ST4070

THE UNIVERSITY OF WARWICK

FOURTH YEAR EXAMINATION: April 2021

MONTE CARLO METHODS

Time allowed: **2 hours**

Full marks may be obtained by correctly answering THREE complete questions.

Candidates may attempt all FOUR questions. Credit will only be given for your THREE best answers.

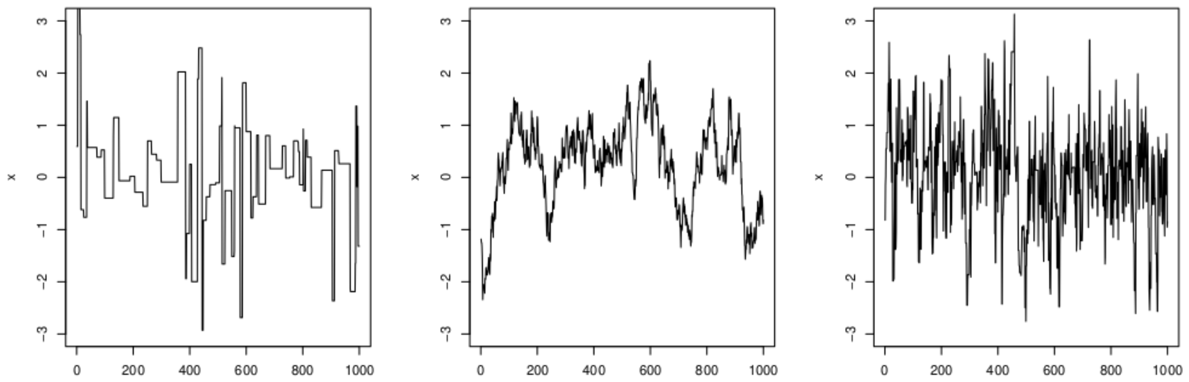
All questions carry an equal weight of 20 marks. There are a total of **60** marks available.

A guideline to the number of marks usually available is shown for each question section.

Calculators may be used in this examination.

Question 1

(a) The figures below show the sample paths of three random-walk Metropolis algorithms, sampling from the same target distribution. Comment qualitatively on the autocorrelation of the chains and explain for each figure the reasons behind the autocorrelation. Comment qualitatively on which chain has the largest effective sample size. [**4 marks**]



(b) Describe *one* method of diagnosing whether a Markov chain generated by an MCMC algorithm has reached its stationary regime. Comment critically upon it (e.g. by giving examples where the method fails). [**5 marks**]

(c) Assume $\text{Var}(h(\mathbf{X}^{(t)})) = \sigma^2$ and $\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho^\tau$, and consider a thinned chain $(\mathbf{Y}^{(t)})_t$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m\times t)}$, for $m \in \mathbb{N}_0$ and $\rho \in (0,1)$.

(i) Show that $\rho(h(\mathbf{Y}^{(t)}), h(\mathbf{Y}^{(t+\tau)})) = \rho^{m\tau}$. Interpret this result. [**3 marks**]

(ii) Argue that we can approximate

$$\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right) \approx \frac{\sigma^2}{T}\frac{1+\rho}{1-\rho},$$

and deduce that

$$\text{Var}\left(\frac{1}{\lfloor T/m \rfloor}\sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)})\right) \approx \frac{m\times\sigma^2}{T}\frac{1+\rho^m}{1-\rho^m}.$$

[**6 marks**]

(iii) Derive from the above an approximation for the ratio

$$\frac{\text{Var}\left(\frac{1}{\lfloor T/m \rfloor}\sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)})\right)}{\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right)}.$$

[**2 marks**]

Question 2

(a) Consider the Laplace(μ, b) distribution with probability density function $f(x|\mu, b)$ defined as

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right), \ x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $b > 0$.

(i) Calculate the corresponding cumulative distribution function $F(x)$. **[3 marks]**

(ii) Calculate the inverse $F^{-1}(x)$ of the cumulative distribution function $F(x)$. **[4 marks]**

(iii) Using the results above and assuming that you can sample efficiently from a uniform distribution $U[0,1]$ provide a simple algorithm (use pseudocode) to generate a sample from the Laplace distribution. **[3 marks]**

(b) Consider the following algorithm for sampling from a distribution with density f such that

$$h(x) \leq f(x) \leq Mg(x),$$

where $h(\cdot)$ is a function that is cheap to evaluate, $M > 0$, and $g(\cdot) > 0$ is the density of a distribution you can easily draw samples from.

1. Draw $X \sim g$.
2. Accept X as a sample from f with probability

$$\frac{h(X)}{Mg(X)}.$$

3. If X was not accepted in Step 2., accept X as a sample from f with probability

$$\frac{f(X) - h(X)}{Mg(X) - h(X)}.$$

(i) Show that the probability of accepting a proposed $X = x$ in either step 2. or 3. is

$$\frac{\int_{\mathcal{X}} f(x)dx}{M}.$$

[6 marks]

(ii) Deduce from this that the above algorithm indeed generates samples from f. **[4 marks]**

Question 3

Consider the following Bayesian model

$$\begin{aligned} Z_1 &\sim \mathsf{N}(\mu_1, \sigma_1^2), \\ Z_2 &\sim \mathsf{N}(\mu_2, \sigma_2^2), \\ Y &= Z_1 + Z_2 \end{aligned}$$

with μ_1 and μ_2 known, Z_1 and Z_2 are independent.[1] The prior distributions for the parameters σ_1^2 and σ_2^2 are:

$$\begin{aligned} \sigma_1^2 &\sim \mathsf{InverseGamma}(a_1, b_1) \\ \sigma_2^2 &\sim \mathsf{InverseGamma}(a_2, b_2). \end{aligned}$$

The density of an $\mathsf{InverseGamma}(a, b)$ distribution is $f^{\text{prior}}_{(a,b)}(t) = \frac{b^a}{\Gamma(a)}(1/t)^{a+1}\exp(-b/t)$ for $t > 0$, where $\Gamma(a) = \int_0^\infty t^{a-1}\exp(-t)\, dt$ is the Gamma function. Assume that Z_1 and Z_2 *cannot* be observed; it is only possible to observe $Y = Z_1 + Z_2$. The probability density function of Y is:

$$l_{(\sigma_1,\sigma_2)}(y) = \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

for $y \in \mathbb{R}$. (You do *not* need to show this.)

(a) Propose a Metropolis-Hastings algorithm to sample from the posterior distribution of $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ given the observation Y, using a proposal $q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})$ restricted on $(0, \infty) \times (0, \infty)$. State the probability of acceptance in terms of the above densities and q. (You do *not* need to bother about the analytical form of q, assume it to be known and provided to you.) **[4 marks]**

(b) One of your friends used a bivariate Normal distribution (with identity matrix as the covariance matrix) as the proposal rather than q. As the bivariate Normal distribution proposes values outside $(0, \infty) \times (0, \infty)$, the range of (σ_1^2, σ_2^2), he gets a lot of rejections. How can you improve your friend's Metropolis-Hastings algorithm by using reparametrization while keeping the same Normal proposal? **[5 marks]**
Hint: A variable $X > 0$ can be transformed to an unbounded variable Y by $Y = \log(X)$ and inverse by $X = \exp(Y)$. Using the change of variable theorem, given the density p_X of X, the density of Y is $p_Y(y) = p_X(\exp(y))\exp(y)$.

(c) Show that if we could observe Z_1, the posterior distribution of σ_1^2 would be

$$\sigma_1^2 | Z_1 = z_1 \sim \mathsf{InverseGamma}(a_1 + 0.5, b_1 + 0.5 \times (z_1 - \mu_1)^2).$$

[3 marks]

(d) Propose a Gibbs sampler to sample from the posterior distributions of σ_1 and σ_2 given the observation Y that does *not* make use of the density of Y. **[5 marks]**
Hint: You might have to introduce an auxiliary variable.

(e) Compare the two algorithms proposed in (b) and (d) by discussing their advantages and disadvantages. **[3 marks]**

[1] Recall the probability density function for a $\mathsf{N}(\mu, \sigma^2)$ random variable is $f(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Question 4

(a) (i) State the importance sampling algorithm for sampling from a distribution with density f using an instrumental distribution with density g (where $\text{supp}(g) \supset \text{supp}(f \times h)$). How would you estimate an expectation $\mathbb{E}_f\left(h(X)\right)$? [**4 marks**]

(ii) Verify that the (not self-normalised) importance sampling estimate of $\mathbb{E}_f\left(h(X)\right)$ is unbiased and consistent. [**4 marks**]

(b) (i) Compute the variance of the importance sampling (not self-normalised) estimate of $\mathbb{E}_f(h)$ with a sampling density $g(x)$. [**4 marks**]

(ii) Suppose that $f(x)$ is the $\mathcal{N}(0, 1)$ distribution, and that $h(x) = \exp\left(-\frac{(x-10)^2}{2}\right)$. Find the optimal importance sampling density $g(x)$ to estimate $\mathbb{E}_f(h)$. [**4 marks**]

(c) Consider a Bayesian model for observed data $\mathbf{y}$ having a single parameter β. Denote by $l(\mathbf{y}|\beta)$ the likelihood and by $f_{\text{prior}}(\beta)$ the (proper) prior distribution of β. The posterior is then

$$f(\beta) = C \times l(\mathbf{z}|\beta) \times f_{\text{prior}}(\beta).$$

If the normalisation constant C is very expensive to compute how would you estimate the mean of the posterior distribution $f(\beta)$ using the prior distribution $f_{\text{prior}}(\beta)$ as the sampling distribution using an importance sampling algorithm. [**4 marks**]

End.

ST4070

THE UNIVERSITY OF WARWICK

FOURTH YEAR EXAMINATION: April 2021

MONTE CARLO METHODS

Time allowed: **2 hours**

Full marks may be obtained by correctly answering THREE complete questions.

Candidates may attempt all FOUR questions. Credit will only be given for your THREE best answers.

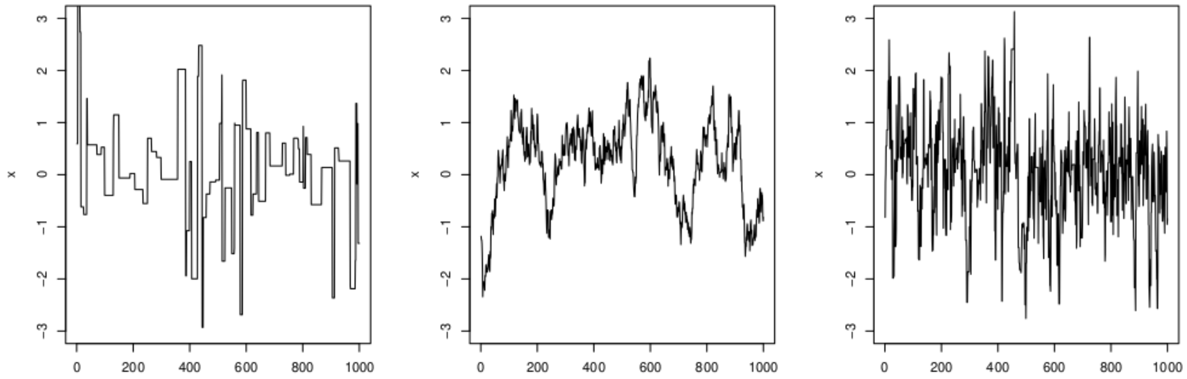
All questions carry an equal weight of 20 marks. There are a total of **60** marks available.

A guideline to the number of marks usually available is shown for each question section.

Calculators may be used in this examination.

Question 1

(a) The figures below show the sample paths of three random-walk Metropolis algorithms, sampling from the same target distribution. Comment qualitatively on the autocorrelation of the chains and explain for each figure the reasons behind the autocorrelation. Comment qualitatively on which chain has the largest effective sample size. [**4 marks**]



(a) (unseen) The first two sample paths show a high autocorrelation. In the left figure this is due to a very low probability of acceptance; in the middle figure this is due to a very small variance of the proposal density. As the right figure appears to have the lowest autocorrelation, it corresponds to the chain with the largest effective sample size.

(b) Describe *one* method of diagnosing whether a Markov chain generated by an MCMC algorithm has reached its stationary regime. Comment critically upon it (e.g. by giving examples where the method fails). [**5 marks**]

(b) (concepts and some examples were presented in the lectures) *Various techniques were presented in the lecture course. Students are expected to produce an answer similar to the following:*
A non-parametric test can be used to diagnose whether the chain has reached its stationary regime. One possibility is splitting the chain into three parts: $(\mathbf{X}^{(t)})_{t=1,\ldots,\lfloor T/3\rfloor}$ (burn in), $(\mathbf{X}^{(t)})_{t=\lfloor T/3\rfloor+1,\ldots,2\lfloor T/3\rfloor}$ (“block 1”), and $(\mathbf{X}^{(t)})_{t=2\lfloor T/3\rfloor+1,\ldots,T}$ (“block 2”). If the Markov chain has reached its stationary regime after $\lfloor T/3\rfloor$ iterations, then the last two blocks should be from the same distribution, which can be tested informally using a Kolmogorov-Smirnov test, whose statistic is

$$K = \sup_{x\in\mathbb{R}} \left| \hat{F}_{(\mathbf{X}^{(t)})_{t=\lfloor T/(3m)\rfloor+1,\ldots,2\lfloor T/(3m)\rfloor}}(x) - \hat{F}_{(\mathbf{X}^{(t)})_{t=2\lfloor T/(3m)\rfloor+1,\ldots,\lfloor T/m\rfloor}}(x) \right|.$$

As the Kolmogorov-Smirnov test is designed for i.i.d. data, it is typically applied to a thinned chain.
The Kolmogorov-Smirnov test suffers from the “you’ve only seen where you’ve been” problem. If the Markov chain fails to explore the entire support of the target distribution (e.g. mixture of two well-separated distributions), it will be impossible to tell this from the KS test.

(c) Assume $\mathrm{Var}(h(\mathbf{X}^{(t)})) = \sigma^2$ and $\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho^\tau$, and consider a thinned chain $(\mathbf{Y}^{(t)})_t$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m\times t)}$, for $m \in \mathbb{N}_0$ and $\rho \in (0, 1)$.

(i) Show that $\rho(h(\mathbf{Y}^{(t)}), h(\mathbf{Y}^{(t+\tau)})) = \rho^{m\tau}$. Interpret this result. **[3 marks]**

(ii) Argue that we can approximate

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right) \approx \frac{\sigma^2}{T}\frac{1+\rho}{1-\rho},$$

and deduce that

$$\mathrm{Var}\left(\frac{1}{\lfloor T/m \rfloor}\sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)})\right) \approx \frac{m \times \sigma^2}{T}\frac{1+\rho^m}{1-\rho^m}.$$

[6 marks]

(iii) Derive from the above an approximation for the ratio

$$\frac{\mathrm{Var}\left(\frac{1}{\lfloor T/m \rfloor}\sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)})\right)}{\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right)}.$$

[2 marks]

(c) This is the answer to (c).

(i) (Seen in lecture) $\rho(h(\mathbf{Y}^{(t)}), h(\mathbf{Y}^{(t+\tau)})) = \rho(h(\mathbf{X}^{(m\times t)}), h(\mathbf{X}^{(m\times t+m\times\tau)}))) = (\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})))^m = \rho^{m\tau}$. Hence, we can notice thinning reduces the autocorrelation.

(ii) (Seen similar) We assume that $(h(\mathbf{X}^{(t)}))_{t=1,\dots,T}$ is a first-order autoregressive time series (AR(1)), i.e. $\mathrm{Var}(h(\mathbf{X}^{(t)})) = \sigma^2$, and $\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho(\tau) = \rho^\tau$. Then

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right) &= \frac{1}{T^2}\left(\sum_{t=1}^{T}\underbrace{\mathrm{Var}(h(\mathbf{X}^{(t)}))}_{=\sigma^2} + 2\sum_{1\le s<t\le T}\underbrace{\mathrm{Cov}(h(\mathbf{X}^{(s)}), h(\mathbf{X}^{(t)}))}_{=\sigma^2\times\rho(t-s)}\right)\\ &= \frac{\sigma^2}{T^2}\left(T + 2\sum_{\tau=1}^{T-1}(T-\tau)\rho(\tau)\right) = \frac{\sigma^2}{T}\left(1 + 2\sum_{\tau=1}^{T-1}\left(1-\frac{\tau}{T}\right)\rho(\tau)\right).\end{aligned}$$

As $\sum_{\tau=1}^{+\infty}|\rho^\tau| < +\infty$, we obtain from the dominated convergence theorem

$$T\times\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right) \longrightarrow \sigma^2\left(1+2\sum_{\tau=1}^{+\infty}\rho(\tau)\right) = \sigma^2\left(1+2\sum_{\tau=1}^{+\infty}\rho^\tau\right) = \sigma^2(1+\rho)/(1-\rho).$$

The result for $(h(\mathbf{Y}^{(t)}))_{t=1,\dots,\lfloor T/m \rfloor}$ can be obtained analogously.

(iii) (unseen) $\dfrac{\mathrm{Var}\left(\frac{1}{\lfloor T/m \rfloor}\sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)})\right)}{\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\mathbf{X}^{(t)})\right)} \approx m \times \dfrac{(1+\rho^m)(1-\rho)}{(1-\rho^m)(1+\rho)}.$

Question 2

(a) Consider the Laplace(μ, b) distribution with probability density function $f(x|\mu, b)$ defined as

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right), \ x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $b > 0$.

(i) Calculate the corresponding cumulative distribution function $F(x)$. **[3 marks]**

(ii) Calculate the inverse $F^{-1}(x)$ of the cumulative distribution function $F(x)$. **[4 marks]**

(iii) Using the results above and assuming that you can sample efficiently from a uniform distribution $U[0, 1]$ provide a simple algorithm (use pseudocode) to generate a sample from the Laplace distribution. **[3 marks]**

(a) This is the solution to (a). (Seen similar)

(i) As we can write the probability density function $f(x|\mu, b), \ x \in \mathbb{R}, \ \mu \in \mathbb{R}, \ b > 0$ as

$$\begin{aligned} f(x|\mu, b) &= \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \\ & \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu-x}{b}\right), & x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right), & x \geq \mu, \end{cases} \end{aligned}$$

the cumulative distribution function would be

$$\begin{aligned} F(x) = \int_{-\infty}^{x} f(u|\mu, b) du &= \begin{cases} \frac{1}{2} \exp\left(\frac{x-\mu}{b}\right), & x < \mu \\ 1 - \frac{1}{2} \exp\left(-\frac{x-\mu}{b}\right), & x \geq \mu, \end{cases} \\ &= \frac{1}{2} + \frac{1}{2} sgn(x-\mu) \left(1 - \exp\left(-\frac{|x-\mu|}{b}\right)\right). \end{aligned}$$

(ii) The inverse of cumulative distribution function would be,

$$F^{-1}(p) = \mu - b \times sgn(p - 0.5) \log\left(1 - 2|p - 0.5|\right).$$

(iii) Comments are only needed when the pseudo-code is not self explanatory.

```
#Sample u from a Uniform distribution U[0,1]
u = runif(1);
#Apply the CDF inversion method
x = mu - b * sign(u-0.5) * log(1- 2*abs(u-0.5));
#x is distributed according to F
```

(b) Consider the following algorithm for sampling from a distribution with density f such that

$$h(x) \leq f(x) \leq Mg(x),$$

where $h(\cdot)$ is a function that is cheap to evaluate, $M > 0$, and $g(\cdot) > 0$ is the density of a distribution you can easily draw samples from.

1. Draw $X \sim g$.
2. Accept X as a sample from f with probability

$$\frac{h(X)}{Mg(X)}.$$

3. If X was not accepted in Step 2., accept X as a sample from f with probability

$$\frac{f(X)-h(X)}{Mg(X)-h(X)}.$$

(i) Show that the probability of accepting a proposed $X = x$ in either step 2. or 3. is

$$\frac{\int_{\mathcal{X}} f(x)dx}{M}.$$

[6 marks]

(ii) Deduce from this that the above algorithm indeed generates samples from f. **[4 marks]**

(b) This is the solution to (b) (Unseen)

(i) We have that

$$\begin{aligned}\mathbb{P}(X \in \mathcal{X} \text{ and is accepted}) &= \mathbb{P}(X \in \mathcal{X} \text{ and is accepted in step 2})\\ &\quad +\mathbb{P}(X \in \mathcal{X} \text{ and is rejected in step 2,}\\ &\quad \text{but accepted in step 3)}\end{aligned}$$

In complete analogy with the above that we have that

$$\begin{aligned}\mathbb{P}(X \in \mathcal{X} \text{ and is accepted in step 2}) &= \int_{\mathcal{X}} g(x) \qquad \underbrace{\frac{h(x)}{M \times g(x)}}_{=\mathbb{P}(X \text{ is accepted in step } 2|X=x)} \qquad dx\\ &= \frac{\int_{\mathcal{X}} h(x)\, dx}{M},\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(X \in \mathcal{X} \text{ and is rejected in step 2, but accepted in step 3})\\ = &\int_{\mathcal{X}} g(x) \qquad \underbrace{1-\frac{h(x)}{M \times g(x)}}_{=\mathbb{P}(X \text{ is rejected in step } 2|X=x)} \qquad \underbrace{\frac{f(x)-h(x)}{M \times g(x)-h(x)}}_{=\mathbb{P}(X \text{ is accepted in step } 3|X=x)} \qquad dx\\ = &\frac{\int_{\mathcal{X}} f(x)-h(x)\, dx}{M}.\end{aligned}$$

Thus

$$\mathbb{P}(X \in \mathcal{X} \text{ and is accepted}) = \frac{\int_{\mathcal{X}} h(x)\, dx}{M} + \frac{\int_{\mathcal{X}} f(x)-h(x)\, dx}{M} = \frac{\int_{\mathcal{X}} f(x)\, dx}{M}$$

(ii) We have that $\mathbb{P}(X \text{ is accepted}) = \frac{\int_E f(x)\, dx}{M} = \frac{1}{M}$, thus

$$\begin{aligned}\mathbb{P}(x \in \mathcal{X}|X \text{ is accepted}) = \frac{\mathbb{P}(X \in \mathcal{X} \text{ and is accepted})}{\mathbb{P}(X \text{ is accepted})} &= \frac{\int_{\mathcal{X}} f(x)\, dx/M}{1/M}\\ &= \int_{\mathcal{X}} f(x)\, dx.\end{aligned}$$

Thus the density of the values accepted by the algorithm is $f(\cdot)$.

Question 3

Consider the following Bayesian model

$$\begin{aligned} Z_1 &\sim \mathsf{N}(\mu_1, \sigma_1^2), \\ Z_2 &\sim \mathsf{N}(\mu_2, \sigma_2^2), \\ Y &= Z_1 + Z_2 \end{aligned}$$

with μ_1 and μ_2 known, Z_1 and Z_2 are independent.[1] The prior distributions for the parameters σ_1^2 and σ_2^2 are:

$$\begin{aligned} \sigma_1^2 &\sim \mathsf{InverseGamma}(a_1, b_1) \\ \sigma_2^2 &\sim \mathsf{InverseGamma}(a_2, b_2). \end{aligned}$$

The density of an $\mathsf{InverseGamma}(a, b)$ distribution is $f_{(a,b)}^{\text{prior}}(t) = \frac{b^a}{\Gamma(a)}(1/t)^{a+1}\exp(-b/t)$ for $t > 0$, where $\Gamma(a) = \int_0^\infty t^{a-1}\exp(-t)\,dt$ is the Gamma function. Assume that Z_1 and Z_2 *cannot* be observed; it is only possible to observe $Y = Z_1 + Z_2$. The probability density function of Y is:

$$l_{(\sigma_1,\sigma_2)}(y) = \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

for $y \in \mathbb{R}$. (You do *not* need to show this.)

(a) Propose a Metropolis-Hastings algorithm to sample from the posterior distribution of $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ given the observation Y, using a proposal $q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})$ restricted on $(0, \infty) \times (0, \infty)$. State the probability of acceptance in terms of the above densities and q. (You do *not* need to bother about the analytical form of q, assume it to be known and provided to you.) [**4 marks**]

(a) This is the solution to (a). (Seen similar)
A random-walk Metropolis algorithm would be:
Starting with $\boldsymbol{\sigma}^{(0)} := (\sigma_1^{(0)}, \sigma_2^{(0)})$, iterate for $t = 1, 2, \ldots$

1. Draw $\boldsymbol{\sigma}^* \sim q(\cdot|\boldsymbol{\sigma}^{(t-1)})$ and set $\boldsymbol{\sigma} = \boldsymbol{\sigma}*$.
2. Compute
$$\alpha(\boldsymbol{\sigma}|\boldsymbol{\sigma}^{(t-1)}) = \min\left\{1, \frac{f(\boldsymbol{\sigma})q(\boldsymbol{\sigma}^{(t-1)}|\boldsymbol{\sigma})}{f(\boldsymbol{\sigma}^{(t-1)})q(\boldsymbol{\sigma}|\boldsymbol{\sigma}^{(t-1)})}\right\},$$
where $f(\boldsymbol{\sigma}) := f_{(a_1,b_1)}^{\text{prior}}(\sigma_1) \times f_{(a_2,b_2)}^{\text{prior}}(\sigma_2) \times l_{(\sigma_1,\sigma_2)}(y)$.
3. With probability $\alpha(\boldsymbol{\sigma}|\boldsymbol{\sigma}^{(t-1)})$ set $\boldsymbol{\sigma}^{(t)} = \boldsymbol{\sigma}$, otherwise set $\boldsymbol{\sigma}^{(t)} = \boldsymbol{\sigma}^{(t-1)}$.

(b) One of your friends used a bivariate Normal distribution (with identity matrix as the covariance matrix) as the proposal rather than q. As the bivariate Normal distribution proposes values outside $(0, \infty) \times (0, \infty)$, the range of (σ_1^2, σ_2^2), he gets a lot of rejections. How can you improve your friend's Metropolis-Hastings algorithm by using reparametrization while keeping the same Normal proposal? [**5 marks**]
Hint: A variable $X > 0$ can be transformed to an unbounded variable Y by $Y = \log(X)$ and inverse by $X = \exp(Y)$. Using the change of variable theorem, given the density p_X of X, the density of Y is $p_Y(y) = p_X(\exp(y))\exp(y)$.

(b) This is the solution to (b). (Unseen)
Using reparametrization and $q(\cdot|\boldsymbol{\nu}) = \mathcal{N}(\cdot|\boldsymbol{\nu}, \mathbf{I})$, a random-walk Metropolis algorithm would be:
Starting with $\boldsymbol{\nu}^{(0)} := (\log(\sigma_1^{(0)})^2, \log(\sigma_2^{(0)})^2)$, iterate for $t = 1, 2, \ldots$

[1] Recall the probability density function for a $\mathsf{N}(\mu, \sigma^2)$ random variable is $f(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

1. Draw $\boldsymbol{\nu}^* \sim q(\cdot|\boldsymbol{\nu}^{(t-1)})$ and set $\boldsymbol{\nu} = \boldsymbol{\nu}*$.
2. $\boldsymbol{\sigma} = (\exp(\nu_1), \exp(\nu_2))$ when $\boldsymbol{\nu} = (\nu_1, \nu_2)$; $\boldsymbol{\sigma}^{(t-1)} = (\exp(\nu_1^{(t-1)}), \exp(\nu_2^{(t-1)}))$ when $\boldsymbol{\nu}^{(t-1)} = (\nu_1^{(t-1)}, \nu_2^{(t-1)})$
3. Compute
$$\alpha(\boldsymbol{\nu}|\boldsymbol{\nu}^{(t-1)}) = \min\left\{1, \frac{f(\boldsymbol{\sigma})q(\boldsymbol{\nu}^{(t-1)}|\boldsymbol{\nu})}{f(\boldsymbol{\sigma}^{(t-1)})q(\boldsymbol{\nu}|\boldsymbol{\nu}^{(t-1)})}\right\},$$
where $f(\boldsymbol{\sigma}) := f^{\text{prior}}_{(a_1,b_1)}(\sigma_1) \times f^{\text{prior}}_{(a_2,b_2)}(\sigma_2) \times l_{(\sigma_1,\sigma_2)}(y)\sigma_1\sigma_2$.
4. With probability $\alpha(\boldsymbol{\nu}|\boldsymbol{\nu}^{(t-1)})$ set $\boldsymbol{\sigma}^{(t)} = \boldsymbol{\sigma}$, otherwise set $\boldsymbol{\sigma}^{(t)} = \boldsymbol{\sigma}^{(t-1)}$.
5. $\boldsymbol{\nu}^{(t)} = (\log(\sigma_1^{(t)})^2, \log(\sigma_2^{(t)})^2)$

(c) Show that if we could observe Z_1, the posterior distribution of σ_1^2 would be

$$\sigma_1^2|Z_1 = z_1 \sim \mathsf{InverseGamma}(a_1 + 0.5, b_1 + 0.5 \times (z_1 - \mu_1)^2).$$

[3 marks]

(c) This is the solution to (c). (Unseen)
$f(\sigma_1^2|z_1) \propto \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(z_1-\mu_1)^2}{2\sigma_1^2}\right) \times (1/\sigma_1^2)^{a_1+1} \exp(-b_1/\sigma_1^2) \propto (1/\sigma_1^2)^{a_1+0.5+1} \exp\left(-\frac{0.5 \times (z_1-\mu_1)^2 + b_1}{\sigma_1^2}\right)$, so $\vartheta_1|Z_1 = z_1 \sim \mathsf{InverseGamma}(a_1 + 0.5, b_1 + 0.5 \times (z_1 - \mu_1)^2)$.

(d) Propose a Gibbs sampler to sample from the posterior distributions of σ_1 and σ_2 given the observation Y that does *not* make use of the density of Y. **[5 marks]**
Hint: You might have to introduce an auxiliary variable.

(d) This is the solution to (d). (Unseen).
If we could observe Z_1, then we could work out $Z_2 = Y - Z_1$, and we could use the posterior distributions from part (c). This suggests introducing Z_1 as a hidden observation into a Gibbs sampler of the form

1. Sample from $Z_1|\sigma_1, \sigma_2$.
2. Sample from $\sigma_1|Z_1$. (σ_1 is given Z_1 conditionally independent of σ_2)
3. Sample from $\sigma_2|Z_2$. (σ_2 is given Z_2 conditionally independent of σ_1)

The distributions of steps 2. and 3. are the InverseGamma distributions derived in part (c) (using $Z_2 = Y - Z_1$).
$Z_1|\sigma_1, \sigma_2$ is

$$\begin{aligned}
f(z_1|\vartheta_1, \vartheta_2, y) &\propto f(z_1, y, \vartheta_1, \vartheta_2) \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(z_1-\mu_1)^2}{2\sigma_1^2}\right) \times (1/\sigma_1^2)^{a_1+1} \exp(-b_1/\sigma_1^2) \\
&\quad \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{((y-z_1)-\mu_2)^2}{2\sigma_2^2}\right) \times (1/\sigma_2^2)^{a_2+1} \exp(-b_2/\sigma_2^2) \\
&\propto \exp\left(-\left[z_1^2\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right) + 2z_1\left(\frac{\mu_2 - y}{2\sigma_2^2} - \frac{\mu_1}{2\sigma_1^2}\right)\right]\right)
\end{aligned}$$

Hence, $f(z_1|\vartheta_1, \vartheta_2, y)$ is $\mathcal{N}(z_1|\frac{\sigma_1^2(\mu_2-y)-\mu_1\sigma_2^2}{\sigma_2^2+\sigma_1^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_2^2+\sigma_1^2})$

(e) Compare the two algorithms proposed in (b) and (d) by discussing their advantages and disadvantages. [**3 marks**]

(e) This is the solution to (e). (Unseen).
$(Z_1, \sigma_1, \sigma_2)$ are highly correlated, so the Gibbs sampler will be moving rather slowly. On the other hand, the steps in the Gibbs sampler are simple and fast to carry out. The Metropolis-Hastings algorithm should yield a less correlated sample, however at the expense of every step requiring a certain amount of computation (though one can avoid having to sample from a InverseGamma distribution that way). In addition, it is not all that easy to come up with a suitable proposal as we have seen.

Question 4

(a) (i) State the importance sampling algorithm for sampling from a distribution with density f using an instrumental distribution with density g (where $\mathrm{supp}(g) \supset \mathrm{supp}(f \times h)$). How would you estimate an expectation $\mathbb{E}_f(h(X))$? [**4 marks**]

(ii) Verify that the (not self-normalised) importance sampling estimate of $\mathbb{E}_f(h(X))$ is unbiased and consistent. [**4 marks**]

(a) This is the solution to (a)

(i) (bookwork) For g such that $\mathrm{supp}(g) \supset \mathrm{supp}(f \times h)$ sample $X_i \sim g$, and set $w(X_i) = \frac{f(X_i)}{g(X_i)}$. $\mathbb{E}_f(h(X))$ can then be estimated by either $\tilde{\mu} = \frac{\sum_{i=1}^n w(W_i)h(X_i)}{n}$ or $\hat{\mu} = \frac{\sum_{i=1}^n w(X_i)h(X_i)}{\sum_{i=1}^n w(X_i)}$.

(ii) (seen in the lectures) We have $\mathbb{E}_g(w(X_1) \times h(X_1)) = \int g(x) \underbrace{\frac{f(x)}{g(x)}}_{=w(x)} h(x)\ dx = \int f(x)h(x)\quad dx = \mathbb{E}_f(h(X_1))$, and thus $\mathbb{E}_g\left(\frac{1}{n}\sum_{i=1}^n w(X_i) \times h(X_i)\right) = \mathbb{E}_g\left(w(X_1) \times h(X_1)\right) = \mathbb{E}_f\left(h(X_1)\right)$.

Provided $\mathbb{E}_g|w(X_1) \times h(X_1)| < +\infty$ we have from the law of large numbers that

$$\tilde{\mu} = \frac{\sum_{i=1}^n w(X_i)h(X_i)}{n} \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_g(w(X_1) \times h(X_1)) = \mathbb{E}_f(h(X_1)),$$

thus $\tilde{\mu}$ is consistent.

(b) (i) Compute the variance of the importance sampling (not self-normalised) estimate of $\mathbb{E}_f(h)$ with a sampling density $g(x)$. [**4 marks**]

(ii) Suppose that $f(x)$ is the $\mathcal{N}(0,1)$ distribution, and that $h(x) = \exp\left(-\frac{(x-10)^2}{2}\right)$. Find the optimal importance sampling density $g(x)$ to estimate $\mathbb{E}_f(h)$. [**4 marks**]

(b) This is the solution to (b). (Seen similar)

(i)

$$\begin{aligned}
Var_g\left(\frac{\sum_{i=1}^n w(X_i)h(X_i)}{n}\right) &= \frac{Var_g(w(X)h(X))}{n} \\
&= \frac{1}{n}Var_g\left(\frac{h(X)f(X))}{g(X)}\right) \\
&= \frac{1}{n}\left[\mathbb{E}_g\left(\left(\frac{h(X)f(X))}{g(X)}\right)^2\right) + \left(\mathbb{E}_g\left(\frac{h(X)f(X))}{g(X)}\right)\right)^2\right] \\
&= \frac{1}{n}\left[\mathbb{E}_g\left(\left(\frac{h(X)f(X))}{g(X)}\right)^2\right) + \left(\mathbb{E}_f(h(X))\right)^2\right]
\end{aligned}$$

(ii) Using (i) an optimal g would minimize $\mathbb{E}_g\left(\left(\frac{h(X)f(X))}{g(X)}\right)^2\right)$. Jensen's inequality gives us,

$$\mathbb{E}_g\left(\left(\frac{h(X)f(X))}{g(X)}\right)^2\right) \geq \left(\mathbb{E}_g\left(\frac{|h(X)|f(X))}{g(X)}\right)\right)^2 = \left(\int_E |h(x)|f(x)dx\right)^2$$

i.e. the estimator obtained by using an importance sampler employing instrumental distribution

$$g^*(x) = \frac{|h(x)|f(x)}{\int_E |h(t)|f(t)dt}$$

attains the minimal possible variance amongst the class of importance sampling estimators. As we have $h(x) = \exp\left(-\frac{(x-10)^2}{2}\right)$ and $f(x) \propto \exp\left(-\frac{x^2}{2}\right)$, the optimal sampling density would be $\mathcal{N}(10, 1/\sqrt{2})$.

(c) Consider a Bayesian model for observed data $\mathbf{y}$ having a single parameter β. Denote by $l(\mathbf{y}|\beta)$ the likelihood and by $f_{\text{prior}}(\beta)$ the (proper) prior distribution of β. The posterior is then

$$f(\beta) = C \times l(\mathbf{z}|\beta) \times f_{\text{prior}}(\beta).$$

If the normalisation constant C is very expensive to compute how would you estimate the mean of the posterior distribution $f(\beta)$ using the prior distribution $f_{\text{prior}}(\beta)$ as the sampling distribution using an importance sampling algorithm. [**4 marks**]

(c) This is the solution to (c) (Seem similar in lecture)
We will use self-normalised importance sampling and the following estimate

$$f(\beta) = \frac{\sum_{i=1}^n l(y|\beta_i)\beta_i}{\sum_{i=1}^n l(y|\beta_i)}$$

where β_i $\forall i = 1, \dots n$ are i.i.d. samples from $f_{\text{prior}}(\beta)$. In this case the normalisation constant C does not need to be evaluated when using self-normalised weights, as it cancels out.

End.