IMPERIAL

MATH70047

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
Summer 2025

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

Stochastic Simulation

Date: Tuesday, May 6, 2025

Time: Start time 10:00 – End time 12:00 (BST)

Time Allowed: 2 hours

This paper has 4 Questions.

Please Answer Each Question in a Separate Answer Booklet

This is a closed book examination.

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.

Credit will be given for all questions attempted.

Allow margins for marking.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO DO SO

1. (a) Let X and Y be continuous random variables. Show that one can obtain $Y \sim p_Y$ with a desired p_Y by choosing $g(x) = F_Y^{-1}(F_X(x))$, where F_X and F_Y are the (invertible) cumulative distribution functions of X and Y, respectively. (3 marks)

(b) Throughout parts (b) and (c), the notation Uniform(a, b) denotes a uniform distribution on $[a, b]$ and $\chi^2(k)$ denotes the chi-squared distribution with k degrees of freedom with density:

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0.$$

Recall that $\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} \mathrm{d}t = (n-1)!$ for $n \in \mathbb{N}$.

(i) State the multivariate transformation of random variables formula. (2 marks)

(ii) Consider the following sampling procedure:

$$r \sim \chi^2(k) \quad \text{and} \quad \theta \sim \text{Uniform}(-\pi, \pi),$$

and $X = \sqrt{r}\cos(\theta)$ and $Y = \sqrt{r}\sin(\theta)$. Using the transformation of random variables formula, derive the joint density of $p_{X,Y}(x, y)$. (3 marks)

(c) Consider a target distribution:

$$p(x) \propto x^2 e^{-x/2}, \quad x > 0.$$

and a proposal:

$$q_\lambda(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

(i) Provide the pseudocode of the rejection sampling algorithm by choosing the appropriate M in terms of λ. Provide the constraint for λ for this to be well defined. (3 marks)

(ii) Find $\lambda_\star$ that maximises the acceptance rate. (2 marks)

(iii) Calculate the value of the acceptance rate using $q_{\lambda_\star}$ as the proposal. (3 marks)

(iv) Suppose we run this rejection sampler for K iterations. Denote the number of *accepted* samples N. Provide the expression for the expected number of accepted samples, i.e. $\mathbb{E}[N]$, in terms of K. (1 mark)

(v) Derive the distribution of N, i.e., $\mathbb{P}(N)$, in terms of K. (3 marks)

(Total: 20 marks)

2. (a) Consider a Bayesian model that consists of a prior $p(x)$ and the likelihood $p(y|x)$. Assume that we only know the prior in an unnormalised form $\bar{p}(x)$ and $p(x) = \bar{p}(x)/Z$ where Z is the unknown (positive) normalising constant. Consider first the task of estimating the marginal likelihood $p(y)$ under this setting.

   (i) Provide the expression of $p(y)$ in terms of the unnormalised prior $\bar{p}(x)$, Z, and the likelihood $p(y|x)$. (1 mark)

   (ii) Derive a self-normalised importance sampling estimator for $p(y)$ assuming only the knowledge of the prior in an unnormalised form and the likelihood. (3 marks)

   (iii) Is this estimator unbiased? Justify your answer. (1 mark)

   (b) Suppose that we have a probabilistic model with a parameter θ, i.e. we have $p_\theta(x, y)$ as the joint distribution. Consider the task of estimating θ by maximising $p_\theta(y)$ where y is fixed.

   (i) Show that

$$\nabla_\theta \log p_\theta(y) = \mathbb{E}_{p_\theta(x|y)} \left[\nabla_\theta \log p_\theta(x, y)\right].$$

   (3 marks)

   Hint: Use the identity $\nabla_\theta \log p_\theta(y) = \nabla_\theta p_\theta(y)/p_\theta(y)$ and that $p_\theta(y) = \int p_\theta(x, y)\mathrm{d}x$. Assume you can swap derivatives and integrals.

   (ii) Provide a self-normalised importance sampling based estimator for estimating this gradient. (5 marks)

   (c) Consider a generic target density

$$p(x) = \mathcal{N}(x; 0, 1).$$

   where $\mathcal{N}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. We are interested in estimating the expectation of a function $\varphi(x) = \mathbf{1}_{\{x \geq K\}}(x)$ for some $K > 0$ via importance sampling (IS). Consider a proposal with fixed variance:

$$q_\mu(x) = \mathcal{N}(x; \mu, 1/2).$$

   We would like to find the best proposal (i.e. μ) that minimises the variance of the importance sampling estimator.

   (i) Describe the optimisation problem for finding the best proposal that minimises the IS estimator variance. (3 marks)

   (ii) Find the minimum variance proposal $q_{\mu^\star}$ by identifying the appropriate $\mu^\star$ via variance minimization. Provide any constraints on μ for this variance to be finite. (4 marks)

(Total: 20 marks)

3. (a) This part is about Markov chain Monte Carlo (MCMC) theory.

(i) Describe the Metropolis-Hastings algorithm targeting a distribution $p_\star(x)$ with an independent proposal $q(x)$ by providing the pseudocode. (2 marks)

(ii) Suppose the chain is at a given point x. Describe the formula for the probability that the next proposed sample is rejected at x for the sampler in 3(a)(i). (2 marks)

(iii) Describe the Metropolis-Hastings algorithm targeting a distribution $p_\star(x)$ with a general proposal $q(x'|x)$ by providing the pseudocode. (2 marks)

(iv) Prove that the Metropolis-Hastings algorithm you provided in Part (iii) satisfies the detailed balance condition. (5 marks)

(b) Suppose that you would like to optimise a function $f : \mathbb{R} \to \mathbb{R}$ globally.

(i) Describe a Metropolis acceptance based algorithm to optimise this function. Outline the pseudocode and relevant parameter choices of the algorithm. (2 marks)

(ii) Describe a Langevin MCMC based algorithm to optimise this function. Outline the pseudocode and relevant parameter choices of the algorithms. (2 marks)

(c) Consider the following modified unadjusted Langevin algorithm (ULA):

$$X_{k+1} = X_k + \gamma \nabla \log p_\star(X_k) + \sigma_q W_k,$$

where $W_k \sim \mathcal{N}(0, 1)$ (assuming the chain is 1D), $\gamma > 0$, and $\sigma_q > 0$. Consider the target distribution:

$$p_\star(x) = \mathcal{N}(x; \mu, \sigma^2).$$

(i) Derive the limiting distribution of the chain $(X_k)_{k\geq 0}$. Provide any necessary constraints on γ for this to be well-defined. (3 marks)

(ii) Can you make ULA unbiased in the limit with an appropriate choice of σ_q? If no, provide your reasoning. If yes, provide the expression of σ_q. (2 marks)

(Total: 20 marks)

4. This question is based on Chapters 8, 11, and 12 of *Bayesian Filtering and Smoothing*, Simo Särkkä, Cambridge University Press, 2013.

(a) Consider the state-space model

$$x_0 \sim \pi_0(x_0), \quad x_k|x_{k-1} \sim f(x_k|x_{k-1}), \quad y_k|x_k \sim g(y_k|x_k),$$

for $k = 1, \ldots, K$, where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$. Assume that the prior $\pi_0(x_0)$, the transition density $f(x_k|x_{k-1})$, and the observation density $g(y_k|x_k)$ are known.

(i) Describe the filtering problem in the context of this state-space model. Similarly, describe the prediction problem. (2 marks)

(ii) Provide an expression of the incremental marginal likelihood $p(y_t|y_{1:t-1})$ in terms of the predictive distribution and any other given quantities in the model. (2 marks)

(iii) Provide the pseudocode of the Bootstrap Particle Filter (BPF) method for approximating the filtering distribution. (3 marks)

(iv) Suppose we run the BPF until time T and would like to compute $p(y_{1:t})$ in a numerically stable way. Provide the appropriate additions to the BPF algorithm to compute this quantity as a byproduct of the BPF. Elaborate on what is required for this computation to be stable. (4 marks)

(b) Consider the state-space model with a parameter θ:

$$x_0 \sim \pi_0(x_0), \quad x_k|x_{k-1} \sim f_\theta(x_k|x_{k-1}) \quad \text{and} \quad y_k|x_k \sim g_\theta(y_k|x_k).$$

Suppose that we are interested in implementing a gradient-based parameter estimation method for this model given the observations $y_{1:T}$.

(i) Derive the expression of the gradient of the log marginal likelihood $\nabla_\theta \log p_\theta(y_{1:T})$ as an expectation. (4 marks)

(ii) Provide the pseudocode for the sequential importance resampling (SIR) smoother with proposal f. (2 marks)

(iii) Using the smoother in Part(b)(ii) as a subroutine, describe a gradient ascent method to estimate θ by maximising the log marginal likelihood. (3 marks)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2025

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) If $X \sim p_X$ and $U = F_X(X)$, then

sim. seen ⇓

$$P(U \leq u) = P(F_X(X) \leq u) = P(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u$$

which means that U is uniform. Thus,

$$P(Y \leq y) = P(F_Y^{-1}(U) \leq y) = P(U \leq F_Y(y)) = F_Y(y)$$

since U is uniform. Thus, $Y \sim F_Y$.

3, B

(b) (i) The multivariate change of variables formula for $Y = g(X)$:

seen ⇓

$$p_Y(y) = p_X(g^{-1}(y)) \left|\det J_{g^{-1}}(y)\right|$$

where $J_{g^{-1}}(y)$ is the Jacobian of g^{-1} evaluated at y.

2, A

(ii) We have the formula for 2D as above and

sim. seen ⇓

$$[x, y] = [g_1(r, \theta), g_2(r, \theta)]$$

where $g_1(r, \theta) = \sqrt{r}\cos\theta$ and $g_2(r, \theta) = \sqrt{r}\sin\theta$. Thus, $r = g_1^{-1}(x, y) = x^2 + y^2$ and $\theta = g_2^{-1}(x, y) = \arctan(y/x)$. The Jacobian has an identical computation to the example in the lecture notes and it evaluates to 2. Next,

$$\begin{aligned} p_{X,Y}(x, y) &= p_{r,\theta}(g_1^{-1}(x, y), g_2^{-1}(x, y))2 \\ &= p_r(x^2 + y^2)p_\theta(\arctan(y/x))2 \\ &= \frac{1}{2^{k/2}\Gamma(k/2)}(x^2 + y^2)^{k/2-1}e^{-(x^2+y^2)/2}\frac{1}{\pi}, \end{aligned}$$

is the asked density.

3, A

(c) (i) The optimal choice of M is given by

sim. seen ⇓

$$M = \sup_x \frac{\bar{p}(x)}{q(x)}.$$

where $\bar{p}(x) = x^2 e^{-x/2}$. By computing the ratio, we obtain

$$R(x) = \frac{1}{\lambda}x^2 e^{-(1/2-\lambda)x}.$$

For this to be well-defined, we need $1/2 - \lambda > 0$ or $\lambda < 1/2$. By taking its log derivative

$$\mathrm{d}\log R(x)/\mathrm{d}x = \frac{2}{x} - (1/2 - \lambda)\,.$$

and setting it to zero, we obtain

$$x^\star = \frac{4}{1 - 2\lambda}.$$

We get:

$$M_\lambda = R(x^\star) = \frac{1}{\lambda}\left(\frac{4}{1-2\lambda}\right)^2 e^{-2}.$$

For fixed λ, the rejection sampler is given below

* $Y \sim q_\lambda$
* $U \sim \text{Uniform}(0,1)$
* If

$$U \leq \frac{\bar{p}(Y)}{M_\lambda q_\lambda(Y)} = \frac{Y^2 e^{-Y/2}}{\frac{1}{\lambda}\left(\frac{4}{1-2\lambda}\right)^2 e^{-2}\lambda e^{-\lambda Y}},$$

then accept Y, otherwise reject.

* Iterate.

3, A

(ii) We compute

$$M_\lambda = R(x^\star) = \frac{1}{\lambda}\left(\frac{4}{1-2\lambda}\right)^2 e^{-2}.$$

Taking the log and derivative of M_λ, it can be shown that $\lambda^\star = 1/6$.

2, B

(iii) The acceptance rate is given by

$$\hat{a} = \frac{Z}{M},$$

as our density is unnormalised. In this case, we can compute

$$Z = \int_0^\infty x^2 e^{-x/2}\mathrm{d}x = 2^3\Gamma(3) = 16.$$

The students can recognise that the density in Part (1)(b) is a special case and normalising constant can be found by just identifying k. Using $q_{\lambda^\star}$, we can evaluate

$$M_{\lambda^\star} = 216e^{-2}.$$

Merging these two together, the acceptance rate is

$$\hat{a} = \frac{16}{216e^{-2}} = \frac{2}{27}e^2.$$

3, B

(iv) Since acceptance rate is $\hat{a} = (2/27)e^2$, for K trials, the expected number of accepted samples will be $\mathbb{E}[N] = K\hat{a}$. The students are not expected to explicitly compute this, the intuitive answer is enough.

1, A

(v) Recall that, we can see the accept/reject as a binary random variable, e.g., a Bernoulli variable with $p = \hat{a}$. Thus, for K trials, we would have a binomial distribution with K trials and $p = \hat{a}$ as the distribution of N. In other words,

$$P(N) = \binom{K}{N}(2/27)^N e^{2N}(1 - 2/27e^2)^{K-N}.$$

3, C

2. (a) (i) Given an unnormalised prior $\bar{p}(x)$ and the likelihood $p(y|x)$, we write

seen ⇓

$$p(y) = \int p(y|x)p(x)\mathrm{d}x = \frac{1}{Z}\int p(y|x)\bar{p}(x)\mathrm{d}x.$$

1, A

(ii) Note that

sim. seen ⇓

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

by using the identity trick, we obtain

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

By sampling from $q(x)$, we can obtain

$$p^N(y) = \frac{\sum_{i=1}^N \mathsf{W}_i p(y|X_i)}{\sum_{i=1}^N \mathsf{W}_i},$$

where $\mathsf{W}_i = \bar{p}(X_i)/q(X_i)$.

3, B

(iii) This estimator is *not* unbiased. This is a ratio of two unbiased estimators, but by itself it will not be unbiased.

1, B

(b) (i) We write

seen ⇓

$$\begin{aligned}\nabla \log p_\theta(y) &= \frac{\nabla p_\theta(y)}{p_\theta(y)} = \frac{\nabla \int p_\theta(x,y)\mathrm{d}x}{\int p_\theta(x,y)\mathrm{d}x} = \frac{\int \nabla p_\theta(x,y)\mathrm{d}x}{\int p_\theta(x,y)\mathrm{d}x},\\ &= \frac{\int \nabla \log p_\theta(x,y) p_\theta(x,y)\mathrm{d}x}{\int p_\theta(x,y)\mathrm{d}x} = \mathbb{E}_{p_\theta(x|y)}[\nabla \log p_\theta(x,y)].\end{aligned}$$

3, C

(ii) This is a classical task for estimating an expectation. Here the test function $\varphi(x) = \nabla_\theta \log p_\theta(x,y)$ since y is fixed and θ is also fixed for a given estimation task (e.g. at θ_k). Since

unseen ⇓

$$p_\theta(x|y) = \frac{p_\theta(x,y)}{\int p_\theta(x,y)\mathrm{d}x},$$

by choosing a proposal $q(x)$ and taking samples $X_i \sim q(x)$, we can estimate the expectation as

$$\mathbb{E}_{p_\theta(x|y)}[\nabla \log p_\theta(x,y)] \approx \frac{\sum_{i=1}^N \nabla \log p_\theta(X_i,y)\mathsf{W}_i}{\sum_{i=1}^N \mathsf{W}_i},$$

where $\mathsf{W}_i = p_\theta(X_i,y)/q(X_i)$.

5, D

(c) (i) To minimize the IS estimator variance for a proposal q_μ with the parameter μ, the problem to solve is

sim. seen ⇓

$$\mu^\star = \arg\min_\mu \mathbb{E}_q\left[\varphi^2(X)\frac{p^2(X)}{q^2(X)}\right]. \tag{1}$$

3, A

(ii) Using the quantities in the question, we compute

$$V(\mu) = \mathbb{E}_{q_\mu}\left[\varphi^2(X)\frac{p^2(X)}{q_\mu^2(X)}\right] = \int_{-\infty}^{\infty} 1^2_{\{x\geq K\}}(x)\frac{p^2(x)}{q_\mu(x)}\mathrm{d}x,$$
$$= \int_K^\infty \frac{\sqrt{\pi}}{2\pi}\frac{e^{-x^2}}{e^{-(x-\mu)^2}}\mathrm{d}x,$$
$$= (1/2\sqrt{\pi})e^{\mu^2}\int_K^\infty e^{-2\mu x}\mathrm{d}x.$$

For this integral to be finite, we impose $\mu > 0$. Then we have

$$V(\mu) = (1/2\sqrt{\pi})e^{\mu^2}\frac{e^{-2\mu K}}{\mu}.$$

Taking the log derivative and setting it to zero

$$\mathrm{d}V(\mu)/\mathrm{d}\mu = 2\mu - 2K - \frac{1}{\mu} = 0. \tag{2}$$

This leads to a quadratic equation:

$$2\mu^2 - 2K\mu - 1 = 0.$$

With solutions

$$\mu = \frac{2K \pm \sqrt{4K^2+8}}{4}.$$

Since $\mu > 0$, we choose the positive root:

$$\mu^\star = \frac{2K + \sqrt{4K^2+8}}{4}.$$

4, D

3. (a) (i) The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method that generates a Markov chain with the target density p as its stationary distribution. The algorithm is as follows: seen ⇓

- Given $X_t = x$, sample $x' \sim q(x')$.
- Compute the acceptance probability

$$\alpha(x, x') = \min\left\{1, \frac{p(x')q(x)}{p(x)q(x')}\right\}.$$

- Accept $X_{t+1} = x'$ with probability $\alpha(x, x')$.
- Otherwise, reject and set $X_{t+1} = x$. 2, A

(ii) We accept with probability $\alpha(x, x')$ where $x' \sim q(x')$. Therefore, the probability of accepting a sample is

$$a(x) = \int_{\mathsf{X}} \alpha(x, x')q(x')\mathrm{d}x'.$$

As a result, the probability of rejecting a sample is $1 - a(x)$. 2, A

(iii) The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method that generates a Markov chain with the target density p as its stationary distribution. The algorithm is as follows:

- Given $X_t = x$, sample $x' \sim q(x'|x)$.
- Compute the acceptance probability

$$\alpha(x, x') = \min\left\{1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right\}.$$

- Accept $X_{t+1} = x'$ with probability $\alpha(x, x')$.
- Otherwise, reject and set $X_{t+1} = x$. 2, A

(iv) Proof from the lecture notes: We first define the kernel induced by the MH algorithm. This is given as

$$K(x'|x) = \alpha(x, x')q(x'|x) + (1 - a(x))\delta_x(x'),$$

where δ_x is the Dirac delta function and 1, A

$$a(x) = \int_{\mathsf{X}} \alpha(x, x')q(x'|x)\mathrm{d}x',$$

is the probability of accepting a sample. Given this, to check detailed balance, we check for a function f

$$\int f(x, x')p_\star(x)K(x'|x)\mathrm{d}x\mathrm{d}x' = \int f(x, x')p_\star(x)q(x'|x)\alpha(x, x')\mathrm{d}x\mathrm{d}x' + \int f(x, x')p_\star(x)(1 - a(x))\delta_x(\mathrm{d}x')\mathrm{d}x.$$

For the first term, we can easily see that 1, B

$$p_\star(x)q(x'|x)\alpha(x, x') = p_\star(x)q(x'|x)\min\left\{1, \frac{p_\star(x')q(x|x')}{p_\star(x)q(x'|x)}\right\} = \min\left\{p_\star(x)q(x'|x), p_\star(x')q(x|x')\right\}$$
$$= \min\left\{\frac{p_\star(x)q(x'|x)}{p_\star(x')q(x|x')}, 1\right\}p_\star(x')q(x|x') = p_\star(x')q(x|x')\alpha(x', x).$$

Since the second term involves a Dirac measure, we will check it against the integral form directly:

$$\begin{aligned}\int f(x,x')p_\star(x)(1-a(x))\delta_x(\mathrm{d}x')\mathrm{d}x &= \int f(x,x)p_\star(x)(1-a(x))\mathrm{d}x \\ &= \int f(x',x')p_\star(x')(1-a(x'))\mathrm{d}x' \\ &= \int f(x,x')p_\star(x')(1-a(x'))\delta_{x'}(\mathrm{d}x)\mathrm{d}x'.\end{aligned}$$

All in all, we can see that

$$\int f(x,x')p_\star(x)K(x'|x)\mathrm{d}x\mathrm{d}x' = \int f(x,x')p_\star(x')K(x|x')\mathrm{d}x\mathrm{d}x',$$

which is the detailed balance condition. 3, C

(b) (i) A Metropolis based algorithm to optimise a function f is simulated annealing. In order to minimise f, simulated annealing targets $p^{\beta_t}(x) \propto \exp(-\beta_t f(x))$ where β_t is the inverse temperature as follows: seen ⇓

- $X' \sim q(\cdot|X_{t-1})$ (symmetric proposal, e.g., random walk)
- Set β_t (e.g. $\beta_t = \sqrt{1+t}$)
- Accept X' with probability

$$\min\left\{1, p^{\beta_t}(X')/p^{\beta_t}(X_{t-1})\right\}.$$

  and set $X_t = X'$ if accepted.
- Otherwise set $X_t = X_{t-1}$ 2, A

(ii) A Langevin MCMC based approach is to target a distribution $p^\beta(x) \propto \exp(-\beta f(x))$ and use the Langevin dynamics to sample from it for large β as this induces concentration. The Langevin dynamics to sample from this can be described as follows:

$$X_{t+1} = X_t - \gamma\nabla f(X_t) + \sqrt{\frac{2\gamma}{\beta}}Z_t,$$

where $Z_t \sim \mathcal{N}(0, I)$ and γ is the step size. 2, A

(c) (i) The solution follows a similar structure to Example 5.14 in the lecture notes. The algorithm described in the question can be written as sim. seen ⇓

$$X_{k+1} = X_k - \gamma\frac{X_k - \mu}{\sigma^2} + \sigma_q W_k.$$

Let

$$a = 1 - \frac{\gamma}{\sigma^2}, \quad b = \frac{\gamma}{\sigma^2}\mu.$$

We can write now the iterates beginning at x_0 as

$$\begin{aligned}x_1 &= ax_0 + b + \sigma_q W_1, \\ x_2 &= \underbrace{a^2x_0 + ab + a\sigma_q W_1}_{ax_1} + b + \sigma_q W_2, \\ x_3 &= \underbrace{a^3x_0 + a^2b + a^2\sigma_q W_1 + ab + a\sigma_q W_2}_{ax_2} + b + \sigma_q W_3, \\ &\vdots \\ x_n &= a^n x_0 + \sum_{k=0}^{n-1} a^k b + \sum_{k=0}^{n-1} a^k \sigma_q W_{n-k}.\end{aligned}$$

We can compute the expected value

$$\mathbb{E}[X_n] = a^n x_0 + \sum_{k=0}^{n-1} a^k b,$$

since W_k are zero mean. We require $|a| < 1$ which implies that $0 < \gamma < 2\sigma^2$. As $n \to \infty$, we have

$$\mu_\infty = \lim_{n\to\infty} \mathbb{E}[X_n] = \sum_{k=0}^{\infty} a^k b = \frac{b}{1-a} = \mu.$$

The variance of the iterates as $n \to \infty$ can also be computed. Note that for finite n, we have

$$\begin{aligned}
\mathsf{var}(x_n) &= \mathsf{var}\left(\sum_{k=0}^{n-1} a^k \sigma_q W_k\right), \\
&= \sigma_q^2 \sum_{k=0}^{n-1} (a^2)^k, \\
&= \sigma_q^2 \frac{1-a^{2n}}{1-a^2}.
\end{aligned}$$

Therefore, we obtain the limiting variance as

$$\begin{aligned}
\lim_{n\to\infty} \mathsf{var}(x_n) &= \sigma_q^2 \frac{1}{1-a^2} \\
&= \sigma_q^2 \frac{1}{1-\left(1-\frac{\gamma}{\sigma^2}\right)^2} \\
&= \sigma_q^2 \frac{1}{\frac{2\gamma}{\sigma^2} - \frac{\gamma^2}{\sigma^4}}, \\
&= \frac{2\sigma^4}{2\sigma^2 - \gamma}.
\end{aligned}$$

Hence

$$\lim_{n\to\infty} \mathsf{var}(X_n) = \sigma_q^2 \frac{\sigma^4}{2\gamma\sigma^2 - \gamma^2}.$$

The limiting distribution is Gaussian as it is a linear combination of Gaussian random variables, with the mean and variance provided above. 3, D

(ii) The choice

$$\sigma_q^2 = \frac{2\gamma\sigma^2 - \gamma^2}{\sigma^2},$$

gives us

$$\lim_{n\to\infty} \mathsf{var}(X_n) = \sigma^2,$$

hence results in an unbiased sampler as the target distribution is now exactly what we want, i.e., $p_\star(x) = \mathcal{N}(x; \mu, \sigma^2)$. 2, B

4. (a) (i) The filtering problem is to sequentially compute the posterior distribution of the state X_t given the observations $y_{1:t} = (y_1, \ldots, y_t)$, i.e., $p(x_t|y_{1:t})$. The prediction problem is to compute distributions $p(x_t|y_{1:t-1})$, i.e., predict the state X_t given the past data $y_{1:t-1}$. seen ⇓ 2, M

(ii) The incremental marginal likelihood can be written as

$$p(y_t|y_{1:t-1}) = \int g(y_t|x_t)p(x_t|y_{1:t-1})\mathrm{d}x_t.$$

2, M

(iii) The bootstrap particle filter can be described as follows. For brevity, we write one step, assuming the access to a sample $X_{k-1}^{(i)}$ for $i = 1, \ldots, N$ from the previous step:

- Sample $\bar{X}_k^{(i)} \sim f(\cdot|X_{k-1}^{(i)})$.
- Compute the weight $W_k^{(i)} = g(y_k|\bar{X}_k^{(i)})$.
- Normalise the weights $w_k^{(i)} = W_k^{(i)} / \sum_{i=1}^N W_k^{(i)}$.
- Resample $X_k^{(i)}$ with probability $w_k^{(i)}$ from the set $\{\bar{X}_k^{(i)}\}_{i=1}^N$.

3, M

(iv) The full marginal likelihood is given as

$$p(y_{1:t}) = \prod_{k=1}^{t} p(y_k|y_{1:k-1}),$$

where $p(y_1|y_0) = p(y_1)$. For a given step k of the algorithm, the incremental marginal likelihood can be approximated as

$$\begin{aligned} p(y_k|y_{1:k-1}) &= \int g(y_k|x_k)p(x_k|y_{1:k-1})\mathrm{d}x_k, \\ &\approx \frac{1}{N}\sum_{i=1}^{N} g(y_k|\bar{X}_k^{(i)}). \end{aligned}$$

The estimator is thus given as

$$p^N(y_{1:t}) = \prod_{k=1}^{t} \frac{1}{N}\sum_{i=1}^{N} g(y_k|\bar{X}_k^{(i)}).$$

A stable computation of this requires a log-sum-exp trick. 4, M

(b) (i) The gradient can be written as unseen ⇓

$$\begin{aligned} \nabla_\theta \log p_\theta(y_{1:t}) &= \frac{\nabla_\theta p_\theta(y_{1:t})}{p_\theta(y_{1:t})} = \frac{\nabla_\theta \int p_\theta(x_{0:t}, y_{1:t})\mathrm{d}x_{0:t}}{p_\theta(y_{1:t})} \\ &= \frac{\int \nabla_\theta p_\theta(x_{0:t}, y_{1:t})\mathrm{d}x_{0:t}}{p_\theta(y_{1:t})} = \frac{\int \nabla_\theta \log p_\theta(x_{0:t}, y_{1:t}) p_\theta(x_{0:t}, y_{1:t})\mathrm{d}x_{0:t}}{p_\theta(y_{1:t})} \\ &= \mathbb{E}_{p_\theta(x_{0:t}|y_{1:t})}[\nabla_\theta \log p_\theta(x_{0:t}, y_{1:t})]. \end{aligned}$$

4, M

(ii) SIR smoother is a straightforward method that accumulates the samples. Given $X_{0:k-1}^{(i)}$ samples,

- Run the BPF sampling and weighting steps to obtain the samples $\bar{X}_k^{(i)}$ and weights $w_k^{(i)}$.
- Construct $\bar{X}_{0:k}^{(i)} = X_{0:k-1}^{(i)} \cup \bar{X}_k^{(i)}$ (append).
- Resample $X_{0:k}^{(i)}$ with probability $w_k^{(i)}$ from the set $\{\bar{X}_{0:k}^{(i)}\}_{i=1}^N$.

At the end of t iterations, the samples $\{X_{0:t}^{(i)}\}_{i=1}^N$ are the smoothed samples, i.e., approximating $p(x_{0:t}|y_{1:t})$. 2, M

(iii) Assume we are at θ_k. The gradient of the log marginal likelihood can be approximated by running an SIR smoother as

$$\begin{aligned}\nabla_\theta \log p_{\theta_k}(y_{1:t}) &= \int \nabla_\theta \log p_{\theta_k}(x_{0:t}, y_{1:t}) p_{\theta_k}(x_{0:t}|y_{1:t}) \mathrm{d}x_{0:t} \\ &\approx \frac{1}{N}\sum_{i=1}^N \nabla_\theta \log p_{\theta_k}(X_{0:t}^{(i)}, y_{1:t}) = \nabla_\theta \widehat{\log p_{\theta_k}}(y_{1:t}).\end{aligned}$$

Note the samples depend on θ_k. We can then define the gradient descent scheme as

$$\theta_{k+1} = \theta_k + \gamma_k \nabla_\theta \widehat{\log p_{\theta_k}}(y_{1:t}).$$

3, 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

MATH70047 Stochastic Simulation Markers Comments

Question 1	There is a good number of students who scored highly in this question. The main confusion points were: 1 (i): This was a question created a common confusion, despite being relatively straightforward in part (c): Many students missed the normalising constant. (c)(v) was attempted fewer times than other questions. Overall, the students provided reasonable answers.
Question 2	Well done on addressing question 2 on this exam. While I was impressed with many answers given the relatively short amount of time and exam conditions, there are a few points that I would like to provide as feedback: in part (a-iii), the correct answer is that self-normalising IS samplers are ratios of two unbiased estimators and so in general not unbiased. It is worth noting this general point. Part (b) was overall felt as the hardest sub-question of Q2. Many students did not see the link to part (a) in that we can only evaluate the joint density p_\theta(x,y) rather than the posterior p_\theta(x | y) =p_\theta(x,y) / Z with unknown normalising constant. With this insight, most of the calculations followed relatively straightforwardly in the spirit of part (a). I thought part (c) was very well done, with a very large proportion of students completing the calculus part correctly.
Question 3	Many students had trouble with writing the acceptance probability that gave the correct conditioning (3aiii), time instances and correct densities (3aii and 3ai); there seemed to be some confusion between variables of target density functions and random variables. Given that the sub questions carried on from previous answers given earlier on, I did not penalise twice for a mistake made in earlier sub questions. This question required quite a few technical details, which students overall did well at; I gave credit showing a clear understanding even if some details were missing along the way but there was a good explanation and demonstration of understanding.
Question 4	There was some confusion about some parts of this question, despite it being relatively straightforward given the lecture and mastery material. Parts 4(a)(i-iv) are relatively straightforward, but in general the level of answers was lower than past years (of similar questions). There were confusions around weight computation steps, BPF algorithm, and general concepts of filtering/marginal likelihoods. Parts 4(b) was about the most basic smoother in the mastery material, SIR

smoother. Despite it being an algorithm, literally, BPF but operated on paths space, there was a confusion about the method.