

1. When you subtract two nearly equal floating point numbers, the result will not have full precision, i.e., there will be less correct digits. This is called “loss of significance” caused by subtraction. For each of the following functions, find the values of  $x$  at which there will be loss of significance, and suggest how to avoid loss of significance.

- (a)  $\sqrt{x^2 + 1} - x$
- (b)  $(1 - \cos x)/x^2$
- (c)  $(\sin x - x)/x^3$
- (d)  $\sqrt{x + 2} - \sqrt{x}$
- (e)  $e^x - e$
- (f)  $\log(x) - 1$
- (g)  $(\cos x - e^{-x})/\sin x$
- (h)  $\sin x - \tan x$
- (i)  $\sinh x - \tanh x$
- (j)  $\ln(x + \sqrt{x^2 + 1})$

2. For the following linear recurrence relationship,

$$a_{n+2} = (a_n - a_{n+1})/6,$$

find the solution satisfying the initial conditions:  $a_0 = \alpha$  and  $a_1 = \beta$ .

- Show that the special solution for  $\alpha = 1$ ,  $\beta = 1/3$  is unstable.
- Show that the special solution for  $\alpha = 1$ ,  $\beta = 1$  is stable.

To analyze stability, you should consider  $\tilde{a}_n$  which satisfies the same recurrence relationship, but with

$$\tilde{a}_0 = \alpha + \epsilon_1, \quad \tilde{a}_1 = \beta + \epsilon_2,$$

for small constants  $\epsilon_1$  and  $\epsilon_2$ . If the relative error

$$R_n = \left| \frac{\tilde{a}_n - a_n}{a_n} \right|$$

remains small for all  $n$ , then  $\{a_n\}$  is stable. If  $R_n \rightarrow \infty$  as  $n \rightarrow \infty$ , then  $\{a_n\}$  is unstable.

3. The standard double precision floating point numbers can be written down as

$$x = \pm(1.b_1b_2\dots b_{52})_2 \times 2^q = \pm \left( 1 + \frac{b_1}{2} + \frac{b_2}{4} \dots + \frac{b_{52}}{2^{52}} \right) \times 2^q$$

for some integer  $q$  (which is also restricted, but this does not concern us here), where  $b_j \in \{0, 1\}$  for  $j = 1, 2, \dots, 52$ . Find  $a < 0$  and  $b > 0$ , such that

$$fl(4 + \epsilon) = 4, \quad a < \epsilon < b.$$

4. Give an example, to show that the computer result for  $x_1 + x_2 + x_3$  is usually not  $fl(x_1 + x_2 + x_3)$ , where  $x_1$ ,  $x_2$  and  $x_3$  are single precision floating point numbers.