TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 16th of January, 2021.

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 $a = 0.08852661$ be an exact value. Round the value a to 4 *correct decimals* to obtain an approximate value $\bar{a}$. Also give a bound for the *relative error* in $\bar{a}$.

b) Let $x = 24.2231$. Give a bound for the *relative error* when x is stored on a computer using the floating point system $(10, 5, -10, 10)$.

c) Explain why the formula $y = \cos x - 1$ 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 $y = \log a$, where $a = 3.87 \pm 0.03$. Compute the approximate value $\bar{y}$ and give an error bound.

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

x	0.4	0.7	0.9
$f(x)$	1.284	1.413	1.475

of correctly rounded function values, be given. Use linear interpolation to approximate the function value $f(0.58)$. Also give a complete error estimate.

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

$$f(x) = 1 - 2x\cos(x)$$

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:** 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 $|g'(x^*)| < C < 1$ and the starting guess x_0 is sufficiently close to the root.

b) The equation $f(x) = \mathrm{e}^{2x} - 2 + 3x = 0$ has a root $x^* \approx 0.19$. Formulate a fixed point iteration for finding a root to $f(x) = 0$ and show that the proposed method is convergent.

c) The equation $f(x) = \mathrm{e}^{2x} - 2 + 3x = 0$ is solved using fixed point iteration and an approximate root $\bar{x} = 0.1845 \approx x^*$ is obtained. Estimate the error in the approximation $\bar{x}$.

(4p) **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) 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.

c) Let

$$A = \begin{pmatrix} 0.3 & -0.9 & -1.3 \\ 2.1 & -0.1 & 0.7 \\ -1.1 & -1.6 & 0.8 \end{pmatrix},$$

and compute $\|A\|_\infty$.

d) Let $r = b - A\widehat{x}$ be the residual for an approximate solution to the linear system $Ax = b$. Prove the formula:

$$\|x - \widehat{x}\| \leq \|A^{-1}\|\|r\|.$$

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

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

a) Draw a clear sketch that shows that x solves the minimization problem if the residual $r = b - Ax$ is orthogonal to all the column vectors from A. Also, clearly, demonstrate how this leads to x satisfying the normal equations.

b) 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

$$y_k \approx c_1 + c_2 x_k + c_3 \cos(\pi x_k) + c_4 \sin(\pi x_k)$$

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.

c) Suppose $A \in \mathbb{R}^{m\times n}$, $m > n$, and that we have the reduced QR decomposition $A = Q_1 R$. Show how the decomposition can be used to find the vector x that minimize $\|Ax - b\|_2$.

d) Suppose the reduced QR decomposition is known. Use the decomposition to write an orthogonal projection P such that the residual is $r = Pb$.

(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)	2.8782	1.8975	1.7885

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)** Let

$$s(x) = \begin{cases} -x^2 + x + 1 & 0 \le x < 1, \\ x^3 - 2x^2 + 2x & 1 \le x < 2. \end{cases}$$

Is $s(x)$ a cubic spline? Motivate your answer

b) Let $P_1 = (1\,,\,0)^T$, $P_2 = (1\,,\,3)^T$, $P_3 = (4\,,\,3)^T$ and $P_4 = (4\,,\,2)^T$. Draw a sketch that clearly shows the convex hull formed by these points. Also use the available information to draw the cubic Beziér curve formed by the four points $P_1, \ldots, P_4$ as accurately as possible.

c) Use the identity $1 = 1^2 = (1 - t + t)^2$ to derive the expression for a quadratic Beziér curve. Also draw a clear sketch that shows an example of a continuously differentiable curve consisting of three different quadratic Beziér curves. The sketch should include all the control points, dashed lines connecting the control points, and also the curve itself. Also state how many control points are needed in total to create the continuous curve.

Answers

(5p) **1:** For **a)** we obtain the approximate value $\bar{a} = 0.0885$ which has 4 correct decimal digits. The absolute error is at most $|\Delta a| \leq 0.5 \cdot 10^{-4}$ and thus the *relative error* is bounded by $\Delta a|/|a| \leq 0.5 \cdot 10^{-4}/0.0885 \leq 0.57 \cdot 10^{-3}$.

In **b)** the unit round off for the floating point system is $\mu = 0.5 \cdot 10^{-5}$. This is an upper bound for the relative error when a number is stored on the computer.

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

$$\cos(x) - 1 = \frac{(\cos(x) - 1)(\cos(x) + 1)}{\cos(x) + 1)} = \frac{\cos^2(x) - 1}{\cos(x) + 1)} = \frac{\sin^2(x)}{\cos^2(x) + 1},$$

where the cancellation is removed.

For **d)** The approximate value is $\bar{y} = \log \bar{a} = \log 3.87 = 1.35$ with $|R_B| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial a}||\Delta a| = |\frac{1}{a}||\Delta a| < 0.008.$$

The total error is $|R_{TOT}| \leq 0.008 + 0.5 \cdot 10^{-2} < 0.013$. Thus $y = 1.35 \pm 0.02$.

(2p) **2:** We use Newtons interpolation formula and the ansatz $p(x) = p_1(x) + R_T(x) = c_0 + c_1(x - 0.4) + c_2(x - 0.4)(x - 0.7)$, where the last term will be used to estimate the truncation error. Inserting the function values from the table leads to $p(0.4) = c_0 = 1.284$ and $p(0.7) = c_0 + c_1(0.3) = 1.413$ which means $c_1 = 0.43$. The last equation is $p(0.9) = c_0 + c_1(0.5) + c_2(0.5)(0.2) = 1.475$ which gives $c_2 = -0.24$. Thus

$$p_1(x) = 1.284 + 0.43(x - 0.4) \text{ and } R_T(x) = -0.24(x - 0.4)(x - 0.7).$$

We obtain $f(0.58) \approx p_1(0.58) = 1.361$ with $|R_B| < 0.5 \cdot 10^{-3}$ and $R_T \leq |-0.24(0.58 - 0.4)(0.58 - 0.7)| < 0.52 \cdot 10^{-2}$. The errors in the function values used also gives an error $R_{XF} < 0.5 \cdot 10^{-3}$ in the result. Thus $f(0.58) = 1.361 \pm 0.62 \cdot 10^{-2} = 1.361 \pm 0.7 \cdot 10^{-2}$.

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

$$f(x) = 1 - 2x\cos(x) = 1 - 2xa + 1 - b = c.$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| + |\frac{\partial f}{\partial c}||\Delta c| = |2x||\Delta a| + |1||\Delta b| + |1||\Delta c| \lesssim$$

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

where we have used $\cos(x) \approx 1$, $f(x) = c \approx 1$ and that 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 μ (since the function value $f(x) \approx 1$).

(3p) **4:** For **a)** we use the mean value theorem and write

$$|x_n - x^*| = |g(x_{n-1}) - g(x^*)| = |g'(\xi)||x_{n-1} - x^*| \leq C|x_{n-1} - x^*|.$$

where $\xi \in [x_{n-1}, x^*]$ which means $|g'(\xi) \leq C$ if x_{n-1} is close enough to the root. We repeat the same argument to obtain $|x_n - x^*| \leq C^n|x_0 - x^*| \to 0$ as $n \to \infty$.

For **b)** we rewrite $f(x) = \mathrm{e}^{2x} - 2 + 3x = 0$ as $x = (2 - \mathrm{e}^{2x})/3$. One possible iteration formula is thus $x_{n+1} = g(x_n) = (2 - \mathrm{e}^{2x_n})/3$. Since

$$g'(x) = -2\mathrm{e}^{2x}/3 \text{ and } g'(0.19) = 0.9749 < 1,$$

the method is convergent (but really slow).

In **c)** the error estimate is given by

$$|x - \bar{x}| \leq \frac{|f(\bar{x})|}{|f'(\bar{x})|} \leq \frac{2.13 \cdot 10^{-4}}{5.89} < 3.7 \cdot 10^{-5}.$$

(4p) **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)** we note that a matrix-vector operation reguires n^2 multiply/additions. Thus $w = Bx + y$ is computed using $2n^2 + n$ operations. The same us true for $z = w + Aw$. The total number of operations is this $4n^2 + 2n$. If the formula were implemented correctly the run time should be given by to $T(n) = cn^2$, or $T(2n)/T(n) = 2^2 = 4$. In the table we have, for instance, $T(4000)/T(2000) = 66300/8360 \approx 7.9$, which is closer to $2^3 = 8$. So likely the formula wasn't implemented correctly but a matrix-matrix multiply was used somewhere.

For **c)** we note that the second row gives the largest sum and $\|A\|_\infty = |-1.1| + |-1.6| + |0.8| = 3.5$.

Finally, **d)** is solved by noting that $r = b - A\widehat{x} = A(A^{-1}b - \widehat{x}) = A(x - \widehat{x})$. Thus

$$\|x - \widehat{x}\| = \|A^{-1}r\| \leq \|A^{-1}\|\|r\|.$$

(4p) **6:** For **a)** the sketch has to make clear that the residual $r = b - Ax$ is orthogonal to the subspace range(A). Since the columns of $A = (a_1, \ldots, a_m)$ for a basis for range(A) then $a_i^T r = 0$, for $i = 1, 2, \ldots, m$, which gives the normal equations $A^T r = A^T(b - Ax) = 0$ or $A^T Ax = A^T b$.

For **b)** we note that each data point (x_i, y_i) gives one row of the over determined system $Ax = b$. The model is $y = c_1 + c_2 x + c_3 \cos(\pi x) + c_4 \sin(\pi x)$. Thus the system $Ax = b$ is

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

For **c)**, we let

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_1 R$$

where $Q = (Q_1, Q_2)$. Since Q is orthogonal we find that

$$\|Ax - b\|_2^2 = \|Q^T(Ax - b)\|_2^2 = \| \begin{pmatrix} R \\ 0 \end{pmatrix} x - \begin{pmatrix} Q_1^T b \\ Q_2^T b \end{pmatrix} \|_2^2 = \|Rx - Q_1^T b\|_2^2 + \|Q_2^T b\|_2^2.$$

The minimum is achived for $x = R^{-1}Q_1^T b$. Thus only the reduced QR decomposition is needed.

Finnally, for **d)**, we observe that $Ax = Q_1 Q_1^T b$ since Q_1 contains the orthogonal basis for range(A). This means that $r = b - Ax = b - Q_1 Q_1^T b = (I - Q_1 Q_1^T)b$, or $P = I - Q_1 Q_1^T$.

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

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

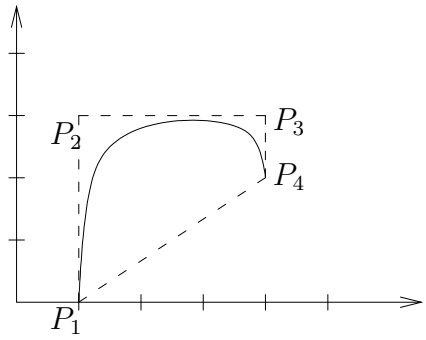
Insert numbers from the table we obtain

$$3^p = \frac{2.8782 - 1.8975}{1.8975 - 1.7885} = 8.9972.$$

Which fits almost perfectly 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 $C = 1.3625$. Finally $R_T = 10^{-3}$ if $h = \sqrt{10^{-3}/1.3625} = 0.0271$. Thus $h < 0.027$ is required.

(3p) **8:** For **a)** $s(x)$ is not a cubic spline since the derivative $s'(x)$ is not continuous at $x = 1$. More precisely $s_1'(1) = -2x|_{x=1} = -2$ and $s_2'(1) = 3x^2 - 4x + 2|_{x=1} = 1$.

For **b)** the sketch is



The convex hull is the area enclosed by the dashed lines. Important features of the Beziér curve is that since both P_1/P_2 and P_3/P_4 have the same x-coordinate the tangent direction of the curve is vertical at both the starting and ending points.

In **c)** the identity $1 = (1-t+t)^2 = (1-t)^2 + 2(1-t)t + t^2$ gives us the weights for the control points. The quadratic Beziér curve is thus

$$p(t) = P_1(1-t)^2 + P_2 2(1-t)t + P_3 t^2, \quad 0 \leq t \leq 1,$$

where the control points P_1, P_2, P_3 are vectors in the plane $\mathbb{R}^2$. The sketch should clearly show that if you have three quadratic Beziér segments then you need a total of $n = 7$ control points.