Examination Computational Statistics

Linköpings Universitet, IDA, Statistik

Course code and name:	732A90 Computational Statistics
Date:	2017/03/22
Assisting teacher:	Krzysztof Bartoszek
Allowed aids:	Printed books and 100 page computer document
Grades:	A= $[18 - 20]$ points
	B= $[15.5 - 18)$ points
	C= $[10.5 - 15.5)$ points
	D= $[8.5 - 10.5)$ points
	E= $[7 - 8.5)$ points
	F= $[0 - 7)$ points **(FAIL)**
Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. 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.

Assignment 1 (10p)

Let

$$f(x) = 1 - (\sin(5x) + \cos(7x))^2 - 5\frac{1}{\sqrt{2\pi \cdot 0.05}} \exp(-x^2/(2 \cdot 0.05)).$$

Question 1.1 (2p)

Plot $f(x)$ and explore (manually) the region around the global minimum. Is it exactly at 0 or slightly off? Justify why the global minimum should be around 0.

Question 1.2 (2p)

Use R's `optim()` function to find the global minimum. Try the methods "Nelder-Mead", "BFGS", "CG" and "SANN". For each method try the following starting values: -10, 0, 10. For which methods and starting points did you arrive at the global minimum? Justify why some work (if they do) and other do not (if they do not).

Question 1.3 (2p)

Providing the gradient of the function to optimizers usually improves the situation. The derivative of this function is

$$f'(x) = -2(\sin(5x)+\cos 7(x))(5\cos(5x)-7\sin(7x))-5\frac{1}{\sqrt{2\pi\cdot 0.05}}\exp(-x^2/(2\cdot 0.05))\cdot\left(\frac{-1}{2\cdot 0.05}\right)2x.$$

Repeat Question 1.3 but now with the gradient included. In what situations did it help? Justify what you observed.

Question 1.4

In some optimization problems it can be recommended to use random search algorithms. One example of an algorithm that randomly explores the state space is a genetic algorithm.
In this part you may choose whether you want to answer Alternative 1 and 2. If you answer both Alternative 1 and Alternative 2 the points will **NOT** sum up. Towards the final grade you will receive the maximum from the two scores (Alternative 1 and Alternative 2) you managed to achieve.

Alternative 1 (4p)

Implement a genetic algorithm to find the global minimum of $f(\cdot)$. Present and briefly justify your choice of state space, individuals to reproduce, individuals to die, crossover operator, mutation operator, population size, initial population and other algorithm parameters. Provide your final found point and plots illustrating how the algorithm searched (e.g. the value of the fittest individual at each time step, the average score of the population).
Tip: Take a population of 100 individuals and 20 generations so that your code finishes quickly. However, should you not obtain a value around the global minimum increase these to check if something is wrong or the algorithm needed more search time.

Alternative 2 (2p)

Discuss how would you implement a genetic algorithm for this problem. What is the state space? Provide examples of crossover and mutation operators. How would you create your initial population? How would you choose individuals to reproduce or to die? What problems could a genetic algorithm run into when searching for the global minimum?

Assignment 2 (10p)

Consider the model $X \sim \text{lognorm}(0,1)$ and $\ln(Y) = 0.5 + 1.5\ln(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$, standard normal, and independent of X. For notational clarity $\ln(X) \sim \mathcal{N}(0,1)$ has mean 0 and variance of 1. In this assignment we are interesting in finding $M = \mathbb{E}[Y/X]$.

Question 2.1 (3p)

Implement the Monte Carlo method for calculating integrals to find M. In your code you may only use a random number generator that generates from the uniform distribution on the interval $[0, 1]$, e.g. `R`'s `runif()`. Point(s) will be deducted for using other generators, in particular (but not limited to) for using `rnorm()` or `rlnorm()`.
Tip: Remember that

$$\mathbb{E}[Y/X] = \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{Y}{X} f_{X,Y}(x, y) \mathrm{d}y \mathrm{d}x,$$

where $f_{X,Y}(x, y)$ is the joint density of the pair (X, Y). To simulate from the density $f_{X,Y}$ you e.g. simulate one component and then the other conditional on the first. Then you estimate $\mathbb{E}[Y/X]$ numerically from the simulated sample.
Tip: Recall that if $\theta \sim \mathrm{Unif}[0, 2\pi]$ and $D \sim \mathrm{Unif}[0, 1]$, then the pair

$$\begin{aligned} X_1 &= \sqrt{-2 \ln D} \cos\theta \\ X_2 &= \sqrt{-2 \ln D} \sin\theta \end{aligned}$$

is a pair of independent and $\mathcal{N}(0, 1)$ distributed random variables.

Question 2.2 (4p)

Study the properties of your Monte Carlo estimator. What is the mean and variance of the estimator for different sample sizes (repeat the procedure many times for each sample size level)? Provide appropriate plots, e.g. the mean of estimator versus sample size, variance of estimator versus sample size. Do not forget to think about e.g. the y–axis limits and point types/sizes so that the graphs are informative.
Tip: If you take too many repeats and/or your sample size is be too large, then this simulation study might take too much time to run. Work with sample sizes up to 100. For testing purposes take about 20 repeats while for your report try with about 100 repeats (try 50 if 100 takes too long).

Question 2.3 (3p)

The true value of M is $M = \exp(9/8) \approx 3.080$. How does your Monte Carlo estimator compare with this value? Repeat appropriate plots from Question 2.2 with this value indicated on them (e.g. as a horizontal line but only on the plot where this makes sense) and discuss. Is your Monte Carlo estimator unbiased? Do an appropriate t–test (function `t.test`) to check. Provide the interpretation of the output.