*Bernoulli* **17**(3), 2011, 987–1014
DOI: 10.3150/10-BEJ307

# On nonlinear Markov chain Monte Carlo

CHRISTOPHE ANDRIEU[1], AJAY JASRA[2], ARNAUD DOUCET[3] and PIERRE DEL MORAL[4]

[1]*Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK. E-mail:* c.andrieu@bris.ac.uk
[2]*Department of Mathematics, Imperial College London, London, SW7 2AZ, UK. E-mail:* a.jasra@ic.ac.uk
[3]*Department of Statistics, University of British Columbia, Vancouver, V6T 1Z2, Canada. E-mail:* a.doucet@cs.ubc.ca
[4]*Centre INRIA & Institut de Mathématiques de Bordeaux, Université de Bordeaux I, 33405, France. E-mail:* pierre.del-moral@inria.fr

Let $\mathcal{P}(E)$ be the space of probability measures on a measurable space $(E, \mathcal{E})$. In this paper we introduce a class of nonlinear Markov chain Monte Carlo (MCMC) methods for simulating from a probability measure $\pi \in \mathcal{P}(E)$. Nonlinear Markov kernels (see [*Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications* (2004) Springer]) $K : \mathcal{P}(E) \times E \to \mathcal{P}(E)$ can be constructed to, in some sense, improve over MCMC methods. However, such nonlinear kernels cannot be simulated exactly, so approximations of the nonlinear kernels are constructed using auxiliary or potentially self-interacting chains. Several nonlinear kernels are presented and it is demonstrated that, under some conditions, the associated approximations exhibit a strong law of large numbers; our proof technique is via the Poisson equation and Foster–Lyapunov conditions. We investigate the performance of our approximations with some simulations.

*Keywords:* Foster–Lyapunov condition; interacting Markov chains; nonlinear Markov kernels; Poisson equation

## 1. Introduction

Monte Carlo simulation is one of the most important elements of computational statistics. This is because of its relative simplicity and computational convenience in constructing estimates of high-dimensional integrals. That is, for a $\pi$-integrable $f : E \to \mathbb{R}$, we approximate:

$$\pi(f) := \int_E f(x)\pi(\mathrm{d}x) \tag{1.1}$$

by

$$S_n^X(f) = \frac{1}{n+1} \sum_{i=0}^{n} f(X_i),$$

where $S_n^X(\mathrm{d}u) := \frac{1}{n+1} \sum_{i=0}^{n} \delta_{X_i}(\mathrm{d}u)$ is the empirical measure based upon random variables $\{X_k\}_{0 \le k \le n}$ drawn from $\pi$. Such integrals appear routinely in Bayesian statistics, in terms of posterior expectations; see [26] and the references therein. In those cases, $E$ is often of very high dimension and complex simulation methods such as MCMC [26] and sequential Monte Carlo (SMC) [10,13] need to be used.

It has long been known by Monte Carlo specialists that standard MCMC algorithms often have difficulties in simulating from complicated distributions – for example, when the target $\pi$ exhibits multiple modes and/or possesses strong dependencies between subcomponents of $X$. In the former case, the Markov chain can take an unreasonable amount of time to jump between these modes and the estimates of (1.1) are very inaccurate.

As a result, there have been a large number of alternative methods proposed in the literature; we detail some of them here. Many of these approaches have relied upon MCMC techniques such as adaptive MCMC [5,20], which, in some instances, attempts to improve the mixing properties of the transition kernel by using the information learned in the past. In addition, there are methods that rely upon the simulation of parallel Markov chains [16] and genetic algorithm type moves; see [22] for a review. These latter methods use the idea of running some of the parallel chains with invariant probability measure $\eta$, where $\eta$ is easier to explore and is related to $\pi$; hence the samples of the parallel chains can provide valuable information for simulating from $\pi$. Extensions to MCMC-based simulation methods have combined MCMC with SMC ideas, see, for example, [2,11]. Such approaches are often more flexible than MCMC.

In this paper, we consider another alternative: nonlinear MCMC via auxiliary or self-interacting approximations. Such methods rely primarily upon the ideas of MCMC. However, it is demonstrated below that the auxiliary/self-interacting approximation idea is similar to that of approximating Feynman–Kac formulae [10] and as such is linked to SMC methodology. It should be noted that related ideas have appeared, directly in [9] and indirectly in [23]; see [4,7] for some theoretical analysis. Subsequent to the first versions of this work [3] a variety of related articles have appeared: [6–8]; we cite these where appropriate, but note the substantial overlap between our work and these papers.

## 1.1. Nonlinear Markov kernels via interacting approximations

Standard MCMC algorithms rely on Markov kernels of the form $K : E \to \mathcal{P}(E)$. These Markov kernels are *linear* operators on $\mathcal{P}(E)$; that is, $\mu(\mathrm{d}y) = \int_E \xi(\mathrm{d}x)K(x, \mathrm{d}y)$, where $\mu, \xi \in \mathcal{P}(E)$. A *nonlinear* Markov kernel $K : \mathcal{P}(E) \times E \to \mathcal{P}(E)$ is defined as a nonlinear operator on the space of probability measures. Nonlinear Markov kernels, $K_\mu$, can often be constructed to exhibit superior mixing properties to ordinary MCMC versions. For example, let

$$K_\mu(x, \mathrm{d}y) = (1-\epsilon)K(x, \mathrm{d}y) + \epsilon \int_E \mu(\mathrm{d}z)K(z, \mathrm{d}y), \tag{1.2}$$

where $K$ is a Markov kernel of invariant distribution $\pi$, $\epsilon \in (0, 1)$ and $\mu \in \mathcal{P}(E)$. Simulating from $K_\pi$ is clearly desirable as we allow regenerations from $\pi$, with $K_\pi$ strongly uniformly ergodic (see [27]). However, in most cases, it is not possible to simulate from $K_\pi$ and, instead, an approximation is proposed.

A self-interacting Markov chain (see [12]) generates a stochastic process $\{X_n\}_{n\geq 0}$ that is allowed to interact with values realized in the past. That is, we might approximate, at time $n+1$, $\mu$ by $S_n^X$. This process corresponds to generating a value from the history of the process, and then a mutation step, via the kernel $K$. In practice, the self-interaction can lead to very poor algorithmic performance [3]; an auxiliary Markov chain is used to approximate the nonlinear kernel.

### 1.2. Motivation and structure of the paper

In the context of stochastic simulation, self-interacting Markov chains (SIMCs), or IMCs, can be thought of as storing modes and then allowing the algorithm to return to them in a relatively simple way. Parametric adaptive MCMC can be thought of as an indirect application of this idea, where parameters of the kernel are optimized via a stochastic approximation algorithm. This approach does not retain all of the features of previously visited states. In other words, SIMCs can be considered as a nonparametric, or infinite-dimensional, generalization of parametric adaptive MCMC. It is thus the attractive idea of being able to fully exploit the information provided by the previous samples that has motivated us to investigate such algorithms.

This paper is structured as follows. We begin by giving our notation in Section 2. In Section 3 our simulation methods are described and several nonlinear Markov kernels and self-interacting approximations are introduced. In Section 4 we introduce some assumptions and some preliminary results, which are used to prove a strong law of large numbers (SLLN). In Sections 5 and 6, some technical proofs and the SLLN are presented; this is for a particular nonlinear kernel introduced in Section 3. This analysis is of interest from a theoretical point of view: it brings together the literature of measure-valued processes and interacting particle systems [10] used in SMC and the relatively recent literature on general state space Markov chains [25] used in MCMC. In Section 7 some algorithms are investigated; our assumptions are verified and some parameter settings are investigated for a toy example. In Section 8 some extensions to our ideas are discussed. The proofs are all given in the Appendices.

## 2. Notation and definitions

### 2.1. Notation

#### 2.1.1. *Probability and measure*

Define a measurable space $(E, \mathcal{E})$. Throughout, $\mathcal{E}$ will be assumed countably generated. $\mathscr{B}(\mathbb{R}^k)$, $k \in \mathbb{N}$ is used to represent the Borel sets with Lebesgue measure denoted by $\mathrm{d}x$.

For a stochastic process $\{X_n\}_{n\geq 0}$ on $(E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}})$, $\mathcal{G}_n^X = \sigma(X_0, \dots, X_n)$ denotes the natural filtration. $\mathbb{P}_\mu$ is taken as a probability law of a stochastic process with initial distribution $\mu$ and $\mathbb{E}_\mu$ the associated expectation. If $\mu = \delta_x$, with $\delta$ the Dirac measure, $\mathbb{P}_x$ (resp., $\mathbb{E}_x$) is used instead of $\mathbb{P}_{\delta_x}$ (resp., $\mathbb{E}_{\delta_x}$). For $\mu \in \mathscr{P}(E)$, the product measure is written $\mu \times \mu = \mu^{\otimes 2}$, with a clear generalization to higher order products. For measurable $f : E \to \mathbb{R}$, $\mu(f) = \int_E f(x)\mu(\mathrm{d}x)$.

If a $\sigma$-finite measure $\pi$ is dominated by another $\eta$ (denoted $\pi \ll \eta$), the Radon–Nikodym derivative is written with the same notation (e.g., if $\pi \ll \eta$, then $\pi(x)/\eta(x) = \mathrm{d}\pi/\mathrm{d}\eta(x)$). For $\sigma$-finite measures $\pi$ and $\eta$, $\pi \sim \eta$ denotes mutual absolute continuity.

#### 2.1.2. *Markov chains*

Let $(E, \mathcal{E})$ be a measurable space. Throughout for a Markov transition kernel $K : E \to \mathscr{P}(E)$ the following standard notation is used: for measurable $f : E \to \mathbb{R}$, $K(f)(x) := \int_E f(y)K(x, \mathrm{d}y)$ and for $\mu \in \mathscr{P}(E)$ $\mu K(f) := \int_E K(f)(x)\mu(\mathrm{d}x)$.

For $K_\mu$, $K : E \times \mathcal{P}(E) \to \mathcal{P}(E)$, given its existence, we will denote by $\omega(\mu)$ ($\omega : \mathcal{P}(E) \to \mathcal{P}(E)$) the invariant distribution of this Markov kernel. Recall that the empirical measure of an arbitrary stochastic process $(E^{\mathbb{N}}, \mathcal{E}^{\mathbb{N}}, \{X_n\}_{n\geq 0}, \mathbb{P})$ is defined, at time $n$, as

$$S_n^X(\mathrm{d}u) := \frac{1}{n+1}\sum_{i=0}^{n}\delta_{X_i}(\mathrm{d}u). \tag{2.1}$$

Throughout this paper, we are concerned with two nonlinear kernels of the form

$$K_\mu(x, \mathrm{d}y) = (1-\epsilon)K(x, \mathrm{d}y) + \epsilon\Phi(\mu)(\mathrm{d}y),$$
$$\Phi(\mu)(f) = \int_E \frac{g(y)f(y)}{\mu(g)}\mu(\mathrm{d}y),$$

where $K : E \to \mathcal{P}(E)$, $F : E \times \mathcal{P}(E) \to \mathcal{P}(E)$ (see [10] for more on $\Phi$) and

$$\begin{aligned} K_\mu(x, \mathrm{d}y) &= (1-\epsilon)K(x, \mathrm{d}y) + \epsilon Q_\mu(x, \mathrm{d}y),\\ Q_\mu(f)(x) &= \int_E \mu(\mathrm{d}u)\alpha(x, u)[f(u) - f(x)] + f(x), \end{aligned} \tag{2.2}$$

where $\alpha(x, u)$ is defined later on.

### 2.1.3. *Norms*

For any $k \in \mathbb{N}$, the Euclidean norm of $x \in \mathbb{R}^k$ is denoted $|x|$. For $f : E \to \mathbb{R}^n$, $n \in \mathbb{N}$, $|f|_\infty := \sup_{x\in E}|f(x)|$. For $f : E \to \mathbb{R}^n$ the $\mathbb{L}_p$-norm is defined, assuming it exists, as $(\int_E |f(x)|^p \,\mathrm{d}\mu)^{1/p}$ for $\mu \in \mathcal{P}(E)$. For $V : E \to [1, \infty)$ and $f : E \to \mathbb{R}^n$

$$|f|_V := \sup_{x\in E}\frac{|f(x)|}{V(x)}.$$

$\mathcal{L}_V$ is the class of functions $f : E \to \mathbb{R}^n$ such that $|f|_V < \infty$. We also use the notions of the $V$-total variation for a signed measure

$$\|\lambda\|_V := \sup_{|f|\leq V}|\lambda(f)|,$$

and the $V$-norm operator between two kernels $K_1, K_2 : E \to \mathcal{P}(E)$

$$|||K_1 - K_2|||_V := \sup_{x\in E}\frac{\|K_1(x,\cdot) - K_2(x,\cdot)\|_V}{V(x)}.$$

### 2.1.4. *Miscellaneous*

The notation $a \vee b := \max\{a, b\}$ (resp., $a \wedge b := \min\{a, b\}$) is adopted. The indicator function of $A \subset E$ is written $\mathbb{I}_A(x)$. $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Throughout the paper we denote a generic finite constant as $M$, that is, the value of $M$ may change from line to line in the proofs and is local to each proof.

# 3. Nonlinear MCMC

## 3.1. Nonlinear Markov kernels

Nonlinear MCMC can be characterised by the following procedure:

- Identify a nonlinear kernel that admits $\pi$ as an invariant distribution and can be expected to mix faster than an ordinary MCMC kernel; for example, (1.2).
- Construct a stochastic process that approximates the kernel, which can be simulated in practice.

Based upon the previous work [3], we consider auxiliary stochastic processes to approximate the nonlinear kernel. That is, it has been found in [3] that using the past history to approximate the nonlinear kernel leads to very poor performance. All of the processes that are simulated in this paper use an auxiliary Markov chain to approximate the nonlinear kernel. The difficulty is then to design sensible nonlinear kernels that may lead to good empirical performance. The two kernels we have designed are below.

## 3.2. Selection/mutation with potential

Let $P$ be an MCMC kernel of invariant distribution $\eta$, and assume $\pi \ll \eta$. Let $g(v) = \frac{\pi(v)}{\eta(v)}$ and set $K$ to be an MCMC kernel of invariant distribution $\pi$. Consider the nonlinear kernel

$$K_\mu(x, \mathrm{d}x') = (1-\epsilon)K(x, \mathrm{d}x') + \epsilon\Phi(\mu)(\mathrm{d}x');$$

clearly, if $\mu = \eta$, then one has $\pi K_\eta = \pi$.

If it is possible to sample exactly from $\eta$, then one could sample exactly from $K_\eta$. However, for efficient algorithms, this will not be the case. The following approximation is adopted at time-step $n+1$ of the simulation:

$$[(1-\epsilon)K(x_n, \mathrm{d}x_{n+1}) + \epsilon\Phi(S_n^Y)(\mathrm{d}x_{n+1})]P(y_n, \mathrm{d}y_{n+1});$$

that is, we are 'feeding' the chain $\{X_n\}_{n\geq 0}$ the empirical measure $S_n^Y$. Intuitively, as $n$ grows large, $S_n^Y(f) \to \eta(f)$ and one samples from the original kernel of interest.

## 3.3. Auxiliary self-interaction with genetic moves

For any $\mu \in \mathcal{P}(E)$ we define a nonlinear Markov kernel $Q_\mu : \mathcal{P}(E) \times E \to \mathcal{P}(E)$

$$Q_\mu(f)(x) = \int_E \mu(\mathrm{d}u)\alpha(x,u)[f(u) - f(x)] + f(x)$$

and for $\pi \sim \eta$

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)\eta(x)}{\pi(x)\eta(y)}.$$

The idea here is to generate a sample from $\mu$ and accept or reject it as the new state on the basis of the probability $\alpha$. Clearly, $\pi Q_\eta = \pi$. Letting $K$ and $P$ be as above, the process is simulated according to

$$\{(1-\epsilon)K(x_n, \mathrm{d}x_{n+1}) + \epsilon Q_{S_n^Y}(x_n, \mathrm{d}x_{n+1})\}P(y_n, \mathrm{d}y_{n+1})$$

at time $n+1$.

### 3.4. Some comments

In the example in Section 3.2 we attempt to use some measure of information, through $g$, to assist the resampling. The example of Section 3.3 provides a way to control the information that is provided by the approximation $S_n^Y$. That is, the kernel $Q_{S_n^Y}$, via $\alpha$ and the possible rejection, will provide a criterion to check the consistency with the target of the value drawn from $S_n^Y$. This may help improve estimation, if $S_n^Y$ converges slowly. Note that the algorithm is related to, but less sophisticated than, that of [23]. This is because we do not consider exchanges to occur between states in equi-energy rings.

It should be remarked that similar kernels are investigated in [7]. The author deduces that for a toy example it is hard to justify the use of such adaptive methods. However, a potential criticism of that study is that it is for a unimodal target; ‘advanced’ methods are seldom necessary for such scenarios. This is discussed further in Section 7.3.

### 3.5. Algorithm

The algorithm is (with the appropriate $\Phi(\mu)$ or $Q_\mu$):

0. (Initialization): Set $n=0$ and $X_0 = x$, $Y_0 = y$, $S_0^Y = \delta_y$.
1. (Iteration): Set $n = n+1$, simulate $Y_n \sim P(Y_{n-1}, \cdot)$ and $X_n \sim K_{S_{n-1}^Y}(X_{n-1}, \cdot)$.
2. (Update): $S_n^Y = S_{n-1}^Y + \frac{1}{n+1}[\delta_{Y_n} - S_{n-1}^Y]$ and return to 1.

## 4. Assumptions

We now seek to prove an SLLN for the nonlinear MCMC algorithm described in Section 3.3. Recall that we simulate a stochastic process on $((E \times E)^{\mathbb{N}}, (\mathcal{E} \otimes \mathcal{E})^{\otimes\mathbb{N}}, \{X_n, Y_n\}_{n\geq 0}, \{\mathcal{G}_n\}_{n\geq 0}, \mathbb{P}_{(x,y)})$, $(x, y) \in E \times E$, with finite-dimensional law:

$$\mathbb{P}_{(x,y),n}(d(x_0, y_0, \dots, x_n, y_n)) = \delta_{(x,y)}(d(x_0, y_0)) \prod_{i=0}^{n-1} K_{S_i^y}(x_i, \mathrm{d}x_{i+1})P(y_i, \mathrm{d}y_{i+1}).$$

Note that the natural filtration is denoted as $\mathcal{G}_n = \mathcal{G}_n^{X,Y}$ for notational simplicity. Since $\{Y_n\}$ is generated independently of $\{X_n\}$, we denote the probability law of the Markov chain $\{Y_n\}$ as $\mathbb{Q}_y$. Note, again, that the proofs are given in the Appendices.

## 4.1. Assumptions

Our assumptions on $K$, used to define our process, are now given. For $\bar{M} \in \mathbb{R}_+$, the notation $\mathcal{P}_{\bar{M}}(E) = \{\mu \in \mathcal{P}(E) : \mu(V) < \bar{M}\}$ is adopted, with $V$ defined below. In the remainder of the paper we say that a set $C \subset E$ is $(1, \theta)$-small if it satisfies a 1-step minorization condition, with parameter $\theta \in (0, 1)$.

(A1) Stability of $K$.

(i) (*Invariance and irreducibility*). $K : E \to \mathcal{P}(E)$ is a $\pi$-invariant and $\phi$-irreducible Markov kernel.

(ii) (*One-step minorization on level sets*). Define $C_d := \{x \in E : V(x) \leq d\}$ for any $d \in (1, \infty)$. We assume that for any $d \geq 1$, $C_d$ is $(1, \theta_d)$-small for some $\theta_d \in (0, 1)$ and $\nu_d \in \mathcal{P}(E)$.

(iii) (*One-step drift condition*). There exist $V : E \to [1, \infty)$ such that $\lim_{|x|\to\infty} V(x) = \infty$, $\lambda < 1$, $b < \infty$, $C \in \mathcal{E}$ such that for any $x \in E$

$$KV(x) \leq \lambda V(x) + b\mathbb{I}_C(x).$$

(A2) *Stability of $P$.*

(i) ($W$*-uniform ergodicity*). $P : E \to \mathcal{P}(E)$ is an $\eta$-invariant Markov kernel. Furthermore, there exists $W : E \to [1, \infty)$ such that $P$ is a $W$-uniformly ergodic Markov transition kernel with a one-step drift condition and one-step minorization condition. In addition, there exists an $r^* \in (0, 1]$ such that $V \in \mathcal{L}_{W^{r^*}}$ (where $V : E \to [1, \infty)$ is defined in (A1)(iii)).

(A3) *State-space constraint*

$$(E, \mathcal{E}) \text{ is Polish.}$$

## 4.2. Discussion of the assumptions

Our proofs of the SLLN will rely upon a martingale approximation via the solution of the Poisson equation (e.g., [17]). For any $\bar{M} < \infty$, (A1) will allow us to establish a drift condition for the kernel $K_\mu$ that is uniform in $\mu \in \mathcal{P}_{\bar{M}}(E)$; see [5]. In turn, one can establish: the existence of a solution to Poisson's equation, the existence of an invariant measure $\omega(\mu)$ for $K_\mu$ and regularity properties uniform in $\mu \in \mathcal{P}_{\bar{M}}(E)$. Then, due to (A2), the following facts are exploited: $\{S_n^Y(V)\}$ is $\mathbb{Q}_y$-a.s. finite and given $\{S_n^Y(V)\}$, $\{X_n\}$ is a Markov chain. (A1) and (A.2) appear quite strong, but can be verified in some important cases such as for random walk Metropolis kernels; see [21], for example.

A key result, relying on both (A2) and (A3), which is of interest in itself, is that of the $\mathbb{Q}_y$-a.s. convergence of $V$-statistics of $\{Y_i\}$. This result will enable us to show that, $\mathbb{Q}_y$-a.s., $\omega(S_i^Y) \to \omega(\eta)$; this is needed for our proof.

# 5. Common properties of $K_\mu$

Using standard drift and minorization conditions, the existence of an invariant probability measure is established for any $\mu \in \mathcal{P}_\infty(E)$ under (A1).

**Proposition 5.1.** *Assume* (A1). *Let* $\epsilon \in (0,1)$ *as in* (2.2), $\bar{M} \in (0,\infty)$, *then for* $d > \epsilon\bar{M}/[(1-\epsilon)(1-\lambda)]$ *with* $\lambda$ *and* $b$ *as in* (A1)(iii):

1. *There exist* $(\theta'_d, \nu_d) \in (0,1) \times \mathcal{P}(E)$ *such that for any* $\mu \in \mathcal{P}_{\bar{M}}(E)$ *and* $(x,A) \in E \times \mathcal{E}$:

$$K_\mu(x,A) \geq \mathbb{I}_{C_d}(x)\theta'_d \nu_d(A),$$

$$K_\mu V(x) \leq \tilde{\lambda} V(x) + \tilde{b}\mathbb{I}_{C_d}(x)$$

*with* $\tilde{\lambda} = (1-\epsilon)\lambda + \epsilon + \frac{\epsilon\bar{M}}{d} < 1$, $\tilde{b} = (1-\epsilon)[\lambda d + b] + \epsilon[\bar{M} + d]$.

2. *There exists a function* $\omega : \mathcal{P}_\infty(E) \to \mathcal{P}_\infty(E)$, *such that for any* $\mu \in \mathcal{P}_\infty(E)$

$$\omega(\mu) = \omega(\mu)K_\mu.$$

3. *There exist constants,* $\rho \in (0,1)$ *and* $M < \infty$ *depending upon* $\bar{M}$, $\epsilon$, $\lambda$, $b$, $V$, $d$, $\theta_d$ *(as defined in equation* (2.2) *and* (A1)), *such that for any* $\mu \in \mathcal{P}_{\bar{M}}(E)$, $r \in (0,1]$ *and* $f \in \mathcal{L}_{V^r}$

$$|K_\mu^n(f) - \omega(\mu)(f)|_{V^r} \leq M|f|_{V^r}\rho^n.$$

Some continuity properties associated with the invariant measures are as follows.

**Proposition 5.2.** *Assume* (A1) *and let* $\bar{M} \in (0,\infty)$. *Then there exists* $M < \infty$ *(depending solely on* $\bar{M}$ *and the constants in* (A1)) *such that for any* $r \in (0,1]$, $\mu, \xi \in \mathcal{P}_{\bar{M}}(E)$,

$$\|\omega(\xi) - \omega(\mu)\|_{V^r} \leq M |||K_\xi - K_\mu|||_{V^r}.$$

Noting that for any $\mu, \xi \in \mathcal{P}(E)$ and $r \in [0,1]$, $|||K_\xi - K_\mu|||_{V^r} = \epsilon |||Q_\xi - Q_\mu|||_{V^r}$ we establish global Lipschitz continuity results for $\mu \mapsto Q_\mu$, which, together with the result above, will allow us to deduce uniform Lipschitz continuity of $\mu \to K_\mu$ on $\mathcal{P}_{\bar{M}}(E)$ for any $\bar{M} \in (0,\infty)$. This is to be used in the proofs of many of the subsequent results.

**Proposition 5.3.** *Let* $\mu, \xi \in \mathcal{P}_\infty(E)$, *then for any* $r \in (0,1]$:

$$|||Q_\mu - Q_\xi|||_{V^r} \leq 2\|\mu - \xi\|_{V^r}.$$

# 6. Law of large numbers

## 6.1. Main result

Our main result is the following SLLN.

**Theorem 6.1.** *Assume* (A1)–(A3). *Let* $r \in [0, 1)$. *Then for any* $f \in \mathcal{L}_{V^r}$, $(x, y) \in E \times E$

$$S_n^X(f) \xrightarrow{a.s.}_{\mathbb{P}_{(x,y)}} \pi(f).$$

The proof is detailed in Appendix B, but we outline its main steps below.

## 6.2. Strategy of the proof

The strategy of the proof is now outlined. Introduce the following sequence of probability distributions $\{S_n^\omega := 1/(n+1)\sum_{i=0}^n \omega(S_i^Y)\}_{n\ge 0}$, where $\omega(\mu)$ is the invariant measure of $K_\mu$ (which, if $\mu = S_m^Y$, exists $\mathbb{Q}_y$-a.s.). This distribution can be used as a re-centering term in the following decomposition,

$$S_n^X(f) - \pi(f) = S_n^X(f) - S_n^\omega(f) + S_n^\omega(f) - \pi(f). \tag{6.1}$$

Let $\mu \in \{S_n^Y(f)\}$ and assume, for now, the almost sure existence of a solution $\hat{f}_\mu$ to Poisson's equation, that is, such that for any $x \in E$

$$f(x) - \omega(\mu)(f) = \hat{f}_\mu(x) - K_\mu(\hat{f}_\mu)(x).$$

Then, the first term on the right-hand side of (6.1) can be rewritten as

$$\begin{aligned}(n+1)[S_n^X - S_n^\omega](f) = M_{n+1} + \sum_{m=0}^{n}[\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})] \\ + \hat{f}_{S_0^Y}(X_0) - \hat{f}_{S_{n+1}^Y}(X_{n+1}),\end{aligned} \tag{6.2}$$

where

$$M_n = \sum_{m=0}^{n-1}[\hat{f}_{S_m^Y}(X_{m+1}) - K_{S_m^Y}(\hat{f}_{S_m^Y})(X_m)]$$

is such that $\{M_n, \mathcal{G}_n^X\}$ will be a martingale conditional upon $\mathcal{G}_\infty^Y$. In addition, critical to our analysis, will be that, $\mathbb{Q}^Y$-a.s., $\{S_n^Y(V)\}$ is finite. This latter fact will enable us to control the various terms in (6.2) on events of the type $\{\sup_{k\ge 0} S_k^Y(V) \le \bar{M}\}$ for $\bar{M} > 0$. This is now elaborated.

## 6.3. $\{M_m\}$ is $\mathbb{L}_p$-bounded

One can establish the following uniform in time $\mathbb{L}_p$-bounds of the solution to Poisson's equation and the sequence $\{M_n\}$, restricted to events $\{\sup_{k\ge 0} S_k^Y(V) \le \bar{M}\}$ for any $\bar{M} > 0$.

**Proposition 6.2.** *Assume* (A1). *Let* $r \in [0, 1]$, $p \in [1, 1/r]$ *for* $r \neq 0$ *and* $p \ge 1$ *otherwise and* $\bar{M} \in (0, \infty)$. *Then there exists* $M < \infty$ *such that for any* $f \in \mathcal{L}_{V^r}$, $(x, y) \in E \times E$ *and any* $m \in \mathbb{N}_0$,

$$\mathbb{E}_{(x,y)}\big[|\hat{f}_{S_m^Y}(X_{m+1})|^p \mathbb{I}_{\{\sup_{k\ge 0} S_k^Y(V) \le \bar{M}\}}\big]^{1/p} \le M V(x)^r.$$

**Proposition 6.3.** *Assume* (A1). *Let* $r \in [0, 1]$, $p \in [1, 1/r]$ *for* $r \neq 0$ *and* $p \geq 1$ *otherwise and* $\bar{M} \in (0, \infty)$. *Then there exists* $M < \infty$ *such that for any* $f \in \mathcal{L}_{V^r}$, $(x, y) \in E \times E$ *and any* $m \in \mathbb{N}_0$,

$$\mathbb{E}_{(x,y)}\big[|M_m|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \leq m^{1/2\vee 1/p} M V(x)^r.$$

This result will allow us to prove the $\mathbb{P}_{(x,y)}$-a.s. convergence of $M_n$ to zero (cf. Appendix B).

### 6.4. Smoothness of the solution to Poisson's equation and $\omega(S_n^Y)$

As can be observed in (6.2), we have to control the fluctuations of the solution of the Poisson equation $\{\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})\}$. Also, in (6.1), the convergence of $\omega(S_m^Y)(f)$ to $\omega(\eta)(f)$ $\mathbb{Q}_y$-a.s. must be established. Both of these issues are now dealt with.

**Proposition 6.4.** *Assume* (A1) *and* (A2). *Let* $r \in [0, 1)$, *then for any* $f \in \mathcal{L}_{V^r}$, $(x, y) \in E \times E$

$$\lim_{m\to\infty} |\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})| = 0 \qquad \mathbb{P}_{(x,y)}\text{-a.s.}$$

**Proposition 6.5.** *Assume* (A1)–(A3). *Let* $f \in \mathcal{L}_V$ *and* $(x, y) \in E \times E$, *then*

$$\lim_{m\to\infty} \omega(S_m^Y)(f) = \omega(\eta)(f) \qquad \mathbb{Q}_y\text{-a.s.}$$

## 7. Examples

In this section we present some applications of our algorithms. Specifically, it is demonstrated that the assumptions hold in some very general scenarios. In addition, a numerical investigation of our approach for a toy problem is given.

### 7.1. Verifying the assumptions

It is now shown that it is possible to verify the assumptions in Section 4.1 in quite general scenarios. Let us concentrate upon the case where, for $k \geq 1$, $(E, \mathcal{E}) = (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ and $K$ (resp., $P$ – recall the invariant measure is $\eta$) is a symmetric random walk Metropolis kernel:

$$K(x, dx') = \alpha_\pi(x, x')q_\pi(x - x')\, dx' + \delta_x(dx')\left\{1 - \int_{\mathbb{R}^k} \alpha_\pi(x, x')q_\pi(x - x')\, dx'\right\}, \qquad (7.1)$$

where (resp., $P$)

$$\alpha_\pi(x, x') = 1 \wedge \frac{\pi(x')}{\pi(x)}$$

and $q_\pi$ (resp. $q_\eta$) is a symmetric density (w.r.t. Lebesgue measure).

7.1.1. *Assumptions*

A set of general conditions is introduced, such that the assumptions in Section 4.1 will hold.

(M1) *Density* $\pi$.

- $\pi$ admits a positive and continuous density w.r.t. Lebesgue measure.

(M2) *Definition of* $\eta$.

- $\eta(x) \propto \pi(x)^{\tilde{\alpha}}$, with $\tilde{\alpha} \in (0, 1)$.

(M3) *Boundedness.*

- $\pi$ is upper bounded and bounded away from 0 on compact sets.

(M4) *Super-exponential densities.*

- $\pi$ is super exponential:

$$\lim_{|x|\to+\infty} \frac{x}{|x|} \cdot \nabla \log(\pi(x)) = -\infty.$$

(M5) *Regularity of contours.*

- The contours of $\pi$ are asymptotically regular:

$$\limsup_{|x|\to+\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0.$$

(M6) *Lower bounds on* $q_\pi$, $q_\eta$.

- Both $q_\pi$ and $q_\eta$ are such that there exists $\tilde{\delta}_{q_\pi} > 0$ (resp., $\tilde{\delta}_{q_\eta} > 0$) and $\epsilon_{q_\pi} > 0$ (resp., $\epsilon_{q_\eta} > 0$) such that

$$q_\pi(x) \geq \epsilon_{q_\pi} \qquad \text{for } |x| < \tilde{\delta}_{q_\pi}$$

(resp., $q_\eta(x) \geq \epsilon_{q_\eta}$ for $|x| < \tilde{\delta}_{q_\eta}$).

## 7.2. Result

**Proposition 7.1.** *Assume* (M1)–(M6), *then* (A1)–(A3) *hold for any* $r^* \in (0, 1)$ *with*

$$W(x) = \left[\frac{|\pi|_\infty}{\pi(x)}\right]^{\tilde{\alpha} s_w}, \qquad s_w \in (0, 1),$$

$$V(x) = \left[\frac{|\pi|_\infty}{\pi(x)}\right]^{s_v}, \qquad s_v \in (0, r^*\tilde{\alpha} s_w).$$

The proof is in Appendix F.

7.2.1. *Some comments*

The conditions presented above are quite general. For example, they are satisfied if $\pi$ is a mixture of normals. More generally, it may be difficult to check the assumptions, but this is due to the underlying nature of the geometric ergodicity assumptions; see [21] for more thorough investigations.

## 7.3. Toy example

Our target distribution is

$$\pi(x) = 0.4\psi(x; 0, 0.5) + 0.6\psi(x; 17.5, 1)$$

with $\psi(x; \mu, \sigma^2)$ the normal density of mean $\mu$ and variance $\sigma^2$.

Our algorithms are run with $K$ as a random walk Metropolis kernel with normal random walk proposal density. The kernel is iterated 500 times (i.e., $K = \widetilde{K}$ with $\widetilde{K}$ as a random walk Metropolis kernel); this is to reduce the amount of interaction, especially for large $\epsilon$. $\eta$ was taken to be:

$$\eta(x) \propto \pi(x)^{0.75}.$$

The algorithms were run for the same CPU time and the results can be found, for 50 runs of the algorithm, in Table 1. The assumptions (M1)–(M6) are satisfied here.

In Table 1, the algorithms in Sections 3.2 and 3.3 both perform reasonably well for small values of $\epsilon$. As expected, from the assumptions, as $\epsilon$ gets larger the accuracy falls. This is due to the fact that the amount of auxiliary information that can enter into the $\{X_n\}$ process is increased. For small $\epsilon$, the example in Section 3.3 appears to work better (more accurate estimation) due to the more sophisticated interaction with the auxiliary chain. The drastic poor performance for the kernel in Section 3.3, for large $\epsilon$, is due to the fact that no transition occurs after the swapping move.

To compare to the results of [7], we ran a random walk algorithm for 1 million iterations 50 times and a nonlinear algorithm (Section 3.3). The nonlinear algorithm was run with $\epsilon = 0.01$ but the random walk kernel was not iterated. The auxiliary chain was run with $\tilde{\alpha} = 0.75$ (as in (M2)). This was run for 110 000 iterations 50 times (which is approximately the same CPU time as for the random walk Metropolis algorithm). Both algorithms are such that all initial values

**Table 1.** Estimates from mixture comparison for nonlinear MCMC. The estimates are for the expectation of $X$; the true value is 10.5. Each algorithm is run 50 times for 2 million iterations after a 50 000 iteration burn-in (Section 3.3; the simulations for Section 3.2 are adjusted for the appropriate CPU time). The brackets are $\pm 2$ standard deviations across the repeats

| Example | $\epsilon = 0.05$ | $\epsilon = 0.25$ | $\epsilon = 0.5$ | $\epsilon = 0.75$ | $\epsilon = 0.95$ |
|---|---|---|---|---|---|
| Section 3.2 | 10.32 ($\pm$0.08) | 10.74 ($\pm$0.12) | 10.89 ($\pm$0.19) | 10.37 ($\pm$0.18) | 10.99 ($\pm$0.20) |
| Section 3.3 | 10.57 ($\pm$0.04) | 10.52 ($\pm$0.09) | 10.96 ($\pm$0.7) | 10.02 ($\pm$0.93) | 11.08 ($\pm$1.20) |

are drawn from a uniform on [0, 10.5]. The estimated value for the first moment is $6.93 \pm 16.96$ ($\pm 2$ standard deviations, across the 50 runs) and $10.41 \pm 2.03$ for the random walk and nonlinear methods, respectively. The random walk algorithm is unable to jump between the modes of the target, while the auxiliary chain is able to do so; hence justifying our earlier intuition. This slightly contradicts the 'cautionary tale' in [7] as it illustrates that such algorithms are potentially useful in cases where random walk algorithms do not work well. We remark however, that one must be careful with allowing too much auxiliary information to enter the chain $\{X_n\}_{n \geq 0}$; this can lead to poor results. This is consistent with Proposition 5.1, which indicates that $d$ grows as $\epsilon$ goes to 1.

# 8. Summary

We have investigated a new approach to stochastic simulation: Nonlinear MCMC via auxiliary/self-interacting approximations. Convergence results for several algorithms were established and the algorithm was demonstrated on a toy example. As extensions to our ideas, the following may be considered.

First, the conditions required for convergence may be relaxed. For example, [17] establishes weaker-than-geometric ergodicity assumptions for the solution to the Poisson equation and functional central limit theorem; also, [15] establishes drift conditions for polynomial ergodicity. It would be of interest to see whether such conditions would be sufficient for the convergence of our algorithms; see [28] for proofs for parametric adaptive MCMC.

Second, it would be interesting to design more elaborate methods to control the evolution of the empirical measure. In our current algorithms, the empirical measure is only updated through the addition of simulated points. It may enhance the algorithm to introduce some mechanisms allowing the improvement of this quantity; for example, we could introduce a death process with a rate associated with the un-normalized target distribution.

# Appendix A: Common properties of $K_\mu$

**Proof of Proposition 5.1.** The second and third statement of the proposition are a direct consequence of the first point from [24], Theorem 2.3 (note the $\phi$-irreducibility and aperiodicity follow immediately). The minorization property is direct from the expression for $K_\mu$ and (A1)(ii) with $\theta'_d = (1-\epsilon) \times \theta_d$. Let us focus on the drift condition.

For any $x \in E$, $\mu \in \mathcal{P}_{\bar{M}}(E)$:

$$K_\mu(V)(x) \leq (1-\epsilon)[\lambda V(x) + b\mathbb{I}_{C_d}(x)] + \epsilon[\mu(V) + V(x)\varphi(x)],$$

where $\varphi(x) = 1 - \int_E \alpha(x, y)\mu(\mathrm{d}y)$. Then as $\mu(V) < \bar{M}$, one has

$$K_\mu(V)(x) \leq (1-\epsilon)[\lambda V(x) + b\mathbb{I}_{C_d}(x)] + \epsilon[\bar{M} + V(x)].$$

Let $x \in C_d^c$, then

$$K_\mu(V)(x) \leq \left[(1-\epsilon)\lambda + \epsilon + \frac{\epsilon \bar{M}}{d}\right] V(x) = \tilde{\lambda} V(x).$$

For $x \in C_d$

$$K_\mu(V)(x) \leq (1-\epsilon)[\lambda d + b] + \epsilon[\bar{M} + d]$$

and hence one concludes that

$$K_\mu(V)(x) \leq \tilde{\lambda} V(x) + \tilde{b}\mathbb{I}_{C_d}(x).$$ □

**Proof of Proposition 5.2.** This is a direct application of Proposition 5.1 and Lemma C.1. □

**Proof of Proposition 5.3.** The proof is given for $r = 1$ only. Let $|f| \leq V$:

$$|[Q_\mu - Q_\xi](f)(x)| = \left| \int_E [\mu - \xi](\mathrm{d}u)[\alpha(x,u)\{f(u) - f(x)\}] \right|.$$

Now it is clear that, for any fixed $x \in E$:

$$|\alpha(x,u)\{f(u) - f(x)\}| \leq [V(u) + V(x)],$$

i.e.,

$$|\alpha(x,u)\{f(u) - f(x)\}| \leq 2V(u)V(x).$$

Thus

$$|[Q_\mu - Q_\xi](f)(x)| \leq 2V(x)\|\mu - \xi\|_V$$

and then the result easily follows. □

# Appendix B: Proof of the main result

**Proof of Theorem 6.1.** Let $r \in [0, 1)$ and $f \in \mathcal{L}_{V^r}$. Recall the strategy of the proof outlined in Section 6.2, which relies on the decomposition:

$$S_n^X(f) - \pi(f) = S_n^X(f) - S_n^\omega(f) + S_n^\omega(f) - \pi(f) \tag{B.1}$$

with

$$\begin{aligned}(n+1)&[S_n^X - S_n^\omega](f)\\ &= M_{n+1} + \sum_{m=0}^{n} [\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})] + \hat{f}_{S_0^Y}(X_0) - \hat{f}_{S_{n+1}^Y}(X_{n+1}),\end{aligned}$$

where $\{M_n\}$ is a martingale conditional upon $\mathcal{G}_\infty^Y$. Proving the almost sure convergence of $[S_n^X - S_n^\omega](f)$ relies on classical arguments. For any $n \geq 1$, $\delta > 0$ and $\bar{M} \in (0, \infty)$,

$$\begin{aligned}
&\mathbb{P}_{(x,y)}\Big(\sup_{k\geq n} |[S_k^X - S_k^\omega](f)| > \delta\Big) \\
&\quad \leq \mathbb{P}_{(x,y)}\Big(\sup_{k\geq n} |M_{k+1}/(k+1)| > \delta/3, \sup_{k\geq 0} S_k^Y(V) < \bar{M}\Big) \\
&\qquad + \mathbb{P}_{(x,y)}\left(\sup_{k\geq n}\left|\sum_{m=0}^{k}[\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})]\right| \Big/ (k+1) > \delta/3, \sup_{k\geq 0} S_k^Y(V) < \bar{M}\right) \\
&\qquad + \mathbb{P}_{(x,y)}\Big(\sup_{k\geq n}[|\hat{f}_{S_0^Y}(X_0)| + |\hat{f}_{S_{k+1}^Y}(X_{k+1})|]/(k+1) > \delta/3, \sup_{k\geq 0} S_k^Y(V) < \bar{M}\Big) \\
&\qquad + \mathbb{Q}_y\Big(\sup_{k\geq 0} S_k^Y(V) \geq \bar{M}\Big).
\end{aligned}$$

Let $\varepsilon > 0$. By assumption there exists $\bar{M} > 0$ such that $\mathbb{Q}_y(\sup_{k\geq 0} S_k^Y(V) \geq \bar{M}) \leq \varepsilon/4$. Now we consider the remaining terms on the right-hand side of the above equation from bottom to top; it is proved that there exists $n_0 > 0$ such that for any $n \geq n_0$ each of these terms is less than $\varepsilon/4$. Let $p \in (1, 1/r)$. By Proposition 6.2, one can apply Markov's inequality and a Borel–Cantelli argument to show that the term on the third line vanishes as $n \to \infty$. By Proposition 6.4 and a Cesáro argument one concludes that the term on the second line goes to zero as $n \to \infty$. The term dependent on $\{M_n\}$ is dealt with by using an adaptation of a Birnbaum–Marshall inequality (see [5]) for $p \in (1, 1/r)$.

Controlling the bias term requires a more novel approach. Note that

$$|S_n^\omega(f) - \pi(f)| = \frac{1}{n+1}\left|\sum_{i=0}^{n}[\omega(S_i^Y) - \omega(\eta)](f)\right|,$$

as $\omega(\eta) = \pi$ in our setup. In Proposition 6.5 it is proved that under our assumptions $[\omega(S_i^Y) - \omega(\eta)](f) \to 0$ $\mathbb{Q}_y$-a.s. as $i \to \infty$. We conclude by invoking a Cesàro average argument. □

**Proof of Proposition 6.2.** Let $\bar{M} \in (0, \infty)$. The proof begins by conditioning upon the filtration $\mathcal{G}_m^Y$ generated by the auxiliary process $\{Y_n\}$; then, using the uniform in $\mu \in \mathscr{P}_{\bar{M}}(E)$, geometric ergodicity is proved in Proposition 5.1. As a result, there exists an $M < \infty$ such that

$$\begin{aligned}
&\mathbb{E}_{(x,y)}\big[|\hat{f}_{S_m^Y}(X_{m+1})|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \\
&\quad \leq M\mathbb{E}_{(x,y)}\big[|V(X_{m+1})^r|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \\
&\quad \leq MV^r(x),
\end{aligned}$$

where we have used Jensen and the uniform drift condition on the set $\{\sup_{k\geq 0} S_k^Y(V) \leq \bar{M}\}$ proved in Proposition 5.1 □

**Proof of Proposition 6.3.** We follow a similar argument to that of [5], Proposition 6. Throughout, denote by $B_p$ a generic constant dependent upon $p$ only. Also recall $pr \leq 1$. The proof begins by applying the Bürkholder–Davis inequality (see, e.g., [30], pages 499–500), which yields for $p \geq 1$

$$\begin{aligned}
&\mathbb{E}_{(x,y)}\big[|M_n|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \\
&\quad\leq B_p \mathbb{E}_y\left[\mathbb{E}_{(x,y)}\left[\left(\sum_{m=0}^{n-1}[\hat{f}_{S_m^Y}(X_{m+1}) - K_{S_m^Y}(\hat{f}_{S_m^Y})(X_m)]^2\right)^{p/2}\Big|\mathcal{G}_\infty^Y\right]\mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\right]^{1/p}.
\end{aligned}$$

In the case $p > 2$, by similar manipulations to those featured in [5]

$$\mathbb{E}_{(x,y)}\big[|M_n|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \leq n^{1/2} B_p M V(x)^r.$$

In the case $p \leq 2$, one may apply the $C_p$-inequality to yield

$$\begin{aligned}
&\mathbb{E}_{(x,y)}\big[|M_n|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \\
&\quad\leq \left[\sum_{m=0}^{n-1}\mathbb{E}_{(x,y)}\big[|\hat{f}_{S_m^Y}(X_{m+1}) - K_{S_m^Y}(\hat{f}_{S_m^Y})(X_m)|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]\right]^{1/p}.
\end{aligned}$$

Application of Minkowski, conditional Jensen and Proposition 6.2 yields

$$\mathbb{E}_{(x,y)}\big[|M_n|^p \mathbb{I}_{\{\sup_{k\geq 0} S_k^Y(V)\leq \bar{M}\}}\big]^{1/p} \leq n^{1/p} M V^r(x)$$

from which we can conclude. □

**Proof of Proposition 6.4.** Our proof is based upon the decomposition of Proposition C.2 (in Appendix C) and then using the Lipschitz continuity properties proved in Propositions 5.2 and 5.3. Let $\bar{M} \in (0, \infty)$ be given; suppose that we are on the set $\{\sup_{k\geq 0} S_k^Y(V) \leq \bar{M}\}$. Then

$$\begin{aligned}
&|\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})| \\
&\quad= \Bigg|\sum_{n\in\mathbb{N}}\sum_{i=0}^{n-1}[K_{S_{m+1}^Y}^i - \omega(S_{m+1}^Y)](K_{S_{m+1}^Y} - K_{S_m^Y})[K_{S_m^Y}^{n-i-1} - \omega(S_m^Y)(f)(X_{m+1})] \qquad \text{(B.2)} \\
&\qquad - \sum_{n\in\mathbb{N}}[\omega(S_{m+1}^Y) - \omega(S_m^Y)]\big(K_{S_m^Y}^n - \omega(S_m^Y)\big)(f)\Bigg|.
\end{aligned}$$

Now, consider the first term. Since, for any $m \geq 0$, the kernel $K_{S_m^Y}$ satisfies:

$$\|[K_{S_m^Y}^n - \omega(S_m^Y)](f)\|_{V^r} \leq M\rho^n V(X_{m+1})^r$$

for some finite $M$ and $\rho \in (0, 1)$ independent of $S_m^Y \in \mathcal{P}_{\bar{M}}(E)$, it follows that:

$$\begin{aligned}&|[K^i_{S^Y_{m+1}} - \omega(S^Y_{m+1})](K_{S^Y_{m+1}} - K_{S^Y_m})[K^{n-i-1}_{S^Y_m} - \omega(S^Y_m)(f)(X_{m+1})]|\\&\quad\le M\rho^i V(X_{m+1})^r |(K_{S^Y_{m+1}} - K_{S^Y_m})[K^{n-i-1}_{S^Y_m} - \omega(S^Y_m)(f)]|_{V^r}.\end{aligned}$$

Then, adopting the continuity result for $K_{S^Y_m}$:

$$|||K_\mu - K_\xi|||_{V^r} \le 2\|\mu - \xi\|_{V^r}$$

for *any* $\mu, \xi \in \mathcal{P}_\infty(E)$, it follows that:

$$|(K_{S^Y_{m+1}} - K_{S^Y_m})[K^{n-i-1}_{S^Y_m} - \omega(S^Y_m)(f)]|_{V^r} \le M\rho^{n-i-1}\|S^Y_{m+1} - S^Y_m\|_{V^r}.$$

Since $\|S^Y_{m+1} - S^Y_m\|_{V^r} \le [V(Y_{m+1})^r + S^Y_m(V^r)]/(m+2)$

$$\begin{aligned}&\sum_{n,i} |[K^i_{S^Y_{m+1}} - \omega(S^Y_{m+1})](K_{S^Y_{m+1}} - K_{S^Y_m})[K^{n-i-1}_{S^Y_m} - \omega(S^Y_m)(f)(X_{m+1})]|\\&\quad\le \frac{M}{(1-\rho)^2}\frac{V(X_{m+1})^r}{m+2}[V(Y_{m+1})^r + S^Y_m(V^r)].\end{aligned}$$

Turning to the second sum on the right-hand side of (B.2), using the continuity result

$$\|\omega(\mu) - \omega(\xi)\|_{V^r} \le M|||K_\mu - K_\xi|||_{V^r}$$

(for $M < \infty$ not depending on $\mu, \xi \in \mathcal{P}_{\bar{M}}(E)$ by Proposition 5.3) and the continuity of the kernel $K_\mu$ (Lemma C.1 ) yields,

$$\big|[\omega(S^Y_{m+1}) - \omega(S^Y_m)]\big(K^n_{S^Y_m} - \omega(S^Y_m)\big)(f)\big| \le M\rho^n \frac{[V(Y_{m+1})^r + S^Y_m(V^r)]}{m+2},$$

from which we obtain a similar bound for the second sum on the right-hand side of (B.2).

We now establish an $\mathbb{L}_p$-bound, for $p > 1$ of this upper bound on $\{\sup_{k\ge 0} S^Y_k(V) \le \bar{M}\}$, which will allow us to use a Borel–Cantelli argument to complete the proof. Note that it is naturally sufficient to consider $\frac{V(X_{m+1})^r}{m+2}V(Y_{m+1})^r$ on $\{\sup_{k\ge 0} S^Y_k(V) \le \bar{M}\}$, and we focus on

$$\begin{aligned}&\mathbb{E}_{(x,y)}\big[V(X_{m+1})^{rp}V(Y_{m+1})^{rp}\mathbb{I}_{\{\sup_{k\ge 0} S^Y_k(V)\le\bar{M}\}}\big]^{1/p}\\&\quad= \mathbb{E}_{(x,y)}\big[\mathbb{E}_{(x,y)}[V(X_{m+1})^{rp}|\mathcal{G}^Y_\infty]V(Y_{m+1})^{rp}\mathbb{I}_{\{\sup_{k\ge 0} S^Y_k(V)\le\bar{M}\}}\big]^{1/p}\\&\quad\le MV(x)^r\mathbb{E}_y\big[V(Y_{m+1})^{rp}\mathbb{I}_{\{\sup_{k\ge 0} S^Y_k(V)\le\bar{M}\}}\big]^{1/p} \le MV(x)^r W^{rr^*}(y),\end{aligned}\tag{B.3}$$

where we have used that, conditional upon $\mathcal{G}^Y_\infty$ and on the event $\{\sup_{k\ge 0} S^Y_k(V) \le \bar{M}\}$, the following bound holds $\mathbb{E}_{(x,y)}[V(X_{m+1})^{pr}|\mathcal{G}^Y_\infty]^{1/p} \le M_0 V(x)^r$ for some deterministic constant $M_0$

depending only on $\bar{M}$ and the parameters of the drift condition in (A1). Similarly, $M \geq M_0$ only depends on $\bar{M}$ and the parameters of the drift conditions in (A1)–(A2). With $p > 1$ we conclude that

$$\sum_{m=0}^{\infty} \mathbb{P}_{(x,y)}\left(\left\{\frac{1}{m+1} V(X_{m+1})^r [V(Y_{m+1})^r + S_m^Y(V^r)] > \varepsilon, \sup_{k\geq 0} S_k^Y(V) \leq \bar{M}\right\}\right) < \infty.$$

The result then follows by using for any $\delta > 0$ the bound,

$$\begin{aligned}
&\mathbb{P}_{(x,y)}\Big(\sup_{k\geq m} |\hat{f}_{S_{k+1}^Y}(X_{k+1}) - \hat{f}_{S_k^Y}(X_{k+1})| > \delta\Big) \\
&\quad \leq \mathbb{Q}_y\Big(\sup_{k\geq 0} S_k^Y(V) \geq \bar{M}\Big) + \mathbb{P}_{(x,y)}\Big(\sup_{k\geq m} |\hat{f}_{S_{k+1}^Y}(X_{k+1}) - \hat{f}_{S_k^Y}(X_{k+1})| > \delta, \sup_{k\geq 0} S_k^Y(V) \leq \bar{M}\Big)
\end{aligned}$$

and using the fact that for any $\varepsilon > 0$ one can find an $\bar{M}$ large enough to ensure that the first term on the right-hand side is less than $\varepsilon/2$ and then $m_0$ such that for any $m \geq m_0$ the second term on the right-hand side is also upper bounded by $\varepsilon/2$. □

**Proof of Proposition 6.5.** Note first that for any $i, j \in \mathbb{N}$, $f \in \mathcal{L}_V$ and $x \in E$ such that $V(x) < \infty$ we have the following bound,

$$\begin{aligned}
&|[\omega(S_i^Y) - \omega(\eta)](f)| \\
&\quad \leq |\omega(S_i^Y)(f) - K_{S_i^Y}^j(f)(x)| + |K_{S_i^Y}^j(f)(x) - K_\eta^j(f)(x)| + |K_\eta^j(f)(x) - \omega(\eta)(f)|.
\end{aligned}$$

Let $\varepsilon, \delta > 0$ and $\bar{M} > \eta(V)$ be such that $\mathbb{Q}_y(\sup_{k\geq 0} S_k^Y(V) \geq \bar{M}) < \varepsilon/4$. On the event $\{\sup_{k\geq 0} S_k^Y(V) < \bar{M}\}$ we have by Proposition 5.2 the existence of $M < +\infty$ and $\rho \in [0, 1)$ (independent of $i$) such that the first and last terms on the right-hand side are bounded by $M\rho^j$. We can therefore fix $m$ such that

$$\mathbb{Q}_y\Big(\sup_{j\geq m, i\geq 0} |\omega(S_i^Y)(f) - K_{S_i^Y}^j(f)(x)| + |K_\eta^j(f)(x) - \omega(\eta)(f)| > \delta/2, \sup_{k\geq 0} S_k^Y(V) < \bar{M}\Big) \leq \varepsilon/2.$$

Now from Lemma D.2 one may conclude that there exists $m_0 > 0$ such that for any $m \geq m_0$

$$\mathbb{Q}_y\Big(\sup_{i\geq m} |K_{S_i^Y}^j(f)(x) - K_\eta^j(f)(x)| > \delta/2, \sup_{k\geq 0} S_k^Y(V) < \bar{M}\Big) \leq \varepsilon/4.$$

The proof is completed by noting that the results above imply that for $m \geq m_0$,

$$\mathbb{Q}_y\Big(\sup_{i\geq m} |[\omega(S_i^Y) - \omega(\eta)](f)| > \delta\Big) \leq \varepsilon. \tag{B.4}$$

□

## Appendix C: Standard technical results on Markov chains

**Lemma C.1.** *Let $(E, \mathcal{E})$ be a measurable space, $\bar{b} < \infty$, $\bar{\lambda} \in (0, 1)$ and $\bar{C} \in \mathcal{E}$. Then for any Markov transition probabilities $P_1, P_2 : E \to \mathscr{P}(E)$ satisfying for $(x, A) \in E \times \mathcal{E}$ and $i = 1, 2$,*

$$P_i V(x) \leq \bar{\lambda} V(x) + \mathbb{I}_{\bar{C}}(x)\bar{b}, \tag{C.1}$$

$$P_i(x, A) \geq \mathbb{I}_{\bar{C}}(x)\bar{\epsilon}\bar{\nu}(A). \tag{C.2}$$

*There exist $\bar{M}(\cdot) < \infty$, $\bar{\rho} \in [0, 1)$, invariant probability measures $\pi_1, \pi_2 \in \mathscr{P}(E)$ (corresponding to $P_1$ and $P_2$, respectively), such that for any $n \geq 1$, $r \in [0, 1]$ and any $|f| \leq V^r$*

$$|[P_1^n - \pi_1](f)|_{V^r} \vee |[P_2^n - \pi_2](f)|_{V^r} \leq \bar{M}(r)\bar{\rho}^n$$

*for any $n \geq 1$,*

$$|||P_1^n - P_2^n|||_{V^r} \leq \bar{M}(r)|||P_1 - P_2|||_{V^r}$$

*and*

$$\|\pi_1 - \pi_2\|_{V^r} \leq \bar{M}(r)|||P_1 - P_2|||_{V^r}.$$

**Proof.** Let $r \in [0, 1]$ and $f \in \mathcal{L}_{V^r}$. We have the following decomposition:

$$|[P_1^n - P_2^n](f)| = \left|\sum_{i=0}^{n-1} P_1^i\bigl([P_1 - P_2]\{[P_2^{n-i-1} - \pi_2](f)\}\bigr)\right|.$$

For any $|f| \leq V^r$, in a similar manner to Proposition 3 of [5]:

$$\begin{aligned}
|[P_1^n - P_2^n](f)| &\leq \bar{M}(r)\sum_{i=0}^{n-1} \bar{\rho}^{n-i-1} P_1^i(\|P_1 - P_2\|_{V^r}) \\
&= \bar{M}\sum_{i=0}^{n-1} \bar{\rho}^{n-i-1} P_1^i\left(\frac{\|P_1 - P_2\|_{V^r}}{V^r} V^r\right) \\
&\leq \bar{M}(r)|||P_1 - P_2|||_{V^r}\sum_{i=0}^{n-1} \bar{\rho}^{n-i-1} P_1^i(V^r).
\end{aligned}$$

From the drift condition (A2) and conditional Jensen one can bound $P_1^i V^r$ by $[\bar{\lambda} + \bar{b}/(1 - \bar{\lambda})]^r V(x)^r$ for $r \in [0, 1]$ and hence conclude that:

$$|[P_1^n - P_2^n](f)| \leq \bar{M}(r)|||P_1 - P_2|||_V^r.$$

Since the right-hand side is independent of $n$, the inequality holds in the limit and hence, by $V$-uniform ergodicity, the result. □

**Proposition C.2.** *Assume* (A1). *Then, for* $r \in [0,1]$, $\xi, \mu \in \mathcal{P}_\infty(E)$, $f \in \mathcal{L}_{V^r}$ *we have the following decomposition for the differences in the solution to the Poisson equation*:

$$\hat{f}_\xi(x) - \hat{f}_\mu(x) = \sum_{n\in\mathbb{N}} \Biggl\{ \sum_{i=0}^{n-1} \bigl([K_\xi^i - \omega(\xi)](K_\xi - K_\mu)\{[K_\mu^{n-i-1} - \omega(\mu)](f)\}(x)\bigr) - [\omega(\xi) - \omega(\mu)]\bigl([K_\mu^n - \omega(\mu)](f)\bigr) \Biggr\}.$$

**Proof.** Adopting the resolvent solution to the Poisson equation (which exists under our assumptions), we have

$$\begin{aligned}
\hat{f}_\xi(x) - \hat{f}_\mu(x) &= \sum_{n\in\mathbb{N}_0} \bigl[\bigl([K_\xi^n - \omega(\xi)](f)(x)\bigr) - \bigl([K_\mu^n - \omega(\mu)](f)(x)\bigr)\bigr] \\
&= \sum_{n\in\mathbb{N}} \Biggl[ \sum_{i=0}^{n-1} K_\xi^i \bigl([K_\xi - K_\mu]\{[K_\mu^{n-i-1} - \omega(\mu)](f)\}\bigr)(x) + \omega(\mu)(f) - \omega(\xi)(f) \Biggr] \\
&= \sum_{n\in\mathbb{N}} \Biggl\{ \sum_{i=0}^{n-1} \bigl([K_\xi^i - \omega(\xi)](K_\xi - K_\mu)\{[K_\mu^{n-i-1} - \omega(\mu)](f)\}(x)\bigr) \\
&\qquad - [\omega(\xi) - \omega(\mu)]\bigl([K_\mu^n - \omega(\mu)](f)\bigr) \Biggr\}
\end{aligned}$$

since

$$-\sum_{i=0}^{n-1} \omega(\xi)[K_\xi - K_\mu](K_\mu^{n-i-1}(f)) = -\omega(\xi)\bigl(f - K_\mu^n(f)\bigr).$$

□

# Appendix D: Convergence of the iterates

The main result of this section is Lemma D.2, where it is established that for any $q \geq 1$, $f \in \mathcal{L}_V$

$$\lim_{n\to\infty} |[K_{S_n^Y}^q - K_\eta^q](f)(x)| = 0, \qquad \mathbb{Q}_y\text{-a.s.}, \tag{D.1}$$

with $K_\mu$ as in (2.2). The proof consists of showing that $K_\mu^q(f)(x)$ can be rewritten as $\mu^{\otimes q}(\overline{g})$ for some function $\overline{g} : E^q \to \mathbb{R}$ to be given below. We will then use results from Appendix E, associated with $V$-statistics for an appropriate class of functions, to complete our argument.

Introduce the following family of Markov transition probabilities on $(E \times E, \mathcal{E} \otimes \mathcal{E})$, indexed by $z_1 \in E$,

$$
\begin{aligned}
&T_{z_1}((w_0, w_0'); \mathrm{d}(w_1, w_1')) \\
&\quad := (1-\epsilon)K(w_0, \mathrm{d}w_1)\delta_{w_0}(\mathrm{d}w_1') \\
&\qquad + \epsilon\big[\alpha(w_0, z_1)\delta_{(z_1, w_0)}(\mathrm{d}w_1, \mathrm{d}w_1') + \big(1-\alpha(w_0, z_1)\big)\delta_{(w_0, w_0')}(\mathrm{d}w_1, \mathrm{d}w_1')\big].
\end{aligned}
$$

For any $w_0, w_0' \in E$ and $z := (z_1, \dots, z_q) \in E^q$, we define the iterates of this family of kernels as follows: for $k = 2, \dots, q$ and any $f \in \mathcal{L}_V$,

$$
T^k_{z_1,\dots,z_k}(f \otimes 1)(w_0, w_0') := T^{k-1}_{z_1,\dots,z_{k-1}}\big(T_{z_k}(f \otimes 1)(\cdot)\big)(w_0, w_0'), \tag{D.2}
$$

where for any $x, x' \in E$, $(f \otimes 1)(x, x') := f(x)$. Let $z := (z_1, \dots, z_q) \in E^q$. Following an argument identical to that developed in the proof of Lemma D.2 it is possible to show that for any $k = 1, \dots, q$ $T^k_{z_1,\dots,z_k}(f \otimes 1)(w_0, w_0')$ belongs to $\mathcal{L}_{\mathcal{V}_{z_1,\dots,z_k}}$ where for $w, w' \in E$,

$$
\mathcal{V}_{z_1,\dots,z_k}(w, w') := V(w) + V(w') + \sum_{i=1}^{k} V(z_i).
$$

**Proposition D.1.** *Assume* (A1). *For any $q \geq 1$, $(z_1, \dots, z_q) \in E^q$, $\mu \in \mathcal{P}_\infty(E)$, $f \in \mathcal{L}_V$, $x, x' \in E$ we have that*

$$
K^q_\mu(f)(x) = \int_{E^q} T^q_{z_1,\dots,z_q}(f \otimes 1)(x, x')\mu^{\otimes q}(\mathrm{d}(z_1, \dots, z_q)).
$$

**Proof.** The result is proved by induction. One immediately checks that for any $z_1 \in E$, $f \in \mathcal{L}_V$, $w_0, w_0' \in E$,

$$
T_{z_1}(f \otimes 1)(w_0, w_0') = (1-\epsilon)K(f)(w_0) + \epsilon\big[\alpha(w_0, z_1)f(z_1) + \big(1-\alpha(w_0, z_1)\big)f(w_0)\big],
$$

and hence

$$
\mu\big(T_{z_1}(f \otimes 1)(w_0, w_0')\big) = \int_E T_{z_1}(f \otimes 1)(w_0, w_0')\mu(\mathrm{d}z_1) = K_\mu(f)(w_0).
$$

Now assume the property is true for $k - 1 \geq 1$. Then

$$
\begin{aligned}
\mu^{\otimes k}\big(T^k_{z_1,\dots,z_k}(f \otimes 1)(w_0, w_0')\big) &= \mu^{\otimes(k-1)}\big(T^{k-1}_{z_1,\dots,z_{k-1}}\big\{\mu\big(T_{z_k}(f \otimes 1)(\cdot)\big)\big\}(w_0, w_0')\big) \\
&= \mu^{\otimes(k-1)}\big(T^{k-1}_{z_1,\dots,z_{k-1}}\big(K_\mu(f) \otimes 1\big)(w_0, w_0')\big),
\end{aligned}
$$

as required. □

Now, to establish (D.1) we need to show that $T^q_{z_1,\dots,z_q}(f)(w_0, w_0')$ lies within the class of functions for which Lemma E.2 applies; this is proved below.

**Lemma D.2.** *Assume* (A1)–(A3). *Let $q \geq 1$ be fixed and $f \in \mathcal{L}_V$. Then for any $x \in E$*

$$
\lim_{n\to\infty} |[K^q_{S^Y_n} - K^q_\eta](f)(x)| = 0 \qquad \mathbb{Q}_y\text{-}a.s.
$$

**Proof.** Our objective is to use the representation established in Proposition D.1 along with the result in Lemma E.2. To that end we show that for any $f \in \mathcal{L}_V$, then $T^q_{z_1,\ldots,z_q}(f \otimes 1)(w_0, w_0') \in \mathcal{L}_{\mathcal{V}_z^{(q)}}$, $z^{(q)} = (z_1, \ldots, z_q)$, where $T^q_{z_1,\ldots,z_q}(f \otimes 1)(w_0, w_0')$ is as in (D.2). The result can be proved by induction. Now, for any $k = 1, \ldots, q$, $w_{k-1}, w'_{k-1} \in E$ and $z = (z_1, \ldots, z_q) \in E^q$

$$\begin{aligned} T_{z_k}\big(\mathcal{V}_{z^{(q)}}\big)(w_{k-1}, w'_{k-1}) := (1-\epsilon)\Bigg[K(V)(w_{k-1}) + V(w_{k-1}) + \sum_{i=1}^{q} V(z_i)\Bigg] \\ + \epsilon\Bigg\{\alpha(w_{k-1}, z_k)\Bigg[V(w_{k-1}) + V(z_k) + \sum_{i=1}^{q} V(z_i)\Bigg] \\ + \big(1-\alpha(w_0, z_1)\big)\Bigg[V(w_{k-1}) + V(w'_{k-1}) + \sum_{i=1}^{q} V(z_i)\Bigg]\Bigg\}. \end{aligned}$$

Since there exists $M < \infty$ such that for any $x \in E$, $K(V)(x) \leq MV(x)$ we conclude that there exists $C_1 > 0$ such that for any $k = 1, \ldots, q$, $w_{k-1}, w'_{k-1} \in E$ and $z^{(q)} \in E^q$

$$T_{z_k}(\mathcal{V}_{z^{(q)}})(w_{k-1}, w'_{k-1}) \leq C_1 \mathcal{V}_{z^{(q)}}(w_{k-1}, w'_{k-1}). \tag{D.3}$$

This implies that for any $g \in \mathcal{L}_{\mathcal{V}_{z^{(q)}}}$ then $T_{z_k}(g)(w_{k-1}, w'_{k-1}) \in \mathcal{L}_{\mathcal{V}_{z^{(q)}}}$. Now we can proceed with the induction. Assume that for some $k-1 \geq 1$, if $g \in \mathcal{L}_{\mathcal{V}_{z^{(q)}}}$, then $T^{k-1}_{z_1,\ldots,z_{k-1}}(g)(w, w') \in \mathcal{L}_{\mathcal{V}_{z^{(q)}}}$. Then by definition

$$T^k_{z_1,\ldots,z_k}(f \otimes 1)(w_0, w_0') = T^{k-1}_{z_1,\ldots,z_{k-1}}\{T_{z_k}(f \otimes 1)(\cdot)\}(w_0, w_0'),$$

and the induction follows. Now, for any fixed $w_0, w_0'$ one has that $T^q_{z_1,\ldots,z_q}(f \otimes 1)(w_0, w_0') \in \mathcal{L}_{W^{(q)}}$ and the result follows from Lemma E.2. □

# Appendix E: Results on $U$ and $V$-statistics for Markov chains

Let $(E, \mathcal{E})$ be a Polish space and $\eta \in \{\mu \in \mathcal{P}(E) : \mu(W) < \infty\}$. Denote $\Omega = E^{\mathbb{N}}$ and $\mathcal{F} = \mathcal{E}^{\otimes\mathbb{N}}$ and consider a time-homogeneous Markov chain $\{X_n\}_{n\geq 0}$ with transition kernel $P$ such that $\eta P = \eta$ with $X_0 = x$. Denote by $\mathbb{P}_x$ the corresponding probability distribution. Note that $\{X_n\}$ should not be confused with the process introduced in Section 3.5.

For any sequence $\{Z_n\}$, $Z_n \in E$, any $q \in \mathbb{N}$ and $f : E^q \to \mathbb{R}$, denote for any $n \geq 1$ the associated $V$-statistic

$$S^{\otimes q}_{n,Z}(f) = \frac{1}{(n+1)^q} \sum_{\vartheta \in (q,n+1)} f\big(Z_{\vartheta(1)}, \ldots, Z_{\vartheta(q)}\big), \tag{E.1}$$

where $(q, n+1)$ is the set of all mappings of $\{0, \ldots, q-1\}$ into $\{0, \ldots, n\}$.

The main result of this section is Lemma E.2, where it is shown that under additional assumptions on $P$ and $f$, that

$$\lim_{n\to\infty} S_{n,X}^{\otimes q}(f) = \eta^{\otimes q}(f),$$

$\mathbb{P}_x$-a.s. The proof relies on a coupling argument with another Markov chain $\{Y_n\}_{n\geq 0}$ defined on $(\Omega, \mathcal{F})$ with the same transition $P$, but initialized at stationarity, that is, $Y_0 \sim \eta$. $\mathbb{P}_\eta$ denotes the corresponding probability distribution.

The conditions on $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 0}$ referred to above are given in (A2), and will, in particular, imply geometric ergodicity. The class of functions to which our results apply is defined as follows. Let $(W^r)^{(q)}(x^{(q)}) := \sum_{i=1}^q W(x_i)^r$ for any $r \in (0,1)$, $x^{(q)} := (x_1, \dots, x_q) \in E^q$; we will consider below the following class of functions

$$\mathcal{L}_{(W^r)^{(q)}} := \left\{ f \in mE^q : \sup_{x^{(q)} \in E^q} \left| f\left(x^{(q)}\right)\right| / (W^r)^{(q)}\left(x^{(q)}\right) < \infty \right\}.$$

For any sequence $\{Z_n\}$, $Z_n \in E$, any $q \in \mathbb{N}$ and $f : E^q \to \mathbb{R}$ denote for any $n \geq 1$ the associated $U$-statistic

$$S_{n,Z}^{\odot q}(f) = \frac{1}{(n+1)_q} \sum_{\vartheta \in \langle q, n+1\rangle} f\left(Z_{\vartheta(1)}, \dots, Z_{\vartheta(q)}\right), \tag{E.2}$$

where $\langle q, n+1\rangle$ is the set of one-to-one mappings from $\{0, \dots, q-1\}$ into $\{0, \dots, n\}$ and $n_q := n!/(n-q)!$. A preliminary result on $U$-statistics is first established, based on the aforementioned coupling.

**Proposition E.1.** *Assume* (A2) *and* (A3). *Let* $\{X_n\}_{n\geq 0}$ *and* $\{Y_n\}_{n\geq 0}$ *be as defined above. Then for any* $q \in \mathbb{N}$, $r \in [0,1)$, $f \in \mathcal{L}_{(W^r)^{(q)}}$ *and* $x \in E$, *there exists a coupling* $\{\check{X}_n, \check{Y}_n\}_{n\geq 0}$ *on some probability space* $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F}, \widetilde{\mathbb{P}})$, *such that*

$$\lim_{n\to\infty} |S_{n,\check{X}}^{\odot q}(f) - S_{n,\check{Y}}^{\odot q}(f)| = 0 \qquad \widetilde{\mathbb{P}}\text{-a.s.}$$

**Proof.** Let $\mathbb{P}_x^{(n)}$ (resp., $\mathbb{P}_\eta^{(n)}$) denote the law of $(X_n, X_{n+1}, \dots)$ (resp., $(Y_n, Y_{n+1}, \dots)$). Then, convergence in total variation of the processes is sufficient to imply that:

$$\lim_{n\to\infty} \left\| \mathbb{P}_x^{(n)} - \mathbb{P}_\eta^{(n)} \right\|_{\mathrm{TV}} = 0.$$

By Theorem 2.1 of Goldstein [19] the coupling exists; that is, there is a probability space $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F}, \widetilde{\mathbb{P}})$ such that $\widetilde{\mathbb{P}}(\Omega \times \cdot) = \mathbb{P}_x(\cdot)$ and $\widetilde{\mathbb{P}}(\cdot \times \Omega) = \mathbb{P}_\eta(\cdot)$ (note the dependence on $x$ of $\widetilde{\mathbb{P}}$ is omitted for notational simplicity). The process on this space is written $\{\check{X}_n, \check{Y}_n\}_{n\geq 0}$ and $T$ is the associated coupling time. Choose $q \in \mathbb{N}$. For any $\delta > 0$, $M \in \mathbb{N}$, $n > M \vee q$, one has that

$$\widetilde{\mathbb{P}}\Big(\sup_{k\geq n} |S_{k,\check{X}}^{\odot q}(f) - S_{k,\check{Y}}^{\odot q}(f)| > \delta\Big) \leq \widetilde{\mathbb{P}}\Big(\sup_{k\geq n} |S_{k,\check{X}}^{\odot q}(f) - S_{k,\check{Y}}^{\odot q}(f)| > \delta, T \leq M\Big) + \widetilde{\mathbb{P}}(T > M)$$

with $S^{\odot q}_{n,\check{X}}(f)$ as defined in (E.2). Now let $\varepsilon > 0$ be given and choose $M$ such that $\widetilde{\mathbb{P}}(T > M) < \varepsilon/2$. The first term on the right-hand side of the above inequality is now dealt with:

$$\widetilde{\mathbb{P}}\Big(\sup_{k\geq n}|S^{\odot q}_{k,\check{X}}(f) - S^{\odot q}_{k,\check{Y}}(f)| > \delta, T \leq M\Big) = \sum_{l=1}^{M}\widetilde{\mathbb{P}}\Big(\sup_{k\geq n}|S^{\odot q}_{k,\check{X}}(f) - S^{\odot q}_{k,\check{Y}}(f)| > \delta, T = l\Big). \quad \text{(E.3)}$$

Then, on the event $\{T = l\}$, one has that the terms involved in the definitions of $S^{\odot q}_{n,\check{X}}(f)$ and $S^{\odot q}_{n,\check{Y}}(f)$ only differ for $\vartheta$'s such that $\vartheta(i) \in \{0, \dots, l-1\}$ for some $i \in \{1, \dots, q\}$. For any $k > m > 0$, introduce the subset of $\langle q, k+1\rangle$

$$\Xi_{m,k} := \big\{\vartheta \in \langle q, k+1\rangle : \exists i \in \{1, \dots, q\} \text{ s.t. } \vartheta(i) < m\big\}.$$

Then for any $l \in \{1, \dots, M\}$, with $\bar{X}_{\vartheta(i)} = \check{X}_{\vartheta(i)}$ and $\bar{Y}_{\vartheta(i)} = \check{Y}_{\vartheta(i)}\mathbb{I}_{\{\vartheta(i)<l\}} + \check{X}_{\vartheta(i)}\mathbb{I}_{\{\vartheta(i)\geq l\}}$ and the notation

$$\Delta(f)_{\bar{X},\bar{Y}}(\vartheta(1), \dots, \vartheta(q)) := f\big(\bar{X}_{\vartheta(1)}, \dots, \bar{X}_{\vartheta(q)}\big) - f\big(\bar{Y}_{\vartheta(1)}, \dots, \bar{Y}_{\vartheta(q)}\big),$$

we have

$$\widetilde{\mathbb{P}}\Big(\sup_{k\geq n}|S^{\odot q}_{k,\check{X}}(f) - S^{\odot q}_{k,\check{Y}}(f)| > \delta, T = l\Big)$$
$$= \widetilde{\mathbb{P}}\left(\sup_{k\geq n}\frac{1}{(k+1)_q}\left|\sum_{\vartheta\in\Xi_{l,k}}\Delta(f)_{\bar{X},\bar{Y}}(\vartheta(1), \dots, \vartheta(q))\right| > \delta, T = l\right).$$

Let us denote for $l, n \in \mathbb{N}$ such that $n > l$

$$A_{l,n} := \left\{\sup_{k\geq n}\frac{1}{(k+1)_q}\left|\sum_{\vartheta\in\Xi_{l,k}}\Delta(f)_{\bar{X},\bar{Y}}(\vartheta(1), \dots, \vartheta(q))\right| > \delta\right\}.$$

It is now shown that $\widetilde{\mathbb{P}}(A_{l,n})$ vanishes as $n \to \infty$, which in turn will prove that the above vanishes as well for any $l \in \{1, \dots, M\}$. Since $f \in \mathcal{L}_{(W^r)^{(q)}}$, there exists some (deterministic) constant $\bar{M} < \infty$ such that

$$\widetilde{\mathbb{P}}(A_{l,n}) \leq \widetilde{\mathbb{P}}\left(\sup_{k\geq n}\frac{\bar{M}}{(k+1)_q}\left|\sum_{\vartheta\in\Xi_{l,k}}\left\{\sum_{i=1}^{q}\big[W\big(\bar{X}_{\vartheta(i)}\big)^r + W\big(\bar{Y}_{\vartheta(i)}\big)^r\big]\right\}\right| > \delta\right).$$

Consequently

$$\widetilde{\mathbb{P}}(A_{l,n}) \leq \mathbb{P}_x\left(\sup_{k\geq n}\frac{\bar{M}}{(k+1)_q}\left|\sum_{\vartheta\in\Xi_{l,k}}\left\{\sum_{i=1}^{q}\big[W\big(X_{\vartheta(i)}\big)^r + W\big(X_{\vartheta(i)}\big)^r\mathbb{I}_{\{\vartheta(i)\geq l\}}\big]\right\}\right| > \delta/2\right)$$
$$+ \mathbb{P}_\eta\left(\sup_{k\geq n}\frac{\bar{M}}{(k+1)_q}\left|\sum_{\vartheta\in\Xi_{l,k}}\left\{\sum_{i=1}^{q}W\big(Y_{\vartheta(i)}\big)^r\mathbb{I}_{\{\vartheta(i)<l\}}\right\}\right| > \delta/2\right).$$

The drift condition on $P$ yields the classical result $\sup_{i\geq 0}\{\mathbb{E}_x[W(X_i)]\vee\mathbb{E}_\eta[W(Y_i)]\}<\infty$. Note in addition that the cardinality of $\Xi_{l,k}$ is

$$l\binom{k}{q-1}q! = (k+1)_q\frac{ql}{k+1}.$$

Hence one may use an $\mathbb{L}_p$-proof similar to that in Proposition 6.4, with $p\in(1,1/r)$ along with a Borel–Cantelli argument via Markov's inequality, to conclude that $\lim_{n\to\infty}\widetilde{\mathbb{P}}(A_{l,n})=0$. This allows us to complete the proof by choosing $n$ such that each of the $M$ terms in the summation (E.3) is less than $\varepsilon/2M$.

□

**Lemma E.2.** *Assume* (A2) *and* (A3). *Let* $q\in\mathbb{N}$, $r\in[0,1)$, $f\in\mathcal{L}_{(W^r)^{(q)}}$, $x\in E$ *and* $\{X_i\}$ *be as defined earlier. Then*,

$$\lim_{n\to\infty}\left|\left[S^{\otimes q}_{n,X}-\eta^{\otimes q}\right](f)\right|\to 0 \qquad \mathbb{P}_x\text{-a.s.}$$

**Proof.** The idea of the proof is to use the almost sure convergence results for $U$-statistics of ergodic stationary processes established in [1]. In order to achieve this, the coupling $\widetilde{\mathbb{P}}$ introduced in Proposition E.1 is utilized. In particular, for any $\delta>0$, consider the following upper bound

$$\begin{aligned}
&\mathbb{P}_x\Bigl(\sup_{k\geq n}\left|\left[S^{\otimes q}_{k,X}-\eta^{\otimes q}\right](f)\right|>\delta\Bigr)\\
&\quad=\widetilde{\mathbb{P}}\Bigl(\sup_{k\geq n}\left|\left[S^{\otimes q}_{k,\check{X}}-S^{\odot q}_{k,\check{X}}\right](f)+\left[S^{\odot q}_{k,\check{X}}-S^{\odot q}_{k,\check{Y}}\right](f)+\left[S^{\odot q}_{k,\check{Y}}-\eta^{\otimes q}\right](f)\right|>\delta\Bigr)\\
&\quad\leq\mathbb{P}_x\Bigl(\sup_{k\geq n}\left|\left[S^{\otimes q}_{k,X}-S^{\odot q}_{k,X}\right](f)\right|>\delta/3\Bigr)+\widetilde{\mathbb{P}}\Bigl(\sup_{k\geq n}\left|\left[S^{\odot q}_{k,\check{X}}-S^{\odot q}_{k,\check{Y}}\right](f)\right|>\delta/3\Bigr)\\
&\qquad+\mathbb{P}_\eta\Bigl(\sup_{k\geq n}\left|\left[S^{\odot q}_{k,Y}-\eta^{\otimes q}\right](f)\right|>\delta/3\Bigr).
\end{aligned}$$

The convergence to zero of terms on the right-hand side of the inequality above from right to left are now considered. Since $\{Y_n\}_{n\geq 0}$ is an homogeneous Markov chain, started in stationarity, it is a stationary ergodic process. In addition, as $f$ is bounded by integrable products, $(E,\mathcal{E})$ is Polish and $\{Y_n\}_{n\geq 0}$ is absolutely regular (or weakly Bernoulli) [14], Theorem U of [1] can be invoked; the last term goes to zero (note that the proofs of [1] extend to Polish spaces). By Proposition E.1, the second term goes to zero.

Let us turn to the first term on the right-hand side of the inequality above. We use an argument similar to that of Theorem 5.1 of [18]. This uses the following identity

$$(n+1)^q\left[S^{\odot q}_{n,X}-S^{\otimes q}_{n,X}\right](f)=\left[(n+1)^q-(n+1)_q\right]S^{\odot q}_{n,X}(f)-\sum_{\vartheta\in\overline{\langle q,n+1\rangle}}f\bigl(X_{\vartheta(1)},\dots,X_{\vartheta(q)}\bigr),$$

where $\overline{\langle q,n+1\rangle}:=(q,n+1)\setminus\langle q,n+1\rangle$. Let $p\in(1,1/r)$. Since $f\in\mathcal{L}_{(W^r)^{(q)}}$, for any $(i_1,\dots,i_q)\in\{0,\dots,n\}^q$ then by Minkowski's inequality, followed by Jensen's inequality and the fact that via the drift condition $\sup_{i\geq 0}\mathbb{E}_x[W(pr)(X_i)] < MW(pr)(x)$ for some $M < \infty$

$$\mathbb{E}_x[|f(X_{i_1},\dots,X_{i_q})|^p]^{1/p} \leq \|f\|_{(W^r)^{(q)}} \sum_{l=1}^{q} \mathbb{E}_x[W(X_{i_l})^{rp}]^{1/p} \leq Mq\|f\|_{W^{(q)}} W^r(x).$$

As a result

$$\mathbb{E}_x[|S^{\odot q}_{n,X}(f)|^p]^{1/p} \leq Mq\|f\|_{W^{(q)}} W^r(x)$$

and

$$\mathbb{E}_x\left[\left|\sum_{\vartheta\in\langle q,n+1\rangle} f\big(X_{\vartheta(1)},\dots,X_{\vartheta(q)}\big)\right|^p\right]^{1/p} \leq M[(n+1)^q-(n+1)_q]q\|f\|_{W^{(q)}} W^r(x),$$

which allows us to conclude that there exists $C_q < \infty$ such that for any $n > q$

$$\mathbb{E}_x[(n+1)^q|[S^{\odot q}_{n,X} - S^{\otimes q}_{n,X}](f)|^p]^{1/p} \leq C_q[(n+1)^q-(n+1)_q]W^r(x).$$

Now since $(n+1)^q - (n+1)_q = \mathrm{O}(n^{q-1})$ and $p > 1$, a Borel–Cantelli argument can be used. The proof of the lemma now follows. □

# Appendix F: Verifying the assumptions

**Proof of Proposition 7.1.** Verifying many of the assumptions (A1) and (A2) is fairly simple and can be found in, for example, [21] (i.e., (A1)(i)(iii) and (A2)). The small-set condition (A1)(ii) can easily be proved in a similar way to the proof of Theorem 2.2 in [29] and is thus omitted. This leaves us with the latter part of (A2) ((A3) is clearly true here).

In our case,

$$V(x) = \left[\frac{|\pi|_\infty}{\pi(x)}\right]^{s_v}$$

for any $s_w \in (0,1)$ (see [21], Theorems 4.1 and 4.3). The expression for $W(x)$

$$\left[\frac{|\pi|_\infty}{\pi(x)}\right]^{\tilde{\alpha}s_w}, \qquad s_w \in (0,1),$$

follows similarly. For the last part of (A2), fix $r^*, s_w \in (0,1)$; then

$$\frac{V(x)}{W(x)^{r^*}} = |\pi|_\infty^{s_v - r^*\tilde{\alpha}s_w}\pi(x)^{r^*\tilde{\alpha}s_w - s_v},$$

which is upper bounded if $s_v \in (0, r^*\tilde{\alpha}s_w)$. □

## Acknowledgements

We thank two referees and the associate editor for valuable comments that vastly improved the paper. We also thank the editor for his patience with the paper. The second and third authors acknowledge the Institute of Statistical Mathematics, Japan, for their support during the writing of the paper. We also thank Adam Johansen for some useful comments on previous versions.

# References

[1] Aaronson, J., Burton, R., Dehling, H., Gilhat, D., Hill, T. and Weiss, B. (1996). Strong laws for $L$- and $U$-statistics. *Trans. Amer. Math. Soc.* **348** 2845–2866. MR1363941
[2] Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **72** 269–342.
[3] Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2008). Non-Linear Markov chain Monte Carlo. *ESIAM Proc*. **19** 79–84. MR2405652
[4] Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2008). On the convergence of the equi-energy sampler. *Stoch. Anal. Appl.* **26** 298–312. MR2399737
[5] Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16** 1462–1505. MR2260070
[6] Atchadé, Y.F. (2009). Resampling from the past to improve MCMC algorithms. *Far East. J. Theor. Stat.* **27** 81–99. MR2514400
[7] Atchadé, Y.F. (2010). A cautionary tale on the efficiency of some adaptive Monte Carlo Schemes. *Ann. Appl. Probab.* **20** 841–868. MR2680550
[8] Atchadé, Y.F., Fort, G., Moulines, É. and Priouret, P. (2011). Adaptive Markov chain Monte Carlo: Theory and methods. In *Inference and Learning in Dynamic Models* (D. Barber, S. Chiappa and A.T. Cemgil, eds.). Cambridge: CUP. To appear.
[9] Brockwell, A.E., Del Moral, P. and Doucet, A. (2011). Sequentially interacting Markov chain Monte Carlo methods. *Ann. Statist.* **38** 3387–3411.
[10] Del Moral, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. New York: Springer. MR2044973
[11] Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *J. Roy. Statist. Soc. Ser. B* **68** 411–436. MR2278333
[12] Del Moral, P. and Miclo, L. (2004). On convergence of chains with occupational self-interactions. *Proc. R. Soc. Lond. A. Math. Phys. Eng. Sci.* **460** 325–346. MR2052266
[13] Doucet, A., De Freitas, J.F.G. and Gordon, N.J. (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer. MR1847783
[14] Doukhan, P. (1994). *Mixing: Properties and Examples. Lecture Notes in Statistics* **85**. Berlin: Springer. MR1312160
[15] Fort, G. and Moulines, É. (2003). Polynomial ergodicity of Markov transition kernels. *Stochastic Process. Appl.* **103** 57–99. MR1947960
[16] Geyer, C. (1991). Markov chain maximum likelihood. In *Computing Science and Statistics: The 23rd Symposium on the Interface* (E. Keramigas, ed.) 156–163. Fairfax, VA: Interface Foundation.
[17] Glynn, P.W. and Meyn, S.P. (1996). A Lyapunov bound for solutions of the Poisson equation. *Ann. Probab.* **24** 916–931. MR1404536
[18] Grams, W.F. and Serfling, R.J. (1973). Convergence rates for $U$-statistics and related statistics. *Ann. Statist.* **1** 153–160. MR0336788

[19] Goldstein, S. (1979). Maximal coupling. *Probab. Theory Related Fields* **46** 193–204. MR0516740
[20] Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504
[21] Jarner, S.F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.* **85** 341–361. MR1731030
[22] Jasra, A., Stephens, D.A. and Holmes, C.C. (2007). On population-based simulation for static inference. *Statist. Comput.* **17** 263–279. MR2405807
[23] Kou, S.C., Zhou, Q. and Wong, W.H. (2006). Equi-energy sampler with applications to statistical inference and statistical mechanics (with discussion). *Ann. Statist.* **34** 1581–1619. MR2283711
[24] Meyn, S.P. and Tweedie, R.L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.* **4** 981–1011. MR1304770
[25] Meyn, S.P. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: CUP. MR2509253
[26] Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer. MR2080278
[27] Roberts, G.O. and Rosenthal, J.S. (1998). Two convergence properties of hybrid samplers. *Ann. Appl. Probab.* **8** 397–407. MR1624941
[28] Roberts, G.O. and Rosenthal, J.S. (2007). Coupling and ergodicity of adaptive MCMC. *J. Appl. Probab.* **44** 458–475. MR2340211
[29] Roberts, G.O. and Tweedie, R.L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110. MR1399158
[30] Shiryaev, A. (1996). *Probability*. New York: Springer. MR1368405

*Received April 2007 and revised March 2010*