Advanced Simulation

Problem Sheet 3

Exercise 1 (Gibbs Sampler)

Let $\pi_{X,Y}(x,y)$ be the density of a distribution of interest. We recall that the systematic scan Gibbs sampler proceeds as follows to sample from $\pi_{X,Y}$.

Systematic Scan Gibbs sampler. Let $X^{(1)}, Y^{(1)}$ be the initial state then iterate for $t = 2, 3, \ldots$

- Sample $Y^{(t)} \sim \pi_{Y|X}\left(\cdot \mid X^{(t-1)}\right)$.
- Sample $X^{(t)} \sim \pi_{X|Y}\left(\cdot \mid Y^{(t)}\right)$.

The random scan Gibbs sampler is an alternative algorithm which proceeds as follows to sample from $\pi_{X,Y}$.

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

- Sample $J \in \{1, 2\}$ where $\mathbb{P}(J=1) = \mathbb{P}(J=2) = 1/2$.
- If $J = 1$, sample $Y^{(t)} \sim \pi_{Y|X}\left(\cdot \mid X^{(t-1)}\right)$ and set $X^{(t)} = X^{(t-1)}$.
- If $J = 2$, Sample $X^{(t)} \sim \pi_{X|Y}\left(\cdot \mid Y^{(t-1)}\right)$ and set $Y^{(t)} = Y^{(t-1)}$.

1. Give the expression of the transition kernel density $K^{\mathrm{S}}_{X,Y}\left((x,y),(x',y')\right)$ of the Markov chain $\left(X^{(t)}, Y^{(t)}\right)_{t\geq 1}$ generated by the systematic Gibbs sampler as a function of $\pi_{X|Y}$ and $\pi_{Y|X}$. Show that $K^{\mathrm{S}}_{X,Y}$ is *not* reversible with respect to $\pi_{X,Y}$.

2. Show that the sequence $\left(X^{(t)}\right)_{t\geq 1}$ associated to the systematic scan Gibbs sampler is a π_X-reversible Markov chain and give the expression of its associated transition kernel density $K^{\mathrm{S}}_{X}(x,x')$ as a function of the two "full" conditional densities $\pi_{Y|X}$ and $\pi_{Y|X}$.

3. Give the expression of the transition kernel density $K^{\mathrm{R}}_{X,Y}\left((x,y),(x',y')\right)$ of the Markov chain $\left(X^{(t)}, Y^{(t)}\right)_{t\geq 1}$ generated by the random scan Gibbs sampler as a function of $\pi_{X|Y}$ and $\pi_{Y|X}$. Show that $K^{\mathrm{R}}_{X,Y}$ is $\pi_{X,Y}-$reversible.

Exercise 2 (Metropolis-within-Gibbs)

On a product space $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$, consider a target distribution of density $\pi(x_1, x_2)$. To sample from π, the Gibbs sampler iterately samples from $\pi_{X_1|X_2}(x_1 \mid x_2)$ and $\pi_{X_2|X_1}(x_2 \mid x_1)$. We consider here a scenario where it is possible to sample from $\pi_{X_2|X_1}(x_2 \mid x_1)$ but impossible to sample from $\pi_{X_1|X_2}(x_1 \mid x_2)$. 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 $q(x_1' \mid x_1, x_2)$ on $\mathbb{X}_1$; i.e. $q(x_1' \mid x_1, x_2) \geq 0$ and $\int_{\mathbb{X}_1} q(x_1' \mid x_1, x_2)\, dx_1' = 1$ for any $(x_1, x_2) \in \mathbb{X}$.

Starting with $X^{(1)} := \left(X_1^{(1)}, X_2^{(1)}\right)$, iterate for $t = 2, 3, \ldots$

- Sample $X_1 \sim q\left(\cdot \mid X_1^{(t-1)}, X_2^{(t-1)}\right)$.

- Compute $\alpha\left(X_1 \mid X_1^{(t-1)}, X_2^{(t-1)}\right) = \min\left\{1, \dfrac{\pi\left(X_1, X_2^{(t-1)}\right) q\left(X_1^{(t-1)} \middle| X_1, X_2^{(t-1)}\right)}{\pi\left(X_1^{(t-1)}, X_2^{(t-1)}\right) q\left(X_1 \middle| X_1^{(t-1)}, X_2^{(t-1)}\right)}\right\}$.

- With probability $\alpha\left(X_1 \mid X_1^{(t-1)}, X_2^{(t-1)}\right)$, set $X_1^{(t)} = X_1$, otherwise set $X_1^{(t)} = X_1^{(t-1)}$.

- Sample $X_2^{(t)} \sim \pi_{X_2|X_1}\left(\cdot \mid X_1^{(t)}\right)$.

1. Show that when $q\left(x_1'\middle| x_1, x_2\right) = \pi_{X_1|X_2}\left(x_1'\middle| x_2\right)$ 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.

Exercise 3 (Metropolis-Hastings and Gibbs Sampler)

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

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

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; i.e. $\delta_x(y) = 1$ if $y = x$ and zero otherwise.

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 $\pi-$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

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

   then, for all functions $\phi : \mathbb{X} \to \mathbb{R}$, the asymptotic variance of MCMC estimators $\widehat{I}_n(\phi) = \frac{1}{n}\sum_{t=0}^{n-1} \phi\left(X^{(t)}\right)$ 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?

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

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

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

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

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

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

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

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

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

Exercise 4 (Metropolis-Hastings)

Consider a target distribution on $\mathbb{X} = \mathbb{R}^d$ of density $\pi(x)$. We propose to sample from it using a Metropolis-Hastings algorithm based on a random proposal. Given $X^{(t-1)}$, the proposal X is sampled as follows: $X^{(t-1)}$ is the input to a stochastic optimization procedure that returns an estimate $\vartheta \in \mathbb{X}$ of the local maximiser of $\pi(x)$ in the vicinity of $X^{(t-1)}$. The random variable ϑ follows an unknown probability density function $f\left(\theta \middle| X^{(t-1)}\right)$. We then sample $X \sim g(\cdot|\vartheta)$ where $g(\cdot|\vartheta)$ is a multivariate normal of mean ϑ and fixed covariance Σ.

1. Express the proposal $q(x'|x)$ as a function of f and g. Is it possible to evaluate the acceptance probability $\alpha_{MH}(x'|x) = \min\left\{1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right\}$ associated to this proposal?
2. Consider now the following randomized Metropolis-Hastings algorithm.

   Set $X^{(1)} = x, \vartheta^{(1)} = \theta$ then iterate for $t = 2, 3, ...$

   (a) Sample $X \sim g\left(\cdot \middle| \vartheta^{(t-1)}\right)$ and $\vartheta \sim f(\cdot|X)$.

   (b) Compute
   $$\alpha_{RMH}\left\{(X,\vartheta)\middle|\left(X^{(t-1)},\vartheta^{(t-1)}\right)\right\} = \min\left\{1, \frac{\pi(X)\,g\left(X^{(t-1)}\middle|\vartheta\right)}{\pi\left(X^{(t-1)}\right)g\left(X\middle|\vartheta^{(t-1)}\right)}\right\}.$$

   (c) With probability $\alpha_{RMH}\left\{(X,\vartheta)\middle|\left(X^{(t-1)},\vartheta^{(t-1)}\right)\right\}$, set $\left(X^{(t)},\vartheta^{(t)}\right) = (X,\vartheta)$ and otherwise set $\left(X^{(t)},\vartheta^{(t)}\right) = \left(X^{(t-1)},\vartheta^{(t-1)}\right)$.

   Show that the transition kernel associated to the above algorithm admits an invariant distribution of density $\overline{\pi}(x,\theta)$ such that $\overline{\pi}(x) = \pi(x)$.
3. Prove that for any real-valued random variables U, V, we have
   $$\mathbb{E}(\min\{U,V\}) \leq \min(\mathbb{E}\{U\}, \mathbb{E}\{V\}).$$
4. Using the result of (3), prove that the expected acceptance probability $\alpha_{RMH}\left\{(X,\vartheta)\middle|\left(X^{(t-1)},\vartheta^{(t-1)}\right)\right\}$ of the randomized Metropolis-Hastings at stationarity is smaller than the expected acceptance probability $\alpha_{MH}\left\{X\middle|X^{(t-1)}\right\}$ of the "ideal" Metropolis-Hastings at stationarity.

Exercise 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,
   $$|\mathbb{E}[YZ]|^2 \leq \mathbb{E}\left[Y^2\right]\mathbb{E}\left[Z^2\right].$$
   (Hint: $(Y-\alpha Z)^2 \geq 0$ *for any* $\alpha \in R$*).*

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

3. Thinning of a Markov chain $\left\{X^{(t)}\right\}_{t\geq 0}$ 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 $\left\{Y^{(t)}\right\}_{t\geq 0}$ defined by
$$Y^{(t)} := X^{(m.t)}.$$
We might hope that $\left\{Y^{(t)}\right\}_{t\geq 0}$ will exhibit lower autocorrelation than the original chain $\left\{X^{(t)}\right\}_{t\geq 0}$ and thus will yield ergodic averages of lower variance.

Consider a stationary Markov chain $\left\{X^{(t)}\right\}_{t\geq 0}$. Let T and m be any two integers such that $T \geq m > 1$ and $T/m \in \mathbb{N}$. Show that
$$\mathbb{V}\mathrm{ar}\left[\frac{1}{T}\sum_{t=0}^{T-1} X^{(t)}\right] \leq \mathbb{V}\mathrm{ar}\left[\frac{1}{T/m}\sum_{t=0}^{T/m-1} Y^{(t)}\right]$$
and briefly explain what this result tells us about the use of thinning.

(Hint: start by writing $\sum_{t=0}^{T-1} X^{(t)} = \sum_{t=0}^{m-1}\sum_{s=0}^{T/m-1} X^{(s.m+t)}$)

Simulation question (Probit model — Gibbs and M-H)

Suppose our dataset is made of binary observations $Y_1,\ldots,Y_n$. For instance Y_i is 1 if student "i" has passed the exam and 0 otherwise. Assume we know p covariates about the students, such as the time spent studying, the number of classes he attended, the ability to cheat without getting caught, etc. We call the covariates "explanatory variables" and store them in a matrix X of size $n \times p$. The *probit model* states that for each $i = 1,\ldots,n$,
$$Y_i = \begin{cases} 1 \text{ with probability } \Phi(X_i^T\beta) \\ 0 \text{ with probability } 1-\Phi(X_i^T\beta) \end{cases}$$
where X_i is the i-th row of X, Φ is the distribution function of a standard Normal distribution, and $\beta \in \mathbb{R}^p$ is the parameter to infer. Inferring β allows to learn and quantify the effect of each covariate on the observation.

1. Generate a synthetic dataset Y from the probit model for an arbitrary value of β and an matrix X.

   (Hint: choose p = 2 and n small, say 50, to make things easier.)

2. Introduce the prior distribution on β:
$$\pi(\beta) = \mathcal{N}(0,B)$$
for a $p\times p$ covariance matrix B. Write a function taking a vector β as argument and returning the log posterior density function evaluated at β.

3. Use it to run a Metropolis-Hastings algorithm and plot the output.

4. For all $i = 1,\ldots,n$, introduce the random variable Z_i distributed as $\mathcal{N}(X_i^T\beta,1)$. Compare the law of $1_{Z_i\geq 0}$ with the law of Y_i.

5. Use Z to design a Gibbs sampler, alternatively sampling from β given Z,Y and from Z given β,Y.

6. Compare the performance of your Gibbs and Metropolis-Hastings samplers.