Numerical Analysis — FMN011 — 120528 Solutions

The exam lasts 4 hours. A minimum of 35 points out of the total 70 are required to get a passing grade. These points will be added to those obtained in your two home assignments, and the final grade is based on your total score.

Justify all your answers and write down all important steps. Unsupported answers will be disregarded.

During the exam you are allowed a pocket calculator, but no textbook, lecture notes or any other electronic or written material.

1. **(5p)** True or false:

   (a) The spectral radius of A is defined as $\rho(A) = \{\max_{\lambda}|\lambda|, Ax = \lambda x, x \neq 0\}$. If $\rho(A) = 0$, then $A = 0$.
   Solution: *False. Counterexample:*

   $$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

   (b) Gauss elimination for very large systems of linear equations can produce large truncation errors.
   Solution: *False. It is a direct method so it has only round-off errors.*

   (c) A spline is a polynomial and therefore has infinitely many continuous derivatives.
   Solution: *False. It is a piecewise polynomial and a spline of degree n has up to $n-1$ continuous derivatives at the nodes.*

   (d) The power method is a fixed point iteration.
   Solution: *True. $g(x) = Ax$.*

   (e) To find the determinant of a matrix, we can use the singular value decomposition (svd) but not the QR algorithm.
   Solution: $\det(A) = \pm r_1 \cdots r_k$.

2. **(4p)** The following reaction occurs when water vapor is heated:

   $$H_2O \rightleftharpoons H_2 + \frac{1}{2}O_2$$

   The fraction x of H_2O that is consumed satisfies

   $$K = \frac{x}{1-x}\sqrt{\frac{2p_t}{2+x}}$$

   where K and p_t are constants.

   Which of these methods can be used to find x?

   (a) Bisection method

   (b) Newton-Raphson method

   (c) Fixed point iteration

(d) Gauss elimination

(e) Interpolation

(f) Least squares fitting

(g) QR algorithm

(h) SOR

(i) svd

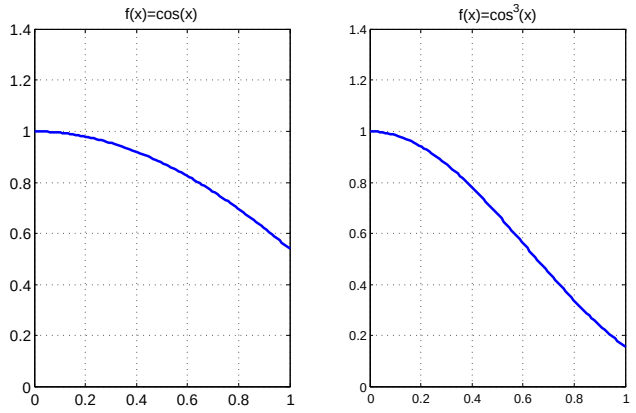
(j) Gram-Schmidt orthogonalization

Solution: *It is a nonlinear equation. Bisection, Newton-Raphson and fixed point iteration.*

Would your suggested method(s) need an initial guess? In that case, what could a reasonable initial guess be? How could the residual be calculated after computing x? Bisection needs an initial interval, the other two need an initial guess.

Solution: *If $K > 0$, as $x \in (0,1)$, you can take an interval [0.1,0.9] for bisection (if they give opposite signs, otherwise try more extreme values) and an initial guess of $x \approx 1/2$ for N-R and fixed point iteration. The residual is calculated by subtracting the right hand side from K after plugging in the calculated solution.*

3. **(4p)** The following figure shows the plots of the functions $f(x) = \cos(x)$ and $f(x) = \cos^3(x)$.



When I applied the fixed point iteration method with $x_0 = 0$, the method converged to the solution only for the first function. Explain why it converged for $f(x) = \cos(x)$ but not for $f(x) = \cos^3(x)$.
Solution: *For $f(x) = cos(x)$, $|f'(x)| < 1$, except for a finite number of points. For $f(x) = \cos^3(x)$, $0 \leq |f'(x)| \leq 3$, and $|f'(x)| > 1$ in an entire subinterval, which caused the method to diverge.*

4. **(4p)** Estimate:

   (a) How many more steps it takes to solve a nonlinear equation with 5 correct decimal figures with the bisection method if the length of the initial interval is divided by 3.
   Solution: *There was a mistake in the formulation. It should have said How many fewer steps. The number of steps it takes to get an error of e is $N = \log_2((b-a)/e) - 1$. If the interval is $(b-a)/3$, the number of steps will be $N_new = \log_2((b-a)/(3e)) - 1 = \log_2((b-a)/e) - 1 - \log_2 3$. As $\log_2 3 = 1.585$, it takes 2 fewer steps.*

   (b) How much longer it takes to solve n equations in n unknowns using Gauss elimination if n is tripled.
   Solution: *As Gauss takes about $4n^3/3$ operations, if n is tripled it needs $3^3 \cdot 4n^3/3 =$, i.e., 27 times more.*

5. **(6p)** Answer in detail:

   (a) Will the Runge phenomenon show up if $f(x) = e^{x^2}$ is interpolated at a large number of evenly spaced points on the interval [-1, 1]?
   Solution: *The Runge phenomenon shows up only for some functions, for example, of the type $f(x) = 1/(1 + 10x^2)$. It doesn't show for $f(x) = e^{x^2}$. If a great number of points are taken for interpolation, the high degree of the polynomial will cause oscillations anyway.*

   (b) What can you do to offset the Runge phenomenon when it appears?
   Solution: *Use Chebyshev points or reduce the number of interpolation points.*

6. **(5p)** Select the most appropriate answer(s).

   (a) If a function is interpolated at n points $\{x_1, x_2, \ldots, x_n\}$, the error of the interpolation at a point $x \neq x_j$ contained in the interval of interpolation depends on

      i. the basis chosen for the interpolation
      ii. the value of the function at x
      iii. the number of data points

   Solution: *Only on the number of points.*

   (b) All previous computations can still be used when new data points are added in the following type of polynomial representations:

      i. Lagrange's
      ii. Bernstein's
      iii. Newton's

   Solution: *Newton's*

(c) How many extra (boundary) conditions are needed for quadratic splines if there are n data points?

i. 1

ii. 2

iii. n

iv. $n+1$

Solution: *1*

(d) What is the structure of the matrix involved in the construction of a cubic spline?

i. an orthogonal matrix

ii. a diagonal matrix

iii. a tridiagonal matrix

Solution: *tridiagonal*

(e) The Jacobi or Gauss-Seidel methods applied to $Ax = b$ will not converge if

i. A is not strictly diagonally dominant

ii. the largest eigenvalue of the iteration matrix has absolute value equal to 1

iii. the initial guess is not close enough to the exact solution

Solution: *the largest eigenvalue of the iteration matrix has absolute value equal to 1*

7. **(5p)** Draw:

(a) The convex hull of the set of control points.

(b) A sketch of de Casteljau's algorithm for finding the point on the Bézier curve for $t = 1/4$.

(c) A sketch of the Bézier curve with control points (0,0), (0,1), (1,1), (1,0).

(d) The convex hull of the new set of control points, where (0,1) is changed to (0,-1).

(e) A sketch of the curve if the point (0,1) is changed to (0,-1).

8. **(6p)** $A = QR$ with

$$Q = \begin{pmatrix} -0.4155 & -0.29842 & -0.69291 & -0.50812 \\ -0.53573 & 0.81006 & 0.12236 & -0.20452 \\ 0.49246 & 0.48842 & -0.68 & 0.23775 \\ -0.54574 & -0.12726 & -0.20617 & 0.80216 \end{pmatrix}, R = \begin{pmatrix} 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where Q is an orthogonal matrix. If this information is sufficient, answer the following questions. Otherwise explain why it cannot be done.

(a) What is the rank of A?

(b) What is the determinant of A?

(c) What are the eigenvalues of A?

Solution: *rank(A)=3,* $\det(A) = 0$. *The eigenvalues cannot be determined: the QR method would be needed.*

9. **(6p)** Matlab has the built-in integrators `trapz` and `quad`.

   (a) On what method is each one of them based on?
   Solution: *`trapz` is based on the Trapezoidal rule and `quad` on Simpson's Rule.*

   (b) Given data points $(t_i, f(t_i)), \quad i = 1, \ldots, 30$, can either one be used to calculate the integral of f in the interval $[t_1, t_{30}]$? Be specific and explain how it could be done. You do not need to use Matlab syntax.
   Solution: *Both can be used. To use `quad` we must first generate a function by doing some piecewise interpolation.*

   (c) Given the function $f(t) = te^t/\sqrt{1+t^2}$, can either be used to calculate the integral of f in the interval $[a, b]$? Be specific and explain how it could be done. You do not need to use Matlab syntax.
   Solution: *Both can be used. To use `trapz` we must first generate vectors x and y by sampling f at a number of points in the interval.*

10. **(4p)** The trigonometric function

$$P(t) = \frac{a_0}{\sqrt{n}} + \frac{2}{\sqrt{n}} \sum_{k=1}^{n/2-1} (a_k \cos 2k\pi t - b_k \sin 2k\pi t) + \frac{a_{n/2}}{\sqrt{n}} \cos n\pi t$$

    can be used to interpolate points (t_i, x_i).

    (a) How many data points are needed?
    Solution: *As there are $2(n/2-1)+2 = n$ coefficients, we need n data points.*

    (b) What are a_k and b_k?
    Solution: *The real and imaginary parts of the Fourier transform of x_k.*

    (c) This formula is valid for $t \in [0, 1]$. How must the formula be modified if we wish to use it for $s \in [a, b]$?
    Solution: *Substitute $t = (s-a)/(b-a)$*

    (d) What assumptions must be made on the set of points $\{t_i\}$?
    Solution: *They must be equally spaced.*

11. **(5p)** A set of data was interpolated using the Fourier transform and the result was

$$\begin{aligned} P(t) \;=\; & 1.61 - 0.13\cos\,\pi\,t - 0.50\sin\,\pi\,t - 0.19\cos\,2\pi\,t - \\ & 0.21\sin\,2\pi\,t - 0.20\cos\,3\pi\,t - 0.09\sin\,3\pi\,t - \\ & 0.10\cos\,4\pi\,t \end{aligned}$$

    (a) In what interval was the interpolation done?
    Solution: *[0, 2]*

    (b) How many data points were available?
    Solution: *8*

(c) Do a least squares fit of order 6.
Solution:

$$P(t) = 1.61 - 0.13\cos\,\pi\,t - 0.50\sin\,\pi\,t - 0.19\cos\,2\pi\,t - 0.21\sin\,2\pi\,t - 0.20\cos\,3\pi\,t$$

12. (a) **(2p)** The Shannon information formula is

$$I = -\sum_{i=1}^{k} p_i \log_2 p_i$$

Calculate the average number of bits needed (minimum) to code the matrix

$$M = \begin{bmatrix} -8 & -2 & 3 & 1 \\ 6 & 0 & -2 & 0 \\ 2 & 1 & 0 & 1 \\ 0 & 0 & -1 & -1 \end{bmatrix}$$

Solution: $-(4 * log2(1/16) + 2 * 2 * log2(2/16) + 3 * log2(3/16) + 5 * log2(5/16))/16 \approx 2.73$

(b) **(2p)** Construct a Huffman tree for M.

Solution:

1	*00*
-8	*0100*
6	*0101*
3	*0110*
2	*0111*
-1	*100*
-2	*101*
0	*11*

(c) **(2p)** What is the average bits/symbol for this coding? What is the average if the standard binary system is used for the matrix entries?
Solution: *Huffman: 44/16=2.75; standard: 8 symbols will require 3 bits/symbol*

13. **(5p)** A is a real matrix and when the *Matlab* command

```
[X,L]=eig(A)
```

is executed, the following result is obtained:

```
X = 0.3004     -0.73463              -0.73463              -0.57735
    0.042914    0.3025  + 0.25928i    0.3025  - 0.25928i   -0.57735
    0.55788     0.10803 - 0.06482i    0.10803 + 0.06482i   -0.57735
    0.77245    -0.51856 - 0.12964i   -0.51856 + 0.12964i   -1.6726e-015

L = 6      0                     0                     0
    0     -2.3592e-016 + 3i      0                     0
    0      0                    -2.3592e-016 - 3i      0
    0      0                     0                    -2.207e-015
```

(a) What algorithm was used by *Matlab*?
Solution: *QR method*

(b) What does matrix X contain?
Solution: *The eigenvectors of A.*

(c) What does matrix L contain?
Solution: *The corresponding eigenvalues of A.*

(d) Is it possible that matrix A is a symmetric matrix? Justify your answer.
Solution: *No, because symmetric matrices have real eigevalues.*

(e) Is it possible to conclude that matrix A is invertible? Justify.
Solution: *No, because the last eigenvalue is 0 up to round-off error.*

14. **(5p)** The svd of a matrix A is $A = USV^T$. The eigenvalues and corresponding eigenvectors of

$$A = \begin{bmatrix} 5 & 6 & -2 & 0 \\ 6 & 0 & 9 & 12 \\ -2 & 9 & -10 & -13 \\ 0 & 12 & -13 & -1 \end{bmatrix}$$

are
1.614 and (0.6741 -0.5458 -0.4961 -0.03837)^T
-27.55 and (0.1306 -0.4854 0.6692 0.5472)^T
5.851 and (-0.5734 -0.2601 -0.5363 0.5621)^T
14.08 and (-0.4470 -0.6315 0.1353 -0.6190)^T.

(a) What are the singular values of A?
Solution: *As the matrix is symmetric, the singular values are $|\lambda_i|$, 1.614, 27.55, 5.851, 14.08*

(b) What are the right and left singular vectors?
Solution: *Right singular vectors: the given eigenvectors. Left singular vectors: the given eigenvectors, except the second one, which is* $(-0.1306, 0.4854, -0.6692, -0.5472)^T$

(c) Find a rank-1 approximation to A. (You need not carry out matrix and vector multiplications.)
Solution: *27.55*$(-0.1306, 0.4854, -0.6692, -0.5472)^T(-0.1306, 0.4854, -0.6692, -0.5472)$

(d) What is the compression rate of this approximation?
Solution: *To store the entire matrix takes 16 elements, to store the rank-1 approximation takes at most 9 elements. The compression is 9/16.*