TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 13th of January, 2024.

Allowed:

1. Pocket calculator

Examiner: Fredrik Berntsson

Marks: 25 points total and 10 points to pass.

Jour: Fredrik Berntsson - (telefon 013 28 28 60)

Good luck!

(5p) **1:** **a)** Let $x = 113.782378$ be an exact value and let $\bar{x} = 113.81$ be an approximation of x. Give a bound for both the *absolute error* and the *relative error* in $\bar{x}$. Also how many *significant digits* do the approximate value $\bar{x}$ have?

b) We want to store the number $x = 18.7892189$ on a computer using the floating point system $(10, 5, -10, 10)$. What approximate number $\bar{x}$ would actually be stored on the machine?

c) Let $\bar{a}$ and $\bar{b}$ be two positive real numbers, with small errors Δa and Δb. Clearly explain why it might be problematic to compute $\bar{a} - \bar{b}$. Also, explain why computing $\bar{a} + \bar{b}$ doesn't cause the same problems.

d) Let $z = (1+y)\exp(x/2)$, where $x = 0.29 \pm 0.02$, and $y = 0.62 \pm 0.03$. Compute the approximate value $\bar{z}$ and give an error bound.

(2p) **2:** Let x_1, x_2, x_3 and x_4 be given interpolation points. In the Lagrange interpolation formula we use basis functions $\ell_i(x)$ such that $\ell_i(x_j) = 1$ if $i = j$ and zero otherwise. Give an explicit expression for the basis function $\ell_2(x)$ for the case with $n = 4$ interpolation points. What is the degree of the basis polynomial?

(3p) **3:** We need to evaluate the function

$$f(x) = \mathrm{e}^x - 3x$$

for small x on a computer with machine precision $\mu = 1.11 \cdot 10^{-16}$. Perform a computational error analysis and find a bound for the relative error in the computed value $f(x)$. When doing the analysis you should assume that all computations are done with a relative error at most μ. Also use your error bound to determine if *catastrophic cancellation* occurs during the computations. If cancellation occurs also suggest an alternative formula that should give better accuracy.

(3p) **4:** Non-linear equations $f(x) = 0$ can be solved using fixed point iteration where the problem is reformulated so that a root x^*, i.e. $f(x^*) = 0$, is a fixed point to the iteration $x_{n+1} = g(x_n)$, that is $x^* = g(x^*)$.

a) Show that the iteration $x_{n+1} = g(x_n)$ is convergent if $|g'(x^*)| \leq C < 1$ and the starting guess x_0 is sufficiently close to the root.

b) The equation $f(x) = \mathrm{e}^{-x^2} - x = 0$ has a root $x^* \approx 0.65$. Formulate a fixed point iteration for finding a root to $f(x) = 0$ and show that the proposed method is convergent.

c) The equation $f(x) = \mathrm{e}^{-x^2} - x = 0$ is solved using the Newton-Raphson method and an approximate root $\bar{x} = 0.652919 \approx x^*$ is obtained. Estimate the error in the approximation $\bar{x}$.

(3p) **5:** Do the following

a) A computer program has computed the decomposition $PA = LU$ and the output is

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -0.7 & 1 & 0 \\ 0.3 & 1.8 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1.7 & -2.3 & -1.4 \\ 0 & 1.2 & -0.5 \\ 0 & 0 & 3.1 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Determine if pivoting was used correctly during the computations.

b) Let A and B be $n \times n$ matrices and x, y, be $n \times 1$ vectors. How many floating point operations are required to implement the formula

$$z = (I + A)(Bx + y),$$

where I is the identity matrix, as efficiently as possible? In a practical test one implementation of the formula was tested on a computer and the following run times were reported

n	1000	2000	4000	8000
time (ms)	1060	8360	66300	529000

Was the implementation done using the most efficient method? Motivate your answer carefully.

(4p) **6:** Do the following:

a) Let $p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$ be a cubic polynomial. We want to find values for the coefficients so that $p(0) = p(1) = 0$ and $p'(0) = p'(1) = 1$. Show how to derive a linear system of equations such that the solution $c = (c_0, c_1, c_2, c_3)^T$ are the coefficients of a cubic polynomial satisfying these conditions. Also find the specific polynomial satisfying all the above conditions.

b) Spline interpolation can be used to approximnate a function $y = f(x)$. We have a table

x	-2	-1	0	1	2
$f(x)$	0	1	3	1	0

We attempt to approximate $f(x)$ by a cubic spline $s(x)$. Clearly state the conditions that have to be satisfied for $s(x)$ to be a cubic spline that interpolates the above table. Also state if the given information sufficient for the spline $s(x)$ to be uniquely determined?

(2p) **7:** A numerical method, depends on a discretization parameter h, and has a truncation error that can be described as $R_T \approx Ch^p$. We use the method to compute a few approximations $T(h)$ of the exact result $T(0)$ and obtain

h	0.9	0.3	0.1
T(h)	2.3721	2.1409	2.1152

Use the table to determine C and p. Also estimate the value of h needed for the error to be of magnitude 10^{-4}.

(3p) **8:** Let A be an $m \times n$, $m > n$, matrix.

a) Suppose that A has full column rank. Use the singular value decomposition $A = U\Sigma V^T$ to give a general formula for the solution to the least squares problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2.$$

Clealy motivate your answer.

b) The least squares solution x, see **a)**, is also a solution to the linear system $Ax = b$ if $b \in \text{Range}(A)$. Use the singular value decomposition to give a basis for the space Range(A). Also formulate a criteria that uses the vector b and also the basis for Range(A) to check if the least squares solutuion x is also a solution to the linear system $Ax = b$. Clearly motivate your answer.

(5p) **1:** For **a)** the absolute error is $|\Delta x| = |113.782378 - 113.81| < 0.03$. The relative error is $|\Delta x|/|x| < 0.03/113.81 < 2.7 \cdot 10^{-4}$. Since the absolute error satisfies $|\Delta x| < 0.05 = 0.5 \cdot 10^{-1}$ we have one correct decimal and thus 4 significant digits.

In **b)** we rewrite the number as $x = 1.87892189 \cdot 10^1$ to see that $\bar{x} = 1.87892 \cdot 10^1$ is actually stored on the computer.

For **c)** problems can occur if $\bar{a}$ and $\bar{b}$ is of approximately equal magnitude since in that case $\bar{a} - \bar{b}$ is much smaller than either of $\bar{a}$ or $\bar{b}$. This means that the resulting *relative error* in the result may be very large. This is called *catastrophic cancellation.* For the addition the result $\bar{a}+\bar{b}$ is always larger than $\bar{a}$ or $\bar{b}$. Thus the result cannot have a large relative error (unless either of $\bar{a}$ or $\bar{b}$ has a large relative error).

For **d)** The approximate value is $\bar{z} = (1+\bar{y})\exp(\bar{x}/2) = (1+0.62)\exp(0.29/2) = 1.87$ with $|R_B| \leq 0.003$. The error propagation formula gives

$$|\Delta z| \lesssim |\frac{\partial z}{\partial x}||\Delta x| + |\frac{\partial z}{\partial y}||\Delta y| = |\frac{1}{2}(1+y)\exp(x/2)||\Delta x| + |\exp(x/2)||\Delta y| < 0.054$$

The total error is $|R_{TOT}| \leq 0.054 + 0.003 < 0.06$. Thus $y = 1.87 \pm 0.06$.

(3p) **2:** The basis function satisfies $\ell_2(x_2) = 1$ and $\ell_2(x_i) = 0$, $i \neq 2$. Thus

$$\ell_2(x) = \frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)}.$$

The degree of $\ell_2(x)$ is $n = 3$.

(2p) **3:** The computational order is

$$f(x) = \mathrm{e}^x - 3x = a - 3x = a - b = c,$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| + |\frac{\partial f}{\partial c}||\Delta c| = |1||\Delta a| + |1||\Delta b| + |1||\Delta c| \lesssim$$

$$\mu(|a| + |b| + |c|) \approx \mu(|1| + |3x| + |1|) \approx 2\mu,$$

where we have used that $\mathrm{e}^x \approx 1$ and $f(x) = c \approx 1$, when x is small.There is no risk of cancellation here. The relative error is bounded by 2μ (since $f(x) \approx 1$).

(3p) **4:** For **a)** we let x^* be the fixed point. Then

$$|x_n - x^*| = |g(x_{n-1}) - g(x^*)| = |g'(\xi_n)||x_{n-1} - x^*|,$$

where $\xi_n \in (x_{n-1}, x^*)$. Since $|g'(x^*) \leq C < 1$ it will hold that $|g'(\xi_n)| \leq C' < 1$, provided that the previous iterate x_{n-1} is close enough to x^*. This means that the iterations will converge.

For **b)** the easiest possible method would be $x_{n+1} = g(x_n) = \mathrm{e}^{-x_n^2}$. We compute $g'(x) = -2x\mathrm{e}^{-x^2}$, and $|g'(0.65)| = |-0.8520| < 1$. Thus the iterations will converge.

For **c)** we let $\bar{x} = 0.652919$ and use the error estimate

$$|\bar{x} - x^*| = \frac{|f(\bar{x})|}{|f'(\bar{x})|} \approx \frac{|-6.67 \cdot 10^{-7}|}{|-1.85|} < 4 \cdot 10^{-7}.$$

(3p) **5:** For **a)** we simply observe that correct pivoting means that $|L_{ij}| \leq 1$ but here $|L_{32}| = 1.8$.

For **b)** we note that we first evaluate $z_1 = Bx + y$. A matrix vector multiply Bx requires approximately $2n^2$ arithmetic operations and the vector addition requires n additions. The remaining computation is $(I+A)z_1 = z_1 + Az_1$ which is exactly the same computation as before. Thus the arithmetic work involved in the computation should be approximately $4n^2$. This counts both multiplications and additions.

With the assumption that computation time is roughly proportional to the amount of arithmetic work the time can be written as $t(n) \approx cn^2$. Thus $t(2n)/t(n) \approx (c(2n)^2)/(cn^2) = 4$. Thus double n should mean 4 times longer computation time. In the table however $t(2000)/t(1000) \approx 8$. We conclude that the formula was not implemented efficiently.

Its likely that a matrix–matrix multiply, e.g. AB, was evaluated at some point since thats an $\mathcal{O}(n^3)$ operation and $2^3 = 8$.

(4p) **6:** For **a)** we note that $p(0) = c_0 = 0$ and $p(1) = c_0 + c_1 + c_2 + c_3 = 0$ gives two equations. Then $p'(x) = c_1 + 2c_2x + 3c_3x^2$ so we also obtain $p'(0) = c_1 = 1$ and $p'(1) = c_1 + 2c_2 + 3c_3 = 1$. Thus the system of equations is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

We can solve the linear system by noting that $c_0 = 0$ and $c_1 = 1$. Then we are left with two equations for c_2 and c_3. The solution is $p(x) = x - 3x^2 + 2x^3$.

For **b)** the conditions for $s(x)$ to be a cubic spline are (i) on each sub interval $[x_i, x_{i+1}]$ the spline $s(x)$ should be given by a cubic polynomial, and (ii) $s(x)$, $s'(x)$ and $s''(x)$ should be continuous on the whole interval $[x_1, x_n]$. Also (iii) the interpolation conditions $s(x_i) = f(x_i)$ needs to be satisfied. The given information is not sufficient since we also need two end point conditions for the spline to be unique.

(2p) **7:** Since $T(h) = T(0) + Ch^p$ we get

$$\frac{T(9h) - T(3h)}{T(3h) - T(h)} \approx \frac{(9^p - 3^p)Ch^p}{(3^p - 1^p)Ch^p} = 3^p$$

Insert numbers from the table we obtain

$$3^p = \frac{2.3721 - 2.1409}{2.1409 - 2.1152} = 8.9767$$

Which fits very well with $p = 2$. In order to determine C we use the last equation $T(3h) - T(h) = (3^2 - 1^2)Ch^2$ and insert $h = 0.1$ to obtain $C = 0.321$. Finally $R_T = 10^{-4}$ if $h = \sqrt{10^{-4}/0.321} = 0.0177$. Thus $h < 0.0017$ is required.

(3p) **8:** For **a)** we write x using the basis given by the columns of the V matrix as

$$x = \sum_{i=1}^{n} c_i v_i.$$

In order to determine x we compute

$$Ax = \sum_{i=1}^{n} c_i A v_i = \sum_{i=1}^{n} c_i \sigma_i u_i,$$

and write b in the U basis,

$$b = \sum_{i=1}^{m} (u_i^T b) u_i.$$

We see that

$$Ax - b = \sum_{i=1}^{n} (c_i \sigma_i - (u_i^T b)) u_i + \sum_{i=n+1}^{m} (u_i^T b) u_i.$$

Thus

$$x = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i,$$

sets the first n coefficients to zero. This will minimize the norm $\|Ax - b\|_2$. The rest of the coefficients we can't influence.

For **b)** we note that the residual is

$$r = Ax - b = \sum_{i=n+1}^{m} (u_i^T b) u_i.$$

If the residual is $r = 0$ then we have a solution to the linear system $Ax = b$. Thus the criteria for existance could be written as

$$u_i^T b = 0, \qquad \text{for } i = n+1, \ldots, m.$$

Since the range Range(A) has basis $\{u_1, u_2, \ldots, u_n\}$ we have to formulate it a bit differently. One way is to say that

$$b - \sum_{i=1}^{n} (u_i^T b) u_i = 0.$$