Advanced Simulation

Problem Sheet 3, with answers

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

$$\pi(x) \propto \frac{q(x)}{\alpha(x)}.$$

Answer. The condition states that the acceptance $\alpha(x)$ has to be large enough in the tails of q. We have

$$\begin{aligned}\int_{\mathbb{X}} \pi(x)\, K(x,y)\, dx &= \int_{\mathbb{X}} \pi(x)\, \alpha(x)\, dx\; q(y) + (1-\alpha(y))\, \pi(y) \\ &\propto \int_{\mathbb{X}} \frac{q(x)}{\alpha(x)} \alpha(x)\, dx\; q(y) + (1-\alpha(y))\, \frac{q(y)}{\alpha(y)} \\ &\propto q(y) + (1-\alpha(y))\, \frac{q(y)}{\alpha(y)} \propto \pi(y)\end{aligned}$$

Alternatively, we can check that $K(x,y)$ is $\pi-$reversible as for $x \neq y$

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

so it is π-invariant.

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 $\mathbb{C}ov(X_1, X_k)$.)*

Answer. In this case, we have $\pi(x) = q(x)$. We have $\mathbb{V}[X_1] = \sigma^2$ and

$$\begin{aligned}\mathbb{C}ov[X_1, X_k] &= \mathbb{E}[X_1 X_k] - \mathbb{E}[X_1]\mathbb{E}[X_k] \\ &= \mathbb{E}[\mathbb{E}[X_1 X_k | X_1, X_{k-1}]] - \mathbb{E}^2[X_1] \\ &= (1-\alpha)\mathbb{E}[X_1 X_{k-1}] + \alpha\mathbb{E}^2[X_1] - \mathbb{E}^2[X_1] \\ &= (1-\alpha)\mathbb{C}ov[X_1, X_{k-1}] = (1-\alpha)^{k-1}\sigma^2.\end{aligned}$$

Hence we have

$$\sigma_X^2 = \sigma^2\left(1 + 2\sum_{k=1}^{\infty}(1-\alpha)^k\right) = \sigma^2\frac{2-\alpha}{\alpha}$$

which, as expected, goes to ∞ as $\alpha \to 0$.

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 $\pi(x, y)$.

Answer. We have

$$\begin{aligned}\pi(y|x) &\propto \frac{\exp\left(-\frac{1}{2}(x-1)^2(y-2)^2\right)}{\int \exp\left(-\frac{1}{2}(x-1)^2(y-2)^2\right)dy} \\ &\propto \exp\left(-\frac{1}{2(x-1)^{-2}}(y-2)^2\right) \\ &= \mathcal{N}\left(y; 2, (x-1)^{-2}\right)\end{aligned}$$

and similarly

$$\pi(x|y) = \mathcal{N}\left(x; 1, (y-2)^{-2}\right).$$

2. Does the resulting Gibbs sampler make any sense?

Answer. No it does not because

$$\int\left[\int \pi(x, y)\, dx\right] dy = \int\sqrt{2\pi(y-2)^{-2}}dy = \infty.$$

Hence the density $\pi(x, y)$ does not correspond to a probability distribution. Note that the algorithm is still easily implementable (which is scary).

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 $p(z_1, \ldots, z_T | \theta_1, \theta_2)$.

Answer. We have

$$p(z_1, \ldots, z_T | \theta_1, \theta_2) = \prod_{i=1}^{T} \left[\sum_{j_i = \max\{0, z_i - n_i\}}^{\min\{m_i, z_i\}} \binom{m_i}{j_i} \binom{n_i}{z_i - j_i} \theta_1^{j_i} (1-\theta_1)^{m_i - j_i} \theta_2^{z_i - j_i} (1-\theta_2)^{n_i - z_i + j_i} \right].$$

Indeed we can use the discrete convolution formula, for variables X, Y and $Z = X + Y$:

$$p_Z(z) = \sum_{x=0}^{z} p_X(x) p_Y(z - x).$$

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)

Answer. We introduce the latent variables X_i, Y_i and propose to sample from $p(\theta_1, \theta_2, x_{1:T}, y_{1:T} | z_{1:T})$. We have

$$p(\theta_1, \theta_2 | x_{1:T}, y_{1:T}, z_{1:T}) = p(\theta_1 | x_{1:T}) p(\theta_2 | y_{1:T})$$

where

$$\begin{aligned} p(\theta_1 | x_{1:T}) &\propto \prod_{i=1}^{T} \theta_1^{x_i} (1-\theta_1)^{m_i - x_i} \\ &= \mathcal{B}eta\left(\theta_1; 1 + \sum_{i=1}^{T} x_i, 1 + \sum_{i=1}^{T} (m_i - x_i)\right) \end{aligned}$$

and

$$p(\theta_2 | y_{1:T}) = \mathcal{B}eta\left(\theta_2; 1 + \sum_{i=1}^{T} y_i, 1 + \sum_{i=1}^{T} (n_i - y_i)\right).$$

Now we have

$$\begin{aligned} p(x_{1:T}, y_{1:T} | z_{1:T}, \theta_1, \theta_2) &= \prod_{i=1}^{T} p(x_i, y_i | z_i, \theta_1, \theta_2) \\ &\propto \prod_{i=1}^{T} \mathcal{B}inomial(x_i; m_i, \theta_1) \mathcal{B}inomial(y_i; n_i, \theta_2) \mathbb{1}_{x_i + y_i = z_i}. \end{aligned}$$

This distribution is intractable but we can still evaluate its probability mass function pointwise, up to a normalizing constant. Thus we can perform a Metropolis-Hastings step within the Gibbs sampler. For instance, we could naively propose from a uniform distribution on $\{0, \ldots, m_i\} \times \{0, \ldots, n_i\}$.

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 $\pi_{X,Y}$.

Answer. We have

$$K^{\mathrm{S}}_{X,Y}((x,y),(x',y')) = \pi_{Y|X}(y'|\, x)\, \pi_{X|Y}(x'|\, y').$$

and

$$\begin{aligned}
&\frac{\pi_{X,Y}(x,y)\, K^{\mathrm{S}}_{X,Y}((x,y),(x',y'))}{\pi_{X,Y}(x',y')\, K^{\mathrm{S}}_{X,Y}((x',y'),(x,y))} \\
&= \frac{\pi_{X,Y}(x,y)\, \pi_{Y|X}(y'|\, x)\, \pi_{X|Y}(x'|\, y')}{\pi_{X,Y}(x',y')\, \pi_{Y|X}(y|\, x')\, \pi_{X|Y}(x|\, y)} \\
&= \frac{\pi_Y(y)\, \pi_{Y|X}(y'|\, x)}{\pi_Y(y')\, \pi_{Y|X}(y|\, x')} \neq 1
\end{aligned}$$

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 $\pi_{Y|X}$.

Answer. We have

$$\int K_{X,Y}((x,y),(x',y'))\, dy' = \int \pi_{Y|X}(y'|\, x)\, \pi_{X|Y}(x'|\, y')\, dy'.$$

As this is expression is independent of y, we can conclude that $\left(X^{(t)}\right)_{t\geq 1}$ is Markov of transition kernel density

$$K_X(x,x') = \int \pi_{Y|X}(y'|\, x)\, \pi_{X|Y}(x'|\, y')\, dy'.$$

Moreover, we have

$$\begin{aligned}
\pi_X(x)\, K_X(x,x') &= \pi_X(x) \int \pi_{Y|X}(y'|\, x)\, \pi_{X|Y}(x'|\, y')\, dy' \\
&= \pi_X(x) \int \frac{\pi_{X|Y}(x|\, y')\, \pi_Y(y')}{\pi_X(x)} \frac{\pi_{Y|X}(y'|\, x')\, \pi_X(x')}{\pi_Y(y')} dy' \\
&= \pi_X(x') \int \pi_{X|Y}(x|\, y')\, \pi_{Y|X}(y'|\, x')\, dy' \\
&= \pi_X(x')\, K_X(x',x).
\end{aligned}$$

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.

Answer. We have

$$K^{\mathrm{R}}_{X,Y}\left((x,y),(x',y')\right)=\frac{1}{2}\pi_{Y|X}\left(y'|\,x\right)\delta_x\left(x'\right)+\frac{1}{2}\pi_{X|Y}\left(x'|\,y\right)\delta_y\left(y'\right).$$

Therefore

$$\begin{aligned}&\pi_{X,Y}\left(x,y\right)K^{\mathrm{R}}_{X,Y}\left((x,y),(x',y')\right)\\&=\frac{1}{2}\pi_{X,Y}\left(x,y\right)\pi_{Y|X}\left(y'|\,x\right)\delta_x\left(x'\right)+\frac{1}{2}\pi_{X,Y}\left(x,y\right)\pi_{X|Y}\left(x'|\,y\right)\delta_y\left(y'\right).\end{aligned}$$

On one hand we have

$$\begin{aligned}\pi_{X,Y}(x,y)\pi_{Y|X}(y'|x)\delta_x(x')&=\pi_{X,Y}(x',y)\pi_{Y|X}(y'|x)\delta_x(x')\\&=\pi_X(x')\pi_{Y|X}(y|x')\pi_{Y|X}(y'|x)\delta_x(x')\\&=\pi_{X,Y}(x',y')\pi_{Y|X}(y|x')\delta_x(x')\end{aligned}$$

and similarly

$$\pi_{X,Y}(x,y)\pi_{X|Y}(x'|y)\delta_y(y')=\pi_{X,Y}(x',y')\pi_{X|Y}(x|y')\delta_y(y')$$

So that we finally obtain

$$\begin{aligned}\pi_{X,Y}\left(x,y\right)K^{\mathrm{R}}_{X,Y}\left((x,y),(x',y')\right)&=\frac{1}{2}\pi_{X,Y}\left(x',y'\right)\pi_{Y|X}\left(y|\,x'\right)\delta_{x'}\left(x\right)+\frac{1}{2}\pi_{X,Y}\left(x',y'\right)\pi_{X|Y}\left(x|\,y'\right)\delta_{y'}\left(y\right)\\&=\pi_{X,Y}\left(x',y'\right)K^{\mathrm{R}}_{X,Y}\left((x',y'),(x,y)\right).\end{aligned}$$

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 $\sup_{\boldsymbol{q}}\|\nabla^2U(\boldsymbol{q})\|\leq L_U$, 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.)

Answer. The Leapfrog steps $(\boldsymbol{q}(\epsilon), \boldsymbol{p}(\epsilon)) = \Psi_\epsilon(\boldsymbol{q}, \boldsymbol{p})$ from a starting position $\boldsymbol{z} = (\boldsymbol{q}, \boldsymbol{p})$ are defined as

$$\begin{aligned}\boldsymbol{p}(\epsilon/2) &= \boldsymbol{p} - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}) \\ \boldsymbol{q}(\epsilon) &= \boldsymbol{q} + \epsilon \cdot \boldsymbol{p}(\epsilon/2). \\ \boldsymbol{p}(\epsilon) &= \boldsymbol{p}(\epsilon/2) - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}(\epsilon)).\end{aligned}$$

Let Ψ_ϵ denote the Leapfrog map, and let $\boldsymbol{z}(k\epsilon) = (\boldsymbol{q}(k\epsilon), \boldsymbol{p}(k\epsilon)) := \Psi_\epsilon^k(\boldsymbol{q}, \boldsymbol{p})$.
Let $\boldsymbol{M}(\boldsymbol{z}) := \frac{\partial \Psi_\epsilon}{\partial \boldsymbol{z}}$, then by the chain rule, the Jacobian of Ψ_ϵ^L can be written as

$$\frac{\partial \Psi_\epsilon^L}{\partial \boldsymbol{z}} = \boldsymbol{M}(\boldsymbol{z})\boldsymbol{M}(\boldsymbol{z}(\epsilon)) \cdot \ldots \cdot \boldsymbol{M}(\boldsymbol{z}((L-1)\epsilon)).$$

Moreover, for any $\boldsymbol{z} = (\boldsymbol{q}, \boldsymbol{p})$, we have

$$\begin{aligned}\boldsymbol{M}(\boldsymbol{z}) &= \begin{pmatrix} \boldsymbol{I}_d & \boldsymbol{0}_d \\ -\frac{\epsilon}{2}\nabla^2 U(\boldsymbol{q}) & \boldsymbol{I}_d \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_d & \epsilon \boldsymbol{I}_d \\ \boldsymbol{0}_d & \boldsymbol{I}_d \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_d & 0 \\ -\frac{\epsilon}{2}\nabla^2 U(\boldsymbol{q}(\epsilon)) & \boldsymbol{I}_d \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{I}_d - \frac{\epsilon^2}{2}\nabla^2 U(\boldsymbol{q}(\epsilon)) & \epsilon \boldsymbol{I}_d \\ -\frac{\epsilon}{2}\left(\nabla^2 U(\boldsymbol{q}) + \nabla^2 U(\boldsymbol{q}(\epsilon))\right) + \frac{\epsilon^3}{4}\nabla^2 U(\boldsymbol{q})\nabla^2 U(\boldsymbol{q}(\epsilon)) & \boldsymbol{I}_d - \frac{\epsilon^2}{2}\nabla^2 U(\boldsymbol{q}) \end{pmatrix}.\end{aligned}$$

Let $\boldsymbol{D}_0 = \boldsymbol{M}(\boldsymbol{z}) - \boldsymbol{I}_{2d}$, and $\boldsymbol{D}_k = \boldsymbol{M}(\boldsymbol{z}(k\epsilon)) - \boldsymbol{I}_d$ for $1 \le k \le L-1$. One can show that for a block matrix $\boldsymbol{E} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$, we have $\|\boldsymbol{E}\| \le \|\boldsymbol{A}\| + \|\boldsymbol{B}\| + \|\boldsymbol{C}\| + \|\boldsymbol{D}\|$. Hence using the fact that $\|\nabla^2 U(\boldsymbol{q})\| \le L_U$, and our assumption on ϵ, for any $0 \le k \le L-1$, we have

$$\|\boldsymbol{D}_k\| \le \epsilon^2 L_U + \epsilon + \epsilon L_U + \frac{\epsilon^3}{4}L_U^2 \le (4/3)\epsilon(1 + L_U).$$

Since

$$\begin{aligned}\frac{\partial \Psi_\epsilon^L}{\partial \boldsymbol{z}} &= (\boldsymbol{I}_d + \boldsymbol{D}_0) \cdot \ldots \cdot (\boldsymbol{I}_d + \boldsymbol{D}_{L-1}) \\ &= \boldsymbol{I}_d + \boldsymbol{D}_0 + \ldots + \boldsymbol{D}_{L-1} + \sum_{0 \le i_1 < i_2 \le L-1} \boldsymbol{D}_{i_1}\boldsymbol{D}_{i_2} + \sum_{0 \le i_1 < i_2 < i_3 \le L-1} \boldsymbol{D}_{i_1}\boldsymbol{D}_{i_2}\boldsymbol{D}_{i_3} + \ldots + \boldsymbol{D}_0 \ldots \boldsymbol{D}_{L-1},\end{aligned}$$

using the above bound on $\|\boldsymbol{D}_k\|$, we have

$$\begin{aligned}&\left\| \frac{\partial \Psi_\epsilon^L}{\partial \boldsymbol{z}} - (\boldsymbol{I}_d + \boldsymbol{D}_0 + \ldots + \boldsymbol{D}_{L-1}) \right\| \\ &\le \binom{L}{2}((4/3)\epsilon(1+L_U))^2 + \binom{L}{3}((4/3)\epsilon(1+L_U))^3 + \ldots + \binom{L}{L}((4/3)\epsilon(1+L_U))^L \\ &\le \frac{[L((4/3)\epsilon(1+L_U))]^2}{2!} + \frac{[L((4/3)\epsilon(1+L_U))]^3}{3!} + \ldots + \frac{[L((4/3)\epsilon(1+L_U))]^L}{L!} \\ &\le \exp\left(L((4/3)\epsilon(1+L_U))\right) - 1 - L((4/3)\epsilon(1+L_U)) \le (L((4/3)\epsilon(1+L_U))^2 \le 2(L\epsilon)^2(1+L_U)^2 \le \frac{L\epsilon}{2},\end{aligned}$$

using the fact that $\exp(x) - 1 - x \le x^2$ for $0 \le x \le 1$, and $L((4/3)\epsilon(1+L_U)) \le 1$ based on our assumption on ϵ. Now the result follows from the fact that $\boldsymbol{I}_d + \boldsymbol{D}_0 + \ldots + \boldsymbol{D}_{L-1}$ has $L\epsilon\boldsymbol{I}_d$ in the top right $d \times d$ block-matrix corresponding to the Jacobian $\frac{\partial \boldsymbol{q}(L\epsilon)}{\partial \boldsymbol{p}(0)}$.

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 $0 \le k \le L$,

$$H(\boldsymbol{q}(k\epsilon), \boldsymbol{p}(k\epsilon)) - H(\boldsymbol{q}, \boldsymbol{p}) \le C_H L\epsilon \cdot H(\boldsymbol{q}, \boldsymbol{p}). \tag{1}$$

Hence the Hamiltonian is approximately preserved for sufficiently short intervals. (Hint: use Taylor expansion with remainder term and the assumption $\sup_{\boldsymbol{q}} \|\nabla^2 U(\boldsymbol{q})\| \le L_U$ to control the change of the Hamiltonian).

Answer. The Leapfrog dynamics consists of 3 steps,

$$\begin{aligned}
\boldsymbol{p}(\epsilon/2) &= \boldsymbol{p} - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}) \\
\boldsymbol{q}(\epsilon) &= \boldsymbol{q} + \epsilon \cdot \boldsymbol{p}(\epsilon/2). \\
\boldsymbol{p}(\epsilon) &= \boldsymbol{p}(\epsilon/2) - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}(\epsilon)).
\end{aligned}$$

Note that by Taylor's expansion with second order raminder term, we have

$$U(\boldsymbol{q}) - U_{\min} \ge U(\boldsymbol{q}) - U(\boldsymbol{q} - \nabla U(\boldsymbol{q})/L_U) \ge \langle \nabla U(\boldsymbol{q})/L_U, \nabla U(\boldsymbol{q})\rangle - \frac{1}{2}L_U\|\nabla U(\boldsymbol{q})/L_U\|^2 = \frac{\|\nabla U(\boldsymbol{q})\|^2}{2L_U},$$

hence $\|\nabla U(\boldsymbol{q})\|^2 \le 2L_U(U(\boldsymbol{q}) - U_{\min})$.

For the first step in the Leapfrog, by the Cauchy-Schwarz inequality, we can show that

$$\begin{aligned}
H(\boldsymbol{q}, \boldsymbol{p}(\epsilon/2)) - H(\boldsymbol{q}, \boldsymbol{p}) &= \frac{\|\boldsymbol{p}(\epsilon/2)\|^2}{2} - \frac{\|\boldsymbol{p}\|^2}{2} = -\frac{\epsilon}{2}\langle \nabla U(\boldsymbol{q}), \boldsymbol{p}\rangle + \frac{\epsilon^2}{4}\|\nabla U(\boldsymbol{q})\|^2 \le \frac{\epsilon+\epsilon^2}{4}\|\nabla U(\boldsymbol{q})\|^2 + \frac{\epsilon}{4}\|\boldsymbol{p}\|^2 \\
&\le \frac{\epsilon(1+\epsilon)}{2}L_U(U(\boldsymbol{q}) - U_{\min}) + \frac{\epsilon}{2}\frac{\|\boldsymbol{p}\|^2}{2} \le \frac{\epsilon(1+\epsilon)(1+L_U)}{2}H(\boldsymbol{q}, \boldsymbol{p}).
\end{aligned}$$

For the second step in the Leapfrog, Taylor expansion with second order remainder term, and Cauchy-Schwarz, we have

$$\begin{aligned}
H(\boldsymbol{q}(\epsilon), \boldsymbol{p}(\epsilon/2)) - H(\boldsymbol{q}, \boldsymbol{p}(\epsilon/2)) &= U(\boldsymbol{q}(\epsilon)) - U(\boldsymbol{q}) \le \langle \nabla U(\boldsymbol{q}), \epsilon\boldsymbol{p}(\epsilon/2)\rangle + \frac{L_U\|\epsilon\boldsymbol{p}(\epsilon/2)\|^2}{2} \\
&\le \frac{(\epsilon^2 L_U + \epsilon)\|\boldsymbol{p}(\epsilon/2)\|^2}{4} + \epsilon L_U(U(\boldsymbol{q}) - U_{\min}) \le \epsilon(1 + L_U + L_U\epsilon)H(\boldsymbol{q}, \boldsymbol{p}(\epsilon/2)).
\end{aligned}$$

A similar result holds for the third step too, and combinging these together leads to (1).

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

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

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).

Answer. Using (1), for any $0 \le k \le L$, we have

$$\frac{\|\boldsymbol{p}(k\epsilon)\|^2}{2} \le H(\boldsymbol{q}(k\epsilon), \boldsymbol{p}(k\epsilon)) \le \left(\frac{\|\boldsymbol{p}(0)\|^2}{2} + U(\boldsymbol{q}(0)) - U_{\min}\right)(1 + C_H k\epsilon),$$

therefore

$$\begin{aligned}
\|\boldsymbol{p}(k\epsilon)\| &\le \sqrt{\left(\frac{\|\boldsymbol{p}(0)\|^2}{2} + U(\boldsymbol{q}(0)) - U_{\min}\right)(1 + C_H L\epsilon)} \\
&\le (1 + C_H L\tau)\left(\sqrt{2(U(\boldsymbol{q}(0)) - U_{\min})} + \|\boldsymbol{p}(0)\|\right).
\end{aligned}$$

Thus for $\boldsymbol{p}(0) = \frac{\boldsymbol{Q}-\boldsymbol{q}(0)}{L\epsilon}$, for any $0 \le k \le L$, we have

$$\|\boldsymbol{q}(k\epsilon) - \boldsymbol{q}(0)\| \le D \text{ for } D = (1 + C_H L\tau)(L\tau\sqrt{2(U(\boldsymbol{q}(0)) - U_{\min})} + \|\boldsymbol{Q} - \boldsymbol{q}(0)\|).$$

The rest of the argument is similar to the proof of Lemma 2 of Chapter 7. Let $B_D(\boldsymbol{q}) = \{\boldsymbol{q}' \in \mathbb{R}^d : \|\boldsymbol{q}' - \boldsymbol{q}\| \le D\}$ denote the ball of radius D centered at $\boldsymbol{q}$. Then from the continuity of ∇U, we have

$$E := \sup_{\boldsymbol{q}' \in B_D(\boldsymbol{q})} \|\nabla U(\boldsymbol{q}') < \infty.$$

Let $\boldsymbol{p}(k\epsilon + \epsilon/2) = \boldsymbol{p}(k\epsilon) - \frac{\epsilon}{2}\nabla U(\boldsymbol{q}(k\epsilon))$ denote the first step of the Leapfrog dynamics started at $(\boldsymbol{q}(k\epsilon), \boldsymbol{p}(k\epsilon))$. Then by induction it follows that for any $0 \le k \le L-1$,

$$\|\boldsymbol{p}(k\epsilon + \epsilon/2) - \boldsymbol{p}(0)\| \le E(k + 1/2)\epsilon,$$

and therefore

$$\|\boldsymbol{q}(L\epsilon) - \boldsymbol{Q}\| = \|\boldsymbol{q}(L\epsilon) - \boldsymbol{q}(0) - L\epsilon\boldsymbol{p}(0)\| = \left\|\sum_{k=0}^{L-1} \epsilon(\boldsymbol{p}(k\epsilon + \epsilon/2) - \boldsymbol{p}(0))\right\|$$
$$\le \sum_{k=0}^{L-1}(k + 1/2)E\epsilon^2 = \frac{\epsilon^2 L^2 E}{2},$$

hence the result.

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

Answer. The proof is essentially same as the proof of Proposition 4 of Chapter 7, except that we use Parts 1 and 3 in place of Lemmas 1 and 2. For completeness, we repeat it here.

Let $\nu(\boldsymbol{q}) \propto \exp(-U(\boldsymbol{q}))$ be the target distribution, and μ be the multivariate normal distribution on $\mathbb{R}^d$ with mean 0 and covariance matrix $\boldsymbol{M}$. Assume without loss of generality that $\boldsymbol{M} = \boldsymbol{I}_d$ (the general case follows by a simple modification of this argument).

Our goal is to show that for any starting point $\boldsymbol{q}(0)$, any measurable set $A \subset \mathbb{R}^d$ with $\nu(A) > 0$, we have $K(\boldsymbol{q}(0), A) > 0$. Since we can fill $\mathbb{R}^d$ into countably many balls of radius δ (for any $\delta > 0$), we can assume without loss of generality that A is a subset of a ball of radius δ centered at some point $\boldsymbol{Q} \in \mathbb{R}^d$. We will denote this closed ball by $B_\delta(\boldsymbol{Q}) = \{\boldsymbol{q} \in \mathbb{R}^d : \|\boldsymbol{q} - \boldsymbol{Q}\| \le \delta\}$.

Then by Part 3, it follows that if $\epsilon \le \frac{\sqrt{\delta}}{L\sqrt{C}}$, and

$$\boldsymbol{p}(0) = \boldsymbol{p}^* = \frac{\boldsymbol{Q} - \boldsymbol{q}(0)}{t},$$

then $\boldsymbol{q}(L\epsilon) \in B_\delta(\boldsymbol{Q})$.

Let $\boldsymbol{\Psi}(\boldsymbol{p}(0))$ denote the position $(\boldsymbol{q})$ component of $(\boldsymbol{\Psi}_\epsilon^L)(\boldsymbol{q}(0), \boldsymbol{p})$ ($\boldsymbol{q}(0)$ is assumed fixed). Hence for a fixed $\boldsymbol{q}(0)$, $\boldsymbol{\Psi}$ maps from the initial momentum $\boldsymbol{p}(0)$ to the position after L leapfrog steps of size ϵ.

It follows from Part 1 that for any vector $\boldsymbol{v}$, we have

$$\|\boldsymbol{\Psi}(\boldsymbol{p}^*) - \boldsymbol{\Psi}(\boldsymbol{p}^* + \boldsymbol{v})\| = \left\|\left(\int_{s=0}^1 \frac{\partial \boldsymbol{q}(L\epsilon)}{\partial \boldsymbol{p}(0)}\Big|_{\boldsymbol{p}(0)=\boldsymbol{p}^*+\boldsymbol{v}s} ds\right)\boldsymbol{v}\right\| \ge \frac{L\epsilon\|\boldsymbol{v}\|}{2}.$$

Therefore, with the choice $r = \frac{6\delta}{t}$ the sphere $S_r(\boldsymbol{p}^*) = \{\boldsymbol{p}' \in \mathbb{R}^d : \|\boldsymbol{p}' - \boldsymbol{p}^*\| = r\}$ satisfies that for any $\boldsymbol{p}' \in S_r(\boldsymbol{p}^*)$,

$$\|\boldsymbol{\Psi}(\boldsymbol{p}^*) - \boldsymbol{\Psi}(\boldsymbol{p})\| \ge 3\delta.$$

Moreover, the map of $\boldsymbol{p}^*$, $\boldsymbol{\Psi}(\boldsymbol{p}^*)$ is also contained in the ball $B_\delta(\boldsymbol{Q})$, thus it follows that map of the ball $B_r(\boldsymbol{p}^*)$ by $\boldsymbol{\Psi}$ will contain the ball $B_\delta(\boldsymbol{Q})$ (see Figure **??**). Hence for any point in $\boldsymbol{Q}' \in B_\delta(\boldsymbol{Q})$, there is at least one point $\boldsymbol{p}' \in B_r(\boldsymbol{p}^*)$ such that $\boldsymbol{\Psi}(\boldsymbol{p}') = \boldsymbol{q}'$. Denote the set of points in $B_r(\boldsymbol{p}^*)$ that get mapped into the set A by $\boldsymbol{\Psi}$ as B_ϵ (preimage of A). Since $\nu(A) > 0$, and ν has a density, it follows that $\mathrm{Vol}(A) > 0$, and by Part 1, it follows that if $0 < \epsilon \leq \frac{1}{4L(1+L_U)^2}$, the determinant of the Jacobian of $\boldsymbol{\Psi}$ on B_ϵ is finite and positive, so we also must have $\mathrm{Vol}(B_\epsilon) > 0$, implying that $\mu(B_\epsilon) > 0$ (μ denotes the standard normal distribution on $\mathbb{R}^d$, corresponding to the distribution of the resampled momentum variable). Since this holds for any $\epsilon \leq \min\left(\frac{\sqrt{\delta}}{L\sqrt{C}}, \frac{1}{4L(1+L_U)^2}\right)$, ϵ is uniformly distributed on $[0, \tau]$, and the proposed steps are accepted with positive probability, it follows by integration that $\boldsymbol{K}(\boldsymbol{q}(0), A) > 0$, thus $\boldsymbol{K}$ is strongly ν-irreducible.

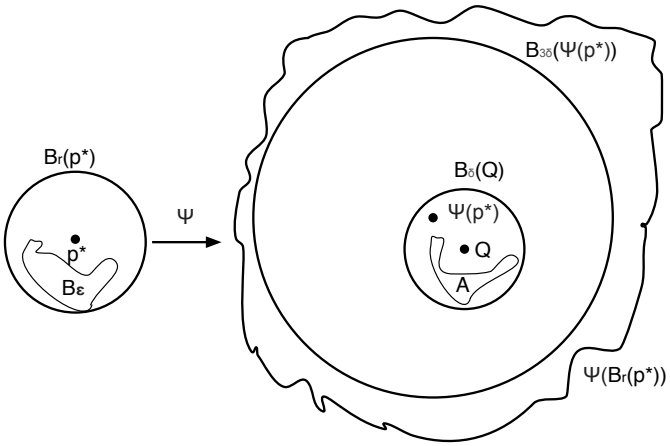

Figure 1: Effect of the map $\boldsymbol{\Psi}$

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?

Answer.

(a) The first part is obvious

$$\begin{aligned}\mathbb{E}[\hat{Z}_n^N] &= \int\cdots\int \nu^{\otimes n}(x_{1:n})\frac{\tilde{\pi}^{\otimes n}(x_{1:n})}{\nu^{\otimes n}(x_{1:n})}\mathrm{d}x_{1:n}\\ &= \int\cdots\int \tilde{\pi}^{\otimes n}(x_{1:n})\mathrm{d}x_{1:n} = Z_n,\end{aligned}$$

by definition.

(b) Since the Monte Carlo samples are i.i.d. the variance of the average is $1/N$ times the variance of one sample. To keep notation to a minimum let us write $\hat{Z}$ for the random variable $\hat{Z}_n := \tilde{\pi}^{\otimes n}(Y_{1:n})/\nu^{\otimes n}(Y_{1:n})$, where $Y_{1:n}\sim\nu^{\otimes n}$, and $\hat{Z} := \tilde{\pi}(Y)/\nu(Y)$, where $Y\sim\nu$, Then the variance of a single sample is then given by,

$$\begin{aligned}\mathrm{var}\left[\hat{Z}_n\right] &= \mathbb{E}\left[\hat{Z}_n^2\right] - \mathbb{E}\left[\hat{Z}_n\right]^2\\ &= \mathbb{E}\left[\hat{Z}_n^2\right] - Z^{2n}\\ &= \left[\int\nu(\mathrm{d}y)\frac{\tilde{\pi}(y)^2}{\nu(y)^2}\right]^n - Z^{2n}\\ &= \mathbb{E}[\hat{Z}^2]^n - \mathbb{E}[\hat{Z}]^{2n}.\end{aligned}$$

(c) Continuing from above we have

$$\mathrm{var}\left[\frac{\hat{Z}_n}{Z_n}\right] = \mathbb{E}\left[\frac{\hat{Z}^2}{Z^2}\right]^n - 1.$$

$$N\,\mathrm{var}\left[\frac{\hat{Z}_n^N}{Z_n}\right] = N\left[\left(\frac{\mathbb{E}[\hat{Z}^2]}{\mathbb{E}[\hat{Z}]^2}\right)^n - 1\right].$$

Let us write

$$\gamma := \frac{\mathbb{E}[\hat{Z}^2]}{\mathbb{E}[\hat{Z}]^2},$$

whence using either Jensen's inequality inequality it follows that $\gamma \geq 1$. Using the assumption that $\tilde{\pi}/\nu$ is not almost everywhere constant, it follows that Jensen's inequality is a strict inequality and therefore $\gamma > 1$. Thus

$$\frac{1}{n}\log\left(N \operatorname{var}\left[\hat{Z}_n^N\right]\right) = \frac{\log N}{n} + \frac{1}{n}\log\left(\gamma^n - 1\right).$$

For the last term notice that

$$\begin{aligned}\frac{1}{n}\log\left(\gamma^n - 1\right) &= \frac{1}{n}\log\left(\gamma^n \times \frac{\gamma^n - 1}{\gamma^n}\right) \\ &= \frac{1}{n}\log\left(\gamma^n\right) + \frac{1}{n}\log\left(\frac{\gamma^n - 1}{\gamma^n}\right) \\ &= \log\gamma + \frac{1}{n}\log\left(1 + \gamma^{-n}\right) \to \log\gamma,\end{aligned}$$

as $n \to \infty$ since $\gamma > 1$. Therefore as $n \to \infty$, with N fixed,

$$\frac{1}{n}\log\left(N \operatorname{var}\left[\hat{Z}\right]\right) \to \log\gamma.$$

(d) From the previous part we can see that the relative variance of $\hat{Z}_n^N$ grows exponentially in the time horizon n and is proportional to $1/N$, that is

$$\operatorname{var}\left(\frac{\hat{Z}_n^N}{Z_n}\right) \propto \frac{C}{N}\gamma^{n(1+o(1))},$$

therefore to have an estimator with bounded relative variance for large n we need a number of samples N proportional to γ^n.

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.

Answer.

(a) The first part is trivial as it is a standard normalized importance sampling algorithm, but just for completeness

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

Using the law of large numbers on the numerator we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N f\left(X_k^{(i)}\right) w_k\left(X_{1:k}^{(i)}\right) &\to \int\cdots\int \nu^{(\otimes k)}(x_{1:k})\frac{\pi(x_{1:k})}{\nu(x_{1:k})} f(x_k)\mathrm{d}x_{1:k} \\ &\to \int\cdots\int f(x_k)\pi(\mathrm{d}x_{1:k}) = \pi_k(f),\end{aligned}$$

whereas the above replacing f with 1, shows that the numerator converges to 1 by the Law of Large Numbers. Combining the two limits we get the result.

(b) For the second part we have

$$\begin{aligned}N^{1/2}\left[\hat{\pi}_k^N(f) - \pi_k(f)\right] &= \left(\frac{1}{N}\sum_{j=1}^N w_k\left(X_{1:k}^{(j)}\right)\right)^{-1} \times \sqrt{N}\left[\frac{1}{N}\sum_{i=1}^N f\left(X_k^{(i)}\right) w_k\left(X_{1:k}^{(i)}\right) - \pi_k(f)\right] \\ &= \left(\frac{1}{N}\sum_{j=1}^N w_k\left(X_{1:k}^{(j)}\right)\right)^{-1} \times \frac{1}{\sqrt{N}}\sum_{i=1}^N w_k\left(X_{1:k}^{(i)}\right)\left[f\left(X_k^{(i)}\right) - \pi_k(f)\right]\end{aligned}$$

since $\pi_k(f)$ does not depend on i.

Notice that

$$\int\cdots\int \nu(\mathrm{d}x_{1:k}) w_k(x_{1:k})\left[f(x_k) - \pi_k(f)\right] = \int\cdots\int \pi_k(\mathrm{d}x_{1:k})\left[f(x_k) - \pi_k(f)\right] = 0,$$

so that the terms of the numerator have zero mean.

Therefore, the standard central limit theorem applied to the numerator gives

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_k\left(X_{1:k}^{(i)}\right)\left[f\left(X_k^{(i)}\right)-\pi_k(f)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma_k^2(f))$$

where $\sigma_k^2(f)$ is given by

$$\begin{aligned}\sigma_k^2(f) &:= \int\cdots\int \nu(\mathrm{d}x_{1:k}) w_k^2(x_{1:k})\left[f(x_k)-\pi_k(f)\right]^2 \\ &= \int\cdots\int \nu(\mathrm{d}x_{1:k-1}) w_{k-1}^2(x_{1:k-1}) \int \nu(\mathrm{d}x_k)\frac{\pi(x_k)^2}{\nu(x_k)^2}\left[f(x_k)-\pi_k(f)\right]^2 \\ &= \left(\int \nu(\mathrm{d}x)\left[\frac{\pi(x)}{\nu(x)}\right]^2\right)^{k-1} \int \nu(\mathrm{d}x_k)\frac{\pi(x_k)^2}{\nu(x_k)^2}\left[f(x_k)-\pi_k(f)\right]^2,\end{aligned}$$

since the first $k-1$ integrals factorise due to the identities

$$\begin{aligned}\nu(\mathrm{d}x_{1:k-1}) &= \prod_{j=1}^{k-1}\nu(x_j)\mathrm{d}x_j \\ w_{k-1}^2(x_{1:k-1}) &= \prod_{j=1}^{k-1}\left[\frac{\pi(x_j)}{\nu(x_j)}\right]^2.\end{aligned}$$

(c) It is obvious that

$$\int \nu(\mathrm{d}x)\frac{\pi(x)}{\nu(x)} = \int \pi(x)\mathrm{d}x = 1.$$

Thus Jensen's inequality gives that

$$\gamma := \int \nu(\mathrm{d}x)\left[\frac{\pi(x)}{\nu(x)}\right]^2 \geq \left[\int \nu(\mathrm{d}x)\frac{\pi(x)}{\nu(x)}\right]^2 = 1.$$

Jensen's inequality is strict unless the random variable is constant almost surely. In our case this translates to π/ν not being equal to a constant or $\pi(\cdot) = c\nu(\cdot)$. Since π and ν are both probability densities, the only c possible here is $c \equiv 1$. We know that π is not identically equal to ν by the assumption and thus Jensen's inequality is strict giving $\gamma > 1$. This proves that the asymptotic variance grows like γ^{k-1}, that is exponentially with the time-horizon.

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.