Topics in Statistics: Markov chain Monte Carlo
Problem Set 2

Carsten Jørgensen

Exercise 1

Suppose that we wish to use the Gibbs sampler on

$$\pi(x, y) \propto \exp\left(-\frac{1}{2}(x-1)^2(y-2)^2\right).$$

1. Write down the two "full" conditional distributions associated to $\pi(x, y)$.
2. Does the resulting Gibbs sampler make any sense?

Solution

Given the target distribution:

$$\pi(x, y) \propto \exp\left(-\frac{1}{2}(x-1)^2(y-2)^2\right)$$

Question 1: Full conditional distributions

To find the full conditional distributions, we examine the distribution of each variable while treating the other as fixed.

Conditional distribution of X given Y:

$$\begin{aligned}\pi(x|y) &\propto \exp\left(-\frac{1}{2}(x-1)^2(y-2)^2\right) \\ &\propto \exp\left(-\frac{(y-2)^2}{2}(x-1)^2\right)\end{aligned}$$

Since $(y-2)^2$ is constant when y is fixed, this is the kernel of a normal distribution with mean 1 and precision $(y-2)^2$. Therefore:

$$\pi(x|y) = \mathcal{N}\left(1, \frac{1}{(y-2)^2}\right) \text{ for } y \neq 2$$

Conditional distribution of Y given X:

Similarly, we have:

$$\pi(y|x) = \mathcal{N}\left(2, \frac{1}{(x-1)^2}\right) \text{ for } x \neq 1$$

Part 2: Does the Gibbs sampler make sense?

Answer: No, the Gibbs sampler does NOT make sense for this distribution.
Here are the fundamental problems:

1. Undefined conditional distributions:

- When $y = 2$: $\text{Var}(X|y = 2) = \frac{1}{(2-2)^2} = \frac{1}{0} = \infty$ (undefined)
- When $x = 1$: $\text{Var}(Y|x = 1) = \frac{1}{(1-1)^2} = \frac{1}{0} = \infty$ (undefined)

2. Improper joint distribution: The joint distribution $\pi(x, y) \propto \exp\left(-\frac{1}{2}(x-1)^2(y-2)^2\right)$ becomes constant along the lines $x = 1$ and $y = 2$:

- Along $x = 1$: $\pi(1, y) \propto \exp(0) = 1$ (constant for all y)
- Along $y = 2$: $\pi(x, 2) \propto \exp(0) = 1$ (constant for all x)

This means the distribution does not integrate to a finite value, making it an improper distribution.

3. Practical failure:

- If the sampler reaches exactly $x = 1$ or $y = 2$, the corresponding conditional distribution becomes improper (uniform over the entire real line), making sampling impossible.
- Even near these values, the conditional variances become extremely large, leading to poor mixing and numerical instability.
- The sampler would exhibit erratic behavior and fail to converge to a proper stationary distribution.

Conclusion: The Gibbs sampler fails because the target distribution is improper and the conditional distributions are not well-defined at the modal lines. This example illustrates the importance of verifying that the target distribution is proper before attempting to use MCMC methods.

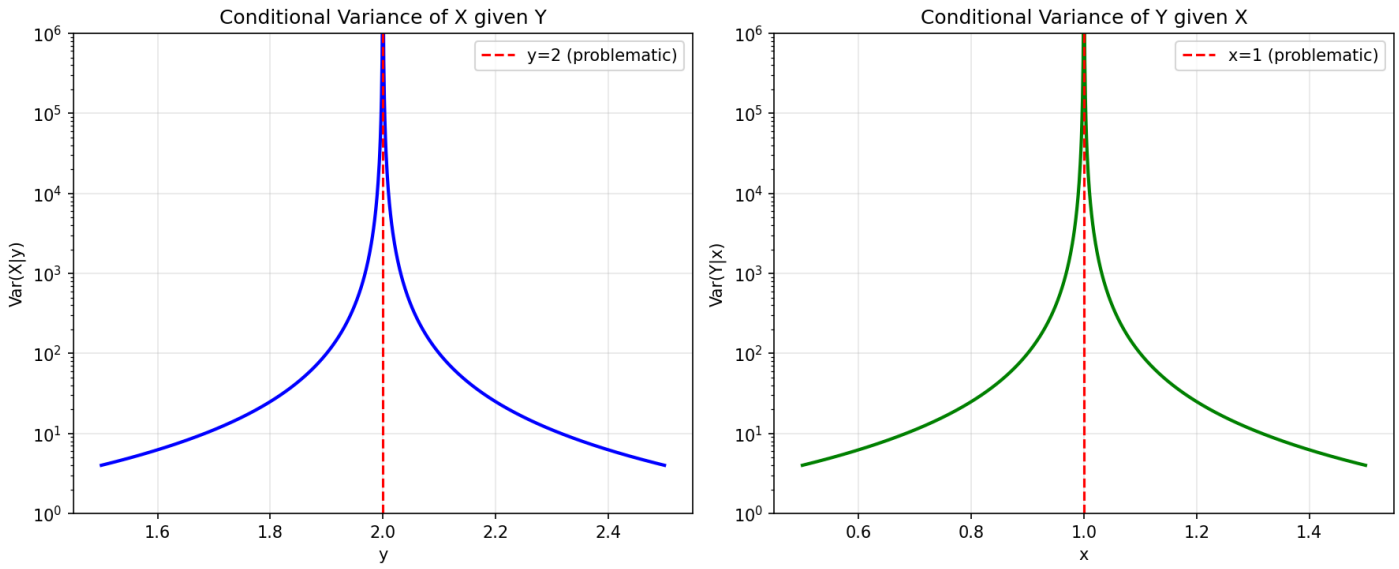

Figure 1: Plot of the conditional variances $\text{Var}(X|Y)$ and $\text{Var}(Y|X)$ showing they become infinite at $y = 2$ and $x = 1$ respectively.

Exercise 2

For $i = 1, \ldots, T$ consider $Z_i = X_i + Y_i$ with independent X_i, Y_i such that

$$X_i \sim \text{Binomial}(m_i, \theta_1), \quad Y_i \sim \text{Binomial}(n_i, \theta_2). \tag{1}$$

1. We assume $0 \leq z_i \leq m_i + n_i$ for $i = 1, \ldots, T$. We observe z_i for $i = 1, \ldots, T$ and the n_i, m_i, for $i = 1, \ldots, T$ are given. Give the expression of the likelihood function $p(z_1, \ldots, z_T | \theta_1, \theta_2)$.

2. Assume we set independent uniform priors $\theta_1 \sim U[0,1]$, $\theta_2 \sim U[0,1]$. Propose a Gibbs sampler to sample from $p(\theta_1, \theta_2 | z_1, \ldots, z_T)$. Recall that the Beta distribution of parameter $\alpha, \beta > 0$ admits a density $f(x) \propto x^{\alpha-1}(1-x)^{\beta-1} I_{[0,1]}(x)$.

   (*Hint: introduce auxiliary variables*)

Solution

Given: For $i = 1, \ldots, T$, consider $Z_i = X_i + Y_i$ with independent X_i, Y_i such that:

$$X_i \sim \text{Binomial}(m_i, \theta_1) \tag{2}$$
$$Y_i \sim \text{Binomial}(n_i, \theta_2) \tag{3}$$

Part 1: Likelihood Function

We observe z_i for $i = 1, \ldots, T$ but not the individual X_i and Y_i values. We need to find $p(z_1, \ldots, z_T | \theta_1, \theta_2)$. For each i, the probability mass function of $Z_i = X_i + Y_i$ is obtained by summing over all possible ways to achieve the observed sum z_i:

$$P(Z_i = z_i) = \sum_k P(X_i = k) \cdot P(Y_i = z_i - k) \tag{4}$$

The summation limits are determined by the constraints:

- $0 \leq k \leq m_i$ (since X_i cannot exceed m_i)
- $0 \leq z_i - k \leq n_i$ (since Y_i cannot exceed n_i)

This gives us $k \in \{\max(0, z_i - n_i), \ldots, \min(z_i, m_i)\}$.
Therefore:

$$P(Z_i = z_i) = \sum_{k=\max(0, z_i - n_i)}^{\min(z_i, m_i)} \binom{m_i}{k} \theta_1^k (1-\theta_1)^{m_i - k} \binom{n_i}{z_i - k} \theta_2^{z_i - k} (1-\theta_2)^{n_i - (z_i - k)} \tag{5}$$

The likelihood function is:

$$\boxed{p(z_1, \ldots, z_T | \theta_1, \theta_2) = \prod_{i=1}^{T} P(Z_i = z_i)} \tag{6}$$

where each $P(Z_i = z_i)$ is given by the formula above.

Part 2: Gibbs Sampler with Auxiliary Variables

Following the hint, we introduce the unobserved values X_i for $i = 1, \ldots, T$ as auxiliary variables. If we knew the X_i values, then $Y_i = Z_i - X_i$ would be determined.

Complete Data Likelihood:

$$p(z_1, \ldots, z_T, x_1, \ldots, x_T | \theta_1, \theta_2) \tag{7}$$

$$= \prod_{i=1}^{T} P(X_i = x_i | \theta_1) \cdot P(Y_i = z_i - x_i | \theta_2) \tag{8}$$

$$= \prod_{i=1}^{T} \binom{m_i}{x_i} \theta_1^{x_i} (1-\theta_1)^{m_i - x_i} \binom{n_i}{z_i - x_i} \theta_2^{z_i - x_i} (1-\theta_2)^{n_i - (z_i - x_i)} \tag{9}$$

Posterior Distribution: With uniform priors $\theta_1 \sim U[0,1]$ and $\theta_2 \sim U[0,1]$:

$$p(\theta_1, \theta_2, x_1, \ldots, x_T | z_1, \ldots, z_T) \tag{10}$$

$$\propto \theta_1^{\sum_{i=1}^T x_i} (1-\theta_1)^{\sum_{i=1}^T (m_i - x_i)} \theta_2^{\sum_{i=1}^T (z_i - x_i)} (1-\theta_2)^{\sum_{i=1}^T (n_i - (z_i - x_i))} \tag{11}$$

Full Conditional Distributions:

1. For θ_1:

$$\theta_1 | \theta_2, x_1, \ldots, x_T, z_1, \ldots, z_T \sim \text{Beta}\left(1 + \sum_{i=1}^{T} x_i, 1 + \sum_{i=1}^{T} m_i - \sum_{i=1}^{T} x_i\right) \tag{12}$$

2. For θ_2:

$$\theta_2 | \theta_1, x_1, \ldots, x_T, z_1, \ldots, z_T \sim \text{Beta}\left(1 + \sum_{i=1}^{T} (z_i - x_i), 1 + \sum_{i=1}^{T} n_i - \sum_{i=1}^{T} (z_i - x_i)\right) \tag{13}$$

3. For each X_i:

$$P(X_i = k | \theta_1, \theta_2, z_i) \propto \binom{m_i}{k} \theta_1^k (1-\theta_1)^{m_i - k} \binom{n_i}{z_i - k} \theta_2^{z_i - k} (1-\theta_2)^{n_i - (z_i - k)} \tag{14}$$

for $k \in \{\max(0, z_i - n_i), \ldots, \min(z_i, m_i)\}$.

Gibbs Sampling Algorithm:

1. **Initialize:** Choose starting values $\theta_1^{(0)}, \theta_2^{(0)}, x_1^{(0)}, \ldots, x_T^{(0)}$
2. **For** $t = 1, 2, 3, \ldots$:
   (a) **Update auxiliary variables:** For each $i = 1, \ldots, T$, sample
   $$X_i^{(t)} \sim P(X_i = \cdot | \theta_1^{(t-1)}, \theta_2^{(t-1)}, z_i) \tag{15}$$
   where the probability mass function is given above.
   (b) **Update θ_1:** Sample
   $$\theta_1^{(t)} \sim \text{Beta}\left(1 + \sum_{i=1}^{T} x_i^{(t)}, 1 + \sum_{i=1}^{T} m_i - \sum_{i=1}^{T} x_i^{(t)}\right) \tag{16}$$
   (c) **Update θ_2:** Sample
   $$\theta_2^{(t)} \sim \text{Beta}\left(1 + \sum_{i=1}^{T} (z_i - x_i^{(t)}), 1 + \sum_{i=1}^{T} n_i - \sum_{i=1}^{T} (z_i - x_i^{(t)})\right) \tag{17}$$

Summary: The key insight is to introduce the unobserved X_i values as auxiliary variables, which makes the conditional distributions conjugate and easy to sample from. The Gibbs sampler alternates between sampling the auxiliary variables from discrete distributions and sampling the parameters from Beta distributions.

Exercise 4

1. Prove the Cauchy-Schwarz inequality which states that for any two real-valued random variables Y and Z,
$$|E[YZ]|^2 \leq E\left[Y^2\right] E\left[Z^2\right].$$
(*Hint:* $(Y - \alpha Z)^2 \geq 0$ *for any* $\alpha \in \mathbb{R}$).

2. Using Cauchy-Schwarz inequality, show that when the marginal distributions of Y and Z are identical then
$$\mathrm{Cov}(Y, Z) \leq \mathrm{Var}(Y).$$

3. Thinning of a Markov chain $\{X^{(t)}\}_{t\geq 0}$ is the technique of retaining a subsequence of the sampled process for purposes of computing ergodic averages. For some $m \in \mathbb{N}$ we retain the "subsampled" chain $\{Y^{(t)}\}_{t\geq 0}$ defined by
$$Y^{(t)} := X^{(m\cdot t)}.$$
We might hope that $\{Y^{(t)}\}_{t\geq 0}$ will exhibit lower autocorrelation than the original chain $\{X^{(t)}\}_{t\geq 0}$ and thus will yield ergodic averages of lower variance.
Consider a stationary Markov chain $\{X^{(t)}\}_{t\geq 0}$. Let T and m be any two integers such that $T \geq m > 1$ and $T/m \in \mathbb{N}$. Show that
$$\mathrm{Var}\left[\frac{1}{T}\sum_{t=0}^{T-1} X^{(t)}\right] \leq \mathrm{Var}\left[\frac{1}{T/m}\sum_{t=0}^{T/m-1} Y^{(t)}\right]$$
and briefly explain what this result tells us about the use of thinning.
(*Hint: start by writing* $\sum_{t=0}^{T-1} X^{(t)} = \sum_{t=0}^{m-1}\sum_{s=0}^{T/m-1} X^{(s\cdot m+t)}$)

Solution

Question 1: Proof of Cauchy-Schwarz Inequality

Theorem 1 *For any two real-valued random variables Y and Z:*
$$|E[YZ]|^2 \leq E[Y^2]E[Z^2]$$

Following the hint, consider $(Y - \alpha Z)^2 \geq 0$ for any $\alpha \in \mathbb{R}$.
Expanding:
$$(Y - \alpha Z)^2 = Y^2 - 2\alpha YZ + \alpha^2 Z^2$$
Taking expectation:
$$E[(Y - \alpha Z)^2] = E[Y^2] - 2\alpha E[YZ] + \alpha^2 E[Z^2] \geq 0$$
This is a quadratic function in α that is always non-negative. For a quadratic $a\alpha^2 + b\alpha + c \geq 0$ for all α, the discriminant must satisfy $b^2 - 4ac \leq 0$.
Here: $a = E[Z^2]$, $b = -2E[YZ]$, $c = E[Y^2]$
Therefore:
$$\begin{aligned}(-2E[YZ])^2 - 4E[Z^2]E[Y^2] &\leq 0\\ 4(E[YZ])^2 - 4E[Z^2]E[Y^2] &\leq 0\\ (E[YZ])^2 &\leq E[Z^2]E[Y^2]\end{aligned}$$
Taking the square root:
$$|E[YZ]|^2 \leq E[Y^2]E[Z^2]$$

Question 2: Covariance Inequality with Identical Marginals

Theorem 2 *When Y and Z have identical marginal distributions:*

$$Cov(Y, Z) \leq Var(Y)$$

When Y and Z have identical marginal distributions:

- $E[Y^2] = E[Z^2]$
- $E[Y] = E[Z]$

From the Cauchy-Schwarz inequality:

$$|E[YZ]|^2 \leq E[Y^2]E[Z^2] = (E[Y^2])^2$$

Therefore: $|E[YZ]| \leq E[Y^2]$, which implies $E[YZ] \leq E[Y^2]$.
Now:

$$\text{Cov}(Y, Z) = E[YZ] - E[Y]E[Z] = E[YZ] - (E[Y])^2$$

We want to show:

$$\text{Cov}(Y, Z) \leq \text{Var}(Y) = E[Y^2] - (E[Y])^2$$

This is equivalent to showing:

$$E[YZ] - (E[Y])^2 \leq E[Y^2] - (E[Y])^2$$

Simplifying:

$$E[YZ] \leq E[Y^2]$$

This follows directly from our Cauchy-Schwarz result above.

Question 3: Thinning and Variance

Theorem 3 *For a stationary Markov chain $\{X^{(t)}\}_{t\geq 0}$ with thinned chain $Y^{(t)} := X^{(m\cdot t)}$, where $T \geq m > 1$ and $T/m \in \mathbb{N}$:*

$$Var\left[\frac{1}{T}\sum_{t=0}^{T-1} X^{(t)}\right] \leq Var\left[\frac{1}{T/m}\sum_{t=0}^{T/m-1} Y^{(t)}\right]$$

Let us define the total sum of the original chain:

$$S_{\text{total}} = \sum_{t=0}^{T-1} X(t)$$

Following the hint, we decompose this sum into m subsampled sequences:

$$S_{\text{total}} = \sum_{i=0}^{m-1}\sum_{s=0}^{T/m-1} X(s\cdot m + i) = \sum_{i=0}^{m-1} S_i$$

where $S_i = \sum_{s=0}^{T/m-1} X(s\cdot m + i)$. The variance of the average of the original chain is:

$$\text{Var}\left[\frac{1}{T}S_{\text{total}}\right] = \frac{1}{T^2}\text{Var}[S_{\text{total}}] = \frac{1}{T^2}\text{Var}\left[\sum_{i=0}^{m-1} S_i\right]$$

Expanding the variance of the sum:

$$\text{Var}\left[\sum_{i=0}^{m-1} S_i\right] = \sum_{i=0}^{m-1} \text{Var}[S_i] + 2\sum_{i=0}^{m-1}\sum_{j=i+1}^{m-1} \text{Cov}(S_i, S_j)$$

Since the chain is stationary, all S_i have the same variance. Let $V_s = \text{Var}[S_i]$ for all i. Then:

$$\sum_{i=0}^{m-1} \text{Var}[S_i] = mV_s$$

Now, we apply the result from Question 2: If two random variables have identical marginal distributions, then $\text{Cov}(Y, Z) \leq \text{Var}(Y)$. Due to stationarity, the marginal distributions of the components of S_i and S_j are identical, and this property extends to the sums S_i and S_j. Therefore:

$$\text{Cov}(S_i, S_j) \leq \text{Var}(S_i) = V_s \quad \text{for all } i \neq j$$

There are $\binom{m}{2} = \frac{m(m-1)}{2}$ covariance pairs. Thus:

$$2\sum_{i=0}^{m-1}\sum_{j=i+1}^{m-1} \text{Cov}(S_i, S_j) \leq 2 \cdot \frac{m(m-1)}{2} \cdot V_s = m(m-1)V_s$$

Putting both parts together:

$$\text{Var}[S_{\text{total}}] \leq mV_s + m(m-1)V_s = m^2 V_s$$

Now, the thinned chain is defined as $Y(t) = X(m \cdot t)$. Notice that $S_0 = \sum_{t=0}^{T/m-1} Y(t)$ is exactly the sum of the thinned chain. Therefore:

$$\text{Var}[S_0] = V_s$$

The variance of the average of the thinned chain is:

$$\text{Var}\left[\frac{1}{T/m}\sum_{t=0}^{T/m-1} Y(t)\right] = \frac{m^2}{T^2}\text{Var}[S_0] = \frac{m^2}{T^2}V_s$$

Finally, comparing the two variances:

$$\text{Var}\left[\frac{1}{T}S_{\text{total}}\right] = \frac{1}{T^2}\text{Var}[S_{\text{total}}] \leq \frac{1}{T^2} \cdot m^2 V_s = \text{Var}\left[\frac{1}{T/m}\sum_{t=0}^{T/m-1} Y(t)\right]$$

0.1 Interpretation

This result shows that **thinning actually increases variance**, contrary to intuition. The full chain average has lower variance than the thinned chain average. This happens because:

1. Thinning discards information that could help reduce variance
2. The averaging effect over the full chain provides more variance reduction than the potential decorrelation benefit from thinning
3. In practice, thinning is often used for computational convenience rather than variance reduction

The key insight is that while thinning may reduce autocorrelation, it also reduces the effective sample size, and the latter effect dominates in terms of variance.

Exercise 6

On a product space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, consider a target distribution of density $\pi(x_1, x_2)$. To sample from π, the Gibbs sampler iterately samples from $\pi_{X_1|X_2}(x_1|x_2)$ and $\pi_{X_2|X_1}(x_2|x_1)$. We consider here a scenario where it is possible to sample from $\pi_{X_2|X_1}(x_2|x_1)$ but impossible to sample from $\pi_{X_1|X_2}(x_1|x_2)$. Then the following algorithm may be useful. Note that this is nothing but a standard Metropolis–Hastings algorithm with a cycle of kernels, each updating only one component of the state; but it is commonly referred to as Metropolis-within-Gibbs (MWG).
We introduce a proposal $q(x_1'|x_1, x_2)$ on $\mathcal{X}_1$; i.e. $q(x_1'|x_1, x_2) \geq 0$ and $\int_{\mathcal{X}_1} q(x_1'|x_1, x_2)dx_1' = 1$ for any $(x_1, x_2) \in \mathcal{X}$.
Starting with $X^{(1)} := \left(X_1^{(1)}, X_2^{(1)}\right)$, iterate for $t = 2, 3, \ldots$

- Sample $\tilde{X}_1 \sim q\left(\cdot|X_1^{(t-1)}, X_2^{(t-1)}\right)$.
- Compute $\alpha\left(\tilde{X}_1|X_1^{(t-1)}, X_2^{(t-1)}\right) = \min\left\{1, \frac{\pi\left(\tilde{X}_1, X_2^{(t-1)}\right) q\left(X_1^{(t-1)}|\tilde{X}_1, X_2^{(t-1)}\right)}{\pi\left(X_1^{(t-1)}, X_2^{(t-1)}\right) q\left(\tilde{X}_1|X_1^{(t-1)}, X_2^{(t-1)}\right)}\right\}$.
- With probability $\alpha\left(\tilde{X}_1|X_1^{(t-1)}, X_2^{(t-1)}\right)$, set $X_1^{(t)} = \tilde{X}_1$, otherwise set $X_1^{(t)} = X_1^{(t-1)}$.
- Sample $X_2^{(t)} \sim \pi_{X_2|X_1}\left(\cdot|X_1^{(t)}\right)$.

1. Show that when $q(x_1'|x_1, x_2) = \pi_{X_1|X_2}(x_1'|x_2)$ then the MWG corresponds to the systematic scan Gibbs sampler.
2. State the transition kernel corresponding to this algorithm and show that it has invariant distribution π.

Question 1: MWG reduces to systematic scan Gibbs sampler

When $q(x_1'|x_1, x_2) = \pi_{X_1|X_2}(x_1'|x_2)$, the acceptance probability becomes:

$$\alpha(X_1|X_1^{(t-1)}, X_2^{(t-1)}) = \min\left\{1, \frac{\pi(X_1, X_2^{(t-1)}) q(X_1^{(t-1)}|X_1, X_2^{(t-1)})}{\pi(X_1^{(t-1)}, X_2^{(t-1)}) q(X_1|X_1^{(t-1)}, X_2^{(t-1)})}\right\}$$

Substituting $q(x_1'|x_1, x_2) = \pi_{X_1|X_2}(x_1'|x_2)$:

$$\alpha(X_1|X_1^{(t-1)}, X_2^{(t-1)}) = \min\left\{1, \frac{\pi(X_1, X_2^{(t-1)}) \pi_{X_1|X_2}(X_1^{(t-1)}|X_2^{(t-1)})}{\pi(X_1^{(t-1)}, X_2^{(t-1)}) \pi_{X_1|X_2}(X_1|X_2^{(t-1)})}\right\}$$

Using $\pi(x_1, x_2) = \pi_{X_1|X_2}(x_1|x_2)\pi_{X_2}(x_2)$:

$$\begin{aligned}
&= \min\left\{1, \frac{\pi_{X_1|X_2}(X_1|X_2^{(t-1)}) \pi_{X_2}(X_2^{(t-1)}) \pi_{X_1|X_2}(X_1^{(t-1)}|X_2^{(t-1)})}{\pi_{X_1|X_2}(X_1^{(t-1)}|X_2^{(t-1)}) \pi_{X_2}(X_2^{(t-1)}) \pi_{X_1|X_2}(X_1|X_2^{(t-1)})}\right\} \\
&= \min\{1, 1\} = 1
\end{aligned}$$

Since the acceptance probability is always 1, every proposed move is accepted. This means:

- $X_1^{(t)}$ is always sampled from $\pi_{X_1|X_2}(\cdot|X_2^{(t-1)})$
- $X_2^{(t)}$ is sampled from $\pi_{X_2|X_1}(\cdot|X_1^{(t)})$

This is exactly the systematic scan Gibbs sampler! $\square$

Question 2: Transition kernel and invariant distribution

The transition kernel $K((x_1, x_2), (x_1', x_2'))$ can be written as:

$$K((x_1, x_2), (x_1', x_2')) = K_{MH}((x_1, x_2), (x_1', x_2)) \cdot \pi_{X_2|X_1}(x_2'|x_1')$$

where K_{MH} is the standard Metropolis-Hastings kernel:

$$\begin{aligned} K_{MH}((x_1, x_2), (x_1', x_2)) = q(x_1'|x_1, x_2)\alpha(x_1'|x_1, x_2) \\ + \delta_{x_1}(x_1') \left(1 - \int_{X_1} q(u_1|x_1, x_2)\alpha(u_1|x_1, x_2)du_1\right) \end{aligned}$$

Proof that π is invariant:
The MWG algorithm consists of two sequential steps:

1. Metropolis-Hastings update for X_1 (keeping X_2 fixed)
2. Gibbs update for X_2 (given the new X_1)

Step 1 preserves π: For any fixed value x_2, the MH algorithm with proposal $q(\cdot|x_1, x_2)$ and the given acceptance probability is designed to have stationary distribution $\pi_{X_1|X_2}(\cdot|x_2)$.
Since we start with $X_1^{(t-1)}|X_2^{(t-1)} \sim \pi_{X_1|X_2}(\cdot|X_2^{(t-1)})$, after the MH step we still have $X_1^*|X_2^{(t-1)} \sim \pi_{X_1|X_2}(\cdot|X_2^{(t-1)})$.
Therefore $(X_1^*, X_2^{(t-1)}) \sim \pi$.
Step 2 preserves π: Starting with $(X_1^*, X_2^{(t-1)}) \sim \pi$, we sample $X_2^{(t)} \sim \pi_{X_2|X_1}(\cdot|X_1^*)$. By the definition of conditional distributions, this gives $(X_1^*, X_2^{(t)}) \sim \pi$.
Therefore, starting with a sample from π and applying one iteration of MWG gives another sample from π, proving that π is the invariant distribution. $\square$

Exercise 7

Let $\mathcal{X}$ be a finite state-space. We consider the following Markov transition kernel

$$T(x,y) = \alpha(x,y)q(x,y) + \left(1 - \sum_{z\in\mathcal{X}} \alpha(x,z)q(x,z)\right)\delta_x(y)$$

where $q(x,y) \geq 0$, $\sum_{y\in\mathcal{X}} q(x,y) = 1$ and $0 \leq \alpha(x,y) \leq 1$ for any $x, y \in \mathcal{X}$. $\delta_x(y)$ is the Kronecker symbol; i.e. $\delta_x(y) = 1$ if $y = x$ and zero otherwise.

1. Let π be a probability mass function on $\mathcal{X}$. Show that if

   $$\alpha(x,y) = \frac{\gamma(x,y)}{\pi(x)q(x,y)}$$

   where $\gamma(x,y) = \gamma(y,x)$ and $\gamma(x,y)$ is chosen such that $0 \leq \alpha(x,y) \leq 1$ for any $x, y \in \mathcal{X}$ then T is π-reversible.

2. Verify that the Metropolis-Hastings algorithm corresponds to $\gamma(x,y) = \min\{\pi(x)q(x,y), \pi(y)q(y,x)\}$. The Baker algorithm is an alternative corresponding to

   $$\gamma(x,y) = \frac{\pi(x)q(x,y)\pi(y)q(y,x)}{\pi(x)q(x,y) + \pi(y)q(y,x)}.$$

   Give the associated acceptance probability $\alpha(x,y)$ for the Baker algorithm.

3. Peskun's theorem (1973) is a very important result in the MCMC literature which states the following.

   Theorem: Let T_1 and T_2 be two reversible, aperiodic and irreducible Markov transition kernels w.r.t π. If

   $$T_1(x,y) \geq T_2(x,y), \text{ for all } x \neq y \in \mathcal{X}$$

   then, for all functions $\phi : \mathcal{X} \to \mathbb{R}$, the asymptotic variance of MCMC estimators $\hat{I}_n(\phi) = \frac{1}{n}\sum_{t=0}^{n-1} \phi\left(X^{(t)}\right)$ of $I(\phi) = E_\pi[\phi(X)]$ is smaller for T_1 than T_2.

   Assume that you are in a scenario where both Metropolis-Hastings and Baker algorithms yield aperiodic and irreducible Markov chains. Which algorithm provides estimators of $I(\phi)$ with the lowest asymptotic variance?

4. Suppose that $X = (X_1, \ldots, X_d)$ where X_i takes $m \geq 2$ possible values and $\pi(x) = \pi(x_1, \ldots, x_d)$ is the distribution of interest. The random scan Gibbs sampler proceeds as follows.

   Random scan Gibbs sampler. Let $\{X_1^{(1)}, \ldots, X_d^{(1)}\}$ be the initial state then iterate for $t = 2, 3, \ldots$

   - Sample an index K uniformly on $\{1, \ldots, d\}$.
   - Set $X_i^{(t)} := X_i^{(t-1)}$ for $i \neq K$ and sample $X_K^{(t)} \sim \pi_{X_K|X_{-K}}\left(\cdot|X_1^{(t)}, \ldots, X_{K-1}^{(t)}, X_{K+1}^{(t)}, \ldots, X_d^{(t)}\right)$.

   Consider now a modified random scan Gibbs sampler where instead of sampling $X_K^{(t)}$ from its conditional distribution, we use the following proposal

   $$q(X_K = x_K^*|x_{-K}, x_K) = \begin{cases} \frac{\pi_{X_K|X_{-K}}(x_K^*|x_{-K})}{1-\pi_{X_K|X_{-K}}(x_K|x_{-K})} & \text{for } x_K^* \neq x_K \\ 0 & \text{otherwise} \end{cases}$$

   where $x_{-K} := (x_1, \ldots, x_{K-1}, x_{K+1}, \ldots, x_d)$ which is accepted with probability

   $$\alpha(x_{-K}, x_K, x_K^*) = \min\left\{1, \frac{1 - \pi_{X_K|X_{-K}}(x_K|x_{-K})}{1 - \pi_{X_K|X_{-K}}(x_K^*|x_{-K})}\right\}.$$

Modified random scan Gibbs sampler. Let $\{X_1^{(0)}, \ldots, X_d^{(0)}\}$ be the initial state then iterate for $t = 2, 3, \ldots$

- Sample an index K uniformly on $\{1, \ldots, d\}$.
- Set $X_i^{(t)} := X_i^{(t-1)}$ for $i \neq K$.
- Sample $\tilde{X}_K$ such that $P(\tilde{X}_K = x_K^*) = q\left(\tilde{X}_K^* = x_K^* | X_{-K}^{(t)}, X_K^{(t-1)}\right)$.
- With probability $\alpha\left(X_{-K}^{(t)}, X_K^{(t-1)}, \tilde{X}_K\right)$, set $X_K^{(t)} = \tilde{X}_K^*$ and $X_K^{(t)} = X_K^{(t-1)}$ otherwise.

Assume that both algorithms provide an irreducible and aperiodic Markov chain. Check that both transition kernels are π-reversible and use Peskun's theorem to show that the modified random scan Gibbs sampler provides estimators of $I(\phi)$ with a lower asymptotic variance than the standard random scan Gibbs sampler.

Question 1: Show T is π-reversible

We need to show that $\pi(x)T(x,y) = \pi(y)T(y,x)$ for all $x, y \in X$.
Given:

- $\alpha(x,y) = \frac{\gamma(x,y)}{\pi(x)q(x,y)}$
- $\gamma(x,y) = \gamma(y,x)$ (symmetry condition)
- $0 \leq \alpha(x,y) \leq 1$

The transition kernel is:

$$T(x,y) = \alpha(x,y)q(x,y) + \left(1 - \sum_{z \in X} \alpha(x,z)q(x,z)\right)\delta_x(y)$$

Case 1: $x \neq y$
In this case, $\delta_x(y) = 0$, so:

$$T(x,y) = \alpha(x,y)q(x,y) = \frac{\gamma(x,y)}{\pi(x)q(x,y)} \cdot q(x,y) = \frac{\gamma(x,y)}{\pi(x)}$$

Similarly:

$$T(y,x) = \alpha(y,x)q(y,x) = \frac{\gamma(y,x)}{\pi(y)q(y,x)} \cdot q(y,x) = \frac{\gamma(y,x)}{\pi(y)}$$

Now checking detailed balance:

$$\pi(x)T(x,y) = \pi(x) \cdot \frac{\gamma(x,y)}{\pi(x)} = \gamma(x,y)$$
$$\pi(y)T(y,x) = \pi(y) \cdot \frac{\gamma(y,x)}{\pi(y)} = \gamma(y,x)$$

Since $\gamma(x,y) = \gamma(y,x)$, we have:

$$\pi(x)T(x,y) = \pi(y)T(y,x)$$

Case 2: $x = y$
For $x = y$, the detailed balance condition $\pi(x)T(x,x) = \pi(x)T(x,x)$ is trivially satisfied.
Therefore, T is π-reversible. $\square$

Question 2: Verify Metropolis-Hastings and find Baker acceptance probability

Part (a): Metropolis-Hastings

For Metropolis-Hastings, we have:

$$\gamma(x, y) = \min\{\pi(x)q(x, y), \pi(y)q(y, x)\}$$

The acceptance probability is:

$$\alpha(x, y) = \frac{\gamma(x, y)}{\pi(x)q(x, y)} = \frac{\min\{\pi(x)q(x, y), \pi(y)q(y, x)\}}{\pi(x)q(x, y)}$$

Case 1: If $\pi(x)q(x, y) \leq \pi(y)q(y, x)$, then:

$$\alpha(x, y) = \frac{\pi(x)q(x, y)}{\pi(x)q(x, y)} = 1$$

Case 2: If $\pi(x)q(x, y) > \pi(y)q(y, x)$, then:

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

Combining both cases:

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}$$

This is exactly the standard Metropolis-Hastings acceptance probability! ✓

Part (b): Baker algorithm

For the Baker algorithm:

$$\gamma(x, y) = \frac{\pi(x)q(x, y)\pi(y)q(y, x)}{\pi(x)q(x, y) + \pi(y)q(y, x)}$$

The acceptance probability is:

$$\begin{aligned}
\alpha(x, y) &= \frac{\gamma(x, y)}{\pi(x)q(x, y)} \\
&= \frac{\pi(x)q(x, y)\pi(y)q(y, x)}{\pi(x)q(x, y)[\pi(x)q(x, y) + \pi(y)q(y, x)]} \\
&= \frac{\pi(y)q(y, x)}{\pi(x)q(x, y) + \pi(y)q(y, x)}
\end{aligned}$$

Verification of symmetry:

$$\begin{aligned}
\gamma(y, x) &= \frac{\pi(y)q(y, x)\pi(x)q(x, y)}{\pi(y)q(y, x) + \pi(x)q(x, y)} \\
&= \frac{\pi(x)q(x, y)\pi(y)q(y, x)}{\pi(x)q(x, y) + \pi(y)q(y, x)} = \gamma(x, y)
\end{aligned}$$

Therefore, the Baker algorithm acceptance probability is:

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y) + \pi(y)q(y, x)}$$