TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 14th of January, 2023.

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 $a = 0.08199237$ be an exact value. Round the value a to 3 *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 $x = 294.37723$ on a computer using the floating point system $(10, 4, -10, 10)$. 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 $y = \alpha(1 + x)^2$, where $x = 0.34 \pm 0.02$, and $\alpha = 2.13 \pm 0.07$. Compute the approximate value $\bar{y}$ and give an error bound.

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

x	1.3	1.5	1.6
$f(x)$	0.6772	0.7251	0.74976

of correctly rounded function values, be given. Use linear interpolation to estimate the function value $f(1.39)$. Also estimate the error in the computed approximation.

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

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

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.

(3p) **4:** Consider the function $f(x) = \cos(x) - xe^x$. 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 function $f(x)$.

b) When Newton-Raphson's method is applied to the function $f(x)$ above with the starting guess $x_0 = 1$ we obtain the following table

k	x_k	$f(x_k)$
0	1.0000000	$-2.2 \cdot 10^0$
1	0.6530794	$-4.6 \cdot 10^{-1}$
2	0.5313434	$-4.2 \cdot 10^{-2}$
3	0.5179099	$-4.6 \cdot 10^{-4}$
4	0.5177574	$-5.9 \cdot 10^{-8}$

We decide to use $\bar{x} = 0.5178$ as an approximation of x^*. Estimate the error in the approximation $\bar{x}$.

c) State the definition of the *order of convergence* for an iterative method for finding the root x^* of an equation $f(x) = 0$. Also use the tabl above to estimate the order of convergence for the Newton-Raphson method.

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

a) Suppose $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times n}$, $m > n$. How many arithmetic operations are required to evaluate the formula $z = (A + I)Bx + y$, where x and y are vectors.

b) Suppose we have a linear system $Ax = b$ where

$$A = \begin{pmatrix} -1 & 0 & 3 \\ 2 & 1 & -2 \\ 1 & 2 & -1 \end{pmatrix} \text{ and } b = \begin{pmatrix} 6 \\ 1 \\ -3 \end{pmatrix}.$$

Find a *Permutation matrix* P_1 and a *Gauss-transformation* L_1 such that $U_1 = L_1P_1A_1$ has zeros below the diagonal in the first column. Pick L_1 and P_1 so that U_1 is the intermediate result you would obtain after the first step in computing the LU decomposition of the matrix A when partial pivoting is used.

(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)	1.57213	1.706951	1.72197

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 $n \times n$ matrix.

a) Suppose that A has full rank. Use the singular valur decomposition $A = U\Sigma V^T$ to give a general formula for the solutions x of the system $Ax = b$. Clealy motivate your answer.

b) 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.

Answers

(5p) **1:** For **a)** we obtain the approximate value $\bar{a} = 0.082$ which has 3 correct decimal digits. The absolute error is at most $|\Delta a| \leq 0.5 \cdot 10^{-3}$ and thus the *relative error* is bounded by $\Delta a|/|a| \leq 0.5 \cdot 10^{-3}/0.082 \leq 0.61 \cdot 10^{-2}$.

In **b)** we rewrite the number as $x = 2.9437723 \cdot 10^2$ to see that $\bar{x} = 2.9438 \cdot 10^2$ 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 $\bar{y} = \bar{\alpha}(1+\bar{x})^2 = 2.13(1+0.34)^2 = 3.8$ with $|R_B| \leq 0.5 \cdot 10^{-1}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial \alpha}||\Delta \alpha| + |\frac{\partial y}{\partial x}||\Delta x| = |(1+x)^2|\Delta \alpha| + |\alpha 2(1+x)||\Delta x| < 0.24$$

The total error is $|R_{TOT}| \leq 0.24 + 0.5 \cdot 10^{-2} < 0.3$. Thus $y = 3.8 \pm 0.3$.

(3p) **2:** We Newtons interpolation formula and let

$$p_1(x) = c_0 + c_1(x-1.3) + c_2(x-1.3)(x-1.5),$$

where the quadratic term is used to obtain the truncation error. The interpolation conditions gives

$$p_1(1.3) = c_0 = 0.6772, p_1(1.5) = c_0 + c_1(1.5-1.3) = 0.7251, \quad \text{and}$$

$$p_1(1.6) = c_0 + c_1(1.6-1.3) + c_2(1.6-1.3)(1.6-1.5) = 0.74976.$$

Solve gives $c = (c_0, c_1, c_2)^T = (0.6772\,,\,0.2395\,,\,0.0237)^T$. This gives us $\bar{f}(1.39) = c_0 + c_1(1.39-1.3) = 0.6988$, with $|R_B| \leq 0.5 \cdot 10^{-4}$. The truncation error is estimated $|R_T| \leq |0.0237(1.39-1.3)(1.39-1.5)| < 2.4 \cdot 10^{-4}$. We also have $|R_{XF}| \leq 0.5 \cdot 10^{-4}$ since the function values in the table are correctly rounded to four decimal digits. Thus the total error is $|R_{TOT}| \leq 4 \cdot 10^{-4}$ and we can answer $f(1.39) = 0.6988 \pm 4 \cdot 10^{-4}$.

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

$$f(x) = f(x) = \cos(2x) - (1-x)^2 = \cos(a) - b^2 = c - d = e.$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial a}||\Delta a| + |\frac{\partial f}{\partial b}||\Delta b| + |\frac{\partial f}{\partial c}||\Delta c| + |\frac{\partial f}{\partial d}||\Delta d| + |\frac{\partial f}{\partial e}||\Delta e| = |\sin(a)||\Delta a| + |2b||\Delta b| +$$

$$|1||\Delta c| + |1||\Delta d| + |1||\Delta e| \lesssim \mu(|a\sin(a)| + |2b^2| + |c| + |d| + |e|) \approx \mu(|0| + |2| + |1| + |1| + |0|) \approx 4\mu,$$

where we have used that $a = 2x$ is small and thus $a\sin(a) \approx 0$. So the absolute error in the computation is bounded by 4μ but since $f(x) \to 0$ as $x \to 0$ the *relative error* can be very large. Thus we have *cancellation* in the computation.

(3p) **4:** For **a)** we recall that given a starting approximation x_0 the Newton-Raphson method computes

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

and we only need to calculate the derivative $f'(x) = -\sin(x) - (x+1)\mathrm{e}^x$.

For **b)** we use the error estimate

$$|\bar{x} - x^*| \lesssim \frac{|f(\bar{x})|}{|f'(\bar{x})|} \approx \frac{1.2971 \cdot 10^{-4}}{3.0433} < 4.3 \cdot 10^{-5}.$$

For **c)** we define the order of convergence as the largest integer p such that

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

Since x^* is unknown we cannot directly apply the definition. The simplest solution is to assume that the iteraton x_4 has a much smaller error than the other iterations x_1, x_2, x_3. Thus we approximate $x^* = 0.5177574$ and compute the errors $|x_0 - x^*| \approx 4.8 \cdot 10^{-1}$, $|x_1 - x^*| \approx 1.4 \cdot 10^{-1}$, $|x_2 - x^*| \approx 1.4 \cdot 10^{-2}$, and $|x_3 - x^*| \approx 1.5 \cdot 10^{-4}$. Since $(|x_1 - x^*|)^2 \approx (1.4 \cdot 10^{-1})^2 \approx 2 \cdot 10^{-2} \approx |x_2 - x^*|$ and $(|x_2 - x^*|)^2 \approx (1.4 \cdot 10^{-2})^2 \approx |x_3 - x^*|$ we conclude that the table shows that $p = 2$ for Newton-Raphsons method.

(3p) **5:** For **a)** we evaluate the expression using the following operations

$$z = (A+I)Bx + y = (A+I)x_1 + y = Ax_1 + x_1 + y = x_2 + x_1 + y = x_3 + y = x_4$$

Computing the matrix vector product $x_1 = Bx$ requires mn multiplications and additions each, i.e. a total of $2mn$ operations. The product $x_2 = Ax_1$ requires $2m^2$ operations. The remaining two vector additons require m additions (as $y, x_1 \in \mathbb{R}^m$). So the operation count is $m(2m + 2n + 2)$.

For **b)** we note that the largest element in the first column is on the second row and thus

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

With this choice we have

$$A_1 = P_1 A = \begin{pmatrix} 2 & 1 & -2 \\ -1 & 0 & 3 \\ 1 & 2 & -1 \end{pmatrix}.$$

For the elimination step the multipliers are $m_{21} = -1/2 = -0.5$ and $m_{31} = 1/2 = 0.5$. Therefore the Gauss transformation is

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.5 & 0 & 1 \end{pmatrix}.$$

where we recall that the elements under the elimination matrix are $-m_{ij}$.

(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

$$3^p = \frac{1.57213 - 1.706951}{1.706951 - 1.72197} = 8.9767$$

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 $C = -0.18774$. Finally $R_T = 10^{-4}$ if $h = \sqrt{10^{-4}/0.18774} = 0.02307$. Thus $h < 0.0023$ is required.

(3p) **8:** For **a)** we write the solution x using the vasis 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

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

Thus

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

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 $x = V^T e_n = v_n$. The solution cannot be unique since x and $-x$ will give the same minimum. However if $\sigma_{n-1} > \sigma_n$ at least thats the only source of non-uniqueness.