TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

# Exam TANA09 Datatekniska beräkningar

**Date:** 14-18, 18th of March, 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 $a = 22.73531443$. Round the value $a$ correctly to 5 *significant digits* to obtain the approximation $\bar{a}$. Give both the approximate value $\bar{a}$ and an upper bound for the absolute error $|\Delta a|$ in the approximation.

**b)** Let $x = -102.232$. Give a bound for the *absolute error* when $x$ is stored on a computer using the floating point system $(10, 3, -10, 10)$.

**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 $z = x^2 y$, where $x = 2.35 \pm 0.01$ and $y = 1.17 \pm 0.02$. Compute the approximate value $\bar{z}$ and an error bound.

*(2p)* **2:** Do the following:

**a)** Use Lagrange interpolation to find the polynomial of degree 2 that interpolates the table

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $f(x)$ | 1.3 | 0.6 | 1.9 |

**b)** Suppose the value $f(2) = 0.6$ has an error and we actually have $f(2) = 0.6 \pm 0.03$. Find the maximum error in the interpolating polynomial, for the interval $1 < x < 3$, due to the error in the function value $f(2) = 0.6$.

*(3p)* **3:** Let $x$, $y$, and $z$ be column vectors of length $n$. We want to implement the formula

$$w = (I + xy^T)(I - yx^T)z$$

where $I$ is the identity matrix as efficiently as possible. Do the following

**a)** How many floating point operations are required to implement the formula? Also how many memory slots are reguired for storing intermediate results?

**b)** 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) | 537 | 4369 | 35721 | 283913 |

Was the implementation done using the most efficient method? Motivate your answer carefully.

*(4p)* **4:** Consider the function $f(x) = 2e^{-x/2} - x^2 - \sqrt{x}$. 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 $x_0 = 1.0$ we obtain the following table

| $k$ | $x_k$ | $f(x_k)$ |
|---|---|---|
| 0 | 1.0000 | $-7.9 \cdot 10^{-1}$ |
| 1 | 0.7466825 | $-4.5 \cdot 10^{-2}$ |
| 2 | 0.7304596 | $-1.7 \cdot 10^{-4}$ |
| 3 | 0.7303989 | $-2.3 \cdot 10^{-9}$ |

We decide to use $\bar{x} = 0.7304$ as an approximation of $x^*$. Estimate the error in the approximation $\bar{x}$.

**c)** Prove that the Newton iteration has quadratic convergence to a single root $x^*$ provided that the starting guess is sufficiently good.

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

**a)** 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 $\|AB\| \leq \|A\|\|B\|$.

**b)** Prove that $\|I\| = 1$ and $\|A\|\|A^{-1}\| \geq 1$ for all matrix norms induced by a vector norm.

**c)** Let $\bar{x} = (1.23\,,\, 0.37\,,\, -2.6)^T$ and assume that the elements $\bar{x}_k$ are correctly rounded. Compute both the absolute and relative error measured in $\|\cdot\|_\infty$.

*(3p)* **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 $m$ points $(x_i, y_i)$ that are supposed to be located on a circle. A model for this situation is that the points $(x_i, y_i)$ satisfy an equation

$$c_1(x_i^2 + y_i^2) + c_2 x_i + c_3 y_i + 1 = 0,$$

where the parameters $c_1$, $c_2$ and $c_3$ uniquely determines the circle. Formulate the problem of identifying a circle from points $(x_i, y_i)$, $i = 1, 2, \ldots, m$, as a least squares problem. Clearly show the $A$ matrix and the $b$ vector for this case.

**b)** Let $A = Q_1 R$ be the reduced $QR$ decomposition of the matrix $A$. Clearly demonstrate how the reduced $QR$ decomposition can be used to compute $\|r\|_2$ where $r = b - Ax$ is the residual and $x$ is the least squares solution.

**c)** Consider the vector $a$ as an $n \times 1$ matrix. Write out its reduced $QR$ decomposition explicitly. Also write down a formula for the solution of the least squares problem $ax \approx b$, where $b$ is a given $n \times 1$ vector.

*(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 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|
| T(h) | 1.7631 | 1.7675 | 1.7786 | 1.8052 | 1.8456 |

Use the table to determine $C$ and $p$.

*(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 $\bar{a} = 22.735$ which has 3 correct decimal digits. The absolute error is at most $|\Delta a| \leq 0.5 \cdot 10^{-3}$.

In **b)** the unit round off for the floating point system is $\mu = 0.5 \cdot 10^{-3}$. This is an upper bound for the relative error. Thus the absolute error is bounded by $|\Delta x| \leq \mu |x| \leq 0.5 \cdot 10^{-3} 103 \leq 0.052$

For **c)** Since $y = \sqrt{1+x} \approx 1$, for small $x$, *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

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

where the cancellation is removed.

For **d)** The approximate value is $\bar{z} = x^2 y = (2.35)^2(1.17) = 6.46$ with $|R_B| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives

$$|\Delta z| \lesssim |\frac{\partial z}{\partial x}||\Delta x| + \frac{\partial z}{\partial y}||\Delta y| = |2xy||\Delta x| + |x^2||\Delta y| \leq 0.17.$$

The total error is $|R_{TOT}| \leq 0.17 + 0.5 \cdot 10^{-2} < 0.2$. Thus $z = 6.46 \pm 0.2$. Possibly it would have been better to use $\bar{z} = 6.5$.

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

$$p_2(x) = 1.3\frac{(x-2)(x-3)}{(1-2)(1-3)} + 0.6\frac{(x-1)(x-3)}{(2-1)(2-3)} + 1.9\frac{(x-1)(x-2)}{(3-1)(3-2)}.$$

There is no reason to simplify the expression further.

For **b)** we note that if the function value $f(2) = f_2 = 0.6$ has en error then the Lagrange polynomial changes as

$$\bar{p}_2(x) = p_2(x) + \Delta f_2 \frac{(x-1)(x-3)}{(2-1)(2-3)}.$$

The function $|(x-1)(x-3)|$ has a maximum for $x = 2$ which means that

$$|\bar{p}_2(x) - p_2(x)| \leq |\Delta f_2| \frac{(2-1)(2-3)}{(2-1)(2-3)} \leq 0.03.$$

*(3p)* **3:** For **a)** we observe that $x^T z$ is a scalar product that requires $n$ multiplications and additions. Thus $(I - yx^T)z = z - (x^T z)y$ requires only $4n$ floating point operations. We also need one slot of temporary storage for the scalar product and also one vector to store the intermediate result $w_1 = (x^T y)z$. The same temporary vector can be overwritten when the subtractions $w_2 = z - w_1$ are computed. The second

component $(I+xy^T)w_2$ similarily needs one more temporary vector and also an extra $4n$ floating point operations. Though it could be argued that this is the memory where we will store the final result $w$. Thus the formula required $8n$ floating point operations and either $n$ or $2n$ memory slots.

For **b)** we remark that if the formula were implemented correctly the run time should be given by to $T(n) = cn$, or $T(2n)/T(n) = 2^1 = 1$. In the table we have, for instance, $T(4000)/T(2000) = 35721/4369 \approx 8.17$, which is closer to $2^3 = 8$. So likely the formula wasn't implemented correctly but rather the expression where implemented by first computing both matrices $A_1 = I + xy^T$ and $A_2 = I - yx^T$ and then computing the matrix-matrix multiply $A_1 A_2$.

*(4p)* **4:** For **a)** Newton Raphsons method is $x_{k+1} = x_k - f(x_k)/f'(x_k)$, where the function $f(x)$ and its derivative $f'(x) = -\mathrm{e}^{-x/2} - 2x - \frac{1}{2}x^{-1/2}$ is needed. There is no need to simplify the formula. For **b)** the error estimate is

$$|x - \bar{x}| \le \frac{|f(\bar{x})|}{|f'(\bar{x})|} \le \frac{2.94 \cdot 10^{-6}}{2.73} < 1.1 \cdot 10^{-6}.$$

In **c)** we recall that Newton-Raphsons method is defined by the iteration function

$$\phi(x) = x - \frac{f(x)}{f'(x)}, \text{ and } \phi'(x) = -\frac{f(x)f''(x)}{(f'(x))^2}.$$

Since $x^*$ is a single root, i.e. $f'(x^*) \neq 0$, we see that $\phi'(x^*) = 0$. A Taylor series expansion shows that

$$\phi(x_k) = \phi(x^*) + \phi'(x^*)(x_k - x^*) + \frac{\phi''(\xi)}{2}(x_k - x^*)^2, \xi \in (x_k, x^*).$$

Since $\phi(x_k) = x_{k+1}$, $\phi(x^*) = x^*$ and $\phi'(x^*) = 0$ we obtain

$$x_{k+1} - x^* = \frac{\phi''(\xi)}{2}(x_k - x^*)^2,$$

which shows that the convergence is quadratic.

*(3p)* **5:** For **a)** 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

$$\|AB\| = \max_{x \neq 0} \frac{\|ABx\|}{\|x\|} = \max_{x \neq 0} \frac{\|ABx\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \le \max_{y \neq 0} \frac{\|Ay\|}{\|y\|}\|B\| \le \|A\|\|B\|.$$

For **b)** from the definition of the matrix norm, and since $Ix = x$ we have

$$\|I\| = \max_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x \neq 0} \frac{\|x\|}{\|x\|} = 1, \text{ so } 1 = \|I\| = \|AA^{-1}\| \le \|A\|\|A^{-1}\|.$$

For **c)** if $\bar{x} = (1.23\,,\,0.37\,,\,-2.6)^T$ is correctly rounded then the error vector satisfies $|\delta x| \leq (0.005\,,\,0.005\,,\,0.05)^T$. Thus $\|x - \bar{x}\|_\infty \leq 0.5 \cdot 10^{-1}$ is the absolute error and $\|x - \bar{x}\|_\infty / \|x\|_\infty \leq 0.05/2.6 < 0.02$ is the relative error.

*(3p)* **6:** For **a)** we note that for each point $(x_i, y_i)$ we get one row of the system $Ax = b$. More precisely the system is

$$\begin{pmatrix} x_1^2 + y_1^2 & x_1 & y_1 \\ x_2^2 + y_2^2 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ x_m^2 + y_m^2 & x_m & y_m \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}.$$

For **b)** there are many options. The simplest is to note that since $Q_1$ is a basis for range$(A)$ then $Ax = Q_1 Q_1^T b$. This means that we need to compute $\|b - Ax\|_2 = \|b - Q_1 Q_1^T b\|_2$. The other option is to simnply compute $x = R^{-1}(Q_1^T b)$ and then compute $r = b - Ax$ directly.

For **c)** the vector $a$ can be seen as a matrix in $\mathbb{R}^{n \times 1}$. This means that

$$a = (a/\|a\|_2)\|a\|_2 = Q_1 R$$

where $Q_1 \in \mathbb{R}^{n \times 1}$ and $R \in \mathbb{R}^{1 \times 1}$. The formula for the least squares solution can be written using the normal equations $a^T a x = a^T b$ or $x = (a^T b)/(a^T a)$. This is the same as $x = R^{-1} Q_1^T b$ with the decomposition above.

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

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

From the table we can insert the numbers for $h = 0.4$, $h = 0.2$ and $h = 0.1$ to obtain

$$2^p = \frac{1.8052 - 1.7675}{1.7675 - 1.7631} = 8.5682.$$

Which fits well with $p = 3$. In order to determine $C$ we use the last equation $T(2h) - T(h) = (2^3 - 1^3)Ch^3$ and insert $h = 0.1$ to obtain $C = 0.6286$.

*(3p)* **8:** For **a)** we note that if rank$(A) = k$ then $\{v_{k+1}, \ldots, v_n\}$ is a basis for 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 $\sigma_n$ is the smallest singular value, with equality if $y = e_n$ which means that $x = V^T e_n = v_n$.