Note.

If two floating point numbers with n significant digits are subtracted, then the result may have fewer than n significant digits. This is called loss of precision or loss of signifant digits due to cancellation of digits.

Ex

$x = \underbrace{0.123}4, \quad y = \underbrace{0.123}3$

$0.1234 - 0.1233 = 0.0001 = (0.1000)_{10} \cdot 10^{-3}$

the result has at most one correct significant digit

Ex (pg. 45) quadratic formula

$$ax^2 + bx + c = 0 \qquad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$0.2\,x^2 - 47.91\,x + 6 = 0$$

$$a = 0.2, \quad b = -47.91, \quad c = 6$$

$\Rightarrow x_1 = \underbrace{239.4}247, \quad x_2 = 0.1253$: computed by Matlab

Now suppose we use 4 digit arithmetic.

$$x = \frac{47.91 \pm \sqrt{47.91^2 - 4(0.2)\cdot 6}}{2\cdot(0.2)} = \frac{47.91 \pm \sqrt{2295 - 4.8}}{0.4} =$$

$$= \frac{47.91 \pm \sqrt{2290}}{0.4} = \frac{47.91 \pm 47.85}{0.4}$$

$\Rightarrow x_1 = \dfrac{47.91 + 47.85}{0.4} = \dfrac{95.76}{0.4} = \underline{239.4}$: all 4 digits are correct

$$x_2 = \frac{47.91 - 47.85}{0.4} = \frac{0.06}{0.4} = \underline{0.15}$$: only 1 digit is correct

The problem is due to cancellation of digits since we subtract two close numbers: 47.91 and 47.85. The remedy is to use a higher precision arithmetic (matlab) or reformulate the problem:

$$x = \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b + \sqrt{b^2 - 4ac}}{-b + \sqrt{b^2 - 4ac}} =$$

conjugate factor

$$(A - B)(A + B) = A^2 - B^2$$

$$= \frac{(-b)^2 - (b^2 - 4ac)}{2a(-b + \sqrt{b^2 - 4ac})} = \frac{\cancel{4}^{2}\cancel{a}c}{\cancel{2}\cancel{a}(-b + \sqrt{b^2 - 4ac})} = \frac{2c}{-b + \sqrt{b^2 - 4ac}}$$

Now: $$x_2 = \frac{2 \cdot 6}{47.91 + 47.85} = \frac{12}{95.76} = \underline{0.1253}$$: now all 4 digits are correct!

Note: $n=4$

$$x = \pm (0.d_1 d_2 d_3 d_4)_{10} \cdot 10^e, \qquad -M \le e \le M$$

$$x = 0.0001253700 \overset{?}{\neq} 0.0001$$

round

correct $fl(x) = 0.1254 \cdot 10^{-3}$

Other methods to eliminate cancellation of digits:

- Taylor expansion
- trigonometric identities
- properties of ln, exp etc.

Ex finite difference approximation of a derivative

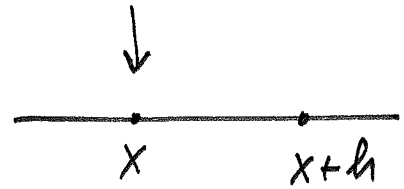
Recall

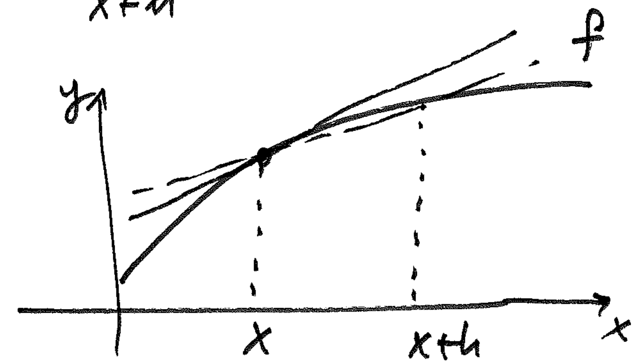
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

forward difference

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

$$= D_+ f(x)$$





Question: how large is the error?

Taylor series of $f(x)$ about $x=a$.

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \dots$$

Equivalent form:

$x \rightarrow x+h$

$a \rightarrow x$

$$\Rightarrow f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \dots$$

$$\Rightarrow \underset{\substack{\uparrow \\ \text{exact}}}{f'(x)} = \underset{\substack{\uparrow \\ \text{approximation}}}{\frac{f(x+h)-f(x)}{h}} \underbrace{- \frac{h}{2!} f''(x) - \dots}_{\text{discretization error}}$$

Hence, the error is proportional to h. We can write this as

$$f'(x) = D_+ f(x) + O(h)$$

$O(h) \simeq C \cdot h$ or

$$|f'(x) - D_+ f(x)| \le C \cdot h$$

where symbol $O(h)$ means "of order of h".

For example, if $f(x) = e^x$, $x = 1$

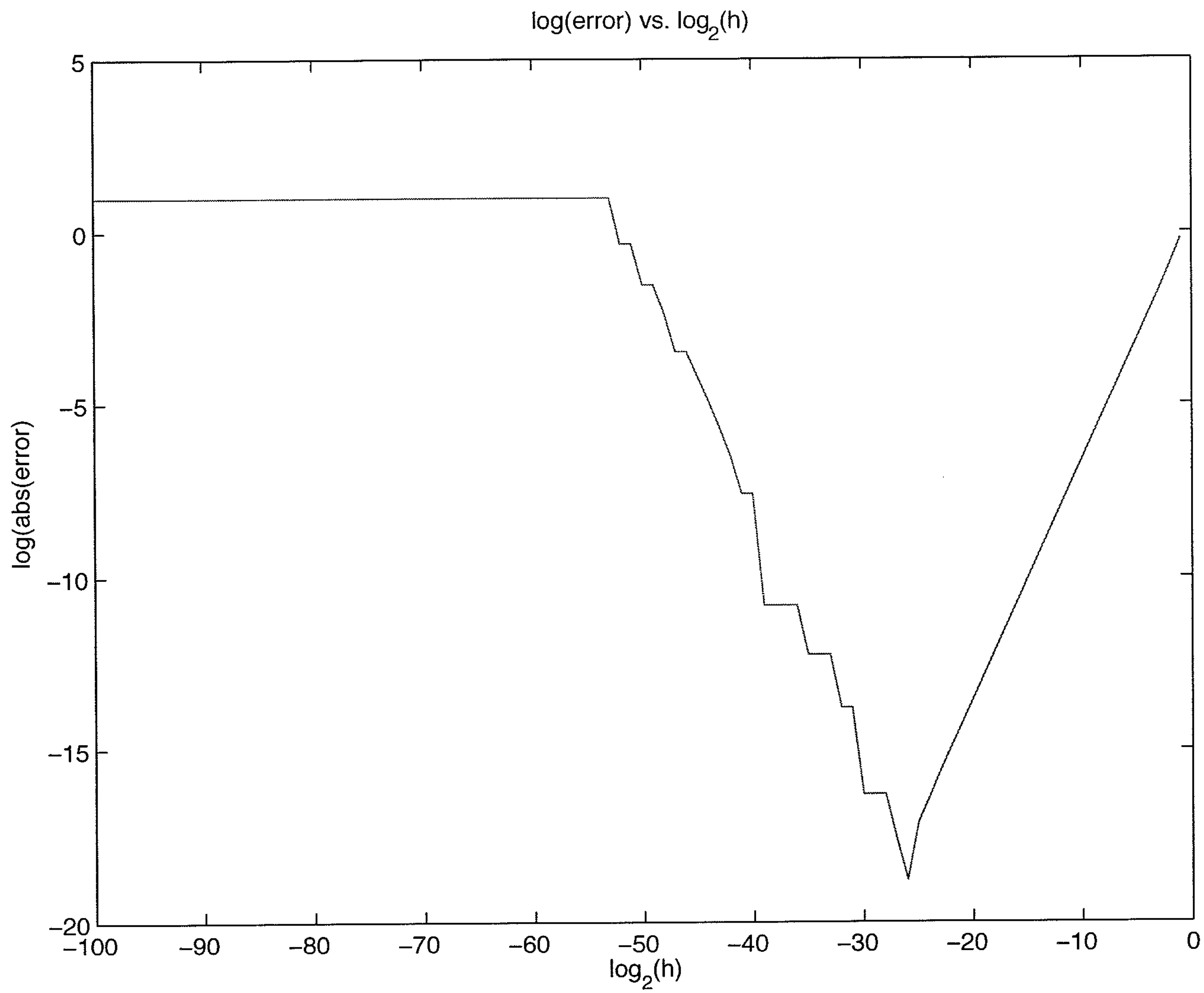
$f'(x) = e^x$, $f'(1) = e^1 = 2.71828\dots$: exact value

h	D_+f	$f'(x) - D_+f$	$\frac{f'(x) - D_+f}{h}$
0.1	2.8588	−0.1406	−1.4056
0.05	2.7874	−0.0691	−1.3821
0.025	2.7525	−0.0343	−1.3705
0.0125	2.7353	−0.0171	−1.3648
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
0	e	0	$-\frac{e}{2} = -\frac{1}{2} f''(1)$

```
x=1.0;
n=100;
for j=1:n,
    h(j)=1/2^j;
    deriv=(exp(x+h(j))-exp(x))/h(j);
    error(j)=deriv-exp(x);
if floor(j/10)==j/10
    disp(['h=',num2str(h(j),'%1.15e'), ', error=',num2str(error(j),'%1.15e')])
end
end
plot(log2(h),(log(abs(error))))
title('log(error) vs. log2(h) ')
xlabel('log2(h)')
ylabel('log(abs(error)')
```

n	$h = 1/2^n$	$D_+f - f'(1)$
10	9.765625000000000e-004	1.327718213427254e-003
20	9.536743164062500e-007	1.295809009427273e-006
30	9.313225746154785e-010	-8.254840100363481e-008
40	9.094947017729282e-013	-5.083909590455349e-004
50	8.881784197001252e-016	-2.182818284590455e-001
60	8.673617379884036e-019	-2.718281828459046e+000
70	8.470329472543003e-022	-2.718281828459046e+000
80	8.271806125530277e-025	-2.718281828459046e+000
90	8.077935669463161e-028	-2.718281828459046e+000
100	7.888609052210118e-031	-2.718281828459046e+000

1. For $h \geq 10^{-10}$, the error decreases as h is reduced, due to the discrete approximation.
2. For $h \geq 10^{-10}$, the error is linearly proportional to h.
3. For $h < 10^{-10}$, the error increases as h is reduced, due to finite precision arithmetic.



log(error) vs. log2(h)
log(abs(error))
5
0
−5
−10
−15
−20
−100
−90
−80
−70
−60
−50
−40
−30
−20
−10
0
log2(h)