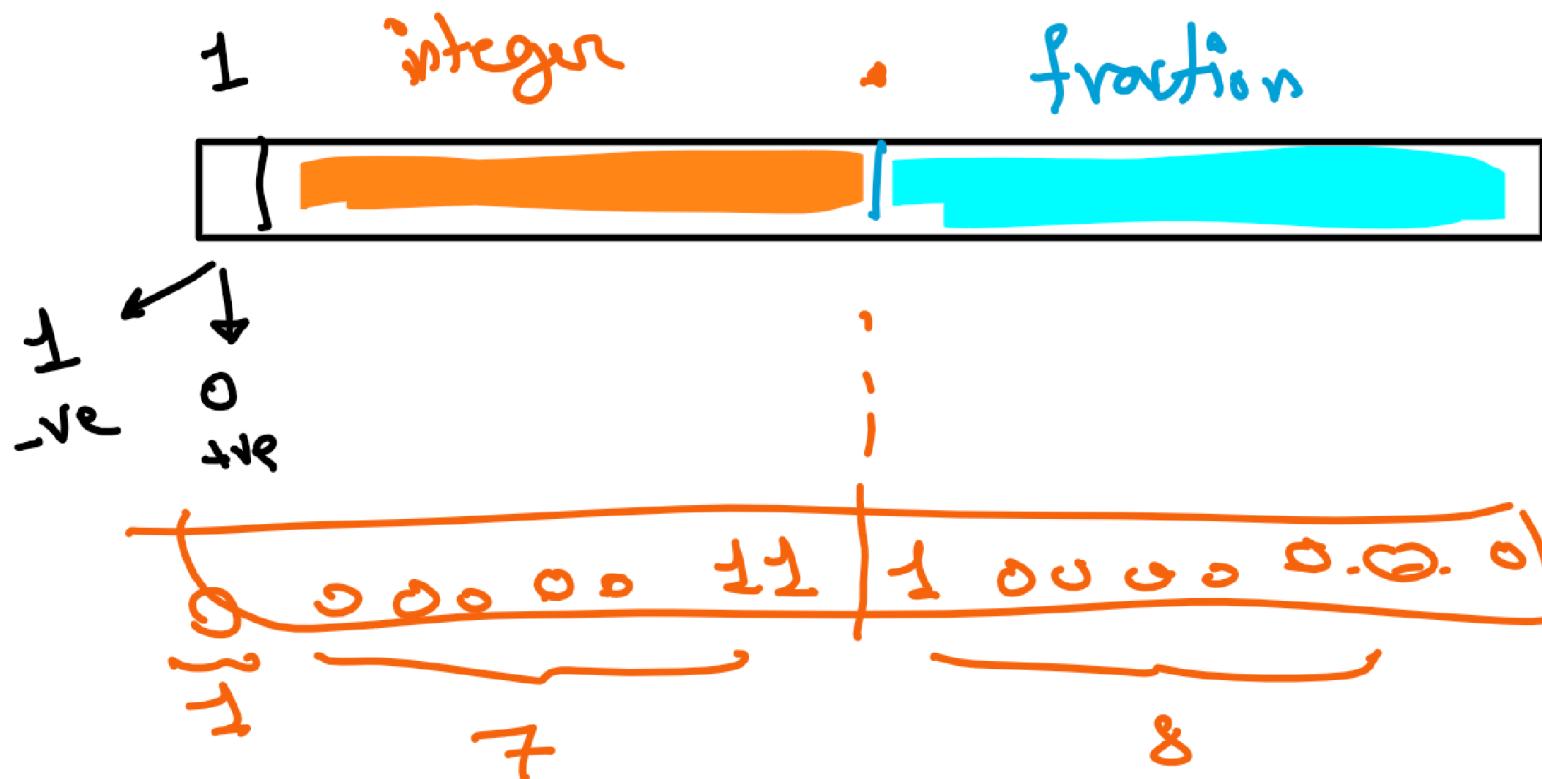
Decimal real \rightarrow Binary

$$\boxed{\frac{1}{3}} \rightarrow 0.\underline{\underline{33333}} \dots$$

$$(\frac{1}{10})_{10} \rightarrow 0.00011001100$$

$+0 =$
 $-C =$

16-bit



Fixed

3.5

- limited & fixed representations
 - ...

Point Representation

- Two distinct vals for \emptyset

Scientific Notation

$$0.\underline{0036525} \downarrow \\ 3.6525 \times 10^{-3}$$

365.25

$\frac{365.25}{1.0}$

$\text{S} = \underline{3.6525} \times 10^{\underline{2}}$

$S \in [1, 10)$

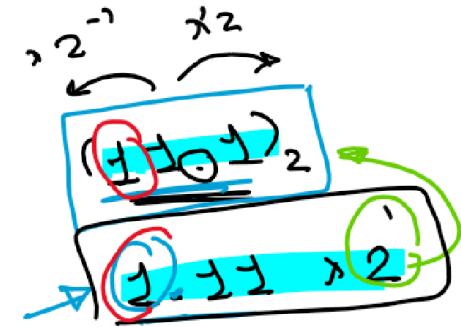
$x = \pm \underline{S} \times 10^{\underline{E}}$

$1 \leq S < 10$

Scientific Notation

$$x = \pm \underline{S} \times 2^{\underline{E}}$$

$$1 \leq S < 2$$



$$1.0$$

$$S = \underline{1.111111111\dots}$$

$$S = \underline{1.0} + \underline{f}$$

$$x = \pm (\underline{1} + \underline{f}) \times 2^{\underline{E}}$$

$$\textcircled{1} \quad \underline{1.011} \times 2^{-3}$$

$$\textcircled{2} \quad \underline{1.001001} \times 2^3$$

$$\textcircled{3} \quad \underline{1.0} \times 2^0 \checkmark$$

Quiz

$$\cdot (0.\underline{001011})_2 =$$

$$\cdot (1001.\underline{001})_2 =$$

$$\cdot (1.0)_2 =$$

Floating Point Representation

Floating Point

$$\rightarrow x = \underline{\pm(1 + f)} \times 2^e$$

1 5bit

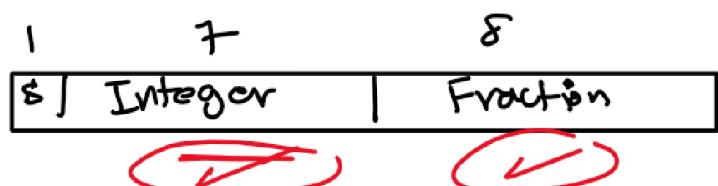


sign bit
+ve -ve Exponent Mantissa
 $\underline{0}$ 011 11001

* No fixed size for integers and fraction

* Exponent increases range

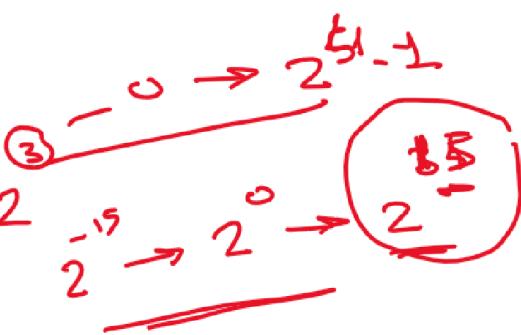
Fixed Point



$-15 \rightarrow +15$

2-

$$\begin{aligned}
 & + (1 + 0.11001) \times 2^{51-1} \\
 &= (1.011001) \times 2^3 \\
 &= \underline{1011.001}
 \end{aligned}$$



IEEE 745: Floating Point

Single Precision
(32-bit)

11	8	23
s	E	Mantissa

Mantissa (F)

128

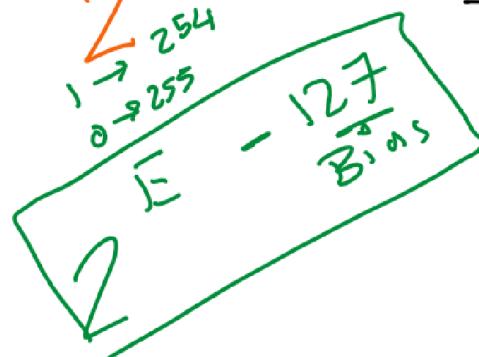
$$X = (-1)^s \times (1 + F) \times 2^{E-127}$$

E: 1 → 254
0 255 special values

Range
-126 → +127

Mantissa:

$$0 \rightarrow 2^3 - 1$$



Double-Precision
(64-bit)

1	11	52
s	E	Mantissa

Quiz

single Precision

$$= (-1)^{1} \times (1 + 0.011001) \times 2^{100001100_2 - 127}$$

$$= (-1)^{1} \times (1 + 0.011001) \times 2^{132 - 127}$$

$$\begin{aligned}
 X &= (-1)^s \times (1 + M) \times 2^{E-127} \\
 &= -1 \times (1 + 0.011001) \times 2^{132 - 127} \\
 &= -1 \times 1111 \times 2^5 \\
 &= -11011001 \times 2^{111001101} \\
 &= -(32+8+4) \cdot 5 = \boxed{-44 \cdot 5} \quad \checkmark
 \end{aligned}$$

$$= -\overline{(32+8+4)} \cdot 5 = \boxed{-44 \cdot 5} \quad \checkmark$$

$\overset{44,5}{\curvearrowleft} \quad \downarrow$

$$\begin{array}{r} 101100.1 \\ \rightarrow 1.\overline{101100} \times 2^5 \\ 5+127 = \boxed{132} \end{array}$$

Quiz

- Smallest positive number can be stored? Single Precision
- Largest " " " " " " " ?
- How to store zero? $E=0, M=0$ $\boxed{+0} \quad \boxed{-0}$
- How to store $+0$, -0 ?

$$(1+0) \times 2^{1-127} = \boxed{2^{-126}}$$

$$\underline{E=255}, \underline{M=0}$$

$\frac{\infty}{\infty}$ = not defined

$\frac{0}{0}$ = not defined

Not a Number

NaN

$$\underline{E=255} \quad \underline{M \neq 0} \quad \underline{S=0}$$

$$0.\overline{111\dots} \approx 1$$

$$254-127$$

$$(1+0.\overline{111\dots}) \times 2$$

$$\frac{1+1}{2+2} \stackrel{127}{=} \frac{(128)}{2} \approx \boxed{\frac{128}{2}}$$

Single

2^{-23} Precision

$$0 \quad i^1 i^2 i^3 \quad f \quad 2^{-2}$$

\$	E	00. - Manfissa	0	2
----	---	---------------------------	---	---

Machine
Epsilon
 ϵ

$$x = (1 + f) \times 2^0 = 1.0$$

$$\underline{x}_2 = (1 + 0.000\dots 01) \times 2^3 = 1.0\dots 01$$

$$x_1 = (1 + \frac{r}{n}) \times 2^{\frac{n}{2}} = (1 + 0) \times 2^{\frac{2}{2}} = 100 = \boxed{100}$$

$$x_2 = (1 + 0.00\ldots 1) \times 2^2 = \cancel{1.000\ldots 1} \times 2^2 = \cancel{100.00\ldots 100} \leftarrow \\ = 4.0 + 1 \times 2$$

$$w_p = \sum x_2^p$$

n.p. spacing (\rightarrow)

jump between two numbers

Dawnlina Fmav

$$x_2 = 1 - 2^{-21} \quad \underline{\hspace{10mm}} \quad 2^{-21}$$

Rounding Error

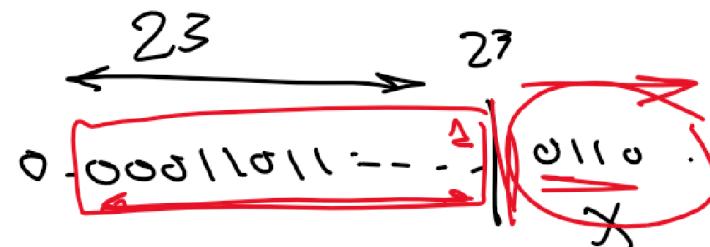
$$x_2 = \frac{11 \cdot 2^{-21}}{100.0 \dots 100}$$

↓

$$\begin{array}{r} 100.0 \dots 100 \\ 1.0 \dots 1 \\ \hline 100. \quad 1000 \end{array}$$

$\times 2^2$

$(0.1)_{10} \rightarrow \underbrace{\text{infinite digits}}$



Truncation Error

Cancellation Error

Rounding Error

Propagation Error

A - B
Absorption Error

$$1.0 \times 2^0$$

$$\begin{aligned} x &= 3.141592653589739 \\ y &= 3.141592653585682 \end{aligned}$$

$$x - y = \frac{0}{25}$$

$$\frac{1}{2} - 0 \times 2^{-26}$$

$$\begin{aligned} &1.000\ldots 1 \\ &\text{Valid} \quad 125 \\ &= 1.00\ldots 01 \quad \text{Valid} \quad 125 \end{aligned}$$

Root finding Problem

$$2x - 4 = 0$$

$$f(x) = 2x - 4 = 0$$

2

$$2 \cancel{x} - 4 = 0$$

$$2x - 4 = 0$$

$$\frac{2x}{2} = \frac{4}{2}$$

$$\boxed{x = 2}$$

solution
"root"

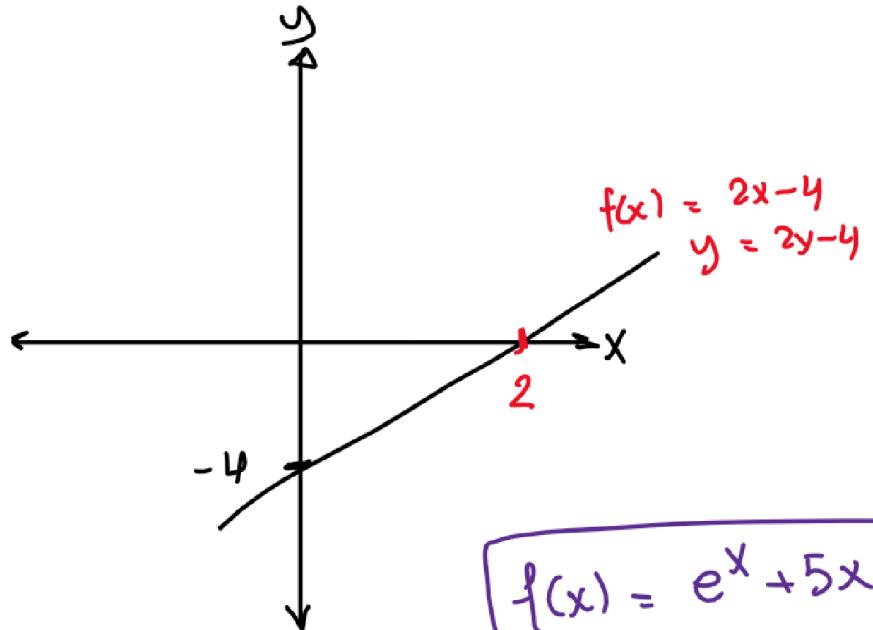
x = 2 is a solution

Qmiz

$$f(x) = 2x^2 + 6x + 4$$

roots?

$$x_1 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$



$$(2x + 2)(x + 2) = 0$$

$$\begin{aligned} 2x + 2 &= 0 \\ x &= -1 \end{aligned}$$

$$\begin{aligned} x + 2 &= 0 \\ x &= -2 \end{aligned}$$



Numerical Algorithms :

① Start initial guess

$$\text{root} = 0$$

② Iterations

→ improve root / guess

③ After N iterations

guess \approx actual root



Loss Function ($\frac{m}{I}, b$)

$$\boxed{\frac{d L}{dm} = 0}$$

Root Finding Problem

Bisection Method

$$f(x) = 2x^2 + 6x + 4$$

^{root}
① Interval

② $f(x)$

③ tolerance ϵ

Algorithm

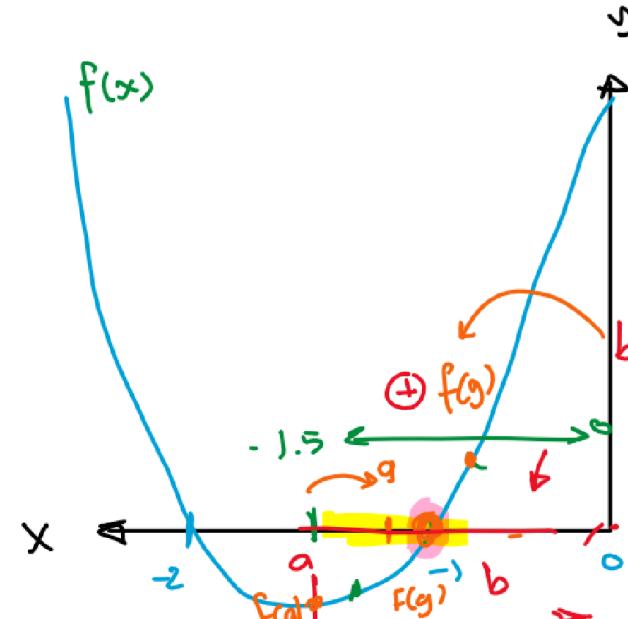
while $|a-b| > \epsilon$
 $g = \frac{a+b}{2}$

root on the left of g ? $f(a)f(g) < 0$

$\hookrightarrow b = g$

root on the right of g ? else

$\hookrightarrow a = g$



$f(g)f(a) < 0$ ~~~~~ on different sides of x-axis
 $f(g)f(a) > 0$

$|a-b|$ so small
 $|a-b| < \epsilon$

10^{-4}

Fixed Point Method

$$f(x) = e^x - 5x + 2 = 0$$

$y = e^x - 5x + 2$

$y = 0$

$$e^x - 5x + 2 = 0$$

$\ln e^x = 5x - 2$

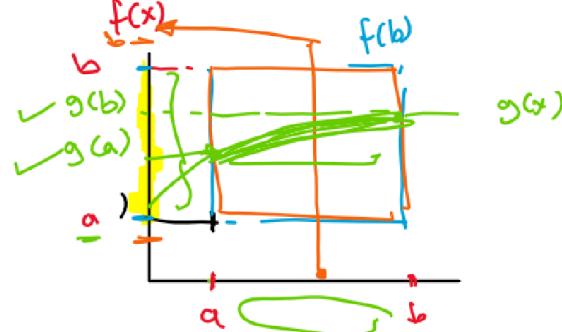
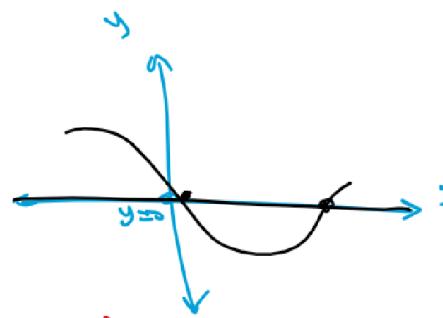
$x = \ln(5x - 2)$

$y = x$

$y = \ln(5x - 2)$

$$f(x) = 0$$

$$g(x) = x$$

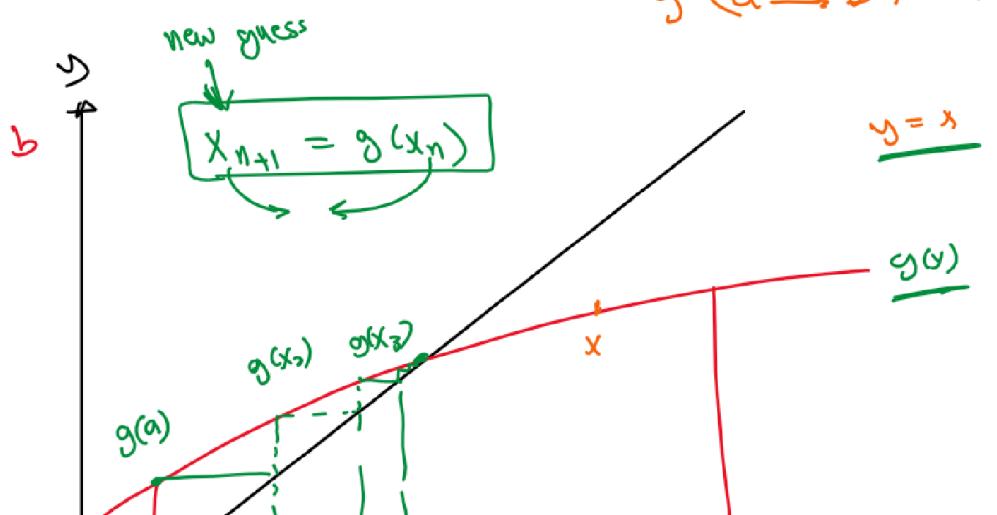


Conditions

$g(x) \rightarrow$ continuous in $[a, b]$

$\rightarrow g(x) \in [a, b] \quad \forall x \in [a, b]$

$g(a) \rightarrow b$



$$f(x) = e^x - 5x + 2$$

$$= e^x - x - 4x + 2 = 0$$

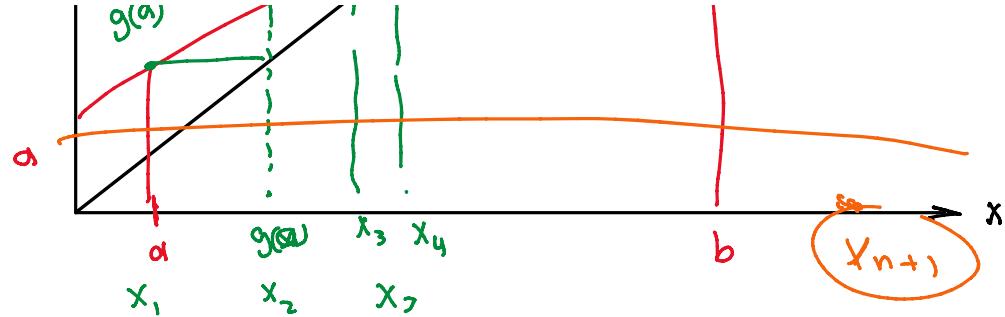
$$e^x - 4x + 2 = x$$

$y = g(x)$

$$e^x + 2 = 5x$$

$$\frac{e^x + 2}{5} = x$$

$y = g(x)$



Alg.

$$P_n = a$$

while $|P_{n+1} - P_n| > \varepsilon$

$$P_{n+1} = g(P_n)$$

Newton's Method

- ① $f(x)$
- ② $\frac{df(x)}{dx}, f'(x)$
- ③ tolerance

Algorithm

$x_1 = \text{guess}$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$\text{while } |x_2 - x_1| > \epsilon$$

$$x_1 = x_2$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$f'(x_1) = \tan \theta = \frac{f(x_1)}{\Delta x}$$

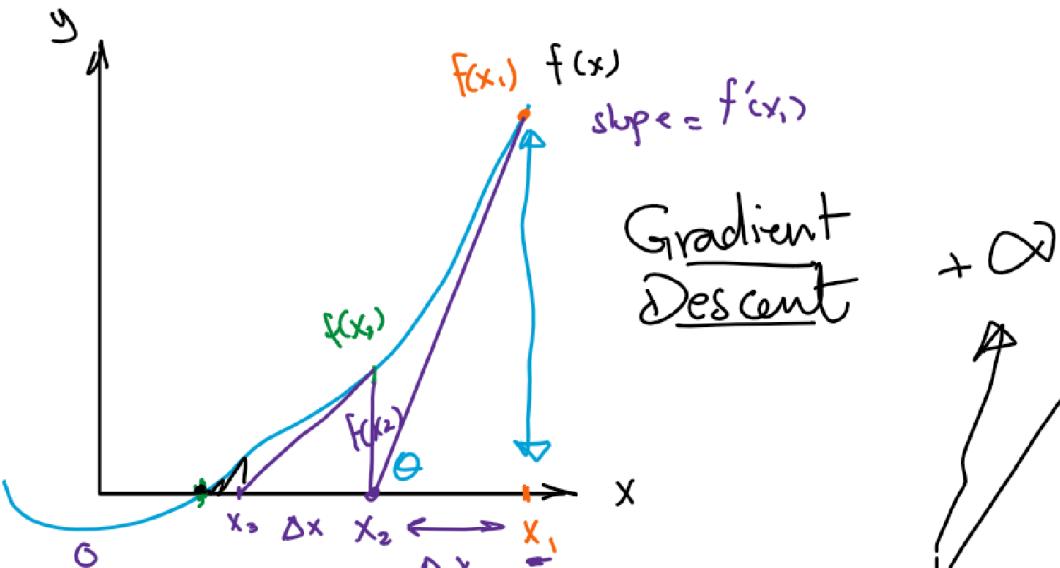
$$\Delta x = \frac{f(x_1)}{f'(x_1)}$$

$$x_1 - x_2 = \frac{f(x_1)}{f'(x_1)}$$

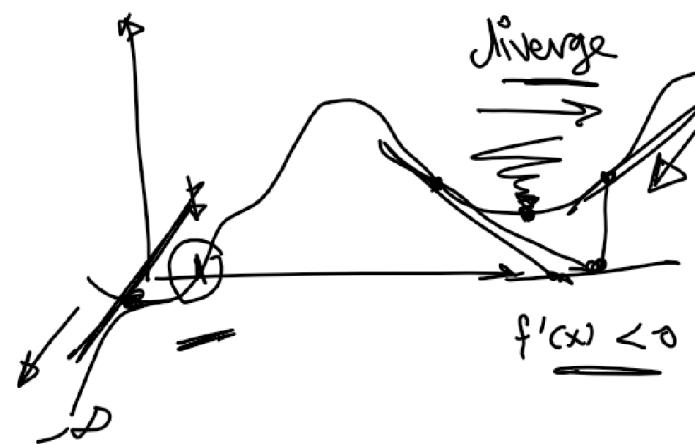
$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$|x_{n+1} - x_n| < \epsilon$$



Gradient Descent



HW

Saturday, May 29, 2021 3:40 PM

⑥ EP

- ① Implement 3 algorithms using

$$f(x) = 2x^2 + 6x + 4$$

How many iterations to reach the solution?

- ② Use different $g(x)$ for

$$f(x) = e^x - 5x + 2$$

other than $g(x) = \ln(5x - 2)$ [Fixed Point]

- ③ Use Newton's method $p=1$, what happens? why?

- ④ Apply Bisection Method on

$$e^x - 5x + 2$$

2, 3