Numerical Methods I – Problem Set 1

Homework due September 12, 2003

Projects due September 15, 2003

1. (a) The IEEE 754 floating point format has the useful property that the ordering of numbers is preserved when all the bits are interpreted as a sign-magnitude integer. Explain briefly why this is the case.

   (b) The introduction of subnormals into the IEEE floating point standard was considered a significant advance. What useful property of was lost if subnormals were not present?

   Hint: Think about the distance between zero and the two smallest positive floating point numbers.

2. (a) Explain the following Octave result:

   ```
   octave:1> log(1+3e-16)/3e-16
   ans = 0.74015
   ```

   (Example due to L. Vandenberghe.)

   (b) Use Octave to compute $\sin(1.0 \times 10^{20}\pi)$. What goes wrong?

   (c) The following Octave Program has a subtle bug. Can you fix it?

   ```
   x = 0.0;
   d = 0.1;
   while x <> 1.0
     x = x + d;
   end
   x
   ```

3. Find the condition number of evaluating $y = \sqrt{x}$ near $x = 1$ and $x = 0$.

4. Given a smooth function $f(x)$ with a simple zero at $x = x_0$, and a smooth bounded function $g(x)$. Let

   $$h(x) = f(x) + \varepsilon\, g(x)\,.$$

   Show that when ε is small, $h(x)$ has a zero at $x = x_0 + \delta$, where

   $$\delta \approx -\varepsilon\, \frac{g(x_0)}{f'(x_0)}\,.$$

   When is the problem well, and when is it ill conditioned?

5. **Project:** Find the roots of the quadratic equation

$$x^2 + p\,x + 1 = 0$$

by using the standard formula.

(a) Show that the zeros are approximately $-p$ and $-1/p$ when p is large.

(b) Let Octave compute the zeros when $p = 10^{10}$. What do you get?

(c) Can you rewrite the solution formula so that the computation is stable?

6. **Project:** (From QS, p. 6) The sequence defined by

$$z_2 = 2\,,$$
$$z_{n+1} = 2^{n-1/2}\sqrt{1 - \sqrt{1 - 4^{1-n} z_n^2}} \qquad \text{for } n = 2, 3, \dots\,,$$

converges to π as $n \to \infty$.

(a) Write an Octave program that plots the logarithm of the error vs. n.

(b) Explain why the error grows when n gets larger than about 16.

(c) Why does this sequence converge to π?

(d) Can you improve the stability of this algorithm?