AMSC/CMSC 660 Quiz 1 , Fall 2006

Show all work. You may leave arithmetic expressions in any form that a calculator could evaluate. By putting your name on this paper, you agree to abide by the university's code of academic integrity in completing the quiz. Use no books, calculators, cellphones, other electronic devices, communication with others, scratchpaper, etc.

1. (8) For each machine-representable number `r`, define `f(r)` to be the next larger machine-representable number. Consider the following statements:

> (a) For fixed point (integer) arithmetic, the distance between `r` and `f(r)` is constant.
> (b) For floating point arithmetic, the relative distance `|(f(r)-r)/r|` is constant (for $\texttt{r} \neq 0$).

Are the statements true or false? Give examples or counterexamples to explain your reasoning.

Answer:

(a) This is true. The integers are equally spaced, with distance equal to 1, and you can easily generate examples.

(b) This is only *approximately* true.

- For example, with a 53-bit mantissa, if $\texttt{r} = 1$, then $\texttt{f(r)} = 1 + 2^{-52}$, for a relative distance of 2^{-52}.
- Similarly, if $\texttt{r} = 8 = 1000_2$, then $\texttt{f(r)} = 1000_2 + 2^{-52+3}$, for a relative distance of $2^{-49}/2^3 = 2^{-52}$.
- But if $\texttt{r} = 1.25 = 1.01_2$, then $\texttt{f(r)} = 1.25 + 2^{-52}$, for a relative distance of $2^{-52}/1.25$.

In general, suppose we have a machine-representable number `r` with positive mantissa z and exponent p. Then $\texttt{f(r)} = (z + 2^{-52}) \times 2^p$, so the relative distance is

$$\frac{(z + 2^{-52}) \times 2^p - (z) \times 2^p}{z \times 2^p} = \frac{2^{-52}}{z}.$$

Because $1 \leq z < 2$, the relative distance is always between 2^{-52} and 2^{-53}, constant within a factor of 2. A similar argument holds for negative mantissas.

2. (6) Consider the following code fragment:

```
x = 1;
delta = 1 / 2^(53);
for j1=1:2^(20),
  x = x + delta;
end
```

Using mathematical reasoning, we expect the final value of `x` to be $1 + 2^{-33}$. Use your knowledge of floating-point arithmetic to predict what it will actually be. Briefly explain your prediction.

Answer: (Note this is Challenge 1.1a.) The number `delta` is represented with a mantissa of 1 and an exponent of -53. When this is added to 1 the first time through the loop, the sum has a mantissa with 54 digits, but only 53 can be stored, so the low-order 1 is dropped and the answer is stored as 1. This is repeated 2^{20} times, and the final value of `x` is still 1.

3. (6) Bound the backward error in approximating the solution to

$$\begin{bmatrix} 2 & 1 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5.244 \\ 21.357 \end{bmatrix} \text{ by } \mathbf{x}_c = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Answer: Notice that $\mathbf{x}_c$ solves the linear system

$$\begin{bmatrix} 2 & 1 \\ 3 & 6 \end{bmatrix} \mathbf{x}_c = \begin{bmatrix} 5 \\ 21 \end{bmatrix},$$

so we have solved a linear system whose right-hand side is perturbed by

$$\mathbf{r} = \begin{bmatrix} 0.244 \\ 0.357 \end{bmatrix}.$$

The norm of $\mathbf{r}$ gives a bound on the change in the data, so it is a backward error bound.

(The true solution is $\mathbf{x}_{true} = [1.123, 2.998]^T$, and a forward error bound would be computed from $\|\mathbf{x}_{true} - \mathbf{x}_c\|$.)