important notes

Given info.

Things to do

ST2195 Coursework Project

Instructions to candidates

This project contains two questions. Answer **BOTH** questions. All questions will be given equal weight (50%).

to obtain sequence of random sampling from random distribution

sampling from a probability distribution

Part 1 In this part, you are asked to work with the Markov Chain Monte Carlo algorithm, in particular the Metropolis-Hastings algorithm. The aim is to simulate random numbers for the distribution with probability density function given below

$$f(x) = \frac{1}{2}\exp\left(-|x|\right),$$

→ find g(x) to help estimate f(x)

where x takes values in the real line and $|x|$ denotes the absolute value of x. More specifically, you are asked to generate $x_0, x_1, \ldots, x_N$ values and store them using the following version of the Metropolis-Hastings algorithm (also known as random walk Metropolis) that consists of the steps below:

Random walk Metropolis

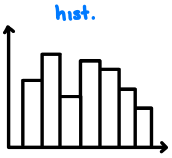
Step 1 Set up an initial value x_0 as well as a positive integer N and a positive real number s.

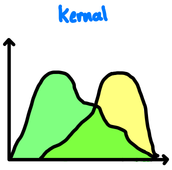
Step 2 Repeat the following procedure for $i = 1, \ldots, N$:

- Simulate a random number x_* from the Normal distribution with mean x_{i-1} and standard deviation s.
- Compute the ratio

$$r\left(x_*, x_{i-1}\right) = \frac{f\left(x_*\right)}{f\left(x_{i-1}\right)}.$$

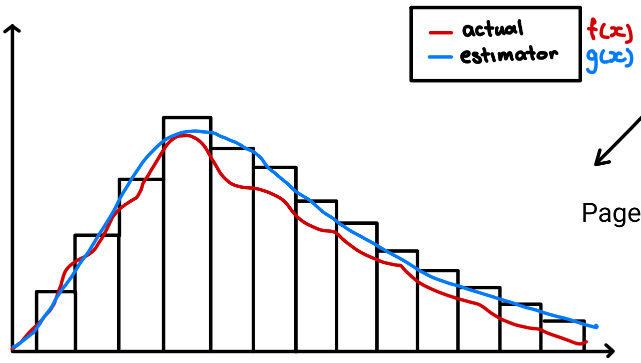
- Generate a random number u from the uniform distribution between 0 and 1.
- If $u < r\left(x_*, x_{i-1}\right)$, set $x_i = x_*$, else set $x_i = x_{i-1}$.





(a) Apply the random walk Metropolis algorithm using $N = 10000$ and $s = 1$. Use the generated samples $(x_1, \ldots x_N)$ to construct a histogram and a kernel density plot in the same figure. Note that these provide estimates of $f(x)$. Overlay a graph of $f(x)$ on this figure to visualise the quality of these estimates. Also, report the sample mean and standard deviation of the generated samples (Note: these are also known as the Monte Carlo estimates of the mean and standard deviation respectively).

Practical tip: To avoid numerical errors, it is better to use the equivalent criterion $\log u < \log r\left(x_*, x_{i-1}\right) = \log f\left(x_*\right) - \log f\left(x_{i-1}\right)$ instead of $u < r\left(x_*, x_{i-1}\right)$.



(b) The operations in part 1(a) are based on the assumption that the algorithm has converged. One of the most widely used convergence diagnostics is the so-called $\widehat{R}$ value. In order to obtain a valued of this diagnostic, you need to apply the procedure below:

- Generate more than one sequence of $x_0, \dots, x_N$, potentially using different initial values x_0. Denote each of these sequences, also known as chains, by $(x_0^{(j)}, x_1^{(j)}, \dots, x_N^{(j)})$ for $j = 1, 2, \dots, J$.
- Define and compute M_j as the sample mean of chain j as

$$M_j = \frac{1}{N} \sum_{i=1}^{N} x_i^{(j)}.$$

and V_j as the within sample variance of chain j as

$$V_j = \frac{1}{N} \sum_{i=1}^{N} (x_i^{(j)} - M_j)^2.$$

- Define and compute the overall within sample variance W as

$$W = \frac{1}{J} \sum_{j=1}^{J} V_j$$

- Define and compute the overall sample mean M as

$$M = \frac{1}{J} \sum_{j=1}^{J} M_j,$$

and the between sample variance B as

$$B = \frac{1}{J} \sum_{j=1}^{J} (M_j - M)^2$$

- Compute the $\widehat{R}$ value as

$$\widehat{R} = \sqrt{\frac{B + W}{W}}$$

In general, values of $\widehat{R}$ close to 1 indicate convergence, and it is usually desired for $\widehat{R}$ to be lower than 1.05. Calculate the $\widehat{R}$ for the random walk Metropolis algorithm with $N = 2000$, $s = 0.001$ and $J = 4$. Keeping N and J fixed, provide a plot of the values of $\widehat{R}$ over a grid of s values in the interval between 0.001 and 1.

Part 2 The 2009 ASA Statistical Computing and Graphics Data Expo consisted of flight arrival and departure details for all commercial flights on major carriers within the USA from October 1987 to April 2008. This is a large dataset; there are nearly 120 million records in total, and it takes up 1.6 gigabytes of space when compressed and 12 gigabytes when uncompressed. The complete dataset, along with supplementary information and variable descriptions, can be downloaded from the *Harvard Dataverse* at `https://doi.org/10.7910/DVN/HG7NV7`

Choose any subset of ten consecutive years and any of the supplementary information provided by the Harvard Dataverse to answer the following questions using the principles and tools you have learned in this course:

(a) What are the best times and days of the week to minimise delays each year?

(b) Evaluate whether older planes suffer more delays on a year-to-year basis.

(c) For each year, fit a logistic regression model for the probability of diverted US flights using as many features as possible from attributes of the departure date, the scheduled departure and arrival times, the coordinates and distance between departure and planned arrival airports, and the carrier. Visualize the coefficients across years.

General Instructions

- All questions should be answered using R and Python for all tasks.

- Your answers should be provided in a separate structured report of no more than 1 page for part 1 and 6 pages for part 2. The page limit excludes title, references and table of contents but includes graphics and tables. The report should be in PDF format and also contain adequate explanations for readers not familiar with programming. In addition to the report, you will also be asked to provide your R and Python code in RMarkdown and Jupyter notebooks, respectively. All the relevant files must be submitted in the designated Atrio or VLE submission portal.

- For part 2, each report should detail all steps you took starting from raw data up to the answer for each question. Any databases you set up, data wrangling/cleaning operations you carry out, and any modelling decisions you make should be clearly described in each structured report. Each report should also include any relevant graphics and tables as part of the answer.

- If you are using elements (e.g. code, databases, graphics, etc) from your answer to a previous question to answer the current one, you will need to refer to those elements.

- You should also supply the code you used to answer each question, in a way that can be used by someone else to replicate your analyses. You can do this either as separate scripts or separate RMarkdown/Jupyter notebooks per question, clearly indicating (both with comments and in the filename) which question each script refers to.

Steps taken from raw data → answer for qns :

Databases setup

data wrangling / cleaning

modeling decisions

relevant graphics and tables

codes used : R and Python (that allows others to replicate)

Codes in seperate Rmarkdown / Jupyter notebooks per Q1

MAX 10 pages excluding titles, references, table of content
including graphics and tables

explanation for people who dk programming

logistic regression

Metropolis - hasting algorithm

```
$p(x) \propto g(x)$

① Select initial value for $x_0$

② for $i = 1, \dots, n$
    repeat: a) Draw $x^* \sim f(x^* | x_{i-1})$
            b) evaluate $\frac{g(x^*) f(x_{i-1} | x^*)}{g(x_{i-1}) f(x^* | x_{i-1})}$
            c) if $\alpha \geq 1$, accept $x^*$ and set $x_i \leftarrow x^*$ ✓
                  $0 < \alpha < 1$, accept $x^*$ and set $x_i \leftarrow x^*$ $\bar{w}$ prob. $\alpha$
                                              or
                        reject $x^*$ and set $x_i \leftarrow x_{i-1}$ $\bar{w}$ prob. $1-\alpha$
```