UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4051/STK9051 — Computational statistics

Day of examination: June 7th, 2024

Examination hours: 15.00 – 19.00.

This problem set consists of 7 pages

Permitted aids: None

Note: the final sheet contains definitions useful formulas relevant in the problems.

Problem 1 Monte Carlo integration

In a risk evaluation we are interested in evaluating a high dimensional integral:

$$\mu_h = \int_{\mathbb{R}^d} h(x) f(x)\, dx \tag{1}$$

Where $x \in \mathbb{R}^d$, $d = 10^8$. Given $\{x_1, x_2, \dots, x_{100}\}$ being 100 independent samples from $f(x)$. We find:

$$m = \frac{1}{100}\sum_{i=1}^{100} h(x_i)\,, \qquad \text{and} \quad v = \frac{1}{99}\sum_{i=1}^{100}(h(x_i) - m)^2\,. \tag{2}$$

a) In terms of m and v, what is the common approximation of μ_h, and how large is the estimate of the uncertainty associated with this approximation? State assumption(s) needed for the approximation and the uncertainty estimate to be valid? How does the dimension of x influence the results? Discuss how to determine whether the set of 100 samples is sufficient.

Problem 2 EM algorithm

A soft sensor is used for continuously monitoring a process. In a period where the process is stationary, we collect the data, $x_1, x_2, \dots, x_T$, and want to investigate the statistical properties of these data. However the system is occasionally disturbed by external forces, which create erroneous readings. To analyze the situation for the soft sensor data we consider a contamination model, where the density for one data is given as:

$$f(x) = (1-p)\phi(x|\mu,\sigma^2) + pf_c(x) \tag{3}$$

Where p is the probability for contamination, $f_c(x)$ is the distribution of a contaminated signal, and the distribution of the uncontaminated signal is assumed to follow a normal distribution with parameters μ, σ^2 denoted $\phi(x|\mu,\sigma^2)$. We will assume that the data collected are independent and identically distributed according to $f(x)$. Also define the unobserved quantities, $C_1, C_2, \dots, C_T$, as:

$$C_i = \begin{cases} 1 & x_i \text{ is contaminated} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

Consider the case where $f_c(x)$ is known. The unknown parameters are $\theta = (\mu, \sigma^2, p)$

a) Give an expression for the likelihood of the data $x_1, x_2, \dots, x_T$, and argue that the likelihood of the complete data, i.e. $(x_i, C_i),\ i = 1, \dots, T$ is given by:

$$l(x, C|\theta) = \tag{5}$$

$$\sum_{i=1}^{T} I(C_i = 0) \cdot \frac{1}{2}\left(2\ln(1-p) - \ln 2\pi - \ln\sigma^2 - \frac{(x_i-\mu)^2}{\sigma^2}\right) + I(C_i = 1)(\ln p\ + \ln f_c(x_i))$$

b) In the context of the EM algorithm what is the interpretation of $Q(\theta|\theta^{(t)})$ function? Identify $Q(\theta|\theta^{(t)})$ for the problem above and derive expressions which updates the parameter estimates in the EM algorithm.

c) We want to quantify how much information is lost due to contamination. Specifically, we want to compare the uncertainty in the estimate of μ and σ^2 from contaminated and clean data. How can you use a parametric bootstrap to quantify this difference?

Problem 3

The stochastic gradient decent (SGD) is used to minimize a function $g(x)$. The standard updating scheme has the form:

$$x_{t+1} = x_t - \alpha_t \cdot z_t. \tag{6}$$

Where $\alpha_t$is the learning rate and z_t, is related to the gradient of $g(x)$. We will analyze a simplified situation with a constant learning rate: $\alpha_t = \alpha,\ g(x) = \frac{1}{2}(x-b)^2$, $z_t = \nabla g(x_t) + \varepsilon_t$, with $\varepsilon_t \sim N(0,1)$. The figure below shows a convergence of SGD using two different learning rates, in both these runs a constant learning rate has been used.

a) Comment on the use of constant learning rate in SGD and how this influence convergence, relate your answer to Figure 1. Analyze the simplified setting above to quantify the impact of the constant learning rate in the example above, starting iterations in x_0. Hint: show first that:

$$(x_{t+1} - b) = (1-\alpha)^2(x_{t-1} - b) - \alpha \cdot \varepsilon'_t, \quad \text{where} \quad (7)$$
$$\varepsilon'_t = (1-\alpha)\varepsilon_{t-1} + \varepsilon_t,$$

and generalize this to show the impact of x_0 and α on the convergence.

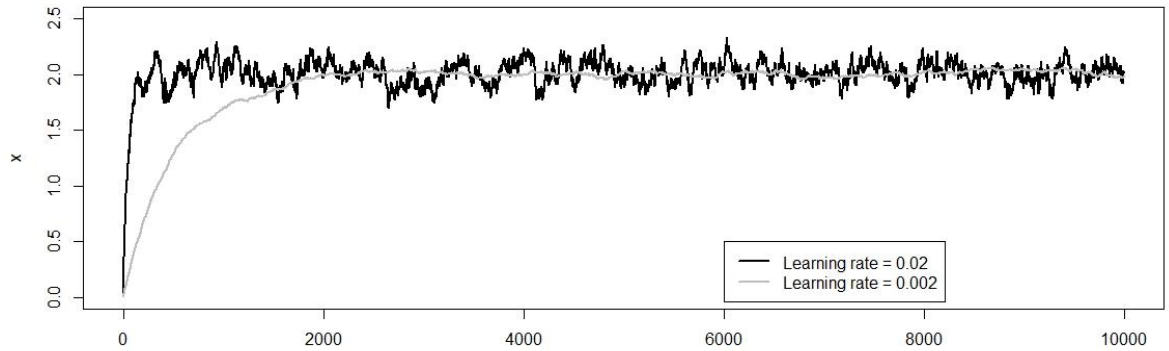

Figure 1: Convergence of SGD. The figure shows the path x_t, for $t = 1, \ldots, 10000$, for two SGD runs with different learning rates.

Problem 4 Slice sampler

We will in this exercise sample from the standard normal distribution by sampling from the uniform distribution corresponding to the area under the density $\phi(x)$, this area is shown in gray in the Figure 1 and is defined $as \quad A = \{(x,u)|u < \phi(x) \text{ and } u > 0\}$. We introduce the additional random variable U which gives the vertical distance. To sample from this distribution, we will use a MCMC approach.

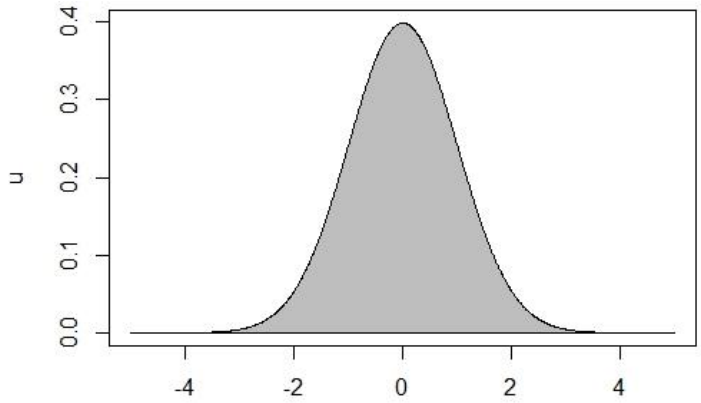

Figure 2: The area A. The area under the standard normal distribution is colored in in gray,

a) State the principles of Gibbs sampling. Argue why the pseudocode below, will give samples which are uniform in Area A.

1) Set $x = 0$
2) Sample $U \sim \text{Uniform}\big(0, f(x)\big)$
3) Sample $X \sim \text{Uniform}(-L, L)$ with $L = \sqrt{-2 \ln u \sqrt{2\pi}}$
4) Repeat from 2 until N iterations have been completed

b) For functions which are more complex than the standard normal distribution, step 3 in the algorithm might be hard to generalize, an alternative is to replace step 3 with a Metropolis-Hastings update of X. Describe the Metropolis-Hastings algorithm. Give details for the update of x if the proposal distribution is: $\text{Uniform}(x - r, x + r)$ with r being the reach of the proposal distribution, and x being the current state of the chain.

c) Which criteria need to be fulfilled for a MCMC algorithm to converge to a target distribution $f(x)$? Comment on the different criteria in relation to the sampling algorithm from b).

d) Table 1 shows the acceptance rate, the effective sample size and the Gelman Rubin statistics for runs using different values of r. Figure 3 shows the sample path of the x chain, and the cumsum diagnostics for the corresponding runs. Comment on the plots and the output, which run would you recommend, and which runs are acceptable? Give arguments for your choices. Is there any additional plots and summary statistics that you would suggest, that could increase the confidence in the results?

Table 1: Summary statistics for five cases. Case number, reach of the proposal distribution, effective sample size and Gelman-Rubin statistics.

Case #	r	p_{acc}	N_{eff}	GR
1	0.1	0.976	14.39	1.145
2	0.2	0.962	37.75	1.018
3	1	0.799	564.77	1.006
4	10	0.165	1326.19	1.001
5	100	0.016	90.77	1.009

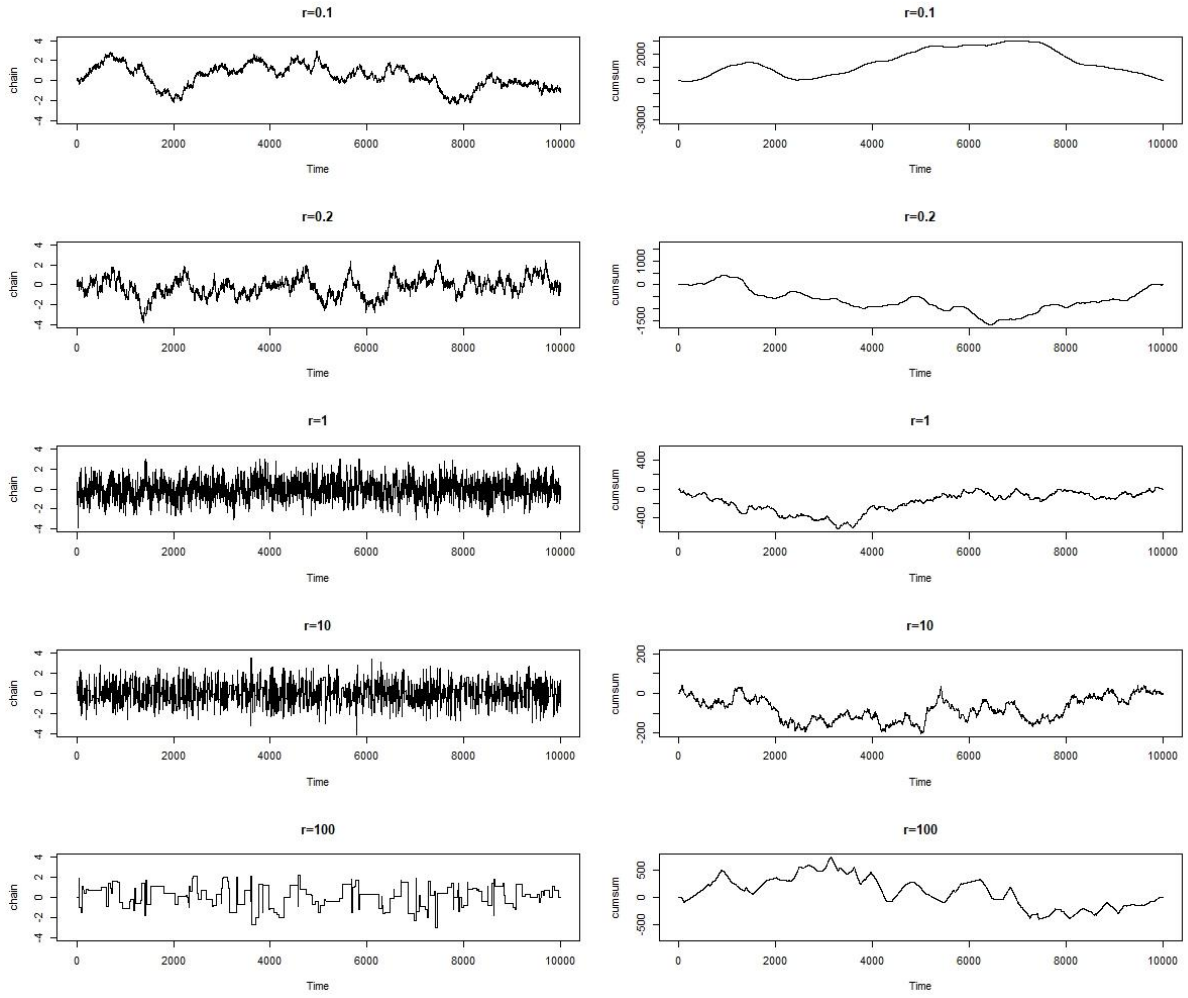

Figure 3: Trace plots and cumsum diagnostics. Each row corresponds to different values of the reach parameter in the proposal distribution. In each row the left panel shows a sample trace of x, the right panel shows the corresponding cumsum diagnostics.

Problem 5 Simulated Annealing

A person who wants to learn about Simulated annealing for discrete optimization, asks a large language model (LLM) the following question: *"In the context of discrete optimization what role does the neighborhood play when using simulated annealing."*

The replay is*: "Certainly! In the context of **discrete optimization** using **Simulated Annealing (SA)**, the neighborhood plays a crucial role. Let's explore how:*

1. ***Objective Function and Solution Space**:*

a) Discrete optimization deals with a finite set of possible solutions.
b) The objective function evaluates the quality of a solution (e.g., minimizing tour length in the Traveling Salesman Problem).
c) The solution space consists of all valid solutions.

2. ***Local Search and Neighborhoods**:*

a) SA iteratively explores neighboring solutions.
b) The neighborhood defines which solutions are adjacent to the current one.
c) For discrete problems, the neighborhood typically involves small changes (e.g., swapping elements in a permutation).

3. ***Escape from Local Optima**:*

a) Local optima are suboptimal solutions better than their immediate neighbors.
b) By restricting the neighborhood, SA avoids getting stuck in local optima.
c) SA occasionally accepts worse solutions (with a probability) to escape local traps.

4. ***Exploration and Exploitation**:*

a) The neighborhood size balances exploration and exploitation.
b) A small neighborhood intensifies the search (exploitation).
c) A larger neighborhood encourages exploration of distant solution regions.

5. ***Computational Efficiency**:*

a) Discrete optimization problems have large solution spaces.
b) Limiting the neighborhood ensures computational efficiency.
c) SA explores a subset of solutions.

6. ***Choosing the Neighborhood**:*

a) Problem-dependent choice.
b) Common neighborhoods include 2-opt (swapping edges) for TSP or changing vertex colors in graph coloring.

In summary, the neighborhood guides SA, allowing it to escape local optima while maintaining efficiency. By defining the neighborhood wisely, SA effectively explores the solution space."

a) Comment on this answer. What is correct (if anything)? What is misleading/wrong (if anything)? What is missing (if anything)? Give further details about how to use simulated annealing for discrete optimization. (give details about the full algorithm not only the neighborhood)

Useful formulas:

Density of Gaussian/Normal distribution:

$$\phi(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), x \in \mathbb{R}$$

Metropolis hastings ratio:

$$R(x_p|x) = \frac{f(x_p)g(x|x_p)}{f(x)g(x_p|x)}$$

$f(x)$ – target distribution.

$g(x_p|x)$ – proposal distribution (for proposing x_p when the current value is x).