IMPERIAL

MATH70047

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2024

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Stochastic Simulation

Date: Wednesday, May 8, 2024
Time: 10:00 – 12:00 (BST)
Time Allowed: 2 hours
This paper has 4 Questions.
Please Answer All Questions in 1 Answer Booklet
Candidates should start their solutions to each question on a new sheet of paper.
Supplementary books may only be used after the relevant main book(s) are full.
Any required additional material(s) will be provided.
Allow margins for marking.
Credit will be given for all questions attempted.
Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

1. (a) Below F_X and p_X denote the cumulative density function (CDF) and probability density function (PDF) of a continuous random variable (r.v.) X, respectively.

   (i) Let $U \sim \text{Unif}(0,1)$ and $Y = F_X^{-1}(U)$. **Using only the transformation of random variables formula**, prove that $p_Y = p_X$. (2 marks)

   (ii) Let X be a Weibull(k, 1) random variable with PDF $p_X(x) = kx^{k-1}\exp(-x^k)$ where $x \in [0, \infty)$. Show that $F_X(x) = 1 - \exp(-x^k)$, where $x \in \mathbb{R}_+$. Derive F_X^{-1}. (2 marks)

   (iii) Describe the inversion sampler for a Weibull(k, 1) variable using Part(ii). (1 mark)

(b) Let $U_1 \sim \text{Unif}(0,1)$ and $U_2 \sim \text{Unif}(-\pi, \pi)$. Consider the following transformation

$$X_1 = \cos(U_2)\sqrt{-2\log(U_1)} \quad \text{and} \quad X_2 = \sin(U_2)\sqrt{-2\log(U_1)}.$$

Show that X_1 and X_2 are independent and standard Normal random variables.
Hint: $\frac{\mathrm{d}\arctan(x)}{\mathrm{d}x} = 1/(1+x^2)$. (3 marks)

(c) Suppose we would like to perform rejection sampling for a target density $p(x)$ using a proposal density $q(x)$.

   (i) Provide the pseudocode of the rejection sampler under the assumption that $p(x)$ can be evaluated. Give the form of **optimal** M in terms of $p(x)$ and $q(x)$ so that $p(x) \leq Mq(x)$. (2 marks)

   (ii) Show that the acceptance rate of this rejection sampler is $\hat{a} = 1/M$. (3 marks)

   (iii) Suppose we can only evaluate an unnormalised version of the target density $\bar{p}(x) \propto p(x)$. Show that the acceptance rate for the resulting rejection sampler is given by $\hat{a} = Z/M$ where Z is the normalising constant of $p(x)$. (2 marks)

   (iv) Consider the following **unnormalised** density function

$$\bar{p}(x) = \mathbf{1}_{\{x \leq a\}}(x)\exp(-\mu x), \quad x \in [0, \infty),$$

   where μ and a are positive known constants. Consider a rejection sampler with a proposal $q(x) = \mu\exp(-\mu x)$, $x \in [0, \infty)$. Choose $M = 1$ and provide the acceptance rate in terms of a and μ. (2 marks)

(d) Consider the following Erlang density function

$$p(x) = \frac{x^{k-1}\exp(-\lambda x)}{(k-1)!} \quad \text{for} \quad x \in [0, \infty),$$

where $k \in \mathbb{N}_+$ and $\lambda > 0$ are parameters of the density. Consider a rejection sampler for this target density with a proposal $q(x) = \mu\exp(-\mu x)$, $x \in [0, \infty)$ with $\mu < \lambda$. Find $\mu^\star$ in terms of λ and k that maximises the acceptance rate of the rejection sampler. (3 marks)
Hint: Treat $\log(k-1)!$ terms as constants w.r.t. x and μ as necessary.

(Total: 20 marks)

2. Consider the following Rayleigh density function:

$$p(x) = x \exp\left(-\frac{x^2}{2}\right), \quad x \in [0, \infty).$$

Suppose we would like to estimate the **mean** of this distribution.

(a) Consider the importance sampling (IS) task for this problem. We would like to use another Rayleigh density function $q_\lambda(x)$ as the proposal:

$$q_\lambda(x) = \frac{x}{\lambda^2} \exp\left(-\frac{x^2}{2\lambda^2}\right), \quad x \in [0, \infty), \quad \lambda > 0.$$

(i) State the test function $\varphi(x)$ for this problem. Show that $\int \varphi(x) p(x) \mathrm{d}x = \sqrt{\pi/2}$.
Hint: Gaussian second moment identity: $\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = 1$. (2 marks)

(ii) State the IS estimator $\hat{\varphi}_{\mathsf{IS}}^N$ in terms of N samples $X_1, \ldots, X_N \sim q_\lambda(x)$. (1 mark)

(iii) State the variance expression $\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N)$ in terms of p, φ, q_λ, and N (You do not have to derive it). (2 marks)

(iv) By explicitly computing the quantity you provided in Part(a)(iii), show that

$$\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N) = \frac{1}{N}\left(2\lambda^6 (2\lambda^2 - 1)^{-2} - \frac{\pi}{2}\right).$$

Hint: You can use integration by parts (or any other technique of your choice). In integration by parts, recognise that the second term can be completed to an integral of a Rayleigh density. (5 marks)

(v) Show that the optimal $\lambda_\star$ that minimises the variance is $\lambda_\star = \sqrt{3/2}$. (2 marks)

(vi) Derive the expression for the variance $\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N)$ with this optimal choice $\lambda_\star$. (1 mark)

(b) Now consider perfect Monte Carlo (MC) sampling for the same problem with the same test function used in Part(a).

(i) Provide the MC estimator $\hat{\varphi}_{\mathsf{MC}}^N$ in terms of N samples $X_1, \ldots, X_N \sim p(x)$. (1 mark)

(ii) State the variance expression $\mathsf{var}(\hat{\varphi}_{\mathsf{MC}}^N)$ in terms of p, φ, and N (You do not have to derive it). (1 mark)

(iii) By explicitly computing the quantity you provided in Part(b)(ii), show that

$$\mathsf{var}(\hat{\varphi}_{\mathsf{MC}}^N) = \frac{1}{N}\left(2 - \frac{\pi}{2}\right).$$

Hint: The computation here is a special case of Part(a)(iv), so you can directly compute the integral using your earlier result. (3 marks)

(iv) Compare $\mathsf{var}(\hat{\varphi}_{\mathsf{MC}}^N)$ and $\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N)$ derived in Part(a)(vi) and discuss why we would perform IS rather than MC even if we have access to i.i.d samples. (2 marks)

(Total: 20 marks)

3. (a) Consider a generic Bayesian model with a prior $p(x)$ and a likelihood $p(y|x)$, where x is the parameter of interest and y is the data.

(i) Provide the expression of the posterior $p(x|y)$ in terms of $p(x)$ and $p(y|x)$. (1 mark)

(ii) Suppose we would like to sample from $p(x|y)$ using Metropolis-Hastings (MH) with a general proposal $q(x'|x)$. Assume that the normalising constant $p(y)$ is not computable. Provide the pseudocode of this algorithm. (2 marks)

(iii) Assume that the proposal is symmetric: $q(x'|x) = q(x|x')$. Provide the simplified MH algorithm with a pseudocode. (1 mark)

(iv) Provide the Metropolis-adjusted Langevin algorithm (MALA) for this problem. (2 marks)

(v) Provide the unadjusted Langevin algorithm (ULA) for this problem. (2 marks)

(vi) Suppose our likelihood has the form $p(y_{1:n}|x) = \prod_{i=1}^{n} p(y_i|x)$, where n is very large. Suggest a modification of ULA that allows approximate sampling of $p(x|y_{1:n})$ and which does not use the full data at each iteration. (2 marks)

(b) In this part, we will consider online inference in Bayesian models.

(i) Consider a single data-point model with a fixed data y:

$$p(x) = \text{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x \in [0, \infty),$$
$$p(y|x) = \text{Poisson}(y; x) = \frac{x^y}{y!} \exp(-x),$$

Show that $p(x|y)$ is a Gamma density with parameters $\alpha + y$ and $\beta + 1$. (2 marks)

(ii) Assume now that we have a sequence of data points $y_{1:n} = (y_1, \dots, y_n)$ where y_i is the i-th data point, which are *conditionally independent* given x. For any prior $p(x)$ and likelihoods $\{p(y_i|x)\}_{i=1}^{n}$ (without using the model in Part (i)), **prove** that the update equation for the posterior $p(x|y_{1:n})$ in terms of the previous posterior $p(x|y_{1:n-1})$ takes the form

$$p(x|y_{1:n}) = \frac{p(y_n|x) p(x|y_{1:n-1})}{p(y_n|y_{1:n-1})}.$$

(4 marks)

(iii) Suppose we have the following prior and likelihoods:

$$p(x) = \text{Gamma}(x; \alpha_0, \beta_0) \quad \text{and} \quad p(y_i|x) = \text{Poisson}(y_i; x) \quad \text{for} \quad i = 1, \dots, n.$$

Using the sequential Bayesian update of Part(b)(ii), derive a recursion for the posterior sufficient statistics, i.e. give a relationship for each of α_n, β_n in terms of $\alpha_{n-1}, \beta_{n-1}$, where $p(x|y_{1:n}) = \text{Gamma}(x; \alpha_n, \beta_n)$. (4 marks)

(Total: 20 marks)

4. This question is based on the mastery material, A. Doucet and A. M. Johansen. "A tutorial on particle filtering and smoothing: Fifteen years later." *Handbook of nonlinear filtering* (2009).

(a) Consider a state space model:

$$x_0 \sim \mu(x_0), \quad \text{and} \quad x_t|x_{t-1} \sim f(x_t|x_{t-1}), \quad \text{and} \quad y_t|x_t \sim g(y_t|x_t),$$

where $t = 1, \ldots, T$.

(i) Provide the joint unnormalised density of the state and observation sequence, i.e., $\bar{\pi}(x_{0:t}, y_{1:t})$ in terms of μ, f, and g. (2 marks)

(ii) Using this unnormalised density, prove that

$$\pi_t(x_{0:t}|y_{1:t}) = \pi_{t-1}(x_{0:t-1}|y_{1:t-1}) \frac{g(y_t|x_t) f(x_t|x_{t-1})}{p(y_t|y_{1:t-1})}.$$

(2 marks)

(iii) Provide the pseudocode of the path space self-normalised importance sampler (SNIS) using a proposal $q(x_{0:t})$ and the unnormalised target $\bar{\pi}(x_{0:t}, y_{1:t})$. (2 marks)

(iv) Assume that your proposal has a Markov structure $q(x_{0:t}) = q(x_0) \prod_{k=1}^{t} q(x_k|x_{k-1})$. Derive a sequential recursion for unnormalised importance weights. (2 marks)

(v) Using the expression in Part(a)(iv), derive the sequential importance sampling (SIS) algorithm and provide the pseudocode. (3 marks)

(b) In this part, we consider the variance behaviour of the SNIS schemes on the path space.

(i) Let $\pi(x)$ and $q(x)$ be the target and proposal densities and let $\pi(x) = \gamma(x)/Z$ where $\gamma(x)$ is the unnormalised density. Provide the SNIS estimator $\widehat{Z}^N$ of Z. (1 mark)

(ii) Show that the *relative* variance of this SNIS estimator is given by

$$\frac{\mathsf{var}(\widehat{Z}^N)}{Z^2} = \frac{1}{N} \left(\int \frac{\pi^2(x)}{q(x)} \mathrm{d}x - 1 \right).$$

(2 marks)

(iii) Consider now distributions on the path space:

$$\pi(x_{0:t}) = \prod_{k=0}^{t} \mathcal{N}(x; 0, 1) \quad \text{and} \quad q(x_{0:t}) = \prod_{k=0}^{t} \mathcal{N}(x_k; 0, \sigma^2),$$

where $\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Let Z_t be the normalising constant of $\pi(x_{0:t})$. Adapting the formula in Part(b)(ii), prove that the relative variance of the SNIS estimator of the normalising constant Z_t is given by

$$\frac{\mathsf{var}(\widehat{Z}_t^N)}{Z_t^2} = \frac{1}{N} \left[\left(\frac{\sigma^4}{2\sigma^2 - 1} \right)^{t/2} - 1 \right].$$

Provide the necessary constraint on σ^2. (5 marks)

(iv) What implication does this result have for the SIS algorithm you derived in Part(a)(v)? Suggest a modification (in words) that would remedy the issue. (1 mark)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2024

This paper is also taken for the relevant examination for the Associateship.

MATH60047/70047

Stochastic Simulation (Solutions)

Setter's signature	Checker's signature	Editor's signature
....	....	....

1. (a) (i) For a generic transformation $Y = g(U)$, 1D transformation of random variables formula is given by $p_Y(y) = p_U(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|$. In this particular case, $g(U) = F_X^{-1}(U)$, therefore $g^{-1}(y) = F_X(y)$. It needs to be recognised that $p_U(F_X(y)) = 1$, since $F(\cdot)$ maps to $[0,1]$ and that the derivative of $g^{-1}(y) = F_X(y)$ is $p_X(y)$, which implies that $p_Y(y) = p_X(y)$. sim. seen ⇓ 2, A

(ii) Consider a change of variable $u = y^k$. With this, we obtain

$$F_X(x) = \int_0^x k y^{k-1} \exp(-y^k) dy = \int_0^{x^k} \exp(-u) du = 1 - \exp(-x^k).$$

Therefore, $F_X^{-1}(u) = (-\log(1-u))^{1/k}$. 2, B

(iii) Using the result from part (ii), we obtain the sampler

$$U_i \sim \mathsf{Unif}(0,1) \qquad X_i = (-\log(1-U_i))^{1/k}.$$

1, A

(b) We first need to compute inverse transformation $(u_1, u_2) = g^{-1}(x_1, x_2) = (g_1^{-1}(x_1,x_2), g_2^{-1}(x_1,x_2))$. For $g_1^{-1}(x_1,x_2)$, we have meth seen ⇓

$$X_1^2 + X_2^2 = -2\log U_1 \implies g_1^{-1}(x_1,x_2) = \exp\left(\frac{-x_1^2 - x_2^2}{2}\right).$$

Similarly, for $g_2^{-1}(x_1,x_2)$, we have

$$\frac{X_2}{X_1} = \tan(U_2) \implies g_2^{-1}(x_1,x_2) = \arctan\left(\frac{x_2}{x_1}\right).$$

We can then write using transformation of random variables formula that

$$p_{x_1,x_2}(x_1,x_2) = p_{u_1,u_2}(g_1^{-1}(x_1,x_2), g_2^{-1}(x_1,x_2)) |\det J_{g^{-1}}(x_1,x_2)|$$

where p_{u_1,u_2} is a 2D uniform distribution on $[0,1] \times [-\pi,\pi]$ where u_1 and u_2 are independent. The Jacobian of g^{-1} is given by

$$\det J_{g^{-1}}(x_1,x_2) = \begin{vmatrix} \frac{\partial g_1^{-1}}{\partial x_1} & \frac{\partial g_1^{-1}}{\partial x_2} \\ \frac{\partial g_2^{-1}}{\partial x_1} & \frac{\partial g_2^{-1}}{\partial x_2} \end{vmatrix} = \begin{vmatrix} -x_1 \exp\left(\frac{-x_1^2-x_2^2}{2}\right) & -x_2 \exp\left(\frac{-x_1^2-x_2^2}{2}\right) \\ -\frac{x_2}{x_1^2}\frac{1}{1+(x_2/x_1)^2} & \frac{1}{x_1}\frac{1}{1+(x_2/x_1)^2} \end{vmatrix}$$

$$= -\exp\left(\frac{-x_1^2-x_2^2}{2}\right)\frac{1}{1+(x_2/x_1)^2} - \exp\left(\frac{-x_1^2-x_2^2}{2}\right)\frac{x_2^2}{x_1^2}\frac{1}{1+(x_2/x_1)^2}$$

$$= -\exp\left(\frac{-x_1^2-x_2^2}{2}\right).$$

Plugging this into the transformations formula (be mindful of the absolute value and the facts that $p_{u_1}(\cdot) = 1$ and $p_{u_2}(\cdot) = 1/2\pi$), we obtain

$$p_{x_1,x_2}(x_1,x_2) = \frac{1}{2\pi}\exp\left(\frac{-x_1^2-x_2^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x_1^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x_2^2}{2}\right) = p_{x_1}(x_1)p_{x_2}(x_2).$$

3, D

(c) (i) The rejection sampler pseudocode is given as seen ⇓

- Sample $X' \sim q(x)$.
- Sample $U \sim \mathsf{Unif}(0,1)$.
- If $U \leq \frac{p(X')}{Mq(X')}$, then accept X', otherwise reject.
- Repeat.

1, A

The optimal M is given by $M = \sup_{x \in \mathsf{X}} \frac{p(x)}{q(x)}$. $\boxed{1, \mathrm{A}}$

(ii) It can be seen that, conditioned on $X' = x'$, the acceptance probability is

$$\mathbb{P}(\text{accept}|X' = x') = \mathbb{P}\left(U \leq \frac{p(x')}{Mq(x')}\right) = \int_0^{\frac{p(x')}{Mq(x')}} 1 \mathrm{d}u = \frac{p(x')}{Mq(x')}.$$

Therefore, the unconditional acceptance probability is given by

$$\mathbb{P}(\text{accept}) = \int_{\mathsf{X}} \mathbb{P}(\text{accept}|X' = x')q(x')\mathrm{d}x' = \int_{\mathsf{X}} \frac{p(x')}{Mq(x')} q(x')\mathrm{d}x' = \frac{1}{M}\int_{\mathsf{X}} p(x')\mathrm{d}x' = \frac{1}{M}.$$

$\boxed{3, \mathrm{B}}$

(iii) The same argument holds exactly when $\bar{p}(x)$ is available where $p(x) = \frac{1}{Z}\bar{p}(x)$, where $Z = \int \bar{p}(x)\mathrm{d}x$ is the normalising constant. In other words,

$$\mathbb{P}(\text{accept}) = \int_{\mathsf{X}} \mathbb{P}(\text{accept}|X' = x')q(x')\mathrm{d}x' = \int_{\mathsf{X}} \frac{\bar{p}(x')}{Mq(x')} q(x')\mathrm{d}x' = \frac{1}{M}\int_{\mathsf{X}} \bar{p}(x')\mathrm{d}x' = \frac{Z}{M}.$$

$\boxed{2, \mathrm{B}}$

(iv) This is a truncated sampling method. Since $M = 1$, the acceptance rate is $Z = \int \bar{p}(x)\mathrm{d}x = \int \mathbf{1}_{x \leq a}(x)\exp(-\mu x)\mathrm{d}x = \int_0^a \exp(-\mu x)\mathrm{d}x = \frac{1}{\mu}(1 - \exp(-\mu a))$.

$\boxed{2, \mathrm{B}}$

(d) In order to find the optimal proposal $q_{\mu^\star}$, we first need to find M_μ. We have $\boxed{\text{meth seen} \Downarrow}$

$$M_\mu = \sup_{x \in [0,\infty)} \frac{p(x)}{q_\mu(x)}$$

where

$$R_\mu(x) = \frac{p(x)}{q_\mu(x)} = \frac{\frac{x^{k-1}\exp(-\lambda x)}{(k-1)!}}{\mu \exp(-\mu x)} = \frac{1}{(k-1)!}\frac{x^{k-1}}{\mu}\exp\left(-(\lambda - \mu)x\right).$$

We would like to maximise $R_\mu(x)$ and solve this optimisation problem by first computing $\log R_\mu(x)$

$$\log R_\mu(x) = (k-1)\log(x) - (\lambda - \mu)x + C$$

where C is a constant that does not depend on x. Computing the derivative and setting it to zero

$$\frac{\mathrm{d}}{\mathrm{d}x}\log R_\mu(x) = \frac{k-1}{x} - (\lambda - \mu) = 0 \implies x^\star = \frac{k-1}{\lambda - \mu}.$$

We can then see $M_\mu = R_\mu(x^\star)$, which is given by

$$M_\mu = \frac{1}{(k-1)!}\frac{(k-1)^{k-1}}{(\lambda-\mu)^{k-1}}\frac{1}{\mu}\exp\left(-(\lambda-\mu)\frac{k-1}{\lambda-\mu}\right) = \frac{(k-1)^{k-1}}{(k-1)!}\frac{\exp(1-k)}{\mu(\lambda-\mu)^{k-1}}.$$

We would like to now minimise M_μ with respect to μ (as this would then maximise the acceptance rate $1/M_\mu$). Taking $\log M_\mu$, differentiating with respect to μ, and setting it to zero, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\log M_\mu = -\frac{1}{\mu} + \frac{k-1}{\lambda - \mu} = 0 \implies \mu^\star = \frac{\lambda}{k}.$$

$\boxed{3, \mathrm{C}}$

2. (a) (i) The test function is $\varphi(x) = x$ as **mean** is to be estimated. Then $\boxed{\text{meth seen} \Downarrow}$

$$\int \varphi(x)p(x)\mathrm{d}x = \int_0^\infty x^2 \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = \frac{1}{2}\int_{-\infty}^\infty x^2 \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = \sqrt{\pi/2},$$

as $x^2\exp(-x^2/2)$ is an even function and given the identity in the hint. $\boxed{2, \mathrm{A}}$

(ii) Given samples $X_1, \ldots, X_N \sim q_\lambda(x)$, the basic IS estimator is given as (for $\varphi(x) = x$):

$$\hat{\varphi}_{\mathsf{IS}}^N = \frac{1}{N}\sum_{i=1}^N \frac{p(X_i)}{q_\lambda(X_i)}X_i.$$

$\boxed{1, \mathrm{A}}$

(iii) The variance of the IS estimator is given by (for a generic φ)

$$\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N) = \frac{1}{N}\left(\mathbb{E}_{q_\lambda}\left[\frac{p^2(X)}{q_\lambda^2(X)}\varphi(X)^2\right] - \bar{\varphi}^2\right).$$

where $\bar{\varphi} = \int \varphi(x)p(x)\mathrm{d}x$. Alternatively, students might use $w_\lambda(x) = p(x)/q_\lambda(x)$ or can just give the expression for $\varphi(x) = x$. $\boxed{2, \mathrm{A}}$

(iv) Among the terms in the variance expression, we know $\bar{\varphi} = \sqrt{\pi/2}$. Therefore, we are interested in computing the expectation (by putting $\varphi(x) = x$)

$$\mathbb{E}_{q_\lambda}\left[\frac{p^2(X)}{q_\lambda^2(X)}X^2\right] = \int_0^\infty \frac{x^4\exp(-x^2)}{\frac{x}{\lambda^2}\exp\left(-\frac{x^2}{2\lambda^2}\right)}\mathrm{d}x = \lambda^2\int_0^\infty x^3\exp\left(-\frac{x^2}{2}\left(\frac{2\lambda^2-1}{\lambda^2}\right)\right)\mathrm{d}x.$$

Write $\sigma^2 = \lambda^2/(2\lambda^2 - 1)$. Then, we have

$$\mathbb{E}_{q_\lambda}\left[\frac{p^2(X)}{q_\lambda^2(X)}X^2\right] = \lambda^2\int_0^\infty x^3\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x.$$

At this point, we perform integration by parts by setting $u = x^2$ and $\mathrm{d}v = x\exp(-x^2/2\sigma^2)\mathrm{d}x$. Then, $\mathrm{d}u = 2x\mathrm{d}x$ and $v = -\sigma^2\exp(-x^2/2\sigma^2)$. Therefore,

$$\begin{aligned}\int_0^\infty x^3\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x &= \int_0^\infty u\mathrm{d}v = [uv]_0^\infty - \int_0^\infty v\mathrm{d}u \\ &= \underbrace{\left[-\sigma^2x^2\exp\left(-\frac{x^2}{2\sigma^2}\right)\right]_0^\infty}_{=0} + 2\sigma^4\underbrace{\int_0^\infty \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x}_{=1 \text{ as int. of Rayleigh PDF}} = 2\sigma^4.\end{aligned}$$

Therefore,

$$\mathbb{E}_{q_\lambda}\left[\frac{p^2(X)}{q_\lambda^2(X)}X^2\right] = 2\lambda^2\sigma^4 = 2\lambda^6(2\lambda^2-1)^{-2}.$$

Plugging this into the expression, we get

$$\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N) = \frac{1}{N}\left(2\lambda^6(2\lambda^2-1)^{-2} - \frac{\pi}{2}\right).$$

$\boxed{5, \mathrm{D}}$

(v) We compute optimal λ by minimising the expression

$$\lambda^\star = \arg\min_\lambda \mathbb{E}_{q_\lambda}\left[\frac{p^2(X)}{q_\lambda^2(X)}X^2\right] = \arg\min_\lambda 2\lambda^6(2\lambda^2-1)^{-2}.$$

Taking the log, then its derivative and setting it to zero, we obtain

$$\frac{6}{\lambda} - \frac{8\lambda}{2\lambda^2-1} = 0 \implies \lambda^\star = \sqrt{3/2},$$

since $\lambda > 0$. $\boxed{2, \mathrm{C}}$

(vi) Plugging $\lambda^\star$ into the variance expression, we obtain

$$\mathsf{var}(\hat{\varphi}_{\mathsf{IS}}^N) = \frac{1}{N}\left(2\lambda_\star^6(2\lambda_\star^2-1)^{-2} - \frac{\pi}{2}\right) = \frac{1}{N}\left(\frac{27}{16} - \frac{\pi}{2}\right).$$

$\boxed{1, \mathrm{B}}$

(b) (i) Given $X_1, \ldots, X_N \sim p(x)$, the MC estimator is given as $\boxed{\text{seen} \Downarrow}$

$$\hat{\varphi}_{\mathsf{MC}}^N = \frac{1}{N}\sum_{i=1}^N X_i.$$

$\boxed{1, \mathrm{A}}$

(ii) The variance of the MC estimator (for any φ) is given by $\mathsf{var}(\hat{\varphi}_{\mathsf{MC}}^N) = \frac{1}{N}\mathsf{var}(\varphi)$ where $\mathsf{var}(\varphi) = \mathbb{E}[\varphi(X)^2] - \mathbb{E}[\varphi(X)]^2$ and $X \sim p$. $\boxed{1, \mathrm{A}}$

(iii) To compute this variance for $\varphi(x) = x$, we need to compute $\mathsf{var}(X)$. We have $\boxed{\text{meth seen} \Downarrow}$

$$\mathsf{var}_p(X) = \int x^2 p(x)\mathrm{d}x - \left(\int xp(x)\mathrm{d}x\right)^2.$$

The second term above can be computed using Part(a)(i), the first term is given by

$$\int x^2 p(x)\mathrm{d}x = \int_0^\infty x^3 \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x.$$

This integral can be computed by integration by parts. Indeed, it needs to be recognised that using the same integration by parts as in Part(a)(iv), we have $\int_0^\infty x^3\exp(-x^2/2)\mathrm{d}x = 2$. Therefore, $\mathsf{var}_p(X) = 2 - (\sqrt{\pi/2})^2 = 2 - \pi/2$. $\boxed{3, \mathrm{B}}$

(iv) This result shows that the variance of the MC estimator is larger than the variance of the IS estimator even though we have access to i.i.d samples. It is therefore preferable in certain risk sensitive situations to use IS estimators to reduce variance. $\boxed{2, \mathrm{B}}$

3. (a) (i) We have $\boxed{\text{seen} \Downarrow}$

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)\mathrm{d}x}.$$

$\boxed{1, \text{A}}$

(ii) We have a generic proposal $q(x'|x)$. Furthermore, $p(y)$ is uncomputable, therefore, we have to consider the unnormalised distribution $p(x, y)$ instead. We can then write the MH sampler initialised at $X_0 = x_0$ as follows. For $t = 1, \dots, T$:

- Sample $X' \sim q(x'|X_{t-1})$.
- Sample $U \sim$ Unif$(0, 1)$.
- If

$$U \leq \frac{p(X', y)q(X_{t-1}|X')}{p(X_{t-1}, y)q(X'|X_{t-1})}$$

then accept $X_t = X'$, otherwise reject and set $X_t = X_{t-1}$.

Alternatively students can use $p(x, y) = p(y|x)p(x)$ in the acceptance ratio. $\boxed{2, \text{A}}$

(iii) We have a symmetric proposal $q(x'|x) = q(x|x')$. We can then write the symmetric MH sampler initialised at $X_0 = x_0$ as follows. For $t = 1, \dots, T$:

- Sample $X' \sim q(x'|X_{t-1})$.
- Sample $U \sim$ Unif$(0, 1)$.
- If

$$U \leq \frac{p(X', y)}{p(X_{t-1}, y)}$$

then accept $X_t = X'$, otherwise reject and set $X_t = X_{t-1}$.

Alternatively students can use $p(x, y) = p(y|x)p(x)$ in the acceptance ratio. $\boxed{1, \text{A}}$

(iv) The MALA algorithm uses a gradient-based proposal. We notice that

$$\nabla_x \log p(x|y) = \nabla_x \log p(y|x) + \nabla_x \log p(x).$$

as $p(y)$ drops from the gradient. Therefore, the MALA proposal is given as

$$q(x'|x) = \mathcal{N}(x'; x + \gamma \nabla_x \log p(y|x) + \gamma \nabla_x \log p(x), 2\gamma I_d).$$

or any equivalent formulation (e.g. $\nabla_x \log p(x, y)$). Then we have

- Sample $X' \sim \mathcal{N}(x'; x + \gamma \nabla_x \log p(y|x) + \gamma \nabla_x \log p(x), 2\gamma I_d)$.
- Sample $U \sim$ Unif$(0, 1)$.
- If

$$U \leq \frac{p(X', y)q(X_{t-1}|X')}{p(X_{t-1}, y)q(X'|X_{t-1})}$$

then accept $X_t = X'$, otherwise reject and set $X_t = X_{t-1}$. $\boxed{2, \text{A}}$

(v) ULA method just uses MALA proposal and skips the accept/reject step, therefore can be compactly written as

$$X_t = X_{t-1} + \gamma \nabla_x \log p(y|X_{t-1}) + \gamma \nabla_x \log p(X_{t-1}) + \sqrt{2\gamma} W_t$$

where $W_t \sim \mathcal{N}(0, I_d)$. $\boxed{2, \text{A}}$

(vi) For this case, we have the algorithm as

$$X_t = X_{t-1} + \gamma \left(\nabla_x \log p(X_{t-1}) + \sum_{k=1}^{n} \nabla_x \log p(y_k|X_{t-1}) \right) + \sqrt{2\gamma} W_t,$$

where $W_t \sim \mathcal{N}(0, I_d)$. This is impractical as n is large. We can instead use subsampling to estimate the second term of the gradient with a subset of the data. We sample uniformly $i_1, \ldots, i_K$ from $\{1, \ldots, n\}$ and then use the following approximation

$$\sum_{k=1}^{n} \nabla_x \log p(y_k|X_{t-1}) \approx \frac{n}{K} \sum_{k=1}^{K} \nabla_x \log p(y_{i_k}|X_{t-1}).$$

Then we obtain the following method which is called stochastic gradient Langevin dynamics (SGLD):

$$X_t = X_{t-1} + \gamma \left(\nabla_x \log p(X_{t-1}) + \frac{n}{K} \sum_{k=1}^{K} \nabla_x \log p(y_{i_k}|X_{t-1}) \right) + \sqrt{2\gamma} W_t,$$

where $W_t \sim \mathcal{N}(0, I_d)$. Note that indices i_k are different for every t, but clear statement of stochastic gradients Langevin dynamics for one iteration is enough here. $\boxed{\text{2, A}}$

(b) (i) We can see that the unnormalised posterior is given by $\boxed{\text{seen} \Downarrow}$

$$p(x|y) \propto x^{y+\alpha-1} \exp(-(\beta+1)x),$$

by dropping all terms that do not depend on x. This can be recognised as another Gamma distribution with parameters $y + \alpha$ and $\beta + 1$. $\boxed{\text{2, A}}$

(ii) We can prove this recursion as follows

$$p(x|y_{1:n}) = \frac{p(x, y_{1:n})}{p(y_{1:n})} = \frac{p(y_n|x, y_{1:n-1})p(x, y_{1:n-1})}{p(y_{1:n})}$$

At this point, we use two key observations. (1) $p(y_n|x, y_{1:n-1}) = p(y_n|x)$ as y_n is conditionally independent of $y_{1:n-1}$ given x, (2) the chain rule implies that $p(y_{1:n}) = p(y_n|y_{1:n-1})p(y_{1:n-1})$. Therefore, we have

$$p(x|y_{1:n}) = \frac{p(y_n|x)}{p(y_n|y_{1:n-1})} \frac{p(x, y_{1:n-1})}{p(y_{1:n-1})} = \frac{p(y_n|x)p(x|y_{1:n-1})}{p(y_n|y_{1:n-1})}.$$

$\boxed{\text{4, D}}$

(iii) We know that (from Part(b)(i)) for fixed data, the posterior is given by $\boxed{\text{meth seen} \Downarrow}$

$$p(x|y) = \text{Gamma}(x; y + \alpha, \beta + 1).$$

where α, β are the prior parameters. We then observe that for the prior we have

$$p(x|y_1) = \text{Gamma}(x; \alpha_1, \beta_1) = \text{Gamma}(x; y_1 + \alpha_0, \beta_0 + 1).$$

Assume as an induction hypothesis that

$$p(x|y_{1:n-1}) = \mathsf{Gamma}(x; \sum_{i=1}^{n-1} y_i + \alpha_0, \beta_0 + n - 1) = \mathsf{Gamma}(x; \alpha_{n-1}, \beta_{n-1}).$$

By Part(b)(i) and Part(b)(ii), we know that

$$p(x|y_{1:n}) = \mathsf{Gamma}(x; \alpha_n, \beta_n),$$

where

$$\alpha_n = \alpha_{n-1} + y_n \quad \beta_n = \beta_{n-1} + 1.$$

$\boxed{\text{4, C}}$

4. (a) (i) The joint unnormalised density is given as $\boxed{\text{seen} \Downarrow}$

$$\bar{\pi}(x_{0:t}, y_{1:t}) = \mu(x_0) \prod_{k=1}^{t} f(x_k|x_{k-1}) g(y_k|x_k).$$

$\boxed{2, \text{M}}$

(ii) We can write

$$\pi_t(x_{0:t}|y_{1:t}) = \frac{\bar{\pi}(x_{0:t}, y_{1:t})}{p(y_{1:t})} = \frac{f(x_t|x_{t-1}) g(y_t|x_t) \bar{\pi}_{t-1}(x_{0:t-1}, y_{1:t-1})}{p(y_{1:t})}.$$

Using chain rule $p(y_{1:t}) = p(y_t|y_{1:t-1}) p(y_{1:t-1})$, we obtain

$$\pi_t(x_{0:t}|y_{1:t}) = \frac{f(x_t|x_{t-1}) g(y_t|x_t) \bar{\pi}_{t-1}(x_{0:t-1}, y_{1:t-1})}{p(y_t|y_{1:t-1}) p(y_{1:t-1})} = \frac{f(x_t|x_{t-1}) g(y_t|x_t) \pi_{t-1}(x_{0:t-1}|y_{1:t-1})}{p(y_t|y_{1:t-1})},$$

which is the desired result. $\boxed{2, \text{M}}$

(iii) Using a proposal $q(x_{0:t})$ and the unnormalised distribution $\bar{\pi}(x_{0:t}, y_{1:t})$, we can define an importance sampler

- Sample $X_{0:t}^{(i)} \sim q(x_{0:t})$ for $i = 1, \ldots, N$.
- Compute the importance weights

$$W_{0:t}^{(i)} = \frac{\bar{\pi}(X_{0:t}^{(i)}, y_{1:t})}{q(X_{0:t}^{(i)})}$$

- Report an approximation

$$\pi_t^N(x_{0:t}|y_{1:t}) \mathrm{d}x_{0:t} = \sum_{i=1}^{N} W_{0:t}^{(i)} \delta_{X_{0:t}^{(i)}}(\mathrm{d}x_{0:t}).$$

The obvious shortcoming of this sampler is that, whenever we get new data points $y_{t+1}, \ldots, y_T$, we need to recompute the importance weights in an increasing space dimension. Another problem (also acceptable as an answer) as this dimensionality increases, then the weights will be hard to compute properly. $\boxed{2, \text{M}}$

(iv) Consider the weight expression above as a function:

$$W_{0:t}(x_{0:t}) = \frac{\bar{\pi}(x_{0:t}, y_{1:t})}{q(x_{0:t})}.$$

If the proposal is Markov, we can suitably decompose

$$W_{0:t}(x_{0:t}) = \frac{f(x_t|x_{t-1}) g(y_t|x_t)}{q(x_t|x_{t-1})} \frac{\bar{\pi}(x_{0:t-1}, y_{1:t-1})}{q(x_{0:t-1})} = \frac{f(x_t|x_{t-1}) g(y_t|x_t)}{q(x_t|x_{t-1})} W_{0:t-1}(x_{0:t-1}).$$

Therefore, the whole computation can be done sequentially. Given $W_0^{(i)}$ for all i, we can compute the weights recursively as

$$W_{0:t}^{(i)} = \frac{f(x_t^{(i)}|x_{t-1}^{(i)}) g(y_t|x_t^{(i)})}{q(x_t^{(i)}|x_{t-1}^{(i)})} W_{0:t-1}^{(i)}.$$

$\boxed{2, \text{M}}$

(v) The SIS sampler is then given as

- Sample $X_t^{(i)} \sim q(x_t | X_{t-1}^{(i)})$ for $i = 1, \dots, N$.
- Compute the importance weights

$$W_{0:t}^{(i)} = \frac{f(x_t^{(i)} | x_{t-1}^{(i)}) g(y_t | x_t^{(i)})}{q(x_t^{(i)} | x_{t-1}^{(i)})} W_{0:t-1}^{(i)}.$$

- Report an approximation

$$\pi_t^N(x_{0:t} | y_{1:t}) \mathrm{d}x_{0:t} = \sum_{i=1}^N W_{0:t}^{(i)} \delta_{X_{0:t}^{(i)}}(\mathrm{d}x_{0:t}).$$

$\boxed{3, \mathrm{M}}$

(b) (i) Given an unnormalised target $\gamma(x)$ and a proposal $q(x)$, the estimator of the normalising constant is given by $\boxed{\text{unseen} \Downarrow}$

$$\widehat{Z}^N = \frac{1}{N} \sum_{i=1}^N \frac{\gamma(X^{(i)})}{q(X^{(i)})},$$

where $X^{(i)} \sim q(x)$. $\boxed{1, \mathrm{M}}$

(ii) We can straightforwardly compute the variance as

$$\mathsf{var}(\widehat{Z}^N) = \frac{1}{N^2} \sum_{i=1}^N \mathsf{var}\left(\frac{\gamma(X^{(i)})}{q(X^{(i)})}\right) = \frac{1}{N^2} \sum_{i=1}^N \left(\mathbb{E}_q\left[\frac{\gamma^2(X^{(i)})}{q^2(X^{(i)})}\right] - \mathbb{E}_q\left[\frac{\gamma(X^{(i)})}{q(X^{(i)})}\right]^2\right),$$

$$= \frac{1}{N}\left(\int \frac{\gamma^2(x)}{q(x)} \mathrm{d}x - \left(\int \frac{\gamma(x)}{q(x)} q(x) \mathrm{d}x\right)^2\right) = \frac{1}{N}\left(Z^2 \int \frac{\pi^2(x)}{q(x)} \mathrm{d}x - Z^2\right).$$

By dividing this expression by Z^2, we obtain the result. $\boxed{2, \mathrm{M}}$

(iii) We have

$$\gamma(x_{0:t}) = \prod_{k=1}^t \exp\left(-\frac{x_k^2}{2}\right) \qquad \text{and} \qquad Z_t = (2\pi)^{t/2}.$$

The formula in (ii) should be used here, by just replacing $\pi(x)$ with $\pi(x_{0:t})$ and $q(x)$ with $q(x_{0:t})$. When done so, the relative variance is given by

$$\frac{\mathsf{var}(\widehat{Z}_t^N)}{Z_t^2} = \frac{1}{N}\left(\int \frac{\pi^2(x_{0:t})}{q(x_{0:t})} \mathrm{d}x_{0:t} - 1\right) = \frac{1}{N}\left[\prod_{k=1}^t \int \frac{(1/2\pi)\exp(-x_k^2)}{(1/\sqrt{2\pi\sigma^2})\exp(-x_k/2\sigma^2)} \mathrm{d}x_k - 1\right]$$

The integral can be computed by recognising its Gaussian form

$$\int \frac{(1/2\pi)\exp(-x_k^2)}{(1/\sqrt{2\pi\sigma^2})\exp(-x_k/2\sigma^2)} \mathrm{d}x_k = \sigma \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_k^2}{2}\left(2 - \frac{1}{\sigma^2}\right)\right) \mathrm{d}x_k.$$

By putting $1/\lambda^2 = \left(2 - 1/\sigma^2\right)$,

$$\sigma \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_k^2}{2}\left(2 - \frac{1}{\sigma^2}\right)\right) \mathrm{d}x_k = \sigma \int \lambda \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{x_k^2}{2\lambda^2}\right) \mathrm{d}x_k = \sigma\lambda = \left(\frac{\sigma^4}{2\sigma^2 - 1}\right)^{1/2}.$$

With this we conclude $\boxed{4, \text{M}}$

$$\frac{\mathsf{var}(\widehat{Z}_t^N)}{Z_t^2} = \frac{1}{N}\left[\prod_{k=1}^{t}\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{1/2} - 1\right] = \frac{1}{N}\left[\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{t/2} - 1\right].$$

A necessary constraint here for variance to be positive and finite is $\sigma^2 > 1/2$.

$\boxed{1, \text{M}}$

(iv) This means that as $t \to \infty$, SIS variance increases exponentially even in simple cases. Resampling resolves this issue (enough for points). In particular, resampling results in an asymptotic relative variance that is only linear in t. $\boxed{1, \text{M}}$

Review of mark distribution:
Total A marks: 24 of 24 marks
Total B marks: 15 of 15 marks
Total C marks: 9 of 9 marks
Total D marks: 12 of 12 marks
Total marks: 80 of 60 marks
Total Mastery marks: 20 of 20 marks

MATH60047 Stochastic Simulation

Question	Marker's comment
1	This question aims at measuring basic knowledge about direct and exact sampling methods. Some parts of this question contained problems that were similar to the ones covered in the course. The question is attempted by almost all students and usually was solved with high accuracy.
2	This questionnbsp;aims at measuring the knowledge of Monte Carlo integration and Importance sampling. The question contained a challenging integration problem. The question in general was attempted by all students and usually, again, done well on average.
3	This question aims at measuring the knowledge of Markov chain Monte Carlo and Bayesian updating. The question was in general well done, there was some confusion about Metropolis Adjusted Langevin algorithm vs Langevin algorithm but overall the questions were answered correctly.

MATH70047 Stochastic Simulation

Question	Marker's comment
1	This question aims at measuring basic knowledge about direct and exact sampling methods. Some parts of this question contained problems that were similar to the ones covered in the course. The question is attempted by almost all students and usually was solved with high accuracy.
2	This questionnbsp;aims at measuring the knowledge of Monte Carlo integration and Importance sampling. The question contained a challenging integration problem. The question in general was attempted by all students and usually, again, done well on average.
3	This question aims at measuring the knowledge of Markov chain Monte Carlo and Bayesian updating. The question was in general well done, there was some confusion about Metropolis Adjusted Langevin algorithm vs Langevin algorithm but overall the questions were answered correctly.
4	This mastery question aimed at measuringnbsp;knowledge related to sequential Monte Carlo filtering. The answers to this question were mixed; some on point but also many struggled, especially about the variance computations, despite similar calculations were in the mastery material.