Computational Statistics
Exam, Fall 2023

Department of Computer and Information Science (IDA), Linköping University

January 9, 2024, 8:00-13:00

Course:	732A90 Computational Statistics
Teachers:	Krzysztof Bartoszek, Frank Miller (examiner)
Allowed aids:	Printed course books, 100 page computer document
Provided aids:	material in zip file exam_material_732A90.zip
Grades:	A = [18, 20] points, B = [16, 18) points, C = [14, 16) points, D = [12, 14) points, E = [10, 12) points, F = [0, 10) points.
Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. If you are unable to include a plot in your solution file clearly indicate the section of R code that generates it. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in an appendix. In a number of questions you are asked to do plots. Make sure that they are informative, have correctly labelled axes, informative axes limits and are correctly described. Points may be deducted for poorly done graphs. If you have problems with creating a pdf you may submit your solutions in text files with unambiguous references to graphics and code that are saved in separate files. There are TWO assignments (with sub–questions) to solve. Provide a separate solution file for each assignment. Include all R code that was used to obtain your answers in your solution files. Make sure it is clear which code section corresponds to which question. If you also need to provide some hand–written derivations please number each page according to the pattern: Question number . page in question number, i.e. Q1.1, Q1.2, . . ., Q2.1, Q2.2, ... Name your solution files as: [your exam account id] [own file description].[format]

Note: If you are not able to solve a part of a question, you can anyway try to show how you would solve the subsequent parts with explaining and providing code examples and you might receive partial points.

Assignment 1 (10p)

We consider a case when a Gumbel distributed observation x is made. The Gumbel distribution has a known scale parameter 1; the location parameter μ is unknown; this means that the likelihood for the parameter given the observation is proportional to

$$g(\mu|x) = \exp(-x + \mu - \exp(-x + \mu)).$$

A Bayesian analysis should be done and the prior distribution for μ is here assumed to be a standard normal distribution; this means, the prior has a density proportional to

$$h(\mu) = \exp(-\mu^2/2).$$

The posterior density $f_{\text{posterior}}(\mu|x)$ is then the normalized product of likelihood and prior and therefore equal to $c \cdot g(\mu|x) \cdot h(\mu)$ with a constant $c > 0$. We assume in this assignment that the observation made is $x = 1.65$.

You can use the following results: The derivative of $g(\mu|x) \cdot h(\mu)$ with respect to μ is

$$g(\mu|x)h'(\mu) + g'(\mu|x)h(\mu),$$

(product rule) with

$$\begin{aligned} h'(\mu) &= -\mu \exp(-\mu^2/2) = -\mu h(\mu), \\ g'(\mu|x) &= g(\mu|x)(1 - \exp(-x + \mu)). \end{aligned}$$

Question 1.1 (2p)

Plot the posterior density $f_{\text{posterior}}(\mu|x = 1.65)$ (you can ignore the constant c, here; this means, you can plot $g(\mu|x) \cdot h(\mu)$). Plot also the derivative.

Question 1.2 (3p)

Write an algorithm to maximize the posterior using the secant method (the value $\hat{\mu}$ which maximizes the posterior is the so-called maximum a posteriori estimate). Note that the constant c is not relevant here. Choose two pairs of starting values: One pair which leads to convergence to the right global maximum at $\hat{\mu}$ and another pair not leading to convergence to that maximum. Explain what happens in the unsuccessful case.

Question 1.3 (4p)

Write an own program using a Metropolis algorithm to generate draws of the posterior. The proposal distribution should be of random-walk type: current generated random number plus a draw of the uniform distribution between $-a$ and a, Unif$[-a, a]$. Use $a = 0.15$ and $a = 1.5$. Discuss which of the a-values is better when you want to generate 1000 observations and why (give at least two different arguments).

Question 1.4 (1p)

Using the better a-value, plot a histogram of the generated random variables following the posterior. Based on your generated sample, compute a Monte Carlo estimate for $P(\mu \geq 0)$.

Assignment 2 (10p)

Your task is to maximize the function

$$f(x) := \frac{x^2}{e^x} - 2\exp(-(9\sin x)/(x^2 + x + 1))$$

for $x \geq 0$ using a genetic algorithm.

Question 2.1 (1p)

Visualize the function to maximize and mark the global maximum.

Question 2.2 (2p)

Decide how you will encode an individual and implement appropriate crossover and mutation functions.

Question 2.3 (2p)

The next step is to implement a selection procedure in order to choose the pool of individuals that will be allowed to contribute to the next generation. Implement a solution that takes advantage of the representation of the numbers and also of the suspected location of the maximum.

Question 2.4 (5p)

Implement all the other necessary components of the genetic algorithm and use it to find the maximum of the function $f(x)$. Take population size of 50. You may not take a constant starting population at the maximum. The starting population has to be random but it may use information on the **approximate (do not use the exact one)** location of the maximum.

It is recommended that you first try your code on a smaller population (say 6) and only after it runs without error, run it on the population of size 50. Follow your population for at least 100 generations. Provide example calls to your code.

Visualize the population's behaviour with the number of generations for mutation probabilities 0.1, 0.5, and 0.9 and provide comments. At which generation was the best value found? Was the maximum found? Can you explain the observed behaviour, especially when taking into account the mutation probability?