



HAL
open science

The importance Markov Chain

Charly Andral, Randal Douc, Hugo Marival, Christian Robert

► **To cite this version:**

Charly Andral, Randal Douc, Hugo Marival, Christian Robert. The importance Markov Chain. Stochastic Processes and their Applications, 2024, 171, pp.2-26. 10.1016/j.spa.2024.104316 . hal-03912132

HAL Id: hal-03912132

<https://hal.science/hal-03912132v1>

Submitted on 26 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Importance Markov Chain

Charly Andral¹, Randal Douc², Hugo Marival², and Christian P. Robert^{1,3}

¹ CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016, Paris, France

² SAMOVAR, Telecom Sudparis, Institut Polytechnique de Paris, 9 rue Charles Fourier, 91820, Evry, France

³ Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
Emails: {andral, xian}@ceremade.dauphine.fr, {randal.douc, hugo.marival}@telecom-sudparis.eu

February 26, 2024

Abstract

The Importance Markov chain is a novel algorithm bridging the gap between rejection sampling and importance sampling, moving from one to the other through a tuning parameter. Based on a modified sample of an instrumental Markov chain targeting an instrumental distribution (typically via a MCMC kernel), the Importance Markov chain produces an extended Markov chain where the marginal distribution of the first component converges to the target distribution. For example, when targeting a multimodal distribution, the instrumental distribution can be chosen as a tempered version of the target which allows the algorithm to explore its modes more efficiently. We obtain a Law of Large Numbers and a Central Limit Theorem as well as geometric ergodicity for this extended kernel under mild assumptions on the instrumental kernel. Computationally, the algorithm is easy to implement and preexisting libraries can be used to sample from the instrumental distribution.

Keywords— Markov chain Monte Carlo, importance sampling, Monte Carlo methods, ergodicity, regeneration

1 Introduction

In Monte Carlo methods [1] and in particular in computational Bayesian statistics, sampling is used to construct estimates for quantities depending on problem-specific distributions. As a first approach, one can simulate independently according to another distribution, called the *instrumental* distribution, and use this sample to build an estimate of the quantity of interest (see, e.g., [2] for a general introduction to Monte Carlo methods). The most well-known example is the importance sampling (IS) technique [3], which produces a weighted sample to approximate $\pi(f) = \int f(x)\pi(x)dx$ where π is a given distribution (by an abuse of notation, we also denote π its density with respect to a dominating measure dx). Importance sampling is based on rewriting the quantity of interest as $\pi(f) = \int f(x)\frac{\pi(x)}{\tilde{\pi}(x)}\tilde{\pi}(x)dx$ for any density $\tilde{\pi}$ that dominates π . Then, $\pi(f)$ can be estimated by sampling independently X_1, X_2, \dots from the instrumental distribution $\tilde{\pi}$ and by returning the estimate $\tilde{I} = n^{-1} \sum_{i=1}^n \frac{\pi(X_i)}{\tilde{\pi}(X_i)} f(X_i)$. It is fundamental to recall here that importance sampling does not deliver a sample distributed from π . In contrast, rejection sampling allows to construct a perfect sample according to π but at the cost that a portion of the sampled points are rejected. To be more specific, if we assume that $\pi \leq M\tilde{\pi}$ for some constant M , then we sample

independently $X_1, X_2, \dots \sim_{iid} \tilde{\pi}$ and $U_1, U_2, \dots \sim_{iid} \mathcal{U}(0, 1)$ until the condition $U_i < \frac{\pi(X_i)}{M\tilde{\pi}(X_i)}$ is met. For the exit index i , setting $Y = X_i$, it turns out that the law of the accepted candidate Y is then exactly π (see [4]).

Another approach is proposed by Markov Chain Monte Carlo (MCMC) methods: instead of constructing an independent and identically distributed (iid) sample, an MCMC algorithm provides a Markov chain (thus a dependent sample), that converges to the distribution of interest. The most common MCMC algorithm is the Metropolis-Hastings algorithm [5, 6]. Note that MCMC and IS are not incompatible, and the idea of using a Markov chain for the *instrumental* distribution appeared as soon as 1963 [7], and is mentioned by Hasting in [6]. More recently, many algorithms combine IS and MCMC [8, 9, 10].

The Importance Markov Chain (IMC) algorithm uses those ideas in a novel way. Indeed, while most MCMC algorithms try to adjust the proposals to explore the support of the target distribution efficiently, the IMC algorithm allows to target a more friendly *instrumental* distribution which is then transformed into the initial target with IS. More specifically, the *instrumental* Markov chain is transformed into an *augmented* Markov chain targetting the distribution of interest on its first marginal. This is different from classic subsampling or thinning of the chain that preserve the distribution [11, 12, 13]. The instrumental distribution is considered as a given. Indeed, our aim in this paper is to establish properties that are preserved by our transformation for a given *instrumental* distribution—namely a law of large numbers (LLN), a Central Limit Theorem (CLT) and geometric ergodicity.

Of course, adding a resampling step to a classical importance sampling based on a $\tilde{\pi}$ -sample $(\tilde{X}_1, \dots, \tilde{X}_n)$ may lead to a random variable with distribution $\hat{\pi}_n$ close to the target distribution π . But the total variation norm between the two distributions $\hat{\pi}_n$ and π is typically of order $O(1/n)$ whereas our Importance Markov chain, under mild assumptions, is geometrically ergodic, showing that the decrease in the total variation norm may be geometrically fast with respect to n .

The Importance Markov chain, in the specific setting with independent proposals, is related to previous works on *Self Regenerative Markov chains* [14, 15] and on Dynamic Weighting Monte Carlo [16, 17]. It was developed in the dependent case in [18] but the framework there was restrained to a semi-Markov formulation.

The article is organized as follows:

1. We first define the *Rejection Markov chain*, a generalization of the rejection sampling in a context of MCMC sampling. This part allows us to define the rejection kernel used further on, and provides some intuition for the algorithm in the specific case where there exists a known constant M such that the density ratio $\frac{\pi}{M\tilde{\pi}}$ is upper-bounded by 1.
2. We then generalize the Rejection Markov chain using repetitions to allow the density ratio $\frac{\pi}{M\tilde{\pi}}$ to be greater than 1, thereby relieving the constraint of the first part. The idea is similar to IS as the number of repetitions is proportional to the density ratio, to the exceptions that: (1) the number of repetitions is a random integer and the constraint is simply that its expectation is proportional to the density ratio; (2) the output is a true random sample and not a weighted one as in classical IS. We use an extended space to construct an augmented Markov chain composed of repetitions of the instrumental chain as its first component and an integer as its second, keeping track of the remaining number of remaining repetitions. We then proceed to establish some theoretical properties, under mild assumptions, notably a law of large numbers, a Central Limit Theorem, a geometric ergodicity property and some uniqueness results.
3. Finally, we illustrate the IMC on two synthetic examples. The first is a multidimensional mixture of Gaussian distributions, using as instrumental distribution a tempered version of the target, and a NUTS kernel. The second focuses on an i.i.d. sample from the instrumental distribution, defined by a normalizing flow trained to approximate a multimodal target up to dimension 25.

2 Notations

Let us denote $\mathbb{R}^+ = [0, \infty)$, $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. We use the standard convention that $\prod_{k=m}^n = 1$ if $m > n$. For integers $k \leq \ell$, the notation $[k : \ell]$ stands for the set $\{k, \dots, \ell\}$ and in case where $k > \ell$, $[k : \ell]$ is the empty set. Moreover, $u_{k:\ell} = (u_k, u_{k+1}, \dots, u_\ell)$ for all $k \leq \ell$ and $u_{k:\infty} = (u_\ell)_{\ell \geq k}$. If a space X is equipped with a σ -field \mathcal{X} , we denote by $F_+(\mathsf{X})$ the set of all nonnegative measurable functions with respect to \mathcal{X} , that is, we make implicit the dependence on the σ -field \mathcal{X} in the notation $F_+(\mathsf{X})$. Similarly $F_b(\mathsf{X})$ is the set of all bounded measurable functions on X and $F_{b+}(\mathsf{X}) = F_b(\mathsf{X}) \cap F_+(\mathsf{X})$. Moreover, we denote by $M_1(\mathsf{X})$ the set of probability measures on $(\mathsf{X}, \mathcal{X})$. For a non-negative real number x , we denote the floor function by $[x]$ and the fractional part by $\langle x \rangle$ and hence $x = [x] + \langle x \rangle$. The positive part of x is written $(x)^+$.

If P, Q are Markov kernels on $\mathsf{X} \times \mathcal{X}$, h a measurable function on X , ν a measure on $(\mathsf{X}, \mathcal{X})$ we define:

- $Ph(x) := \int_{\mathsf{X}} P(x, dy)h(y)$, for $x \in \mathsf{X}$,
- $\nu P(A) := \int_{\mathsf{X}} \nu(dx)P(x, A)$ for $A \in \mathcal{X}$,
- $PQ(x, A) := \int_{\mathsf{X}} P(x, dy)Q(y, A)$ for $x \in \mathsf{X}$ and $A \in \mathcal{X}$,
- $\nu(h) := \int_{\mathsf{X}} h(x)\nu(dx)$, also denoted νh if the context is clear.

Furthermore, we simply denote $P^0 := I$ and for $k \in \mathbb{N}$, $P^k := PP^{k-1} = P^{k-1}P$. Last, a kernel P is said *sub-Markovian* if for all $x \in \mathsf{X}$, $P(x, \mathsf{X}) \leq 1$.

3 The Rejection Markov chain

Rejection sampling is a standard in Monte Carlo simulation. In this algorithm, we create samples distributed according to the target π by subsampling among a batch of iid random variables distributed according to the instrumental $\tilde{\pi}$.

Our Markov rejection algorithm closely resembles the rejection algorithm except that we no longer need to subsample from an i.i.d. batch of random variables exactly distributed according to $\tilde{\pi}$, which can be restrictive. Instead, we rely on a Markov chain targeting $\tilde{\pi}$, i.e. generated by a Markov kernel with invariant probability $\tilde{\pi}$, which can be achieved for example via a Metropolis Hastings algorithm. Note that it is sufficient to know $\tilde{\pi}$ up to a normalizing constant. Subsampling is done by accepting each candidate sample X with probability $\rho(X)$ where $\rho : \mathsf{X} \rightarrow [0, 1]$ is a well-chosen function.

3.1 Formal definition

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. For a given Markov kernel Q on $\mathsf{X} \times \mathcal{X}$, we denote by \mathbb{P}_ξ^Q the probability measure induced on $(\mathsf{X}^{\mathbb{N}_0}, \mathcal{X}^{\otimes \mathbb{N}_0})$ by the Markov kernel Q and the initial distribution ξ , and by \mathbb{E}_ξ^Q the associated expectation operator. If $\xi = \delta_x$ for some $x \in \mathsf{X}$, we simply use $\mathbb{E}_x^Q := \mathbb{E}_{\delta_x}^Q$. On $(\mathsf{X}^{\mathbb{N}_0}, \mathcal{X}^{\otimes \mathbb{N}_0})$, we define X_ℓ as the projection on ℓ^{th} -component, i.e., for any $w = (w_\ell)_{\ell \in \mathbb{N}_0} \in \mathsf{X}^{\mathbb{N}_0}$, $X_\ell(w) = w_\ell$, and θ the shift operator on $\mathsf{X}^{\mathbb{N}_0}$ such that $\theta : (w_0, w_1, \dots) \mapsto (w_1, w_2, \dots)$. For any measurable function $\rho : \mathsf{X} \rightarrow [0, 1]$, the rejection kernel S is defined as follows:

$$Sh(x) = \sum_{k=1}^{\infty} \mathbb{E}_x^Q \left[h(X_k) \rho(X_k) \prod_{i=1}^{k-1} (1 - \rho(X_i)) \right], \quad (1)$$

where $x \in \mathsf{X}$ and h is any nonnegative measurable function on $(\mathsf{X}, \mathcal{X})$. A transition according to S is obtained by generating a Markov chain $\{X_k : k \in \mathbb{N}_0\}$ according to the kernel Q and by selecting the first accepted candidate among $\{X_k : k \in \mathbb{N}_0\}$ with the success probability sequence $\{\rho(X_k) : k \in \mathbb{N}_0\}$. More precisely, define $\mathsf{Y} = \mathsf{X} \times [0, 1]$ and $\mathcal{Y} = \mathcal{X} \otimes \mathcal{B}([0, 1])$ and let G be the Markov kernel on $\mathsf{Y} \times \mathcal{Y}$ such that for all $y = (x, u) \in \mathsf{Y}$ and all $A \in \mathcal{Y}$,

$$G(y, A) = \int_{\mathsf{Y}} \mathbf{1}_A(x', u') Q(x, dx') du'. \quad (2)$$

Therefore if $Y' = (X', U') \sim G(y, \cdot)$, then X' and U' are independent and marginally, $X' \sim Q(x, \cdot)$ and $U' \sim \mathcal{U}(0, 1)$. For the Markov chain $\{Y_k = (X_k, U_k) : k \in \mathbb{N}_0\}$ with Markov kernel G , define the first return time to the set $D = \{y = (x, u) \in Y : u \leq \rho(x)\}$ by

$$\sigma_D = \inf \{k \geq 1 : Y_k \in D\}. \quad (3)$$

Then, $Sh(x) = \mathbb{E}_{\delta_x \otimes \gamma}^G[\mathbf{1}_{\{\sigma_D < \infty\}} h(X_{\sigma_D})]$ where γ is any probability measure on $[0, 1]$, showing on the side that $S\mathbf{1}_X(x) = \mathbb{P}_{\delta_x \otimes \gamma}^G(\sigma_D < \infty) \leq 1$, which implies that the kernel S is sub-Markovian.

Proposition 1. *Let Q be a Markov kernel on $X \times X$ with invariant probability measure $\mu \in \mathbf{M}_1(X)$ and let $\rho : X \rightarrow [0, 1]$ be a measurable function. Provided that $\mu(\rho) > 0$, the probability measure ν defined by*

$$\nu(A) = \frac{\int_A \rho(x) \mu(dx)}{\int_X \rho(x) \mu(dx)}, \quad A \in \mathcal{X}, \quad (4)$$

is invariant with respect to S , i.e. $\nu S = \nu$.

Proof. Define $\bar{\rho} := 1 - \rho$. For any bounded function $h \in F_+(X)$ and any $x \in X$,

$$\begin{aligned} Sh(x) &= \mathbb{E}_x^Q \left[\sum_{k=1}^{\infty} \rho(X_k) h(X_k) \prod_{i=1}^{k-1} \bar{\rho}(X_i) \right] \\ &= \mathbb{E}_x^Q [\rho(X_1) h(X_1)] + \mathbb{E}_x^Q \left[\bar{\rho}(X_1) \sum_{\ell=1}^{\infty} \rho(X_{\ell+1}) h(X_{\ell+1}) \prod_{j=1}^{\ell-1} \bar{\rho}(X_{j+1}) \right] \\ &= Q(\rho h)(x) + Q(\bar{\rho} Sh)(x). \end{aligned} \quad (5)$$

Integrating with respect to μ yields

$$\mu Sh = \mu Q(\rho h) + \mu Q(\bar{\rho} Sh) = \mu(\rho h) + \mu(\bar{\rho} Sh),$$

where we used $\mu Q = \mu$ in the last equality. Since h is bounded, $\mu Sh < \infty$. Retrieving μSh on both sides, we finally obtain $\mu(\rho Sh) = \mu(\rho h)$. Hence $\nu(Sh) = \nu(h)$. \square

3.2 Application to sampling

Let $\pi \in \mathbf{M}_1(X)$ be the *target* distribution and denote by $\tilde{\pi} \in \mathbf{M}_1(X)$ an *instrumental* distribution. As for rejection or importance sampling, the goal is to produce a sample targeting π using a sample of $\tilde{\pi}$, here obtained by using a Markov kernel Q . We denote $\{\tilde{X}_i : i \in \mathbb{N}_0\}$ a Markov chain with transition kernel Q and make the following hypothesis on Q :

(H1) The Markov kernel Q admits $\tilde{\pi}$ as invariant probability measure.

We also need the following domination assumption, which is compulsory for rejection sampling.

(H_{rej}) There exists $M > 0$ such that $\pi \leq M\tilde{\pi}$.

Then, we can use (4) to define ρ such that π is the invariant probability measure for the Markov kernel S . Indeed, if $\mu = \tilde{\pi}$, we get $\nu = \pi$ in (4) by defining

$$\rho \propto \frac{d\pi}{d\tilde{\pi}}.$$

If in addition, we want ρ to take values in $[0, 1]$, we may pick

$$\rho(x) = \frac{1}{M} \frac{d\pi}{d\tilde{\pi}}(x), \quad (6)$$

for $\tilde{\pi}$ -almost all $x \in X$. From Proposition 1, we deduce immediately:

Theorem 2. *Assume (H1) and (H_{rej}) and take ρ as defined in (6). Then S is π -invariant.*

Note that standard rejection sampling consists in applying a transition according to S to the particular case where $Q(x, \cdot) = \tilde{\pi}(\cdot)$, $\mu = \tilde{\pi}$, ρ as in (6) and hence $\nu = \pi$.

4 The Importance Markov chain

As for classical rejection sampling, the Markov chain rejection sampling suffers from the drawback that $(\mathbf{H}_{\text{rej}})$ may be satisfied only with a prohibitively large M (or worse, $(\mathbf{H}_{\text{rej}})$ may not even be satisfied). The sampling is in that case inefficient since the average acceptance ratio is equal to $1/M$. Actually, ρ can be interpreted in the following way: given \tilde{X}_{k-1} , we draw a new point $\tilde{X}_k \in \mathcal{X}$ according to Q and insert $\tilde{N}_k \sim \text{Ber}(\rho(\tilde{X}_k))$ replica in current sample. Therefore $\rho(\tilde{X}_k) = \mathbb{E}[\tilde{N}_k | \tilde{X}_k]$. This can be viewed as the equivalent of the weight of \tilde{X}_k in a importance sampling context.

The idea with $\varrho : \mathcal{X} \rightarrow \mathbb{R}^+$ is similar but we now offer to replicate \tilde{X}_k a random number of times \tilde{N}_k where the conditional expectation of the random integer $\tilde{N}_k \in \mathbb{N}_0$ w.r.t. \tilde{X}_k is equal to $\varrho(\tilde{X}_k)$, as sketched in Algorithm 1. This relates to [15] where the author conditions the weights of his estimate to be nonnegative integers.

4.1 The extended Markov chain

Let π and $\tilde{\pi}$ be two probability measures on $(\mathcal{X}, \mathcal{X})$ and let κ be a positive real number. Assume that π is dominated by $\tilde{\pi}$ and let $\varrho_\kappa : \mathcal{X} \rightarrow \mathbb{R}^+$ be a measurable function such that

$$\varrho_\kappa(x) = \kappa \frac{d\pi}{d\tilde{\pi}}(x), \quad (7)$$

for $\tilde{\pi}$ -almost all $x \in \mathcal{X}$. For $\kappa = 1$ let us simply denote

$$\varrho := \varrho_1. \quad (8)$$

Let \tilde{R} be a Markov kernel on $\mathcal{X} \times \mathcal{P}(\mathbb{N}_0)$ and by an abuse of notation, let us write $\tilde{R}(x, n) = \tilde{R}(x, \{n\})$ for any $(x, n) \in \mathcal{X} \times \mathbb{N}_0$. The distribution $\tilde{R}(\tilde{X}_k, \cdot)$ will be used to draw the number \tilde{N}_k of replications of \tilde{X}_k under the *unbiasedness assumption*:

(H2) For all $x \in \mathcal{X}$,

$$\sum_{n=0}^{\infty} \tilde{R}(x, n)n = \varrho_\kappa(x).$$

In this paper, we will face two cases : (1) \tilde{R} and ϱ_κ are available in closed form up to a normalizing constant, (2) \tilde{R} can be generated under (H2) without the explicit knowledge of ϱ_κ . Furthermore, κ can be chosen arbitrarily even though we will discuss some interesting choices later on (see Section 4.6.2 and Section 6.1.2).

Algorithm 1

- 1: $X = []$
 - 2: Set an arbitrary \tilde{X}_0 .
 - 3: **for** $k \leftarrow 1$ to n **do**
 - 4: Draw $\tilde{X}_k \sim Q(\tilde{X}_{k-1}, \cdot)$ and $\tilde{N}_k \sim \tilde{R}(\tilde{X}_k, \cdot)$
 - 5: Append \tilde{N}_k replicas of \tilde{X}_k to X
 - 6: **end for**
 - 7: **output:** X
-

As seen below, if the Markov kernel Q is $\tilde{\pi}$ -invariant, then the output sequence $X = (X_0, X_1, \dots)$ of Algorithm 1 targets π . However, X is not a Markov chain *per se*, and in order to study its ergodic properties, we need add a second component N to the sequence X so that the augmented sequence (X, N) becomes a Markov chain. This is done by rewriting Algorithm 1 as Algorithm 2.

Let us describe the transition of the extended Markov chain $\{(X_\ell, N_\ell) : \ell \in \mathbb{N}_0\}$. From Algorithm 2, we can see that (X_ℓ, N_ℓ) is updated according to two different moves. Either we are already inside the while loop described in lines 6-9 of Algorithm 2, in which case, $N_\ell \geq 1$ and the update is simply $(X_{\ell+1}, N_{\ell+1}) = (X_\ell, N_\ell - 1)$, or we are outside the while loop, in which case,

Algorithm 2 Importance Markov chain (IMC)

```

1:  $\ell \leftarrow 0$ 
2: Set an arbitrary  $\tilde{X}_0$ .
3: for  $k \leftarrow 1$  to  $n$  do
4:   Draw  $\tilde{X}_k \sim Q(\tilde{X}_{k-1}, \cdot)$  and  $\tilde{N}_k \sim \tilde{R}(\tilde{X}_k, \cdot)$ 
5:   Set  $N_\ell = \tilde{N}_k$ 
6:   while  $N_\ell \geq 1$  do
7:     Set  $(X_\ell, N_\ell) \leftarrow (\tilde{X}_k, N_\ell - 1)$ 
8:     Set  $\ell \leftarrow \ell + 1$ 
9:   end while
10: end for

```

$N_\ell = 0$ and $X_\ell = \tilde{X}_k$ for some $k \in \mathbb{N}_0$. Then, the update of (X_ℓ, N_ℓ) happens when we enter again the while loop, in which case $(X_{\ell+1}, N_{\ell+1}) = (\tilde{X}_T, \tilde{N}_T - 1)$ where $T = \inf \{n > k : \tilde{N}_n \neq 0\}$.

The associated Markov kernel P on $(\mathsf{X} \times \mathbb{N}_0) \times (\mathcal{X} \otimes \mathcal{P}(\mathbb{N}_0))$ is then defined by: for all $h \in \mathsf{F}_+(\mathsf{X} \times \mathbb{N}_0)$,

$$Ph(x, n) = \mathbf{1}_{\{n \geq 1\}} h(x, n-1) + \mathbf{1}_{\{n=0\}} \cdot \sum_{k, \ell=1}^{\infty} \mathbb{E}_x^Q \left[h(X_k, \ell-1) \tilde{R}(X_k, \ell) \prod_{i=1}^{k-1} \tilde{R}(X_i, 0) \right]. \quad (9)$$

To simplify this expression, let us introduce additional notation. Write $\rho_{\tilde{R}}(x) = \tilde{R}(x, [1 : \infty))$ and let S be the Markov kernel on $\mathsf{X} \times \mathcal{X}$ defined by

$$Sf(x) = \sum_{k=1}^{\infty} \mathbb{E}_x^Q \left[f(X_k) \rho_{\tilde{R}}(X_k) \prod_{i=1}^{k-1} (1 - \rho_{\tilde{R}}(X_i)) \right], \quad (10)$$

for $f \in \mathsf{F}_+(\mathsf{X})$.

The kernel S is of the same form as in (1) except that ρ is now replaced by $\rho_{\tilde{R}}$. Note that by construction, $\rho_{\tilde{R}}$ is $[0, 1]$ -valued and $\rho_{\tilde{R}}(x)$ can be interpreted as the probability for an $\tilde{X}_i = x$ drawn from Q to be accepted for the chain (X_ℓ) , in which case, we keep at least one replica of \tilde{X}_i .

Then, the extended Markov kernel P writes, for $h \in \mathsf{F}_+(\mathsf{X} \times \mathbb{N}_0)$:

$$Ph(x, n) = \mathbf{1}_{\{n \geq 1\}} h(x, n-1) + \mathbf{1}_{\{n=0\}} \sum_{n'=0}^{\infty} \int_{\mathsf{X}} S(x, dx') R(x', n') h(x', n'), \quad (11)$$

where R is the Markov kernel on $\mathsf{X} \times \mathcal{P}(\mathbb{N}_0)$ defined by

$$R(x, n) := \tilde{R}(x, n+1) / \rho_{\tilde{R}}(x), \quad (x, n) \in \mathsf{X} \times \mathbb{N}_0, \quad (12)$$

and where we, again, make the abuse of notation $R(x, n) := R(x, \{n\})$.

Note that (9) and (11) give two different but equivalent decompositions of P , for the sampling step (when $n = 0$). In (9), we sample \tilde{X}_i according to Q and then use $\tilde{R}(\tilde{X}_i, \cdot)$ to draw a number of replicas \tilde{N}_i , until it is larger than 1, in which case we retain \tilde{X}_i and $\tilde{N}_i - 1$, the number of remaining replicas. In (11), we bypass the rejection step by drawing directly a new accepted point X_i using S and then the number of remaining replicas from $R(X_i, \cdot)$, which corresponds to the law of $\tilde{N} - 1$ conditionnally on $\{\tilde{N} \geq 1\}$ when $\tilde{N} \sim \tilde{R}(X_i, \cdot)$.

Remark 1. *The unbiasedness assumption (H2) is closely related to the notion of correctly weighted density developed in [16, 17]. Write $\hat{\pi}(dx dn) = \hat{\pi}(dx) \tilde{R}(x, dn)$ a joint distribution on $\mathsf{X} \times \mathbb{N}_0$. Then, under (H2), $\sum_{n \in \mathbb{N}} \hat{\pi}(dx) \tilde{R}(x, n) n = \kappa \pi(dx)$ so $\hat{\pi}$ is correctly weighted. And by construction, the kernel $Q(x, dy) \tilde{R}(y, dn)$ that generates the samples $(\tilde{X}_i, \tilde{N}_i)_{i \in \mathbb{N}}$ of Algorithm 1 admits $\hat{\pi}$ as an invariant probability distribution (see Lemma 6).*

Remark 2. While importance sampling requires exact simulations from $\tilde{\pi}$, the IMC method only relies on a Markov kernel Q targeting $\tilde{\pi}$. This allows us to extend the set of usable instrumental distributions.

Remark 3. Perhaps surprisingly, Metropolis-Hastings (MH) algorithms can be cast into the framework of importance Markov chains. Indeed, take a MH algorithm with proposal kernel $A(x, dy)$ and acceptance rate $\alpha(x, y)$, targeting π . Following the framework of [19], the accepted points $\{\tilde{X}_i : i \in \mathbb{N}_0\}$ form a Markov chain with Markov kernel $Q(x, dy)$ proportional to $\alpha(x, y)A(x, dy)$. Before moving to a new accepted point, $\tilde{X}_i = x$ is repeated a random number of times \tilde{N}_i that follows (conditionally on $\tilde{X}_i = x$) a geometric distribution with success probability $p(x) := \int_{\mathbf{X}} \alpha(x, y)A(x, dy)$. Then it can be shown that Q is $\tilde{\pi}$ -invariant where $\tilde{\pi}(dx) \propto p(x)\pi(dx)$, hence (H1) holds. Define $\tilde{R}(x, \cdot)$ as the geometric distribution with parameter $p(x)$. Choosing $\kappa = 1 / \int_{\mathbf{X}} p(x)\pi(dx)$, ϱ_κ defined in (7) writes $\varrho_\kappa(x) = 1/p(x)$, hence (H2) holds. Then, with these choices of Q and \tilde{R} , $(\tilde{X}_i, \tilde{N}_i)$ corresponds to the output of the IMC algorithm defined in Algorithm 1.

4.2 Invariant probability measure

4.2.1 Existence

Let $\tilde{\pi}$ be the measure on $\mathbf{X} \times \mathbb{N}_0$ defined by: for any $h \in \mathbf{F}_+(\mathbf{X} \times \mathbb{N}_0)$,

$$\begin{aligned} \tilde{\pi}(h) &= \kappa^{-1} \sum_{n=1}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \tilde{R}(x, n) \sum_{k=0}^{n-1} h(x, k) \\ &= \kappa^{-1} \sum_{\ell=0}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) R(x, \ell) \sum_{k=0}^{\ell} h(x, k), \end{aligned} \quad (13)$$

where the last equality follows from (12) and the change of variable $\ell = n - 1$.

Proposition 3. Assume (H1) and (H2). Let P be the Markov kernel defined in (11) and let $\tilde{\pi}$ be the probability measure on $\mathbf{X} \times \mathbb{N}_0$ defined in (13). Then,

- (i) the Markov kernel P is $\tilde{\pi}$ -invariant,
- (ii) the marginal of $\tilde{\pi}$ on the first component is π .

Proof. We start with (i). Let $h \in \mathbf{F}_+(\mathbf{X} \times \mathbb{N}_0)$. Interchanging the sum in ℓ and the sum in k in (13) yields

$$\tilde{\pi}(h) = \kappa^{-1} \sum_{k=0}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) R(x, [k : \infty)) h(x, k). \quad (14)$$

We now replace h by Ph and combine with the expression of Ph given in (11), we then obtain

$$\begin{aligned} \tilde{\pi}(Ph) &= \kappa^{-1} \sum_{k=1}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) R(x, [k : \infty)) h(x, k-1) + \kappa^{-1} \sum_{n'=0}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) \int_{\mathbf{X}} S(x, dx') R(x', n') h(x', n') \\ &= \kappa^{-1} \sum_{n'=0}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) R(x, [n'+1 : \infty)) h(x, n') + \kappa^{-1} \sum_{n'=0}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) R(x, n') h(x, n'), \end{aligned}$$

where the last equality follows (a) from the change of variable $n' = k - 1$ for the first term of the rhs and (b) from Proposition 1 applied, under (H1), to $\rho = \rho_{\tilde{R}}$ and $\mu = \tilde{\pi}$ for the second term. Noting that $R(x, [n'+1 : \infty)) + R(x, n') = R(x, [n' : \infty))$, we finally get

$$\tilde{\pi}(Ph) = \kappa^{-1} \sum_{n'=0}^{\infty} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\tilde{R}}(x) R(x, [n' : \infty)) h(x, n') = \tilde{\pi}(h), \quad (15)$$

where (14) is used to obtain the last equality.

We now turn to (ii). For any $A \in \mathcal{X}$, applying (13) with $h(x, k) = \mathbf{1}_A(x)$ yields under (H2)

$$\bar{\pi}(A \times \mathbb{N}_0) = \kappa^{-1} \int_{\mathbf{X}} \bar{\pi}(dx) \left(\sum_{n=1}^{\infty} \tilde{R}(x, n) n \right) \mathbf{1}_A(x) = \kappa^{-1} \bar{\pi}(\varrho_{\kappa} \mathbf{1}_A) = \pi(A). \quad (16)$$

□

4.3 Uniqueness

Proposition 4. *Assume (H1) and (H2). If any invariant measure for Q is proportional to $\bar{\pi}$ (defined in (H1)), then $\bar{\pi}$ defined in (13) is the unique invariant probability measure for P .*

Proof. See A. □

The uniqueness of the invariant probability measure $\bar{\pi}$ for P , as stated in Theorem 4 allows to obtain the Birkhoff ergodic theorem ([20, Theorem 5.2.9]): for any measurable function $g : \mathbf{X} \times \mathbb{N}_0 \rightarrow \mathbb{R}$ such that $\bar{\pi}(|g|) < \infty$,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} g(X_k, N_k) = \bar{\pi}(g), \quad \mathbb{P}_{\bar{\pi}}^P - a.s. \quad (17)$$

Although reassuring, the law of large numbers holds $\mathbb{P}_{\bar{\pi}}^P - a.s.$, i.e. the initial distribution is set to be the invariant probability measure $\bar{\pi}$, which is not realistic from a practical point of view. Consequently, we will now turn to conditions under which the law of large numbers holds, irrespective to the initial distribution.

4.4 Law of large numbers

To establish a strong law of large numbers for the kernel P , we rely on the single hypothesis that the instrumental kernel Q satisfies a law of large numbers. More precisely if the instrumental kernel Q satisfies a law of large numbers for any initial distribution $\xi \in \mathbf{M}_1(\mathbf{X})$, Theorem 5 will show that it is also the case for the importance Markov kernel P .

(H_{lln}) For every $\xi \in \mathbf{M}_1(\mathbf{X})$ and measurable function $g : \mathbf{X} \rightarrow \mathbb{R}$ such that $\bar{\pi}(|g|) < \infty$,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} g(\tilde{X}_k) = \bar{\pi}(g), \quad \mathbb{P}_{\xi}^Q - a.s.$$

Theorem 5. *Assume (H1) and (H_{lln}). Then, for every $\xi \in \mathbf{M}_1(\mathbf{X} \times \mathbb{N}_0)$ and measurable function $g : \mathbf{X} \times \mathbb{N}_0 \rightarrow \mathbb{R}$ such that $\bar{\pi}(|g|) < \infty$,*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} g(X_k, N_k) = \bar{\pi}(g), \quad \mathbb{P}_{\xi}^P - a.s. \quad (18)$$

Proof. The proof relies on [21, Proposition 3.5], which relates (H_{lln}) to a property on the harmonic functions for Q (i.e. measurable functions h such that $Qh = h$). More precisely, it states that for any Markov kernel Q satisfying $\bar{\pi}Q = \bar{\pi}$ for some $\bar{\pi} \in \mathbf{M}_1(\mathbf{X})$, (H_{lln}) is equivalent to (H_{hnm}) defined as follows:

(H_{hnm}) Any bounded harmonic function $h : \mathbf{X} \rightarrow \mathbb{R}$ for Q is constant.

Hence, proving Theorem 5 is equivalent to showing that any bounded harmonic function for P is constant.

Let $\bar{h} : \mathbf{X} \times \mathbb{N}_0 \rightarrow \mathbb{R}$ be a bounded harmonic function for P . Then for $n > 0$,

$$\bar{h}(x, n) = P\bar{h}(x, n) = \bar{h}(x, n-1), \quad (19)$$

where the last equality comes from (11). Thus \bar{h} does not depend on its second argument and we can define a measurable function $h : \mathsf{X} \rightarrow \mathbb{R}$ such that

$$h(x) = \bar{h}(x, n) \quad (20)$$

for all $(x, n) \in \mathsf{X} \times \mathbb{N}_0$. Now,

$$h(x) = \bar{h}(x, 0) = P\bar{h}(x, 0) = \int_{\mathsf{X} \times \mathbb{N}_0} S(x, dx') R(x', dn) \bar{h}(x, n) = Sh(x)$$

using the expression of P in (11) as well as (20). From the recursive expression for S in (5) we have:

$$\begin{aligned} h(x) &= Sh(x) \\ &= Q(\rho h)(x) + Q((1 - \rho)Sh)(x) \\ &= Q(\rho h)(x) + Q((1 - \rho)h)(x) \\ &= Qh(x). \end{aligned}$$

Therefore h is harmonic for Q , and since it is also bounded, (\mathbf{H}_{Π_n}) implies that it is constant. Then (20) shows that \bar{h} is constant which concludes the proof. \square

4.5 Central Limit Theorem

Let us now establish a Central Limit Theorem (CLT) associated to the Importance Markov chain for a particular function h , based on a similar hypothesis on the instrumental kernel Q with the function ϱh . More precisely, the Central Limit Theorem for Q stems from the existence of a solution to the Poisson equation for this kernel. This is a quite common sufficient condition for a CLT, and although other conditions exist (references are given for example in [20, Chap 21]), we choose this one for its simplicity in the proofs. To be more specific, we are interested in measurable functions $h : \mathsf{X} \rightarrow \mathbb{R}$ such that if we define

$$h_0 := h - \pi h, \quad (21)$$

the following condition holds:

$(\mathbf{H}_{\text{Poiss}})$ The Poisson equation associated to ϱh_0 for the kernel Q on X admits a $\tilde{\pi}$ -square integrable solution H , i.e. for all $x \in \mathsf{X}$,

$$H(x) - QH(x) = \varrho h_0(x) \text{ and } \tilde{\pi}H^2 < \infty.$$

In addition, $\int_{\mathsf{X} \times \mathbb{N}_0} n^2 h(x)^2 \tilde{\pi}(dx) \tilde{R}(x, dn) < \infty$.

Remark 4. Note that $\tilde{\pi}(\varrho h_0) = 0$ since $\varrho = \frac{d\tilde{\pi}}{d\pi}$, hence this term does not appear in the Poisson equation.

Under $(\mathbf{H}_{\text{Poiss}})$, [20, Theorem 21.2.5] ensures that the Markov chain (\tilde{X}_i) generated by the kernel Q satisfies a Central Limit Theorem for the function ϱh_0 :

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \varrho(\tilde{X}_i) h_0(\tilde{X}_i) \stackrel{\mathbb{P}_{\tilde{\pi}}^Q\text{-law}}{\rightsquigarrow} \mathcal{N}(0, \tilde{\sigma}^2(\varrho h_0)), \quad (22)$$

where $\tilde{\sigma}^2(\varrho h_0) = 2\tilde{\pi}(\varrho h_0 H) - \tilde{\pi}((\varrho h_0)^2)$. Lemma 11 of A.3 combined with (\mathbf{H}_{Π_n}) then extends the weak convergence under $\mathbb{P}_{\tilde{\pi}}^Q$ in the equation above to a weak convergence under \mathbb{P}_{ξ}^Q for any $\xi \in \mathbf{M}_1(\mathsf{X})$. We can now state the CLT for the Importance Markov chain in a formal manner.

Theorem 6. Assume $(\mathbf{H1})$, $(\mathbf{H2})$, (\mathbf{H}_{Π_n}) and let $h : \mathsf{X} \rightarrow \mathbb{R}$ be a measurable function that satisfies $(\mathbf{H}_{\text{Poiss}})$. Then there exists a constant $\sigma^2(h) > 0$ such that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (h(X_i) - \pi h) \stackrel{\mathbb{P}_x^P\text{-law}}{\rightsquigarrow} \mathcal{N}(0, \sigma^2(h)),$$

where the distribution χ is defined by $\chi(f) = \int \xi(dx)S(x, dx')R(x', dn')f(x', n')$. Moreover, we have the following expression of $\sigma^2(h)$:

$$\sigma^2(h) = \kappa \tilde{\sigma}^2(\varrho h_0) + \kappa^{-1} \hat{\sigma}^2(h_0, \kappa), \quad (23)$$

where

- $\tilde{\sigma}^2(\varrho h_0)$ is the variance obtained in (22),
- $\hat{\sigma}^2(h_0, \kappa) := \int_{\mathcal{X}} h_0^2(x) \text{Var}_x^{\tilde{R}}[N] \tilde{\pi}(dx)$,
- $\text{Var}_x^{\tilde{R}}[N] := \int_{\mathbb{N}_0} \tilde{R}(x, dn) n^2 - \left(\int_{\mathbb{N}_0} \tilde{R}(x, dn) n \right)^2$.

Proof. See A.3. □

Remark 5. Note that the variance $\sigma^2(h)$ can be decomposed into two terms: (1) $\tilde{\sigma}^2(\varrho h)$ is the variance coming from the instrumental chain, while (2) $\hat{\sigma}^2(h, \kappa)$ is the variance brought by the random number of repetitions of the instrumental chain.

4.6 Minimizing the asymptotic variance

4.6.1 Optimal choice of the kernel \tilde{R}

Following Remark 5, one can notice that the expression $\hat{\sigma}^2(h, \kappa) = \int_{\mathcal{X}} h^2(x) \text{Var}_x^{\tilde{R}}[N] \tilde{\pi}(dx)$ directly depends on the variance of N under $\tilde{R}(x, \cdot)$. Therefore, minimizing the variance associated to $\tilde{R}(x, \cdot)$, for $x \in \mathcal{X}$, leads to minimization of the asymptotic variance of the chain as defined in Theorem 6. To help tuning \tilde{R} , we state the following lemma:

Lemma 1. Let N be an integer-value random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[N] = \rho$ for a fixed $\rho \in \mathbb{R}^+$. Then,

$$\text{Var}(N) \geq \langle \rho \rangle (1 - \langle \rho \rangle). \quad (24)$$

This bound is reached for $N = \lfloor \rho \rfloor + S$, where $S \sim \text{Ber}(\langle \rho \rangle)$

Proof. Using $0 = \mathbb{E}[N] - \rho = \mathbb{E}[(N - \rho)^+] - \mathbb{E}[(N - \rho)^-]$ and $(N - \rho)^2 \geq (1 - \langle \rho \rangle)(N - \rho)^+ + \langle \rho \rangle (N - \rho)^-$, we get

$$\mathbb{E}[(N - \rho)^2] \geq \mathbb{E}[(N - \rho)^+] = \mathbb{E}[(N - \rho)^-].$$

- If $\mathbb{P}(N > \rho) \geq \langle \rho \rangle$, then

$$\mathbb{E}[(N - \rho)^2] \geq \mathbb{E}[(N - \rho)^+] \geq (1 - \langle \rho \rangle) P(N > \rho) \geq \langle \rho \rangle (1 - \langle \rho \rangle).$$

- If $\mathbb{P}(N > \rho) < \langle \rho \rangle \Leftrightarrow \mathbb{P}(N \leq \rho) > 1 - \langle \rho \rangle$, then

$$\mathbb{E}[(N - \rho)^2] \geq \mathbb{E}[(N - \rho)^-] \geq \langle \rho \rangle P(N \leq \rho) \geq \langle \rho \rangle (1 - \langle \rho \rangle).$$

□

In the case where ϱ_κ can be computed, we can use Lemma 1 to define \tilde{R} as:

$$\tilde{R}_{\text{opt}} = (1 - \langle \varrho_\kappa(x) \rangle) \delta_{\lfloor \varrho_\kappa(x) \rfloor} + \langle \varrho_\kappa(x) \rangle \delta_{\lfloor \varrho_\kappa(x) \rfloor + 1}. \quad (25)$$

This \tilde{R}_{opt} implies the following expression for R_{opt} :

$$R_{\text{opt}} = (1 - \langle \varrho_\kappa(x) \rangle) \delta_{(\lfloor \varrho_\kappa(x) \rfloor - 1)^+} + \langle \varrho_\kappa(x) \rangle \delta_{\lfloor \varrho_\kappa(x) \rfloor}.$$

4.6.2 Optimal upper bound

With the optimal choice of \tilde{R} given in (25), we have $\text{Var}_x^{\tilde{R}_{\text{opt}}} [N] = \langle \varrho_\kappa(x) \rangle (1 - \langle \varrho_\kappa(x) \rangle) \leq 1/4$. Hence,

$$\begin{aligned} \sigma^2(h) &= \kappa \tilde{\sigma}^2(\varrho h_0) + \kappa^{-1} \hat{\sigma}^2(h_0, \kappa) \\ &\leq \kappa \tilde{\sigma}^2(\varrho h_0) + \kappa^{-1} \tilde{\pi}(h_0^2)/4. \end{aligned}$$

Therefore, for a given function h , optimizing the rhs of the above inequality yields:

$$\kappa = \frac{1}{2} \sqrt{\frac{\tilde{\pi}(h_0^2)}{\tilde{\sigma}^2(\varrho h_0)}} \quad (26)$$

from which we deduce the upper bound

$$\sigma^2(h) \leq \sqrt{\tilde{\pi}(h_0^2) \tilde{\sigma}^2(\varrho h_0)}.$$

Note that the choice of κ in (26) depends on the function h which is usually not given beforehand in practice. We propose another way of choosing κ in Section 6.1.2, more adapted to practical concerns.

4.7 Geometric ergodicity

For any set $A \in \mathcal{X}$, we use the notation $\bar{A} = \mathbf{X} \setminus A$ to denote its complement. Define the set $C_\eta := \{x \in \mathbf{X} : \rho_{\tilde{R}}(x) \geq \eta\}$ for $\eta \in [0, 1]$ and note that

$$x \in C_\eta \iff 1 - \rho_{\tilde{R}}(x) \leq 1 - \eta. \quad (27)$$

Finally we denote by $\sigma_C = \inf \{k \geq 1 : X_k \in C\}$ the first return time of the set C .

Lemma 2. *Assume that for some $\eta \in (0, 1)$, C_η is a $(1, \varepsilon\nu)$ -small set for the kernel Q . Then, there exists a probability measure $\tilde{\nu}$ on $\mathbf{X} \times \mathbb{N}_0$ satisfying*

- (i) $C_\eta \times \{0\}$ is a $(1, \varepsilon\tilde{\nu})$ -small set for the kernel P .
- (ii) if $\nu(C_\eta \cap \{\tilde{R}(\cdot, 1) > 0\}) > 0$, then

$$\tilde{\nu}(C_\eta \times \{0\}) > 0.$$

Proof. See A.4.1. □

We now introduce the following assumption:

(H3) There exists $\beta_0 > 1$ such that

$$\sup_{x \in \mathbf{X}} \int_{\mathbb{N}_0} \beta_0^n \tilde{R}(x, dn) < \infty.$$

(H3) is a relatively weak condition, and it is in fact necessary for geometric ergodicity of a Metropolis-Hastings algorithm, seen as an instance of the IMC algorithm (see Remark 3). Indeed, in this case, $\tilde{R}(x, \cdot)$ is a geometric distribution with success parameter $p(x)$ and therefore (H3) is equivalent to p being lower bounded away from zero. [22, Proposition 5.1] proves that this last condition is a necessary condition for geometric ergodicity of a Metropolis-Hastings algorithm.

Also, (H3) holds whenever the support of $\tilde{R}(x, \cdot)$ is uniformly bounded on \mathbf{X} and in particular when $\frac{d\pi}{d\tilde{\pi}}$ is upper-bounded, and $\tilde{R}(x, \cdot)$ is the distribution of $\lfloor \varrho_\kappa(x) \rfloor + U$ with $U \sim \text{Ber}(\langle \varrho_\kappa(x) \rangle)$.

Remark 6. *Actually, although condition (ii) of Lemma 2 is not verified for any kernel \tilde{R} , it is always possible to transform it slightly into a new kernel \tilde{R}' satisfying this condition while keeping its other properties untouched, namely assumptions (H2) and (H3). Indeed, define $C_\eta^- = \{x \in C_\eta : \tilde{R}(x, 0) > 0\}$ and observe that $\nu(C_\eta \cap \{\tilde{R}(\cdot, 1) > 0\}) > \nu(C_\eta^- \cap \{\tilde{R}(\cdot, 1) > 0\})$ since $C_\eta^- \subset C_\eta$. We will now construct a kernel \tilde{R}' such that for all $x \in C_\eta^-$, $\tilde{R}'(x, 1) > 0$, which satisfies the desired condition as $\nu(C_\eta^- \cap \{\tilde{R}'(\cdot, 1) > 0\}) = \nu(C_\eta^-) > 0$. Let $x \in C_\eta^-$, implying that $\tilde{R}(x, 0) > 0$. Due to the unbiasedness assumption, there exists $k \in \mathbb{N}_0, k > 1$ such that $\tilde{R}(x, k) > 0$. Now define \tilde{R}' such that:*

- $\tilde{R}'(x, n) = \tilde{R}(x, n)$ if $n \notin \{0, 1, k\}$,
- $\tilde{R}'(x, 1) = \epsilon \tilde{R}(x, 0)$,
- $\tilde{R}'(x, 0) = \tilde{R}(x, 0) - \frac{(k-1)\epsilon}{k} \tilde{R}(x, 0)$,
- $\tilde{R}'(x, k) = \tilde{R}(x, k) - \frac{\epsilon}{k} \tilde{R}(x, 0)$.

Note that ϵ can be chosen small enough to guarantee $\tilde{R}'(x, k) > 0$. One can easily check that

- $\sum_{n \geq 0} \tilde{R}'(x, n) = \sum_{n \geq 0} \tilde{R}(x, n) = 1$,
- $\sum_{n \geq 0} \tilde{R}'(x, n)n = \sum_{n \geq 0} \tilde{R}(x, n)n = \varrho_\kappa(x)$.

Hence (H2) and (H3) hold for \tilde{R}' .

Recall that $A \in \mathcal{X}$ is accessible for the kernel Q if and only if for all $x \in X$, there exists $n \in \mathbb{N}_0$ such that $Q^n(x, A) > 0$ (see for example Lemma 3.5.2 of [20]). We then have the following lemma:

Lemma 3. Assume (H3). Let $A \in \mathcal{X}$ be an accessible set for Q such that $\inf_{x \in A} \rho_{\tilde{R}}(x) > 0$. Then A is accessible for S and $A \times \{0\}$ is accessible for P .

Proof. See A.4.1. □

Let us now assume a drift condition on Q in order to deduce an upper bound for $\sup_{x \in C_\eta} \mathbb{E}_x^S [\beta^{\sigma c_\eta}]$.

(H_{drift}) There exist constants $(\eta_0, \lambda) \in (0, 1)^2$ and a measurable function $V : X \rightarrow [1, \infty)$ such that

- (a) for any $x \notin C_{\eta_0}$, we have $QV(x) \leq \lambda V(x)$,
- (b) $b_\infty = \sup_{x \in C_{\eta_0}} \frac{QV}{V}(x) < \infty$,
- (c) $\sup_{x \in C_{\eta_0}} V(x) < \infty$.

Remark 7. Under (H_{drift}), we have $QV(x) \leq (b_\infty \vee \lambda)V(x)$ for any $x \in X$. A straightforward induction yields for any $n \in \mathbb{N}_0$, $1 \leq Q^n V(x) \leq (b_\infty \vee \lambda)^n V(x)$. This implies that $b_\infty \vee \lambda \geq V(x)^{-1/n}$. Since n is arbitrary, $b_\infty \vee \lambda \geq 1$. Finally, $b_\infty \geq 1 > \lambda$ and for all $x \in X$,

$$QV(x) \leq b_\infty V(x). \quad (28)$$

Lemma 4. Assume (H3) and (H_{drift}). Then there exist constants $(\eta, \beta_r, \beta_\star) \in (0, \eta_0) \times (1, \infty) \times (0, \infty)$ such that for all $(x, n) \in X \times \mathbb{N}_0$:

$$\mathbb{E}_{(x, n)}^P \left[\beta_r^{\sigma c_\eta \times \{0\}} \right] \leq \beta_\star \beta_r^n V(x). \quad (29)$$

Proof. See A.4.2. □

Before stating the main theorem, let us introduce a smallness assumption:

(H_{sm1}) For any $\eta \in (0, 1)$ there exist a probability measure $\nu \in M_1(X)$ and a constant $\epsilon > 0$ such that, C_η is a $(1, \epsilon\nu)$ -small set and $\nu(C_\eta \cap \{\tilde{R}(\cdot, 1) > 0\}) > 0$.

Theorem 7. Assume (H1), (H2), (H3), (H_{sm1}) and (H_{drift}). Then P has a unique invariant probability measure $\bar{\pi}$ and there exist constants $\delta, \beta_r > 1, \zeta < \infty$, such that for all $\xi \in M_1(X \times \mathbb{N}_0)$,

$$\sum_{k=1}^{\infty} \delta^k d_{TV}(\xi P^k, \bar{\pi}) \leq \zeta \int_{X \times \mathbb{N}_0} \beta_r^n V(x) \xi(dx dn). \quad (30)$$

Proof. According to Theorem 11.4.2 of [20], there exists $\zeta_0 < \infty$ such that:

$$\sum_{k=1}^{\infty} \delta^k d_{TV}(\xi P^k, \bar{\pi}) \leq \zeta_0 \mathbb{E}_\xi^P [\beta_r^{\sigma c_\eta \times \{0\}}], \quad (31)$$

provided that:

(i) $C_\eta \times \{0\}$ is an accessible $(1, \varepsilon\tilde{\nu})$ -small set for P satisfying

$$\tilde{\nu}(C_\eta \times \{0\}) > 0,$$

(ii) $\sup_{x \in C_\eta} \mathbb{E}_{(x,0)}^P[\beta_r^{\sigma_{C_\eta \times \{0\}}}] < \infty$ for some $\beta_r > 1$.

Let us start by proving (i). By Lemma 2, there exists $\tilde{\nu} \in M_1(X)$ such that $C_\eta \times \{0\}$ is a $(1, \varepsilon\tilde{\nu})$ -small set for P satisfying $\tilde{\nu}(C_\eta \times \{0\}) > 0$. Moreover it is accessible for P , by Lemma 3 since C_η is accessible for Q and $\inf_{x \in C_\eta} \rho_{\tilde{R}}(x) \geq \eta > 0$. Hence (i).

It remains to show (ii). We can apply Lemma 4 to get, for $\beta > 1$:

$$\sup_{x \in C_\eta} \mathbb{E}_{(x,0)}^P[\beta_r^{\sigma_{C_\eta \times \{0\}}}] \leq \beta_\star \sup_{x \in C_\eta} V(x) < \infty,$$

which shows (31). Then (30) is obtained from (31) by noting that

$$\mathbb{E}_\xi^P[\beta_r^{\sigma_{C_\eta \times \{0\}}}] = \int_{X \times \mathbb{N}_0} \mathbb{E}_{(x,n)}^P[\beta_r^{\sigma_{C_\eta \times \{0\}}}] \xi(\mathrm{d}x \mathrm{d}n)$$

and applying Lemma 4. □

5 Pseudo-marginal IMC

We develop in this section two different frameworks for pseudo-marginal Importance Markov chain [23]. The first, simplest one is valid if we want to replace $\pi(x)$ by an unbiased estimate $\hat{\pi}(x)$. This can be written as a specific kernel \tilde{R} using the same space as classic IMC.

The second framework tackles the issue of having the instrumental chain (\tilde{X}_i) being itself a pseudo-marginal chain, i.e. in this case, both π and $\hat{\pi}$ are computed through two unbiased estimates.

5.1 Pseudo-marginal within IMC

5.1.1 Adaptation of the kernel \tilde{R} in the pseudo-marginal setting

The first pseudo-marginal approach of the Importance Markov chain can be directly implemented and fits within the framework we develop in this article. Indeed, knowledge of the density of π is never assumed, only the unbiasedness assumption (H2) is needed (and the geometric control of hypothesis (H3) for geometric ergodicity).

Assume that for $x \in X$, the density $\pi(x)$ (with respect to some measure μ) is not directly computable but a nonnegative estimate $\hat{\pi}(x)$ is available, drawn from a kernel $T_\pi(x, \cdot)$ such that $\int_{\mathbb{R}^+} T_\pi(x, \mathrm{d}w)w = \pi(x)$. Then, one can replace $\varrho_\kappa(x)$ in (25) by $\hat{\varrho}_\kappa(x) = \kappa \frac{\hat{\pi}(x)}{\tilde{\pi}(x)}$ to get a plug-in kernel \tilde{R}_{pm} that satisfies (H2).

This can be formalized as follows. First, define an extended kernel \tilde{R}_ψ on $X \times \mathbb{R}^+ \times \mathcal{P}(\mathbb{N}_0)$ by

$$\tilde{R}_\psi(x, w, \mathrm{d}n) = (1 - \langle \kappa w / \tilde{\pi}(x) \rangle) \delta_{\lfloor \kappa w / \tilde{\pi}(x) \rfloor}(\mathrm{d}n) + \langle \kappa w / \tilde{\pi}(x) \rangle \delta_{\lfloor \kappa w / \tilde{\pi}(x) \rfloor + 1}(\mathrm{d}n).$$

Therefore, $\tilde{R}_\psi(x, \hat{\pi}(x), \cdot)$ corresponds to the plug-in random kernel of \tilde{R}_{opt} of (25) using the estimate $\hat{\pi}(x)$. By construction, $\int_{\mathbb{N}_0} n \tilde{R}_\psi(x, w, \mathrm{d}n) = \kappa w / \tilde{\pi}(x)$. We can now define the integrated kernel \tilde{R}_{pm} by

$$\tilde{R}_{\text{pm}}(x, \mathrm{d}n) = \int_{\mathbb{R}^+} \tilde{R}_\psi(x, w, \mathrm{d}n) T_\pi(x, \mathrm{d}w),$$

and R_{pm} using (12).

Lemma 5. \tilde{R}_{pm} satisfies (H2).

Proof. Let $x \in \mathsf{X}$,

$$\begin{aligned} \int_{\mathbb{N}_0} n \tilde{R}_{\text{pm}}(x, dn) &= \int_{\mathbb{N}_0} \int_{\mathbb{R}^+} n \tilde{R}_\psi(x, w, dx) T_\pi(x, dx) = \int_{\mathbb{R}^+} \left(\int_{\mathbb{N}_0} n \tilde{R}_\psi(x, w, dn) \right) T_\pi(x, dw) \\ &= \int_{\mathbb{R}^+} \frac{\kappa w}{\tilde{\pi}(x)} T_\pi(x, dw) = \varrho_\kappa(x). \end{aligned}$$

□

As \tilde{R}_{pm} satisfies (H2), the whole methodology developed in the paper applies to the pseudo marginal case. In particular, the geometric ergodicity result (see next section) still holds if the estimator $\hat{\pi}$ is bounded and under (H_{rej}), as (H3) will be satisfied.

5.1.2 Variance of \tilde{R}_{pm}

The variance of \tilde{R}_{pm} can be expressed as a function of the variance of \tilde{R} . We have the following setup: $W \sim T_\pi(X, \cdot)$ and $N \sim \tilde{R}_\psi(X, W, \cdot)$. Then

$$\begin{aligned} \text{Var}(N|X) &= \text{Var}(\mathbb{E}[N|X, W]|X) + \mathbb{E}[\text{Var}(N|X, W)|X] \\ &= \text{Var}\left(\frac{\kappa W}{\tilde{\pi}(X)} \middle| X\right) + \mathbb{E}[\text{Var}(N|X, W)|X] \\ &= \frac{\kappa^2}{\tilde{\pi}^2(X)} \text{Var}(W|X) + \mathbb{E}\left[\left\langle \frac{\kappa W}{\tilde{\pi}(X)} \right\rangle \left(1 - \left\langle \frac{\kappa W}{\tilde{\pi}(X)} \right\rangle\right) \middle| X\right]. \end{aligned}$$

The variance can be decomposed into two terms: while the second one is similar to the variance obtained from \tilde{R} and can also be upper-bounded by 1/4, the first one is the direct contribution of the variance of the kernel T that generates the estimate. In particular, if $T_\pi(X, \cdot) = \delta_{\pi(X)}$, we recover the same expression as in the previous case.

5.2 Fully pseudo-marginal IMC

In this section, we will write that $\pi(dx) = \pi(x)\mu(dx)$ and $\tilde{\pi}(dx) = \tilde{\pi}(x)\mu(dx)$ for a common measure μ .

We suppose here that (\tilde{X}_k) is itself a pseudo-marginal chain where the estimates of $\tilde{\pi}$ are drawn from a kernel $T_{\tilde{\pi}}(x, \cdot)$ such that for all $x \in \mathsf{X}$, $\int_{\mathbb{R}^+} T_{\tilde{\pi}}(x, du)u = \tilde{\pi}(x)$. This constructs a two-component Markov chain $(\tilde{X}_k, \tilde{U}_k)$ on $\mathsf{X} \times \mathbb{R}^+$ that targets $\mu(dx) T_{\tilde{\pi}}(x, du)u$ [24].

In order to extend the Importance Markov chain to this case, we need once again to increase the dimension of the chain. The space $\mathsf{X} \times \mathbb{R}^+$ is replaced by $\mathsf{X} \times \mathbb{R}^+ \times \mathbb{R}^+$. In that case, the second (resp. third) marginal corresponds to the nonnegative estimates of respectively $\tilde{\pi}$ (resp. π). The third component \tilde{V}_k is drawn from a kernel $T_\pi(\tilde{X}_k, \cdot)$ such that for all $x \in \mathsf{X}$, $\int_{\mathbb{R}^+} T_\pi(x, dv)v = \pi(x)$.

This constructs a Markov chain $(\tilde{X}_k, \tilde{U}_k, \tilde{V}_k)$ that targets the probability measure $\tilde{\Pi}$ defined by

$$\tilde{\Pi}(dx du dv) = \mu(dx) T_{\tilde{\pi}}(x, du)u T_\pi(x, dv).$$

The distribution $\tilde{\Pi}$ is the *instrumental* distribution for the extended space. Its first marginal is $\tilde{\pi}$ as $\tilde{\Pi}(A \times \mathbb{R}^+ \times \mathbb{R}^+) = \tilde{\pi}(A)$ for any $A \in \mathcal{X}$. We can define our *target* distribution Π on the same extended space by

$$\Pi(dx du dv) = \mu(dx) \tilde{T}_{\tilde{\pi}}(x, du) T_\pi(x, dv)v.$$

We can perform an Importance Markov chain using the instrumental chain $(\tilde{X}_k, \tilde{U}_k, \tilde{V}_k)$ targeting the instrumental density $\tilde{\Pi}$ and the target distribution Π . In this setting, ϱ_κ becomes:

$$\varrho_\kappa(x, u, v) = \kappa \frac{d\Pi}{d\tilde{\Pi}}(x, u, v) = \kappa \frac{v}{u}.$$

It remains to draw some random integer \tilde{N}_k with conditional expectation:

$$\varrho_\kappa(\tilde{X}_k, \tilde{U}_k, \tilde{V}_k) = \kappa \frac{\tilde{V}_k}{\tilde{U}_k},$$

using for instance \tilde{R}_{opt} .

6 Numerical experiments

6.1 Toy example: mixture of Gaussians

6.1.1 Setting

For starters, we apply Algorithm 2 on an multidimensional Gaussian mixture. The target distribution π writes

$$\pi(x) \propto \sum_{i=1}^n \phi_d(x; \mu_i, I_d),$$

where $\phi_d(x; \mu, \Sigma)$ is the density of a Gaussian distribution in dimension d with mean μ and covariance matrix Σ . The means of these distributions are random, i.i.d., $\mu_i \sim \mathcal{N}(0, 10^2 I_d)$. The instrumental distribution $\tilde{\pi}$ is chosen to be $\tilde{\pi}(x) \propto \pi(x)^\beta$ for a fixed $\beta \in (0, 1)$. The kernel Q targeting $\tilde{\pi}$ is a No-U-turn Sampler (NUTS) [25].

The parameter β flattens the instrumental distribution, thus easy out the way for the instrumental chain to move from one mode to another, when compared with the original targetted π . Conversely, extremely small values of β lead to a very flat instrumental distribution from which it is hard to reconstruct the original target π . Therefore, we test different values of β towards finding an optimal tradeoff. The kernel \tilde{R} used to draw the number of replicas is set to be the same as in (25), i.e. a shifted Bernoulli.

Note that with such a choice for $\tilde{\pi}$, the ratio $\frac{\pi}{\tilde{\pi}}$ becomes proportional to $\pi^{1-\beta}$ so that the computation of ϱ_κ is just a pointwise evaluation of π plus basic float operations. As the choice of R also requires basic float operations (floor, fractional part and multiplication) and a Bernoulli draw, the complexity of Algorithm 2 is close to the complexity of directly running Q targeting π .

To assess the influence of β we estimate the mean squared error in approximating the expectation of π by running 200 chains for each $\beta \in \{0.004, 0.01, 0.04, 0.1, 1\}$. The results are presented in Table 1. As expected, for an untempered instrumental distribution, i.e. $\beta = 1$, the MSE is high due to a low exploration of the space, and is minimal for $\beta = 0.04$.

| β | 0.004 | 0.010 | 0.040 | 0.100 | 1.000 |
|---------|--------|-------|--------------|--------|--------|
| MSE | 16.150 | 6.123 | 0.544 | 17.863 | 33.982 |

Table 1: MSE for different values of β for the Gaussian mixture

6.1.2 Choice of κ and analysis of the Effective Sample Size

The parameter κ that appears in (7) is somewhat more arbitrary as the other parameters as its value is directly impacted by the normalizing constants of π and $\tilde{\pi}$. Indeed, assume $\pi_U = \pi Z$ (resp. $\tilde{\pi}_U$) is an unnormalized density and write Z (resp. \tilde{Z}) the unknown normalization constant. We can then compute the ratio $\varrho_{\kappa,U} := \kappa \frac{\pi_U}{\tilde{\pi}_U} = \kappa \frac{Z}{\tilde{Z}} \frac{\pi}{\tilde{\pi}} = \varrho_{\kappa \frac{Z}{\tilde{Z}}}$. Thus, ignoring the normalizing constants leads to a multiplier term κ compared to the case where both densities are normalized.

However it proves possible to overcome this issue by noticing that $\mathbb{E}[\sum_{i=1}^n \tilde{N}_i] = \kappa n$, i.e. that, for an instrumental chain of length n , the (expected) length final chain (X_i) is proportional to κ . So, one way to deal with this issue is to tune κ such that the length of the final chain is approximately αn , where α is fixed. This is easily solved by taking $\kappa = \frac{\alpha n}{\sum_{i=1}^n \varrho_U(\tilde{X}_i)}$, for $\varrho_U = \frac{\pi_U}{\tilde{\pi}_U}$.

Remark 8. *The diagnosis of the choice of κ can be easily done for a fixed instrumental chain via the computation of the number of replicas for different values of κ in parallel is cheap once the vector of the values taken by ϱ is stored.*

The problem of tuning α remains. We define a metric that may help to this effect, namely, the effective sample size (denoted ESS_κ , as it depends on κ). The ESS for an importance Markov chain is similar to the ESS defined for importance sampling, at the difference that here the weights are discretized. Formally,

$$\text{ESS}_\kappa := \frac{(\sum_{i=1}^n \tilde{N}_i)^2}{\sum_{i=1}^n \tilde{N}_i^2}$$

and we also define the usual importance sampling ESS:

$$\text{ESS}_{\text{IS}} := \frac{(\sum_{i=1}^n \varrho(\tilde{X}_i))^2}{\sum_{i=1}^n \varrho(\tilde{X}_i)^2}.$$

Remark 9. Note that the new definition of ESS_κ only considers the replication step of the IMC algorithm, and does not take into account the convergence of the instrumental chain to the instrumental target. See [9] for a recent work on the effective sample size for IS with dependent proposals. The authors add a term to the ESS to take into account the correlation of the sample.

Writing $w_i = \frac{\varrho(\tilde{X}_i)}{\sum_{i=1}^n \varrho(\tilde{X}_i)}$ for the self-normalized ϱ , we have that $\mathbb{E}[\tilde{N}_i | \tilde{X}_i] = \kappa w_i$, so conditionally on $\tilde{X}_{1:n}$,

$$\text{ESS}_\kappa \xrightarrow[\kappa \rightarrow \infty]{} \text{ESS}_{\text{IS}}, \quad \mathbb{P} - \text{a.s.} \quad (32)$$

As expected by the definition of kernel R , as κ increases, the stochastic part of the number of replicas vanishes and estimates built with the chain will behave as with importance sampling, in the sense that, for $M_n := \sum_{i=1}^n \tilde{N}_i$:

$$M_n^{-1} \sum_{i=1}^{M_n} h(X_i) = \sum_{i=1}^n \frac{\tilde{N}_i}{M_n} h(\tilde{X}_i) \xrightarrow[\kappa \rightarrow \infty]{} \sum_{i=1}^n w_i h(\tilde{X}_i).$$

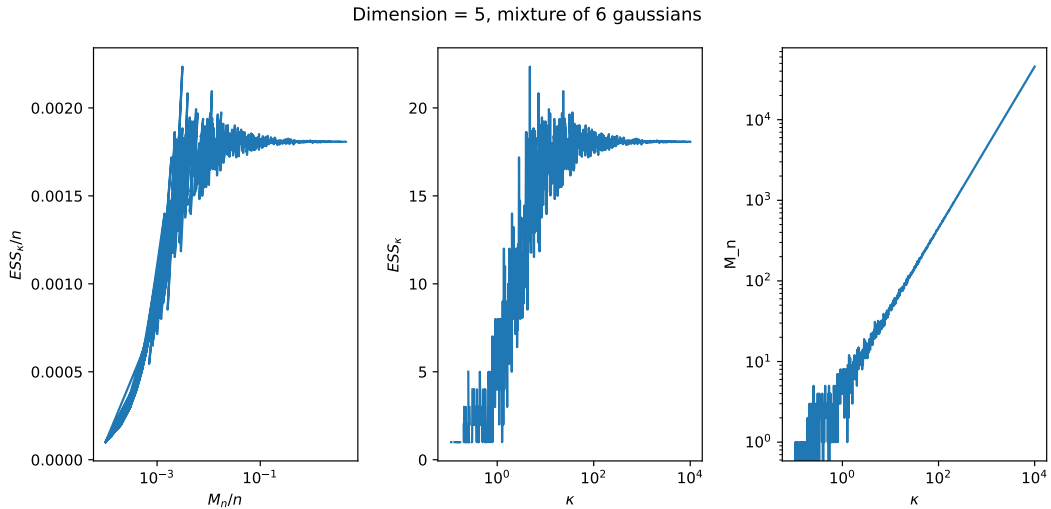


Figure 1: Analysis of the effect of κ on the ESS and the length of the chain.

Following Remark 8, we plot the effect of κ on ESS_κ in Figure 1. For a fixed instrumental chain of length n and a fixed $\beta = 0.04$, we computed the ESS_κ for 10^3 values of κ varying on log scale from 10^{-1} to 10^4 . The plot on the right confirms a linear dependence between the length of the final chain and κ . Both other plots (ESS_κ/n as a function of M_n/n and ESS_κ as a function of κ) have the same shape. Overall, the ESS is increasing with κ and reaches a stationary regime for κ large enough: it converges to ESS_{IS} . Therefore, taking κ too large will not increase the quality of the estimate. In that specific case, this diagnosis leads us to choose $\alpha \simeq 1$.

6.2 Independent IMC with normalizing flows

6.2.1 Settings

In this section, we compare the Metropolis - Hastings algorithm with independent proposals with the Importance Markov chain algorithm in the specific case where $Q(x, \cdot) = \tilde{\pi}(\cdot)$ for all $x \in \mathbf{X}$.

The target π in dimension d is defined, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, by

$$\pi(x) \propto \exp \left\{ -\frac{1}{2} \left(\frac{\|x\| - 2}{0.1} \right)^2 + \sum_{i=1}^d \log \left(e^{-\frac{1}{2} \left(\frac{x_i+3}{0.6} \right)^2} + e^{-\frac{1}{2} \left(\frac{x_i-3}{0.6} \right)^2} \right) \right\}.$$

This distribution suffers from multimodality as it has 2 modes per marginal, for a total of 2^d modes.

The instrumental distribution $\tilde{\pi}$ is obtained by training a normalizing flow targeting π . We recall that a normalizing flow is an invertible map T from \mathbb{R}^d to \mathbb{R}^d . Taking a base distribution μ on \mathbb{R}^d (chosen in that case to be the standard Gaussian), the map T should be chosen such that the pushforward measure of μ through T denoted $T_{\#}\mu := \mu(T^{-1}(\cdot))$ is close to π . This can be done by optimizing T in a family of maps, for instance rational quadratic splines (RQSplines) using neural networks [26]. The training is done by minimizing the forward Kullback-Leibler divergence between $T_{\#}\mu$ and π . See [27] for further details on the training. To get a sample from the flow, one can generate $x \sim \mu$ and derive $T(x)$. The density ρ of $T_{\#}\mu$ is given by, for $x \in \mathbb{R}^d$:

$$\rho(x) = \mu(T^{-1}(x)) |\det J_{T^{-1}}(x)|.$$

The flows are designed such that T^{-1} and $J_{T^{-1}}(x)$ are easily computable.

We used the Python package FlowMC [28] with a RQSpline model to train the flow. Every training of a flow yields a different $\tilde{\pi} = T_{\#}\mu$ as the training is stochastic. For details of implementation, see B.1.

The Self Regenerative Markov Chain Monte Carlo (with no adaptation) [14, 15] is close to the Independent IMC, for the special case where the distribution of the number of replicas \tilde{N}_i is written as

$$\mathbb{P}(\tilde{N}_i = n | \tilde{X}_i = x) = \mathbb{P}(VS = n),$$

where V is a Bernoulli random variable with parameter $q(x)$ and S is geometric with parameter $\alpha(x)$. In [15], the author suggests $q(x) = \min(1, \varrho_{\kappa}(x))$ and $\alpha(x) = \min(1, 1/(\varrho_{\kappa}(x)))$. This method, called *optimal self-regenerative chain* (OSR), is the one we use as a comparison benchmark, with the same tuning of κ . We compare three methods: the independent Metropolis-Hastings (see [2] for details), the independent importance Markov chain with kernel \tilde{R}_{opt} of (25) and the OSR chain defined above. In this case, the IMC algorithm is close to the rejection sampling chain of [29].

The parameter κ is tuned such that the length of the final chain is equal to the length of the instrumental chain. The computational cost (and running time) of all three algorithms is similar.

6.2.2 Results

Comparison with Metropolis Hastings and OSR For each dimension $d \in \{5, 10, 15, 20, 25\}$, we trained 10 different flows on the same target. For each flow, 30 i.i.d. samples of size $n = 3 \cdot 10^4$ were generated. We display in Figure 2 the boxplot of the effective sample size (ESS) of the first marginal for each algorithm, computed using the bulk method.

While the ESS of the OSR is a bit higher than the one obtained with IMH, both are outperformed by the importance Markov chain. Regardless of the dimension, the ESS of the IMC is approximately twice that of MH. Performances across dimensions are quite similar, even if it is worth noting that the ESS is slightly lower for the dimension 5, probably due to the training of the flow and the fact that the hyperparameters of the flow are not optimized for each dimension.

Comparison with importance sampling We compare in Table 2 the mean squared error (MSE) of the first four odd moments of the first marginal obtained by either the importance

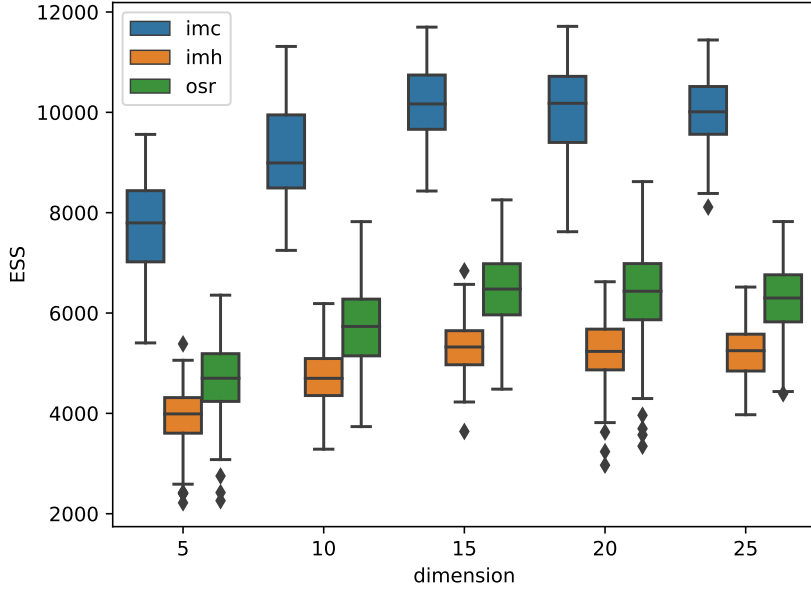


Figure 2: Comparison of the effective sample size between MH, OSR and IMC.

Markov chain or with the importance sampling (IS) estimate defined by

$$\hat{I}_{IS}(h) = \frac{\sum_{i=1}^n \varrho_{\kappa}(\tilde{X}_i) h(\tilde{X}_i)}{\sum_{i=1}^n \varrho_{\kappa}(\tilde{X}_i)}$$

against the independent IMC estimate defined by

$$\hat{I}_{IMC}(h) = \frac{1}{k} \sum_{i=1}^k h(X_i) = \frac{1}{k} \sum_{i=1}^n \tilde{N}_i h(\tilde{X}_i),$$

where we denote $k = \sum_{i=1}^n \tilde{N}_i$ the length of the final chain. The performances of the IMC are very close to the ones of IS, while the gap increases with the dimension. The last column shows the mean number points that are replicated once or more by the IMC, which is the length of the chain in the case of importance sampling. This has a strong implication: by storing the importance chain under the representation $(\tilde{X}_i, \tilde{N}_i)$, only (around) 16400 points (and their associated number of replicas) are needed to be stored for the dimension 25, instead of 30000 for importance sampling. If the dimension and the number of points are large, this can be useful to reduce the memory usage. In some contexts, the slight loss in the MSE can be compensated by the gain in memory usage. Moreover, the IMC outputs an actual sample and not a weighted one, sample that approaches the target distribution at mostly a geometric rate.

| Dimension | Method | Mean | Third | Fifth | Seventh | # positive copies |
|-----------|--------|-----------|-----------|-----------|-----------|-------------------|
| 5 | IMC | 9.955e-05 | 2.057e-04 | 8.372e-04 | 4.638e-03 | 1.388e+04 |
| | IS | 9.822e-05 | 2.024e-04 | 8.220e-04 | 4.552e-03 | 3.000e+04 |
| 10 | IMC | 5.176e-05 | 4.146e-05 | 6.799e-05 | 1.653e-04 | 1.512e+04 |
| | IS | 4.890e-05 | 3.976e-05 | 6.571e-05 | 1.607e-04 | 3.000e+04 |
| 15 | IMC | 3.449e-05 | 1.403e-05 | 1.279e-05 | 1.830e-05 | 1.595e+04 |
| | IS | 3.229e-05 | 1.341e-05 | 1.254e-05 | 1.804e-05 | 3.000e+04 |
| 20 | IMC | 2.560e-05 | 7.580e-06 | 5.402e-06 | 5.940e-06 | 1.609e+04 |
| | IS | 2.486e-05 | 7.218e-06 | 5.068e-06 | 5.425e-06 | 3.000e+04 |
| 25 | IMC | 2.356e-05 | 5.585e-06 | 2.962e-06 | 2.483e-06 | 1.646e+04 |
| | IS | 2.273e-05 | 5.180e-06 | 2.631e-06 | 2.112e-06 | 3.000e+04 |

Table 2: Mean square error for the first four odd moments of the first marginal for both the IMC chain and the importance sampling estimate

7 Conclusion

The Importance Markov Chain is a meta-algorithm, in the sense that the produced Markov chain $\{X_k : k \in \mathbb{N}_0\}$ is built upon another one, namely $\{\tilde{X}_k : k \in \mathbb{N}_0\}$. This allows the practitioner to change the target of the MCMC kernel used for the sampling: instead of targeting the distribution of interest directly, our algorithm targets another distribution that may have better properties, and then transforms - with relatively low cost - the obtained sample into an output sample following the distribution of interest.

A range of applications obviously includes the different instances of Bayesian sensitivity analysis and robustness, such as a switch of prior distributions in Bayesian robustness [30, 31], the inclusion or exclusion of a particular data point for detecting influential observations [32], delayed acceptance MCMC [33], safe Bayes modifications of the likelihood [34, 35]. A practical illustration is found in the statistical analysis of the standard cosmological model [36], where accounting for further satellite observation experiments proves very costly in computing the modified likelihoods.

Acknowledgments

Charly Andral is supported by a grant from Région Ile-de-France. Christian Robert is funded by the European Union under the GA 101071601, through the 2023-2029 ERC Synergy grant OCEAN.

A Postponed proofs

The following appendix contains supplementary information that either does not constitute an essential part of the paper, but is helpful in providing a more comprehensive understanding of the research problem, or is too cumbersome to be included in the body of the paper.

A.1 Uniqueness of the invariant probability measure

(*Proof of Theorem 4*). Let π_0 be an invariant probability measure for P and denote by π_0^k the measure defined by $\pi_0^k(A) = \pi_0(A \times \{k\})$ for any $A \in \mathcal{X}$. To obtain that $\pi_0 = \bar{\pi}$, we will first express π_0 using the measure π_0^0 only. Let $f \in F_{b^+}(X)$. Applying (11) with $h_k(x, n) = f(x)\mathbf{1}_{\{n \neq k\}}$ and integrating with respect to π_0 yields for any $k \geq 0$,

$$\pi_0^k f = \pi_0 h_k = \pi_0 P h_k = \pi_0^{k+1} f + \int_{\mathcal{X}} \pi_0^0 S(dx) R(x, k) f(x). \quad (33)$$

To simplify this equation, we first show that $\pi_0^0 = \pi_0^0 S$. Indeed, using again $\pi_0 = \pi_0 P$ and (11) with the function $h(x, n) = f(x)$,

$$\int_{\mathbf{X} \times \mathbb{N}_0} \pi_0(\mathrm{d}x \mathrm{d}n) f(x) = \mathbb{E}_{\pi_0^0}^P[f(X_0)] = \mathbb{E}_{\pi_0^0}^P[f(X_1)] = \int_{\mathbf{X} \times \mathbb{N}_0} \pi_0(\mathrm{d}x \mathrm{d}n) \mathbf{1}_{\{n \geq 1\}} f(x) + \int_{\mathbf{X} \times \mathbb{N}_0} \pi_0(\mathrm{d}x \mathrm{d}n) \mathbf{1}_{\{n=0\}} S f(x),$$

which can be equivalently written as $\pi_0^0 f = \pi_0^0 S f$. Plugging $\pi_0^0 = \pi_0^0 S$ into (33), we get $\pi_0^k f = \pi_0^{k+1} f + \int_{\mathbf{X}} \pi_0^0(\mathrm{d}x) R(x, k) f(x)$ and by straightforward induction,

$$\pi_0^k f = \pi_0^0 f - \sum_{\ell=0}^{k-1} \int_{\mathbf{X}} \pi_0^0(\mathrm{d}x) R(x, \ell) f(x) = \int_{\mathbf{X}} \pi_0^0(\mathrm{d}x) f(x) \left[1 - \sum_{\ell=0}^{k-1} R(x, \ell) \right] = \int_{\mathbf{X}} \pi_0^0(\mathrm{d}x) f(x) \sum_{\ell=k}^{\infty} R(x, \ell).$$

Hence, for any $h \in \mathbf{F}_{b+}(\mathbf{X} \times \mathbb{N}_0)$,

$$\begin{aligned} \pi_0 h &= \int_{\mathbf{X} \times \mathbb{N}_0} \pi_0(\mathrm{d}x \mathrm{d}n) h(x, n) = \sum_{k=0}^{\infty} \int_{\mathbf{X}} \pi_0^k(\mathrm{d}x) h(x, k) \\ &= \sum_{k=0}^{\infty} \int_{\mathbf{X}} \pi_0^0(\mathrm{d}x) h(x, k) \sum_{\ell=k}^{\infty} R(x, \ell) = \sum_{\ell=0}^{\infty} \int_{\mathbf{X}} \pi_0^0(\mathrm{d}x) R(x, \ell) \sum_{k=0}^{\ell} h(x, k). \end{aligned} \quad (34)$$

Combining with (13), we can conclude the proof of Theorem 4 (ie $\pi_0 = \bar{\pi}$) provided that

$$\pi_0^0(\mathrm{d}x) = \kappa^{-1} \bar{\pi}(\mathrm{d}x) \rho_{\bar{R}}(x). \quad (35)$$

All that follows consists in proving this identity. Denote by π_1 the measure on $(\mathbf{X} \times [0, 1], \mathcal{X} \otimes \mathcal{B}([0, 1]))$ defined by: for any function $h \in \mathbf{F}_{b+}(\mathbf{X} \times [0, 1])$,

$$\pi_1 h = \int_{\mathbf{X} \times [0, 1]} \pi_0^0(\mathrm{d}x) \mathrm{d}u \mathbf{1}_{[0, \rho_{\bar{R}}(x)]}(u) \rho_{\bar{R}}(x)^{-1} \mathbb{E}_{(x, u)}^G \left[\sum_{k=0}^{\sigma_D - 1} h(X_k, U_k) \right], \quad (36)$$

where $D = \{(x, u) \in \mathbf{X} \times [0, 1] : u \leq \rho_{\bar{R}}(x)\}$ and G is defined in (2). We first show that $\pi_1 = \pi_1 G$.

$$\begin{aligned} \pi_1 G h &= \int_{\mathbf{X} \times [0, 1]} \pi_0^0(\mathrm{d}x) \mathrm{d}u \mathbf{1}_{[0, \rho_{\bar{R}}(x)]}(u) \rho_{\bar{R}}(x)^{-1} \mathbb{E}_{(x, u)}^G \left[\sum_{k=0}^{\sigma_D - 1} G h(X_k, U_k) \right] \\ &= \sum_{k=0}^{\infty} \int_{\mathbf{X} \times [0, 1]} \pi_0^0(\mathrm{d}x) \mathrm{d}u \mathbf{1}_{[0, \rho_{\bar{R}}(x)]}(u) \rho_{\bar{R}}(x)^{-1} \mathbb{E}_{(x, u)}^G \left[h(X_{k+1}, U_{k+1}) \mathbf{1}_{\{k+1 \leq \sigma_D\}} \right] \\ &= \int_{\mathbf{X} \times [0, 1]} \pi_0^0(\mathrm{d}x) \mathrm{d}u \mathbf{1}_{[0, \rho_{\bar{R}}(x)]}(u) \rho_{\bar{R}}(x)^{-1} \mathbb{E}_{(x, u)}^G \left[\sum_{\ell=1}^{\sigma_D} h(X_{\ell}, U_{\ell}) \right]. \end{aligned}$$

This implies

$$\begin{aligned} \pi_1 G h &= \pi_1(h) + \int_{\mathbf{X} \times [0, 1]} \pi_0^0(\mathrm{d}x) \mathrm{d}u \mathbf{1}_{[0, \rho_{\bar{R}}(x)]}(u) \rho_{\bar{R}}(x)^{-1} \mathbb{E}_{(x, u)}^G \left[h(X_{\sigma_D}, U_{\sigma_D}) \mathbf{1}_{\{\sigma_D < \infty\}} \right] \\ &\quad - \int_{\mathbf{X} \times [0, 1]} \pi_0^0(\mathrm{d}x) \left(\int_0^{\rho_{\bar{R}}(x)} h(x, u) \mathrm{d}u \right) \rho_{\bar{R}}(x)^{-1}. \end{aligned} \quad (37)$$

Now, write

$$\begin{aligned}
\mathbb{E}_{(x,u)}^G [h(X_{\sigma_D}, U_{\sigma_D}) \mathbf{1}_{\{\sigma_D < \infty\}}] &= \sum_{\ell=1}^{\infty} \mathbb{E}_{(x,u)}^G [h(X_{\ell}, U_{\ell}) \mathbf{1}_{\{U_{\ell} \leq \rho_{\bar{R}}(X_{\ell})\}} \mathbf{1}_{\{\sigma_D \geq \ell\}}] \\
&= \sum_{\ell=1}^{\infty} \mathbb{E}_{(x,u)}^G \left[\left(\int_0^{\rho_{\bar{R}}(X_{\ell})} h(X_{\ell}, u') du' \right) \mathbf{1}_{\{\sigma_D \geq \ell\}} \right] \\
&= \sum_{\ell=1}^{\infty} \mathbb{E}_{(x,u)}^G \left[\left(\int_0^{\rho_{\bar{R}}(X_{\ell})} h(X_{\ell}, u') du' \right) \rho_{\bar{R}}(X_{\ell})^{-1} \mathbf{1}_{\{U_{\ell} \leq \rho_{\bar{R}}(X_{\ell})\}} \mathbf{1}_{\{\sigma_D \geq \ell\}} \right] \\
&= \mathbb{E}_{(x,u)}^G \left[\left(\int_0^{\rho_{\bar{R}}(X_{\sigma_D})} h(X_{\sigma_D}, u') du' \right) \rho_{\bar{R}}(X_{\sigma_D})^{-1} \right] \\
&= \int_{\mathbf{X}} S(x, dx') \left(\int_0^{\rho_{\bar{R}}(x')} h(x', u') du' \right) \rho_{\bar{R}}(x')^{-1}.
\end{aligned}$$

Plugging this expression into (37) yields:

$$\pi_1 Gh = \pi_1(h) + \int_{\mathbf{X}} \pi_0^0 S(dx') \left(\int_0^{\rho_{\bar{R}}(x')} h(x', u') du' \right) \rho_{\bar{R}}(x')^{-1} - \int_{\mathbf{X}} \pi_0^0(dx) \left(\int_0^{\rho_{\bar{R}}(x)} h(x, u) du \right) \rho_{\bar{R}}(x)^{-1} = \pi_1 h,$$

where we have used that $\pi_0^0 S = \pi_0^0$. Hence π_1 is an invariant measure for G . Since any invariant measure for Q is proportional to $\tilde{\pi}$, we deduce from (2) that G admits a unique invariant measure (up to a multiplicative constant) proportional to $\tilde{\pi}(dx) \mathbf{1}_{[0,1]}(u) du$ and hence, $\tilde{\pi}(dx) \mathbf{1}_{[0,1]}(u) du \propto \pi_1$. Now, taking $h(x, n) = \mathbf{1}_D(x, u) f(x)$ for any arbitrary $f \in \mathbf{F}_{b+}(\mathbf{X})$, we get, using (36),

$$\begin{aligned}
\int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\bar{R}}(x) f(x) &= \int_{\mathbf{X} \times [0,1]} \tilde{\pi}(dx) \mathbf{1}_{[0,1]}(u) du h(x, u) \\
&\propto \pi_1 h = \int_{\mathbf{X} \times [0,1]} \pi_0^0(dx) du \mathbf{1}_{[0, \rho_{\bar{R}}(x)]}(u) \rho_{\bar{R}}(x)^{-1} \mathbf{1}_D(x, u) f(x) = \pi_0^0 f.
\end{aligned}$$

Finally, there exists a constant γ such that for any $f \in \mathbf{F}_{b+}(\mathbf{X})$,

$$\pi_0^0 f = \gamma \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\bar{R}}(x) f(x). \tag{38}$$

Applying (34) with $h = \mathbf{1}$ and using successively (12), (H2) and the identity above, we get

$$1 = \sum_{\ell=0}^{\infty} \int_{\mathbf{X}} \pi_0^0(dx) R(x, \ell) (\ell+1) = \sum_{k=1}^{\infty} \int_{\mathbf{X}} \pi_0^0(dx) \frac{\tilde{R}(x, k) k}{\rho_{\bar{R}}(x)} = \int_{\mathbf{X}} \pi_0^0(dx) \frac{\varrho_{\kappa}(x)}{\rho_{\bar{R}}(x)} = \gamma \int_{\mathbf{X}} \tilde{\pi}(dx) \varrho_{\kappa}(x) = \gamma \kappa.$$

Combined with (38) we finally obtain $\pi_0^0 f = \kappa^{-1} \int_{\mathbf{X}} \tilde{\pi}(dx) \rho_{\bar{R}}(x) f(x)$ which proves (35) and concludes the proof. \square

A.2 A martingale weak convergence result with random indexes.

Theorem 8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let (\mathcal{F}_n) be a filtration on Ω such that $\mathcal{F}_n \subset \mathcal{F}$ for any $n \in \mathbb{N}_0$. Let (M_n) be a square-integrable (\mathcal{F}_n) -martingale such that*

$$\frac{M_n}{\sqrt{n}} \overset{\mathbb{P}\text{-law}}{\rightsquigarrow} G \quad \text{and} \quad \frac{\mathbb{E}[M_n^2]}{n} \rightarrow \sigma^2$$

and let (k_n) be a sequence of random integers such that $\frac{k_n}{n} \overset{\mathbb{P}\text{-prob}}{\rightarrow} \lambda \in (0, \infty)$. Then

$$(n\lambda)^{-1/2} M_{k_n} \overset{\mathbb{P}\text{-law}}{\rightsquigarrow} G.$$

Proof. Let $\alpha, \lambda^-, \lambda^+$ be positive constants such that $\lambda^- < \lambda < \lambda^+$. Define

$$\begin{aligned} B_n &= (n\lambda)^{-1/2} M_{k_n}, \\ C_n &= (n\lambda)^{-1/2} M_{\lfloor n\lambda^- \rfloor}, \end{aligned}$$

and note that $C_n \xrightarrow{\mathbb{P}} (\lambda^-/\lambda)^{1/2} G$. Then we have for any $u \in \mathbb{R}$,

$$\begin{aligned} |\mathbb{E}[e^{iuB_n}] - \mathbb{E}[e^{iuG}]| &\leq |\mathbb{E}[e^{iuB_n}] - \mathbb{E}[e^{iuC_n}]| + |\mathbb{E}[e^{iuC_n}] - \mathbb{E}[e^{iuG}]| \\ &\leq 2\mathbb{P}(|B_n - C_n| > \alpha) + |\mathbb{E}[e^{iuB_n} - e^{iuC_n} | \mathbf{1}_{\{|B_n - C_n| \leq \alpha\}}]| + |\mathbb{E}[e^{iuC_n}] - \mathbb{E}[e^{iuG}]| \\ &\leq 2\mathbb{P}(|B_n - C_n| > \alpha) + \sup_{|\beta| \leq \alpha} |e^{iu\beta} - 1| + |\mathbb{E}[e^{iuC_n}] - \mathbb{E}[e^{iuG}]|. \end{aligned} \quad (39)$$

Since $(M_i - M_{\lfloor n\lambda^- \rfloor})_{i \geq \lfloor n\lambda^- \rfloor}$ is a martingale and $x \rightarrow x^2$ is convex, $((M_i - M_{\lfloor n\lambda^- \rfloor})^2)_{i \geq \lfloor n\lambda^- \rfloor}$ is a non-negative submartingale. Then, the first term of the right hand side in (39) may be bounded by applying Doob's maximal inequality to the non-negative submartingale $((M_i - M_{\lfloor n\lambda^- \rfloor})^2)_{i \geq \lfloor n\lambda^- \rfloor}$,

$$\begin{aligned} \mathbb{P}(|B_n - C_n| > \alpha) &\leq \mathbb{P}(k_n \notin [n\lambda^-, n\lambda^+]) + \mathbb{P}(|B_n - C_n| > \alpha, k_n \in [n\lambda^-, n\lambda^+]) \\ &\leq \mathbb{P}(k_n \notin [n\lambda^-, n\lambda^+]) + \mathbb{P}\left(\sup_{i \in [n\lambda^-, n\lambda^+]} |M_i - M_{\lfloor n\lambda^- \rfloor}| > (n\lambda)^{1/2} \alpha\right) \\ &\leq \mathbb{P}(k_n \notin [n\lambda^-, n\lambda^+]) + \frac{\mathbb{E}\left[(M_{\lfloor n\lambda^+ \rfloor} - M_{\lfloor n\lambda^- \rfloor})^2\right]}{n\lambda\alpha^2} \\ &= \mathbb{P}(k_n \notin [n\lambda^-, n\lambda^+]) + \frac{\sum_{k=\lfloor n\lambda^- \rfloor}^{\lfloor n\lambda^+ \rfloor - 1} \mathbb{E}\left[(M_{k+1} - M_k)^2\right]}{n\lambda\alpha^2}. \end{aligned}$$

Note that since $(M_n)_{n \in \mathbb{N}}$ is a square-integrable martingale, we have

$$D_n = \frac{\mathbb{E}[M_n^2]}{n} = \frac{\mathbb{E}[M_0^2] + \sum_{k=0}^{n-1} \mathbb{E}[(M_{k+1} - M_k)^2]}{n},$$

and the previous bound writes:

$$\mathbb{P}(|B_n - C_n| > \alpha) \leq \mathbb{P}(k_n \notin [n\lambda^-, n\lambda^+]) + \frac{\lfloor n\lambda^+ \rfloor D_{\lfloor n\lambda^+ \rfloor} - \lfloor n\lambda^- \rfloor D_{\lfloor n\lambda^- \rfloor}}{n\lambda\alpha^2}.$$

Finally letting n go to infinity and using successively that $\frac{k_n}{n} \xrightarrow{\mathbb{P}\text{-prob}} \lambda$, $D_n \xrightarrow[n \rightarrow \infty]{} \sigma^2$ and $C_n \xrightarrow{\mathbb{P}} (\frac{\lambda^-}{\lambda})^{1/2} G$, we obtain

$$\limsup_{n \rightarrow \infty} |\mathbb{E}[e^{iuB_n}] - \mathbb{E}[e^{iuG}]| \leq 2\sigma^2 \frac{\lambda^+ - \lambda^-}{\lambda\alpha^2} + \sup_{|\beta| \leq \alpha} |e^{iu\beta} - 1| + |\mathbb{E}[e^{iu(\lambda^-/\lambda)^{1/2} G}] - \mathbb{E}[e^{iuG}]|.$$

Letting $\lambda^+ \searrow \lambda$ and $\lambda^- \nearrow \lambda$, we get

$$\limsup_{n \rightarrow \infty} |\mathbb{E}[e^{iuB_n}] - \mathbb{E}[e^{iuG}]| \leq \sup_{|\beta| \leq \alpha} |e^{iu\beta} - 1|,$$

and letting $\alpha \rightarrow 0$, we finally obtain $\limsup_{n \rightarrow \infty} |\mathbb{E}[e^{iuB_n}] - \mathbb{E}[e^{iuG}]| = 0$. Therefore $B_n \xrightarrow{\mathbb{P}\text{-law}} G$ which concludes the proof. \square

A.3 Central Limit Theorem

A.3.1 Preliminary results

Let \bar{Q} be the Markov kernel on $(\mathbb{X} \times \mathbb{N}_0) \times (\mathcal{X} \otimes \mathcal{P}(\mathbb{N}_0))$ defined by

$$\bar{Q}(x, n; dx' dn') = Q(x, dx') \tilde{R}(x', dn'),$$

and $\hat{\pi}$ be the probability measure on $\mathsf{X} \times \mathbb{N}$ defined by

$$\hat{\pi}(dx dn) = \tilde{\pi}(dx) \tilde{R}(x, dn).$$

Let $S_n = \sum_{i=1}^n \tilde{N}_i$ and define $k_n = \max \{k \in \mathbb{N}^* : S_k \leq n\}$ ensuring that $S_{k_n} \leq n < S_{k_n+1}$.

Lemma 6. Assume (H1). Then \bar{Q} admits $\hat{\pi}$ as invariant probability measure.

Proof. Let $A \in \mathcal{X} \otimes \mathcal{P}(\mathbb{N}_0)$. Then

$$\begin{aligned} \hat{\pi} \bar{Q}(A) &= \int_{(\mathsf{X} \times \mathbb{N}_0)^2} \tilde{\pi}(dx) \tilde{R}(x, dn) Q(x, dx') \tilde{R}(x', dn') \mathbf{1}_A(x', n') \\ &= \int_{\mathsf{X} \times \mathbb{N}_0} \tilde{\pi}(dx') \tilde{R}(x', dn') \mathbf{1}_A(x', n') \\ &= \hat{\pi}(A). \end{aligned}$$

□

Lemma 7. Assume (H1) and (H_{lin}). Then, for every $\xi' \in \mathbf{M}_1(\mathsf{X} \times \mathbb{N}_0)$ and measurable function $g : \mathsf{X} \times \mathbb{N}_0 \rightarrow \mathbb{R}$ such that $\hat{\pi}(|g|) < \infty$,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} g(\tilde{X}_k, \tilde{N}_k) = \hat{\pi}(g), \quad \mathbb{P}_{\xi'}^{\bar{Q}} - a.s. \quad (40)$$

Proof. The proof follows that of Theorem 5 and also relies on [21, Proposition 3.5]. Let \bar{h} be a harmonic function for the kernel \bar{Q} , i.e. for all $(x, n) \in \mathsf{X} \times \mathbb{N}$,

$$\bar{Q} \bar{h}(x, n) = \bar{h}(x, n),$$

and let us prove that \bar{h} is constant. For all $(x, n) \in \mathsf{X} \times \mathbb{N}$,

$$\bar{h}(x, n) = \bar{Q} \bar{h}(x, n) = \int_{\mathsf{X} \times \mathbb{N}_0} Q(x, dx') \tilde{R}(x', dn') \bar{h}(x', n'). \quad (41)$$

The integral on the right hand side of the equation above does not depend on n , therefore \bar{h} is also independant of n and we can write for all $(x, n) \in \mathsf{X} \times \mathbb{N}$,

$$\bar{h}(x, n) = \bar{h}(x, 0) =: h_0(x).$$

Then from (41) we have for all $x \in \mathsf{X}$,

$$h_0(x) = \int_{\mathsf{X}} Q(x, dx') h_0(x') = Q h_0(x). \quad (42)$$

which proves that h_0 is harmonic for Q . Using [21, Proposition 3.5], we get that h_0 is constant. Therefore so is \bar{h} and using [21, Proposition 3.5] again, we have completed the proof of the lemma. □

Lemma 8. For all $\zeta \in \mathbf{M}_1(\mathsf{X} \times \mathbb{N}_0)$, $\frac{k_n}{n} \xrightarrow{\mathbb{P}_{\zeta}^{\bar{Q}}\text{-prob}} \kappa^{-1}$.

Proof. Let $\beta > \kappa^{-1}$, we will prove that $\mathbb{P}_{\zeta}^{\bar{Q}} \left(\frac{k_n}{n} \geq \beta \right) \xrightarrow{n \rightarrow \infty} 0$. From the definition of k_n we have that $\{k_n \geq \beta n\} = \left\{ \sum_{i=1}^{\lfloor \beta n \rfloor} \tilde{N}_i \leq n \right\}$, hence

$$\mathbb{P}_{\zeta}^{\bar{Q}} \left(\frac{k_n}{n} \geq \beta \right) = \mathbb{P}_{\zeta}^{\bar{Q}} \left(\frac{1}{\lfloor \beta n \rfloor} \sum_{i=1}^{\lfloor \beta n \rfloor} \tilde{N}_i \leq \frac{n}{\lfloor \beta n \rfloor} \right),$$

which converges to 0 as n goes to infinity since by applying Lemma 7 with $g(x, \ell) = \ell$ we have

$$\frac{1}{[\beta n]} \sum_{i=1}^{[\beta n]} \tilde{N}_i \xrightarrow{\mathbb{P}_\zeta^{\bar{Q}}-a.s.} \int_{\mathbf{X} \times \mathbb{N}_0} \tilde{\pi}(dx) \tilde{R}(x, d\ell) \ell = \int_{\mathbf{X}} \tilde{\pi}(dx) \varrho_\kappa(x) = \kappa > \beta^{-1}.$$

Similarly we prove that for any $\beta < \kappa^{-1}$,

$$\mathbb{P}_\zeta^{\bar{Q}} \left(\frac{k_n}{n} < \beta \right) \xrightarrow{n \rightarrow \infty} 0,$$

by using that $\{k_n < \beta n\} = \left\{ \sum_{i=1}^{[\beta n]} \tilde{N}_i > n \right\}$, which completes the proof. \square

Lemma 9. Assume (H1) and (H_{lin}). Let $\zeta \in \mathbf{M}_1(\mathbf{X} \times \mathbb{N}_0)$, $f : \mathbf{X} \times \mathbb{N} \rightarrow \mathbb{R}$ be a measurable function, and $(k_n)_n \in \mathbb{N}^{\mathbb{N}}$ be a sequence of random variables such that $\frac{k_n}{n} \xrightarrow{\mathbb{P}_\zeta^{\bar{Q}}-\text{prob}} \kappa^{-1}$ and $\hat{\pi} f^2 < \infty$. Then,

$$\frac{f(\tilde{X}_{k_n}, \tilde{N}_{k_n})}{\sqrt{n}} \xrightarrow{\mathbb{P}_\zeta^{\bar{Q}}-\text{prob}} 0.$$

Proof. Let $\epsilon > 0$, we will prove that

$$\mathbb{P}_\zeta^{\bar{Q}} \left(g(\tilde{X}_{k_n}, \tilde{N}_{k_n}) > \epsilon^2 n \right) \xrightarrow{n \rightarrow \infty} 0,$$

where $g = f^2$. Let $\alpha, \beta \in \mathbb{R}^+$ be two nonnegative numbers such that $\alpha < \kappa^{-1} < \beta$ and $\beta - \alpha < \frac{\epsilon^2}{\hat{\pi}(g)}$.

$$\mathbb{P}_\zeta^{\bar{Q}} \left(g(\tilde{X}_{k_n}, \tilde{N}_{k_n}) > \epsilon^2 n \right) \leq \mathbb{P}_\zeta^{\bar{Q}} \left(\frac{k_n}{n} \notin [\alpha, \beta] \right) + \mathbb{P}_\zeta^{\bar{Q}} \left(g(\tilde{X}_{k_n}, \tilde{N}_{k_n}) > \epsilon^2 n, \frac{k_n}{n} \in [\alpha, \beta] \right).$$

From $\frac{k_n}{n} \xrightarrow{\mathbb{P}_\zeta^{\bar{Q}}-\text{prob}} \kappa^{-1}$ we get $\limsup_n \mathbb{P}_\zeta^{\bar{Q}} \left(\frac{k_n}{n} \notin [\alpha, \beta] \right) = 0$ and therefore,

$$\limsup_n \mathbb{P}_\zeta^{\bar{Q}} \left(g(\tilde{X}_{k_n}, \tilde{N}_{k_n}) > \epsilon^2 n \right) \leq \limsup_n \mathbb{P}_\zeta^{\bar{Q}} \left(g(\tilde{X}_{k_n}, \tilde{N}_{k_n}) > \epsilon^2 n, \frac{k_n}{n} \in [\alpha, \beta] \right).$$

Defining $A_n = \frac{1}{n} \sum_{i=1}^n g(\tilde{X}_i, \tilde{N}_i)$, we have

$$\begin{aligned} \mathbb{P}_\zeta^{\bar{Q}} \left(g(\tilde{X}_{k_n}, \tilde{N}_{k_n}) > \epsilon^2 n, \frac{k_n}{n} \in [\alpha, \beta] \right) &\leq \mathbb{P}_\zeta^{\bar{Q}} \left(\sum_{i=[n\alpha]}^{[n\beta]} g(\tilde{X}_i, \tilde{N}_i) > \epsilon^2 n \right) \\ &= \mathbb{P}_\zeta^{\bar{Q}} \left(\frac{[n\beta]}{n} A_{[n\beta]} - \frac{[n\alpha]}{n} A_{[n\alpha]} > \epsilon^2 \right) \xrightarrow{n \rightarrow \infty} 0, \end{aligned}$$

since $A_n \xrightarrow{\mathbb{P}_\zeta^{\bar{Q}}-a.s.} \hat{\pi} g$ by Lemma 7 and $\hat{\pi}(g)(\beta - \alpha) < \epsilon^2$. This concludes the proof. \square

Lemma 10. Let $(Y_k)_k$ be a Markov chain on \mathbf{Y} generated by a kernel T on $\mathbf{Y} \times \mathbf{Y}$ with invariant measure μ . Assume that there exists a measurable function $j : \mathbf{Y} \rightarrow \mathbb{R}$ such that the Poisson equation on \mathbf{Y} for the kernel T associated to the function j admits a solution J , i.e. for all $y \in \mathbf{Y}$

$$J(y) - TJ(y) = j(y) - \mu(j).$$

Then, considering the filtration $\mathcal{F}_n = \sigma(Y_{0:n})$,

$$\sum_{i=0}^{n-1} j(Y_i) - \mu(j) = M_n - J(Y_n) + J(Y_0),$$

where $M_n := \sum_{i=1}^n J(Y_i) - TJ(Y_{i-1})$ is a \mathcal{F}_n -martingale.

Proof. A simple index shift gives that

$$\sum_{i=0}^{n-1} j(Y_i) - \mu(j) = \sum_{i=1}^{n-1} J(Y_i) - TJ(Y_i) = \sum_{i=1}^n J(Y_i) - TJ(Y_{i-1}) - J(Y_n) + J(Y_1) = M_n - J(Y_n) + J(Y_1).$$

□

Let us state as a reminder the following theorem from [37] around which the proof of the following lemma will revolve.

Theorem 9 (Theorem A.3. of [38]). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{F}_{n,i})_{i \leq n}$ be a filtration on Ω . Assume $\mathbb{E}[U_{n,i}^2 | \mathcal{F}_{n,i-1}] < \infty$ for any $n \in \mathbb{N}_0$ and any $i = 1, \dots, n$, and*

$$\sum_{i=1}^n \left(\mathbb{E}[U_{n,i}^2 | \mathcal{F}_{n,i-1}] - \mathbb{E}[U_{n,i} | \mathcal{F}_{n,i-1}]^2 \right) \longrightarrow \sigma^2 \quad \text{for some } \sigma > 0 \quad (\text{H3})$$

$$\sum_{i=1}^n \mathbb{E}[U_{n,i}^2 \mathbf{1}_{\{|U_{n,i}| \geq \varepsilon\}} | \mathcal{F}_{n,i-1}] \longrightarrow 0 \quad \text{for any } \varepsilon > 0 \quad (\text{H4})$$

Then, for any real u ,

$$\mathbb{E} \left[\exp \left(iu \sum_{i=1}^n (U_{n,i} - \mathbb{E}[U_{n,i} | \mathcal{F}_{n,i-1}]) \right) \middle| \mathcal{F}_{n,0} \right] \longrightarrow \exp(-(u^2/2)\sigma^2).$$

Lemma 11. *Let $(Y_k)_k$ be a Markov chain on \mathcal{Y} with kernel T on $\mathcal{Y} \times \mathcal{Y}$ admitting an invariant probability measure μ . Assume that for every $\nu \in \mathcal{M}_1(\mathcal{Y})$ and any measurable function $g : \mathcal{Y} \rightarrow \mathbb{R}$ such that $\mu(|g|) < \infty$,*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} g(Y_k) = \mu(g), \quad \mathbb{P}_\nu^T - a.s. \quad (43)$$

Let $J : \mathcal{Y} \rightarrow \mathbb{R}$ be a measurable function such that $\mu(J^2) < \infty$. Consider the filtration $\mathcal{F}_n = \sigma(Y_{0:n})$ and the \mathcal{F}_n -martingale $M_n = \sum_{i=1}^n J(Y_i) - TJ(Y_{i-1}) = \sum_{i=1}^n \Delta M_i$. Then,

$$n^{-1/2} M_n \stackrel{\mathbb{P}_\nu^M - \text{law}}{\rightsquigarrow} \mathcal{N}(0, \sigma_J^2(h)),$$

with $\sigma_J^2(h) = \mathbb{E}_\mu^T[\Delta M_1^2]$.

Proof. Define $U_{n,i} = \frac{\Delta M_i}{\sqrt{n}}$ for any $n \geq i \geq 1$ and $\mathcal{F}_{n,i} = \mathcal{F}_i = \sigma(Y_{0:i})$ for any $i, n \in \mathbb{N}$. Let us verify the hypotheses of Theorem 9 :

- For (H3), we have

$$\begin{aligned} \sum_{i=1}^n \left(\mathbb{E}_\nu^T[U_{n,i}^2 | \mathcal{F}_{n,i-1}] - \mathbb{E}_\nu^T[U_{n,i} | \mathcal{F}_{n,i-1}]^2 \right) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\nu^T[\Delta M_i^2 | \mathcal{F}_{i-1}] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y_{i-1}}^T[\Delta M_1^2] \xrightarrow{n \rightarrow \infty} \mathbb{E}_\mu^T[\Delta M_1^2], \end{aligned}$$

where the limit is obtained from (43) with the function $g : y \mapsto \mathbb{E}_y^T[\Delta M_1^2]$.

- For (H4), let $A > 0$ be a positive integer,

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}_\nu^T[U_{n,i}^2 \mathbf{1}_{\{|U_{n,i}| \geq \varepsilon\}} | \mathcal{F}_{n,i-1}] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\nu^T[\Delta M_i^2 \mathbf{1}_{\{|\Delta M_i| \geq \varepsilon \sqrt{n}\}} | \mathcal{F}_{i-1}] \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\nu^T[\Delta M_i^2 \mathbf{1}_{\{|\Delta M_i| \geq A\}} | \mathcal{F}_{i-1}], \end{aligned}$$

for large enough values of n . Then, the Markov property gives us that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_\nu^T [\Delta M_i^2 \mathbf{1}_{\{|\Delta M_i| \geq A\}} | \mathcal{F}_{i-1}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y_{i-1}}^T [\Delta M_1^2 \mathbf{1}_{\{|\Delta M_1| \geq A\}}].$$

Applying (43) on the right hand side with $g : y \mapsto \mathbb{E}_y^T [\Delta M_1^2 \mathbf{1}_{\{|\Delta M_1| \geq A\}}]$ gives that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y_{i-1}}^T [\Delta M_1^2 \mathbf{1}_{\{|\Delta M_1| \geq A\}}] \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-a.s.} \mu g = \mathbb{E}_\mu^T [\Delta M_1^2 \mathbf{1}_{\{|\Delta M_1| \geq A\}}].$$

Now let $A \rightarrow \infty$, $(\mathbb{E}_\mu^T [\Delta M_1^2 \mathbf{1}_{\{|\Delta M_1| \geq A\}}])_A$ converges to 0 by monotone convergence. Hence

$$\frac{M_n}{\sqrt{n}} \xrightarrow{\mathbb{P}_\nu^T-law} \mathcal{N}(0, \sigma_J^2(h)) \text{ with } \sigma_J^2(h) = \mathbb{E}_\mu^T [\Delta M_1^2].$$

□

A.3.2 Proof of Theorem 6

Let h be a bounded measurable function and denote $h_0 := h - \pi h$. We want to prove that :

$$\frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} h_0(X_i) \xrightarrow{\mathbb{P}_\chi^P-law} \mathcal{N}(0, \sigma^2(h)).$$

Let $U_n = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} h_0(X_i)$. Let $\xi' = \xi \otimes \mu$ where μ is any probability measure on \mathbb{N} . The probability measure associated to the trajectories $(X_i)_i$ obtained from the Markov chains $(\tilde{X}_i, \tilde{N}_i)_i$ generated by \bar{Q} starting from ξ' is the same as the one associated to the sequence $(X_i)_i$ produced as the first component of the Markov chain $(X_i, N_i)_i$ with kernel P starting from χ defined by $\chi(f) = \int \xi(dx) S(x, dx') R(x', n') f(x', n')$. We will work with the former probability distribution (i.e. under $\mathbb{P}_{\xi'}^{\bar{Q}}$) as it is best suited for our proof. Denote $V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{k_n-1} \tilde{N}_i h_0(\tilde{X}_i)$ and let us

start by proving that $U_n - V_n \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-prob} 0$. By definition of k_n , $h_0(\tilde{X}_{k_n})$ appears less than \tilde{N}_{k_n} times in U_n and therefore,

$$|U_n - V_n| = (n - S_{k_n}) \frac{|h_0(\tilde{X}_{k_n})|}{\sqrt{n}} \leq \tilde{N}_{k_n} \frac{|h_0(\tilde{X}_{k_n})|}{\sqrt{n}}.$$

From Lemma 8, $\frac{k_n}{n} \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-prob} \kappa^{-1}$ and we can apply Lemma 9 to the function $(x, n) \mapsto n |h_0(x)|$ to obtain $\tilde{N}_{k_n} \frac{|h_0(\tilde{X}_{k_n})|}{\sqrt{n}} \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-prob} 0$. Let us write $V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{k_n-1} f(\tilde{X}_i, \tilde{N}_i)$ where $f : (x, n) \mapsto n h_0(x)$.

It now suffices to prove that $V_n \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-law} \mathcal{N}(0, \sigma^2(h))$. We will proceed in two steps :

(i) rewrite $V_n = \frac{1}{\sqrt{n}} M_{k_n} + \delta_n$ using a solution to the Poisson equation associated to f , where

$$(M_n)_{n \in \mathbb{N}} \text{ is a martingale and } \delta_n \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-prob} 0 ;$$

(ii) prove that $\frac{1}{\sqrt{n}} M_n \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-law} \mathcal{N}(0, \sigma_M^2(h))$ and apply Theorem 8 to obtain $\frac{1}{\sqrt{n}} M_{k_n} \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}}-law} \mathcal{N}(0, \kappa^{-1} \sigma_M^2(h))$.

Starting with (i), let H be the solution to the Poisson equation associated to ρh_0 for the kernel Q on \mathbb{X} given by $(\mathbf{H}_{\text{Pois}})$, i.e. for all $x \in \mathbb{X}$,

$$H(x) - QH(x) = \rho(x) h_0(x).$$

Then, $H_\kappa := \kappa H$ is a solution to the Poisson equation associated to $\varrho_\kappa h_0$ for the kernel Q on X , and

$$F(x, n) := H_\kappa(x) + nh_0(x) - \varrho_\kappa(x)h_0(x) \quad (44)$$

is a solution to the Poisson equation associated to f for the Markov kernel \bar{Q} such that $\hat{\pi}F^2 < \infty$. Indeed, for $(x, n) \in \mathsf{X} \times \mathbb{N}$ we have

$$\begin{aligned} \bar{Q}F(x, n) &= \int_{\mathsf{X} \times \mathbb{N}_0} Q(x, dx') \tilde{R}(x', dn') H(x') + \int_{\mathsf{X} \times \mathbb{N}_0} Q(x, dx') \tilde{R}(x', dn') n' h(x') \\ &\quad - \int_{\mathsf{X} \times \mathbb{N}_0} Q(x, dx') \tilde{R}(x', dn') \varrho_\kappa(x') h(x') \\ &= QH_\kappa(x) \end{aligned}$$

since $\int_{\mathbb{N}_0} \tilde{R}(x', dn') n' = \varrho_\kappa(x')$, and therefore

$$F(x, n) - \bar{Q}F(x, n) = H_\kappa(x) + nh_0(x) - \varrho_\kappa(x)h_0(x) - QH_\kappa(x) = nh_0(x) = f(x, n).$$

Moreover, $\hat{\pi}F^2 \leq 4(\kappa^2 \hat{\pi}H^2 + \hat{\pi}f^2 + \hat{\pi}(\varrho_\kappa h_0)^2) < \infty$ since $\hat{\pi}H^2 = \tilde{\pi}H^2 < \infty$ and $\hat{\pi}f^2 < \infty$ from **(H_{Pois})**. Then,

$$\hat{\pi}f^2 = \int_{\mathsf{X}} \left(\int_{\mathbb{N}_0} n^2 \tilde{R}(x, dn) \right) h_0(x)^2 \tilde{\pi}(dx) \geq \int_{\mathsf{X}} \left(\int_{\mathbb{N}_0} n \tilde{R}(x, dn) \right)^2 h_0(x)^2 \tilde{\pi}(dx) = \int_{\mathsf{X}} \varrho_\kappa(x)^2 h_0(x)^2 \tilde{\pi}(dx),$$

proving that $\hat{\pi}(\varrho_\kappa h_0)^2 = \tilde{\pi}(\varrho_\kappa h_0)^2 < \infty$. Now let $M_n = \sum_{i=2}^n F(\tilde{X}_i, \tilde{N}_i) - \bar{Q}F(\tilde{X}_{i-1}, \tilde{N}_{i-1})$ and consider the filtration $\mathcal{F}_n = \sigma(\tilde{X}_{1:n}, \tilde{N}_{1:n})$. From Lemma 10, $(M_n)_n$ is a \mathcal{F}_n -martingale, and

$$\sum_{i=1}^{n-1} f(\tilde{X}_i, \tilde{N}_i) = M_n - F(\tilde{X}_n, \tilde{N}_n) + F(\tilde{X}_1, \tilde{N}_1). \quad (45)$$

Therefore, $V_n = \frac{1}{\sqrt{n}} M_{k_n} + \delta_n$ with $\delta_n = \frac{F(\tilde{X}_{k_n}, \tilde{N}_{k_n})}{\sqrt{n}} + \frac{F(\tilde{X}_1, \tilde{N}_1)}{\sqrt{n}}$. Note that $\frac{F(\tilde{X}_1, \tilde{N}_1)}{\sqrt{n}} \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}-\text{prob}}} 0$ trivially, and $\frac{F(\tilde{X}_{k_n}, \tilde{N}_{k_n})}{\sqrt{n}} \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}-\text{prob}}} 0$ as a consequence of Lemma 9. Using Lemma 11 combined to Lemma 7 with $Y_i = (\tilde{X}_i, \tilde{N}_i)$,

$$\frac{M_n}{\sqrt{n}} \xrightarrow{\mathbb{P}_{\xi'}^{\bar{Q}-\text{law}}} \mathcal{N}(0, \sigma_F^2(h)) \text{ with } \sigma_F^2(h) = \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} [\Delta M_1^2].$$

which proves (ii) and concludes the first part of the proof. Let us now turn our attention to the expression of the variance claimed by the theorem :

$$\sigma^2(h) = \kappa \tilde{\sigma}^2(\varrho h_0) + \kappa^{-1} \hat{\sigma}^2(h_0, \kappa),$$

with $\tilde{\sigma}^2(\varrho h_0) = 2\tilde{\pi}(\varrho h_0 H) - \tilde{\pi}((\varrho h_0)^2)$ and $\hat{\sigma}^2(h_0, \kappa) = \int_{\mathsf{X}} h_0(x)^2 \text{Var}^{\tilde{R}(x, \cdot)}[N] \tilde{\pi}(dx)$. From the expression of F in (44) and denoting $\Delta H_1 = H(\tilde{X}_1) - QH(\tilde{X}_0)$ we have

$$\begin{aligned} \sigma^2(h) &= \kappa^{-1} \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[\left(\kappa \Delta H_1 + (\tilde{N}_1 - \varrho_\kappa(\tilde{X}_1)) h(\tilde{X}_1) \right)^2 \right] \\ &= \kappa \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} [\Delta H_1^2] + \kappa^{-1} \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[(\tilde{N}_1 - \varrho_\kappa(\tilde{X}_1))^2 h(\tilde{X}_1)^2 \right] + 2 \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[\Delta H_1 (\tilde{N}_1 - \varrho_\kappa(\tilde{X}_1)) h(\tilde{X}_1) \right]. \quad (46) \end{aligned}$$

Let us take a look at the first term of the rhs :

$$\begin{aligned} \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} [\Delta H_1^2] &= \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[H(\tilde{X}_1)^2 \right] + \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[QH(\tilde{X}_0)^2 \right] - 2 \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[H(\tilde{X}_1) QH(\tilde{X}_0) \right] \\ &= \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[H(\tilde{X}_0)^2 \right] - \mathbb{E}_{\tilde{\pi}}^{\bar{Q}} \left[QH(\tilde{X}_0)^2 \right], \end{aligned}$$

where we used that $\hat{\pi}$ is a stationary probability measure. Noting that $H^2 - QH^2 = (H - QH)(H + QH) = \varrho h_0(2H - \varrho h_0)$, we finally obtain using **(H_{Poiss})**

$$\mathbb{E}_{\hat{\pi}}^Q [\Delta H_1^2] = 2\hat{\pi}(\varrho h_0 H) - \hat{\pi}((\varrho h_0)^2). \quad (47)$$

Let us now rewrite the second term of **(46)** :

$$\begin{aligned} \mathbb{E}_{\hat{\pi}}^Q [(\tilde{N}_1 - \varrho_{\kappa}(\tilde{X}_1))^2 h_0(\tilde{X}_1)^2] &= \int_{\mathbf{X}} \left(\int_{\mathbb{N}_0} (n - \varrho_{\kappa}(x))^2 \tilde{R}(x, dn) \right) h_0(x)^2 \hat{\pi}(dx) \\ &= \int_{\mathbf{X}} h_0(x)^2 \mathbb{V}\text{ar}^{\tilde{R}(x, \cdot)}[N] \hat{\pi}(dx), \end{aligned}$$

where $\mathbb{V}\text{ar}^{\tilde{R}(x, \cdot)}[N] = \int (n - \varrho_{\kappa}(x))^2 \tilde{R}(x, dn)$ due to **(H2)**. For the last term of **46**, the same hypothesis gives that

$$\mathbb{E}_{\hat{\pi}}^Q [\Delta H_1(\tilde{N}_1 - \varrho_{\kappa}(\tilde{X}_1))h_0(\tilde{X}_1)] = \mathbb{E}_{\hat{\pi}}^Q \left[\mathbb{E}_{\hat{\pi}}^Q [\tilde{N}_1 - \varrho_{\kappa}(\tilde{X}_1) | \tilde{X}_1, \tilde{X}_0] \Delta H_1 h_0(\tilde{X}_1) \right] = 0,$$

which concludes the second part of the proof.

A.4 Geometric ergodicity

A.4.1 Proof of Lemma 2 and Lemma 3

Proof of Lemma 2. We start with the first point **(i)**. From the definition of S in **(10)** and noting that \mathbf{C}_{η} is a $(1, \varepsilon\nu)$ -small set for Q , we have for all $x \in \mathbf{C}_{\eta}$ and $\mathbf{A} \in \mathcal{X}$,

$$\begin{aligned} S(x, \mathbf{A}) &= \sum_{k=1}^{\infty} \mathbb{E}_x^Q \left[\rho_{\tilde{R}}(X_k) \mathbf{1}_{\mathbf{A}}(X_k) \prod_{i=1}^{k-1} (1 - \rho_{\tilde{R}}(X_i)) \right] \\ &\geq Q(x, \rho_{\tilde{R}} \mathbf{1}_{\mathbf{A}}) \geq \varepsilon\nu(\rho_{\tilde{R}} \mathbf{1}_{\mathbf{A}}). \end{aligned}$$

Applying **(11)** with $n = 0$, we deduce that for all $x \in \mathbf{C}_{\eta}$ and all $\mathbf{B} \in \mathcal{X} \otimes \mathcal{P}(\mathbb{N}_0)$,

$$\begin{aligned} P(x, 0; \mathbf{B}) &= \int_{\mathbf{B}} S(x, dx') R(x', du') \\ &\geq \varepsilon \int_{\mathbf{B}} \nu(dx') \rho_{\tilde{R}}(x') R(x', du') \\ &=: \varepsilon \tilde{\nu}(\mathbf{B}), \end{aligned}$$

which shows **(i)**. Let us now turn to **(ii)**. Noting that $\mathbf{C}_{\eta}^+ := \mathbf{C}_{\eta} \cap \{\tilde{R}(\cdot, 1) > 0\} = \mathbf{C}_{\eta} \cap \{R(\cdot, 0) > 0\}$ we have,

$$\begin{aligned} \tilde{\nu}(\mathbf{C}_{\eta} \times \{0\}) &= \int_{\mathbf{C}_{\eta}} \nu(dx) \rho_{\tilde{R}}(x) R(x, 0) \\ &\geq \eta \int_{\mathbf{C}_{\eta}} \nu(dx) R(x, 0) \\ &= \eta \int_{\mathbf{C}_{\eta}^+} \nu(dx) R(x, 0) > 0, \end{aligned}$$

where the last inequality stems from $\nu(\mathbf{C}_{\eta}^+) > 0$ and $R(x, 0) > 0$ for all $x \in \mathbf{C}_{\eta}^+$. Hence **(ii)**. \square

Proof of Lemma 3. Let \mathbf{A} be accessible for Q such that $\epsilon_{\mathbf{A}} := \inf_{x \in \mathbf{A}} \rho_{\tilde{R}}(x) > 0$. We first show that \mathbf{A} is accessible for S . Let $x \in \mathbf{X}$ and $n \in \mathbb{N}_0$ such that $Q^n(x, \mathbf{A}) > 0$. Let us consider the representation of S using the kernel G defined in **(2)**. Define $\mathbf{D} := \{(x, u) \in \mathbf{X} \times [0, 1] : u \leq \rho_{\tilde{R}}(x)\}$

and $\sigma_{\mathbf{D}}^{(m)}$ the m -th return time to the set \mathbf{D} . Then, for any probability measure μ on $[0, 1]$,

$$\begin{aligned} \int_{\mathbf{X}} Q^n(x, dy) \rho_{\tilde{R}}(y) \mathbf{1}_{\mathbf{A}}(y) &= \mathbb{P}_{\delta_x \otimes \mu}^G((X_n, U_n) \in \mathbf{D}, X_n \in \mathbf{A}) \\ &= \mathbb{P}_{\delta_x \otimes \mu}^G(\exists m \in [1 : n], n = \sigma_{\mathbf{D}}^{(m)}, X_{\sigma_{\mathbf{D}}^{(m)}} \in \mathbf{A}) \\ &\leq \mathbb{P}_{\delta_x \otimes \mu}^G(\exists m \in [1 : n], X_{\sigma_{\mathbf{D}}^{(m)}} \in \mathbf{A}) \\ &= \mathbb{P}_x^S(\exists m \in [1 : n], X_m \in \mathbf{A}) \\ &\leq \sum_{m=1}^n \mathbb{P}_x^S(X_m \in \mathbf{A}). \end{aligned}$$

Hence, $\sum_{m=1}^n S^m(x, \mathbf{A}) \geq \epsilon_{\mathbf{A}} Q^n(x, \mathbf{A}) > 0$ and there exists an integer $m \leq n$ such that $S^m(x, \mathbf{A}) > 0$. Thus, \mathbf{A} is an accessible set for S .

We now show that $\mathbf{A} \times \{0\}$ is accessible for P . Let $(x, k) \in \mathbf{X} \times \mathbb{N}_0$. From the first part of the proof, there exists $m \in \mathbb{N}_0$ such that $S^m(x, \mathbf{A}) > 0$. Indeed, let $\mathbf{B} := \mathbf{A} \times \{0\}$, $\mathbf{F} := \mathbf{X} \times \{0\}$ and $(Y_n = (X_n, N_n))_{n \in \mathbb{N}_0}$ be a Markov chain of kernel P . Start by noting that if $x \in \mathbf{X}$ and $N \sim R(x, \cdot)$,

$$\mathbb{P}_{(x,0)}^P(\sigma_{\mathbf{F}} < \infty) = \mathbb{P}(N < \infty) = 1$$

since (H3) ensures that $\mathbb{E}[N] < \infty$. Then, using the strong Markov inequality, we obtain by induction that

$$\mathbb{P}_{(x,0)}^P(\sigma_{\mathbf{F}}^{(m)} < \infty) = 1.$$

Hence,

$$\begin{aligned} \mathbb{P}_{(x,0)}^P(Y_{\sigma_{\mathbf{F}}^{(m)}} \in \mathbf{B}) &= \mathbb{P}_{(x,0)}^P(Y_{\sigma_{\mathbf{F}}^{(m)}} \in \mathbf{B}, \sigma_{\mathbf{F}}^{(m)} < \infty) \\ &= \sum_{\ell=1}^{\infty} \mathbb{P}_{(x,0)}^P(Y_{\sigma_{\mathbf{F}}^{(m)}} \in \mathbf{B}, \sigma_{\mathbf{F}}^{(m)} = \ell) \\ &= \sum_{\ell=1}^{\infty} \mathbb{P}_{(x,0)}^P(Y_{\ell} \in \mathbf{B}, \sigma_{\mathbf{F}}^{(m)} = \ell) \\ &\leq \sum_{\ell=1}^{\infty} \mathbb{P}_{(x,0)}^P(Y_{\ell} \in \mathbf{B}) \\ &= \sum_{\ell=1}^{\infty} P^{\ell}(x, 0; \mathbf{B}). \end{aligned}$$

Since by definition of P in (11) we have

$$\mathbb{P}_{(x,0)}^P(Y_{\sigma_{\mathbf{F}}^{(m)}} \in \mathbf{B}) = \mathbb{P}_x^S(X_m \in \mathbf{A}) > 0,$$

at least one of the terms in the sum above is positive and therefore there exists $\ell \in \mathbb{N}_0$ such that

$$P^{\ell}(x, 0; \mathbf{A} \times \{0\}) > 0.$$

Using the definition of P in (11) once more, we have that $P^k(x, k; \{(x, 0)\}) = 1$ and so

$$P^{k+\ell}(x, k; \mathbf{A} \times \{0\}) > 0,$$

which concludes the proof. □

A.4.2 Proof of Lemma 4

(Proof of Lemma 4). Let $\mathbf{B} := \mathbf{C}_{\eta} \times \{0\}$ and $\mathbf{F} := \mathbf{X} \times \{0\}$. Let $\beta \in (1, \infty)$ be an arbitrary constant and let $D < \infty$ be any positive constant (assuming it exists) such that

$$\sup_{x \in \mathbf{X}} \int_{\mathbb{N}_0} \beta^{n+1} R(x, dn) \leq D.$$

We first show that for all $(x, n) \in \mathbf{X} \times \mathbb{N}_0$ and $\beta > 1$ we have:

$$\mathbb{E}_{(x,n)}^P [\beta^{\sigma_B}] \leq \beta^n \mathbb{E}_x^S [D^{\sigma_{C_\eta}}]. \quad (48)$$

Let us start with the case $n = 0$:

$$\mathbb{E}_{(x,0)}^P [\beta^{\sigma_B}] = \sum_{\ell=1}^{\infty} g^{(\ell)}(x), \quad (49)$$

where $g^{(\ell)}(x) := \mathbb{E}_{(x,0)}^P \left[\beta^{\sigma_F^{(\ell)}} \mathbf{1}_{\{\sigma_F^{(\ell)} = \sigma_B\}} \right]$.

- For $\ell = 1$,

$$g^{(1)}(x) = \mathbb{E}_{(x,0)}^P \left[\beta^{N_1+1} \mathbf{1}_{C_\eta}(X_{N_1+1}) \right] = \int_{\mathbf{X}} \left(\int_{\mathbb{N}_0} \beta^{n+1} R(x, dn) \right) \mathbf{1}_{C_\eta}(x') S(x, dx') \leq D \times \mathbb{P}_x^S (\sigma_{C_\eta} = 1), \quad (50)$$

where the second equality holds from the definition of P in (11) ensuring that $X_{N_1+1} = X_1$ $\mathbb{P}_{(x,0)}^P$ -a.s.

- For $\ell > 1$, let us note that $\{\sigma_F^{(\ell)} = \sigma_B\} = \{\sigma_F^{(\ell-1)} \circ \theta^{\sigma_F} = \sigma_B \circ \theta^{\sigma_F}, X_{\sigma_F} \notin C_\eta\}$ as $\mathbb{P}_{(x,0)}^P$ -a.s. we have $\sigma_F^{(\ell)} = \sigma_F^{(\ell-1)} \circ \theta^{\sigma_F} + \sigma_F$, and $\sigma_B = \sigma_B \circ \theta^{\sigma_F} + \sigma_F$ under the event $\{\sigma_B > \sigma_F\} \supset \{\sigma_F^{(\ell)} = \sigma_B\}$ for $\ell > 1$. Hence,

$$\begin{aligned} g^{(\ell)}(x) &= \mathbb{E}_{(x,0)}^P \left[\beta^{\sigma_F^{(\ell)}} \mathbf{1}_{\{\sigma_F^{(\ell)} = \sigma_B\}} \right] \\ &\leq \mathbb{E}_{(x,0)}^P \left[\beta^{\sigma_F} \mathbf{1}_{\bar{C}_\eta}(X_{\sigma_F}) \mathbb{E}_{(x,0)}^P \left[\beta^{\sigma_F^{(\ell-1)} \circ \theta^{\sigma_F}} \mathbf{1}_{\{\sigma_F^{(\ell-1)} \circ \theta^{\sigma_F} = \sigma_B \circ \theta^{\sigma_F}\}} | \mathcal{F}_{\sigma_F} \right] \right] \\ &= \mathbb{E}_{(x,0)}^P \left[\beta^{\sigma_F} \mathbf{1}_{\bar{C}_\eta}(X_{\sigma_F}) \mathbb{E}_{(X_{\sigma_F},0)}^P \left[\beta^{\sigma_F^{(\ell-1)}} \mathbf{1}_{\{\sigma_F^{(\ell-1)} = \sigma_B\}} \right] \right] \quad (51) \end{aligned}$$

$$= \mathbb{E}_{(x,0)}^P \left[\beta^{N_1+1} \mathbf{1}_{\bar{C}_\eta}(X_1) \mathbb{E}_{(X_1,0)}^P \left[\beta^{\sigma_F^{(\ell-1)}} \mathbf{1}_{\{\sigma_F^{(\ell-1)} = \sigma_B\}} \right] \right], \quad (52)$$

where (51) comes from the strong Markov property applied to the Markov chain $(X_i, N_i)_{i \in \mathbb{N}_0}$ with the stopping time σ_F and (52) comes from the definition of the kernel P since X_1 is repeated $N_1 + 1$ times while the second component decreases by one at each iteration until reaching zero. The definition of P in (11) gives

$$\begin{aligned} \mathbb{E}_{(x,0)}^P \left[\beta^{N_1+1} \mathbf{1}_{\bar{C}_\eta}(X_1) \mathbb{E}_{(X_1,0)}^P \left[\beta^{\sigma_F^{(\ell-1)}} \mathbf{1}_{\{\sigma_F^{(\ell-1)} = \sigma_B\}} \right] \right] &= \int_{\mathbf{X}} \left(\int_{\mathbb{N}_0} \beta^{n+1} R(x, dn) \right) \mathbf{1}_{\bar{C}_\eta}(x') g^{(\ell-1)}(x') S(x, dx') \\ &\leq D \int_{\mathbf{X}} \mathbf{1}_{\bar{C}_\eta}(x') g^{(\ell-1)}(x') S(x, dx') \\ &= D \times \mathbb{E}_{(x,0)}^P \left[\mathbf{1}_{\bar{C}_\eta}(X_1) g^{(\ell-1)}(X_1) \right]. \end{aligned}$$

Hence,

$$g^{(\ell)}(x) = \mathbb{E}_{(x,0)}^P \left[\beta^{\sigma_F^{(\ell)}} \mathbf{1}_{\{\sigma_F^{(\ell)} = \sigma_B\}} \right] \leq D \times \mathbb{E}_{(x,0)}^P \left[\mathbf{1}_{\bar{C}_\eta}(X_{\sigma_F}) g^{(\ell-1)}(X_{\sigma_F}) \right],$$

which used in conjunction with (50) gives

$$\begin{aligned} g^{(\ell)}(x) &\leq D^\ell \times \mathbb{E}_{(x,0)}^P \left[\mathbf{1}_{\bar{C}_\eta}(X_{\sigma_F}) \dots \mathbf{1}_{\bar{C}_\eta}(X_{\sigma_F^{(\ell-1)}}) \mathbb{P}_{X_{\sigma_F^{(\ell-1)}}}^S (\sigma_{C_\eta} = 1) \right] \\ &= D^\ell \times \mathbb{E}_x^S \left[\mathbf{1}_{\bar{C}_\eta}(X_1) \dots \mathbf{1}_{\bar{C}_\eta}(X_{\ell-1}) \mathbb{E}_{X_{\ell-1}}^S \left[\mathbf{1}_{C_\eta}(X_1) \right] \right] \quad (53) \\ &= D^\ell \times \mathbb{E}_x^S \left[\mathbf{1}_{\bar{C}_\eta}(X_1) \dots \mathbf{1}_{\bar{C}_\eta}(X_{\ell-1}) \mathbb{E}_x^S \left[\mathbf{1}_{C_\eta}(X_\ell) | \mathcal{F}_{\ell-1} \right] \right] \\ &= D^\ell \times \mathbb{E}_x^S \left[\mathbf{1}_{\bar{C}_\eta}(X_1) \dots \mathbf{1}_{\bar{C}_\eta}(X_{\ell-1}) \mathbf{1}_{C_\eta}(X_\ell) \right] \\ &= D^\ell \times \mathbb{P}_x^S (\sigma_{C_\eta} = \ell), \end{aligned}$$

where (53) comes from the definition of S and P in (10) and (11). Plugging the above into (49) leads to the following upper bound,

$$\mathbb{E}_{(x,0)}^P [\beta^{\sigma_{\mathbb{B}}}] \leq \sum_{\ell=1}^{\infty} D^{\ell} \mathbb{P}_x^S (\sigma_{\mathbb{C}_{\eta}} = \ell) = \mathbb{E}_x^S [D^{\sigma_{\mathbb{C}_{\eta}}}] .$$

If $x \in \mathbb{C}_{\eta}$, then $\mathbb{E}_{(x,n)}^P [\beta^{\sigma_{\mathbb{B}}}] = \beta^n$, showing (48) for $x \in \mathbb{C}_{\eta}$. If $x \notin \mathbb{C}_{\eta}$, then we have $\sigma_{\mathbb{B}} = \sigma_{\mathbb{B}} \circ \theta^n + n$, $\mathbb{P}_{(x,n)}$ -a.s. Thus,

$$\mathbb{E}_{(x,n)}^P [\beta^{\sigma_{\mathbb{B}}}] = \beta^n \mathbb{E}_{(x,n)}^P [\beta^{\sigma_{\mathbb{B}} \circ \theta^n}] = \beta^n \mathbb{E}_{(x,0)}^P [\beta^{\sigma_{\mathbb{B}}}] \leq \beta^n \mathbb{E}_x^S [D^{\sigma_{\mathbb{C}_{\eta}}}] ,$$

which completes the proof of the inequality (48).

For any $\eta \in (0, \eta_0)$,

$$\mathbb{P}_x^S (\sigma_{\mathbb{C}_{\eta}} > k) = \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \mathbb{E}_x^Q \left[\prod_{j=1}^k \left[\prod_{i=s_{j-1}+1}^{s_j-1} (1 - \rho_{\bar{R}}(X_i)) \right] \rho_{\bar{R}}(X_{s_j}) \mathbf{1}_{\bar{\mathbb{C}}_{\eta}}(X_{s_j}) \right] ,$$

where we have set $s_j = \sum_{i=1}^j \ell_i$ for $j \geq 1$ and $s_0 = 0$. Using that $\rho_{\bar{R}} \mathbf{1}_{\bar{\mathbb{C}}_{\eta}} \leq \eta \mathbf{1}_{\bar{\mathbb{C}}_{\eta}}$ and $1 - \rho_{\bar{R}} \leq (1 - \eta_0) \mathbf{1}_{\mathbb{C}_{\eta_0}}$,

$$\begin{aligned} \mathbb{P}_x^S (\sigma_{\mathbb{C}_{\eta}} > k) &\leq \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \eta^k \mathbb{E}_x^Q \left[(1 - \eta_0)^{\sum_{j=1}^k \sum_{i=s_{j-1}+1}^{s_j-1} \mathbf{1}_{\mathbb{C}_{\eta_0}}(X_i)} \prod_{j=1}^k \mathbf{1}_{\bar{\mathbb{C}}_{\eta}}(X_{s_j}) \right] \\ &\leq \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \eta^k \mathbb{E}_x^Q \left[(1 - \eta_0)^{M_{[1:s_k-1] \setminus \{s_1, \dots, s_{k-1}\}}} \prod_{j=1}^{k-1} \mathbf{1}_{\bar{\mathbb{C}}_{\eta}}(X_{s_j}) \right] , \end{aligned} \quad (54)$$

where $M_I := \sum_{i \in I} \mathbf{1}_{\mathbb{C}_{\eta_0}}(X_i)$ for any $I \subset \mathbb{N}_0$. Moreover, since $\eta < \eta_0$, we have $\mathbb{C}_{\eta_0} \subset \mathbb{C}_{\eta}$, hence

$$\mathbf{1}_{\bar{\mathbb{C}}_{\eta}}(x) = (1 - \eta_0) \mathbf{1}_{\mathbb{C}_{\eta_0}}(x) \mathbf{1}_{\bar{\mathbb{C}}_{\eta}}(x) \leq (1 - \eta_0) \mathbf{1}_{\mathbb{C}_{\eta_0}}(x) . \quad (55)$$

Finally, setting for any arbitrary $\alpha \in (0, 1)$,

$$\begin{aligned} N_{\ell} &= \sum_{i=1}^{\ell-1} \mathbf{1}_{\mathbb{C}_{\eta_0}}(X_i) , \\ R_{\ell} &:= \mathbb{E}_x^Q [(1 - \eta_0)^{N_{\ell}}] , \\ R_{\ell,1} &= \mathbb{E}_x^Q [(1 - \eta_0)^{N_{\ell}} \mathbf{1}_{\{N_{\ell} > \alpha(\ell-k)\}}] , \\ R_{\ell,2} &= \mathbb{E}_x^Q [(1 - \eta_0)^{N_{\ell}} V(X_{\ell}) \mathbf{1}_{\{N_{\ell} \leq \alpha(\ell-k)\}}] , \end{aligned}$$

and plugging (55) into (54) combined with $V \geq 1$, we get

$$\mathbb{P}_x^S (\sigma_{\mathbb{C}_{\eta}} > k) \leq \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \eta^k R_{s_k} \leq \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \eta^k \{R_{s_k,1} + R_{s_k,2}\} \leq \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \eta^k \left[(1 - \eta_0)^{\alpha(s_k-k)} + R_{s_k,2} \right] . \quad (56)$$

We now give an explicit upper bound for $R_{\ell,2}$ for $\ell \geq k$. Under (\mathbf{H}_{df}) , $QV(x) \leq \lambda V(x)$ for $x \notin \mathbb{C}_{\eta_0}$ and $QV(x) \leq b_{\infty} V(x)$ for $x \in \mathbb{C}_{\eta_0}$. Therefore, for any $x \in \mathbb{X}$,

$$(1 - \eta_0) \mathbf{1}_{\mathbb{C}_{\eta_0}}(x) QV(x) \leq (b_{\infty} (1 - \eta_0) \lambda^{-1}) \mathbf{1}_{\mathbb{C}_{\eta_0}}(x) \lambda V(x) \leq A \mathbf{1}_{\mathbb{C}_{\eta_0}}(x) \lambda V(x) ,$$

where $A := 1 \vee b_{\infty} (1 - \eta_0) \lambda^{-1}$. This implies by applying the tower rule conditionally on $X_{1:\ell-1}$, then $X_{1:\ell-2}$ and so on,

$$\mathbb{E}_x^Q \left[V(X_{\ell}) \frac{(1 - \eta_0)^{N_{\ell}}}{A^{N_{\ell}} \lambda^{\ell-1}} \right] = \mathbb{E}_x^Q \left[V(X_1) \prod_{i=1}^{\ell-1} \frac{(1 - \eta_0) \mathbf{1}_{\mathbb{C}_{\eta_0}}(X_i) V(X_{i+1})}{A \mathbf{1}_{\mathbb{C}_{\eta_0}}(X_i) \lambda V(X_i)} \right] \leq \mathbb{E}_x^Q [V(X_1)] = QV(x) \leq b_{\infty} V(x) , \quad (57)$$

where the last inequality follows from (28). Since $A \geq 1$, we have $1 \leq A^{\alpha(\ell-k)}/A^{N_\ell}$ on $\{N_\ell \leq \alpha(\ell-k)\}$ and hence

$$R_{\ell,2} \leq A^{\alpha(\ell-k)} \lambda^{\ell-1} \mathbb{E}_x^Q \left[V(X_\ell) \frac{(1-\eta_0)^{N_\ell}}{A^{N_\ell} \lambda^{\ell-1}} \right] \leq A^{\alpha(\ell-k)} \lambda^{\ell-1} b_\infty V(x),$$

where the last inequality follows from (57). Pick α small enough so that $\lambda A^\alpha < 1$. Plugging the inequality above (with ℓ replaced by $s_k = \ell_1 + \dots + \ell_k$) into (56) yields

$$\begin{aligned} \mathbb{P}_x^S(\sigma_{C_\eta} > k) &\leq \sum_{\ell_1, \dots, \ell_k=1}^{\infty} \eta^k \left[(1-\eta_0)^{\alpha(\ell_1+\dots+\ell_k-k)} + [\lambda A^\alpha]^{\ell_1+\dots+\ell_k} A^{-\alpha k} \lambda^{-1} b_\infty V(x) \right] \\ &= \eta^k \left[\left(\frac{(1-\eta_0)^\alpha}{1-(1-\eta_0)^\alpha} \right)^k \frac{1}{(1-\eta_0)^{\alpha k}} + \left(\frac{\lambda A^\alpha}{1-\lambda A^\alpha} \right)^k \frac{1}{A^{\alpha k}} \lambda^{-1} b_\infty V(x) \right] \\ &= \eta^k \left[\left(\frac{1}{1-(1-\eta_0)^\alpha} \right)^k + \left(\frac{\lambda}{1-\lambda A^\alpha} \right)^k \lambda^{-1} b_\infty V(x) \right]. \end{aligned}$$

Now set $\gamma := \max\left(\frac{1}{1-(1-\eta_0)^\alpha}, \frac{\lambda}{1-\lambda A^\alpha}\right)$ and choose $\eta < \eta_0$ sufficiently small so that $\eta\gamma < 1$. Then,

$$\mathbb{P}_x^S(\sigma_{C_\eta} > k) \leq \eta^k \gamma^k (1 + \lambda^{-1} b_\infty V(x)),$$

and if $D \in (1, \eta^{-1}\gamma^{-1})$,

$$\begin{aligned} \mathbb{E}_x^S \left[\frac{D^{\sigma_{C_\eta}} - 1}{D - 1} \right] &= \mathbb{E}_x^S \left[\sum_{k=0}^{\infty} D^k \mathbf{1}_{\{k < \sigma_{C_\eta}\}} \right] = \sum_{k=0}^{\infty} D^k \mathbb{P}_x^S(\sigma_{C_\eta} > k) \\ &\leq \sum_{k=0}^{\infty} D^k \gamma^k \eta^k (1 + \lambda^{-1} b_\infty V(x)) = \frac{1 + \lambda^{-1} b_\infty V(x)}{1 - D\gamma\eta}. \end{aligned}$$

From (H3) there exists $\beta_0 \in (1, \infty)$ and $D_0 < \infty$ such that

$$\sup_{x \in \mathbf{X}} \int_{\mathbb{N}_0} \beta_0^{n+1} R(x, dn) \leq D_0.$$

Let $r \in (0, 1)$ and consider $\beta_r = \beta_0^r$. From Hölder's inequality,

$$\int_{\mathbb{N}_0} \beta_r^{n+1} R(x, dn) \leq \left(\int_{\mathbb{N}_0} \beta_0^{n+1} R(x, dn) \right)^r \leq D_0^r. \quad (58)$$

Choose r such that $D_r := D_0^r \in (1, \eta^{-1}\gamma^{-1})$. We can now apply the above inequality in combination with (48) using the couple (β_r, D_r) instead of (β, D) and obtain for all $(x, n) \in \mathbf{X} \times \mathbb{N}_0$,

$$\mathbb{E}_{(x,n)}^P[\beta_r^{\sigma_{C_\eta}}] \leq \beta_r^n \mathbb{E}_x^S[D_r^{\sigma_{C_\eta}}] \leq \beta_r^n \left[1 + (D_r - 1) \frac{1 + \lambda^{-1} b_\infty V(x)}{1 - D_r \gamma \eta} \right] \leq \beta_\star \beta_r^n V(x)$$

since $V \geq 1$, and where

$$\beta_\star = 1 + (D_r - 1) \frac{1 + \lambda^{-1} b_\infty}{1 - D_r \gamma \eta}.$$

The proof is completed. □

B Numerical experiments

B.1 Details of the hyperparameters for the normalizing flows

The normalizing flow is a RQSpline with 10 layers, 8 bins, and a (128, 128) hidden size. The local sampler is a MALA algorithm with step size 0.1. Training consists of a total of 30 loops with a

unique epoch each time. 10 global steps are implemented with a further 10 local steps between each, and the optimizer (Adam) has a learning rate of $8 \cdot 10^{-4}$ and a momentum of 0.9. The seed is 1250. The code generating the figures is available at https://github.com/charlyandral/importance_markov_chain.

References

- [1] N. Metropolis, S. Ulam, The Monte Carlo Method, *Journal of the American Statistical Association* 44 (247) (1949) 335–341.
- [2] C. P. Robert, G. Casella, *Monte Carlo Statistical Methods*, 2nd Edition, Springer Texts in Statistics, Springer New York, New York, NY, 2010. doi:10.1007/978-1-4757-4145-2.
- [3] H. Kahn, "Modification of the Monte Carlo Method," in *Proceedings, seminar on scientific computation*, November, 1949, IBM, New York, NY, 1950.
- [4] L. Devroye, *Non-Uniform Random Variate Generation*, Springer, New York Heidelberg, 1986.
- [5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics* 21 (6) (1953) 1087–1092. doi:10.1063/1.1699114.
- [6] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* 57 (1) (1970) 97–109. doi:10.1093/biomet/57.1.97.
- [7] L. Fosdick, Monte Carlo computations on the Ising lattice, *Methods in computational physics* 1 (1963) 245–280.
- [8] Z. I. Botev, P. L'Ecuyer, B. Tuffin, Markov chain importance sampling with applications to rare event probability estimation, *Statistics and Computing* 23 (2) (2013) 271–285. doi:10.1007/s11222-011-9308-2.
- [9] I. Raices Cruz, J. Lindström, M. C. M. Troffaes, U. Sahlin, Iterative importance sampling with Markov chain Monte Carlo sampling in robust Bayesian analysis, *Computational Statistics & Data Analysis* 176 (2022) 107558. doi:10.1016/j.csda.2022.107558.
- [10] I. Schuster, I. Klebanov, Markov Chain Importance Sampling—A Highly Efficient Estimator for MCMC, *Journal of Computational and Graphical Statistics* 30 (2) (2021) 260–268. doi:10.1080/10618600.2020.1826953.
- [11] S. N. MacEachern, L. M. Berliner, Subsampling the Gibbs Sampler, *The American Statistician* 48 (3) (1994) 188. doi:10.2307/2684714.
- [12] W. A. Link, M. J. Eaton, On thinning of chains in MCMC, *Methods in Ecology and Evolution* 3 (1) (2012) 112–115. doi:10.1111/j.2041-210X.2011.00131.x.
- [13] A. B. Owen, Statistically Efficient Thinning of a Markov Chain Sampler, *Journal of Computational and Graphical Statistics* 26 (3) (2017) 738–744. doi:10.1080/10618600.2017.1336446.
- [14] S. K. Sahu, A. A. Zhigljavsky, Self-regenerative Markov chain Monte Carlo with adaptation, *Bernoulli* 9 (3) (2003) 395–422. doi:10.3150/bj/1065444811.
- [15] J. Gåsemyr, Markov chain monte carlo algorithms with independent proposal distribution and their relationship to importance sampling and rejection sampling, *Preprint series. Statistical Research Report* (2002) 1–25.
- [16] W. H. Wong, F. Liang, Dynamic weighting in Monte Carlo and optimization, *Proceedings of the National Academy of Sciences* 94 (26) (1997) 14220–14224. doi:10.1073/pnas.94.26.14220.
- [17] J. S. Liu, F. Liang, W. H. Wong, A Theory for Dynamic Weighting in Monte Carlo Computation, *Journal of the American Statistical Association* 96 (454) (2001) 561–573. doi:10.1198/016214501753168253.

- [18] S. Malefaki, G. Iliopoulos, On convergence of properly weighted samples to the target distribution, *Journal of Statistical Planning and Inference* 138 (4) (2008) 1210–1225. doi:[10.1016/j.jspi.2007.05.030](https://doi.org/10.1016/j.jspi.2007.05.030).
- [19] R. Douc, C. P. Robert, A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms, *The Annals of Statistics* 39 (1) (2011) 261 – 277. doi:[10.1214/10-AOS838](https://doi.org/10.1214/10-AOS838).
- [20] R. Douc, E. Moulines, P. Priouret, P. Soulier, *Markov Chains, Operation Research and Financial Engineering*, Springer, 2018. doi:[10.1007/978-3-319-97704-1](https://doi.org/10.1007/978-3-319-97704-1).
- [21] R. Douc, A. Durmus, A. Enfroy, J. Olsson, Boost your favorite Markov Chain Monte Carlo sampler using Kac’s theorem: The Kick-Kac teleportation algorithm, arXiv:2201.05002 [cs, math, stat] (2022) 1–35 arXiv:[2201.05002](https://arxiv.org/abs/2201.05002).
- [22] G. O. Roberts, R. L. Tweedie, Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms, *Biometrika* 83 (1) (1996) 95–110.
- [23] C. Andrieu, G. O. Roberts, The pseudo-marginal approach for efficient Monte Carlo computations, *The Annals of Statistics* 37 (2) (2009) 697–725. doi:[10.1214/07-AOS574](https://doi.org/10.1214/07-AOS574).
- [24] C. Andrieu, M. Vihola, Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms, *The Annals of Applied Probability* 25 (2) (2015) 1030–1077. doi:[10.1214/14-AAP1022](https://doi.org/10.1214/14-AAP1022).
- [25] M. D. Hoffman, A. Gelman, The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo., *J. Mach. Learn. Res.* 15 (1) (2014) 1593–1623.
- [26] C. Durkan, A. Bekasov, I. Murray, G. Papamakarios, *Neural Spline Flows*, in: *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 2019.
- [27] M. Gabrié, G. M. Rotskoff, E. Vanden-Eijnden, Adaptive Monte Carlo augmented with normalizing flows, *Proceedings of the National Academy of Sciences* 119 (10) (2022) e2109420119.
- [28] K. W. Wong, M. Gabrié, D. Foreman-Mackey, flowMC: Normalizing-flow enhanced sampling package for probabilistic inference in Jax, arXiv preprint arXiv:2211.06397 (2022) 1–3 arXiv:[2211.06397](https://arxiv.org/abs/2211.06397).
- [29] L. Tierney, Markov chains for exploring posterior distributions, *the Annals of Statistics* (1994) 1701–1728 doi:[10.1214/aos/1176325750](https://doi.org/10.1214/aos/1176325750).
- [30] J. Berger, An overview of robust bayesian analysis, *Test* 3 (1) (1994) 5–124.
- [31] R. Weiss, An approach to bayesian sensitivity analysis, *Journal of the Royal Statistical Society Series B (Methodological)* 58 (4) (1996) 739–50.
- [32] D. Jackson, I. R. White, J. Carpenter, Identifying influential observations in bayesian models by using markov chain monte carlo, *Statistics in Medicine* (2012) 1238–48.
- [33] M. Banterle, C. Grazian, A. Lee, C. P. Robert, Accelerating metropolis-hastings algorithms by delayed acceptance, *Foundations of Data Science* 1 (2) (2019) 103–128.
- [34] P. Grünwald, T. van Ommen, Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It, *Bayesian Analysis* 12 (4) (2017) 1069 – 1103.
- [35] P. Grünwald, The e-posterior, *Philosophical Transactions of the Royal Society A* 381 (2023).
- [36] J. M. Hilbe, R. S. de Souza, E. o. bayesian models for astrophysical data, Cambridge University Press (2017).
- [37] R. Douc, E. Moulines, Limit theorems for weighted samples with applications to sequential Monte Carlo methods, *The Annals of Statistics* 36 (5) (2008) 101–107. doi:[10.1214/07-AOS514](https://doi.org/10.1214/07-AOS514).
- [38] R. Douc, E. Moulines, Limit theorems for weighted samples with applications to sequential monte carlo methods, *The Annals of Statistics* 36 (5) (2008) 2344–2376.