Importance sampling and Sequential Monte Carlo

Exercise 1: Choice of proposal distribution

Let X be a random variable with probability density g with respect to the Lebesgue measure on $\mathbb{R}$. Let $\kappa_X : t \mapsto \log(\mathbb{E}[\mathrm{e}^{tX}])$. We want to estimate $\mathbb{P}(X \geq x)$ for $x \in \mathbb{R}$ using the proposal distribution $h_t : x \mapsto \mathrm{e}^{xt-\kappa_X(t)}g(x)$ for $t \in \mathbb{R}$.

1. Propose a naive Monte Carlo estimator of $\mathbb{P}(X \geq x)$ for $x \in \mathbb{R}$.

   Let $(Y_i)_{1\leqslant i\leqslant n}$ be i.i.d. random variables with probability density h_t with respect to the Lebesgue measure. A naive Monte Carlo estimate is:

$$I_n(t,x) = \frac{1}{n}\sum_{i=1}^{n}\frac{h_t(Y_i)}{g(Y_i)}\mathbb{1}_{Y_i\geqslant x}\,.$$

2. Show that
$$\mathbb{E}\left[\mathbb{1}_{Y\geq x}\mathrm{e}^{-2Yt+2\kappa_X(t)}\right] \leq \exp(-xt+\kappa_X(t))\,.$$
   where Y has density h_t for $t \geq 0$.

   By definition,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\{Y\geqslant x\}}\mathrm{e}^{-2Yt+2\kappa_X(t)}\right] = \mathbb{E}\left[\mathbb{1}_{\{X\geqslant x\}}\mathrm{e}^{-2Xt+2\kappa_X(t)}\frac{h_t(X)}{g(X)}\right] &= \mathbb{E}\left[\mathbb{1}_{\{X\geqslant x\}}\mathrm{e}^{-Xt+\kappa_X(t)}\right]\,,\\ &\leqslant \mathbb{E}\left[\mathbb{1}_{\{X\geqslant x\}}\mathrm{e}^{-xt+\kappa_X(t)}\right]\,,\\ &\leqslant \mathrm{e}^{-xt+\kappa_X(t)}\,.\end{aligned}$$

3. Propose a choice t_x to select the proposal distribution h_t.

   lt is enough to choose t_x such that $\kappa'_X(t_x) = x$.

4. Apply this result when $X \sim \mathcal{N}(\mu,\sigma^2)$.

   For all $t \in \mathbb{R}$, $\kappa_X(t) = \mu t + \sigma^2t^2/2$ since

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{tX}] &= (2\pi\sigma^2)^{-1/2}\int_{\mathbb{R}}\mathrm{e}^{tx-(x-\mu)^2/(2\sigma^2)}\mathrm{d}x\,,\\ &= (2\pi\sigma^2)^{-1/2}\int_{\mathbb{R}}\exp\left(-\frac{(x-(t\sigma^2+\mu))^2}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}+\frac{(t\sigma^2+\mu)^2}{2\sigma^2}\right)\mathrm{d}x\,,\\ &= \exp\left(-\frac{\mu^2}{2\sigma^2}+\frac{(t\sigma^2+\mu)^2}{2\sigma^2}\right)\,,\\ &= \exp\left(\mu t+\sigma^2t^2/2\right)\,.\end{aligned}$$

   Therefore, we choose $t_x = (x-\mu)/\sigma^2$.

5. Apply this result when $X \sim \mathcal{P}(\lambda)$.

   If $X \sim \mathcal{P}(\lambda)$, with $\lambda > 0$, for all t,

$$\mathbb{E}[\mathrm{e}^{tX}] = \mathrm{e}^{-\lambda}\sum_{k\geqslant 0}\frac{\lambda^k}{k!}\mathrm{e}^{tk} = \exp\left(-\lambda+\lambda\mathrm{e}^t\right)\,.$$

   Therefore, we choose $t_x = \log(x/\lambda)$.

Exercise 2: Optimal kernel

We consider a linear ang Gaussian hidden Markov model given for $k \geq 0$ by

$$\begin{aligned} X_{k+1} &= \phi X_k + \sigma U_k \,, \\ Y_k &= X_k + \eta V_k \,, \end{aligned}$$

where $(U_k)_{k\geq 0}$ and $(V_k)_{k\geq 0}$ are independent standard Gaussian random variables independent of X_0. The distribution ν of X_0 is the stationnary distribution of the Markov Chain.

1. Write the joint probability density function of $(X_{0:n}, Y_{0:n})$.

   The joint probability density function of $(X_{0:n}, Y_{0:n})$ is

   $$(x_{0:n}, y_{0:n}) \mapsto p(x_{0:n}, y_{0:n}) = \nu(x_0) \prod_{k=1}^{n} m(x_{k-1}, x_k) \prod_{k=0}^{n} g(x_k, y_k) \,,$$

   where $m(x_{k-1}, x_k)$ is the Gaussian probability density function with mean ϕx_{k-1} and variance σ^2, evaluated at x_k, and $g(x_k, y_k)$ is the Gaussian probability density function with mean x_k and variance η^2, evaluated at y_k.

2. Write the recursion defining the filtering distributions, i.e. the distributions of X_n given $Y_{0:n}$ for $n \geq 0$.

   Write ϕ_n the distribution of X_n given $Y_{0:n}$. Note that

   $$p(x_{0:n}|y_{0:n}) \propto p(x_{0:n}, y_{0:n}) \propto p(x_{0:n-1}|y_{0:n-1}) m(x_{n-1}, x_n) g(x_n, y_n) \,.$$

   By integrating over $x_{0:n-1}$, this yields

   $$p(x_{0:n}|y_{0:n}) \propto g(x_n, y_n) \int p(x_{n-1}|y_{0:n-1}) m(x_{n-1}, x_n) \mathrm{d}x_{n-1} \,.$$

   Therefore,

   $$\phi_n(x_n) \propto g(x_n, y_n) \int \phi_{n-1}(x_{n-1}) m(x_{n-1}, x_n) \mathrm{d}x_{n-1} \,.$$

3. Propose a sequential Monte Carlo method to estimate the filtering distribution at time $n+1$ using weighted samples $\{(\xi_n^i, \omega_n^i)\}_{i=1}^N$ targetting the filtering distribution at time n. New particles are proposed using the prior kernel, i.e. the distribution of X_{n+1} given X_n.

   Using that $\{(\xi_n^i, \omega_n^i)\}_{i=1}^N$ approximate ϕ_n, The recursion of the previous question allows to approximate ϕ_{n+1} by

   $$\hat{\phi}_{n+1}(x_{n+1}) \propto \sum_{i=1}^{N} \omega_n^i m(\xi_n^i, x_{n+1}) g(x_{n+1}, y_{n+1}) \,.$$

   This distribution can be estimated using weigthed samples : for all $1 \leq i \leq N$, sample I_{n+1}^i in $\{1, \ldots, N\}$ with probabilities $\{\omega_n^i\}_{i=1}^N$, then sample $\xi_{n+1}^i \sim m(\xi_n^{I_{n+1}^i}, \cdot)$ and set $\omega_{n+1}^i \propto g(\xi_{n+1}^i, y_{n+1})$.

4. The *optimal kernel* to propose new particles is defined as the distribution of X_{n+1} given (X_n, Y_{n+1}). Compute the optimal kernel and the weights $(\omega_{n+1}^i)_{i=1}^N$.

   If we use the optimal kernel, it means that the sampling step of the previous question is replaced by $\xi_{n+1}^i \sim p(x_{n+1}|\xi_n^{I_{n+1}^i}, y_{n+1})$ where

   $$p(x_{n+1}|x_n, y_{n+1}) \propto m(x_n, x_{n+1}) g(x_{n+1}, y_{n+1}) \,.$$

This is a Gaussian distribution. Then, the weights are computed as follows:

$$\omega_{n+1}^i \propto \frac{m(\xi_n^{I_{n+1}^i}, \xi_{n+1}^i) g(\xi_{n+1}^i, y_{n+1})}{p(\xi_{n+1}^i | \xi_n^{I_{n+1}^i}, y_{n+1})} \propto 1\,.$$

Each new samples is associated with the weight $1/N$.

5. In other settings than linear and Gaussian HMM, the optimal kernel is usually not tractable. Propose an accept-reject mechanism to sample from the optimal kernel for general HMM.

   As

   $$p(x_{n+1}|x_n, y_{n+1}) \propto m(x_n, x_{n+1}) g(x_{n+1}, y_{n+1})\,,$$

   if $x_{n+1} \mapsto g(x_{n+1}, y_{n+1})$ *is bounded by c, we can use an accept-reject mechanism by proposing samples with the prior kernel (as in question 3) and use that* $m(x_n, x_{n+1}) g(x_{n+1}, y_{n+1}) \leq cm(x_n, x_{n+1})$.

Exercise 3: Smoothing distribution

Let $\{(X_k, Y_k)\}_{k\geq 0}$ be a HMM where $(X_k)_{k\geq 0}$ is a Markov chain with initial distribution ν and Markov transition density m. For all $k \geq 0$, the conditional distribution of Y_k given $X_{0:n}$ depends on X_k only and its probability density function is written $g(X_k, \cdot)$.

1. Prove that for all $0 \leq k \leq n-1$, the conditional distribution of X_k given X_{k+1} and $Y_{0:k}$ is proportional to $\phi_k(\cdot) m(\cdot, X_{k+1})$ where ϕ_k is the filtering distribution at time k. We write $b_k(X_{k+1}, \cdot)$ this distribution.

   For all $0 \leq k \leq n-1$,

   $$p(X_k | X_{k+1}, Y_{0:k}) \propto p(X_k, X_{k+1}, Y_{0:k}) \propto \phi_k(X_k) m(X_k, X_{k+1})$$

2. Prove that the joint density of $X_{0:n}$ given $Y_{0:n}$ can be written $x_{0:n} \mapsto \phi_n(x_n) \prod_{k=0}^{n-1} b_k(x_{k+1}, x_k)$.

   Note that

   $$p(x_{0:n}|y_{0:n}) = p(x_n|y_{0:n}) \prod_{k=0}^{n-1} p(x_k|x_{k+1:n}, y_{0:n})$$

   and that, by the Markov property,

   $$p(x_k|x_{k+1:n}, y_{0:n}) = p(x_k|x_{k+1:n}, y_{0:k}) = b_k(x_{k+1}, x_k)\,.$$

3. Assume that at each time k, $\{(\xi_k^i, \omega_k^i)\}_{i=1}^N$ is a particle-based approximation of ϕ_k. Propose a particle-based approximation of $b_k(X_{k+1}, \cdot)$.

   If we replace the filtering distribution by its particle-based approximation, we obtain the following approximation of $b_k(X_{k+1}, \cdot)$:

   $$\hat{b}_k^N(X_{k+1}, \mathrm{d}x_k) \propto \sum_{i=1}^N \omega_k^i m(x_k, X_{k+1}) \delta_{\xi_k^i}(\mathrm{d}x_k)\,.$$

4. Deduce from the previous questions an algorithm to approximately sample from the joint distribution $X_{0:n}$ given $Y_{0:n}$.

   It is enough to sample ξ_{n+1} *from* ϕ_n *ans then to recursively sample backward in time using the approximated kernels* b_k^N, $0 \leq k \leq n-1$.