Stochastic Simulation

Autumn Semester 2024

Prof. Fabio Nobile *Assistant: Matteo Raviola*

Project - 5

Submission deadline: 16 January 2025

Unbiased MCMC using coupling

1 Introduction and background

Let X be a separable metric space (e.g. $\mathsf{X} \subset \mathbb{R}^d$) with Borel σ-algebra $\mathcal{B}(\mathsf{X})$, and let μ, μ^0 be two probability measures on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Markov Chain Monte Carlo (MCMC) algorithms produce samples that are asymptotically distributed according to the target measure μ, by generating an ergodic Markov Chain $\{X^n\}_n \sim \text{Markov}(\mu^0, P)$ started at $\mathsf{X}^0 \sim \mu^0$ and having μ as invariant measure. If we consider now a given μ-integrable function $h : \mathsf{X} \to \mathbb{R}$, its expected value with respect to the target measure μ

$$\mathbb{E}_\mu[h] = \int_{\mathsf{X}} h(x)\mu(\mathrm{d}x) \tag{1}$$

can be estimated by the ergodic estimator

$$\hat{h}_N = \frac{1}{N}\sum_{n=1}^{N} h(X^n).$$

Such estimator is however biased since the chain is not started at stationarity ($\mu^0 \neq \mu$), in general, and the states X^n are only asymptotically distributed according to the target measure μ. It is customary to reduce this bias by discarding the first few, say b, states of the chain (so-called *burn-in* period), thus obtaining the estimator

$$\hat{h}_{N,b} = \frac{1}{N-b+1}\sum_{n=b}^{N} h(X^n).$$

In this project we will introduce and implement an alternative method for generating *unbiased* MCMC estimators for $\mathbb{E}_\mu[h]$ using samples obtained from *coupled* Markov chains. Although this project is sufficiently self-contained so that it can be completed without relying on external references, we invite the student to review also the material in [2], which first proposed this idea.

1.1 Constructing unbiased MCMC estimators

Let us introduce the product space of X with itself, denoted $\mathsf{X}^2 = \mathsf{X} \times \mathsf{X}$, with associated Borel σ-algebra $\mathcal{B}(\mathsf{X}^2)$. The unbiased estimator for $\mathbb{E}_\mu[h]$ proposed in [2] is based on a coupled pair of Markov chains $\{X^n\} \sim \text{Markov}(\mu^0, P)$ and $\{Y^n\} \sim \text{Markov}(\mu^0, P)$, both started from

μ^0, and evolved according to the same Markov transition kernel P. To that end, suppose that one can construct a *joint* Markov transition kernel $\boldsymbol{P} : \mathsf{X}^2 \times \mathcal{B}(\mathsf{X}^2) \to [0,1]$ such that $\boldsymbol{P}((u,v), A \times \mathsf{X}) = P(u,A)$ and $\boldsymbol{P}((u,v), \mathsf{X} \times A) = P(v,A)$, $\forall u,v \in \mathsf{X}$, $A \in \mathcal{B}(\mathsf{X})$, i.e., a Markov transition kernel for which each marginal chain, is a Markov chain generated by P, and consider the following algorithm that generates the coupled chains:

Algorithm 1 Coupled-chain

procedure COUPLED-CHAIN-MCMC(μ^0, P)
  Sample $X^0, Y^0 \sim \mu^0$ and $X^1 \sim P(X^0, \cdot)$.
  for $n \geq 1$ **do**
    if $X^n \neq Y^{n-1}$ **then**
      generate $(X^{n+1}, Y^n) \sim \boldsymbol{P}((X^n, Y^{n-1}), \cdot)$.
    else
      generate $X^{n+1} \sim P(X^n, \cdot)$ and set $Y^n = X^{n+1}$
    end if
  end for
end procedure

It is clear from this algorithm that, after the first meeting time $\tau := \inf\{n \geq 1 : X^n = Y^{n-1}\}$, the two chains will evolve together, i.e. $X^{n+1} = Y^n$, $\forall n \geq \tau$. It is also clear from the construction that each marginal chain $\{X^n\}$, $\{Y^n\}$ is a Markov chain Markov(μ^0, P) with invariant measure μ, so both X^n and Y^n are asymptotically distributed as μ. A practical way to construct a coupling joint kernel $\boldsymbol{P}$ is described in the next section. The idea behind the unbiased estimator of [2] is to rewrite $\mathbb{E}_\mu[h]$ using a telescoping sum argument inspired by [1] in the following way: for any fixed $k \geq 0$

$$\mathbb{E}_\mu[h] = \mathbb{E}[h(X^k)] + \sum_{n=k+1}^{\infty} \mathbb{E}[h(X^n)] - \mathbb{E}[h(X^{n-1})] \tag{2}$$

$$= \mathbb{E}[h(X^k)] + \sum_{n=k+1}^{\infty} \mathbb{E}[h(X^n)] - \mathbb{E}[h(Y^{n-1})] \tag{3}$$

$$= \mathbb{E}\Big[\underbrace{h(X^k) + \sum_{n=k+1}^{\tau-1} \big(h(X^n) - h(Y^{n-1})\big)}_{\hat{H}_k}\Big], \tag{4}$$

which shows that the quantity $\hat{H}_k$ is an unbiased estimator of $\mathbb{E}_\mu[h]$. Since k can be taken arbitrarily, we can further construct a *time-average estimator* $\hat{H}_{b:N} = \frac{1}{N-b+1}\sum_{k=b}^{N} \hat{H}_k$ for fixed integers $0 < b < N$, which can be equivalently written as

$$\hat{H}_{b:N} = \frac{1}{N-b+1}\sum_{n=b}^{N} h(X^n) + \sum_{n=b+1}^{\tau-1} \min\left\{1, \frac{n-b}{N-b+1}\right\}\left[h(X^n) - h(Y^n)\right]. \tag{5}$$

The first term in the sum can be understood as a standard (biased) MCMC estimator $\hat{h}_{N,b}$ with burn-in period b, while the second term can be understood as a bias correction. Finally, for fixed values of N, b, R, one can generate R independent realizations of $\hat{H}^{(r)}_{b:N}$, $r = 1, \ldots, R$, to estimate the variance of the estimator $\hat{H}_{b:N}$ and produce suitable confidence intervals.

In [2], the authors show that under the assumptions that

1. $\mathbb{E}[h(X^n)] \xrightarrow{n\to\infty} \mathbb{E}_\mu[h]$ and $\mathbb{E}[|h(X^n)|^{2+\epsilon}] \leq D$, $\forall n > 0$, for some $\epsilon, D > 0$,
2. the meeting time τ satisfies $\mathbb{P}(\tau \geq n) \leq C\delta^n$, for some $C < +\infty$, $\delta \in (0,1)$,

the estimator $\hat{H}_{b,N}$ is indeed unbiased, with finite variance and finite expected computing time.

1.2 Generating coupled chains

We start by giving the definition of *maximal coupling*. Given two probability measures π, ρ on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$, we recall that the *Total Variation (TV) distance* between π and ρ is given by $d_{\mathsf{TV}}(\pi,\rho) = 2\sup_{A\in\mathcal{B}(\mathsf{X})} |\pi(A) - \rho(A)|$. We say that a probability measure γ on $(\mathsf{X}^2, \mathcal{B}(\mathsf{X}^2))$ is a *coupling* between μ, ρ if $(V,W) \sim \gamma$ implies $V \sim \pi$ and $W \sim \rho$. It can be shown that

$$d_{\mathsf{TV}}(\pi,\rho) \leq 2\mathbb{P}_\gamma(V \neq W), \quad (V,W) \sim \gamma. \tag{6}$$

We say that such a coupling is *maximal* if equation (6) holds with equality.

For simplicity, set now $\mathsf{X} = \mathbb{R}^d$ and assume that the measures π and ρ have Lebesgue densities $p, r : \mathsf{X} \to \mathbb{R}_+$, respectively. Algorithm 2 presents a procedure to generate samples from a maximal coupling between π and ρ.

Algorithm 2 Maximal coupling by acceptance-rejection

```
procedure MAXIMAL-COUPLING(p, r)
    Generate V ~ p and U ~ U([0,1]).
    if U ≤ r(V)/p(V) then
        Set W = V and return (V, W)
    else
        Generate W ~ r and U ~ U([0,1]) until r(W)U > p(W).
        Return (V, W)
    end if
end procedure
```

This maximal coupling can be applied in the context of Metropolis Hastings MCMC by coupling the proposal states. Suppose we want to generate coupled chains $\{X^n\}, \{Y^n\} \sim \text{Markov}(\mu^0, P)$ with invariant measure μ, as in Algorithm 1, using a Metropolis Hastings algorithm with proposal kernel $Q : \mathsf{X} \times \mathcal{B}(\mathsf{X}) \to [0,1]$. Then, at each step n we can propose a (joint) state (V,W) from the maximal coupling between $Q(X^n, \cdot)$ and $Q(Y^{n-1}, \cdot)$ as described in Algorithm 2, and accept or reject V and W separately using the standard Metropolis-Hastings acceptance criterion for the two chains $\{X^n\}$ and $\{Y^n\}$, however using the same uniform random number.

2 Goals of this project

1. Prove equation (6). Then, show that Algorithm 2 produces indeed a maximal coupling of p and r. (Recall that the TV-distance between two probability measures π,ρ with (Lebesgue) densities p and r, respectively, is $d_{\mathsf{TV}}(\pi,\rho) = \int_{\mathsf{X}} |p(x) - r(x)|dx$)

2. Given some fixed N, obtain an expression for the total expected cost (in terms of calls to P) of estimating $\mathbb{E}_\mu[h]$. You may assume that the computational cost of sampling from $\boldsymbol{P}$ is twice that of P. **Hint:** See [2].

3. Consider the probability measure $\mathcal{N}(4,1)$. Let p_μ be the μ-invariant Markov transition kernel induced by a Random walk Metropolis proposal with variance $\sigma^2_{RWM} = 1$, and let $\mu^0 = \mathcal{N}(10,1)$. For different values of b, N compute the expected value of $h(u) = \mathbb{1}_{\{u>3\}}$ using (a) the standard estimator $\hat{h}_{N,b}$ and (b) the *time-averaged estimator* $\hat{H}_{b:N}$ discussed above. Compare your results in terms of cost and accuracy, and in terms of $\mathbb{V}\mathrm{ar}[\mathrm{H_{b:N}}]$ vs V_∞, where V_∞ is the asymptotic variance of the standard MCMC estimator. Present your experimental setup in as much detail as possible.

4. Repeat the previous point with $\mu = \frac{1}{2}\mathcal{N}(-4,1) + \frac{1}{2}\mathcal{N}(4,1)$ and the same hyper-parameters.

5. Propose an adaptive Monte Carlo algorithm to estimate $\mathbb{E}_\mu[h]$ by the estimator $\hat{H}_{b,N}$ with prescribed accuracy and confidence level. Test your adaptive algorithm on the problems of points 3 and 4 for which the exact value $\mathbb{E}_\mu[h]$ is known (or can be easily computed) and assess the robustness of your algorithm.

6. Let us consider the Ising model on a 2D uniform square-lattice of dimension $m \times m$, with atoms placed at each vertex. The atoms can have an upward $(+1)$ or a downward (-1) pointing *spin*. The spin of the atom at position (i,j) in the lattice is denoted with x_{ij}, $1 \le i,j \le m$, so that $x_{ij} \in \{-1,+1\}$. A specific system configuration is hence described by $x = (x_{ij}) \in \{-1,+1\}^{m\times m}$, containing the spin of each of the m^2 atoms. The energy of a given system state is given by

$$H(x) = -\sum_{i,j=1}^{m} \frac{1}{2} J x_{ij}(x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1}), \tag{7}$$

where J is a magnetic coupling constant. To account for boundary effects, we set periodic boundary conditions, i.e., using in (7) $x_{0,j} = x_{m,j}$, $x_{m+1,j} = x_{1,j}$, $x_{j,0} = x_{j,m}$, $x_{j,m+1} = x_{j,1}$. For simplicity, we also assume that $J = 1$. The probability of obtaining a specific system state is then given by the *Boltzmann* distribution with probability mass function

$$f(x) \equiv f_\beta(x) = \frac{1}{Z_\beta} e^{-H(x)\beta}\ , \tag{8}$$

where $\beta = 1/(k_B T)$ denotes the so-called inverse-temperature (or thermodynamic beta) with k_B being the Boltzmann constant and T the absolute temperature. Here, Z_β denotes the normalization constant that makes the target distribution $f_\beta \colon \{-1,+1\}^{m\times m} \to \mathbb{R}_+$ a proper probability mass function. Let us denote by $M(x) = \sum_{i,j=1}^{m} x_{ij}/m^2$ the system's magnetic moment corresponding to the configuration x. Notice that the random realizations of the configuration matrix x depend on the inverse temperature β. The expected value of the magnetic moment $\overline{M}(\beta)$ as a function of the inverse temperature β thus reads

$$\overline{M}(\beta) = \sum_{x\in\mathcal{K}} M(x) f_\beta(x) = \frac{1}{Z_\beta} \sum_{x\in\mathcal{K}} M(x) e^{-H(x)\beta}\ , \tag{9}$$

where $\mathcal{K} = \{-1,1\}^{m\times m}$ is the set of all possible system configurations. Propose an MCMC algorithm to sample from f_β and use the described unbiasing technique in

Algorithm 1 to estimate the expected value of the magnetic moment. Argue that $\mathbb{E}[\overline{M}(\beta)] = 0$ and assess the quality of your estimates. Use lattice size $m = 32$ and experiment different values of $\beta \in (0.2, 0.45)$.

References

[1] P. W. Glynn and C.-H. Rhee. Exact estimation for markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

[2] Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased Markov Chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020.