Uppsala universitet
Institutionen för informationsteknologi
Avdelningen för beräkningsvetenskap

Old exam questions, Scientific computing for Data Analysis

(Some questions are updated to fit the course and some are new.)

Time: 3 hours
Tools: Formula sheet (attached), pocket calculator, Python?.
Course goals: Concepts, Algorithm, Analysis

Part A questions

1. (Concepts)
   At least one correct answer is required in (a) and (b), respectively.

   (a) Explain briefly what we mean by the concepts

      i. Order of accuracy

      ii. Markov process

   (b) What concepts are we asking for below

      i. The maximum relative error when storing a real number as a floating point number.

      ii. The method that minimizes the 2-norm difference between right-hand-side b and Ax (A has more rows than columns)

2. (Algorithm)
   The table show thermal conductivity (unit W/cmK) for iron at different temperatures T ($\circ K$).

T	100	200	300	400
Conductivity	1.3	0.9	0.8	0.7

   Fit a 2nd degree polynomial to the data and find a approximative value for conductivity at temperature $T = 150K$.

3. A company makes (and sell) salads with certain ingredients like tomatoes, spinach etc. Now, over time the price of the ingredients (we call them ingredient A, B, ...) used in each batch of salads vary according to

Ingredient	Price
A	Normaldistributed with mean μ_A and standard deviation σ_A
B	Normaldistributed with mean μ_B and standard deviation σ_B
$\vdots$	$\vdots$

The company can sell the salads for a certain price, and want to know if they make a profit or not.

(a) (Algorithms)
Write an algorithm that perform a Monte Carlo simulation to answer the question "in average, what is the cost for making a batch of salads". For simplicity, you can presume three ingredients (A, B and C).
Also you can presume that the function `randn( )` is available, which generates one normal distributed random number, $N(0, 1)$, i.e. mean 0 and standard dev 1. The code `x = mu + s*rand()` will generate a random number `x` with mean `mu` and standard deviation `s`.

(b) (Analysis)
Assume you have done the simulation in (a), but you would like to repeat it with double accuracy. How do you achieve that? Motivation required.

4. Regression analysis leads to an overdetermined system $Ax = b$ ($m \times n$ and $m > n$). The least squares solution can be achieved in different ways and one such solution method is based on the QR-decomposition of A.

(a) (Algorithms)
Show how the QR-decomposition is used to solve the system $Ax = b$. It must be clear from your explanation where and how Q and R are used, and also how you get from the expression $Ax = b$ to the expression where Q and R are involved (show the steps!).

(b) (Analysis)
It is said that orthogonal matrices preserve length. Prove that it is true!

5. (Algorithm)
Assume you've used a Python function `MCqueuesim(T, N)` to do a queue simulation. Here, `N` is the number of realisations in the Monte Carlo method and `T` is the final time in the simulation. A call to this function, will perform a Monte Carlo simulation, simulate a queue and output the resulting queue at time `T`. Now you would like to get an error estimate (an error bound) of the Monte Carlo calculation. Write the code or algorithm that would result in such an error bound (or explain how to do it). The code can be written as pseudocode, and it does not have to be correct Python code.

6. (Algorithm)
Regression analysis leads to the least squares problem $\|b - Ax\|_2$. There are different methods for solving the least squares problem. List three methods and explain the pros and cons with each method.

7. (Algorithm)
Do two iterations with the Power method, when the matrix is

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}.$$

Use initial vector $x^{(0)} = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$. According to the theory, which eigenvalue and eigenvector would the method find if you continued the process?

8. A (tiny) data set has 5 samples and two variables and the data is stored in the matrix A:

$$A = \begin{pmatrix} 11 & 10 \\ 4 & 6 \\ 12 & 18 \\ 9 & 8 \\ 4 & 8 \end{pmatrix}$$

   (a) Write the matrix as a series of rank-1 matrices. Use Python as a "pocket calculator" (to do the computations). Write down the result, but you can round the numbers to two decimal places.

   (b) If you throw away the last rank-1 matrix in the series and form a new matrix A_1 (which also will be rank-1 in this case), how big would the 2-norm difference $A - A_1$ be. Motivate your answer.

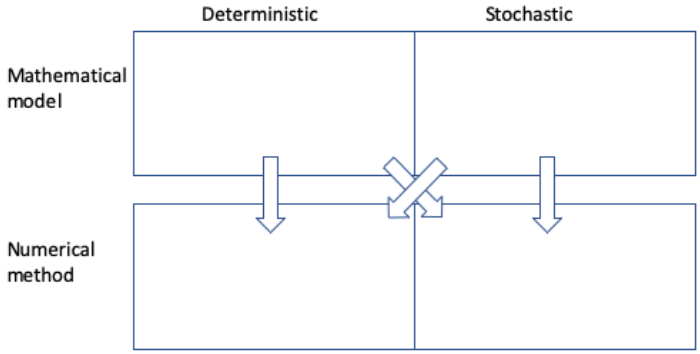
   (c) From the result in (a), form the pseudoinvers A^+ and solve the problem $Ax = b$ using A^+, when

$$b = \begin{pmatrix} 20 \\ 8 \\ 25 \\ 20 \\ 10 \end{pmatrix}$$

   Show on paper explicitly how the SVD in (a) is used to form A^+, but you can do the calculation using Python. Will the result be the exact solution to $Ax = b$?

Part B questions

1. In this course we have been dealing with deterministic/stochastic problems, described as mathematical models. These models can be solved with deterministic and/or stochastic numerical methods. The figure below is supposed to represent the relationships between the models and the methods.
   Replicate the figure to your answer sheet and exemplify with models (or types of models) and methods, i.e. fill in the boxes. Also, exemplify when a deterministic/stochastic model can be solved with a deterministic/stochastic method (represented by the arrows in the figure). Note that not all relationships represented by the arrows are necessarily possible in reality.



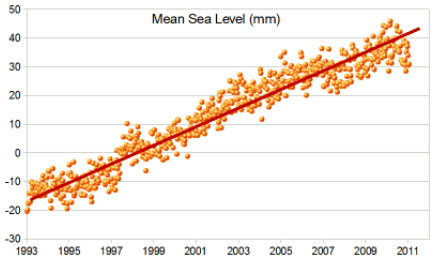
   (For example, if a certain deterministic mathematical model can be solved by a specific deterministic numerical method, write the model in the upper left box and specify the method in the lower left box. You don't need to cover all metods and models, just give examples.)

2. Below is a stochastic model of gene transcription and translation:

$$
\begin{aligned}
&G \xrightarrow{v \cdot G} G + M, \quad \text{transcription} \\
&M \xrightarrow{x \cdot M} M + P, \quad \text{translation} \\
&M \xrightarrow{y \cdot M} \emptyset, \quad \text{degradation} \\
&P \xrightarrow{z \cdot P} \emptyset, \quad \text{degradation}
\end{aligned}
$$

   where G denotes a gene, M m-RNA, P a protein and v, x, y, z the propensities. This problem can be simulated using the SSA algorithm (Gillespies algorithm) as you did in the mini project.
   Assume that, at a certain step in the SSA algorithm, we have $v \cdot G = 1.6$, $x \cdot M = 1.2$, $y \cdot M = 0.4$ and $z \cdot P = 0.8$. Also, a random number $u = 0.55$ from a uniform distribution has been generated, and that the time step has been calculated. Under these premises, show and explain what this particular step in the SSA algorithm will result in.

3. The Figure shows data on Earths average sea level over a number of years, and a trend line (red line).



Explain how such a trend line is computed, the algorithms behind it. Also explain why (for some algorithms) there might be problems with accuracy, and how you can improve the situation. Exemplify your explanation by using the below subset of the data (the numbers are ”simplified” to make the calculation more human). You don't need to solve the problem fully, i.e. to find the final trend line, but you must set up the equations and show the steps for finding the trend line.

year	1994	1998	2000	2005	2011
mean sea level (mm)	-18	0	5	22	38