TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 18th of January, 2025.

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 $a = 712.6623$ be an exact value. Round the value a to 6 *significant digits* to obtain an approximate value $\bar{a}$. Also give a bound for the *absolute error* in $\bar{a}$.

b) Let $x = 37.119875$. Write x in *normalized* form and 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 $y = \sqrt{1 + a/2}$, where $a = 1.27 \pm 0.02$. Compute the approximate value $\bar{y}$ and give an error bound.

(2p) **2:** We have the following table

x	1.0	1.3	1.5
$f(x)$	1.284	1.413	1.475

with correctly rounded function values. Use linear interpolation to approximate the function value $f(1.18)$. 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.

(4p) **4:** The non-linear equation $f(x) = 1 - x^2 + \cos(x/2) = 0$ has a root $x^* \approx 1.34$. Do the following:

a) The equation Formulate the Newton-Raphson method for finding approximatate solutuins $\bar{x} \approx x^*$. Also perform $k = 3$ iterations with the method using the initial guess $x_0 = 1.34$.

b) Estimate the error in the approximate root $\bar{x} = 1.336056 \approx x^*$.

c) An iterative method $x_{k+1} = \varphi(x_k)$ has at least *quadratic convergence* if the iteration function satisfies $\varphi'(x^*) = 0$, where x^* is the fixed point, or the root of the equation $f(x) = 0$. Present a definition that clearly shows what *quadratic convergence* means and also show that the Newton-Raphson method has quadratic convergence if x^* is a *single root.*

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

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

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1.7 & 1 & 0 \\ 0.3 & 0.8 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1.7 & -2.3 & -1.4 \\ 0 & 1.2 & -0.5 \\ 0 & 0 & 3.1 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Determine if pivoting was used correctly during the computations. Motivate your answer!

b) Let

$$A = \begin{pmatrix} 2.2 & -0.9 & -1.3 \\ 1.1 & -0.7 & 1.2 \\ -1.1 & 0.5 & 0.8 \end{pmatrix},$$

and find a Guass transformation matrix M_1 so that $M_1 A$ has zeroes below the diagonal in the first column.

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 $\|A\|_\infty$.

(4p) **6:** Let $A \in \mathbb{R}^{m\times n}$ be a matrix and $A = U\Sigma V^T$ be the *singular value decomposition.* Do the following:

a) Suppose $m = n$ and $\text{rank}(A) = n$. Show that the formula

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

provides a solution to $Ax = b$. Is the solution unique?

b) Suppose $\text{rank}(A) = n$. Clearly demonstrate how the matrices U and V provides basis vectors for the spaces $\text{Range}(A)$ and $\text{null}(A)$. What are the dimension of the range and null space respectively.

c) Show that $\|A\|_2 = \sigma_1$ and if A^{-1} exists then $\|A^{-1}\|_2 = 1/\sigma_n$.

(2p) **7:** The Trapezoidal method computes an approximation

$$T(h) \approx I = \int_a^b f(x)dx,$$

where the accuracy depends on the step size h used. The truncation error of the method can be described as $R_T \approx Ch^p$. We compute a few approximations $T(h)$ of the exact integral I and obtain

h	0.4	0.2	0.1
T(h)	1.5826	1.5672	1.5635

Use the table to determine C and p. Also estimate the step size h needed for the error to be of magnitude 10^{-4}. *Present your calculations.*

(3p) **8:** **a)** Let $s(x)$ be defined by two cubic polynomials,

$$s(x) = \begin{cases} s_1(x) = 0.9 + 0.1x + 0.6x^2 + 0.4x^3, & 0 \le x < 1, \\ s_2(x) = 2.0 + c_1(x-1) + c_2(x-1)^2 + 0.4(x-1)^3, & 1 \le x \le 2. \end{cases}$$

Find the appropriate values for the constants c_1 and c_2 so that $s(x)$ is a cubic spline.

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 $1 = 1^3 = (1 - t + t)^3$ to derive the expression for a cubic Beziér curve. Also draw a clear sketch that shows an example of a continuously differentiable curve consisting of two different cubic 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 $\bar{a} = 712.662$ which has 6 significant digits. The absolute error is at most $|\Delta a| \leq 0.5 \cdot 10^{-3}$.

In **b)** the $x = 3.7119875 \cdot 10^1$ in normalized form and 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 $\bar{y} = \sqrt{1+\bar{a}/2} = \sqrt{1.635} = 1.28$ with $|R_B| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial a}||\Delta a| = |\frac{1}{2\sqrt{1+a/2}}\frac{1}{2}||\Delta a| < 0.004.$$

The total error is $|R_{TOT}| \leq 0.004 + 0.5 \cdot 10^{-2} < 0.009 < 0.01$. Thus $y = 1.27 \pm 0.01$.

(2p) **2:** We use Newtons interpolation formula and the ansatz $p(x) = p_1(x) + R_T(x) = c_0 + c_1(x-1.0) + c_2(x-1.0)(x-1.3)$, where the last term will be used to estimate the truncation error. Inserting the function values from the table leads to $p(1.0) = c_0 = 1.284$ and $p(1.3) = c_0 + c_1(0.3) = 1.413$ 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

$$p_1(x) = 1.284 + 0.43(x-1.0) \text{ and } R_T(x) = -0.24(x-1.0)(x-1.3).$$

We obtain $f(1.18) \approx p_1(1.18) = 1.361$ with $|R_B| < 0.5 \cdot 10^{-3}$ and $R_T \leq |-0.24(1.18-1.0)(1.18-1.3)| < 0.52 \cdot 10^{-2}$. The errors in the function values used also gives an error $R_{XF} < 0.5 \cdot 10^{-3}$ in the result. Thus $f(1.18) = 1.361 \pm 0.62 \cdot 10^{-2} = 1.361 \pm 0.7 \cdot 10^{-2}$.

(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 $f(x) \approx 1$).

(4p) **4:** For **a)** we write the Newton-Raphson method as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

where $f(x)$ is given in the exercise and $f'(x) = -2x - \sin(x/2)/2$. There is no need to simplify anything. Just inserting into the formula gives

$$x_1 = 1.3360614, \quad x_2 = 1.3360557, \text{ and } x_3 = 1.3360557$$

For **b)** the error estimate is given by

$$|x - \bar{x}| \leq \frac{|f(\bar{x})|}{|f'(\bar{x})|} \leq \frac{9.1 \cdot 10^{-7}}{2.98} < 3.1 \cdot 10^{-7}.$$

For **c)** we state that the sequence $\{x_n\}$ has quadratic convergence (to x^*) if

$$\lim_{n\to\infty} \frac{|x_{n-1} - x^*|}{|x_{n-1} - x^*|^2} = C.$$

where C is a constant. The same limit also cannot exist for $p = 3$ or we would have cubic convergence. For the Newton-Raphson method $\varphi(x) = x - f(x)/f'(x)$. Thus,

$$\varphi'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Since for a single root x^* we have $f(x^*) = 0$ and $f'(x^*) \neq 0$ we see that $\varphi'(x^*) = 0$. This means we have quadratic convergence.

(3p) **5:** For **a)** we just observe that one of the multipliers (i.e. $\ell_{21} = -1.7$) is larger than one in magnitude. Thus pivoting wasn't used correctly.

For **b)** the Gauss transformation matrix should have the structure

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -m_1 & 1 & 0 \\ -m_2 & 0 & 1 \end{pmatrix},$$

where $m_1 = 2.1/0.3 = 7$ and $m_2 = -1.1/0.3 = -3.667$.

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

(4p) **6:** For **a)** we simply compute

$$Ax = A(\sum i = 1n\frac{u_i^T b}{\sigma_i} v_i) = \sum i = 1n\frac{u_i^T b}{\sigma_i} A v_i = \sum i = 1n\frac{u_i^T b}{\sigma_i}\sigma_i u_i = \sum i = 1n(u_i^T b) u_i = b.$$

The last equality holds since $m = n$ so $b \in \mathbb{R}^n$ and $\{u_i\}_{i=1}^n$ is an orthonormal basis for $\mathbb{R}^n$. The solution is uniqie since A has full rank so A^{-1} exists.

For **b)** we write the decomposition $A = U\Sigma V^T$ as

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

where $\sigma_n > 0$ as $\mathrm{rank}(A) = n$. This means that $Av_i = \sigma_i u_i \neq 0$ for $i = 1, \ldots, n$. So the null space is only the trivial one $\mathrm{null}(A) = \{0\}$ with dimension 0. Similarily, if y belongs to the range then there is an x such that $y = Ax$, or

$$y = Ax = \sum_{i=1}^{n} \sigma_i (v_i^T x) u_i,$$

so the y is a linear combination of $\{u_1, \ldots, u_n\}$. Thus $\mathrm{range}(A) = \mathrm{span}(u_1, \ldots, u_n)$ and the dimension of the range is n.

For **c)** we use $A = U\Sigma V^T$ where U, V are orthogonal and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$. Since U, V are orthogonal we obtain $\|A\|_2 = \|U\Sigma V^T\|_2 = \|\Sigma\|_2$. The norm of a diagonal matrix can be computed by

$$\|\Sigma\|_2 = \max_{y \in \mathbb{R}^n} \frac{\|\Sigma y\|_2}{\|y\|_2} = \max_{y \in \mathbb{R}^n} \sqrt{\frac{\sum \sigma_i^2 y_i^2}{\sum y_i^2}} \leq \sigma_1 \max_{y \in \mathbb{R}^n} \sqrt{\frac{\sum y_i^2}{\sum y_i^2}} = \sigma_1,$$

with equality for $y = e_1$. Thus $\|A\|_2 = \sigma_1$. If A^{-1} exists then $A^{-1} = V\Sigma^{-1}U^T$ and $\|A^{-1}\|_2 = \|\Sigma^{-1}\|_2$. Since the diagonal elements of Σ^{-1} are $1/\sigma_i$ the largest diagonal element is $1/\sigma_n$ and $\|A^{-1}\|_2 = 1/\sigma_n$.

(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

$$2^p = \frac{1.5826 - 1.5672}{1.5672 - 1.5635} = 4.1622.$$

Which fits almost perfectly 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 $C = 0.3700$. Finally $R_T = 10^{-4}$ if $h = \sqrt{10^{-4}/0.3700} = 0.0164$. Thus $h < 0.016$ is required.

(3p) **8:** For **a)** can use $s_1'(1) = s_2'(1)$, or

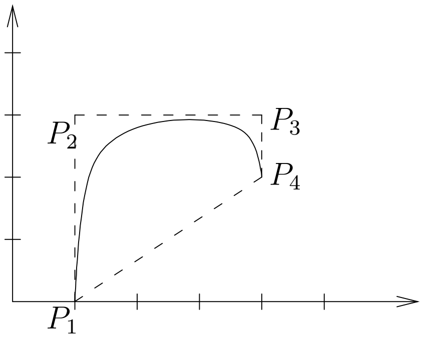
$$0.1 + 2 \cdot 0.6 \cdot 1 + 3 \cdot 0.4 \cdot 1^2 = c_1,$$

to find $c_1 = 2.5$. Similarily $s_1''(1) = s_2''(1)$, or

$$2 \cdot 0.6 + 3 \cdot 2 \cdot 0.4 \cdot 1 = 2c_2,$$

to obtain $c_2 = 1.8$.

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 $1 = (1-t+t)^3 = (1-t)^3 + 3(1-t)^2t + 3(1-t)t^2 + t^3$ gives us the weights for the control points. The cubic Beziér curve is thus

$$p(t) = P_1(1-t)^3 + P_2 3(1-t)^2 t + P_3 3(1-t)t^2 + P_4 t^3, \quad 0 \le t \le 1,$$

where the control points P_1, P_2, P_3, P_4 are vectors in the plane $\mathbb{R}^2$. The sketch should clearly show that if you have two cubic Beziér segments then you need a total of $n = 7$ control points.