Stochastic Simulation

Autumn Semester 2024

Prof. Fabio Nobile *Assistant: Matteo Raviola*

Lab 11 – 28 November 2024

Markov Chain Monte Carlo

Exercise 1

In many applications of interest, it is not uncommon to encounter the need for sampling from a multi-modal distribution f. The theory developed so far can be directly applicable to these types of distributions. However, in practice, sampling from these distributions using MCMC can be computationally challenging, as we will investigate in this problem. Throughout this exercise, we will consider the bi-modal distribution

$$f(x;\gamma,x_0)=\frac{e^{-\gamma(x^2-x_0)^2}}{Z},\quad \gamma>0, \tag{1}$$

where Z is some normalizing constant. Depending on the values of γ and x_0, designing a sampling strategy to properly sample from (1) can become challenging. Intuitively, if both peaks are too far apart, using a random walk Metropolis (RWM) might not work, as it is possible for the sampler to get stuck on one of the peaks if the *step-size* is too small. Conversely, a RWM with very large *steps* might tend to reject quite often, thus rendering the whole sampling procedure inefficient. We begin by verifying this. Implement the RWM algorithm using as proposal distribution $q(x,y)=\mathcal{N}(x,\sigma^2)$ and target distribution $f(x;\gamma,x_0)$ for $\gamma=1$, $x_0=1,4,9,25$ and different choices of σ. Discuss the quality of your samples by analyzing the trace-plots (one realization of the chain), autocorrelation functions and histograms of the chains obtained.

Exercise 2

Ideally, we would like to obtain (approximately) i.i.d samples from a target distribution f using Markov Chain Monte Carlo (MCMC) algorithms. One practical way of doing so is via *sub-sampling* (also called *batch sampling*), which is implemented to reduce or eliminate correlation between the successive values in the Markov chain. That is, instead of considering the entire chain $\{X_n\colon n\geq 0\}$, say, this technique sub-samples the chain with a batch size $k>1$, so that only the values $\{X_{kn}\colon n\geq 0\}$ are considered. If the covariance $\mathrm{Cov}_f(X_0,X_n)$ vanishes as $n\to\infty$, then the idea of sub-sampling is quite natural since X_{kn} and $X_{k(n+1)}$ can be considered to be approximately independent for k sufficiently big; estimating such a k may be difficult in practice though. While sub-sampling provides a way of generating (approx.) i.i.d. samples from f and may thus be useful assessing the convergence of a MCMC method, it necessarily leads to an efficiency loss. Let $\{X_n\in\mathbb{R}^d\colon n\geq 0\}$ be a Markov chain with a unique stationary distribution f, and $X_0\sim f$ (i.e., the chain is at equilibrium). Take

$\phi\colon \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}_f\big(|\phi|^2\big) < \infty$ and consider two estimators for $\mu = \mathbb{E}_f(\phi)$, namely one that uses the entire Markov chain ($\hat{\mu}$) and one based on sub-sampling ($\hat{\mu}_k$) using only every k-th value:

$$\hat{\mu} = \frac{1}{Nk}\sum_{n=1}^{Nk}\phi(X_n)\,, \quad \text{and} \quad \hat{\mu}_k = \frac{1}{N}\sum_{n=1}^{N}\phi(X_{nk})\,.$$

Show that the variance of $\hat{\mu}$ satisfies $\mathrm{Var}_f(\hat{\mu}) \leq \mathrm{Var}_f(\hat{\mu}_k)$ for every $k > 1$.

Exercise 3

Let $X \subset \mathbb{R}^d$ and $P_i : X \times \mathcal{B}(X) \to [0,1]$, $i = 1\ldots, m$ be a Markov transition kernels on X with $\mathcal{B}(X)$ the associated $\sigma-$algebra.

(a) Given $a_1, \ldots, a_m \in \mathbb{R}^+$, such that $\sum_{i=1}^m a_i = 1$, show that $P(x, A) = \sum_{i=1}^m a_i P_i(x, A)$ is a Markov kernel.

(b) Suppose that a measure $\pi : \mathcal{B} \to [0,1]$ is invariant for each kernel P_i. Show that it is also invariant for $P = \sum_{i=1}^m a_i P_i$, where $a_1, \ldots, a_m \in \mathbb{R}^+$, such that $\sum_{i=1}^m a_i = 1$. If each P_i is reversible, is P reversible?

(c) Under the same assumptions for point (b), define the Markov operator $\mathcal{P}_i$ associated to P_i (i.e., $\pi\mathcal{P}_i = \int P(x,\cdot)d\pi(x)$). Then, show that π is also invariant for $\mathcal{P} = \mathcal{P}_{i_1} \circ \cdots \circ \mathcal{P}_{i_k}$, for any choice of $i_1, \ldots, i_k$. If each P_i is reversible, for which choice of $i_1, \ldots, i_k$ is $\mathcal{P}$ reversible?

Exercise 4

At every iteration of the general Metropolis–Hastings algorithm, a new candidate state $\boldsymbol{Y}_{n+1}$ is proposed by sampling $\boldsymbol{Y}_{n+1} \sim q(\boldsymbol{X}_n, \cdot)$, given the current state $\boldsymbol{X}_n$. Here, $q(\boldsymbol{x}, \boldsymbol{y})$ is the so-called proposal density. Consider now the case where the proposal does not depend on the current state, that is $q(\boldsymbol{x}, \boldsymbol{y}) \equiv q(\boldsymbol{y})$, so that the proposed candidate is $\boldsymbol{Y}_{n+1} \sim q$. This particular Markov Chain Monte Carlo (MCMC) variant is sometimes called *independent Metropolis–Hastings algorithm* with fixed proposal (or simply *independence sampler*). Let's denote the target density by f. As such, this MCMC variant appears very similar to the Accept–Reject method for sampling from f (cf. Lab 02).

1. Suppose there exists a positive constant C such that $f(\boldsymbol{x}) \leq Cq(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathrm{supp}(f) = \{\boldsymbol{x} \in \mathbb{R}^d\colon f(\boldsymbol{x}) > 0\}$. Show that the expected acceptance probability of the independent Metropolis–Hastings algorithm is *at least* $\frac{1}{C}$ whenever the chain is stationary. How does this compare to the expected acceptance probability of an Accept–Reject method?

2. Let us compare the independent Metropolis–Hastings algorithm and the Accept–Reject method in some more detail by an example. Specifically, the goal is to sample from a Gamma distribution with shape parameter α and scale parameter β, denoted by $\mathrm{Gamma}(\alpha, \beta)$, so that the target PDF reads $f(x) \equiv f(x; \alpha, \beta) = \beta^\alpha x^{\alpha-1} e^{-\beta x}/\Gamma(\alpha)\mathbb{I}_{\{x \geq 0\}}$, where $\Gamma(\cdot)$ denotes the Gamma function.

(a) Implement the Accept–Reject method to sample from Gamma$(\alpha, 1)$ for $\alpha > 1$, using the PDF of the Gamma(a, b) distribution with $a = [\alpha]$ as auxiliary density (here $[\alpha]$ denotes the integer part of α).[1] Show that $b = [\alpha]/\alpha$ is the optimal choice for b.

(b) Use your Accept–Reject method to generate m random numbers $X_1, \ldots, X_m$ with each $X_i \sim$ Gamma$(\alpha, 1)$, when using $n = 5000$ random variables $Y_1, \ldots, Y_n$ from the auxiliary Gamma$([\alpha], [\alpha]/\alpha)$ distribution. Notice that m is a random variable, which is smaller than n due to rejections. Perform the simulations for $\alpha = 4.85$.

(c) Implement the independent Metropolis–Hastings algorithm using as proposal q the PDF of the Gamma$([\alpha], [\alpha]/\alpha)$ distribution.

(d) Use the same sample $Y_1, \ldots, Y_n$ used within the Accept–Reject method, now in the corresponding Metropolis–Hastings algorithm to generate $n = 5000$ realizations of the target distribution Gamma$(\alpha, 1)$ with $\alpha = 4.85$.

(e) Compare both methods with respect to:

i. their acceptance rates,

ii. their estimates for the mean of the Gamma$(4.85, 1)$ distribution, which is 4.85,

iii. the correctness of the target distribution,

Discuss your results.

Exercise 5 (Optional)

Consider a Markov chain $\{X_n\} \sim$ Markov(π, P) on a discrete state space $\mathcal{X}$ at equilibrium, with P irreducible, and π the unique invariant probability measure of P. Let l^2_π be the Hilbert space $l^2_\pi = \{f : \mathcal{X} \to \mathbb{R} : \sum_{i\in\mathcal{X}} f(i)^2\pi_i < \infty\}$ with inner product $(f, g)_{l^2_\pi} = \sum_{i\in\mathcal{X}} f(i)g(i)\pi_i$, and $l^2_{\pi,0} = \{f \in l^2_\pi : \mathbb{E}_\pi[f] = 0\}$.

1. Show that if (P, π) are in detailed balance, then $(Pf, g)_{l^2_\pi} = (f, Pg)_{l^2_\pi}$ for any $f, g \in l^2_\pi$

2. Show that $\mathbb{E}[f(X_n)f(X_m)] = (P^{m-n}f, f)_{l^2_\pi}$ for any $f \in l^2_\pi$ and $m > n$.

3. Consider now the estimator
$$\hat{\mu}_N = \frac{1}{N}\sum_{n=1}^{N} f(X_n)$$
of $\mu = \mathbb{E}_\pi[f]$ under the assumption that $f \in l^2_\pi$. Show that $\mathbb{E}_\pi[\hat{\mu}_N] = \mu$, and
$$\mathbb{V}ar[\hat{\mu}_N] = \frac{1}{N}\sum_{l=0}^{N} c_l(P^l\tilde{f}, \tilde{f})_{l^2_\pi},$$
with $\tilde{f} = f - \mathbb{E}_\pi[f] \in l^2_{\pi,0}$ and
$$c_{l,N} = \begin{cases} 1, & l = 0 \\ 2(1 - \frac{l}{N}), & l > 0 \end{cases} \tag{2}$$

[1] *Hint:* *Recall that* $\sum_{k=1}^{K} \xi_k \sim$ Gamma(K, β) *for* $K \in \mathbb{N}$, *if* $\xi_k \overset{\text{i.i.d.}}{\sim}$ Gamma$(1, \beta) \equiv$ Exp(β).

4. Conclude that the asymptotic variance $\mathbb{V}(f,p) \coloneqq \lim_{N\to\infty} N\mathbb{V}ar_\pi(\hat{\mu}_N)$ satisfies $\mathbb{V}(f,p) = ((2(I-P)^{-1}-I)\tilde{f},\tilde{f})_{l^2_\pi}$ if

$$\sup_{g\in l^2_{\pi,0}} \frac{(Pg,g)_{l^2_\pi}}{\|g\|_{l^2_\pi}} = \beta < 1. \tag{3}$$

5. Consider now the two irreducible transition matrices P_1 and P_2, both in detailed balance with π and satisfying (3) for some β_1, β_2. Show that if $(P_1)_{ij} \geq (P_2)_{ij} \forall i \neq j$, then

$$\mathbb{V}(f,P_1) \leq \mathbb{V}(f,P_2), \tag{4}$$

for any $f \in l^2_\pi$.

Hint: Take $P(\lambda) = (1-\lambda)P_1 + \lambda P_2, \lambda \in [0,1]$ and show that $\frac{d}{d\lambda}\mathbb{V}(f,P(\lambda)) \geq 0$.