Projects - Quantile estimation

- R is the only programming language allowed in answers. Questions requiring R code are indicated by the symbol ♠
- The report (file name: groupe_rapport_noms, e.g., 001_rapport_robert_duche)
  - to be returned in .pdf format and must contain your answers and comments. Careful drafting is expected. It is important to justify/comment on the theoretical and numerical results
  - To integrate all or part of your code and outputs into your report, you can use the dedicated tools Notebook, Rmarkdown or LATEX+ knitr. However, it is forbidden to copy and paste raw code into the body of the text. Graphics must be carefully annotated and presented (title, colour, captions, etc.).
- A version of the code that can be tested must be provided (same file name). This code must:
  - run without errors and reproduce all the results presented in the report. You specify the seed used for the results obtained.
  - be well commented. You may be asked for an oral explanation.
  - use as much as possible the specificities of the language (bonus for the most efficient codes).
- The project will be emailed at the address stanislasduche@gmail.com with the object **groupe_name1_name2 - Monte Carlo Project Part 1**
- The project will be released in 3 times, with blocks of 3/4 exercises. **This is the first block of the project** to be returned for **Sunday November 24^{th} 2024 at 23:59:59. Every hour late starting from November 25^{th}, 00:00:01, will be penalized by a point.**
- The project can be carried out in groups of two or three. For groups of 3, a requirement for detail and initiative will be included in the marking of the project.
- I would remind you that plagiarism is strictly forbidden and any suspicion will result in a 0 for the exercise. The Large Language Models chatbots can be used as an aid; but any code suspected to be written by an assistant will not be corrected.

Exercise 1: Negative weighted mixture

Consider a random variable of a negative weighted mixture X following a density law proportional to

$$\forall x \in \mathbb{R}, \quad f(x) \propto f_1(x) - af_2(x)$$

with $f_1(x) = \mathcal{N}(x; \mu_1, \sigma_1^2)$ and $f_2(x) = \mathcal{N}(x; \mu_2, \sigma_2^2)$ density laws of two normal distributions; and $a > 0$. The objective of this exercise is to understand the behavior of the random variable and to compute statistics.

PREPARED BY Stanislas du Ché

Definition

Question 1: Recall the conditions for a function f to be a probability density. Considering the tail behavior of f, derive necessary conditions on (σ_1^2, σ_2^2) for f to be a density.
Question 2: For given parameters $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_2 = (\mu_2, \sigma_2^2)$, determine a bound a^* on a for $f(x) \propto f_1(x) - af_2(x)$ to be a well-defined density. Provide its normalization constant.
Question 3: (♠) Create an R function `f(a, mu_1, mu_2, s_1, s_2, x)` that produces the pdf of f as a function of x and of the parameters $a, \mu_1, \mu_2, \sigma_1, \sigma_2$. Plot on the same graph the pdf of f for different values of a, especially for $a = a^*$. In another graph, do the same for different values of σ_2, for σ_1 fixed. Examine the impact of a and σ_2 on the shape of f.
From now on, we set the following numerical values: $\mu_1 = 0, \mu_2 = 1, \sigma_1^2 = 9, \sigma_2^2 = 1$, and $a = 0.2$. Check they are compatible with the constraint derived above.

Inverse c.d.f Random Variable simulation

Question 4: (♠) Show that the cumulative density function associated with f is available in closed form. Create an R function `F(a, mu_1, mu_2, s_1, s_2, x)` that produces the cdf of f as a function of x and of the parameters $a, \mu_1, \mu_2, \sigma_1, \sigma_2$. Construct an algorithm that returns the value of the inverse function method as a function of $u \in (0, 1)$, of the parameters $a, \mu_1, \mu_2, \sigma_1, \sigma_2$, and of an approximation precision ε. Deduce an algorithm that implements the inverse function method for the generation of random variables from F.
Question 5: (♠) Write an R function `inv_cdf(n)` that generates n samples from f using the inverse function method. Generate `n = 10000` samples, and graphically check that `inv_cdf()` is correct.

Accept-Reject Random Variable simulation.

Question 6: Describe a method to simulate under f using the accept-reject algorithm and give the expression theoretical acceptance rate according to the parameters of f.
Question 7: (♠) Write a function `accept_reject(n)` that generates n samples from f using the accept-reject method. Generate `n=10000` samples, and graphically check that `accept_reject()` is correct. Compute the empirical acceptance rate and check whether it agrees with its theoretical value.
Question 8: (♠) Vary the value of a and plot the acceptance rate for different values of a. Describe its impact when $a \to a^*$.

Random Variable simulation with stratification

Consider a partition $\mathcal{P} = (D_0, D_1, \dots, D_k), k \in \mathbb{N}$ of $\mathbb{R}$ such that D_0 covers the tails of f_1 and f_1 is upper bounded and f_2 lower bounded in $D_1, \dots, D_k$. We consider the following dominating function g conditioned on the partition:

$$\begin{cases} g(x) = \frac{1}{Z} f_1(x) \text{ if } x \in D_0 \\ g(x) = \frac{1}{Z} \left(\sup_{D_i} f_1(x) - a \inf_{D_i} f_2(x) \right) \text{ if } x \in D_i, i \in \{1, \dots, k\} \end{cases}$$

where Z is the normalization constant of f.
Question 9: Describe an accept-reject algorithm using the dominating function g and the partition $\mathcal{P}$, and calculate the acceptance rate of this new algorithm.
Question 10: For $\delta \in [0, 1]$, prove that there exists a partition $\mathcal{P} = (D_0, D_1, \dots, D_{n_\delta})$, such that the acceptance rate is greater than δ.
Question 11: (♠) Write a function `stratified(n)` that generates n samples from f using the accept-reject method. Generate `n=10000` samples, and graphically check that `stratified()` is correct. Compute the empirical acceptance rate and check whether it agrees with its theoretical value.

PREPARED BY Stanislas du Ché

Question 12: (♠) Write a function `stratified(n,delta)` that generates n samples from f using the accept-reject method with an acceptance rate of δ. You will return the n samples generated and the partition $\mathcal{P}$ used. Generate `n=10000` samples and compute the empirical acceptance rate and check that your code is correct.

Cumulative density function.

Question 13: Write down the cumulative density function $F_X(x)$ for $x \in \mathbb{R}$; and for a given x, write a Monte Carlo estimator $F_n(x)$ using n random variables $(X_i)_{i=1}^n$, *i.i.d.* following the law of X.
Question 14: Prove the strong consistency of the estimator $F_n(x)$ for a given x.
Remark 14: In fact, Glivenko-Cantelli theorem asserts that $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0$ *almost surely. Hence* F_n *is a good estimate of* F *as a function of* x.
Question 15: (♠) Write a function `empirical_cdf(x,Xn)` that returns an estimate of the empirical cdf of f at value x, using n *i.i.d.* random draws `Xn`. Illustrate graphically strong consistency using increasing values of n.
Question 16: Remind the Central Limit Theorem. For $x \in \mathbb{R}$, deduce a 95 % confidence interval for $F(x)$ using $F_n(x)$.
Question 17: (♠) Using a R code, give n, the number of simulation needed to have a 95% confidence interval of $F(x)$ for $x = 1$, and for $x = -15$. What do you notice ?

Empirical quantile function

We define the empirical quantile function defined on $(0, 1)$ by :

$$Q_n(u) := \inf\{x \in \mathbb{R} : u \le F_n(x)\}$$

By the Glivenko-Cantelli theorem, we can deduce the almost surely convergence of $Q_n(u) \xrightarrow{\text{a.s.}} Q(u)$
Question 18: From *i.i.d.* samples $(X_i)_{i=1}^n$ following the same law as X, give the value of $Q_n(u)$ for $u \in (0, 1)$.
Reminder: Lindeberg-Levy Central Limit theorem: At n fixed, consider independent random variables $(X_{n,j})_{1\le j\le n}$ uniformly bounded in j and in n, $\mathrm{Var} X_{n,j} = \sigma_{n,j}^2 < \infty$. If $s_n^2 = \sum_{j=1}^n \sigma_{n,j}^2 \to \infty$ and if $t_n \to t$, then :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{j=1}^n (X_{n,j} - \mathbb{E}(X_{n,j}))}{s_n} < t_n\right) = \int_{-\infty}^t \frac{\exp(-u^2/2)}{\sqrt{2\pi}} du$$

Question 19: Notice that $Y_{j,n} := \mathbb{1}_{X_{n,j} < Q(u) + \frac{t}{\sqrt{n}} \frac{\sqrt{u(1-u)}}{f(Q(u))}}$ are Bernoulli random variable, using the Lindeberg-Levy CLT, deduce a Central Limit Theorem for $Q_n(u)$.
Question 20: What do you notice when $u \to 0$ or $u \to 1$?
Question 21: (♠) Write a function `empirical_quantile(u,Xn)` that returns the empirical quantile using n *i.i.d.* random draws `Xn`. Check numerically the intuition of the previous question for different values of u.
Question 22: (♠) Using a R code, give n, the number of simulation needed to have a 95% confidence interval of $Q(u)$ for $u \in \{0.5, 0.9, 0.99, 0.999, 0.9999\}$.

Quantile estimation Naïve Reject algorithm

Question 23: Describe a method using accept-reject method to simulate a random variable X conditional to the event $\{X \in A\}$, $A \subset \mathbb{R}$. Justify theoretically

Question 24: (♠) Using the previous question algorithm, propose a way to simulate $\delta = \mathbb{P}(X \geq q)$. Compute a function `accept_reject_quantile(q,n)` that returns a Monte Carlo estimate $\hat{\delta}_n^{Reject}$ of the target probability using n random variable simulations.
Question 25: (♠) Compute a confidence interval for δ at level 95% for a required precision ϵ.

Importance Sampling

Question 26: Propose a sampling distribution g to realise importance sampling and remind the importance sampling Monte Carlo estimator $\hat{\delta}_n^{IS}$ of δ for $n \in \mathbb{N}$. On which condition, the importance sampling estimator is preferred to the classical Monte Carlo estimator.
From now on, g will be a Cauchy distribution of the parameters (μ_0, γ)
Question 27: Remind the density of a Cauchy distribution. Choose parameters (μ_0, γ), explain your reasoning.
Question 28: (♠) Compute a R function `IS_quantile(q,n)` that gives, for $n = 10000$, the estimator $\hat{\delta}_n^{IS}$ using n simulated random variables. Compute a confidence interval for δ at level 95% at precision ϵ.

Control Variate

Question 29: Remind the definition of the score. Derive the partial derivative of the log-likelihood $\log f(x|\theta_1, \theta_2)$ under μ_1. We note $s_{\mu_1}(x|\theta) = \frac{\partial \log f(x|\theta_1,\theta_2)}{\partial \mu_1}$. (Remind that $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_2 = (\mu_2, \sigma_2^2)$)
Question 30: Propose a control variate Monte Carlo Estimator $\hat{\delta}_n^{CV}$ using the control random variable $s_{\mu_1}(X|\theta)$
Question 31: (♠) Compute a R function `CV_quantile(q,n)` that gives for $n = 10000$, the estimator $\hat{\delta}_n^{CV}$ using n simulated random variables. Compute a confidence interval for δ at level 95 % at precision ϵ
Question 32: Compare the 3 methods: Naive, Control Variate and Importance Sampling. You can asses their algorithmic complexity, their computational cost for a required precision, etc.