TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 15th of January, 2022.

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 $a = 0.008755661$ be an exact value. Round the value a to 5 *correct decimals* to obtain an approximate value $\bar{a}$. Also give a bound for the *relative error* in $\bar{a}$.

b) We want to store the number $x = 117.2277634$ 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) Explain why the formula $y = \sqrt{1+x} - 1$ can give poor accuracy when evaluated, for small x, on a computer. Also propose an alternative formula that can be expected to work better.

d) Let $y = e^{-2x}$, where $x = 0.95 \pm 0.02$. Compute the approximate value $\bar{y}$ and give an error bound.

(3p) **2:** Let the table,

x	0.6	0.8	1.0
$f(x)$	1.3	1.1	1.2

of correctly rounded function values, be given. Do the following

a) Use Lagrange interpolation formula to write an explicit expression for the polynomial that interpolates the above table.

b) Suppose the value $f(0.6) = 1.3$ has an error and we actually have $f(0.6) = 1.3 \pm 0.03$. Find the maximum error in the interpolating polynomial, for the interval $0.6 < x < 1.0$, due to the error in the function value $f(0.6)$.

(2p) **3:** We compute the function

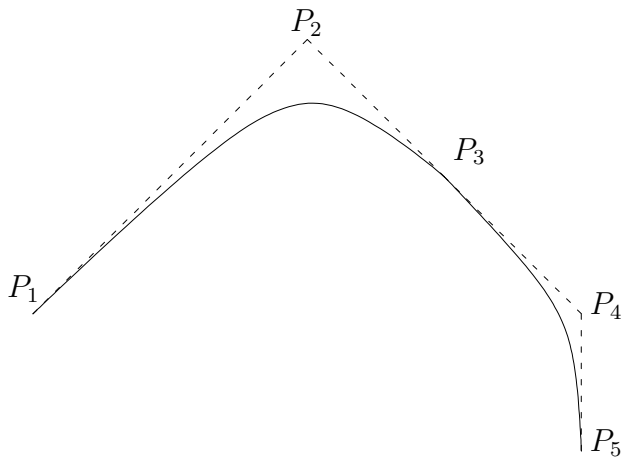
$$f(x) = \mathrm{e}^x - 3x$$

for small x values on a computer with unit round off $\mu = 1.11 \cdot 10^{-16}$. Preform an analysis of the computational errors to obtain a bound for the relative error in the computed results $f(x)$. For the analysis you may assume that all computations are performed with a relative error at most μ. Also, use the obtained bound to argue if *cancellation* occurs during the computations. In case of cancellation also suggest an alternative formula that can be expected to give better accuracy.

(3p) **4:** A quadratic Beziér curve is given by the expression

$$p(t) = (1-t)^2 P_1 + 2(1-t)tP_2 + t^2 P_3, \quad 0 < t < 1,$$

where P_1, P_2 and P_3 are control points. Suppose we want to combine two quadratic Beziér curves to one single corve. For this purpose we chose five control points as follows



The point P_3 is common for both curves. We have chosen $P_2 = (2\,,\,6)^T$, $P_3 = (3\,,\,5)^T$ and $P_5 = (6\,,\,1)$. Find coordinates for the point P_4 such that the tangent direction of the combined curve is continuous at the point P_3 and that the tangent is vertical at the endpoint P_5. Motivate your choice for P_4 carefully.

(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. Motivate your answer!

b) Find a Gauss transformation M such that

$$M \begin{pmatrix} 2 \\ 3 \\ 0.6 \\ -1.8 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \\ 0 \end{pmatrix}.$$

c) Explain what is ment by a matrix norm beeing *induced* from a vector norm. Also show that if A and B are matrices then for an induced norm $\|AB\| \leq \|A\|\|B\|$.

(4p) **6:** Suppose $A \in \mathbb{R}^{m\times n}$, $m > n$. The least squares method can be used to minimize

$$\|Ax - b\|_2.$$

Do the following:

a) Suppose we have a set of measurements (x_k, y_k) for $k = 1, \ldots, m$. We want to adapt a function of the type

$$y_k \approx c_1 + c_2 \mathrm{e}^{-x_k} + c_3 \sin(\pi x_k) + c_4 x_k \sin(\pi x_k)$$

to the measurements by using the least squares method. Clearly show what the matrix A and the right hand side b is for this particular case.

b) Let A be an $m \times n$, $m > n$, matrix, and let $A = Q_1 R$ be the *reduced* QR decomposition. Give the dimensions for Q_1 and R. Also give a formula for computing the solution to the least squares problem $Ax = b$ using the reduced QR decomposition. Finally estimate the amount of arithmetic work required to compute the least squares solution (not counting the work needed to compute the QR decomposition itself).

c) Show that if $\|\cdot\|$ is an *induced norm* and Q is orhogonal then $\|AQ\| = \|A\|$.

(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	0.4	0.2	0.1
T(h)	4.0272	3.9240	3.8970

Use the table to determine C and p. Also estimate the value of h needed for the error to be of magnitude 10^{-3}.

(3p) **8:** **a)** Suppose the $n \times n$ matrix A has rank $k < n$ and that the linear system of equations $Ax = b$ has a solution. Use the singular valur decomposition $A = U\Sigma V^T$ to give a general formula for all solutions x of the system $Ax = b$. Clealy motivate your answer.

b) Let A be an $m \times n$ matrix, $m > n$. Show how the singular value decomposition $A = U\Sigma V^T$ can be used for solving the minimization problem

$$\min_{\|x\|_2=1} \|Ax\|_2.$$

Give both the minmizer x and the minimum in terms of singular values and singular vectors.

(5p) **1:** For **a)** we obtain the approximate value $\bar{a} = 0.00876$ which has 5 correct decimal digits. The absolute error is at most $|\Delta a| \leq 0.5 \cdot 10^{-5}$ and thus the *relative error* is bounded by $\Delta a|/|a| \leq 0.5 \cdot 10^{-5}/0.00876 \leq 0.58 \cdot 10^{-3}$.

In **b)** we rewrite the number as $x = 1.172277634 \cdot 10^2$ to see that $\bar{x} = 1.17228 \cdot 10^2$ is actually stored on the computer.

For **c)** Since $\sqrt{1+x} \approx 1$, for small x, we *catastrophic cancellation* will occur when $\sqrt{1+x)} - 1$ is computed resulting in a large relative error in the result. A better formula would be

$$\sqrt{1+x)} - 1 = \frac{(\sqrt{1+x)} - 1)(\sqrt{1+x)} + 1)}{\sqrt{1+x)} + 1} = \frac{x}{\sqrt{1+x)} + 1}$$

where the cancellation is removed.

For **d)** The approximate value is $\bar{y} = e^{-2\bar{x}} = \exp(-2 \cdot 0.95) = 0.15$ with $|R_B| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial x}||\Delta x| = |-2 \cdot \exp(-2 \cdot 0.95)||\Delta a| < 0.006.$$

The total error is $|R_{TOT}| \leq 0.006 + 0.5 \cdot 10^{-2} < 0.011$. Thus $y = 0.15 \pm 0.02$.

(3p) **2:** For **a)** the polynomial is

$$p_2(x) = 1.3\frac{(x-0.8)(x-1.0)}{(0.6-0.8)(0.6-1.0)} + 1.1\frac{(x-0.6)(x-1.0)}{(0.8-0.6)(0.8-1.0)} + 1.2\frac{(x-0.6)(x-0.8)}{(1.0-0.6)(1.0-0.8)}.$$

There is no reason to simplify the expression further.

For **b)** we note that if the function value $f(0.6) = f_1 = 1.3$ has en error then the Lagrange polynomial changes as

$$\bar{p}_2(x) = p_2(x) + \Delta f_1 \frac{(x-0.8)(x-1.0)}{(0.6-0.8)(0.6-1.0)}.$$

The function $|(x-0.8)(x-1.0)|$ has a local maximum for $x = 0.9$ and also a local maximum at $x = 0.6$. The largest absolute value is achived for $x = 0.6$ which means that

$$|\bar{p}_2(x) - p_2(x)| \leq |\Delta f_1| \frac{(0.6-0.8)(0.6-1.0)}{(0.6-0.8)(0.6-1.0)} = |\Delta f_1| \leq 0.03.$$

(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 $\mathrm{e}^x \approx 1$, $f(x) = c \approx 1$ since x is small. There is no cancellation present in these calculations. Everything turns out fine and both the absolute and relative errors are bounded by 2μ (since the function value $f(x) \approx 1$).

(3p) **4:** We note that for the tangent to be vertical at P_5 the x–coordinate need to be the same at P_4 and P_5. Thus $P_4 = (6, , \alpha)^T$ for some real number α. In order to get a continuous tangent direction at P_3 we need the vectors $P_3 - P_2 = (1\,, -1)^T$ to be parallell with $P_4 - P_3 = (3\,, \alpha - 5)^T$ which only works out if $\alpha = 2$. Thus $P_4 = (6\,, 2)^T$.

(3p) **5:** For **a)** we just observe that one of the multipliers (i.e. $\ell_{32} = 1.8$) is larger than one. Thus pivoting wasn't used correctly.

For **b)** The multipliers are $m_3 = 0.6/3 = 0.2$ and $m_4 = -1.8/3 = -0.6$. Therefore the Gauss transformation is

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.2 & 1 & 0 \\ 0 & -0.6 & 0 & 1 \end{pmatrix}.$$

For **c)** A matrix norm is *induced* if its definition is based on a vector norm, i.e.

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

For such norms we have

$$\|AB\| = \max_{x \neq 0} \frac{\|ABx\|}{\|x\|} = \max_{x \neq 0} \frac{\|ABx\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \leq \left(\max_{x \neq 0} \frac{\|ABx\|}{\|Bx\|}\right)\left(\max_{x \neq 0} \frac{\|Bx\|}{\|x\|}\right) \leq \max_{y \neq 0} \frac{\|Ay\|}{\|y\|}\|B\| \leq \|A\|\|$$

(4p) **6:** For **a)** we note that each data point (x_i, y_i) gives one row of the over determined system $Ax = b$. The model is $y = c_1 + c_2\mathrm{e}^{-x} + c_3 \sin(\pi x) + c_4 x \sin(\pi x)$. Thus the system $Ax = b$ is

$$\begin{pmatrix} 1 & \mathrm{e}^{-x_1} & \sin(\pi x_1) & x_1 \sin(\pi x_1) \\ 1 & \mathrm{e}^{-x_2} & \sin(\pi x_2) & x_2 \sin(\pi x_2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{e}^{-x_m} & \sin(\pi x_m) & x_m \sin(\pi x_m) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

For **b)** The dimensions are $m \times n$ for Q_1 and $n \times n$ for R. The formula is $x = R^{-1}(Q_1^T b)$ and the matrix vector multiply $y = Q_1^T b$ requires approimately mn multiplications and additions. Since R is triangular computing $R^{-1}y$ by backwards substitution requires $n^2/2$ multiplications and additions. So the operation count is $2mn + n^2 \approx 2mn$ if $m >> n$.

Finnally, for **c)** we have

$$\|AQ\| = \max_{x \neq 0} \frac{\|AQx\|}{\|x\|} = \{ \text{ set } y = Qx \text{ and note } \|Qx\| = \|y\| \} = \max_{y \neq 0} \frac{\|Ay\|}{\|y\|} = \|A\|$$

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

$$\frac{T(4h) - T(2h)}{T(2h) - T(h)} \approx \frac{(4^p - 2^p)Ch^p}{(2^p - 1^p)Ch^p} = 2^p$$

Insert numbers from the table we obtain

$$2^p = \frac{4.0272 - 3.9240}{3.9240 - 3.8970} = 3.8224$$

Which fits resonably well with $p = 2$. In order to determine C we use the last equation $T(2h) - T(h) = (2^2 - 1^2)Ch^2$ and insert $h = 0.1$ to obtain $C = 0.9$. Finally $R_T = 10^{-3}$ if $h = \sqrt{10^{-3}/0.9} = 0.035$. Thus $h < 0.035$ is required.

(3p) **8:** For **a)** we note that if $\text{rank}(A) = k$ then $\{v_{k+1}, \ldots, v_n\}$ is a basis for $\text{null}(A)$ and $\{v_1, \ldots, v_k\}$ is a basis for its orthogonal complement $(\text{null}(A))^\perp$. Thus for evey x we can write

$$x = x_1 + x_2 = (\sum_{i=1}^{k} c_i v_i) + (\sum_{i=k+1}^{n} c_i v_i).$$

In order to determine x_1 we compute

$$Ax = A(x_1 + x_2) = Ax_1 + 0 = \sum_{i=1}^{k} c_i \sigma_i u_i = b = \sum_{i=1}^{n} (u_i^T b) u_i.$$

Where $(u_i^T b) = 0$, for $i = k+1, \ldots, n$, since it is said that the solution exists. Thus

$$x_1 = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i \text{ and } x_2 = \sum_{i=k+1}^{n} c_i v_i,$$

where $c_i, i = k+1, \ldots, n$, are undetermined parameters.

For **b)** we use the singular value decomposition to write $\|Ax\|_2 = \|U\Sigma V^T x\|_2 = \|\Sigma y\|_2$, where $y = V^T x$. Since V is orthogonal $\|x\|_2 = \|y\|_2$. Thus the minimization problem is equivalent to

$$\min_{\|y\|_2 = 1} \|\Sigma y\|_2^2 = . \min_{\|y\|_2 = 1} \sum_{i=1}^{n} \sigma_i^2 y_i^2 \geq \sigma_n^2 \sum_{i=1}^{n} y_i^2 = \sigma_n^2,$$

since σ_n is the smallest singular value, with equality if $y = e_n$ which means that $x = V^T e_n = v_n$.