Question 1: Computations with Metropolis–Hastings

Consider the following probability density function:

$$f(x) \propto x^5 e^{-x}, \quad x > 0.$$

You can see that the distribution is known up to some constant of proportionality. If you are interested (**NOT** part of the Lab) this constant can be found by applying integration by parts multiple times and equals 120.

1. Use Metropolis–Hastings algorithm to generate samples from this distribution by using proposal distribution as log–normal $LN(X_t, 1)$, take some starting point. Plot the chain you obtained as a time series plot. What can you guess about the convergence of the chain? If there is a burn–in period, what can be the size of this period?

2. . Perform Step 1 by using the chi–square distribution $\chi^2(\lfloor X_t + 1 \rfloor)$ as a proposal distribution, where $\lfloor x \rfloor$ is the floor function, meaning the integer part of x for positive x, i.e. $\lfloor 2.95 \rfloor = 2$

3. Compare the results of Steps 1 and 2 and make conclusions.

4. Generate 10 MCMC sequences using the generator from Step 2 and starting points $1, 2, \ldots,$ or 10. Use the Gelman–Rubin method to analyze convergence of these sequences.

5. Estimate
$$\int_0^\infty x f(x) \mathrm{d}x$$
using the samples from Steps 1 and 2.

6. The distribution generated is in fact a gamma distribution. Look in the literature and define the actual value of the integral. Compare it with the one you obtained.

Question 2: Gibbs sampling

A concentration of a certain chemical was measured in a water sample, and the result was stored in the data `chemical.RData` having the following variables:

- `X`: day of the measurement
- `Y`: measured concentration of the chemical.

The instrument used to measure the concentration had certain accuracy; this is why the measurements can be treated as noisy. Your purpose is to restore the expected concentration values.

1. Import the data to `R` and plot the dependence of `Y` on `X`. What kind of model is reasonable to use here?

2. A researcher has decided to use the following (random–walk) Bayesian model (n=number of observations, $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ are unknown parameters):
$$Y_i \sim \mathcal{N}(\mu_i, \text{ variance} = 0.2), \quad i = 1. \ldots, n$$
where the prior is
$$p(\mu_1) = 1$$
$$p(\mu_{i+1}|\mu_i) = \mathcal{N}(\mu_i, 0.2), i = 1, \ldots, n1 \quad .$$
Present the formulae showing the likelihood $p(\vec{Y}|\vec{\mu})$ and the prior $p(\vec{\mu})$. **Hint**: a chain rule can be used here $p(\vec{\mu}) = p(\mu_1)p(\mu_2|\mu_1)p(\mu_3|\mu_2)\ldots p(\mu_n|\mu_{n1})$.

3. Use Bayes' Theorem to get the posterior up to a constant proportionality, and then find out the distributions of $(\mu_i|\vec{\mu}_{-i}, \vec{Y})$, where $\vec{\mu}_{-i}$ is a vector containing all μ values except of μ_i.

Hint A: consider for separate formulae for $(\mu_1|\vec{\mu}_{-1}, \vec{Y})$, $(\mu_n|\vec{\mu}_{-n}, \vec{Y})$ and then a formula for all remaining $(\mu_i|\vec{\mu}_{-i}, \vec{Y})$.

Hint B:
$$\exp\left(-\frac{1}{d}\left((x-a)^2 + (x-b)^2\right)\right) \propto \exp\left(-\frac{(x-(a+b)/2)^2}{d/2}\right)$$

Hint C:
$$\exp\left(-\frac{1}{d}\left((x-a)^2 + (x-b)^2 + (x-c)^2\right)\right) \propto \exp\left(-\frac{(x-(a+b+c)/3)^2}{d/3}\right)$$

4. Use the distributions derived in Step 3 to implement a Gibbs sampler that uses $\vec{\mu}^0 = (0, \ldots, 0)$ as a starting point. Run the Gibbs sampler to obtain 1000 values of $\vec{\mu}$ and then compute the expected value of $\vec{\mu}$ by using a Monte Carlo approach. Plot the expected value of $\vec{\mu}$ versus X and Y versus X in the same graph. Does it seem that you have managed to remove the noise? Does it seem that the expected value of $\vec{\mu}$ can catch the true underlying dependence between Y and X?

5. Make a trace plot for μ_n and comment on the burn–in period and convergence.

Question 1: Hypothesis testing

In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether or not the draft numbers were randomly selected. The draft numbers (`Y=Draft_No`) sorted by day of year (`X=Day_of_year`) are given in the file `lottery.xls`.

1. Make a scatterplot of Y versus X and conclude whether the lottery looks random.

2. Compute an estimate $\hat{Y}$ of the expected response as a function of X by using a loess smoother (use `loess()`), put the curve $\hat{Y}$ versus X in the previous graph and state again whether the lottery looks random.

3. To check whether the lottery is random, it is reasonable to use test statistics

$$T = \frac{\hat{Y}(X_b) - \hat{Y}(X_a)}{X_b - X_a}, \quad \text{where } X_b = \operatorname{argmax}_X Y(X), X_a = \operatorname{argmin}_X Y(X)$$

   If this value is significantly greater than zero, then there should be a trend in the data and the lottery is not random. Estimate the distribution of T by using a non–parametric bootstrap with $B = 2000$ and comment whether the lottery is random or not. What is the p–value of the test?

4. Implement a function depending on *data* and B that tests the hypothesis
   H_0: Lottery is random
   versus
   H_1: Lottery is non–random
   by using a permutation test with statistics T. The function is to return the p–value of this test. Test this function on our data with $B = 2000$.

5. Make a crude estimate of the power of the test constructed in Step 4:

   (a) Generate (an obviously non–random) dataset with $n = 366$ observations by using same X as in the original data set and $Y(x) = \max(0, \min(\alpha x + \beta, 366))$, where $\alpha = 0.1$ and $\beta \sim \mathcal{N}(183, \text{sd} = 10)$.

   (b) Plug these data into the permutation test with $B = 200$ and note whether it was rejected.

   (c) Repeat Steps 5a–5b for $\alpha = 0.2, 0.3, \ldots, 10$.

   What can you say about the quality of your test statistics considering the value of the power?

Question 2: Bootstrap, jackknife and confidence intervals

The data you are going to continue analyzing is the database of home prices in Albuquerque, 1993. The variables present are `Price`; `SqFt`: the area of a house; `FEATS`: number of features such as dishwasher, refrigerator and so on; `Taxes`: annual taxes paid for the house. Explore the file `prices1.xls`.

1. Plot the histogram of `Price`. Does it remind any conventional distribution? Compute the mean price.

2. Estimate the distribution of the mean price of the house using bootstrap. Determine the bootstrap bias–correction and the variance of the mean price. Compute a 95% confidence interval for the mean price using bootstrap percentile, bootstrap BCa, and first–order normal approximation
   (**Hint**: use `boot()`,`boot.ci()`,`plot.boot()`,`print.bootci()`)

3. Estimate the variance of the mean price using the jackknife and compare it with the bootstrap estimate

4. Compare the confidence intervals obtained with respect to their length and the location of the estimated mean in these intervals.

Question 1: Genetic algorithm

In this assignment, you will try to perform one-dimensional maximization with the help of a genetic algorithm.

1. Define the function
$$f(x) := \frac{x^2}{e^x} - 2\exp(-(9\sin x)/(x^2+x+1))$$
2. Define the function `crossover()`: for two scalars x and y it returns their "kid" as $(x+y)/2$.
3. Define the function `mutate()` that for a scalar x returns the result of the integer division $x^2 \mod 30$. (Operation mod is denoted in `R` as `%%`).
4. Write a function that depends on the parameters `maxiter` and `mutprob` and:
   (a) Plots function f in the range from 0 to 30. Do you see any maximum value?
   (b) Defines an initial population for the genetic algorithm as $X = (0, 5, 10, 15, \ldots, 30)$.
   (c) Computes vector `Values` that contains the function values for each population point.
   (d) Performs `maxiter` iterations where at each iteration
      i. Two indexes are randomly sampled from the current population, they are further used as parents (use `sample()`).
      ii. One index with the smallest objective function is selected from the current population, the point is referred to as victim (use `order()`).
      iii. Parents are used to produce a new kid by crossover. Mutate this kid with probability `mutprob` (use `crossover()`, `mutate()`).
      iv. The victim is replaced by the kid in the population and the vector `Values` is updated.
      v. The current maximal value of the objective function is saved.
   (e) Add the final observations to the current plot in another colour.
5. Run your code with different combinations of `maxiter`$= 10, 100$ and `mutprob`$= 0.1, 0.5, 0.9$. Observe the initial population and final population. Conclusions?

Question 2: EM algorithm

The data file `physical.csv` describes a behavior of two related physical processes $Y = Y(X)$ and $Z = Z(X)$.

1. Make a time series plot describing dependence of Z and Y versus X. Does it seem that two processes are related to each other? What can you say about the variation of the response values with respect to X?
2. Note that there are some missing values of Z in the data which implies problems in estimating models by maximum likelihood. Use the following model
$$Y_i \sim \exp(X_i/\lambda), \qquad Z_i \sim \exp(X_i/(2\lambda))$$
where λ is some unknown parameter.
The goal is to derive an EM algorithm that estimates λ.
3. Implement this algorithm in `R`, use $\lambda_0 = 100$ and convergence criterion "stop if the change in λ is less than 0.001". What is the optimal λ and how many iterations were required to compute it?
4. Plot $\mathrm{E}[Y]$ and $\mathrm{E}[Z]$ versus X in the same plot as Y and Z versus X. Comment whether the computed λ seems to be reasonable.