

Solutions to Homework #3

1(c)

In the following table, δ denotes the absolute error and ϵ the relative error.

y	$fl(y)$	Chopping error	$fl(y)$	Rounding error
$\sqrt{2}$	1.414	$\delta = 2.136 \times 10^{-4}$ $\epsilon = 1.510 \times 10^{-4}$	1.414	$\delta = 2.136 \times 10^{-4}$ $\epsilon = 1.510 \times 10^{-4}$

2

Consider the floating point system $\mathbf{F}(\beta, k, m, M)$ with rounding. Let y be a real number whose expansion is given by

$$y = \pm(0.d_1d_2d_3 \cdots d_kd_{k+1} \cdots)_\beta \times \beta^e$$

with $d_1 \neq 0$ and $m \leq e \leq M$. If we let d denote $\beta/2$, then a bound on the absolute size of the roundoff error is

$$\begin{aligned} |fl_{\text{round}}(y) - y| &\leq (0.d)_\beta \times \beta^{e-k} \\ &= \frac{1}{2}\beta^{e-k}. \end{aligned}$$

Provided $y \neq 0$, given the restriction on d_1 ,

$$\begin{aligned} |y| &= (0.d_1d_2d_3 \cdots)_\beta \times \beta^e \\ &\geq (0.1)_\beta \times \beta^e = \beta^{e-1}. \end{aligned}$$

Therefore, the relative error in $fl_{\text{round}}(y)$ is bounded by

$$\frac{|fl_{\text{round}}(y) - y|}{|y|} \leq \frac{\frac{1}{2}\beta^{e-k}}{\beta^{e-1}} = \frac{1}{2}\beta^{1-k}.$$

4

- (a) Assuming the floating point system uses rounding, here is an algorithm to determine machine precision. Multiplication by β is performed in the output step because the while loop terminates when one too many divisions by β have been carried out.

GIVEN: base β

STEP 1: initialize $u = 1/2$

STEP 2: while ($1 + u > 1$)
 replace u by u/β

OUTPUT: $\beta \cdot u$

Here is an algorithm to determine the smallest positive number, assuming that underflow is handled by setting the value to zero.

GIVEN: base β

STEP 1: initialize $temp = 1$

STEP 2: while ($temp > 0$)

STEP 3: set $sm = temp$
 replace $temp$ by $temp/\beta$

OUTPUT: sm

- (b) Answers will of course vary. On a SunBlade 100, machine precision in both single and double precision is 2.22045×10^{-16} . The smallest positive number in single precision is 1.4013×10^{-45} and in double precision is 4.94066×10^{-324} .

6(c)

Because

$$\left| \frac{\sqrt{10002} - \sqrt{10001}}{\sqrt{10001}} \right| = 4.999 \times 10^{-5}$$

and

$$10^{-5} < 4.999 \times 10^{-5} \leq 10^{-4},$$

it follows that $\sqrt{10002}$ and $\sqrt{10001}$ agree to at least 4 and at most 5 decimal digits. Since

$$2^{-15} = 3.052 \times 10^{-5} < 4.999 \times 10^{-5} < 6.103 \times 10^{-5} = 2^{-14},$$

we see that $\sqrt{10002}$ and $\sqrt{10001}$ agree to at least 14 and at most 15 binary digits.

13.

With $\beta = 10$ and $k = 10$, machine precision with rounding is

$$u = \frac{1}{2} 10^{1-10} = 5 \times 10^{-10}.$$

Accordingly, there are between 9 and 10 significant decimal digits available in $F(10, 10, -98, 100)$. The smallest positive number is

$$(0.1)_{10} \times 10^{-98} = 10^{-99},$$

while the largest positive number is

$$(1 - 10^{-10})10^{100} = 9.999999999 \times 10^{99}.$$