1. In the following transcript of an Octave session, all output has been deleted.

```
octave:1> a=1e308; b=1.05e308; c=-1.5e308;
octave:2> a+(b+c)
ans = ..................
octave:3> (a+b)+c
ans = ..................
octave:4> a/((1+a)-a)
ans = ..................
octave:5> x=1e-15;
octave:6> 1-cos(x)
ans = ..................
octave:7> sin(x)^2/(1+cos(x))
ans = ..................
```

(d)

(b)

(e)

(a)

(c)

Identify which of the following answers belongs to each of the dotted lines.

(a) `ans = 0`

(b) `ans = Inf`

(c) `ans = 5.0000e-31`

(d) `ans = 5.5000e+307`

(e)
```
warning:  division by zero
ans = Inf
```

(10)

Comments: (not required for the exam)

- ② and ③ should give the same answer, however a+b produces an exponent overflow.
- a is so large that $\mathrm{fl}(a+1) = \mathrm{fl}(a)$
  ⇒ the denominator in ⑥ is zero
- $$1-\cos x = \frac{(1-\cos x)(1+\cos x)}{1+\cos x} = \frac{\sin^2 x}{1+\cos x}$$

  So ⑥ and ⑦ should give the same answer, but ⑥ has cancellation of 2 significant digits while ⑦ is numerically stable when x is small.

2. Suppose that a function has a zero in the interval $[0, 1]$. Show that the bisection method is guaranteed to approximate the zero within a specified tolerance ε after

$$k \geq \frac{\ln 1/\varepsilon}{\ln 2} - 1$$

iterations. (10)

After 0 iterations the zero is contained in an interval $I_0 = [0,1]$, so at most a distance $\frac{1}{2} = 2^{-(0+1)}$ from the midpoint of that interval.

Each successive iteration halves the width of the interval, i.e. halves the maximum error, so that after k iteration the maximum error is

$$2^{-(k+1)}$$

⇒ We need

$$2^{-(k+1)} \leq \varepsilon$$

$$\Rightarrow 2^{k+1} \geq \frac{1}{\varepsilon}$$

$$\Rightarrow (k+1)\ln 2 \geq \ln\frac{1}{\varepsilon}$$

$$\Rightarrow k \geq \frac{\ln\frac{1}{\varepsilon}}{\ln 2} - 1$$

□

Note: Depending on how an "iteration" is defined, the result may differ by 1.

3. Consider the matrix

$$A = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}.$$

(a) Compute the condition number of A in a matrix norm of your choice. What happens if ε is close to 1?

(b) Show that the Jacobi method for solving $Ax = b$ converges when $|\varepsilon| < 1$.

(c) Show that the Gauss–Seidel method for solving $Ax = b$ converges when $|\varepsilon| < 1$. Which method converges faster?
Hint: Recall that the Gauss–Seidel method is based on the splitting $A = P + (A - P)$ where P contains the right upper triangular entries of A. So

$$P = \begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad P^{-1} = \begin{pmatrix} 1 & -\varepsilon \\ 0 & 1 \end{pmatrix}.$$

(10+10+10)

(a) Let's take the 2-norm for simplicity. Since A is symmetric,

$$\|A\|_2 = \max_i |\lambda_i|$$

where λ_i are the eigenvalues of A.

Here: $p_A(\lambda) = (1-\lambda)^2 - \varepsilon^2$, so the eigenvalues are

$$\lambda_{1,2} = 1 \pm \varepsilon$$

Assuming that ε is positive (with obvious modification for $\varepsilon < 0$),

$$\|A\|_2 = 1 + \varepsilon$$

$$\|A^{-1}\|_2 = \frac{1}{|1-\varepsilon|}$$

(If λ_i is an eigenvalue of A, then $\frac{1}{\lambda_i}$ is an eigenvalue of A^{-1} ∇)

$$\Rightarrow \text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{1+\varepsilon}{|1-\varepsilon|}$$

So $\text{cond}_2(A) \to \infty$ as $\varepsilon \to 1$.

(b) For the Jacobi method, need to look at the eigenvalues of

$$B_J = I - P_J^{-1} A$$

where $P_J = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

$$\Rightarrow B_J = \begin{pmatrix} 0 & -\varepsilon \\ -\varepsilon & 0 \end{pmatrix} \quad \text{with eigenvalues } \lambda_{1,2} = \pm\varepsilon$$

$$\Rightarrow \rho(B_J) < 1 \quad \text{iff} \quad |\varepsilon| < 1 .$$

(c) $B_{GS} = I - P_{GS}^{-1} A$ where $P_{GS} = \begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix}$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \underbrace{\begin{pmatrix} 1 & -\varepsilon \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}}_{= \begin{pmatrix} 1-\varepsilon^2 & 0 \\ \varepsilon & 1 \end{pmatrix}}$$

$$= \begin{pmatrix} \varepsilon^2 & 0 \\ -\varepsilon & 0 \end{pmatrix}$$

The eigenvalues are clearly 0 and ε^2

$$\Rightarrow \rho(B_{GS}) = \varepsilon^2 < 1 \quad \text{iff} \quad |\varepsilon| < 1 .$$

The Gauss-Seidel method converges faster, because for $|\varepsilon| < 1$

$$\rho(B_{GS}) = \varepsilon^2 < |\varepsilon| = \rho(B_J) .$$

4. Compute the LU decomposition of the matrix

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

(10)

Step 1: Add -1 times row 1 to row 2:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Step 2: Add row 2 to row 3:

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

$$\Rightarrow A = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}}_{= L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}} \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 2 \end{pmatrix}}_{=: U}$$

5. Consider the sequence

$$I_0 = 1 - e^{-1},$$
$$I_n = 1 - n I_{n-1}.$$

(a) What is the condition number of computing I_n in terms of I_{n-1}?
What is the condition number of recursively computing I_n in terms of I_0?

(b) Is this recursion a good method to compute I_n? Explain!
Hint: $I_n \to 0$ as $n \to \infty$; see below.

(c) **Extra credit:** Show that $I_n \to 0$ as $n \to \infty$.
Hint: Show that

$$I_n = \int_0^1 (1-x)^n e^{-x}\, dx.$$

(10+10+*10*)

(a) $\Delta I_n = -n\, \Delta I_{n-1}$

$\Rightarrow \kappa_{abs} = n$

$\Delta I_n = \pm\, n(n-1) \cdot \ldots \cdot 1\ \Delta I_0$

$\Rightarrow \kappa_{abs} = n!$

(If you like relative condition numbers: For n large,

$I_n \sim \frac{1}{n+1}$

$\Rightarrow \frac{I_n}{I_{n-1}} \sim 1$

$\Rightarrow \kappa_{rel} \sim \kappa_{abs}$ ☺)

(b) No, for two reasons:

- For large n, the condition number is extremely large, so rounding errors are massively amplified
- Since $I_n \to 0$ as $n \to \infty$, the algorithm repeatedly subtracts numbers of almost equal size
  $\Rightarrow$ cancellation of significant digits
  $\Rightarrow$ massive introduction of rounding errors.

(c)
$$\int_0^1 e^{-x}\,dx = -e^{-x}\Big|_0^1 = 1 - \frac{1}{e}$$

$$\underbrace{\int_0^1 (1-x)^n e^{-x}\,dx}_{\equiv I_n} = \underbrace{-(1-x)^n e^{-x}\Big|_0^1}_{=1 \text{ for } n>0} - n\underbrace{\int_0^1 (1-x)^{n-1} e^{-x}\,dx}_{\equiv I_{n-1}}$$

$\Rightarrow$ The integral satisfies the given recursion relation.

Furthermore,

$$0 \le I_n = \int_0^1 (1-x)^n e^{-x}\,dx \le \int_0^1 (1-x)^n\,dx = \frac{1}{n+1} \to 0$$

as $n \to \infty$

□

6. The following root finding method is a modification of the bisection method. It is called *regula falsi*.

$$x_{k+1} = x_0 - \frac{x_k - x_0}{f(x_k) - f(x_0)} f(x_0).$$

(a) Show that the *regula falsi* is consistent.
(Recall that a method is consistent if every fixed point ξ of this iteration solves the equation $f(\xi) = 0$.)

(b) Give an argument using Taylor expansion that the *regula falsi* is convergent with order 1.

(c) **Extra credit:** Show that if f is continuous, and two starting values x_0 and x_1 are chosen so that $f(x_0)$ and $f(x_1)$ have opposite sign, then the *regula falsi* will always converge. (*)
Hint: Try to find a geometric interpretation of the method, then note that the sequence ~~is monotonic~~. of bracketing intervals is monotonic.

(10+10+*10*)

(a) Assume that ξ is a fixed point of this iteration and that $\xi \neq x_0$ (otherwise the algorithm will fail!)

$$\Rightarrow \xi = x_0 - \frac{\xi - x_0}{f(\xi) - f(x_0)} f(x_0)$$

$$\Rightarrow (\xi - x_0)(f(\xi) - f(x_0)) = -(\xi - x_0) f(x_0)$$

$$\Rightarrow f(\xi) \underbrace{(\xi - x_0)}_{\neq 0} = 0$$

$$\Rightarrow f(\xi) = 0.$$

(b) As usual, write $\delta_k = x_k - \xi$.

$$\Rightarrow \delta_{k+1} = \delta_0 - \frac{\delta_k - \delta_0}{f(\xi + \delta_k) - f(x_0)} f(x_0)$$

Now Taylor expand to order 1:

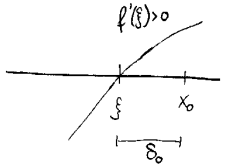
$$f(\xi + \delta_k) = \underbrace{f(\xi)}_{=0 \text{ because of consistency}} + f'(\xi)\,\delta_k + \dots$$

$$\Rightarrow \left(\delta_{k+1} - \delta_0\right)\left(f'(\xi)\,\delta_k - f(x_0)\right) \approx -(\delta_k - \delta_0)\,f(x_0)$$

$$\Rightarrow -\delta_{k+1} f(x_0) - \delta_0 f'(\xi)\,\delta_k - \delta_0 f(x_0) \approx -\delta_k f(x_0) - \delta_0 f(x_0)$$

(Note that the term $\delta_{k+1}\,\delta_k\,f'(\xi)$ is very small near the fixed point and is therefore neglected.)

$$\Rightarrow \delta_{k+1} \approx \left(1 - \frac{\delta_0 f'(\xi)}{f(x_0)}\right)\delta_k$$



Note that if $|\delta_0|$ is small enough, then

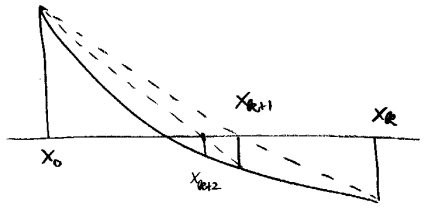
$$0 < \frac{\delta_0 f'(\xi)}{f(x_0)} \approx 1 \qquad \text{(see picture)}$$

$$\Rightarrow \left| 1 - \frac{\delta_0 f'(\xi)}{f(x_0)} \right| < 1$$

$$\Rightarrow \delta_k \to 0 \quad \text{at least linearly.}$$

Note: This is not a proof, but can be turned into a proof by using the MVT instead of truncated Taylor expansion.

(c) The geometric picture is this: x_{k+1} is the zero of the straight line connecting $(x_0, f(x_0))$ and $(x_k, f(x_k))$:



Note: x_0 and x_{k+1} must always bracket the zero of f. If $f(x_k)$ and $f(x_{k+1})$ have opposite sign, we must re-initialize by setting $x_0 := x_k$.

Thus, the zero remains bracketed by a nested sequence of intervals with monotonically increasing lower boundary and monotonically decreasing upper boundary. The length of the interval decreases as follows:

Case 1 (no re-initialization): $x_{k+1} - x_0 = (x_k - x_0)\,\underbrace{\frac{f(x_0)}{f(x_0) - f(x_k)}}_{\in (0,1)}$

$$\Rightarrow L_{k+1} = L_k \frac{1}{1 - \frac{y_k}{y_0}}$$

Case 2 (with re-initialization): $x_{k+1} - x_k = (x_0 - x_k)\,\underbrace{\left(1 - \frac{f(x_0)}{f(x_0) - f(x_k)}\right)}_{\in (0,1)}$

$$\Rightarrow L_{k+1} = L_k \frac{1}{1 - \frac{y_0}{y_k}}$$

where L_k is the length of the bracketing interval at step k.

If there is only a finite number of re-initializations, one end of the sequence of intervals is fixed after a finite number of steps, and the other "loose" end converges by virtue of being a bounded monotonic sequence.

Otherwise there must be an infinite number of re-initializations, which we can group into pairs. Let's look at a single such pair with one re-initialization at step 1, the other at step k. Then

$$L_{k+1} = L_1 \frac{1}{1-\frac{y_0}{y_1}} \underbrace{\frac{1}{1-\frac{y_2}{y_1}} \frac{1}{1-\frac{y_3}{y_1}} \cdots \frac{1}{1-\frac{y_{k-1}}{y_1}}}_{\text{no re-init; each term is} \leq 1} \frac{1}{1-\frac{y_1}{y_k}}$$

$$\leq L_1 \frac{1}{\left(1-\frac{y_0}{y_1}\right)\left(1-\frac{y_1}{y_k}\right)}$$

Since y_0 and y_k are at the same end of the sequence of intervals, y_1 is at the other end, we have for the sequence of pairs of re-initializations that

$$\frac{y_0^{(p)}}{y_0^{(p)}} \to \lambda < 0 \qquad \frac{y_1^{(p)}}{y_k^{(p)}} \to \frac{1}{\lambda} < 0 \qquad \text{for } p \to \infty$$

$$\Rightarrow L_{k+1}^{(p)} \leq L_1^{(p)} \left(\frac{1}{2} + \varepsilon\right) \qquad \text{for any } \varepsilon > 0 \text{ and } p \text{ sufficiently large}$$

$\Rightarrow L_{k+1}^{(p)}$, the length of the bracketing interval, converges to zero as $p \to \infty$, i.e. x_k converges as well. □