Problem 1 Suppose that X is a continuously distributed random variable with density

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 2x & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } x > 1. \end{cases}$$

a) How can X be simulated using the inversion method? Suppose that we wanted to estimate $E(X)$ using Monte-Carlo integration. Find the exact variance of the Monte-Carlo estimate of $E(X)$ based on $n = 1000$ realisations of X.

b) Let $U \sim \text{unif}(0, 1)$. Show that $E(\sqrt{U(1-U)}) = \pi/8$. Find the exact variance of a Monte-Carlo estimate of $E(X)$ based on an additional $n = 1000$ antithetic realisations of X generated via the inversion method.

Problem 2 Suppose that X is continuously distributed random variable with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}x} e^{-\frac{1}{2}(\ln x)^2}(1 + s\sin(2k\pi \ln x))$$

for $x > 0$ where $0 \leq s \leq 1$ and $k \in \mathbb{Z}$ are parameters. A graph of f for $s = 1/2$ and $k = 2$ is shown in Fig. 1.

a) What is the name of this distribution and its parameters in the special case of $s = 0$?

Find an algorithm for simulating realisations of X. If you choose rejection sampling as a method, find the acceptance rate (that is, the long-run or 'unconditional' acceptance probability).

Problem 3

Suppose that we observe a random sample $y_1, y_2, \ldots, y_n$ (see Fig. 2) from a Gamma distribution with density

$$f(y) = \frac{\lambda^a}{\Gamma(a)} y^{a-1} e^{-\lambda y}$$

for $y > 0$ and where a and λ are positive parameters.

Representing our prior beliefs about a and λ with a prior density

$$\pi(a, \lambda) \propto \frac{e^{-a}}{\lambda},$$

suppose we want to sample from the posterior distribution of a and λ using Gibbs sampling, updating a and λ in separate steps.

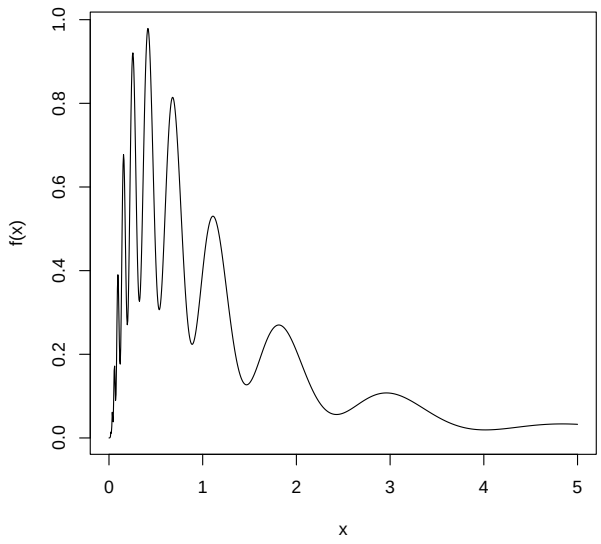

Figure 1: A graph of $f(x)$ in problem 2 for $s = 1/2$ and $k = 2$.

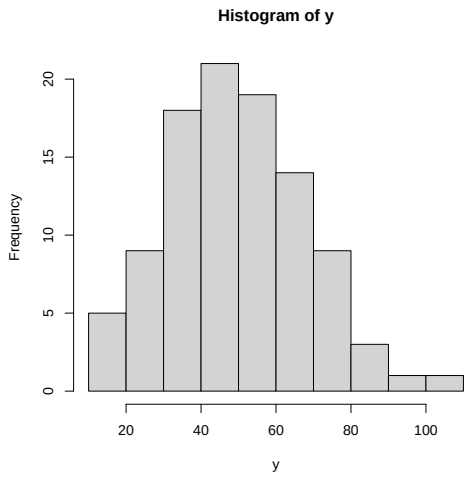

Figure 2: Observations $y_1, y_2, \ldots, x_{100}$ from the density described in problem 3.

a) Show that the full conditional of λ is a Gamma distribution. Find its parameters.

b) Find the full conditional density of a up to a constant of proportionality. To simulate from this distribution we may use Metropolis-Hastings within Gibbs. Using a uniform proposal for the next value a' of a centered around the current value a with maximal step size d, that is, $Q(a'|a) = \frac{1}{2d} I(a - d < a' < a + d)$, find the log of the acceptance probability.

Figs. 3 show trace plots, kernel density estimates of the marginal posteriors of a and λ, and samples from their joint posterior when running the Gibbs sampler in point a) and b) after tuning the step size d to an optimal value of $d = 0.8$ to achieve an acceptance rate of 0.424.

c) Briefly comment on why the convergence of the Markov chain is somewhat slow.

Construct and write an expression (up to a normalizing constant) for the proposal density $Q(a', \lambda' \mid a, \lambda)$ for an alternative Metropolis-Hastings algorithm in which a and λ are updated in a single block. The joint proposal should involve a random walk proposal for a of the same form as in point b) but should also exploit the fact that the full conditional of λ is known explicitly.

How would you expect the optimal step size d and the rate of convergence of the resulting Markov chain to change when using the single-block sampler as compared to using separate Gibbs steps for a and λ? Explain why.

Problem 4

Let $(X_1, X_2, \ldots, X_{10})$ be an iid sample of size $n = 10$ from the exponential distribution with cumulative distribution function $F(x) = 1 - e^{-\lambda x}$. The MLE of λ is then given by

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Suppose we obtain an estimate $\hat{\lambda} = 2.0$ based on an observed sample $\mathbf{x} = (x_1, x_2, \ldots, x_{10})$

As a toy problem, suppose we want to examine the bias of $\hat{\lambda}$ by bootstrapping and to this end we simulate $B = 1000$ parametric bootstrap samples $\mathbf{x}^{1*}, \mathbf{x}^{2*}, \ldots, \mathbf{x}^{B*}$ from $F(\hat{\lambda})$. Let $\hat{\lambda}^{1*}, \hat{\lambda}^{2*}, \ldots, \hat{\lambda}^{B*}$ denote the corresponding bootstrap replicates of $\hat{\lambda}$ and suppose that the observed value of $\frac{1}{B}\sum_{b=1}^{B} \hat{\lambda}^{b*} = 2.24$.

a) Compute an estimate of the bias of $\hat{\lambda}$, that is, $E(\hat{\lambda}) - \lambda$, from the information given above. Given the same information, compute a bias corrected estimate of λ.

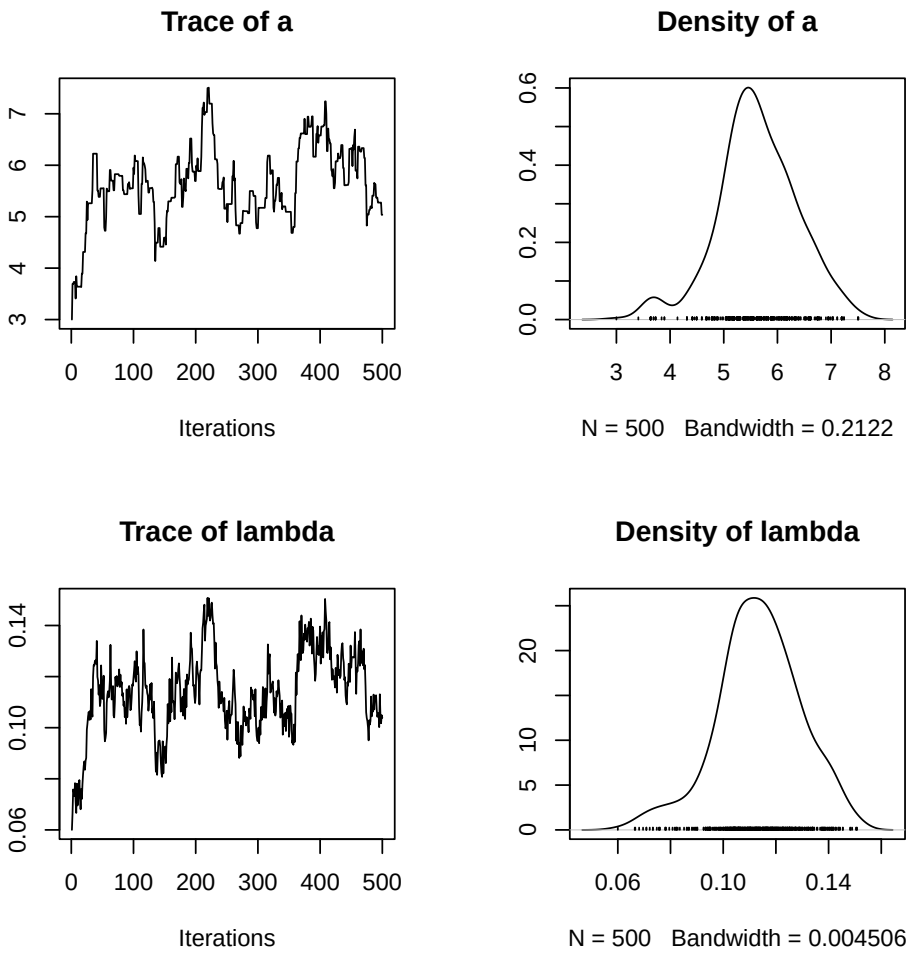

Figure 3: Trace plots and kernel density estimates of the marginal posteriors of a and λ when running the Gibbs sampler in problem 3a-b for 1000 iterations for the data shown in Fig. 2.

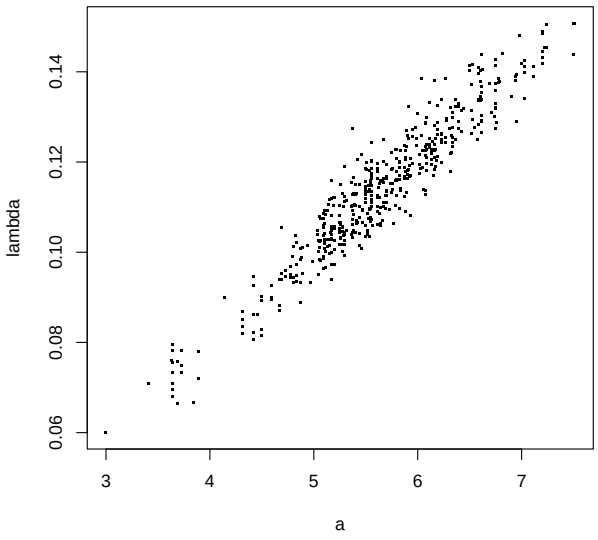

Figure 4: Samples from the joint posterior of a and λ based on the same Gibbs sampling run shown in Fig. 3.

Show that the bias corrected estimator $\hat{\lambda}_c$ can be expressed explicitly in terms of $\hat{\lambda}$ and the bootstrap replicates $\hat{\lambda}^{1*}, \hat{\lambda}^{2*}, \ldots, \hat{\lambda}^{B*}$ as

$$\hat{\lambda}_c = 2\hat{\lambda} - \frac{1}{B}\sum_{b=1}^{B} \hat{\lambda}^{b*}.$$

b) The expected value of $\hat{\lambda}$ is known analytically to be $E(\hat{\lambda}) = \frac{n}{n-1}\lambda$. Use this to find the conditional expectations $E(\hat{\lambda}^{b*} \mid \hat{\lambda})$ and $E(\hat{\lambda}_c \mid \hat{\lambda})$. Also, find $E(\hat{\lambda}_c)$. Hint: You may need to use the law of total expectation. Is the bias-corrected estimator $\hat{\lambda}_c$ unbiased? If not, by which percentage does $\hat{\lambda}_c$ over- or underestimate λ for $n = 10$? How does the bias of $\hat{\lambda}_c$ compare to that of $\hat{\lambda}$?

Problem 5

Suppose that we sample n individuals randomly from a population consisting of three genotypes AA, Aa and aa. Assuming that the population is in so called Hardy-Weinberg equilibrium, the proportions of the three genotypes in the population are p^2, $2p(1-p)$ and $(1-p)^2$, respectively, where $0 < p < 1$ is an unknown parameter. Assuming also that the population is large, the number of each genotype in the sample, Z_{AA}, Z_{Aa}, Z_{aa} will be approximately multinomially distributed.

a) Show that the log likelihood can be written as

$$\begin{aligned} l(p; Z_{AA}, Z_{Aa}, Z_{aa}) = \ln n! - \ln Z_{AA}! - \ln Z_{Aa}! - \ln Z_{aa}! + Z_{Aa} \ln 2 \\ + (2Z_{AA} + Z_{Aa}) \ln p + (Z_{Aa} + 2Z_{aa}) \ln(1-p) \end{aligned}$$

In the following suppose that we are unable to distinguish individuals of genotypes AA and Aa because allele A is dominant, that is, we only observe $X_{A-} = Z_{AA} + Z_{Aa}$ and $X_{aa} = Z_{aa}$. We would like to fit the model using the EM-algorithm. Let $p_{(t)}$ denote the value of the parameter p at the tth iteration of the algorithm.

b) Find

$$\begin{aligned} Z^*_{AA} &= E(Z_{AA}|X_{A-}, X_{aa}, p_{(t)}), \\ Z^*_{Aa} &= E(Z_{Aa}|X_{A-}, X_{aa}, p_{(t)}), \\ Z^*_{aa} &= E(Z_{aa}|X_{A-}, X_{aa}, p_{(t)}), \end{aligned}$$

that is, conditional expectations given the observed data and given that the parameter p is equal to $p_{(t)}$.

Next, express

$$Q(p|p_{(t)}) = E(l(p; Z_{AA}, Z_{Aa}, Z_{aa})|X_{A-}, X_{aa}, p_{(t)})$$

in terms of $Z^*_{AA}, Z^*_{Aa}, Z^*_{aa}$ up to a constant that does not depend of p.

Finally, find

$$p_{(t+1)} = \underset{p}{\operatorname{argmax}}\, Q(p|p_{(t)}).$$