Advanced Simulation

Problem Sheet 3

Exercise 1

Consider the following $\mathbb{X}$-valued Markov chain $(X_t)_{t\geq 1}$. It evolves over time as follows. At time t, with probability $\alpha(X_{t-1})$ sample

$$X_t \sim q(\cdot)$$

where $q(x)$ is a probability density function and otherwise set $X_t := X_{t-1}$. Hence its transition kernel is given by

$$K(x,y) = \alpha(x)\, q(y) + (1-\alpha(x))\, \delta_x(y)$$

where $\delta_x(y)$ is the Dirac mass located at x.

1. Show that if
$$\int_{\mathbb{X}} \frac{q(x)}{\alpha(x)} dx < \infty$$
then K admits a stationary distribution of density
$$\pi(x) \propto \frac{q(x)}{\alpha(x)}.$$

2. Assume that $0 \leq \alpha(x) = \alpha < 1$ then it can be easily shown that a central limit theorem holds for $\frac{1}{t}\sum_{i=1}^{t} X_i$ as long as $\sigma^2 := \mathbb{V}_q[X_1] < \infty$. Compute the asymptotic variance $\sigma_X^2 = \mathbb{V}[X_1] + 2\sum_{k=2}^{\infty} \mathbb{C}ov[X_1, X_k]$ in the stationary regime as a function of α and σ^2.
*(**Hint.** First prove that the marginal distribution of X_k is q for all k, then find a recursion formula for $\mathbb{C}ov(X_1, X_k)$.)*

Exercise 2

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?

Exercise 3: (Optional)

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

$$X_i \sim \mathcal{B}inomial(m_i, \theta_1),\ Y_i \sim \mathcal{B}inomial(n_i, \theta_2).$$

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 $\vartheta_1 \sim \mathcal{U}_{[0,1]}$, $\vartheta_2 \sim \mathcal{U}_{[0,1]}$. 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 $f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}\mathbb{I}_{[0,1]}(x)$.
(Hint: introduce auxiliary variables)

Exercise 4: 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 $t = 2, 3, ...$

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

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 $t = 2, 3, ...$

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

1. Give the expression of the transition kernel density $K^{\mathrm{S}}_{X,Y}((x,y),(x',y'))$ 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 $K^{\mathrm{R}}_{X,Y}((x,y),(x',y'))$ 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 5: ν-irreducibility of HMC (Optional)

Let $\nu(\boldsymbol{q}) \propto \exp(-U(\boldsymbol{q}))$ be the d-dimensional target density of HMC. In this exercise, we are going to show that HMC with leapfrog steps and randomized step size ϵ is ν-irreducible. The goal of the exercise is to show the following proposition.

Proposition 1 *Let ϵ be distributed uniformly on an interval $[0,\tau]$ for some $\tau > 0$. Let the number of steps L be fixed. Let $\boldsymbol{K}$ denote the Markov kernel for the position variables on $\mathbb{R}^d$ corresponding to sampling a random momentum $\boldsymbol{p}$ from $N(0,\boldsymbol{M})$, then doing L Leapfrog steps started at $(\boldsymbol{q},\boldsymbol{p})$ with step size ϵ sampled from $[0,\tau]$ uniformly (independently before each sequence of L leapfrog steps, but constant during the sequence of L steps), and finally discarding the momentum variable. Suppose that U is continuously differentiable on $\mathbb{R}^d$, and satisfies that $\sup_{\boldsymbol{q}} \|\nabla^2 U(\boldsymbol{q})\| \leq L_U$, and $U_{\min} := \inf_{\boldsymbol{q}\in\mathbb{R}^d} U(\boldsymbol{q}) > -\infty$. Then $\boldsymbol{K}$ is strongly ν-irreducible.*

The proof of this result is similar to the continuous time case. It consists of the following steps.

1. Assume that $\boldsymbol{M} = \boldsymbol{I}_d$ (the general case follows similarly). Let Ψ_ϵ denote the Leapfrog map as defined in the lectures. Let $(\boldsymbol{q}(L\epsilon),\boldsymbol{p}(L\epsilon)) = \Psi^L_\epsilon(\boldsymbol{q}(0),\boldsymbol{p}(0))$ denote the new position and momentum after L leapfrog steps started from $(\boldsymbol{q}(0),\boldsymbol{p}(0))$. Show that if $\epsilon L \leq 1/(4(1+L_U)^2)$, then the Jacobian $\frac{\partial \boldsymbol{q}(L\epsilon)}{\partial \boldsymbol{p}(0)}$ satisfies that
$$\frac{\epsilon L \boldsymbol{I}_d}{2} \preceq \frac{\partial \boldsymbol{q}(L\epsilon)}{\partial \boldsymbol{p}(0)} \preceq \frac{3\epsilon L \boldsymbol{I}_d}{2}.$$
(Hint: the Jacobian of a product of maps is the product of the Jacobians of each map.)

2. Let $H(\boldsymbol{q},\boldsymbol{p}) = U(\boldsymbol{q}) - U_{\min} + \frac{\|\boldsymbol{p}\|^2}{2}$. Show that there is a constant C_H only depending on L, L_U and τ such that for any $0 \leq k \leq L$,
$$H(\boldsymbol{q}(k\epsilon),\boldsymbol{p}(k\epsilon)) - H(\boldsymbol{q},\boldsymbol{p}) \leq C_H L\epsilon \cdot H(\boldsymbol{q},\boldsymbol{p}). \tag{1}$$
Hence the Hamiltonian is approximately preserved for sufficiently short intervals. (Hint: use Taylor expansion with remainder term and the assumption $\sup_{\boldsymbol{q}} \|\nabla^2 U(\boldsymbol{q})\| \leq L_U$ to control the change of the Hamiltonian).

3. Let $\boldsymbol{Q}, \boldsymbol{q}(0) \in \mathbb{R}^d$ be arbitrary points, and let $\boldsymbol{p}(0) = \frac{\boldsymbol{Q}-\boldsymbol{q}(0)}{L\epsilon}$. Show that $(\boldsymbol{q}(L\epsilon), \boldsymbol{p}(L\epsilon)) = \Psi_\epsilon^L(\boldsymbol{q}(0), \boldsymbol{p}(0))$ satisfies that

$$\|\boldsymbol{q}(L\epsilon) - \boldsymbol{Q}\| \leq C \cdot (L\epsilon)^2,$$

where C is a constant only depending on L, L_U, τ, $\boldsymbol{q}(0)$, $\boldsymbol{Q}$ and U but independent of ϵ (Hint: use the approximate conservation of the Hamiltonian by (1) and try to do a similar argument as in Lemma 2 of Chapter 7).

4. Using the above results, show Proposition 1 (Hint: the argument is similar to the continuous time case we have studied during the lecture).

Exercise 6: Sequential Importance Sampling

In this example we will carefully study the phenomenon of weight degeneracy of the sequential importance sampling algorithm (SIS) in a simplified situation where the exponential growth of the asymptotic variance with the time-horizon is easy to capture. This in turn provides motivation for the use of a resampling step that results in the sequential importance resampling algorithm (SIR).

Let $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$ be a fixed distribution on $\mathbb{R}$, with density $\pi(\cdot)$ known up to a normalizing constant, that is $\pi(x) = \tilde{\pi}(x)/Z$, where $\tilde{\pi}(x)$ can be evaluated for every x whereas Z is unknown.

Consider a sequence of target distributions of increasing dimension

$$\pi_n(\mathrm{d}x_{1:n}) = \pi^{\otimes n}(\mathrm{d}x_{1:n}) = \prod_{i=1}^n \pi(\mathrm{d}x_i) = \frac{1}{Z_n}\prod_{i=1}^n \tilde{\pi}(\mathrm{d}x_i),$$

where of course the normalizing constant $Z_n = Z^n$ is unknown. Let $\nu(\mathrm{d}x) = \nu(x)\mathrm{d}x$, with known density $\nu(\cdot)$, be another distribution on $\mathbb{R}$, absolutely continuous wrt π and similarly define $\nu_n := \nu^{\otimes n}$.

We want to estimate the unknown normalizing constant of π_n, that is we want to estimate Z_n. One approach, perhaps a bit contrived, is to use Sequential Importance Sampling to construct an estimate $\hat{Z}_n^N$ as follows: sample N, i.i.d. samples $X_{1:n}^{(i)}$ from $\nu^{\otimes n}$ and compute

$$\hat{Z}_n^N := \frac{1}{N}\sum_{j=1}^N \frac{\tilde{\pi}^{\otimes n}\left(X_{1:n}^{(j)}\right)}{\nu^{\otimes n}\left(X_{1:n}^{(j)}\right)} = \frac{1}{N}\sum_{j=1}^N \prod_{i=1}^n \frac{\tilde{\pi}(X_i^{(j)})}{\nu(X_i^{(j)})}.$$

(a) Show that $\hat{Z}_n^N$ is an unbiased estimator of Z_n.

(b) Show that the variance of $\hat{Z}_n^N$ is given by

$$N \times \mathrm{var}\left(\hat{Z}_n^N\right) = \mathbb{E}_{Y\sim\nu}\left[\frac{\tilde{\pi}(Y)^2}{\nu(Y)^2}\right]^n - Z^{2n} = \mathbb{E}_{Y\sim\nu}\left[\frac{\tilde{\pi}(Y)^2}{\nu(Y)^2}\right]^n - \mathbb{E}_{Y\sim\nu}\left[\frac{\tilde{\pi}(Y)}{\nu(Y)}\right]^{2n}.$$

(c) The *relative variance*, that is the variance of $\hat{Z}_n^N/Z_n$ is a useful measure of the efficiency of an estimator, as it measures the variability of an estimator relative to the size of the quantity being estimated. Show that if $\tilde{\pi}/\nu$ is not almost everywhere constant we have

$$\lim_{n\to\infty}\frac{1}{n}\log\left[\mathrm{var}\left(\frac{\hat{Z}_n^N}{Z_n}\right)\right] = \log\mathbb{E}_{Y\sim\nu}\left[\frac{\pi(Y)^2}{\nu(Y)^2}\right] > 0.$$

Hint: When is Jensen's inequality a strict inequality?

(d) How does the number N of Monte-Carlo samples required to estimate Z_n efficiently depend on the time-horizon n? Is this an efficient algorithm?

Exercise 7: Sequential Importance Sampling 2 (Optional)

All notation is the same as in the previous exercise.

In the previous exercise we proved that the estimator of the normalizing constant produced by sequential importance sampling grows exponentially with the time-horizon, at least in the trivial, i.i.d. scenario. In this exercise we will use the same algorithm, but we will be estimating the expectation of a function of the k-th marginal of the state process. That is, for some function $f : \mathbb{R} \to \mathbb{R}$, that is not constant almost everywhere, we want to estimate the expectation

$$\pi_k(f) := \int \cdots \int \pi_k(\mathrm{d}x_{1:k}) f(x_k).$$

Of course one may point out that $\pi_k(f)$ is simply $\pi(f)$, but the proposed method is a perfectly valid approach and is instructive to study the performance of SIS in this simplified, albeit contrived, scenario as it does capture its performance in more complicated models in the presence of observations.

As in the previous exercise sample N, i.i.d. samples $X_{1:n}^{(i)}$ from $\nu^{\otimes n}$ and compute the following self-normalized IS estimator, where we now assume that π can be computed exactly rather than up to a normalizing constant,

$$\hat{\pi}_k^N(f) := \sum_{i=1}^{N} f\left(X_k^{(i)}\right) \frac{w_k\left(X_{1:k}^{(i)}\right)}{\sum_{j=1}^{N} w_k\left(X_{1:k}^{(j)}\right)},$$

where

$$w_k(x_{1:k}) = \frac{\pi^{\otimes k}(x_{1:k})}{\nu^{\otimes k}(x_{1:k})} = \prod_{l=1}^{k} \frac{\pi(x_l)}{\nu(x_l)}.$$

(a) Using the Law of Large Numbers show that the estimator is consistent, that is show that $\hat{\pi}_k^N(f) \to \pi_k(f)$ as the number of samples $N \to \infty$.

(b) The central limit theorem, and Slutsky's lemma, guarantee that

$$N^{1/2}\left[\hat{\pi}_k^N(f) - \pi_k(f)\right] \to \mathcal{N}(0, \sigma_k^2(f)).$$

Show that

$$\sigma_k^2(f) = \left(\int \nu(\mathrm{d}x)\left[\frac{\pi(x)}{\nu(x)}\right]^2\right)^{k-1} \cdot \int \nu(\mathrm{d}x)\left[\frac{\pi(x)}{\nu(x)}\right]^2 \left[f(x) - \int f(x)\pi(\mathrm{d}x)\right]^2.$$

(c) Using Jensen's inequality argue that if π/ν is not equal to one almost everywhere, then $\sigma_k^2(f)$ grows exponentially with k.

Programming Questions

Suppose we 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. Compute the gradient of the log-posterior density from 2.) Use this to implement Hamiltonian Monte Carlo with leapfrog steps for β, with mass matrix $M = B$ (the covariance of the prior). Choose the step size ϵ to be uniformly distributed on some interval $[0, m]$, and the step size L fixed. Experiment with different choices of m and L to obtain good performance by making sure that the acceptance rate remains high.

5. 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.

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

7. Compare the performance of your Gibbs, HMC and Metropolis-Hastings samplers.