

## Pre-class assignment discussion

1) Discuss 2 required reading + code assignments

- Compare Qs you have about readings: what questions remain?
- Compare code results: do you get the same answers?  
What's the answer to Q3 of the PCA?

2) What are the implications of floating-point arithmetic?

Differentiation: Taylor expansion!

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2}f''(x_0) + \dots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0)$$

$$\text{1st deriv: } f'(x_i) = \lim_{\Delta x \rightarrow 0} \frac{f(x_i + \Delta x) - f(x_i)}{\Delta x}$$

"two-point  
formula"

$$x_{i+1} - x_i = \Delta x = h, \text{ so } f'_i = \frac{f_{i+1} - f_i}{h} + \underbrace{\mathcal{O}(h)}_{\text{error}} \quad \left. \vphantom{\frac{f_{i+1} - f_i}{h}} \right\} \text{1st-order solution}$$

"three-point  
formula"

$$\text{better solution: } f'(x_i) = \frac{f(x_i+h) - f(x_i-h)}{2h}$$

$$f(x_i \pm h) \approx f(x_i) \pm h f'(x_i) + \frac{h^2}{2} f''(x_i) \pm \frac{h^3}{6} f^{(3)}(x_i)$$

$$\text{end up with: } f'(x_i) \approx \frac{f(x_i+h) - f(x_i-h)}{2h} + \underbrace{\mathcal{O}(h^2)}_{\text{error}} \quad \left. \vphantom{\frac{f(x_i+h) - f(x_i-h)}{2h}} \right\} \text{2nd order solution}$$

## Numerical integration

$$F = \int_a^b f(x) dx$$

$f(x)$

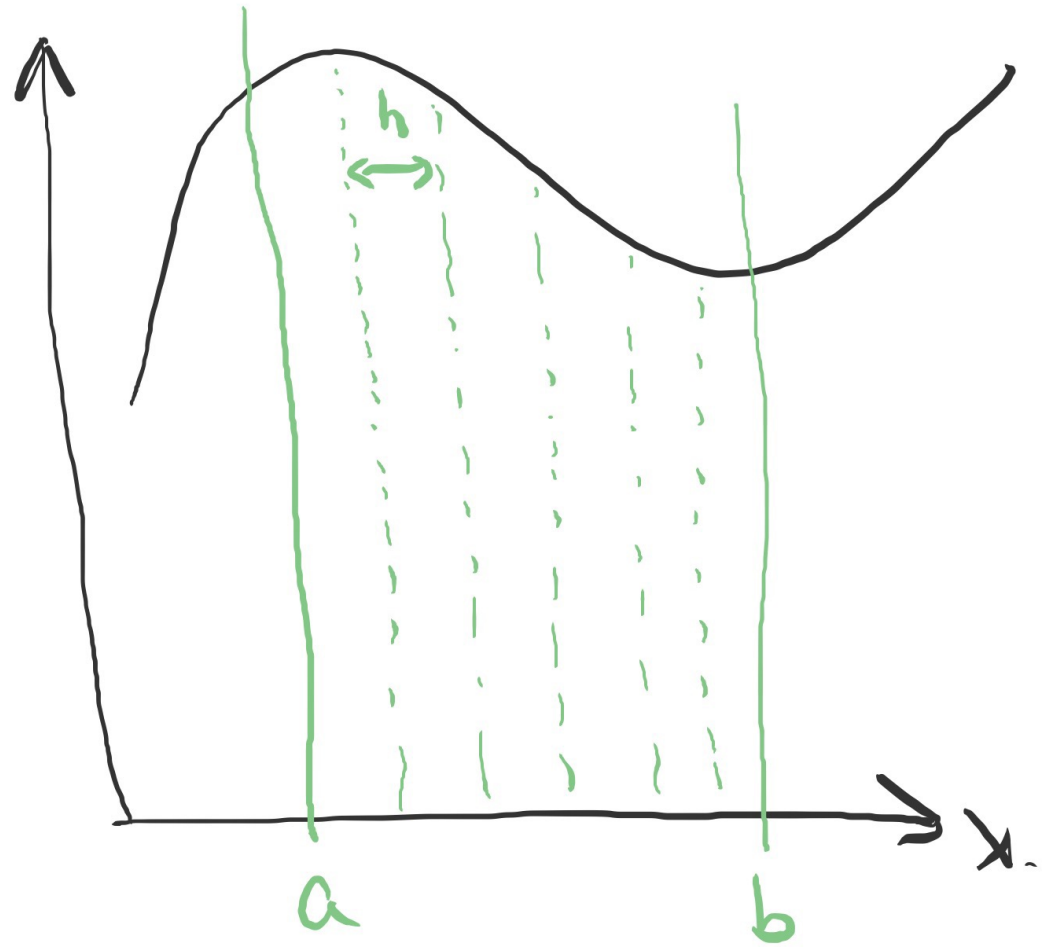
Can approximate area over  
small intervals and sum:

for evenly-sized intervals:

rectangles:  $O(h)$  error

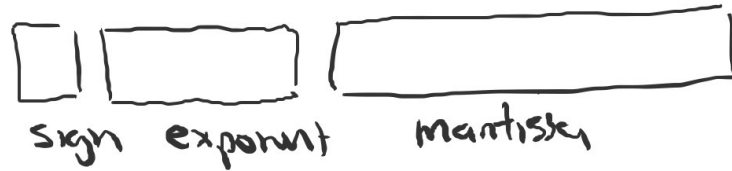
trapezoids:  $O(h^2)$  error

parabolas:  $O(h^4)$  error  
(Simpson's)



# Floating-point errors

IEEE standard:



32 bit: exponent:  $+127$  to  $-126$  mantissa  $\sim 7$  digits

64 bit: exponent:  $+1023$  to  $-1022$  mantissa  $\sim 16$  digits

Example:  $-1.234 =$ 

-
---

-3
----

1234
------

$+\pi =$ 

+
---

-7
----

31415927
----------

 — not exact; rounding

## types of floating-point errors

truncation - truncating  $\infty$  sum (represented by finite sum)

roundoff - difference between calculated approx. of a number and exact value (quantization error)

summation - accumulation of errors from computing sums (or generally from arithmetic operations)

representation - errors coming from numbers of diff. scales (e.g.,  $1.23 \times 10^{17}$  and  $0.11321197$ )

+ additional errors coming from compilers, "minimum ex", hardware, type conversion (float 64  $\rightarrow$  float 32, etc.)