Lecture January 16:

Solutions to practice problems double precision:

$$52.0 = 2^5 + 2^4 + 2^2 = 1.101 \times 2^5 \quad (1)$$
$$c - 1023 = 5 \quad c = 1028 = 2^{10} + 2^2 \quad (2)$$
$$52.0 = 0 \quad 10000000100 \quad 1010\ldots 0 \quad (3)$$
$$19.5 = 2^4 + 2^1 + 2^0 + 2^{-1} = 1.00111 \times 2^4 \quad (4)$$
$$c - 1023 = 4 \quad c = 1027 = 2^{10} + 2^1 + 2^0 \quad (5)$$
$$19.5 = 0 \quad 10000000011 \quad 001110\ldots 0 \quad (6)$$
$$3.1 = 2^1 + 2^0 + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + \ldots = 1.1000110011\ldots \times 2^1 \quad (7)$$
$$c - 1023 = 1 \quad c = 1024 = 2^{10} \quad (8)$$
$$3.1 = 0 \quad 10000000000 \quad 1000110011001100110011 0011\ldots \quad (9)$$

single precision:

$$52.0 = 2^5 + 2^4 + 2^2 = 1.101 \times 2^5 \quad (10)$$
$$c - 127 = 5 \quad c = 132 = 2^7 + 2^2 \quad (11)$$
$$52.0 = 0 \quad 10000100 \quad 1010\ldots 0 \quad (12)$$
$$19.5 = 2^4 + 2^1 + 2^0 + 2^{-1} = 1.00111 \times 2^4 \quad (13)$$
$$c - 127 = 4 \quad c = 131 = 2^7 + 2^1 + 2^0 \quad (14)$$
$$19.5 = 0 \quad 10000011 \quad 001110\ldots 0 \quad (15)$$
$$3.1 = 2^1 + 2^0 + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + \ldots = 1.1000110011\ldots \times 2^1 \quad (16)$$
$$c - 127 = 1 \quad c = 128 = 2^7 \quad (17)$$
$$3.1 = 0 \quad 10000000 \quad 10001100110011 00110011\ldots \quad (18)$$

Loss of precision error when evaluating

$$N(h) = \frac{f(x+h) - f(x)}{h}. \quad (19)$$

x is a constant in (19).

example

$$N(h) = \frac{\log(x+h) - \log(x)}{h}. \quad (20)$$

$$\lim_{h \to 0} N(h) = \lim_{h \to 0} \frac{\frac{1}{x+h}}{1} = \frac{1}{x}. \quad (21)$$

evaluate $N(h)$ using 4 digit chopping and $x = 2$

h	p^*	p	$\lvert(p - p^*)/p\rvert$
0.01	$(0.6981 - 0.6931)/0.01 = 0.5000$	0.4987542	0.00249
0.001	$(0.6936 - 0.6931)/0.001 = 0.5000$	0.499875	0.00025
0.0001	$(0.6931 - 0.6931)/0.001 = 0.0$	0.49999	1.0

replace numerator of $N(h)$ with $P_2(h)$ and simplify; Define $M(h)$ to be the numerator of $N(h)$:

$$M(h) = \log(x + h) - \log(x) \tag{22}$$

$$M'(h) = \frac{1}{x + h} \tag{23}$$

$$M''(h) = \frac{-1}{(x + h)^2} \tag{24}$$

expand about $h_0 = 0$:

$$P_2(h) = 0 + \frac{1}{x}h - \frac{1}{2x^2}h^2 \tag{25}$$

$$N(h) \approx \frac{P_2(h)}{h} = \frac{1}{x} - \frac{1}{2x^2}h \tag{26}$$

evaluate $P_2(h) = \frac{1}{x} - \frac{1}{2x^2}h$ using 4 digit chopping and $x = 2$

h	p^*	p	$\lvert(p - p^*)/p\rvert$
0.01	$0.5000 - 0.00125 = 0.4987$	0.4987542	0.00011
0.001	$0.5000 - 0.000125 = 0.4998$	0.499875	0.00015
0.0001	$0.5000 - 0.0000125 = 0.4999$	0.49999	0.00018

predict floating point precision

$$N(h) = \frac{f(x+h) - f(x)}{h} \tag{27}$$

$$fl(N(h)) = fl(\frac{fl(f(fl(x+h))) - fl(f(x))}{h}) = \tag{28}$$

$$\frac{f(x+h) + \epsilon_1 - f(x) - \epsilon_2}{h} \tag{29}$$

expand the numerator of (27) in a Taylor series:

$$M(h) = f(x+h) - f(x) \tag{30}$$

$$M'(h) = f'(x+h) \tag{31}$$

$$M''(h) = f''(x+h) \tag{32}$$

$$P_2(h) = 0 + f'(x)h + f''(x)h^2/2 \tag{33}$$

$$\frac{f(x+h) - f(x)}{h} \approx \frac{P_2(h)}{h} = f'(x) + f''(x)\frac{h}{2} \tag{34}$$

$$fl(\frac{f(x+h) - f(x)}{h}) = \frac{f(x+h) + \epsilon_1 - f(x) - \epsilon_2}{h} = \tag{35}$$

$$\frac{f(x+h) - f(x)}{h} + \frac{\epsilon_1 - \epsilon_2}{h} \approx \tag{36}$$

$$f'(x) + f''(x)\frac{h}{2} + \frac{\epsilon_1 - \epsilon_2}{h} \tag{37}$$

$$|N(h) - f'(x)| < M\frac{h}{2} + \frac{2\epsilon}{h} \quad M = \max_{x \le \xi \le x+h} |f''(\xi)| \tag{38}$$

optimal h? Define,

$$g(h) = M\frac{h}{2} + \frac{2\epsilon}{h} \tag{39}$$

$\min_h |g(h)|$ is found by checking critical points:

$$g'(h) = \frac{M}{2} - \frac{2\epsilon}{h^2} = 0 \tag{40}$$

$$\frac{M}{2} = \frac{2\epsilon}{h^2} \tag{41}$$

$$h^2 = \frac{4\epsilon}{M} \tag{42}$$

$$h = \sqrt{\frac{4\epsilon}{M}} \tag{43}$$

finding ϵ given h: In order to predict ϵ do the following steps:

1. Make a plot of $\log h$ versus $\log |f'(x) - N(h)|$ where $N(h)$ was found on a computer.

2. identify the value h_c that corresponds to the minimum point on the plot.
3. Solve (43) for ϵ.