TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 18th of Mars, 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 $a = 0.0390267$ 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) Let $x = 13.245$ and $y = 7.8802$ be two numbers that belong to the floating point system $(10, 4, -9, 9)$. What would be the result of $z = x + y$ if the computations where carried out on a computer using the floating point system $(10, 4, -9, 9)$?

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 $y = \sqrt{2x}$, where $x = 0.35 \pm 0.02$. Compute the approximate value $\bar{y}$ and give an error bound.

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

x	1.3	1.4	1.5	1.6	1.7
$f(x)$	0.917	1.031	1.183	1.129	1.056

of correctly rounded function values, be given. Use linear interpolation to find an approximation of the function value $f(1.57)$. Also estimate the error in the obtained result.

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

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

for small x values on a computer with unit round off $\mu = 1.11 \cdot 10^{-16}$. We find that the results are quite poor and that the *relative error* in the result tends to grow as $x \to 0$. Explain the poor accuracy by performing an analysis of the computational errors and give a bound for the relative error in the computed result $f(x)$. For the analysis you may assume that all computations are performed with a relative error at most μ.

(3p) **4:** A cubic Beziér curve is given by an expression

$$p(t) = c_1(t)P_1 + c_2(t)P_2 + c_3(t)P_3 + c_4(t)p_4, \quad 0 < t < 1,$$

where P_1, P_2 P_3 and P_4 are control points, and $c_i(t)$, $i = 1, 2, 3, 4$ are the weights.

a) Use the identity $1 = (1-t+t)^3$ to dervie the weights $c_i(t)$ for the cubic Beziér curve.

b) Suppose we want to combine two cubic Beziér curves to one single curve. The combined curve should start in the point $(0, 3)$, pass through the point $(1, 1)$ and end in the point $(2, 0)$. The curves tanget should be horizontal at start and end points and also vertical at the interpolation point $(1, 1)$. Answer the following queastions: Are these requirements enough to make the curve unique? Provide a set of control points and clearly argue that all the requirements are satisfied. Finally draw a sketch illustrating the two curve segments.

(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) Let L be given as above and compute $\|L\|_\infty$.

c) Suppose we want to solve a linear system $Ax = b$ but have errors in the vector b. Show the error estimate

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|},$$

where $\|\cdot\|$ is any induced norm and $\kappa(A)$ is the condition number.

(4p) **6:** Consider the cubic polynomial $f(x) = x^3 - 9x^2 + 24x - 20$. We want to use Newton-Raphsons method for finding a root. Do the following

a) Formulate the Newton-Raphson method and derive the resulting iteration formula when the method is applied to the above cubic polynomial.

b) When Newton-Raphson's method is applied to the function $f(x)$ above with the starting guess $x_0 = 1.8$ we obtain the following table

k	x_k	$\lvert x_k - x^* \rvert$
0	1.8000	0.2000
1	1.8970	0.1030
2	1.9476	0.0524
3	1.9736	0.0264
4	1.9867	0.0133

State the definition of order of convergence p for an iterative method. Also use the table to determine the order of convergence when Newton-Raphson's method is applied to this specific function $f(x)$.

c) Use the results from **b)** and known properties of Newton-Raphson's method to determine if $x^* = 2$ is a double or single root. Explain briefly why you reach the conclusion.

(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)	2.923	3.172	3.201

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 $m \times n$ matrix A, with $m > n$, has rank n, and that the linear system $Ax = b$ has a solution. Use the singular value decomposition $A = U\Sigma V^T$ to give the a formula for the solution x of the system $Ax = b$. Is the solution unique? Clealy motivate your answers.

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 $\bar{a} = 0.03903$ 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 $\Delta a|/|a| \leq 0.5 \cdot 10^{-5}/0.03903 \leq 0.128 \cdot 10^{-3}$.

In **b)** we first compute the exact result $z = 13.245 + 7.8802 = 21.1252$ and round the result to fit the number system $(10, 4, -9, 9)$. to obtain $\bar{z} = 21.125 = 2.1125 \cdot 10^1$.

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

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

where the cancellation is removed.

For **d)** The approximate value is $\bar{y} = \sqrt{2\bar{x}} = \sqrt{2 \cdot 0.35} = 0.84$ with $|R_B| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial x}||\Delta x| = |(2 \cdot 0.35)^{-1/2}|0.02| < 0.024.$$

The total error is $|R_{TOT}| \leq 0.024 + 0.5 \cdot 10^{-2} < 0.03$. Thus $y = 0.84 \pm 0.03$.

(3p) **2:** First we note that the closest points to $x = 1.57$ in the table are $x_1 = 1.5$, $x_2 = 1.6$, and $x_3 = 1.7$.

Making the anzatz $p_1(x) = c_0 + c_1(x - 1.5) + c_2(x - 1.5)(x - 1.6)$, where the last term will be used to estimate the truncation error R_T, we obtain

$$p_1(1.5) = c_0 = 1.183, \quad p_1(1.6) = c_0 + c_1(1.6 - 1.5) = 1.129 \Longrightarrow c_1 = -0.54.$$

For the truncation error we also compute

$$p_1(1.7) = c_0 + c_1(1.7 - 1.5) + c_2(1.7 - 1.5)(1.7 - 1.6) = 1.056 \Longrightarrow c_2 = -0.95.$$

So $p_1(x) = 1.183 - 0.54 \cdot (x - 1.5)$ and $R_T(x) = -0.95(x - 1.5)(x - 1.6)$. Thus we obtain $p_1(1.57) = 1.145$, $R_B \leq 0.5 \cdot 10^{-3}$, $|R_T(1.57)| \leq 2 \cdot 10^{-3}$. We also have to remember the errors in the table giving $R_{XF} \leq 0.5 \cdot 10^{-3}$. Thus $f(1.57) = 1.145 \pm (0.5 \cdot 10^{-3} + 2 \cdot 10^{-3} + 0.5 \cdot 10^{-3}) = 1.145 \pm 3 \cdot 10^{-3}$.

(3p) **3:** We first determine the computational order as

$$f(x) = \sqrt{1 + x} - \sqrt{1 - x} = \sqrt{a} - \sqrt{b} = c - d = e.$$

The relative errors in the intermediate results, e.g. $|\Delta a|/|a|$, are boudned by μ. The error propagation formula gives

$$|\Delta f| \lesssim |\frac{1}{2\sqrt{a}}||\Delta a| + |\frac{1}{2\sqrt{b}}||\Delta b| + |\Delta c| + |\Delta d| + |\Delta e|.$$

In order to simplify the result we use $a \approx b \approx c \approx d \approx 1$ for small x. Also

$$f(x) = \frac{(\sqrt{1+x} - \sqrt{1-x})(\sqrt{1+x} + \sqrt{1-x})}{\sqrt{1+x} - \sqrt{1-x}} = \frac{2x}{\sqrt{1+x} - \sqrt{1-x}} \approx x,$$

for small x. We obtain

$$|\Delta f| \lesssim \mu(\frac{1}{2} + \frac{1}{2} + 1 + 1 + |x|) \approx 3\mu.$$

Since $f(x) \approx x$ for small x the bound for the relative error is $|\Delta f|/|f| \leq 3|x|^{-1}\mu$.

(3p) **4:** For **a)** we compute $(1-t+t)^3 = (1-t)^3 + 3(1-t)^2 t + 3(1-t)t^2 + t^3 = c_1(t) + c_2(t) + c_3(t) + c_4(t)$.

For **b)** we recall that the tangent on a Beziér cubic curve satisfies $s'(0) = 3(P_2 - P_1)$ and $s'(1) = 3(P_4 - P_3)$. Thus the curve is horizontal at the start point if P_1 and P_2 has the same y-coordinate. This does not put any requirements on the x-coordinate of P_2. Thus we can pick points $P_1 = (0,3)$, $P_2 = (0.5,3)$, $P_3 = (1,1.5)$, and $P_4 = (1,1)$ for the first curve segment. The curve will have a vertical tangent at the end point since P_3 and P_4 has the same x-coordinate. The y-coordinate can be chosen freely. Similarily the second curve segment is determined by the control points $p_4 = (1,1)$, $P_5 = (1,0.5)$, $P_6 = (1.5,0)$ and $P_7 = (2,0)$. As explained above the curve is not unique.

(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)** we note that the largest row-sum is given by the last row so $\|L\|_\infty = 0.3 + 1.8 + 1 = 3.1$.

In **c)** we note the solution with the noisy b vector can be written $A(x + \Delta x) = b + \Delta b$ and thus $A(\Delta x) = \Delta b$, or $\|\Delta x\| = \|A^{-1}\Delta b\| \leq \|A^{-1}\|\|\Delta b\|$. We also have $\|b\| = \|Ax\| \leq \|A\|\|x\|$. Thus

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\Delta b\|}{\|b\|/\|A\|} = \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|},$$

where $\kappa(A) = \|A\|\|A^{-1}\|$ is the condition number.

(4p) **6:** For **a)** we give the definition of the Newton-Raphson method as $x_{k+1} = x_k - f(x_k)/f'(x_k)$, where for our case $f'(x) = 3x^2 - 18x + 24$. There is no reason at all to simplify the resulting iteration formula.

For **b)** we state the order of convergence as the p value such that $|x_k - x^*| \approx C|x_{k-1} - x^*|^p$, or

$$\frac{|x_k - x^*|}{|x_{k-1} - x^*|^p} = C.$$

The goal is to select the value p so that the quotient is approximately a constant independently of the iteration number k. If we use $p = 1$ we see that $k = 1$ gives

$C = 0.2/1.03 = 1.9417$, $k = 2$ gives $C = 0.1030/0.0524 = 1.9656$, etc. This means that $p = 1$ and we have linear convergence.

In **c)** we recall that Newtons method is supposed to have quadratic convergence if the root x^* is a single root. Also the convergence is linear for double roots. Thus the function $f(x)$ the root $x^* = 2$ is a double root.

(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

$$3^p = \frac{2.9122 - 3.1689}{3.1689 - 3.1974} = 9.0070$$

Which fits perfectly with $p = 2$. In order to determine C we use the last equation $T(2h) - T(h) = (3^2 - 1^2)Ch^2$ and insert $h = 0.1$ to obtain $C = -0.356$. Finally $R_T = 10^{-3}$ if $h = \sqrt{10^{-3}/0.356} = 0.053$. Thus $h < 0.053$ is required.

(3p) **8:** For **a)** we note that if $\text{rank}(A) = n$ then the matrix only has the trivial null space. Thus any solution we find is unique. Thus the solution can be written as

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

In order to determine the coefficients c_i we compute

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

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

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

is the sought after unique solution.

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 $x = V^T e_n = v_n$.