$$flp(x) = x(1 \pm \varepsilon)$$

Consequence :

ARITHMETICS: $(x+y)+z = x+(y+z)$

F.P. ARITHMETICS: $flp(x+y)+flp(z) \neq flp(x)+flp(y+z)$

EXAMPLE : $x = 1 \quad y = 10^{-16} \quad z = 10^{-16}$

in F.P. : $\underbrace{\underbrace{(x+y)}_{1}+z}_{1} = 1 \quad \Big| \quad \underbrace{x}_{1} + \underbrace{(y+z)}_{2\cdot 10^{-16}} = 1.0000000000000002$

IEEE 754 STANDARD

→ FLOATING Point implementation

GOAL: reproducibility of results on different machines

FEATURES

① Binary format + HIDDEN BIT TECHNIQUE

$[\beta = 2]$

Ex: $F(\beta=2, t=3, L=-1, U=2)$

$$F: \{0\} \cup \left\{ \left(\sum_{k=0}^{t-1} d_k \beta^{-k} \right) \beta^{p} \right\} \qquad p \in [-1, 0, 1, 2]$$

MANTISSA : $d_0 \neq 0$

	d_0		d_1	d_2
$\beta=2$	1	.	0	0
	1	.	0	1
	1	.	1	0
	1	.	1	1

$\beta = 10 \quad d_0 \beta^{-k}$

VALUE of LAST BIT

SPACING $\Delta = 0,25$

$\Rightarrow$

$1\cdot 2^{0} + 0\cdot 2^{-1} + 0\cdot 2^{-2} = 1$

$1\cdot 2^{0} + 0\cdot 2^{-1} + 1\cdot 2^{-2} = 1,25$

$1\cdot 2^{0} + 1\cdot 2^{-1} + 0 = 1,5$

$1\cdot 2^{0} + 1\cdot 2^{-1} + 1\cdot 2^{-2} = 1,75$

Hidden bit technique: since $d_0 \neq 0$, with $\beta = 2$

$$d_0 \begin{cases} \cancel{0} \\ 1 \end{cases} \Rightarrow d_0 \text{ ALWAYS} = 1$$

EXPONENT PART

$$\begin{matrix} L = -1 \\ U = 2 \end{matrix} \rightarrow \beta^p = \{ 2^{-1} ; 2^{0} ; 2^{1} ; 2^{2} \}$$

$F(\beta = 2, t = 3, L = -1, U = 2)$

$p = -1 \rightarrow \beta^{-1} = 2^{-1} = 0{,}5$

$\vdots$

$= 1 \rightarrow \beta^{1} = 2$

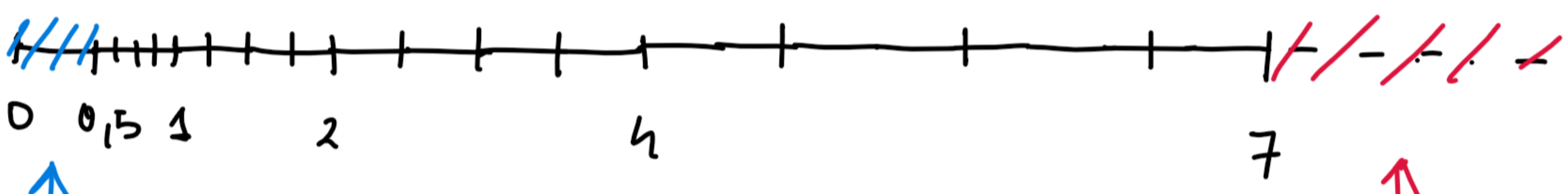
$p=-1$	$p=0$	$p=1$	$p=2$
0,5	1	2	4
0,625	1,25	2,5	5
0,75	1,5	3	6
0,875	1,75	3,5	7
$\Delta = 0{,}125$	$\Delta = 0{,}25$	$\Delta = 0{,}5$	$\Delta = 1$

$$\varepsilon = \frac{1}{2}\beta^{1-\overset{=3}{t}} = \frac{1}{2}\underset{=2}{\beta}^{-2} = 0{,}125$$

$x \Rightarrow x(1 \pm \varepsilon)$

REL. ERR

12,5% UNCERTAINTY over numbers



UNDERFLOW REGION

if x in underflow $x \rightarrow 0$

ALSO presented by standard

OVERFLOW REGION

if x in OVERFLOW $x \rightarrow \inf$

② PRÉCISION LEVELS

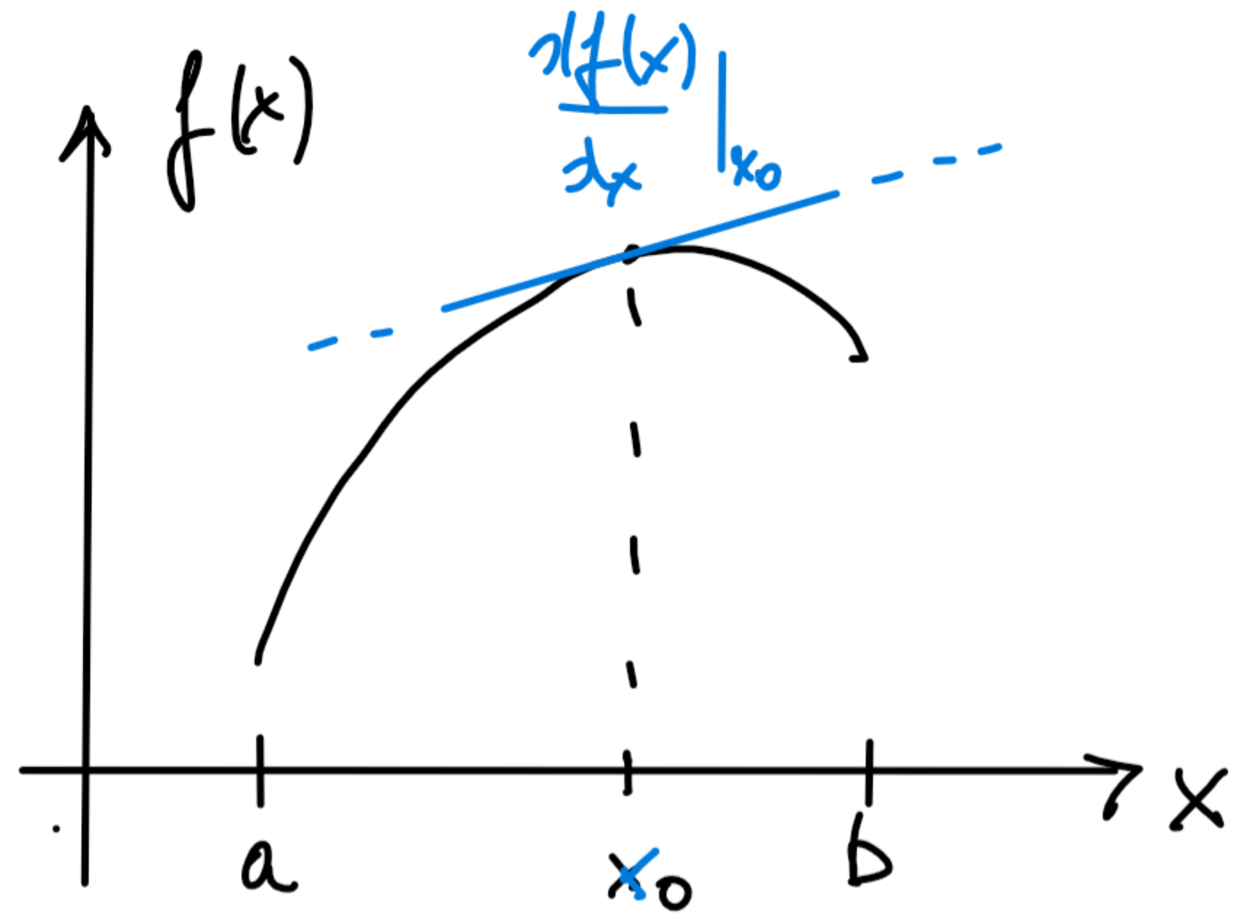
	SINGLE PRECISION 32 - BIT	DOUBLE PRECISION 64 - BIT
SIGN	1 BIT	1 BIT
EXPONENT	8 BIT $(2^8 = 256 \quad p \in [-126; 127])$	11 BIT $(2^{11} = 2048 \quad p \in [-1022, 1023])$
MANTISSA	23 BIT	52 BIT
ε	$\beta^{1-(23+1)} = \beta^{-23} \sim 10^{-7}$	$\beta^{1-(52+1)} = \beta^{-52} \sim 10^{-16}$

ε ↑ β^{1-t} ↑ # of digits for mantissa

NUMERICAL DIFFERENTIATION

$f(x) \in [a,b]$

$$\left.\frac{df(x)}{dx}\right|_{x_0} = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}$$



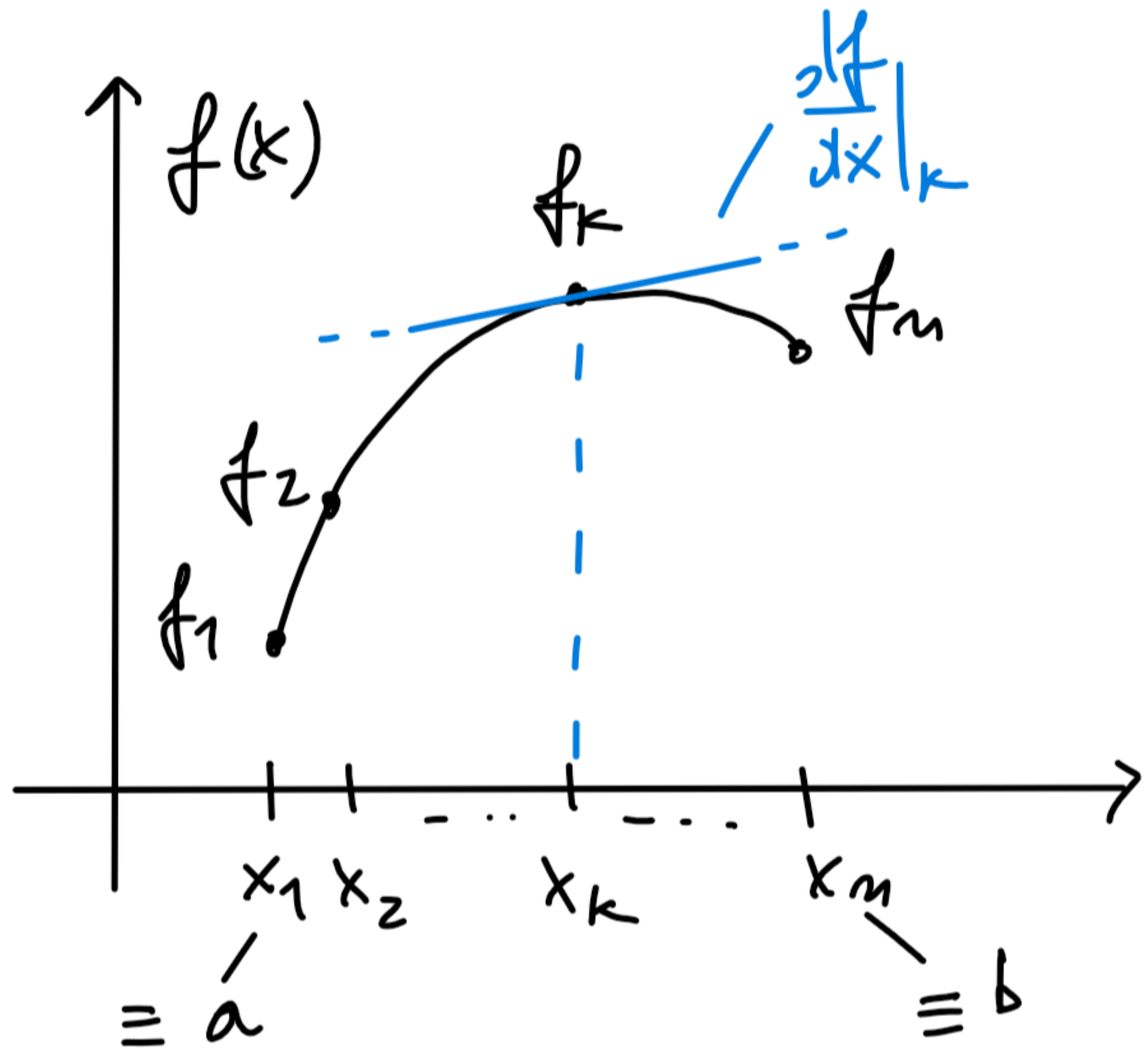
We need ALGEBRAIC APPROXIMATION of derivative

- Divide $[a,b]$ into set of n - nodes

$[x_1, x_2, x_3, \ldots, x_k, \ldots, x_n]$

$x_1 \equiv a \qquad x_n \equiv b$

$f(x_1) = f_1 \;;\; f(x_k) = f_k \;;\; f(x_n) = f_n$



TAYLOR SERIES around x_k

$$f(x) = f_k + \left.\frac{df(x)}{dx}\right|_k (x-x_k) + \frac{1}{2}\frac{d^2 f(x)}{dx^2}(x-x_k)^2 + \underbrace{\cdots\cdots}_{O(x-x_k)^3}$$

$\exists M \in \mathbb{R}^+ ,\; |err| \leq M(x-x_k)^3 ,$ when $(x-x_k) \to 0$

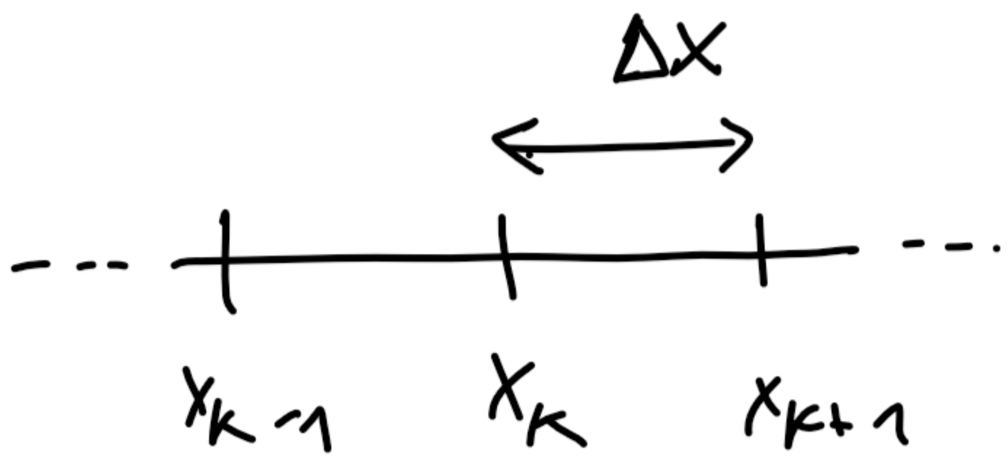
$\downarrow$

$$|err| = \left| f(x) - \left[f_k + \left.\frac{df}{dx}\right|_k (x-x_k) + \frac{1}{2}\left.\frac{df}{dx}\right|_k (x-x_k) \right] \right|$$

" the error SCALES AS $(x-x_k)^3$ when $(x-x_k)$ goes to zero

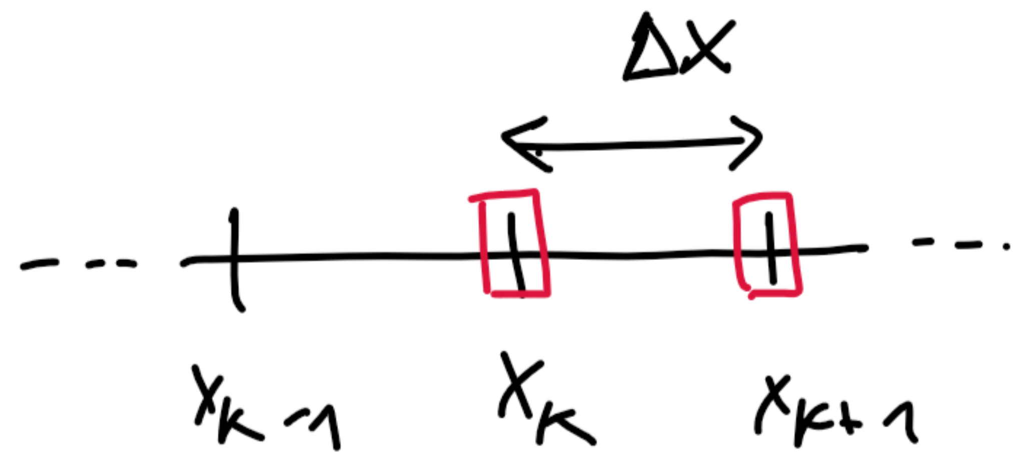
Assume <u>equal spacing</u>:

$$\Delta x = x_{k+1} - x_k = \frac{b-a}{n-1}$$



from x_k to x_{k+1}

$$f_{k+1} = f_k + \left.\frac{df(x)}{dx}\right|_k \overbrace{(x_{k+1} - x_k)}^{\Delta x} + \frac{1}{2}\frac{d^2 f(x)}{dx^2}\Delta x^2 + O(\Delta x^3)$$



TRUNCATION TO SECOND-ORDER

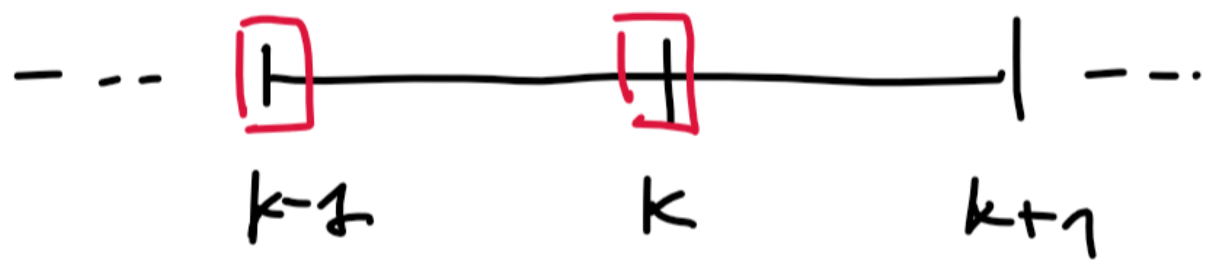
$$f_{k+1} = f_k + \left.\frac{df(x)}{dx}\right|_k \Delta x + O(\Delta x^2) \qquad (1)$$

$$\boxed{\left.\frac{df(x)}{dx}\right|_k = \frac{f_{k+1} - f_k}{\Delta x} + O(\Delta x)} \approx \frac{f_{k+1} - f_k}{\Delta x}$$

FORWARD FINITE difference formula

FIRST-ORDER ACCURATE $\rightarrow$ if $\Delta x' = \frac{\Delta x}{2}$ $\rightarrow$ $|err'| = \left|\frac{err}{2}\right|$

exponent of $O(\Delta x^1)$



TAYLOR for x_k to x_{k-1}

$$f_{k-1} = f_k - \left.\frac{df(x)}{dx}\right|_k \Delta x + \frac{1}{2}\frac{d^2 f(x)}{dx^2}\Delta x^2 + O(\Delta x^3)$$

$(x_{k-1} - x_k)$ $\qquad$ $(\underbrace{x_{k-1} - x_k}_{<0})^2$

2nd order.

$$f_{k-1} = f_k - \left.\frac{df(x)}{dx}\right|_k \Delta x + O(\Delta x^2) \qquad (2)$$

$$\boxed{\left.\frac{df(x)}{dx}\right|_k = \frac{f_k - f_{k-1}}{\Delta x} + O(\Delta x)} \approx \frac{f_k - f_{k-1}}{\Delta x}$$

BACKWARD Finite difference formula

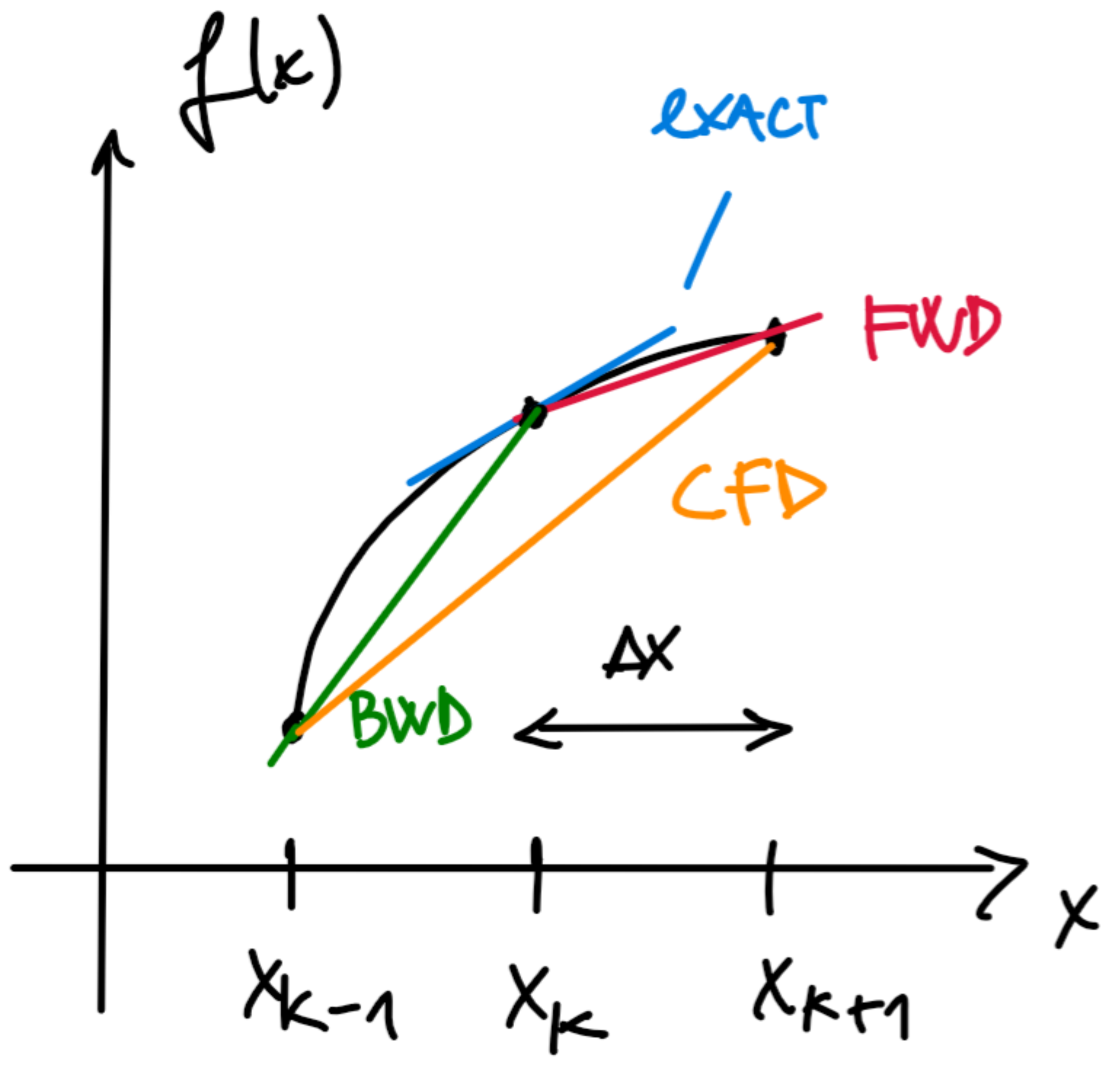
[ACCURACY: First-order]

if subtract (1) − (2), using THIRD-ORDER accurate expansions

$$f_{k+1} - f_{k-1} = \cancel{f_k} + \left.\frac{df(x)}{dx}\right|_k \Delta x + \cancel{\frac{1}{2}\left.\frac{d^2 f(x)}{dx^2}\right|_k \Delta x^2} - \left[\cancel{f_k} - \left.\frac{df(x)}{dx}\right|_k \Delta x + \cancel{\frac{1}{2}\frac{d^2 f(x)}{dx^2}\Delta x^2}\right] + O(\Delta x^3)$$

$$\boxed{\left.\frac{df(x)}{dx}\right|_k = \frac{f_{k+1} - f_{k-1}}{2\Delta x} + O(\Delta x^2)}$$

CENTERED Finite difference formula [ACCURACY: 2nd-order]



FORWARD F.D.

$$\left.\frac{df}{dx}\right|_k = \frac{f_{k+1} - f_k}{\Delta x}$$

BACKWARD F.D

$$\left.\frac{df}{dx}\right|_k = \frac{f_k - f_{k-1}}{\Delta x}$$

CENTERED F.D.

$$\left.\frac{df}{dx}\right|_k = \frac{f_{k+1} - f_{k-1}}{2\Delta x}$$