TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA09 Datatekniska beräkningar

Date: 14-18, 15th of January, 2020.

Allowed:

1. Pocket calculator

Examiner: Fredrik Berntsson

Marks: 25 points total and 10 points to pass.

Jour: Fredrik Berntsson - (telefon 013 28 28 60)

Vists at around 15.00 and 16.30.

Good luck!

(5p) **1:** **a)** Let $a = 0.05347352661$ be an exact value. Round the value a to 4 *significant digits* to obtain an approximate value $\bar{a}$. Also give a bound for the *absolute error* in $\bar{a}$.

b) Let $x = 12.7152$. Give a bound for the *relative error* when x is stored on a computer using the floating point system $(10, 3, -10, 10)$.

c) Let $f(x) = (\mathrm{e}^x - 1)/x$. For small values of x we make the approximation $f(x) \approx \bar{f}(x) = 1 + \frac{x}{2}$. The *truncation error* when $f(x)$ is approximated by $\bar{f}(x)$ can be written $|R_T| \lesssim Cx^p$. What is the value of the integer p? Clearly present calculations that motivates your answer.

d) Let $y = \sqrt{a}$, where $a = 2.48 \pm 0.04$. Compute the approximate value $\bar{y}$ and give an error bound.

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

x	1	3	5
$f(x)$	1.5	2.2	3.4

be given. Use Lagrange interpolation formula to write the second degree polynomial that interpolates the above table.

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

$$f(x) = \frac{\cos(x) - 1}{\sin(x)}$$

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. 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 $f(x) = 1 + x^2 - 3\sqrt{x} = 0$ has a root $x^* \approx 0.11$. Formulate a fixed point iteration for finding a root to $f(x) = 0$ and show that the proposed method is convergent.

c) The equation $f(x) = 1 + x^2 - 3\sqrt{x}$ is solved using fixed point iteration and an approximate root $\bar{x} = 0.1140 \approx x^*$ is obtained. Estimate the error in the approximation $\bar{x}$.

(3p) **5:** Consider 4×4 matrices.

a) The Gauss transformation used in the first step of computing the LU decomposition can be written as

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{pmatrix}.$$

Write down the matrix M_1^{-1}. Also give a formal proof that shows that your proposed matrix actually is the correct inverse.

b) Let

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -1 & 1.2 \\ -1 & 1.7 & 0.9 \end{pmatrix},$$

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

c) Suppose we want to solve a linear system $Ax = b$, but where only an approximate right hand side b_δ, satisfying an error bound $\|\Delta b\| \leq \delta$, is available. Show that

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

where $\kappa(A)$ is the condition number, and where $\|\cdot\|$ denotes any of the norms $\|\cdot\|_2$, $\|\cdot\|_1$, or $\|\cdot\|_\infty$, *(4p)* **6:** Points (x_i, y_i) on an ellipse satisfy the equation $c_1x^2+c_2xy+c_3y^2+c_4x+c_5y+1=0$. Let the following data be given

x_i	0.2366	0.3375	0.2286	0.1145	-0.0866	-0.5321	-0.6732
y_i	1.2995	1.4137	1.6712	1.8473	2.0610	2.3195	2.4774

and do the following

a) Formulate the problem of finding the coefficients $c_1, c_2, \ldots,$ and c_5 as a least squares problem $Ax = b$. Give the matrix A, the solution x and the right hand side b explicitly.

Hint Give A and b in terms of the data (x_i, y_i) symbolically. Don't write numbers.

b) We compute the reduced QR decomposition, i.e. $QR = A$, of the above matrix A. Give the dimensions of the matrices Q and R.

c) In the general case where A is an $m \times n$ matrix, b is a vector of length m, and the reduced QR decomposition is given. Clearly show how many floating point operations that are required to compute the solution x to the least squares problem $\min \|Ax - b\|_2$.

(2p) **7:** To compute the derivative $f'(2)$ we can use the formula

$$Df(2) = \frac{1}{2h}(-f(x+2h) + 4f(x+h) - 3f(x)).$$

When the formula is applied for a few different h values we obtain the results

h	0.2	0.1	0.05
error	0.342	0.0861	0.0209

Assume that the error is proportional to h^p and use the table to determine p.

(3p) **8:** **a)** Let

$$s(x) = \begin{cases} x+1 & 0 \le x < 1, \\ x^3 - 3x^2 + 4x & 1 \le x < 2. \end{cases}$$

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) Let $h > 0$ be a step size. The B–spline basis function $B(x)$ is the unique natural cubic spline that interpolates the table

$B(x)$	0	1/6	2/3	1/6	0
x	-2h	-h	0	h	2h

Introduce the functions $B_k(x) = B(x-kh)$, and a uniform grid $x_1 < x_2 < \ldots < x_n$, with $h = x_i - x_{i-1}$. Answer the following questions: What is the dimension of the space consisting of all the cubic splines defined on the grid $\{x_i\}_{i=1}^n$? Which of the basis functions $B_k(x)$ are non-zero on the interval $[x_1, x_n]$? Also show that the functions $\{B_k(x)\}$ are linearly independent. Write down a basis for for the linear space consisting of all cubic splines defined on the grid $\{x_i\}_{i=1}^n$. Motivate your choice carefully.

Answers

(5p) **1:** For **a)** we obtain the approximate value $\bar{a} = 0.05347$ which has 4 significant digits. The absolute error is at most $|\Delta a| \leq 0.5 \cdot 10^{-5}$.

In **b)** the unit round off for the floating point system if $\mu = 0.5 \cdot 10^{-3}$. This is an upper bound for the relative error when a number is stored on the computer.

For **c)** we insert the Taylor series for e^x into the expression for $f(x)$ to obtain

$$f(x) = \frac{(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots) - 1}{x} = 1 + \frac{x}{2} + \frac{x^2}{6} + \ldots = \bar{f}(x) + R_T.$$

Thus the leading term in the truncation error is $x^2/6$ and $p = 2$.

d) The approximate value is $\bar{y} = \sqrt{\bar{a}} = \sqrt{2.48} = 1.57$ with $|R_B| \leq 0.5 \cdot 10^{-2}$. The error propagation formula gives

$$|\Delta y| \lesssim |\frac{\partial y}{\partial a}||\Delta a| = |\frac{1}{2\sqrt{a}}||\Delta a| < 0.013.$$

The total error is $|R_{TOT}| \leq 0.013 + 0.5 \cdot 10^{-2} < 0.02$. Thus $y = 1.57 \pm 0.02$.

(2p) **2:** The interpolating polynomial is

$$p(x) = 1.5\frac{(x-3)(x-5)}{(1-3)(1-5)} + 2.2\frac{(x-1)(x-5)}{(3-1)(3-5)} + 3.4\frac{(x-1)(x-3)}{(5-1)(5-3)}.$$

There is no need to simplify the expression.

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

$$f(x) = \frac{1-\cos(x)}{\sin(x)} = \frac{1-c}{s} = \frac{d}{s} = e.$$

The error propagation formula gives us

$$|\Delta f| \lesssim |\frac{\partial f}{\partial c}||\Delta c| + |\frac{\partial f}{\partial s}||\Delta s| + |\frac{\partial f}{\partial d}||\Delta d| + |\frac{\partial f}{\partial c}||\Delta e| = |\frac{1}{s}||\Delta c| + |\frac{1-c}{s^2}||\Delta s| + |\frac{1}{s}||\Delta d| + |\Delta e| \leq$$

$$\mu(|\frac{c}{s}| + |\frac{1-c}{s}| + |\frac{d}{s}| + |e|) \approx \frac{\mu}{x},$$

where we have used $c \approx 1$, $s \approx x$ and $d/s = e = f \approx x/2$. An alternate formula that avoids the cancellation is

$$f(x) = \frac{\sin(x)}{1+\cos(x)}.$$

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

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

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

For **b)** we rewrite $f(x) = 1 + x^2 - 3\sqrt{x} = 0$ as $(1+x^2)^2 = 9x$. One possible iteration formula is thus $x_{n+1} = g(x_n) = (1 + x_n^2)^2/9$. Since

$$g'(x) = \frac{2}{9}(1 + x^2)2x \text{ and } g'(0.11) = 0.0495 < 1,$$

the method is convergent.

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

$$|x - \bar{x}| \le \frac{|f(\bar{x})|}{|f'(\bar{x})|} \le \frac{7.95 \cdot 10^{-5}}{4.2} < 1.9 \cdot 10^{-5}.$$

(3p) **5:** For **a)** we propose the inverse

$$M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -m_{21} & 1 & 0 & 0 \\ -m_{31} & 0 & 1 & 0 \\ -m_{41} & 0 & 0 & 1 \end{pmatrix}.$$

There are several ways to show that this is indeed the inverse. The simplest to write down is that

$$M_1^{-1} M_1 x = M_1^{-1} \begin{pmatrix} x_1 \\ x_2 + m_{21}x_1 \\ x_3 + m_{31}x_1 \\ x_4 + m_{41}x_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 + m_{21}x_1 - m_{21}x_1 \\ x_3 + m_{31}x_1 - m_{31}x_1 \\ x_4 + m_{41}x_1 - m_{41}x_1 \end{pmatrix} = x,$$

for every vector x. Thus $M_1^{-1} M_1 = I$.

For **b)** we note that the second row gives the largest sum and $\|A\|_\infty = |2| + |-1| + |1.2| = 4.2$.

Finally, **c)** is solved by noting that the systems $A(x + \Delta x) = b + \Delta b$ and $Ax = b$ both holds. Subtracting gives $A\Delta x = \Delta b$ or $\Delta x = A^{-1}\Delta b$. Taking norms we find that $\|\Delta x\| \le \|A^{-1}\|\|\Delta b\|$. Also $\|b\| = \|Ax\| \le \|A\|\|x\|$. Thus

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\|A^{-1}\|\|\Delta b\|}{\|b\|/\|A\|} = \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|}.$$

(4p) **6:** For **a)** we remark that each data point (x_i, y_i) gives one row of the over determined system $Ax = b$. The model is $c_1x_i^2 + c_2xy + c_3y_i^2 + c_4x_i + c_5y_i = -1$. Thus the system $Ax = b$ is

$$\begin{pmatrix} x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 \\ x_2^2 & x_2y_2 & y_2^2 & x_2 & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_7^2 & x_7y_7 & y_7^2 & x_7 & y_7 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}.$$

For **b)**, we note that A is 7×5 and Q has the same dimension. Also R is 5×5 upper triangular.

Finally for **c)** we note that the solution is computed using the formula $x = R^{-1}Q^Tb$, where R is $n \times n$ upper triangular and Q is $m \times n$. The matrix vector multiplication requires mn multiplications and additions, or $2mn$ floating point operations. Multiplication by R^{-1} is equivalent to solving the upper triangular system using backwards substitution. The formula for a general step is

$$x_i = (\sum_{j=i+1}^{n} r_{ij}x_j)/r_{ii},$$

which requires $n - i - 1$ multiplications and additions, and also one division. The total amount of work is approximately

$$\sum_{i=1}^{n} 2(n-i) \approx n^2.$$

Thus computing the solution requires $2mn + n^2$ floating point operations.

(2p) **7:** We denote the error by $\epsilon_h \approx Ch^p$. Then

$$\frac{\epsilon_{h_1}}{\epsilon_{h_2}} \approx \frac{Ch_1^p}{Ch_2^p} = (\frac{h_1}{h_2})^p.$$

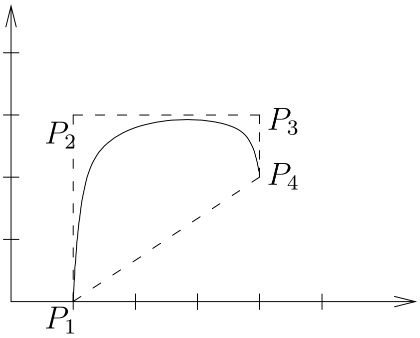
Insert numbers from the table we obtain

$$2^p = (\frac{0.2}{0.1})^p \approx \frac{\epsilon_{0.2}}{\epsilon_{0.1}} = \frac{0.342}{0.0861} \approx 3.97 \text{ and } 2^p \approx \frac{\epsilon_{0.1}}{\epsilon_{0.05}} = \frac{0.0861}{0.0209} \approx 4.11.$$

We see that $2^p = 4$ which means $p = 2$.

(3p) **8:** For **a)** the function $s(x)$ is a cubic spline since $s(x)$, $s'(x)$ and $s''(x)$ are continuous at $x = 1$.

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 dimension of the space of cubic splines defined on the grid $\{x_i\}_{i=1}^{n}$ is $n+2$ since adding n interpolation conditions and two end point conditions makes the spline unique. The function $B_k(x)$ is non-zero on the interval $(k-2)h < x < (k+2)h$. Also $x_n = (n-1)h$. Thus the functions $B_k(x)$, for $k = -1, 0, 1, \ldots, n$ are non-zero on $[x_1, x_n]$. This is a total of $n+2$ functions. The easiest way to show that the functions are linearly independent is to start with $\{B_k(x)\}_{k=-1}^{j-1}$ and add the next function $B_j(x)$. Since $B_j(x)$ is non-zero on the last interval $(j+1)h < x < (j+2)h$, where all the previous functions are zero, the function B_j has to be linearly independent of B_k for $k < j$. Repeat the argument and all the basis functions are linealy independent. Finally the basis is $\{B_k(x)\}_{k=-1}^{n}$ since the functions are linearly independent and the number of functions match the dimension of the space.