Computational statistics 2021.2
School of Applied Mathematics, Fundação Getulio Vargas
Professor Luiz Max de Carvalho

Problem sheet 3

Lucas Machado Moschen

Exercício 1.

Exercício 2.

Exercício 3. *(Metropolis-Hastings and Gibbs Sampler) Let $\mathbb{X}$ be a finite state-space. We consider the following Markov transition kernel*

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

where $q(x,y) \geq 0$, $\sum_{y\in\mathbb{X}} q(x,y) = 1$ and $0 \leq \alpha(x,y) \leq 1$ for any $x, y \in \mathbb{X}$, $\delta_x(y)$ is the Kronecker symbol.

1. *Let π be a probability mass function on $\mathbb{X}$. 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 $x, y \in \mathbb{X}$ then T is π-reversible.

By Proposition 2.3 from the notes, we have to show that π satisfies detailed balance with respect to T, that is,

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

If $x = y$, this is clearly true. If $x \neq y$, we have that

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

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.

Setting $\gamma(x,y) = \min\{\pi(x)q(x,y), \pi(y)q(y,x)\}$, we will have that

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

exactly as in the Metropolis-Hastings algorithm. Besides that, it is clearly that $\gamma(x,y) = \gamma(y,x)$.

We will have for Baker algorithm that

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

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 $T_1(x,y) \geq T_2(x,y)$, for all $x \neq y \in \mathbb{X}$ then, for all funtions $\phi : \mathbb{X} \to \mathbb{R}$, the asymptotic variance of MCMC estimators $\hat{I}_n(\phi) = \frac{1}{n}\sum_{t=0}^{n-1}\phi(X^{(t)})$ of $I(\phi) = \mathbb{E}_\pi[\phi(X)]$ 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?

The ideia is to use Peskun's theorem. Therefore, we compute both transition kernels. For the Metropolis-Hastings algorithm,

$$T_1(x,y) = \alpha(x,y)q(x,y) = \frac{\min\{\pi(x)q(x,y), \pi(y)q(y,x)\}}{\pi(x)}.$$

Baker algorithm yields

$$T_2(x,y) = \frac{\gamma(x,y)}{\pi(x)} = \frac{1}{\pi(x)}\frac{\pi(x)q(x,y)\pi(y)q(y,x)}{\pi(x)q(x,y)+\pi(y)q(y,x)}.$$

Suppose that $\pi(x)q(x,y) \leq \pi(y)q(y,x)$. We will have that

$$\begin{aligned}\pi(x)^2q(x,y)^2 \geq 0 \to \pi(x)q(x,y)(\pi(x)q(x,y)+\pi(y)q(y,x)) &\geq \pi(x)q(x,y)\pi(y)q(y,x) \\ \to \pi(x)q(x,y) &\geq \frac{\pi(x)q(x,y)\pi(y)q(y,x)}{\pi(x)q(x,y)+\pi(y)q(y,x)} \\ \to T_1(x,y) &\geq T_2(x,y).\end{aligned}$$

If $\pi(x)q(x,y) \geq \pi(y)q(y,x)$, a similar result is achieved by the symmetry of the problem. Therefore, by Peskun's theorem, Metropolis-Hastings have smaller asymptotic variance.

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

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

   - *Sample an index* K *uniformly on* $\{1, \dots, d\}$
   - *Set* $X_i^{(t)} = X_i^{(t-1)}$ *for* $i \neq K$ *and sample*

   $$X_K^{(t)} \sim \pi_{X_k|X_{-K}}(\cdot \mid X_1^{(t)}, \dots, X_{K-1}^{(t)}, X_{K+1}^{(t)}, \dots, X_d^{(t)}).$$

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

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

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

   $$\alpha(x_{-K}, x_K, x_K^*) = \min\left\{1, \frac{1-\pi(x_K \mid x_{-K})}{1-\pi(x_K^* \mid x_{-K})}\right\}.$$

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

   - *Sample an index* K *uniformly on* $\{1, \dots, d\}$
   - *Set* $X_i^{(t)} = X_i^{(t-1)}$ *for* $i \neq K$
   - *Sample* $X_K^{(t)}$ *such that* $\mathbb{P}(X_K = x_K^*) = q(X_K = x_K^* \mid x_{-K}, x_K)$
   - *With probability* $\alpha(x_{-K}, x_K, x_K^*)$, *set* $X_K^{(t)} = X_K^*$ *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.*

   We have to verify the detailed balance for π. The kernel for random scan Gibbs sampler is for $x_j \neq y_j$ para apenas um j,

   $$T_1(x, y) = d^{-1} \sum_{j=1}^{d} \pi_{X_j|X_{-j}}(y_j \mid x_{-j}) \delta_{X_{-j}}(y_{-j}),$$

   then, we have

   $$\begin{aligned} \pi(x) T_1(x, y) &= d^{-1} \sum_{j=1}^{d} \pi(x) \pi(y_j \mid x_{-j}) \delta_{X_{-j}}(y_{-j}) \\ &= d^{-1} \sum_{j=1}^{d} \pi(x_j \mid x_{-j}) \pi(x_{-j}) \pi(y_j \mid x_{-j}) \delta_{X_{-j}}(y_{-j}) \\ &= d^{-1} \sum_{j=1}^{d} \pi(x_j \mid y_{-j}) \pi(y_j, x_{-j}) \delta_{X_{-j}}(y_{-j}) = \pi(y) T_1(y, x), \end{aligned}$$

which proves that π is invariant for T_1 and the chain is reversible with respect to π. Now we shall derive T_2. We will have the sum over the index to space with probability d^{-1} for every point. Moreover, is $y \neq x$, the transition from x to y is following q distribution and the probability of acceptance α. Therefore,

$$\begin{aligned}
T_2(x,y) &= d^{-1} \sum_{j=1}^{d} q(y_j \mid x_{-j}, x_j)\alpha(x, y_j)\delta_{X_{-j}}(y_{-j}) \\
&= d^{-1} \sum_{j=1}^{d} \frac{\pi_{X_j \mid X_{-j}}(y_j \mid x_{-j})}{1 - \pi(x_j \mid x_{-j})} \min\left\{1, \frac{1 - \pi(x_j \mid x_{-j})}{1 - \pi(y_j \mid x_{-j})}\right\} \delta_{X_{-j}}(y_{-j}) \\
&= d^{-1} \sum_{j=1}^{d} \pi(y_j \mid x_{-j}) \min\left\{\frac{1}{1 - \pi(x_j \mid x_{-j})}, \frac{1}{1 - \pi(y_j \mid x_{-j})}\right\} \delta_{X_{-j}}(y_{-j})
\end{aligned}$$

$$\begin{aligned}
\pi(x)T_2(x,y) &= d^{-1} \sum_{j=1}^{d} \pi(x)\pi(y_j \mid x_{-j}) \\
&\qquad\qquad \times \min\left\{\frac{1}{1 - \pi(x_j \mid x_{-j})}, \frac{1}{1 - \pi(y_j \mid x_{-j})}\right\} \delta_{X_{-j}}(y_{-j}) \\
&= d^{-1} \sum_{j=1}^{d} \pi(x_j \mid y_{-j})\pi(y) \\
&\qquad\qquad \times \min\left\{\frac{1}{1 - \pi(x_j \mid x_{-j})}, \frac{1}{1 - \pi(y_j \mid x_{-j})}\right\} \delta_{X_{-j}}(y_{-j}) \\
&= \pi(y)T_2(y,x),
\end{aligned}$$

which proves the detailed balance relation. Moreover, notice that, since $1 > 1 - \pi(x_j \mid x_{-j}) > 0$, we have that $T_2(x,y) \geq T_1(x,y)$ for every $x \neq y$. We conclude that the Modified random scan Gibbs sampler has lower asymptotic variance.

Exercício 4.

Exercício 5. *(Thinning of a Markov chain)*

1. *Prove the Cauchy-Schwarz inequality which states that for any two real-valued random variables Y and Z,*
$$|\mathbb{E}[YZ]|^2 \leq \mathbb{E}[Y^2]\mathbb{E}[Z^2]$$

   Notice that for any $\alpha \in \mathbb{R}$, $(Y - \alpha Z)^2 \geq 0$ implying that $\mathbb{E}[(Y - \alpha Z)^2] \geq 0$, that is,
$$\mathbb{E}[Y^2] - 2\alpha\mathbb{E}[YZ] + \alpha^2\mathbb{E}[Z^2] \geq 0,$$
   which is a quadratic in α. By this relation, the quadratic can at most reach 0 is one point, that is, there exists at most one solution when this is zero. Because of that, the discriminant is non-positive,
$$4\mathbb{E}[YZ]^2 \leq 4\mathbb{E}[Y^2]\mathbb{E}[Z^2] \implies \mathbb{E}[YZ]^2 \leq \mathbb{E}[Y^2]\mathbb{E}[Z^2].$$

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

Notice that

$$\mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] = \mathbb{E}[YZ] - \mathbb{E}[Y]^2 \leq \sqrt{\mathbb{E}[Y^2]\mathbb{E}[Z^2]} - \mathbb{E}[Y]^2 = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2,$$

which implies that $\mathrm{Cov}(Y, Z) \leq \mathrm{Var}(Y)$.

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*

$$Y^{(t)} = X^{(mt)}.$$

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.

Following the tip, notice that

$$\sum_{t=0}^{T-1} X^{(t)} = \sum_{t=0}^{m-1}\sum_{s=0}^{T/m-1} X^{(sm+t)}.$$

Then,

$$\begin{aligned}
\mathrm{Var}\left[\sum_{t=0}^{T-1} X^{(t)}\right] &= \mathrm{Var}\left[\sum_{t=0}^{m-1}\sum_{s=0}^{T/m-1} X^{(sm+t)}\right] \\
&= \sum_{t=0}^{m-1} \mathrm{Var}\left[\sum_{s=0}^{T/m-1} X^{(sm+t)}\right] \\
&+ \quad \sum_{t=0}^{m-1}\sum_{r\neq t} \mathrm{Cov}\left[\sum_{s=0}^{T/m-1} X^{(sm+t)}, \sum_{s=0}^{T/m-1} X^{(sm+r)}\right] \\
&\leq m\,\mathrm{Var}\left[\sum_{s=0}^{T/m-1} X^{sm}\right] + (m-1)\sum_{t=0}^{m-1}\mathrm{Var}\left[\sum_{s=0}^{T/m-1} X^{(sm+t)}\right] \\
&\leq m\,\mathrm{Var}\left[\sum_{s=0}^{T/m-1} X^{sm}\right] + m(m-1)\,\mathrm{Var}\left[\sum_{s=0}^{T/m-1} X^{sm}\right] \\
&= m^2\,\mathrm{Var}\left[\sum_{s=0}^{T/m-1} X^{sm}\right] = \mathrm{Var}\left[m\sum_{t=0}^{T/m-1} Y^{(t)}\right].
\end{aligned}$$

Dividing each side of the inequality per $1/T^2$ yields the result.

Exercício 6. *(Simulation question - Paper sheet 4 (Reversible jump MCMC)) Consider two models. For model 1 the toy target distribution is given*

$$\pi(\theta \mid k = 1) = \exp(-\theta^2/2)$$

whereas for model 2 it is given By

$$\pi(\theta \mid k = 2) = \exp(-\theta_1^2/2 - \theta_2^2/2)$$

We want to design a transdimensional sampler to sample from the distribution of (k, θ).

- *Implement standard Metropolis-Hastings kernels K1 for model 1 and K2 for model 2. Check that they work before going further.*
- *Implement trans-dimensional moves to go from model 1 to model 2. That is, for* $\theta \in \mathbb{R}$, *propose an auxiliary variable* $u \in \mathbb{R}$ *following the distribution of your choice and a deterministic mapping* $G_{1\to2}(\theta, u)$ *to obtain a point in* $\mathbb{R}^2$ *which you will then accept or reject with the appropriate acceptance probability.*
- *Implement trans-dimensional moves to go from model 2 to model 1. That is, for* $\theta \in \mathbb{R}^2$, *propose a deterministic mapping* $G_{2\to1}(\theta)$ *to obtain a point in* $\mathbb{R}$ *which you will then accept or reject with the appropriate acceptance probability.*
- *Put these kernels together to obtain a valid Reversible Jump algorithm. What is the proportion of visits to each model? What should it be in the limit of the number of iterations?*