Lecture January 21:

Notes from last class Next larger representable number after 1.0 for a single precision representation?

$$1.0 = 1.0 \times 2^0$$
$$c - 127 = 0$$
$$c = 01111111$$
$$1.0 = 0 \;\; 01111111 \quad 00000000000000000000000$$
$$1.0 + 2^{-23} = 0 \;\; 01111111 \quad 00000000000000000000001$$

The difference between the two is 1.2×10^{-7}. The immediate previous representable number is *NOT* $1.0 - 2^{-23}$, because

$$1.0 - 2^{-23} = 0 \;\; 01111110 \quad 11111111111111111111110$$

which is smaller than,

$$1.0 - 2^{-24} = 0 \;\; 01111110 \quad 11111111111111111111111 \tag{1}$$

The mantissa in (1) is

$$f = \sum_{i=1}^{23} 2^{-i}$$

In general, the sum of a geometric series has

$$S = a + ar + ar^2 + \ldots + ar^n$$
$$rS = S - a + ar^{n+1} \quad (r-1)S = a(r^{n+1} - 1) \quad S = \frac{a(1 - r^{n+1})}{1 - r}$$

so that

$$f = \frac{(1/2)(1 - (1/2)^{23})}{1 - 1/2} = 1 - 2^{-23}$$

The previous representable number is

$$2^{-1}(1 + 1 - 2^{-23}) = 1 - 2^{-24}$$

Loss of Precision 1 Loss of precision error when finding the root of:

$$f(x) = \epsilon x^2 + x - 1 \tag{2}$$

Loss of Precision 2 Loss of precision error when evaluating

$$N(h) = \frac{f(x+h) - f(x)}{h}. \tag{3}$$

x is a constant in (3).

example

$$N(h) = \frac{\log(x+h) - \log(x)}{h}. \tag{4}$$

$$\lim_{h\to 0} N(h) = \lim_{h\to 0} \frac{\frac{1}{x+h}}{1} = \frac{1}{x}. \tag{5}$$

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 $N(h)$:

$$M(h) = \log(x+h) - \log(x) \tag{6}$$

$$M'(h) = \frac{1}{x+h} \tag{7}$$

$$M''(h) = \frac{-1}{(x+h)^2} \tag{8}$$

expand about $h_0 = 0$:

$$P_2(h) = 0 + \frac{1}{x}h - \frac{1}{2x^2}h^2 \tag{9}$$

$$N(h) \approx \frac{P_2(h)}{h} = \frac{1}{x} - \frac{1}{2x^2}h \tag{10}$$

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

$$N(h) = \frac{f(x+h) - f(x)}{h} \tag{11}$$

$$fl(N(h)) = fl(\frac{fl(f(fl(x+h))) - fl(f(x))}{h}) = \tag{12}$$

$$\frac{f(x+h) + \epsilon_1 - f(x) - \epsilon_2}{h} \tag{13}$$

expand the numerator of (11) in a Taylor series:

$$M(h) = f(x+h) - f(x) \tag{14}$$

$$M'(h) = f'(x+h) \tag{15}$$

$$M''(h) = f''(x+h) \tag{16}$$

$$P_2(h) = 0 + f'(x)h + f''(x)h^2/2 \tag{17}$$

$$\frac{f(x+h) - f(x)}{h} \approx \frac{P_2(h)}{h} = f'(x) + f''(x)\frac{h}{2} \tag{18}$$

$$fl(\frac{f(x+h) - f(x)}{h}) = \frac{f(x+h) + \epsilon_1 - f(x) - \epsilon_2}{h} = \tag{19}$$

$$\frac{f(x+h) - f(x)}{h} + \frac{\epsilon_1 - \epsilon_2}{h} \approx \tag{20}$$

$$f'(x) + f''(x)\frac{h}{2} + \frac{\epsilon_1 - \epsilon_2}{h} \tag{21}$$

$$|N(h) - f'(x)| < M\frac{h}{2} + \frac{2\epsilon}{h} \quad M = \max_{x \le \xi \le x+h} |f''(\xi)| \tag{22}$$

optimal h? Define,

$$g(h) = M\frac{h}{2} + \frac{2\epsilon}{h} \tag{23}$$

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

$$g'(h) = \frac{M}{2} - \frac{2\epsilon}{h^2} = 0 \tag{24}$$

$$\frac{M}{2} = \frac{2\epsilon}{h^2} \tag{25}$$

$$h^2 = \frac{4\epsilon}{M} \tag{26}$$

$$h = \sqrt{\frac{4\epsilon}{M}} \tag{27}$$

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 (27) for ϵ.