Homework 2 Foundations of Computational Math 1 Fall 2017

Problem 2.1

Suppose the n-bit 2's complement representation is used to encode a range of integers, $-2^{n-1} \leq x \leq 2^{n-1} - 1$.

2.1.a. If $x \geq 0$ then $-x$ is represented by bit pattern obtained by complementing all of the bits in the binary encoding of x, adding 1 and ignoring all bits in the result beyond the n-th place, i.e., the bit with weight 2^{n-1}. This procedure is also used when $x < 0$ to recover the encoding of $-x \geq 0$. What is the relationship between the binary encoding of $-2^{n-1} \leq x \leq 2^{n-1} - 1$ and the binary encoding of $-x$ in terms of the number of bits n?

2.1.b. Show that simple addition modulo 2^n on the encoded patterns is identical to integer addition (subtraction) for $-2^{n-1} \leq x,\ y \leq 2^{n-1} - 1$. You may ignore results that are out of range, i.e., overflow.

2.1.c. Show how overflow in addition (subtraction) can be detected efficiently.

2.1.d. Multiplying an unsigned binary number by 2 or 1/2 corresponds to shifting the binary representation left and right respectively (a so-called logical shift). Show how multiplying signed integers encoded via 2's complement representation by 2 or 1/2 can be done via a shifting operation (an arithmetic shift).

Problem 2.2

Consider the following numbers:

- 122.9572
- 457932
- 0.0014973

2.2.a. Express the numbers as floating point numbers with $\beta = 10$ and $t = 4$ using rounding to even and using chopping.

2.2.b. Express the numbers as floating point numbers with in single precision IEEE format using rounding to even. It is strongly recommended that you implement a program to do this rather than computing the representation manually.

2.2.c. Calculate the relative error for each number and verify it satisfies the bounds implied by the floating point system used.

Problem 2.3

2.3.a. Suppose $x \in \mathbb{R}$ and $y \in \mathbb{R}$ with $x < y$. Is it always true that $fl(x) < fl(y)$ in any standard model floating point system?

2.3.b. Suppose x, y and z are floating point numbers in a standard model floating point arithmetic system. Is floating point arithmetic associative, i.e., is it true that

$$(x \boxed{op} (y \boxed{op} z)) = ((x \boxed{op} y) \boxed{op} z) \text{ ?}$$

2.3.c. Is floating point arithmetic distributive, i.e., is it true that

$$fl(fl(x+z)*y) = fl(fl(fl(x*y)+fl(y*z)))?$$

Problem 2.4

Consider the function

$$f(x) = \frac{1.01 + x}{1.01 - x}$$

2.4.a. Find the absolute condition number for $f(x)$.

2.4.b. Find the relative condition number for $f(x)$.

2.4.c. Evaluate the condition numbers around $x = 1$.

2.4.d. Check the predictions of the condition numbers by examining the relative error and the absolute error

$$err_{rel} = \frac{|f(x_1) - f(x_0)|}{|f(x_0)|}$$
$$err_{abs} = |f(x_1) - f(x_0)|$$

with $x_0 = 1$, $x_1 = x_0(1 + \delta)$ and δ small.

Problem 2.5

Let $f(\xi_1,\ \xi_2, \dots, \xi_k)$ be a function of k real parameters ξ_i, $1 \le i \le k$. Recall, the relative condition number of f with respect to ξ_1 can be expressed

$$\kappa_{rel} = \max(1, c(\xi_1, \xi_2, \dots, \xi_k))$$

where $0 \le c(\xi_1, \xi_2, \dots, \xi_k)$ is a value that indicates the sensitivity of f to small relative perturbations to ξ_1 as a function of the parameters ξ_i, $1 \le i \le k$. If $c(\xi_1, \xi_2, \dots, \xi_k) \le 1$

then f is considered well-conditioned. Additionally, however, when $c < 1$ its value gives important information. The smaller c is the less sensitive f is to a relative perturbations in ξ_1.

Let $n \geq 2$ be an integer and $\beta > 0$. Consider the polynomial equation

$$p(x) = x^n + x^{n-1} - \beta = 0.$$

2.5.a. Show that the equation has exactly one positive root $\rho(\beta)$.

2.5.b. Derive a formula for $c(\beta, n)$ that indicates the sensitivity of $\rho(\beta)$ to small relative perurturbations to β.

2.5.c. Derive a upper bound on $c(\beta, n)$.

2.5.d. Comment on the conditioning of $\rho(\beta)$ with respect to β.

Problem 2.6

The evaluation of

$$f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$$

encounters cancellation for $x \gg 0$.

Rewrite the formula for $f(x)$ to give an algorithm for its evaluation that avoids cancellation.

Problem 2.7

2.7.a

Suppose that x and y are two floating point numbers in a system that supports gradual underflow and satisfies the standard model. Show that if $y/2 \leq x \leq 2y$ then

$$fl(x - y) = x - y$$

2.7.b

Suppose a triangle has sides with lengths $a \geq b \geq c$. Heron's formula for its area is

$$A = \sqrt{s(s-a)(s-b)(s-c)}, \quad s = \frac{a+b+c}{2}$$

Kahan has suggested the following formula

$$A = \frac{1}{4}\sqrt{(a + (b + c))(c - (a - b))(c + (a - b))(a + (b - c))}$$

(i) What happens with the Heron's formula with needle-shaped triangles?

(ii) Give an informal proof that Kahan's formula is reliable numerically. You may consult the literature of course.

(iii) Compare the accuracy of the two formulae in single-precision for several examples to illustrate your points.