ON NON-LINEAR MARKOV CHAIN MONTE CARLO VIA SELF-INTERACTING APPROXIMATIONS

BY CHRISTOPHE ANDRIEU, AJAY JASRA,
ARNAUD DOUCET & PIERRE DEL MORAL
UNIVERSITY OF BRISTOL, IMPERIAL COLLEGE LONDON,
UNIVERSITY OF BRITISH COLUMBIA & UNIVERSITY OF BORDEAUX

ABSTRACT. Let $\mathscr{P}(E)$ be the space of probability measures on a measurable space $(E, \mathcal{E})$. In this paper we introduce a class of non-linear Markov Chain Monte Carlo (MCMC) methods for simulating from a probability measure $\pi \in \mathscr{P}(E)$. Non-linear Markov kernels (e.g. Del Moral (2004); Del Moral & Doucet (2003)) $K : \mathscr{P}(E) \times E \to \mathscr{P}(E)$ can be constructed to admit π as an invariant distribution and have superior mixing properties to ordinary (linear) MCMC kernels. However, such non-linear kernels cannot be simulated exactly, so, in the spirit of particle approximations of Feynman-Kac formulae (Del Moral 2004), we construct approximations of the non-linear kernels via *Self-Interacting Markov Chains* (Del Moral & Miclo 2004) (SIMC). We present several non-linear kernels and demonstrate that, under some conditions, the associated self-interacting 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, combining the methodology with population-based Markov chain Monte Carlo (e.g. Jasra et al. (2007)). We also provide a comparison of our methods with sequential Monte Carlo samplers (Del Moral et al. 2006) when applied to a continuous-time stochastic volatility model.

CONTENTS

Date: October 21, 2007.

[1]**AMS 2000 Subject Classification**: Primary 82C80; Secondary 60F99, 62F15

Key words: Foster-Lyapunov Condition, Non-Linear Markov kernels, Poisson Equation, Population-based Simulation, Self Interacting Markov chains. **Short Title**: Non-Linear MCMC.

1. Introduction

Monte Carlo simulation is one of the most important elements of computational statistics and statistical physics. 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:

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

by

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

where $S_n^X(du) := \frac{1}{n+1}\sum_{i=0}^{n}\delta_{X_i}(du)$ is the empirical measure based upon random variables $\{X_k\}_{0\leq k\leq n}$ drawn from π. As an example, such integrals appear routinely in Bayesian statistics, in terms of posterior expectations; see for example Robert & Casella (2004) and the references therein. In such contexts, E is often of very high dimension, and indirect simulation methods such as Markov chain Monte Carlo (Robert & Casella, 2004) and sequential Monte Carlo (SMC) (Doucet et al., 2001; Del Moral 2004) need to be used.

It has long been known by Monte Carlo specialists that standard MCMC algorithms, such as the Metropolis-Hastings method, often have difficulties in simulating from complicated distributions. For example, when they exhibit multiple modes and/or possess strong dependencies between sub-elements of x (when it is vector-valued). In the former case, and despite its theoretical validity, 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, generic, methods proposed in the literature; we detail some of them here. Many of these approaches have relied upon MCMC techniques such as adaptive MCMC (Andrieu & Atchadé 2005; Andrieu & Moulines 2006; Haario et al. 2001), which, in some instances, attempts to improve the mixing properties of the transition kernel by using the information learnt in the past. In addition, there are methods which rely upon the simulation of parallel Markov chains (Geyer, 1991) and genetic algorithm type moves; see Jasra et al. (2007) for a review. These latter methods use the idea of running some of the parallel chains with invariant distribution $\eta \in \mathscr{P}(E)$, where η is easier to explore and related to π; hence the samples of the parallel chains can provide valuable information for traversing the support of π. Extensions to MCMC-based simulation methods have combined MCMC with SMC ideas, see for example Del Moral et al. (2006). Such approaches are often more flexible than MCMC as they do not rely, heavily, upon the ergodicity properties of any Markov kernel.

In this paper, we consider another alternative: non-linear MCMC via self-interacting approximations. Such methods rely primarily upon the ideas of MCMC. However, it is demonstrated below that the self-interacting approximation idea is similar to that of approximating a Feynman-Kac formulae (Del Moral 2004) and as such is linked to SMC methodology. It should be noted that related self-interacting ideas have appeared, directly in Brockwell (2005) and indirectly in Kou et al. (2006). An algorithm closely related to the work presented here is the resampling from the past algorithm of Atchadé (2006). However, the framework presented here is far more general both methodologically and theoretically. Methodologically, non-linear MCMC allows us to create a large class of new stochastic simulation algorithms; some which are presented here. In addition, the proofs presented in (Atchadé, 2006) are technically correct but do not correspond to the algorithm implemented; see Andrieu et al. (2007) for further details.

1.1. **Non-Linear Markov Kernels via Self Interacting Approximations.** Standard MCMC algorithms rely on Markov kernels of the form $K : E \to \mathscr{P}(E)$. These Markov kernels are *linear* operators on $\mathscr{P}(E)$; that is $\mu(dy) = \int_E \xi(dx)K(x, dy)$ where $\mu, \xi \in \mathscr{P}(E)$. A *non-linear* Markov kernel $K : \mathscr{P}(E) \times E \to \mathscr{P}(E)$ is defined as a non-linear operator on the space of probability measures. Non-linear Markov kernels, K_μ, can often be constructed to exhibit superior mixing properties to ordinary MCMC versions. For example, let

$$K_\mu(x, dy) = (1-\epsilon)K(x, dy) + \epsilon\Phi(\mu)(dy) , \tag{1.2}$$

where K is a Markov kernel of invariant distribution π, $\epsilon \in (0, 1)$, $\Phi : \mathscr{P}(E) \to \mathscr{P}(E)$ is a selection/mutation operator (Del Moral 2004), with $\Phi(\mu)(dy) := \mu(gK)/\mu(g)(dy)$ and g is a potential function. The potential function is taken so that it is bounded and measurable with $\Phi(\pi) = \pi$ (In this simple case, $g \equiv 1$ to ensure that π is an invariant distribution. In the following Sections, we discuss more elaborate algorithms where the function g is not constant). Simulating from K_π is clearly desirable as we allow regenerations from π, with K_π strongly uniformly ergodic (e.g. Roberts & Rosenthal (1998)). However, in most cases, we will be unable to simulate from K_π and instead we propose a self-interacting approximation.

A self-interacting Markov chain generates a stochastic process $\{X_n\}_{n\geq 0}$ which is allowed to interact with values realized in the past. That is, we might approximate (at time $n+1$ of the process) the selection/mutation operator by:

$$\Phi(S_n^X)(dy) = \frac{\sum_{i=0}^n g(X_i)K(X_i, dy)}{\sum_{i=0}^n g(X_i)} .$$

This process corresponds to a backward in time selection step (that is, generating a value from the history of the process, based upon the fitness (potential) g) and then a mutation step, via the kernel K. Selection allows previous values with high potential to return.

1.2. **Motivation and Structure of the Article.** In the context of stochastic simulation, SIMCs can be thought of as storing modes and then allowing the algorithm to return to them in a relatively simple way. Such a property, with the exception of Atchadé (2006), Brockwell (2005) and Kou et al. (2006), is not explicitly present in any of the above mentioned methodologies. Adaptive MCMC can be thought of as an indirect application of this idea, where parameters of the kernel are optimized via the Robbins-Monro algorithm. This approach does not retain all of the features of previously visited states. In other words, SIMCs can be considered as a *nonparameteric*, 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 non-linear Markov kernels and self-interacting approximations are introduced; we demonstrate that an algorithm closely related to the equi-energy sampler of Kou et al. (2006) is a special case of non-linear MCMC. In Section 4 we introduce the assumptions and discuss some preliminary results. In Sections 5 and 6, convergence results associated to the the marginals. and strong law of large numbers (SLLN) are presented. This analysis is of interest from a theoretical point of view: it brings together the literature of measure-valued processes and interacting particle systems (Del Moral 2004) used in SMC and the relatively recent literature on general state space Markov chains (Meyn & Tweedie 1993) used in MCMC. In Sections 7 and 8 some SIMCs algorithms are presented, based upon population Monte Carlo and demonstrated on toy and complex examples. In Section 9 some extensions to our ideas are discussed. The proofs are all given in the appendices; it should be noted that the various proofs for the respective algorithms are divided up in the appendix.

2. Notation and Definitions

2.1. **Notation.**

2.1.1. *Probability and Measure.* Define a measurable space $(H, \mathcal{H})$. Throughout, $\mathcal{H}$ 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 dx.

For a stochastic process $\{X_n\}_{n\geq 0}$ on $(H^{\mathbb{N}}, \mathcal{H}^{\otimes\mathbb{N}})$, $\mathcal{G}_n^X = \sigma(X_0, \ldots, X_n)$ denotes the natural filtration. $\mathbb{P}_\mu$ is taken as a probability law of a stochastic process with initial distribution μ and $\mathbb{E}_\mu$ the associated expectation. If $\mu = \delta_x$ (with δ the Dirac measure) we use $\mathbb{P}_x$ (resp. $\mathbb{E}_x$) instead of $\mathbb{P}_{\delta_x}$ (resp. $\mathbb{E}_{\delta_x}$).

If a ($\sigma-$finite) measure is dominated by another (denoted $\pi \ll \eta$), we sometimes abuse the notation and denote the Radon-Nikodym derivative with the same notation (e.g. if $\pi \ll \eta$ then $\pi(x)/\eta(x) = d\pi/d\eta(x)$). For $\sigma-$finite measures π and η we use $\pi \sim \eta$ to denote mutual absolute continuity. For $\mu \in \mathscr{P}(H)$, the notations $\mu^{(1)} \in \mathscr{P}(H\times H)$ (resp. $\mu^{(2)} \in \mathscr{P}(H\times H)$) $\mu^{(1)}(A\times B) = \mu(A)$ (resp. $\mu^{(2)}(A \times B) = \mu(B)$) are adopted.

2.1.2. *Markov and Self-Interacting Markov chains.* Let $(F, \mathcal{F})$ be a measurable space. Throughout for a Markov transition kernel $\Pi : F \to \mathscr{P}(F)$ the following standard notation is used: for measurable $f : F \to \mathbb{R}$, $\Pi(f)(x) := \int_F f(y)\Pi(x, dy)$ and for $\mu \in \mathscr{P}(F)$ $\mu\Pi(f) := \int_F \Pi(f)(x)\mu(dx)$.

Let $\Pi : F \to \mathscr{P}(F)$ be a transition kernel, $\Psi : \mathscr{P}(F) \times F \to \mathscr{P}(F)$ be a non-linear Markov kernel and $\epsilon \in (0, 1)$. The definition of our non-linear Markov process is based upon the following family of Markov transition kernels, given for any $\mu \in \mathscr{P}(F), x \in F$ and $A \in \mathcal{F}$ by

$$\Pi_\mu(x, A) = (1 - \epsilon)\Pi(x, A) + \epsilon\Psi(\mu)(x, A) \; . \tag{2.3}$$

Given its existence, we will denote by $\omega(\mu)$ ($\omega : \mathscr{P}(F) \to \mathscr{P}(F)$) the invariant distribution of this Markov kernel.

Recall that the empirical measure of an arbitrary stochastic process $(F^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \{X_n\}_{n\geq 0}, \mathbb{P})$ is defined, at time n, as:

$$S_n^X(du) := \frac{1}{n+1} \sum_{i=0}^{n} \delta_{X_i}(du) \; . \tag{2.4}$$

The class of self-interacting Markov chains (note that the term Markov chain is used as $\{X_n, S_n^X\}_{n\geq 0}$ forms a Markov chain) we study in this paper are defined as follows

Definition 2.1. *Let $\Pi : F \to \mathscr{P}(F)$ be a Markov kernel, $\Psi : \mathscr{P}(F) \times F \to \mathscr{P}(F)$ be a non-linear Markov kernel and $\epsilon \in (0, 1)$. A self-interacting Markov chain $(F^{\mathbb{N}}, \mathcal{F}^{\otimes\mathbb{N}}, \{X_n\}_{n\geq 0}, \{\mathcal{G}_n^X\}_{n\geq 0}, \mathbb{P}_x)$ is a stochastic process characterised by $X_0 = x$ and the conditional probability measures $\{\mathbb{P}_x(\cdot|\mathcal{G}_{n-1}^X)\} \in \mathscr{P}(F)$, such that $\mathbb{P}_x(X_n \in A|\mathcal{G}_{n-1}^X) := \Pi_{S_{n-1}^X}(X_{n-1}, A)$ which, for any $A \in \mathcal{F}$, is $\mathcal{F}^{\otimes n}-$measurable (where S_{n-1}^X is as Eq. (2.4) and $\Pi_{S_{n-1}^X}$ is as Eq. (2.3) with $\mu = S_{n-1}^X$).*

In this paper we explore various choices of spaces F, Markov transition Π and operator Ψ associated to self-interacting approximations of kernels Π_μ in (2.3).

2.1.3. *Norms.* For any $k \in \mathbb{N}$ the Euclidian norm of $x \in \mathbb{R}^k$ is denoted $|x|$. For $f : H \to \mathbb{R}^n$, $n \in \mathbb{N}$, we define $|f|_\infty := \sup_{x\in H} |f(x)|$. For a measurable function $f_1 : E \to \mathbb{R}^n$ the $\mathbb{L}_p$-norm is defined, assuming it exists, as $\left(\int_H |f_1(x)|^p d\mu\right)^{1/p}$ for $\mu \in \mathscr{P}(H)$. For $\eta, \mu \in \mathscr{P}(H)$ the total variation distance between them is $\|\eta - \mu\|_{TV} := \sup_{A\in\mathcal{H}} |\eta(A) - \mu(A)|$. For $U : H \to [1, \infty)$ and $f : H \to \mathbb{R}^n$

$$|f|_U \quad := \quad \sup_{x\in H} \frac{|f(x)|}{U(x)} .$$

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

$$\|\lambda\|_U \quad := \quad \sup_{|f|\le U} |\lambda(f)| ,$$

and the $U-$norm operator between two kernels $K_1, K_2 : H \to \mathscr{P}(H)$:

$$|||K_1 - K_2|||_U \quad := \quad \sup_{x\in H} \frac{\|K_1(x, \cdot) - K_2(x, \cdot)\|_U}{U(x)}.$$

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 $\mathbb{I}_A(x)$. Note also that $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathbb{T}_m = \{1, \ldots, m\}$ and $\mathscr{I}_\mu = \{f \in m(E) : \mu(gK(|f|)) < \infty\}$, with $m(E)$ the class of real-valued measurable functions on E.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. Non-Linear MCMC

3.1. **Non-Linear Markov kernels.** Non-linear MCMC can be characterised by the following procedure:

- Identify a non-linear kernel, Π_μ, that admits π as an invariant distribution and can be expected to mix faster than an ordinary MCMC kernel e.g. (1.2).
- Construct a stochastic process that approximates the kernel, which can be simulated in practice.

Based upon the representation (2.3), the following non-linear kernels are studied in this paper; the motivation for such kernels will be explained in the following Sections.

(NL1): Self Interacting Approximation. Let K be a Markov kernel of invariant distribution π, and $\Phi : \mathscr{P}(E) \to \mathscr{P}(E)$ a selection/mutation operator, with:

$$\Phi(\mu)(f) := \frac{\mu(gK(f))}{\mu(g)} \tag{3.5}$$

for $g : E \to (0, \infty)$ such that $|g|_\infty < \infty$ and $f : E \to \mathbb{R} \in \mathscr{I}_\mu$. In this case, at time $n+1$, (2.3) becomes

$$K_{S_n^X}(x_n, dx_{n+1}) := (1-\epsilon)K(x_n, dx_{n+1}) + \epsilon\Phi(S_n^X)(dx_{n+1}). \tag{3.6}$$

In the framework above $\Pi_\mu = K_\mu$, $\Pi = K$ and $\Psi = \Phi$.

Example 1. Selection/Mutation, Identity Potential. *Let $g(x) \equiv 1$ and K be an MCMC kernel of invariant distribution π. Then we may simulate (3.6) to estimate (1.1).*

In (NL1) we have presented a 'standard' self-interacting approximation (that is, as presented in Del Moral & Miclo (2004)). Our objective is to study how well such a method can perform, given that no information about π is used in the selection step. As a result, this may lead to very slow convergence and misleading results as illustrated by the following cautionary example.

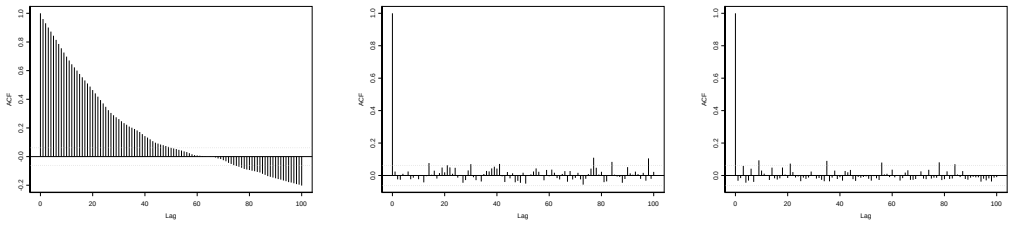
A Cautionary Example

Suppose that we are to simulate from $X \sim 0.4\mathcal{N}(0, 0.5) + 0.6\mathcal{N}(17.5, 1)$, with $\mathcal{N}(\mu, \sigma^2)$ the normal distribution of mean μ and variance σ^2.

In this example we ran a normal Random Walk Metropolis (RWM) algorithm for 50000 iterations with proposal variance adjusted to yield an acceptance rate of 0.25; the algorithm was initialized with a draw from a $\mathcal{N}(0, 0.5)$ distribution. We also ran two self-interacting algorithms (NL1, example 1) with self-interaction allowed every 50th and 500th step (that is, the RWM kernel K is iterated as K^{50} or K^{500} and the empirical measure is based upon the samples generated after the (potential) selection step).

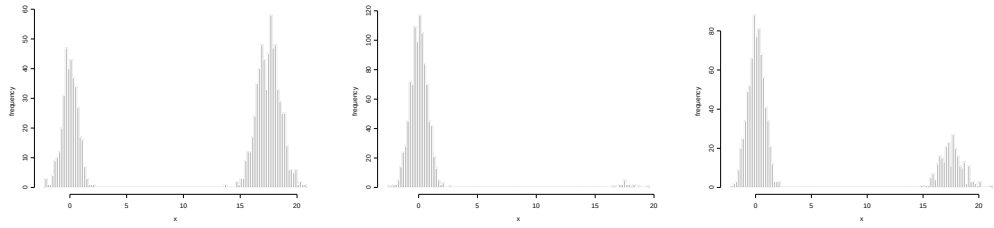
In Figure 1, we present the estimates of the autocorrelation function for the three algorithms. Clearly the RWM chain mixes poorly with a slow decay of the autocorrelations (see Figure 1a). Conversely the self-interacting algorithms appear to mix very quickly (see Figures 1b and 1c). However, we have to be very cautious. In Figure 2, we display the Monte Carlo estimates of the target distribution for the three algorithms. It appears that the RWM (see Figure 2a) significantly outperforms the self-interacting algorithms (see Figures 2b and 2c). We attribute the poor performance of the self-interacting algorithms to the fact that the process has started in a minor mode and spent a significant amount of time there; both due to the slow mixing of K and the selection which forces us back to the minor mode. If the number of self-interactions is reduced then the estimate is improved but the resulting algorithm still does not provide satisfactory results compared to a standard RWM.

This example makes two important points:



(a) RWM algorithm (b) Self-interacting 50th (c) Self-interacting 500th

Figure 1: Estimates of the autocorrelation function.



(a) RWM algorithm (b) Self-interacting 50th (c) Self-interacting 500th

Figure 2: Estimates of the target distribution.

- Self-interacting algorithms can appear to perform very well, with low autocorrelations of sample paths, but can give very poor answers as the interactions slow down convergence to the stationary regime; see Atchadé & Rosenthal (2005) for theoretical evidence in the parametric adaptive case.
- Simple self-interacting mechanisms as described above are expected to be helpful only in situations where the target is unimodal and the chain is initialized in regions of high probability masses.

This cautionary example has motivated the development of non-linear kernels based upon auxiliary self-interactions.

(NL2): Auxiliary Self-Interaction. We introduce the following family of kernels $\{\Pi_\mu, \mu \in \mathscr{P}(F)\}$ which is such that $(F := E \times E, \mathcal{F} := \mathcal{E} \times \mathcal{E})$. The intention is to improve the exploration

ability of the simulated kernel and the selection mechanism,

$$\begin{aligned}\Pi_{\mu^{(2)}}((x,y),d(x',y')) &:= (1-\epsilon)K(x,dx')\times P(y,dy') + \epsilon\Psi(\mu^{(2)})((x,y),d(x',y'))\,,\\ \Psi(\mu^{(2)})((x,y),d(x',y')) &:= \Phi(\mu)(dx)\times P(y,dy')\,.\end{aligned} \tag{3.7}$$

with $P : E \to \mathscr{P}(E)$ a Markov kernel g and Φ as defined in (3.5).

Example 2. Selection/Mutation with Potential. *Let P be an MCMC kernel of invariant distribution η, and assume $\pi \ll \eta$. Let $g(x) = \frac{\pi(x)}{\eta(x)}$ and set K to be an MCMC kernel of invariant distribution π. If we were able to sample exactly from η then one could sample exactly from $\Pi_{\pi\times\eta}$ which has invariant distribution $\pi\times\eta$. However, as we shall see, for efficient algorithms, this will not be the case and instead we suggest using the following approximation, here given at time $n+1$:*

$$\Pi_{\pi\times S_n^y}((x_n,y_n),d(x_{n+1},y_{n+1})) = \left[(1-\epsilon)K(x_n,dx_{n+1}) + \epsilon\Phi(S_n^y)(dx_{n+1})\right]P(y_n,dy_{n+1})$$

that is, we are 'feeding' the chain $\{X_n\}_{n\geq 0}$ the empirical measure S_n^y.

In (NL2) we attempt to circumvent the problem of identity potential. We extend the space to allow us to use information related to π in the selection.

(NL3): Auxiliary Self-Interaction with Genetic Moves. For any $\mu \in \mathscr{P}(E)$ we define a non-linear Markov kernel $Q_\mu : \mathscr{P}(E)\times E \to \mathscr{P}(E)$ with potential $g : E\times E \to (0,\infty)$ ($|g|_\infty < \infty$) as

$$Q_\mu(x,dx') := \frac{\int_{E\times E} g(x,v)\widetilde{K}((x,v),dx')\mu(dv)}{\int_E g(x,v)\mu(dv)}\,,$$

where

$$\begin{aligned}\widetilde{K}((u,v),dx) &:= \alpha(u,v)K(v,dx) + [1-\alpha(u,v)]K(u,dx)\\ \alpha(u,v) &:= 1\wedge\frac{\pi(v)\eta(u)}{\pi(u)\eta(v)}\end{aligned}$$

and $\pi\sim\eta$. We now define the following non-linear kernel

$$\begin{aligned}\Pi_{\mu^{(2)}}((x,y),d(x',y')) &= (1-\epsilon)K(x,dx')P(y,dy') + \epsilon\Psi(\mu^{(2)})((x,y),d(x',y'))\\ \Psi(\mu^{(2)})((x,y),d(x',y')) &:= Q_\mu(x,dx')\times P(y,dy').\end{aligned}$$

It should be noted that changing the input $\mu \in \mathscr{P}(E)$ does not change the function α.

Example 3. Simplified Equi-Energy Sampling (Kou et al. 2006) with Identity Potential. *Let $g(x,y) \equiv 1$, $\eta \sim \pi$, K (resp. P) be an MCMC kernel of invariant distribution π (resp. η). Then we have $(\pi \times \eta)(\Pi)_{\pi\times\eta} = \pi \times \eta$; that is, via Fubini:*

$$\begin{aligned} \pi Q_\eta(dx') &= \int_{E\times E}\left[\int_{E\times E} \pi(dx)\eta(dy)K^S((x,y),d(u,v))\right]K(u,dx') \\ &= \pi K(dx'). \end{aligned}$$

We can then simulate the self-interacting Markov chain $\Pi_{\pi\times S_n^Y}$ at time n, where S_n^Y is the empirical measure that has been built by the chain with invariant distribution η.

(NL3) provides a way to control the information that is provided by the approximation S_n^Y. That is, the exchange step will allow us a criterion to check the consistency with the target of the selected value. This may help improve estimation, if S_n^Y converges slowly. We note that the algorithm is less sophisticated than that of Kou et al. (2006) as we do not consider exchanges to occur between states in equi-energy rings.

3.2. **The Algorithms.** To summarize our algorithm for (NL1) is:

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

From herein, for (NL2) and (NL3), we abuse the notation and use $K_\mu = \Pi_{\pi\times\mu}$, for $\mu \in \mathscr{P}(E)$; the algorithm is:

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 and preliminary results

Our objective is to now prove convergence of the marginals; i.e. convergence to zero of:

$$|\mathbb{E}_{(x,y)}\big[f(X_k) - \pi(f)\big]|$$

for some suitable f. In addition, we seek to prove a strong law of large numbers for the sample path:

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

Before either result is considered, we give our assumptions and a series of technical results used to prove the convergence results. We will prove our results for (NL1) with $g \equiv 1$ as in example 1 and for (NL3) with g as in example 3 ($g(u,v) \equiv 1$).

Remark. *It was noted by É. Moulines (personal communication) that the SLLN and convergence of the marginals for (NL2) can be proved by standard regeneration arguments from Markov chain theory. As a result, we do not give the proofs or statements for this algorithm using our approach.*

Recall that (NL1) simulated a process $(F^{\mathbb{N}} = E^{\mathbb{N}}, \mathcal{F}^{\otimes\mathbb{N}} = \mathcal{E}^{\otimes\mathbb{N}}, \{X_n\}_{n\geq 0}, \{\mathcal{G}_n^X\}_{n\geq 0}, \mathbb{P}_x)$, $x \in E$ with finite-dimensional law:

$$\mathbb{P}_{x,n}(d(x_0,\ldots,x_n)) \quad = \quad \delta_x(dx_0) \times \Pi_{S_0^x}(x_0, dx_1) \times \cdots \times \Pi_{S_{n-1}^x}(x_{n-1}, dx_n)\ .$$

Similarly, (NL2) and (NL3) simulated a stochastic process on $(F^{\mathbb{N}} = (E\times E)^{\mathbb{N}}, \mathcal{F}^{\otimes\mathbb{N}} = (\mathcal{E}\otimes\mathcal{E})^{\otimes\mathbb{N}}, \{X_n, Y_n\}_{n\geq 0},\ \{\mathcal{G}_n\}_{n\geq 0}, \mathbb{P}_{(x,y)})$, $(x,y) \in F$, (here we have denoted the natural filtration $\mathcal{G}_n^{X,Y} = \mathcal{G}_n$ for notational simplicity) with finite-dimensional law as above (with appropriate notational changes). For (NL2-3), since $\{Y_n\}$ is generated independently of $\{X_n\}$, we denote the probability law of the Markov chain $\{Y_n\}$ as $\mathbb{Q}_y$.

We recall that the proofs associated to (NL1) & (NL3) are given their respective appendices at the end of the paper.

4.1. **Assumptions.** We now give our assumptions on K and g used to define our non-linear Markov chains (NL1-3). The assumptions apply to all (NL1-3) except when preceded with **(NLZ)** with $\mathbf{Z} \in \mathbb{T}_3$, in which case the assumption is specific to the algorithm considered. Let $V : E \to [1,\infty)$, for any $M_1, M_2 > 0$ the notation $\mathscr{P}_{M_1}(E) := \{\mu \in \mathscr{P}(E) : \mu(gV)/\mu(g) \leq M_1\}$, $\mathscr{P}^{M_2}(E) := \{\mu \in \mathscr{P}(E) : \mu(g) \geq M_2\}$ and $\mathscr{P}_{M_1}^{M_2}(E) := \mathscr{P}_{M_1}(E) \cap \mathscr{P}^{M_2}(E)$ is used. In an abuse of notation: $\mathscr{P}_\infty(E) := \{\mu \in \mathscr{P}(E) : \mu(V) < \infty\}$.

(**A1**) *Stability of the algorithm.*

(NL1) For V above, there exists a universal constant $M_1^* < \infty$, such that for any $n \geq 0$ we have

$$S_n^X(V) \quad \leq \quad M_1^*\ , \mathbb{P}_x - \text{a.s.}\ .$$

(**A2**) *Stability of K and P.*

Transition K

(i) (*Invariance*). $K : E \to \mathscr{P}(E)$ is a $\pi-$invariant Markov kernel.

(ii) (*One-step Drift Condition*). There exists $V : E \to [1, \infty)$, $\lambda < 1$, $b < \infty$ and $C \in \mathcal{E}$ such that for any $x \in E$

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

(iii) (*One-step Minorization on level set C*). There exists $\theta > 0$ such that C in (ii) is a $(1, \theta)$-small set for K, i.e. there exists $\theta > 0$ and a non-trivial probability measure $\nu \in \mathscr{P}(E)$ satisfying $\nu(C) > 0$ such that for all $(x, A) \in E \times \mathcal{E}$,

$$K(x, A) \quad \geq \quad \theta\mathbb{I}_C(x)\nu(A)\,.$$

(iv) *(Small set constraint.)* The small set C in (ii)-(iii) is of the form $C_d := \{x \in E : V(x) \leq d\}$ for $d \in (0, \infty)$ satisfying the constraint

(NL1) $d > \epsilon(\lambda M_1^* + b)/(1 - (1 - \epsilon)\lambda) \vee 1$

with λ and b defined in (ii), ϵ defined in (2.3) and M_1^* defined in (A1).

The small set C_d is such that for any $d \geq 1$, C_d is $(1, \theta_d)-$small with minorizing measure ν_d and two possibilities associated to the minorization condition:

(NL3) (a) For any $d \geq 1$, $\theta_d > 0$, $\nu_d(C_d) = 1$. In addition, if for any $m \geq 0$, $d \equiv d(S_m^Y)$, then $\theta_d > \varphi_d > 0$, $\varphi_d \geq \Lambda(S_m^Y)$ $\mathbb{Q}_y-$a.s. and for any $r \in (0, 1/4)$, $p \in [1, 1/4r)$ $\mathbb{E}_y[\Lambda(S_m^Y)^p]^{1/p} < \infty$.

(NL3) (b) As above, without the final condition.

(v) (*Convergence rate constraint*). **(NL1)** The assumptions (i), (ii) and (iii) above imply the existence of $M < \infty$ and $\rho \in (0, 1)$ such that for any $r \in (0, 1)$ and $f \in \mathcal{L}_{V^r}$, $|K^n(f) - \pi(f)|_{V^r} \leq \widetilde{M}|f|_{V^r}\rho^n$ (see Theorem 2.3. in Meyn & Tweedie (1994)). We further impose that

$$\widetilde{M} \quad < \quad \frac{1 - (1 - \epsilon)\rho}{\epsilon\rho}\,.$$

Transition P ($W-$*Uniform Ergodicity*). **(NL3)** $P : E \to \mathscr{P}(E)$ is an $\eta-$invariant Markov kernel. Furthermore, there exist $W : E \to [1, \infty)$ such that P is a W-uniformly ergodic Markov transition kernel with a one-step drift condition and one-step minorization condition. In addition $V \in \mathscr{L}_W$ (where $V : E \to [1, \infty)$ is defined in (ii)).

(**A3**) *State-Space Constraint*

(NL3) $(E, \mathcal{E})$ is polish (separable complete metrisable topological space).

In some of our results below, it will sometimes be convenient to express the dependence of some constants ($M(\cdot)$ say), on some of the quantities above. For the parameters in the drift and

minorization conditions as well as the non-linear kernel $(\lambda, b, \theta, \nu(C), d, \epsilon)$ we use the notation $M(\lambda, b, \theta, \nu(C), d, \epsilon) = M(\mathbb{G})$. When a constant depends only on the parameters of the drift condition, $\mathbb{D}$ and $\mathbb{D}_y$ (for the drift associated to P) are adopted.

4.2. **Discussion of the assumptions.** Our proofs of the SLLN will rely upon a Martingale approximation and the Poisson equation (e.g. Glynn & Meyn (1996)). (A1) in conjunction with (A2) will allow us to establish a drift condition for the collection $\{K_\mu\}$ for any fixed $\mu \in \mathscr{P}(E)$. This will allow us to verify the existence of the (resolvent) solution to the Poisson equation and quantitative bounds.

For both (A1) and (A2), with respect to the kernel (NL1), the assumptions are quite strong. The condition (v) in (A2) will be difficult to check and has the interpretation that we would like to both take $\epsilon \approx 0$ and iterate K to improve $\widetilde{M}$ and ρ. In effect, the assumption requires very fast mixing of K and suggests to us that fully self-interacting algorithms are likely to converge very slowly. The condition in (A1) for upper bounded V, will hold, but in more general cases, it is unlikely to be true for every starting point.

(A2), for (NL3), appears quite strong, but can be verified in some important cases such as for RWM kernels; see Andrieu et al. (2001) for example.

(A3) will be used for the SLLN for (NL3). In this case we have dealt with the perturbation between the average of the invariant measures of the non-linear kernels (see the discussion of the strategy of the proof in Section 6.1) and π using $U-$statistics (e.g. Hoeffding (1948)). Essentially, for the algorithm to converge, it appears that we require that the iterated kernel $K^j_{S^Y_n}(x, A)$ converges (almost surely) to $K^j_\eta(x, A)$, which is a difficult task to establish. We adopt a decomposition, via the Von-Mises statistic and then use the well-known relationship between Von-Mises and $U-$statistics (e.g. Grams & Hoeffding (1973)). We then need to prove a SLLN for $U-$statistics for Markov chains (established in Aaronson et al. (1996)), for which the assumption (A3) is required.

In essence, the absence of a stability condition, in terms of the empirical measure, for (NL3) (versus (NL1) (A1)) is due in part to the stability that is provided by the auxiliary Markov chain. For the fully self-interacting algorithm, there is no apparent stability of the process; thus it is imposed by the assumption in (A1). That is, for (NL3) we utilize the fact that $\sup_{m\geq 0} \mathbb{E}_y[V(Y_m)^{r^*}] < \infty$ with r^* as in (A2) against, for (NL1), $S_m(V) < M_1^*$.

4.3. **Invariant probability and geometric ergodicity.** Using standard drift and minorization conditions we establish the existence of an invariant probability measure for K_μ for any $\mu \in \mathscr{P}^{M_2}_{M_1}(E)$ for $M_1 \in (0, \infty)$ and $M_2 > 0$ (NL1) or $\mu \in \mathscr{P}_\infty(E)$ (NL3) under (A2), and the

V-geometric convergence of the Markov chain associated to K_μ, with uniform upper bound on the rate of convergence for all $\mu \in \mathscr{P}_{M_1}^{M_2}(E)$ (NL1) or a rate dependent upon $\mu(V)$ (NL3). This latter property will provide a 'stochastic drift condition' which will require more intricate proof techniques (than for (NL1)).

Proposition 4.1. *For both (NL1) and (NL3) assume (A2-i-ii-iii) with C of the form $C_d := \{x \in E : V(x) \leq d\}$*

- **(NL1)** *for some $d > \epsilon(\lambda M_1 + b)/(1 - (1-\epsilon)\lambda) \geq 1$ for some $M_1 \in (0, \infty)$,*
- **(NL3)** *for all d of the form $d(\mu) = 1 + \frac{\bar{b}(\mu)\alpha}{1-\lambda}$ for $\mu \in \mathscr{P}_\infty(E)$ some $\alpha > 1$ and $\bar{b}(\mu) = b + \epsilon[\lambda\mu(V) + b]$*

where λ, b, V defined in (A2) and ϵ in (2.3). In addition we assume that (A2-iv) (a) or (b) holds for (NL3). Then

(1) **(NL1)** *there exist $\lambda^*, \theta^* \in (0,1)$ and $b^* \in (0, \infty)$ such that for any $\mu \in \mathscr{P}_{M_1}(E)$ and $(x, A) \in E \times \mathcal{E}$*

$$K_\mu V(x) \leq \lambda^* V(x) + b^* \mathbb{I}_{C_d}(x) \,, \tag{4.8}$$

$$K_\mu(x, A) \geq \theta^* \mathbb{I}_{C_d}(x) \nu(A) \,, \tag{4.9}$$

where $\nu \in \mathscr{P}(E)$ is defined in (A2),

(NL3) *for any $\mu \in \mathscr{P}_\infty(E)$ there exist $(\theta'_{d(\mu)}, \nu_{d(\mu)}) \in (0,1) \times \mathscr{P}(E)$ such that for any $r \in (0, 1]$, and $(x, A) \in E \times \mathcal{E}$:*

$$\begin{aligned} K_\mu V^r(x) &\leq \tilde{\lambda}^r V(x)^r + \tilde{b}(\mu)^r \mathbb{I}_{C_{d(\mu)}}(x) \\ K_\mu(x, A) &\geq \mathbb{I}_{C_{d(\mu)}}(x) \theta'_{d(\mu)} \nu_{d(\mu)}(A) \end{aligned}$$

with

$$\begin{aligned} \tilde{\lambda} &= \lambda + \frac{1-\lambda}{\alpha} \\ \tilde{b}(\mu) &= \lambda d(\mu) + \bar{b}(\mu). \end{aligned}$$

(2) there exists a function $\omega : \mathscr{P}_{M_1}(E) \to \mathscr{P}(E)$ **(NL1)** *(resp. $\omega : \mathscr{P}_\infty(E) \to \mathscr{P}_\infty(E)$* **(NL3)***), such that for any $\mu \in \mathscr{P}_{M_1}(E)$ (resp. $\mu \in \mathscr{P}_\infty(E)$)*

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

(3) **(NL1)** *there exist constants, $\rho \in (0,1)$, $M(\cdot) < \infty$ depending upon ϵ, λ, b, θ, $\nu_d(C)$ (as defined in Eq. (2.3) and (A2)), M_1 and d such that for any $\mu \in \mathscr{P}_{M_1}(E)$, $r \in (0, 1]$ and*

$f \in \mathscr{L}_{V^r}$

$$|K_\mu^n(f) - \omega(\mu)(f)|_{V^r} \quad \leq \quad M(M_1, \mathbb{G})|f|_{V^r}\rho^n \,.$$

(NL3) *for any $\mu \in \mathscr{P}_\infty(E)$ there exist constants, $\rho(\cdot) \in (0,1)$, $M(\cdot) < \infty$ depending upon ϵ, λ, b, V, θ, $\nu(C)$ (as defined in Eq. (2.3) and (A2)), such that for any $r \in (0,1]$ and $f \in \mathscr{L}_{V^r}$*

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

For (NL1) by arguments along the lines of those of the proof of Theorem 2 of Breyer & Roberts (2001) (see also Corollary 1 of Hobert & Robert (2004)) one can easily establish the following expression for $\omega(\mu)(A)$ for any $\mu \in \mathscr{P}_{M_1}(E)$ and $A \in \mathcal{E}$

$$\omega(\mu)(A) \quad = \quad \sum_{n \in \mathbb{N}} \epsilon(1-\epsilon)^{n-1}\Phi(\mu)K^{n-1}(A) \,. \tag{4.10}$$

This property will be useful in proving some of our results for (NL1) (for (NL3) the expression is too complicated to be useful). We note that we expect that it is possible to prove our results without using (4.10), but that it does allow for simple arguments in our proofs. Some continuity properties associated to the invariant measures are as follows.

Proposition 4.2. *For both (NL1) and (NL3) assume (A2-i-ii-iii) with C of the form $C_d := \{x \in E : V(x) \leq d\}$*

- **(NL1)** *for some $d > \epsilon(\lambda M_1 + b)/(1-(1-\epsilon)\lambda) \geq 1$ for some $M_1 \in (0, \infty)$,*
- **(NL3)** *for all d of the form $d(\mu, \xi) = 1 + \frac{(\bar{b}(\mu) \vee \bar{b}(\xi))\alpha}{1-\lambda}$ with $\mu, \xi \in \mathscr{P}_\infty(E)$, $\alpha > 1$*

In addition we assume that (A2-iv) (a) or (b) holds for (NL3).

(NL1) *Then there exists $M(\cdot) < \infty$ (depending on M_1 and the constants in (A2)) such that for any $\xi, \mu \in \mathscr{P}_{M_1}(E)$ we have for any $r \in (0,1]$*

$$\|\omega(\xi) - \omega(\mu)\|_{V^r} \quad \leq \quad M(M_1, \mathbb{G})|||K_\xi - K_\mu|||_{V^r} \,.$$

(NL3) *Then for any $\mu, \xi \in \mathscr{P}_\infty(E)$ there exists $M(\cdot) < \infty$ depending on $\mu, \xi \in \mathscr{P}_\infty(E)$, V and the constants in (A2)) such that for any $r \in (0,1]$*

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

The proof is in Appendix A.1. This result motivates the work of the following section, where we show that (A1) and (A2) imply the Lipschitz continuity of $\mu \mapsto K_\mu$.

4.4. **Lipschitz Continuity.** Noting that for any $\mu, \xi \in \mathscr{P}(E)$ and $r \in [0,1]$, $|||K_\xi - K_\mu|||_{V^r} = \epsilon\|\Phi(\xi) - \Phi(\mu)\|_{V^r}$ for (NL1) and $|||K_\xi - K_\mu|||_{V^r} = \epsilon|||Q_\xi - Q_\mu|||_{V^r}$ for (NL3) we establish hereafter local Lipschitz continuity results for $\mu \mapsto \Phi(\mu)$ and $\mu \mapsto Q_\mu$ that will imply the local Lipschitz continuity of $\mu \to K_\mu$ which will be used in the proofs of many of the subsequent results.

4.4.1. *Case (NL1).* We do not present any continuity results, but state without proof that the invariant measure satisfies a Lipschitz contraction condition with respect to V^r-total variation. We will see in the proof of Theorem 6.4 that a result similar to a contraction (as opposed to continuity) will be instrumental in proving a strong law of large numbers. This will be used in a similar manner to Hypothesis 4.1 of Del Moral & Miclo (2004) where the fact that $M < 1$ (the contraction coefficient) allowed them to obtain optimal rates. We remark that we cannot claim that it is necessary to have contraction (i.e. $M < 1$ for the Lipschitz continuity and to ensure a SLLN), but that it does appear to be the case, under our proof technique. If we were able to establish that the contraction is a necessary condition, then this would imply that fully self-interacting algorithms are not of substantial use in the stochastic simulation tasks we are concerned with.

4.4.2. *Case (NL3).* We now consider an analogous result for the kernel Q_μ that appears in the definition of (NL3).

Proposition 4.3. *Assume (A2-ii). Let $\mu, \xi \in \mathscr{P}_\infty(E)$, then for any $r \in (0,1]$:*

$$|||Q_\mu - Q_\xi|||_{V^r} \leq 2(\lambda + b)^r \|\mu - \xi\|_{V^r}$$

where $\lambda \in (0,1)$ and $b < \infty$ as in (A2-ii).

5. Convergence of the marginals

We now present the convergence of the marginals. The difficulties for providing explicit bounds are essentially the dependence structure of (NL1) and the non-linearity of the kernel in (NL3). Below, we introduce a constant $M^\star < 1$ which will be defined in Theorem 6.4. In addition (A2-b) is used to note the fact that option (b) is assumed in (A2) for (NL3).

Lemma 5.1. *Assume (A1, 2-b, 3). Let $k \in \mathbb{N}$ and that $f : E \to \mathbb{R}$:*

- **(NL1)** *$r \in (0, \frac{1}{4} \wedge (1 - M^\star))$ and $|f| \leq V^r$*
- **(NL3)** *$r \in (0, 1/3r)$ and $f \in \mathscr{L}_{V^r}$.*

Then we have, for (NL1) and (NL3):

$$\lim_{k \to \infty} |\mathbb{E}_\cdot[f(X_k) - \pi(f)]| = 0.$$

The result establishes the convergence of the marginals, however, the proof in Appendix A.4 for (NL1) and (NL3) relies upon the convergence results established for the SLLN.

6. Law of large numbers

6.1. **Strategy of the Proof.** The strategy of the proof is now outlined, in the context of (NL1), the case (NL3) being similar. We aim, where possible, to establish vanishing bounds (as $n \to \infty$) for quantities of the type

$$\mathbb{E}_x \left[|S_n^X(f) - \pi(f)|^p \right]^{1/p}$$

for $f \in \mathscr{L}_{V^r}$ for some $r \in (0,1]$ and p's that depend on both r and the algorithm considered. Let us introduce the following sequence of probability distributions $\{S_n^\omega := 1/(n+1) \sum_{i=0}^n \omega(S_i^Y)\}_{n \geq 0}$ where $\omega(\mu)$ is the invariant probability distribution of K_μ. This distribution can be used as a recentering term in the following decomposition,

$$S_n^X(f) - \pi(f) \quad = \quad S_n^X(f) - S_n^\omega(f) + S_n^\omega(f) - \pi(f)\,. \tag{6.11}$$

The analysis of the first term on the RHS of (6.11) relies upon a classical Martingale argument which exploits the existence in our setup (and for specified $\mu \in \mathscr{P}(E)$ which will include $\{S_i^Y\}$) of a solution $\hat{f}_\mu$ to Poisson's equation, i.e. such that for any $x \in E$

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

Indeed, the first term on the RHS of (6.11) can be rewritten

$$\begin{aligned}(n+1)[S_n^X - S_n^\omega](f) = M_{n+1} \\ + \sum_{m=0}^n [\hat{f}_{S_{m+1}^X}(X_{m+1}) - \hat{f}_{S_m^X}(X_{m+1})] + \hat{f}_{S_0^X}(X_0) - \hat{f}_{S_{n+1}^X}(X_{n+1})\,,\end{aligned} \tag{6.12}$$

where

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

is such that $\{M_n, \mathcal{G}_n\}$ is a martingale. In our case, provided that K_μ is geometrically ergodic for example,

$$\hat{f}_\mu(x) = \sum_{n \in \mathbb{N}_0} [K_\mu^n(f)(x) - \omega(\mu)(f)]\,, \tag{6.13}$$

can be show to be a solution to Poisson's equation. This will hold $\mathbb{P}_x$ or $\mathbb{Q}_y-$a.s. for $\mu = S_i^Y$ or $\mu = S_i^X$ under our assumptions; see Proposition 4.1. For (NL3), due to the stochastic drift condition in Proposition 4.1, a slightly different approach is adopted. A quantitative bound on

the solution of the Poisson equation in Glynn & Meyn (1996) is derived in the appendix and used in an $\mathbb{L}_p-$bound on this function.

We will seek to bound the Martingale, M_n, in $\mathbb{L}_p$ via the Burkholder-Gundy-Davis inequality (e.g. Burkholder (1973)) and the fluctuations of the solution to Poisson's equation via Lipschitz continuity of either Φ (NL1-2) or the nonlinear kernel Q_μ (NL3) (see Del Moral (2004) for an account of Lipschitz continuity of semi-groups and Markov kernels).

6.2. $\{M_m\}$ **is $\mathbb{L}_p$-bounded.** We first establish uniform in time $\mathbb{L}_p$-bounds of the solution to Poisson's equation, and then similarly establish bounds on the sequence $\{M_n\}$.

Proposition 6.1. (NL1) *Assume (A1) and (A2-i-ii-iii-iv), let $r \in [0,1)$ and $p \in [1, 1/r)$. Then there exists $M(\cdot) < \infty$ such that for any $f \in \mathscr{L}_{V^r}$ and any $m \in \mathbb{N}_0$,*

$$\mathbb{E}_x\left[|\hat{f}_{S_m^X}(X_{m+1})|^p\right]^{1/p} \leq M(r, M_1^*, \mathbb{G})|f|_{V^r}V(x)^r \ .$$

(NL3) *Assume (A2-a-i-ii-iii-iv), let $r \in (0, 1/3)$ and $p \in [1, 1/3r)$. Then there exists $M(\cdot) < \infty$ such that for any $f \in \mathscr{L}_{V^r}$ and any $m \in \mathbb{N}_0$,*

$$\mathbb{E}_{(x,y)}[|\hat{f}_{S_m^Y}(X_{m+1})|^p]^{1/p} \leq M(\epsilon, \alpha, r, \mathbb{D}, \mathbb{D}_y)V(x)W(y).$$

Proposition 6.2. (NL1) *Assume (A1-2), let $r \in (0,1)$ and $p \in [1, 1/r)$. Then there exists $M(\cdot) < \infty$ such that for any $f \in \mathscr{L}_{V^r}$ and any $m \in \mathbb{N}$,*

(NL1)

$$\mathbb{E}_x\left[|M_m|^p\right]^{1/p} \leq m^{\frac{1}{2}} M(p, r, M_1^*, \mathbb{G})|f|_{V^r}V(x)^r \ .$$

(NL3) *Assume (A2-a-i-ii-iii-iv), let $r \in (0, 1/2)$ and $p \in [1, 1/2r)$. Then there exists $M(\cdot) < \infty$ such that for any $f \in \mathscr{L}_{V^r}$ and any $m \in \mathbb{N}_0$,*

$$\mathbb{E}_{(x,y)}\left[|M_m|^p\right]^{1/p} \quad \leq \quad m^{\frac{1}{2}} M(p, r, \epsilon, \alpha, \mathbb{D}, \mathbb{D}_y)V(x)W(y).$$

6.3. **Bounding the variations of the solution to Poisson's equation.** Finally we establish uniform in time $\mathbb{L}_p$-bounds on the fluctuations of the solution of the Poisson equation $\{\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})\}$ (resp. $\{\hat{f}_{S_{m+1}^X}(X_{m+1}) - \hat{f}_{S_m^X}(X_{m+1})\}$) which are the result of the evolution of the empirical measures $\{S_m^Y\}$ (resp. $\{S_m^X\}$). For notational simplicity we will use the notation

$$\hat{f}(S_{m:m+1}, X_{m+1}) := \hat{f}_{S_{m+1}}(X_{m+1}) - \hat{f}_{S_m}(X_{m+1}) \ , \tag{6.14}$$

for a sequence of generic empirical measures $\{S_m\}$.

Proposition 6.3. *Assume (A1-2-i-ii-iii) and*

(NL1) *let $r \in (0,1)$ and $p \in (1, 1/r)$. Then there exists $M(\cdot) < \infty$ such that for any $x \in E$, $f \in \mathscr{L}_{V^r}$ and $m \in \mathbb{N}_0$,*

$$\mathbb{E}_x\left[|\hat{f}(S^X_{m:m+1}, X_{m+1})|^p\right]^{1/p} \leq \frac{M(r, M_1^*, \mathbb{G})|f|_{V^r} V(x)^r}{m+2} .$$

with W as in (A2).

(NL3) *let $r \in [0, 1/2)$ then for any $x, y \in E$, $f \in \mathscr{L}_{V^r}$, $m \in \mathbb{N}_0$,*

$$\lim_{m\to\infty} |\hat{f}_{S^Y_{m+1}}(X_{m+1}) - \hat{f}_{S^Y_m}(X_{m+1})| = 0$$

$\mathbb{P}_{(x,y)}-a.s.$

6.4. **Main Results.** We now combine the above results to prove the SLLN for (NL1). In this case we give an $\mathbb{L}_p$-bound for $[S^X_n - S^\omega_n](f)$ and use it to establish a SLLN.

6.4.1. *Case (NL1).*

Theorem 6.4. *Assume (A1-2). Then there exists $M^\star < 1$ such that for any $r \in (0, 1/4 \wedge \widetilde{r})$ with $\widetilde{r} = \frac{1}{2} \wedge (1 - M^\star)$ and $p \in [1, 1/r - 1)$ such that there exists $M(\cdot) < \infty$ so that for any $f : E \to \mathbb{R}$ satisfying $|f| \leq V^r$,*

$$\mathbb{E}_x\left[|[S^X_n - S^\omega_n](f)|^p\right]^{1/p} \leq \frac{M(r, p, M_1^*, \mathbb{G})V(x)^r}{(n+1)^{\frac{1}{2}}} .$$

In addition we have

$$S^X_n(f) \xrightarrow{a.s}_{\mathbb{P}_x} \pi(f) .$$

6.4.2. *Case (NL3).* For (NL3) we have the following convergence result.

Theorem 6.5. *Assume (A2-a-3). Let $r \in [0, 1/4)$. Then for any $f \in \mathscr{L}_{V^r}$,*

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

7. A Practical Self-Interacting Algorithm

We now introduce an algorithm where the empirical measure driving the self-interactions is constructed using population-based MCMC algorithms. Our intention is to run a population of chains (for NL2-3) that admit η as a marginal and use the samples corresponding to η to construct the empirical measure. This is selected to minimise the storage of the algorithm compared to equi-energy samplers and the coding effort required by the potential users. Note that we have found that the convergence proofs for such algorithms are similar to those in the previous Sections and have thus omitted them.

7.1. **Population MCMC.** We are interested in simulating from a probability measure $\psi_1 \in \mathscr{P}(E)$. The first step is often to use an MCMC kernel K. However, as noted in the introduction, this may not always work well. A simple way (from a coding point of view) to improve the exploration of the state-space is to use population MCMC.

Consider $(E^{\tilde{m}}, \mathcal{E}^{\otimes \tilde{m}})$ and a sequence of related probability measures $\{\psi_i\}_{i\in\{2,\ldots,\tilde{m}\}}$ easier to simulate than ψ_1. The idea of population MCMC is to build a time-homogeneous Markov kernel $Q : E^{\tilde{m}} \to \mathscr{P}(E^{\tilde{m}})$ of invariant measure:

$$\psi^{\tilde{m}}(d(x_1, \ldots, x_{\tilde{m}})) \quad = \quad \prod_{i=1}^{\tilde{m}} \psi_i(dx_i).$$

The intuition is that the easier to simulate $\psi_2, \ldots, \psi_{\tilde{m}}$ can provide information on ψ_1 and that we can use this information to produce a faster mixing Markov kernel Q with quicker convergence to $\psi^{\tilde{m}}$ than K for ψ_1.

The population kernel adopted in this paper is:

$$P(y_{n-1}^{(\tilde{m})}, dy_n^{(\tilde{m})}) \quad = \quad \sum_{i=1}^{\tilde{m}} \beta_i \Big\{ \phi_i \delta_{y_{n-1}^{(-i)}}(dy_n^{(-i)}) P_i(y_{n-1}^i, dy_n^i) + (1-\phi_i) \times P_i^S(y_{n-1}^{(\tilde{m})}, dy_n^{(\tilde{m})}) \Big\} \tag{7.15}$$

where P_i is an MCMC kernel of invariant measure ψ_i, $\beta_i, \phi_i \in (0,1)$, $Y^{(\tilde{m})} = (Y^1, \ldots, Y^{\tilde{m}})$, $\sum_{i=1}^{\tilde{m}} \beta_i = 1$, $Y^{(-i)} = (Y^1, \ldots, Y^{i-1}, Y^{i+1}, \ldots, Y^{\tilde{m}})$ and P_i^S is an exchange kernel that proposes to swap, with equal probability, the i^{th} population member with any other.

7.2. **Algorithms.** To use this idea for the self-interacting approximation (NL1) we introduce a non-linear kernel (with $\tilde{m} = m + 1$):

$$K_{S_{n-1}}((x_{n-1}, y_{n-1}^{(m)}), d(x_n, y_n^{(m)})) \quad = \quad (1-\epsilon) P((x_{n-1}, y_{n-1}^{(m)}), d(x_n, y_n^{(m)})) + \epsilon \Phi(S_{n-1})(d(x_n, y_n^{(m)}))$$

with $P : E \times E^m \to \mathscr{P}(E \times E^m)$ is a Markov kernel of invariant distribution $\psi_1 = \pi$, $\psi_i = \eta_{i-1} \in \mathscr{P}(E)$, $i = 2, \ldots, \widetilde{m}$, $S_n \in \mathscr{P}(E^{\tilde{m}})$ is the empirical measure, $\Phi : \mathscr{P}(E^{\tilde{m}}) \to \mathscr{P}(E^{\tilde{m}})$ (that is, the selection/mutation operator on an extended space) and $g \equiv 1$.

Our algorithm for (NL1) is:

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

For (NL2-3), our objective is to use the population kernel to construct the empirical measure S_n^Y (here $\tilde{m} = m$). In this way, our algorithm can retain information (from the past) generated by the population kernel. We will see that this will be particularly effective when the population

kernel is slowly mixing. Below, $P : E^m \to \mathscr{P}(E^m)$ is a Markov kernel of invariant distribution $\eta^m = \eta_1 \times \cdots \times \eta_m$, i.e. $\psi_i = \eta_i$ for all i.

For (NL2) and (NL3) we use the algorithm (with the particular K_μ described in examples 2-3):

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

7.3. **Simulations.** We now present some simulations; we begin by describing the target, then the population kernels, simulation parameters and then the results.

7.3.1. *Target Measure.* We consider a sequence of probability measures on $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$:

$$\begin{aligned} \pi(dx) &= \frac{1}{Z}\exp\Big\{-h(x)\Big\}dx \\ \eta_i(dx) &= \frac{1}{Z_i}\exp\Big\{-\gamma_i h(x)\Big\}dx \\ h(x) &= -\log(f(x)) \\ f(x) &= \sum_{l=1}^{20} w_l \phi_2(x; \mu_l, \Sigma_l) \end{aligned}$$

where $i \in \mathbb{T}_4$, $1 > \gamma_1 > \cdots > \gamma_m > 0$, Z_i is the normalizing constant, $w_l = 1/20\ \forall l$, $\phi_2(\cdot; \mu, \Sigma)$ is the bivariate Gaussian density of mean μ and covariance Σ, which is assumed diagonal. We adopt the $\{\mu_l\}$ and $\{\Sigma_l\}$ used by Kou et al. (2006).

7.3.2. *Population Kernel.* For (NL2-3) our population kernel, P, is taken as:

$$P(y_{n-1}^{(m)}, dy_n^{(m)}) = \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{3}{4}\delta_{y_{n-1}^{(-i)}}(dy_n^{(-i)})P_i(y_{n-1}^i, dy_n^i) + \frac{1}{4}P_i^S(y_{n-1}^{(m)}, dy_n^{(m)})\right\}$$

with $m = 4$. For (NL1) we adopt a similar kernel with $m = 5$; see Section 7.2 for further details.

7.3.3. *Simulation Parameters.* Our objective is to demonstrate that our algorithms can improve the performance of a slowly mixing Markov kernel, for similar computational cost.

We take $\{\gamma_m\}$ to be equally spaced on $(0, 1)$ and the proposal variance in the RWM steps to yield acceptance rates around 0.3. The MCMC step in the non-linear kernels was taken as 200 iterates of the random walk/population kernel. It should be noted that the results for lower number of iterates of the Markov kernels are still quite reasonable on average. However, a few of the runs perform poorly emphasizing the fact that, as expected, combining slowly mixing kernel with self-interactions can be inefficient if the initial exploration of the target is quite poor.

7.3.4. *Results.* We ran the algorithms five times for 2 million iterations after burn-in (50000 iterations, that is, we did not use the selection step until this time), storing only 10000 samples (all results averaged over the chains). The chains were run for a similar CPU time, as reported in Table 1.

We estimated $\mathbb{E}[X]$ using the approximation S_{9999} for each non-linear MCMC method (using 5 different settings of ϵ); the results are in Table 1.

In Table 1 we can observe that the fully self-interacting approximation (NL1) has performed quite poorly for all ϵ; the estimates of the means become even more inaccurate as ϵ goes to 1. This is consistent with our assumption (A2) (v) which implies that we would want ϵ to be small. In addition, we can intuitively explain this poor performance as follows. Despite the fact that an approximation of π seems optimal, no property of π is used in the selection step and hence the algorithm suffers from very slow convergence properties.

Conversely, (NL2-3) both perform reasonably well, with quite similar parameter estimates for all values of ϵ. The algorithms (NL2-3) are able to avoid most of the difficulties of (NL1) because they rely on more sophisticated selection schemes (NL2-3) and an exchange step (NL3). This allows them to exploit the information of the empirical measure more efficiently.

A final point is that when we ran the population MCMC kernel to sample from the target for a similar CPU time (110sec), the results were significantly poorer than for the self-interacting algorithms with estimates of 4.00 for $\mathbb{E}[X_1]$ and 3.73 for $\mathbb{E}[X_2]$. This emphasizes the importance of self-interacting mechanisms.

7.3.5. *Discussion.* We have demonstrated that our algorithms improve the performance of slowly mixing population algorithms. This is of interest when population algorithms cannot be calibrated to perform satisfactorily (e.g. in the trans-dimensional case (Jasra et al. 2005)).

In experiments not reported here for fast mixing population algorithms, we have found that for (NL1-2) convergence speed was slowed down whereas for (NL3) convergence speed was similar. We attribute this to the fact that (NL3) is allowed (most efficiently) to exploit the information from the empirical measure through the exchange step. Note that these experimental results are consistent with the parametric (adaptive) case, as raised by Atchadé & Rosenthal (2005). It would be of interest to verify this theoretically; see Lemma 5.1 for some evidence for (NL2).

An important remark is that the selection step should only be used infrequently: In our experiments the MCMC kernel is iterated 200 times and a limited number of iterations significantly degrades performance. The reason is to ensure that we do not bias the algorithm too often, that is the empirical measure may be quite far from convergence. The selection step is ultimately (as

(NL1), example 1	True	$\epsilon = 0.05$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 1.0$
$\mathbb{E}[X_1]$	4.478	4.703	4.900	4.890	4.949	4.763
$\mathbb{E}[X_2]$	4.905	5.113	5.503	5.324	5.606	5.470
CPU (sec)		107	107	109	107	107
(NL2), example 2		4.318	4.391	4.332	4.327	4.399
		4.754	4.734	4.801	4.734	4.691
		108	110	109	114	118
(NL3), example 3		4.423	4.731	4.515	4.418	4.277
		4.628	5.124	4.933	4.936	4.554
		108	107	108	108	108

Table 1: Estimates from mixture comparison for Non-Linear MCMC. We ran each algorithm 5 times for 2 million iterations after a 50000 iteration burn-in and allowed the possibility of self-interaction every 200^{th} iteration.

$n \to \infty$) a draw from π, but we may not have such good performance for n finite. In addition, too many selection steps can make it difficult to diagnose poor performance of the algorithm; e.g. see our cautionary example in Section 3.1. Thus our recommendation would be to allow a relatively low number of selection steps, as demonstrated above. Due to the above discussion, we would recommend that (NL3) be used in complex scenarios.

8. Application

We end the paper with an application of our approach to a complicated statistical model in finance. We apply (NL3) (the population version) to the Bayesian analysis of continuous-time stochastic volatility models (Roberts et al. 2004). We compare our approach to a recently proposed technique in stochastic simulation; SMC samplers (Del Moral et al. 2006).

8.1. **Model.** The model is that of Roberts et al. (2004), which we briefly review. The data are the log-returns of an asset X_t at time $t \in [0, T]$ modelled via the stochastic differential equation (SDE):

$$dX_t = v_t^{1/2} dW_t$$

where $\{W_t\}_{t\in[0,T]}$ is standard Brownian motion The volatility v_t is modelled via the following SDE:

$$dv_t = -\mu\, v_t dt + dZ_t \tag{8.16}$$

where $\{Z_t\}_{t\in[0,T]}$ is a pure jumps Lévy process; see Applebaum (2004) for example.

It is well known (e.g. Applebaum 2004) that for any self-decomposable random variable, there exists a unique Lévy process that satisfies (8.16); we assume that v_t has a Gamma marginal, $\mathcal{G}a(\nu,\theta)$, where $\mathcal{G}a(a,b)$ is the Gamma distribution of mean a/b. In this case Z_t is a compound Poisson process:

$$Z_t = \sum_{j=1}^{K_t} \varepsilon_j$$

where $\{K_t\}_{t\in[0,T]}$ is a Poisson process of rate $\nu\mu$ and the $\{\varepsilon_j\}$ are i.i.d. random variables of distribution $\mathcal{E}x(\theta)$ (where $\mathcal{E}x$ is the exponential distribution). Denote the jump times of the compound Poisson process as $0 < c_1 < \cdots < c_{k_t} < t$.

Since $X_t \sim \mathcal{N}(0, v_t^*)$, where $v_t^* = \int_0^t v_s ds$ is the integrated volatility, it is easily seen that $Y_{t_i} \sim \mathcal{N}(0, v_i^*)$ with $Y_{t_i} = X_{t_i} - X_{t_{i-1}}$, $0 < t_1 < \cdots < t_u = T$ are regularly spaced observation times and $v_i^* = v_{t_i}^* - v_{t_{i-1}}^*$. Additionally, the integrated volatility is:

$$v_t^* = \frac{1}{\mu}\Big(\sum_{j=1}^{K_t}[1-\exp\{-\mu(t-c_j)\}]\varepsilon_j - v_0[\exp\{-\mu t\}-1]\Big)$$

To summarize the likelihood is:

$$f(y_{t_1:t_u}|\{v_t^*\}) = \prod_{i=1}^{u}\phi(y_{t_i}; v_i^*)$$

with $\phi(\cdot;a)$ the density of normal distribution of mean zero and variance a and the notation $x_{1:n} := (x_1,\ldots,x_n)$ is adopted. The priors are exactly as Roberts et al. (2004):

$$\begin{aligned} v_0|\theta,\nu &\sim \mathcal{G}a(\nu,\theta) \\ \nu &\sim \mathcal{G}a(\alpha_\nu,\beta_\nu) \\ \mu &\sim \mathcal{G}a(\alpha_\mu,\beta_\mu) \\ \theta &\sim \mathcal{G}a(\alpha_\theta,\beta_\theta) \end{aligned}$$

The density of the compound Poisson process is:

$$\begin{aligned} p_T(c_{1:k_T},\varepsilon_{1:k_T},k_T) = {} & \frac{k_T!}{T^{k_T}}\mathbb{I}_{\{0<c_1<\cdots<c_{k_T}<T\}}(c_{1:k_T})\theta^{k_T}\exp\{-\theta\sum_{j=1}^{k_T}\varepsilon_j\}\times \\ & \frac{(T\mu\nu)^{k_T}}{k_T!}\exp\{-T\mu\nu\}. \end{aligned}$$

8.2. **Simulation Parameters.** The MCMC kernel we used in our simulations is the CA (centered algorithm) of Roberts et al. (2004). It is demonstrated in that paper that such an algorithm does not always perform well. We will see that we are able to use this kernel and still obtain reasonable results in our simulations.

We ran (NL3) for 1.25 million iterations after a 2000 iteration burn-in. We ran a population MCMC kernel with 5 chains (with targets $\eta_i \propto f^{\gamma_i} p_T$) for the auxiliary process with temperatures starting at 0.99 and falling by 1/6. Selection was allowed to occur with $\epsilon = 0.5$ and the Markov kernel K is the CA algorithm iterated 250 times. For illustration we only use 5000 samples in our empirical measure.

For SMC samplers, we used the foward and backward kernels adopted in Section 4 of Del Moral et al. (2006) (with the difference of using the CA as the MCMC kernel). We ran the algorithm with the same coincidental proposal variances (in MH steps) to the non-linear MCMC algorithm. The sequence of targets was of the same functional form as for the population MCMC except we had 300 densities with a uniform cooling schedule (we found that the resampling schedule was quite reasonable in this case). We simulated 5000 particles.

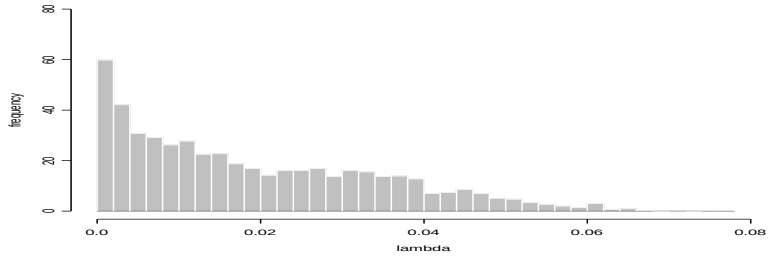
8.3. **Illustration.** Our data is part of the S&P 500 data found in Gander & Stephens (2007) (and kindly provided by Dr. M. Gander) which we standardized for analysis. The data consists of the daily share index returns at the opening of trading of the S&P 500. We reduced the data to only 500 observations, however, we found that for longer series, the algorithms still performed quite well. The prior parameters were as for Roberts et al. (2004) Section 4.

We ran both algorithms for approximately the same amount of CPU time and the estimates of the posterior of $\lambda = \mu\nu$ (as considered by Roberts el al. (2004)) can be seen in Figure 3. In Figures 3a and 3b, we can observe that both the non-linear MCMC scheme and SMC samplers yield rather similar results.

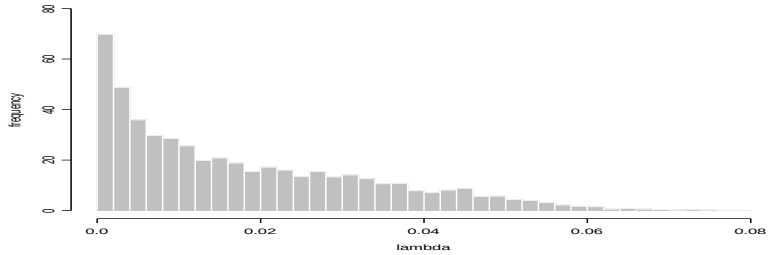
8.4. **Summary.** In this complex example we have seen that our non-linear MCMC method provides comparable results to SMC samplers. One advantage, however, of the non-linear approach against SMC samplers is the *iterative* nature of the procedure. In the above example, if we wanted to improve the estimates of quantities of interest we could simply run the sampler for longer. However, for SMC samplers we would not be able to do this.

9. Summary

We have investigated a new approach in stochastic simulation: Non-Linear MCMC via self-interacting approximations. We established convergence results for several algorithms. Further,



(a) SMC



(b) Non-Linear

Figure 3: Estimates of the posterior of $\lambda = \nu\mu$ for our stochastic volatility example.

we demonstrated the algorithms on a multimodal example, showing that the approach can drastically improve slowly mixing algorithms, but noting that it may not help in cases when an MCMC algorithm mixes quickly. As extensions to our ideas, we may consider the following.

Firstly, to relax conditions required to ensure the convergence of the algorithms. For example, Glynn & Meyn (1996) establish weaker than geometric ergodicity assumptions for the solution to the Poisson equation and functional central limit theorem (for Markov processes), in addition Jarner & Roberts (2002) establish 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 Roberts & Rosenthal (2006) for proofs for parametric adaptive MCMC).

Secondly, 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 rate associated to the unnormalized target distribution.

Thirdly, one aspect of our theoretical analysis that may seem unrealistic is the assumption (A1) for NL1. As noted in Andrieu & Robert (2001), adaptive MCMC algorithms (and hence NL1

in our case) have a direct link to stochastic approximation (SA) algorithms. (A1) can be thought of as an analogue to a boundedness assumption on the space of probability measures which can be compared to bounded parameter spaces in SA: the techniques required, in this case, to prove convergence are much simpler for SA algorithms (see Kushner & Yin (1997)). One way to deal with this difficulty (as used in Andrieu et al. (2005) (see also the references therein) and Andrieu & Moulines (2006) in the context of adaptive MCMC) is the approach of reprojections. Here, when the parameter escapes some compact set, the parameter is reinitialized to some compact set, and the compact set is enlarged. This idea may be pursued, in the non-linear MCMC case, to weaken the assumption (A1).

Acknowledgement. We would like to thank Éric Moulines for his comment pointing out that the convergence proofs for NL2 could be obtained via standard regeneration arguments from Markov chain theory. The second author would like to thank Matthew Gander for providing the data in Section 8. We also thank Adam Johansen for some useful comments on previous versions.

Appendix A. Main proofs

A.1. Common properties of K_μ.

Proof of Proposition 4.2. This is a direct application of Proposition 4.1 and Lemma B.1. □

A.2. Case NL1.

Proof of Proposition 4.1. The second and third statement of the proposition are a direct consequence of the first point from Meyn & Tweedie (1994), Theorem 2.3. The minorization property is direct from the expression for $K_\mu = \Pi_\mu$ in Eq. (2.3) and (A2-iii), with $\theta^* = (1-\epsilon) \times \theta$. We hence focus on the drift condition.

Let $\mu \in \mathscr{P}_{M_1}(E)$ and $x \in C_d$. Then from (A2-ii) and the expression for K_μ in Eq. (2.3) we have,

$$\begin{aligned} K_\mu V(x) &\leq \lambda(1-\epsilon)V(x) + b + \epsilon\lambda M_1 \\ &\leq \lambda(1-\epsilon)d + b + \epsilon\lambda M_1 \, . \end{aligned}$$

Now if $x \in C_d^c$ we have again from (A2)

$$\begin{aligned} K_\mu V(x) &\leq \left[\lambda(1-\epsilon) + \frac{\epsilon(\lambda M_1 + b)}{V(x)}\right] V(x) \\ &\leq \lambda^* V(x) \end{aligned}$$

with $\lambda^* < 1$ by the condition $d > \epsilon[\lambda M_1 + b]/(1 - \lambda(1-\epsilon))$ in (A2). To summarise,

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

with $b^* = \lambda(1-\epsilon)d + b + \epsilon\lambda M_1$, which completes the proof. □

Proof of Proposition 6.1. Let $r \in [0,1)$, $p \in [1, 1/r)$, $f \in \mathscr{L}_{V^r}$ and $m \in \mathbb{N}_0$. There exists $M(\cdot)$ as in Proposition 4.1 such that

$$\begin{aligned}
\mathbb{E}_x\big[|\hat{f}_{S_m^X}(X_{m+1})|^p\big]^{1/p} &\leq \sum_{n\in\mathbb{N}_0} \mathbb{E}_x\big[|K^n_{S_m^X}(f)(X_{m+1}) - \omega(S_m^X)(f)|^p\big]^{1/p} \\
&= \sum_{n\in\mathbb{N}_0} \mathbb{E}_x\left[\mathbb{E}_x\left(V(X_{m+1})^{pr}\frac{|K^n_{S_m^X}(f)(X_{m+1}) - \omega(S_m^X)(f)|^p}{V(X_{m+1})^{pr}}\middle|\mathcal{G}_m\right)\right]^{1/p} \\
&\leq M(M_1^*, \mathbb{G})|f|_{V^r}\left(\sum_{n\in\mathbb{N}_0}\rho^n\right)\mathbb{E}_x\big[V(X_{m+1})^{pr}\big]^{1/p}
\end{aligned}$$

where we have applied Minkowski's inequality and noted that, conditional upon $\mathcal{G}_m$, $K^n_{S_m^X}$ is a Markov kernel that is geometrically ergodic (via Proposition 4.1 which follows from (A1) and (A2)). Jensen's inequality and the condition on p yield

$$\mathbb{E}_x\big[|\hat{f}_{S_m^X}|^p\big]^{1/p} \leq M(M_1^*, \mathbb{G})|f|_{V^r}\left(\sum_{n\in\mathbb{N}_0}\rho^n\right)\mathbb{E}_x\big[V(X_{m+1})\big]^r.$$

Repeated application of the drift condition proved in Proposition 4.1 gives:

$$\mathbb{E}_x\big[V(X_{m+1})\big] \leq \left(1 + \frac{b}{1-\lambda}\right)V(x)$$

with λ and b as in (A2-ii). This completes the proof. □

Proof of Proposition 6.2. We follow a similar argument to that of Andrieu & Moulines (2006), Proposition 6. Throughout, we denote by B_p a generic constant dependent upon p only. We begin by applying the Burkholder-Gundy-Davis inequality (e.g. Burkholder (1973)) which yields:

$$\mathbb{E}_x\big[|M_n|^p\big]^{1/p} \leq B_p\mathbb{E}_x\big[(\sum_{m=0}^{n-1}[\hat{f}_{S_m^X}(X_{m+1}) - K_{S_m^X}(\hat{f}_{S_m^X})(X_m)]^2)^{p/2}\big]^{1/p}.$$

Let $p \in [2, 1/r)$; application of Minkowski's inequality leads to:

$$\mathbb{E}_x\big[|M_n|^p\big]^{1/p} \leq B_p\left(\sum_{m=0}^{n-1}\mathbb{E}_x\big[|\hat{f}_{S_m^X}(X_{m+1}) - K_{S_m^X}(\hat{f}_{S_m^X})(X_m)|^p\big]^{2/p}\right)^{1/2}.$$

Resorting to Minkowski's inequality again we obtain

$$\begin{aligned}
&\mathbb{E}_x\big[|\hat{f}_{S_m^X}(X_{m+1}) - K_{S_m^X}(\hat{f}_{S_m^X})(X_m)|^p\big] \\
&\qquad\qquad \leq \left(\mathbb{E}_x\big[|\hat{f}_{S_m^X}(X_{m+1})|^p\big]^{1/p} + \mathbb{E}_x\big[|K_{S_m^X}(\hat{f}_{S_m^X})(X_m)|^p\big]^{1/p}\right)^p.
\end{aligned}$$

From Proposition 6.1

$$\mathbb{E}_x\big[|\hat{f}_{S_m^X}(X_{m+1})|^p\big]^{1/p} \quad \leq \quad M(r, M_1^*, \mathbb{G})|f|_{V^r}V(x)^r$$

and via conditional Jensen, we have that

$$\begin{aligned}\mathbb{E}_x\big[|M_n|^p\big]^{1/p} &\leq M(p, r, M_1^*, \mathbb{G})\bigg(\sum_{m=0}^{n-1}|f|_{V^r}^2 MV(x)^{2r}\bigg)^{1/2} \\ &\leq n^{1/2}M(p, r, M_1^*, \mathbb{G})|f|_{V^r}V(x)^r\end{aligned}$$

If $p \in [1, 2)$ we have:

$$\begin{aligned}\mathbb{E}_x\big[|M_n|^p\big]^{1/p} &\leq B_p\bigg(\mathbb{E}_x\bigg[|2\sum_{m=0}^{n-1}\big[\hat{f}_{S_m^X}(X_{m+1})^2 + K_{S_m^X}(\hat{f}_{S_m^X})(X_m)^2\big]|^{p/2}\bigg]\bigg)^{1/p} \\ &\leq B_p\bigg(\mathbb{E}_x\bigg[|2\sum_{m=0}^{n-1}\big[\hat{f}_{S_m^X}(X_{m+1})^2 + K_{S_m^X}(\hat{f}_{S_m^X})(X_m)^2\big]|\bigg]\bigg)^{1/2} \\ &\leq B_p n^{1/2}M(r, M_1^*, \mathbb{G})|f|_{V^r}V(x)^r\end{aligned}$$

where we have applied $(a-b)^2 \leq 2[a^2+b^2]$, Jensen twice, Proposition 6.1 as well as conditional Jensen. The result thus follows. □

Proof of Proposition 6.3. Since:

$$\mathbb{E}_x\big[|\hat{f}(S_{m:m+1}^X, X_{m+1})|^p\big]^{1/p} =$$

$$\mathbb{E}_x\big[|\sum_{n=0}^{\infty}\big\{K_{S_{m+1}^X}^n(f)(X_{m+1}) - K_{S_m^X}^n(f)(X_{m+1}) - \omega(S_{m+1}^X)(f) + \omega(S_m^X)(f)\big\}|^p\big]^{1/p}$$

we may apply Proposition B.4 (in Appendix B), Minkowski and use the representation of the invariant measure (4.10) to yield:

$$\mathbb{E}_x\big[|\hat{f}(S_{m:m+1}^X, X_{m+1})|^p\big]^{1/p} \quad \leq \quad \sum_{n=0}^{\infty}\sum_{l=n+1}^{\infty}\epsilon(1-\epsilon)^{l-1}\mathbb{E}_x\big[|[\Phi(S_{m+1}^X) - \Phi(S_m^X)](K^{l-1}(f)|^p\big]^{1/p}.$$

Application of (A1) yields:

$$\begin{aligned}\mathbb{E}_x\big[|\hat{f}(S_{m:m+1}^X, X_{m+1})|^p\big]^{1/p} &\leq M(M_1^*, \mathbb{G})\sum_{n=0}^{\infty}\sum_{l=n+1}^{\infty}\epsilon(1-\epsilon)^{l-1}|K^l(f)|_{V^r} \times \\ &\qquad \mathbb{E}_x\big[|[S_{m+1}^X - S_m^X]\bigg(\frac{K^l(f)}{|K^l(f)|_{V^r}}\bigg)|^p\big]^{1/p} \\ &\leq M\sum_{n=0}^{\infty}\sum_{l=n+1}^{\infty}\epsilon(1-\epsilon)^{l-1}|K^l(f)|_{V^r}\mathbb{E}_x\big[||S_{m+1}^X - S_m^X||_{V^r}^p\big]^{1/p} \\ &\leq M|f|_{V^r}\mathbb{E}_x\big[||S_{m+1}^X - S_m^X||_{V^r}^p\big]^{1/p}\end{aligned}$$

where the fact that $|K^l(f)|_{V^r} \leq |f|_{V^r}M$ and that the double sum is equal to a finite constant has been used.

Noting that for any f we have:

$$\mathbb{E}_x\big[|S^X_{m+1}(f) - S^X_m(f)|\big] \quad = \quad \frac{1}{m+2}\mathbb{E}_y\big[|f(x_{m+1}) - S^X_m(f)|\big]$$

and thus

$$\mathbb{E}_x\big[||S^X_{m+1} - S^X_m||_{V^r}\big] \quad \leq \quad \frac{1}{m+2}\mathbb{E}_x\big[(V(X_{m+1})^r + S^X_m(V^r))\big] \tag{A.17}$$

$|f| \leq V^r$ has been used. The proof may be completed by using the drift condition established in Proposition 4.1. □

Proof of Theorem 6.4. We begin by noting, for the $\widetilde{M}$ in (A2) (v) we have:

$$\begin{aligned} M^\star &= \widetilde{M}\sum_{l\in\mathbb{N}}\epsilon(1-\epsilon)^{l-1}\rho^l \\ &= \frac{\widetilde{M}\rho\epsilon}{1-\rho(1-\epsilon)} < 1 \end{aligned} \tag{A.18}$$

by (A2) (v).

It is straightforward to establish that for $f \in \mathscr{L}_{V^r}$:

$$\mathbb{E}_x\big[|S^X_n(f) - S^\omega_n(f)|^p\big]^{1/p} \quad \leq \quad \frac{|f|_{V^r}B_pV(x)^r}{(n+1)^{\frac{1}{2}}}$$

Noting (6.12); in Proposition 6.2 we bounded M_n in $\mathbb{L}_p$ and in Proposition 6.3 the fluctuations due to the evolution of the empirical measure. Also, we note that our assumptions ensure the existence of the solution to the Poisson equation, so we need not worry about $\hat{f}_{S^Y_0}(X_0) - \hat{f}_{S^Y_{n+1}}(X_{n+1})$. Consider:

$$\begin{aligned} \mathbb{E}_x\big[|\sum_{m=0}^n[\hat{f}_{S^X_{m+1}}(X_{m+1}) - \hat{f}_{S^X_m}(X_{m+1})]|^p\big]^{1/p} &\leq \sum_{m=0}^n \frac{M(r,M_1^*,\mathbb{G})|f|_{V^r}}{m+2}V(x)^r \\ &\leq M(r,M_1^*,\mathbb{G})|f|_{V^r}\log(n+2)V(x)^r \end{aligned}$$

where we have used Minkowski and bounded the sum with an integral. Straightforward manipulations give:

$$\mathbb{E}_x\big[|S^X_n(f) - S^\omega_n(f)|^p\big]^{1/p} \quad \leq \quad \frac{M(r,p,M_1^*,\mathbb{G})|f|_{V^r}V(x)^r}{(n+1)^{\frac{1}{2}}}.$$

To establish an $\mathbb{L}_p$-bound on $S^X_n(f) - \pi(f)$ we follow the proof of Proposition 4.2 of Del Moral & Miclo (2004). To fix some conventions, a sequence (a_n) is said to be of rate $\widetilde{r}$ if:

$$\widetilde{r} \quad = \quad \limsup_{n\to\infty}\frac{\log(a_n)}{\log(n)}.$$

Also define:

$$I^{(p)}_n \quad = \quad \sup_{|f|\leq V^r}(n+1)^p\mathbb{E}_x\big[|[S^\omega_n - \pi](f)|^p\big].$$

As in Del Moral & Miclo (2004), we seek to establish that $I_n^{(p+1)}$ is of rate $(1-\widetilde{r})(p+1)$ with $\widetilde{r}=\frac{1}{2}\wedge(1-M^\star)$, then $\mathbb{E}_x\left[|S_n^X-S_n^\omega|^p\right]^{1/p}$ is of rate $-1/2$; we may follow the proofs of Proposition 4.2 and Theorem 4.3 of Del Moral & Miclo (2004).

Let $p\in[2,1/r-1)$ and assume the hypothesis $I_n^{(p)}$ is of rate $(1-\widetilde{r})p$ (the initialization is discussed below) then as in the proof of Proposition 4.2 of Del Moral & Miclo (2004) we can reach equation (4.1) of that paper and we need only deal with, for $|f|\leq V^r$:

$$(p+1)\mathbb{E}_x\left[(n+1)^p|[S_n^\omega-\pi](f)|^p|[\omega(S_{n+1}^X)-\pi](f)|\right].$$

Then we have that:

$$(p+1)\mathbb{E}_x\left[(n+1)^p|[S_n^\omega-\pi](f)|^p|[\omega(S_{n+1}^X)-\pi](f)|\right]\leq$$

$$(p+1)\mathbb{E}_x\Big[(n+1)^p|[S_n^\omega-\pi](f)|^p\big(|[\omega(S_{n+1}^X)-\omega(S_{n+1}^\omega)](f)|+ \tag{A.19}$$

$$|[\omega(S_{n+1}^\omega)-\omega(S_n^\omega)](f)|+|[\omega(S_n^\omega)-\omega(\pi)](f)|\big)\Big]. \tag{A.20}$$

We deal with each of the three terms separately. The latter term is dealt with via Hölder:

$$(p+1)\mathbb{E}_x\left[(n+1)^p|[S_n^\omega-\pi](f)|^p|[\omega(S_n^\omega)-\omega(\pi)](f)|\right]\leq$$

$$\frac{p+1}{n+1}\sum_{l\in\mathbb{N}}\epsilon(1-\epsilon)^{l-1}\mathbb{E}_x\left[(n+1)^p|[S_n^\omega-\pi](f)|^p(n+1)|[S_n^\omega-\pi](K^l(f))|\right]\leq$$

$$\frac{p+1}{n+1}(I_n^{(p+1)})^{p/p+1}\sum_{l\in\mathbb{N}}\epsilon(1-\epsilon)^{l-1}|[K^l-\pi](f)|_{V^r}\mathbb{E}_x\left[(n+1)^{p+1}\Big|[S_n^\omega-\pi]\left(\frac{[K^l-\pi](f)}{|[K^l-\pi](f)|_{V^r}}\right)\Big|^{p+1}\right]^{1/p+1}\leq$$

$$\frac{p+1}{n+1}I_n^{(p+1)}M^\star$$

with $M^\star<1$ from equation (A.18). The first two terms can be bounded via using the inequality, for any $\phi>0$:

$$\forall x,y\geq 0\quad x^py\leq\phi x^{p+1}+A(\phi,p)y^{p+1}$$

with $A(\phi,p)$ a constant dependent upon p,ϕ.

Now consider (A.19); denoting $K(V^r,\pi,l)=\frac{[K^l-\pi](f)}{|[K^l-\pi](f)|_{V^r}}$ we obtain in a similar manner to the above manipulations (recall that $M^\star$ is defined in (A.18)):

$$(p+1)\mathbb{E}_x\left[(n+1)^p|[S_n^\omega-\pi](f)|^p|[\omega(S_{n+1}^X)-\omega(S_{n+1}^\omega)](f)|\right]\leq$$

$$(p+1)\sum_{l\in\mathbb{N}}\epsilon(1-\epsilon)^{l-1}\widetilde{M}\rho^l\mathbb{E}_x\left[(n+1)^p|[S_n^\omega-\pi](f)|^p|[S_{n+1}^X-S_{n+1}^\omega](K(V^r,\pi,l))|\right]\leq$$

$$(p+1)\sum_{l\in\mathbb{N}}\epsilon(1-\epsilon)^{l-1}\widetilde{M}\rho^l\left[\frac{\phi}{n+1}I_n^{(p+1)}+A_1(\phi,p)B_pV(x)^{r(p+1)}(n+1)^{\frac{1}{2}(p-1)}\right]\leq$$

$$(p+1)M^{\star}\left\{\frac{\phi}{n+1}I_n^{(p+1)}+A_1(\phi,p)B_pV(x)^{r(p+1)}(n+1)^{\frac{1}{2}(p-1)}\right\}$$

where we have used the $\mathbb{L}_p$ bound on $S_n^X(f)-S_n^{\omega}(f)$ and noted that $|K(V^r,\pi,l)|_{V^r}\leq 1$ and $p\geq 1$ (this latter point is needed in the initialization).

Consider the first part of (A.20), using the above manipulations we have:

$$(p+1)\mathbb{E}_x\left[(n+1)^p|[S_n^{\omega}-\pi](f)|^p|[\omega(S_{n+1}^{\omega})-\omega(S_n^{\omega})](f)|\right]\leq$$

$$(p+1)\sum_{l\in\mathbb{N}}\epsilon(1-\epsilon)^{l-1}\widetilde{M}\rho^l\left[\frac{\phi}{n+1}I_n^{(p+1)}+\frac{A_2(\phi,p)}{n+1}\mathbb{E}_x\left[(n+1)^{p+1}|[S_{n+1}^{\omega}-S_n^{\omega}](K(V^r,\pi,l))|^{p+1}\right]\right].$$

Since $|[S_{n+1}^{\omega}-S_n^{\omega}](f)|=\frac{1}{n+1}|\omega(S_{n+1})(f)-S_n^{\omega}(f)|$, we derive the following property, for $|f|\leq V^r$:

$$\mathbb{E}_x\left[\omega(S_{n+1})(|f|)^{p+1}\right] \quad\leq\quad M\mathbb{E}_x\left[S_{n+1}(V^r)^{p+1}\right]$$

as $S_{n+1}(K^l(|f|))\leq MS_{n+1}(V^r)$ (drift condition) and we have used the representation of the invariant measure (4.10). Now applying Minkowski's inequality, we have for any $|f|\leq V^r$:

$$\begin{aligned}\mathbb{E}_x\left[S_{n+1}^{\omega}(|f|)^{p+1}\right] &\leq \frac{M}{(n+2)^{p+1}}\Big(\sum_{i=0}^{n+1}\mathbb{E}_x[V(X_i)^{r(p+1)}]^{\frac{1}{p+1}}\Big)^{p+1}\\ &\leq MV(x)^{r(p+1)}\end{aligned}$$

via Jensen. Due to the above arguments:

$$\begin{aligned}(p+1)\mathbb{E}_x\left[(n+1)^p|[S_n^{\omega}-\pi](f)|^p|[\omega(S_{n+1}^{\omega})-\omega(S_n^{\omega})](f)|\right] &\leq (p+1)M^{\star}\left\{\frac{\phi}{n+1}I_n^{(p+1)}+\right.\\ &\left.\frac{A_2(\phi,p)MV(x)^{r(p+1)}}{n+1}\right\}\end{aligned}$$

and note that the latter expression on the RHS is of rate -1. Thus by the proof of Del Moral & Miclo (2004) we have the desired rate for $I_n^{(p+1)}$. The initialization, for $p=1,2$ can be performed by the above manipulations. The proof is then completed in a similar manner to Propositions 4.2 and Theorem 4.3 of Del Moral & Miclo (2004) and is thus omitted. □

A.3. **Case NL3.**

Proof of Proposition 4.1. The second and third statement of the proposition are a direct consequence of the first point from Meyn & Tweedie (1994), Theorem 2.3. The minorization property is direct from the expression for K_{μ} in Eq. (2.3) and (A2-iii) with $\theta'_{d(\mu)}=(1-\epsilon)\times\theta_{d(\mu)}$. Let us focus on the drift condition.

It is straightforward to prove, for any $x\in E$:

$$K_{\mu}V(x)\leq\lambda V(x)+\bar{b}(\mu).$$

Now, let $x \in C^c_{d(\mu)}$, then clearly, via conditional Jensen:

$$K_\mu V^r(x) \leq \left(\lambda + \frac{\bar{b}(\mu)}{d(\mu)}\right)^r V(x)^r$$

and since $d(\mu) \geq \bar{b}(\mu)\alpha/(1-\lambda)$:

$$K_\mu V^r(x) \leq \tilde{\lambda}^r V(x)^r.$$

Suppose further that $x \in C_{d(\mu)}$, then:

$$K_\mu V^r(x) \leq (\lambda d(\mu) + \bar{b}(\mu))^r.$$

As a result:

$$K_\mu V^r(x) \leq \tilde{\lambda}^r V(x)^r + \tilde{b}(\mu)^r \mathbb{I}_{C_{d(\mu)}}(x).$$

as required.

□

Proof of Proposition 4.3. The proof is given for $r = 1$ only. Let $|f| \leq V$:

$$|[Q_\mu - Q_\xi](f)(x)| = \Big| \int_E [\mu - \xi](du)[\alpha(x,u)\{K(f)(u) - K(f)(x)\}] \Big|.$$

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

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

i.e.

$$\alpha(x,u)\{K(f)(u) - K(f)(x)\} \leq 2(\lambda + b)]V(u)V(x).$$

As a result:

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

and then the result easily follows.

□

Proof of Proposition 6.1. The proof begins by conditioning upon the filtration $\mathcal{G}^Y$ generated by the auxiliary process $\{Y_n\}$ then, via Proposition 4.1 (where $\tilde{\lambda}$ is defined and a V^r drift condition is proved), applying Lemma B.3 followed by Minkowski's inequality to yield

$$\mathbb{E}_{(x,y)}[|\hat{f}_{S^Y_m}(X_{m+1})|^p]^{1/p} \leq \Big\{(1+\tilde{\lambda}^r)\mathbb{E}_x[|\bar{f}_{S^Y_m}|^p_{V^r} V(X_{m+1})^{rp}]^{1/p} \tag{A.21}$$

$$+\mathbb{E}_y[|\bar{f}_{S^Y_m}|^p_{V^r}\left(\frac{(1-\bar{\theta}_{d(S^Y_m)})b'(S^Y_m)}{\theta_{d(S^Y_m)} - \bar{\theta}_{d(S^Y_m)}}\right)^p]^{1/p}\Big\} \tag{A.22}$$

where $\bar{f}_{S^Y_m} := f - \omega(S^Y_m)(f)$. Note that

$$b'(S^Y_m) = \nu_{d(S^Y_m)}(V^r) \vee \check{b}(S^Y_m)$$

with $\bar{\epsilon}$ (as in Lemma B.3) equal to $\bar{\theta}_{d(S_m^Y)}$ for any $\bar{\theta}_{d(S_m^Y)} \in (0, \theta_{d(S_m^Y)})$

$$\check{b}(S_m^Y) = \frac{\tilde{\lambda}^r d(S_m^Y)^r + \tilde{b}^r(S_m^Y) - \bar{\theta}_{d(S_m^Y)} \nu_{d(S_m^Y)}(V^r)}{1 - \bar{\theta}_{d(S_m^Y)}} .$$

We first establish some intermediate results that are used later in the proof. Using the drift condition, it can be seen that almost surely

$$|\bar{f}_{S_m^Y}|_{V^r} \leq |f|_{V^r}[1 + \frac{\tilde{b}(S_m^Y)^r}{1 - \tilde{\lambda}^r}]$$

where $\tilde{b}(S_m^Y)$ is defined in Proposition 4.1. In order to bound the expectation of this term we seek an upper bound of $\mathbb{E}_y[\bar{b}(S_m^Y)]$ which in turn, since $\tilde{b}(S_m^Y)$ is a linear function of $\bar{b}(S_m^Y) = b + \epsilon[\lambda S_m^Y(V) + b]$, requires one to bound

$$\mathbb{E}_y[\bar{b}(S_m^Y)] = b + \epsilon[\lambda \mathbb{E}_y[S_m^Y(V)] + b] .$$

Using that $V \in \mathscr{L}_W$ and applying the drift for P:

$$\mathbb{E}_y[\bar{b}(S_m^Y)] \leq b + \epsilon[M(\mathbb{D}_y)W(y) + b]$$

that is, there exists a finite $M(\epsilon, \mathbb{D}, \mathbb{D}_y)$ such that

$$\mathbb{E}_y[\bar{b}(S_m^Y)] \leq M(\epsilon, \mathbb{D}, \mathbb{D}_y)W(y) . \tag{A.23}$$

Applying the Cauchy-Schwarz inequality to the first term in the upper bound in (A.21) applying Jensen's inequality since by assumption $2pr < 1$ and using (A.23) we just require an upper-bound on the term

$$\mathbb{E}_{(x,y)}[V(X_{m+1})^{2pr}]^{1/2p}.$$

We again use Jensen's inequality followed by the drift inequality (for K_μ associated to V, with parameters $\tilde{\lambda}$ and $\tilde{b}$) and obtain

$$\left(\tilde{\lambda}^{m+1} V(x) + \sum_{j=1}^{m+1} (\tilde{\lambda})^{m+1-j} \mathbb{E}_y[\tilde{b}(S_j^Y)] \right)^r .$$

We therefore focus on

$$\mathbb{E}_y[\tilde{b}(S_m^Y)] \leq \lambda \mathbb{E}_y[d(S_m^y)] + M(\mathbb{D}, \mathbb{D}_y)W(y)$$

which is obtained through the definition of $\tilde{b}(\cdot)$ and (A.23). Using (A.23) and the definition of $d(\mu)$ in Proposition 4.1 it is clear that

$$\begin{aligned} \mathbb{E}_y[\tilde{b}(S_m^Y)] &\leq \lambda[1 + \frac{\alpha}{1-\lambda} M(\epsilon, \mathbb{D}, \mathbb{D}_y)W(y)] + M(\epsilon, \mathbb{D}, \mathbb{D}_y)W(y) \\ &\leq M(\epsilon, \alpha, r, \mathbb{D}, \mathbb{D}_y)W(y) . \end{aligned}$$

Consequently for all $m \geq 0$,

$$\mathbb{E}_{(x,y)}[V(X_{m+1})^{2pr}]^{1/2p} \leq M(\epsilon, \alpha, \mathbb{D}, \mathbb{D}_y)V(x)W(y)\,.$$

We now consider the second term in the RHS term of (A.21). Again from the application of the Cauchy-Schwarz inequality and (A.23) we focus on

$$\mathbb{E}_y\left[\left(\frac{(1-\bar{\theta}_{d(S_m^Y)})b'(S_m^Y)}{\theta_{d(S_m^Y)} - \bar{\theta}_{d(S_m^Y)}}\right)^{3p/2}\right]^{2/3p}.$$

Since we can set $\bar{\theta}_{d(S_m^Y)} = \theta_{d(S_m^Y)} - \varphi_{d(S_m^Y)}$, and apply (A2-iv), we need only concentrate upon $b'(S_m^Y)$:

$$\mathbb{E}_y\left[b'(S_m^Y)^{3p}\right]^{1/3p}.$$

Clearly:

$$\mathbb{E}_y\left[b'(S_m^Y)^{4p}\right]^{1/3p} \leq \mathbb{E}_y[\nu_{d(S_m^Y)}(V^r)^{3p}]^{1/3p} + \mathbb{E}_y[\check{b}(S_m^Y)^{3p}]^{1/3p}$$

For the first expectation, we have, for some constant $M < \infty$:

$$\mathbb{E}_y[\nu_{d(S_m^Y)}(V^r)^{3p}]^{1/3p} \quad \leq \quad \mathbb{E}_y[d(S_m^Y)^{3pr}]^{1/3p}$$

(note that, from (A2-iv) $\nu_d(C_d) = 1 \Rightarrow \nu_d(V) \leq d$) by the above argument (for bounding d) we have:

$$\mathbb{E}_y[\nu_{d(S_m^Y)}(V^r)^{3p}]^{1/3p} \quad \leq \quad M(\epsilon, \alpha, \mathbb{D}, \mathbb{D}_y)W(y).$$

In addition, these arguments can be adopted for $\mathbb{E}_y[\check{b}(S_m^Y)^p]^{1/p}$ and thus:

$$\mathbb{E}_y\left[b'(S_m^Y)^{3p}\right]^{1/3p} \leq M(\epsilon, \alpha, \mathbb{D}, \mathbb{D}_y)W(y).$$

□

Proof of Proposition 6.2. The proof is as for case (NL1) with only notational changes. □

Proof of Proposition 6.3. Our proof is based upon the decomposition of Proposition B.5 (in Appendix B) and then using the Lipschitz continuity properties proved in Propositions 4.2 and 4.3.

$$\begin{aligned} |\hat{f}_{S_{m+1}^Y}(X_{m+1}) - \hat{f}_{S_m^Y}(X_{m+1})| \quad = \quad & |\sum_{n\in\mathbb{N}_0}\sum_{i=0}^{n-1}[K^i_{S_{m+1}^Y} - \omega(S_{m+1}^Y)](K_{S_{m+1}^Y} - K_{S_m^Y}) \\ & [K^{n-i-1}_{S_m^Y} - \omega(S_m^Y)(f)(X_{m+1})] - \\ & \sum_{n\in\mathbb{N}_0}[\omega(S_{m+1}^Y) - \omega(S_m^Y)](K^n_{S_m^Y} - \omega(S_m^Y))(f))|. \end{aligned} \tag{A.24}$$

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

$$\|[K^n_{S_m^Y} - \omega(S_m^Y)](f)\|_{V^r} \leq M(r, S_m^Y, V, \mathbb{G})\rho(S_m^Y, V, \mathbb{G})^n$$

for some finite $M(S_m^Y, V, \mathbb{G})$ and $\rho(S_m^Y, V, \mathbb{G}) \in (0,1)$, it follows that (a.s.):

$$|[K^i_{S^Y_{m+1}} - \omega(S^Y_{m+1})](K_{S^Y_{m+1}} - K_{S^Y_m})[K^{n-i-1}_{S^Y_m} - \omega(S^Y_m)(f)(X_{m+1})]| \leq$$

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

Then, adopting the continuity result for K_{S_m}:

$$|||K_\mu - K_\lambda|||_{V^r} \leq 2(1-\epsilon)(\lambda + b)^r \|\mu - \lambda\|_{V^r}$$

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

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

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

$$\sum_{n,i} |[K^i_{S^Y_{m+1}} - \omega(S^Y_{m+1})](K_{S^Y_{m+1}} - K_{S^Y_m})[K^{n-i-1}_{S^Y_m} - \omega(S^Y_m)(f)(X_{m+1})]| \leq$$

$$M(r, S^Y_{m+1}, V, \mathbb{G})M(r, S^Y_m, V, \mathbb{G})\frac{V(X_{m+1})^r}{m+2}[V(Y_{m+1})^r + S^Y_m(V^r)].$$

As $\sup_m M(r, S^Y_{m+1}, V, \mathbb{G})M(r, S^Y_m, V, \mathbb{G})$ is a.s. finite, the geometric rate of convergence of the Markov chain and the bounds in Meyn & Tweedie (1994), $M(r, S^Y_{m+1}, V, \mathbb{G})M(r, S^Y_m, V, \mathbb{G})$ converges to a finite constant. In addition, by establishing an $\mathbb{L}_p-$bound for $\frac{V(X_{m+1})^r}{m+2}[V(Y_{m+1})^r + S^Y_m(V^r)]$, (using the techniques in the proof of Proposition 6.1) it follows by the first Borel-Cantelli lemma that the first term of the RHS of (A.24) goes to zero as $m \to \infty$ (a.s.).

Turning to the second expression of (A.24), the continuity of the invariant measure:

$$\|\omega(\mu) - \omega(\lambda)\|_{V^r} \leq M(r, \lambda, \mu, V, \mathbb{G})|||K_\mu - K_\lambda|||_{V^r}$$

and the kernel K_μ yields (a.s.):

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

application of the above arguments and the SLLN for Markov chains yields the desired result. □

Proof of Theorem 6.5. The Martingale is dealt with as in Theorem 6.4 and a Cesàro average argument for the fluctuations of the solution to the Poisson equation can also be used.

The difficulty is when considering the bias term $\bar{S}^\omega_n(f)$,

$$|S^\omega_n(f) - \pi(f)| = \frac{1}{n+1}|\sum_{i=0}^n [\omega(S^Y_i) - \omega(\eta)](f)|\,,$$

as $\omega(\eta) = \pi$ in our setup. In order to prove that this term vanishes, we establish pointwise or $\mathbb{Q}_y-$a.s. convergence to zero of $[\omega(S^Y_i) - \omega(\eta)](f)$ as $i \to \infty$ and invoke a Cesàro average argument to conclude.

Let $i, j \in \mathbb{N}$ and introduce the following upper bound

$$|[\omega(S_i^Y) - \omega(\eta)](f)| \leq |\omega(S_i^Y)(f) - K^j_{S_i^Y}(f)(x)| + |K^j_{S_i^Y}(f)(x) - K^j_\eta(f)(x)| + |K^j_\eta(f)(x) - \omega(\eta)(f)|. \tag{A.25}$$

We now consider the three terms. The first and third terms are easily dealt with (Proposition 4.1) and the fact that the geometric bounds are uniform (a.s.) in i. That is, for any fixed i, both terms go to zero $\mathbb{Q}_y-$a.s.. The proof is completed via Lemma B.2, the proof of Theorem 9 of Roberts et al. (1998) and using Cesàro averages.

□

A.4. **Convergence of the Marginals.**

Proof. Consider (NL3) (the proof of (NL1) is much the same except it uses the SLLN proof and is simpler) and introduce the following simple decomposition:

$$\begin{aligned} \left|\mathbb{E}_{(x,y)}\left[f(X_k) - \pi(f)\right]\right| \leq & \left|\mathbb{E}_{(x,y)}\left[f(X_k) - K^{n(k)}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)})\right]\right| + \\ & \left|\mathbb{E}_{(x,y)}\left[K^{n(k)}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)}) - \omega(S^Y_{k-n(k)})(f)\right]\right| + \\ & \left|\mathbb{E}_{(x,y)}\left[\omega(S^Y_{k-n(k)})(f) - \pi(f)\right]\right|. \end{aligned} \tag{A.26}$$

We will let $n(k) = k^\phi$, $\phi \in (0, 1/2)$. The proof is to adopt a dominated convergence argument.

From Lemma B.2 the third expression goes to zero as $k \to \infty$ (by the SLLN for $U-$statistics (as applied in Theorem 6.5) and some arguments below, for (NL1) by SLLN proof).

Consider:

$$\mathbb{E}[\omega(S_n^Y)(V)]$$

In order to apply the dominated convergence theorem for the third term, we note for $\mu(V^r) < \infty$ we have that:

$$\begin{aligned} \omega(\mu)(V^r) &= \omega(\mu)K_\mu(V^r) \\ &\leq \tilde{\lambda}^r \omega(\mu)(V^r) + \tilde{b}(\mu)^r \end{aligned}$$

that is:

$$\omega(\mu)(V^r) \leq \frac{\tilde{b}(\mu)^r}{1 - \tilde{\lambda}^r}.$$

That is, it is integrable (by the arguments of Proposition 6.1). A similar argument may be made for $K^{n(k)}_{S^Y_{k-n(k)}}(f)$ and we may apply the result of Proposition 4.1 to ensure that the second term goes to zero, via dominated convergence.

We now consider the first expression in the decomposition (A.26). To simplify the notation in the subsequent arguments, we adopt the following convention, for $k \leq j$:

$$K_{S^Y_{k:j}}(x, dy) := \int_{E^{j-k}} K_{S^Y_k}(x, dx_1) \dots K_{S^Y_j}(x_{j-k}, dy).$$

As in Haario et al.(2001) we adopt the following argument, for $k - n(k) \geq 1$:

$$\left| \mathbb{E}_{(x,y)} \left[f(X_k) - K^{n(k)}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)}) \right] \right| =$$

$$\left| \mathbb{E}_{(x,y)} \left\{ \mathbb{E}_{(x,y)} \left[f(X_k) - K^{n(k)}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)}) \middle| \mathcal{G}_{k-n(k)} \right] \right\} \right| =$$

$$\left| \mathbb{E}_{(x,y)} \Big\{ \sum_{j=0}^{n(k)-1} [K_{S^Y_{k-n(k):k-(j+1)}} K^j_{S^Y_{k-n(k)}}(f)(X_{k-n(k)}) - K_{S^Y_{k-n(k):k-(j+2)}} K^{j+1}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)})] \Big\} \right|.$$

We refer to the decomposition $|\sum_{j=0}^{n(k)-1} [K_{S^Y_{k-n(k):k-(j+1)}} K^j_{S^Y_{k-n(k)}}(f)(X_{k-n(k)}) - K_{S^Y_{k-n(k):k-(j+2)}} K^{j+1}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)})]|$ as $(*)$. We can adopt the following manipulations for $(*)$:

$$\Big| \sum_{j=0}^{n(k)-1} [K_{S^Y_{k-n(k):k-(j+1)}} K^j_{S^Y_{k-n(k)}}(f)(X_{k-n(k)}) - K_{S^Y_{k-n(k):k-(j+2)}} K^{j+1}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)})]| \leq$$

$$\sum_{j=0}^{n(k)-1} |K^j_{S^Y_{k-n(k)}}(f)|_{V^r} K_{S^Y_{k-n(k):k-(j+2)}} \left[\frac{V^r}{V^r} |(K_{S^Y_{k-(j+1)}} - K_{S^Y_{k-n(k)}})| \left(\frac{K^j_{S^Y_{k-n(k)}}(f)}{|K^j_{S^Y_{k-n(k)}}(f)|_{V^r}} \right) \right]$$

where we remark, conditional upon $\mathcal{G}^Y$ (the natural filtration of the auxiliary chain) all elements are finite a.s. The expression in the sum is bounded by:

$$|K^j_{S^Y_{k-n(k)}}|_{V^r} K_{S^Y_{k-n(k):k-(j+2)}}(V^r)(X_{k-n(k)}) \| |K_{S^Y_{k-(j+1)}} - K_{S^Y_{k-n(k)}}| \|_{V^r}.$$

By the drift condition:

$$|K^{j+1}_{S^Y_{k-n(k)}}|_{V^r} \leq \frac{\tilde{b}(S^Y_{k-n(k)})^r}{1 - \tilde{\lambda}^r}$$

and

$$K_{S^Y_{1:l}}(V^r)(x) \leq \tilde{\lambda}^{rl} V(x)^r + \sum_{j=1}^{l} (\tilde{\lambda})^{l-j} \tilde{b}(S^Y_j)^r.$$

In addition, by Proposition 4.3:

$$\| |K_{S^Y_{k-(j+1)}} - K_{S^Y_{k-n(k)}}| \|_{V^r} \leq 2(\lambda + b)^r \epsilon \| S^Y_{k-(j+1)} - S^Y_{k-n(k)} \|_{V^r}.$$

As a result, (*) is bounded by (a.s., call this (**)):

$$M(\epsilon, \alpha, \mathbb{D}) \tilde{b}(S^Y_{k-n(k)})^r \sum_{j=0}^{n(k)-1} \| S^Y_{k-(j+1)} - S^Y_{k-n(k)} \|_{V^r} \times$$

$$\left\{ \tilde{\lambda}^{rl} V(X_{k-n(k)})^r + \sum_{s=1}^{n(k)-j-1} (\tilde{\lambda})^{n(k)-j-1-s} \tilde{b}(S^Y_{s+k-n(k)-1})^r \right\}$$

Now consider the expectation:

$$\mathbb{E}_{(x,y)}[\tilde{b}(S^Y_{k-n(k)})^r V(X_{k-n(k)})^r \|S^Y_{k-(j+1)} - S^Y_{k-n(k)}\|_{V^r}].$$

Applying the Hölder inequality twice yields the bound:

$$\mathbb{E}_y[\tilde{b}(S^Y_{k-n(k)})^{3r}]^{1/3}\mathbb{E}_{(x,y)}[V(X_{k-n(k)})^{3r}]^{1/3}\mathbb{E}_y[\|S^Y_{k-(j+1)} - S^Y_{k-n(k)}\|^3_{V^r}]^{1/3}$$

by the arguments in the proof of Proposition 6.1 we know how to bound the first two terms. The third term can be dealt with as follows.

$$\begin{aligned}
&\mathbb{E}[\|S_{n(k)+k} - S_{n(k)}\|^3_{V^r}]^{1/3} \\
&\leq \sum_{j=0}^{k-1} \mathbb{E}[\|S^Y_{n(k)+j+1} - S^Y_{n(k)+j}\|^3_{V^r}]^{1/3} \\
&\leq \sum_{j=0}^{k-1} \frac{|V|_W}{n(k)+j+1}\left\{\mathbb{E}[W(Y_{n(k)+j+2})]^{1/3} + \mathbb{E}_y[S^Y_{n(k)+j}(W)]^{1/3}\right\} \\
&\leq \frac{kM(\mathbb{D}_y)W(y)}{n(k)+j+1}.
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}_{(x,y)}[\tilde{b}(S^Y_{k-n(k)})^r V(X_{k-n(k)})^r \|S^Y_{k-(j+1)} - S^Y_{k-n(k)}\|_{V^r}]. \\
&\leq M(\epsilon,\alpha,\mathbb{D},\mathbb{D}_y)\left\{\frac{n(k)-j-1}{k-n(k)+2}\right\}V(x)W(y)^{2+r}.
\end{aligned}$$

A similar argument may be applied to the second part of (**) thus we may conclude that:

$$\left|\mathbb{E}_{(x,y)}\left[f(X_k) - K^{n(k)}_{S^Y_{k-n(k)}}(f)(X_{k-n(k)})\right]\right| \leq M(\epsilon,\alpha,\mathbb{D},\mathbb{D}_y)\left\{\sum_{j=0}^{n(k)-1}\frac{n(k)-j-1}{k-n(k)+2}\right\}V(x)W(y)^{2+2r}.$$

The proof is completed by recalling that $n(k) = k^\phi$. □

Appendix B. Standard Technical Results on Markov chains

Lemma B.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 $i = 1,2$,*

$$P_i V(x) \leq \bar{\lambda} V(x) + \mathbb{I}_{\bar{C}}(x)\bar{b}\,, \tag{B.27}$$

$$P_i(x,A) \geq \mathbb{I}_{\bar{C}}(x)\bar{\epsilon}\bar{\nu}(A) \tag{B.28}$$

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

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

Proof. Let $r \in [0,1]$ and $f \in \mathscr{L}_{V^r}$. It is straightforward to determine the following decomposition (e.g. the proof of Proposition 3 of Andrieu & Moulines (2006)):

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

Consider, for any $|f| \leq V^r$:

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

$$\sum_{i=0}^{n-1} |[P_2^{n-i-1} - \pi_2](f)|_{V^r} \times P_1^i\left(\left| [P_1 - P_2]\left(\frac{[P_2^{n-i-1} - \pi_2](f)}{|[P_2^{n-i-1} - \pi_2](f)|_{V^r}} \right) \right| \right)$$

thus we have:

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

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:

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

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

Lemma B.2. *Consider (NL3) and Assume (A2-ii). Let $f \in \mathscr{L}_V$, $x \in E$ then:*

$$\lim_{i \to \infty} |K_{S_i^Y}^j(f)(x) - K_\eta^j(f)(x)| = 0 \, , \; \mathbb{Q}_\eta - a.s. \, . \tag{B.29}$$

Proof. First, consider the case $\epsilon = 1$; the general case is considered below. Our strategy is to express the iterates of $Q_{S_i^Y}$ in terms of a product empirical measure (Von-Mises statistic) of the feeding Markov chain and then to use the well-known link between Von-Mises statistics and $U-$statistics (e.g. Hoeffding (1948)) and then finally the strong law of large numbers for $U-$statistics for ergodic stochastic processes (Aaronson et al. 1996; Borovkova et al. 1999).

Let $\mu \in \mathscr{P}(E)$, then we will prove that:

$$Q_\mu^j(f)(x) = \mu^{\otimes j}(\tilde{K}^j(f))(x) \tag{B.30}$$

where $\tilde{K}((x,y),\cdot) = \alpha(x,y)K(y,\cdot) + [1-\alpha(x,y)]K(x,\cdot)$ and

$$\tilde{K}^j(f)(x,x_{1:j}) = \int_{E^{j+1}} \tilde{K}((x,x_1),dy_1)\tilde{K}((y_1,x_2),dy_2)\ldots\tilde{K}((y_{j-1},x_j),dy_j)f(y_j).$$

Since, for $j=1$, the result holds, assume for $j-1$, then:

$$\begin{aligned} Q_\mu^j(f)(x) &= Q_\mu^{j-1}(Q_\mu(f))(x) \\ &= \mu^{\otimes(j-1)}(\tilde{K}^{j-1}(\mu(\tilde{K}(f))))(x) \\ &= \mu^{\otimes j}(\tilde{K}^j(f))(x) \end{aligned}$$

where we have applied Fubini's Theorem. As a result of (B.30), we have:

$$|K_{S_n^Y}^j(f)(x) - K_\eta^j(f)(x)| = |[S_n^{\otimes j} - \eta^{\otimes j}](\tilde{K}^j(f))(x)|$$

where $\tilde{K}^j(f)(x,x_{1:j}) \in \mathscr{L}_V$ for any fixed x.

Now, introduce the $U-$ statistic (dropping the superscript Y):

$$S_n^{\odot j}(f) := \frac{1}{(n+1)_j} \sum_{\alpha \in <j,n+1>} f(Y_{\alpha(1)},\ldots,Y_{\alpha(j)})$$

where $<j,n+1>$ is the set of one-to-one mappings of $\mathbb{T}_j$ into $\mathbb{T}_{n+1}$ and $(n+1)_j = \frac{(n+1)!}{(n+1-j)!}$.

Now application of Theorem 5.1 of Grams & Serfling (1973) (note the assumptions of the Theorem are satisfied here, and the Markov structure does not invalidate the result) yields:

$$\lim_{n\to\infty} |S_n^{\otimes j}(f) - S_n^{\odot j}(f)| = 0 \ , \ \mathbb{Q}_\eta - \text{a.s.} \ .$$

Thus we require that:

$$\lim_{n\to\infty} |S_n^{\odot j}(f) - \eta^{\otimes j}(f)| = 0 \ , \ \mathbb{Q}_\eta - \text{a.s.} \ .$$

but this is a direct consequence of Theorem U of Aaronson et al. (1996) (as noted in Borovkova (1999) the result holds for polish spaces). An important remark is that we are able to apply the result as geometrically ergodic Markov chains are $\beta-$mixing with coefficient $O(\rho^n)$ and $P^j(f)$ is bounded by $V^{(j)}$ where:

$$V^{(j)} := \underbrace{V \otimes \cdots \otimes V}_{\text{j times}}.$$

In addition, by Proposition C.1 in Appendix C, that we do not require the auxiliary chain to be in initialized in stationarity.

To complete the proof for $\epsilon \in (0,1)$, we note the following decomposition for iterates of mixtures of Markov kernels K and P:

$$((1-\epsilon)K + \epsilon P)^n(x, dy) = \sum_{l=0}^{n} \epsilon^l (1-\epsilon)^{n-l} \sum_{(\alpha_1,\ldots,\alpha_n)\in\mathcal{S}_l} K^{1-\alpha_1}P^{\alpha_1}\ldots K^{1-\alpha_n}P^{\alpha_n}(x, dy).$$

where $\mathcal{S}_l = \{(\alpha_1,\ldots,\alpha_n) : \sum_{j=1}^n \alpha_j = l\}$; there is no difficulty to extend the result, using the dominated convergence theorem where required. □

Some simple results for the solution to the Poisson equation in Glynn & Meyn (1996) are now proved. Define, for any $A \in \mathcal{E}$, $\sigma_A = \inf\{n \geq 0 : X_n \in A\}$. Then for any Markov chain with an atom $\alpha \in \mathcal{E}$ the solution to the Poisson equation can be written:

$$\mathbb{E}_x\left[\sum_{i=0}^{\sigma_\alpha}[f(X_k) - \pi(f)]\right]$$

when this exists; write $\bar{f} = f - \pi(f)$ from herein. Our objective is to use the split-chain construction for a geometrically ergodic Markov chain to obtain explicit quantitative bounds on the solution of the Poisson equation (in terms of the parameters in the drift and minorization conditions). We have the following result.

Lemma B.3. *Let $(E, \mathcal{E})$ be a measurable space, $V : E \to [1,\infty)$, $b < \infty$, $\lambda, \epsilon \in (0,1)$, $1 < d < \infty$ and $C_d \in \mathcal{E}$, with $C_d = \{x : V(x) \leq d\}$. Then for any Markov transition probability $K : E \to \mathscr{P}(E)$ satisfying for $(x, A) \in E \times \mathcal{E}$,*

$$KV(x) \leq \lambda V(x) + \mathbb{I}_{C_d}(x)b\,,$$

$$K(x, A) \geq \mathbb{I}_{C_d}(x)\epsilon\nu(A)$$

, with $\nu(C_d) > 0$, then for any $f \in \mathscr{L}_V$:

$$\left|\bar{\mathbb{E}}_z\left[\sum_{i=0}^{\sigma_\alpha}[f(X_k) - \pi(f)]\right]\right| \leq |\bar{f}|_V\left[\big(1+\lambda\big)V(x) + \frac{\bar{b}(1-\bar{\epsilon})}{(\epsilon-\bar{\epsilon})\nu(C_d)}\right] \tag{B.31}$$

with $\bar{\mathbb{E}}$, α defined in the proof and

$$\bar{b} \quad = \quad \nu(V) \vee \frac{\lambda d + b - \bar{\epsilon}\nu(V)}{(1-\bar{\epsilon})}$$

for any $\epsilon > \bar{\epsilon} > 0$.

Proof. The strategy of the proof is as follows: In Glynn & Meyn (1996), there are bounds on the solution to Poisson's equation, in the strongly aperiodic case, the bounds are obtained assuming only a drift toward a petite set. We seek to adapt the proofs in the Foster-Lyapunov case. The proof is constructed by using the split chain to introduce an atom, and the underlying drift condition (on K) to yield such a result for the split chain.

We begin by noting that (iii) of Theorem 2.2 of Glynn & Meyn (1996), in the case of C_d small and the kernel admitting a drift condition, yields for any set $A \in \mathcal{E}$ with positive invariant measure (of K)

$$\mathbb{E}_x[\sum_{k=0}^{\tau_A - 1} V(X_k)] \leq (1+\lambda)V(x) + \frac{b}{\epsilon\nu(A)}. \tag{B.32}$$

To complete the proof, we simply seek to establish that under the assumption of geometric ergodicity of K, we can show that the split chain kernel has its own drift and minorization conditions; this property and the bound above will allow us to conclude. Our split chain construction follows that in Nummelin (184): Let $Z_n = (X_n, Y_n)$, with $Y_n \in \{0,1\}$, be the Markov chain with transition probability (and associated expectation denoted $\bar{\mathbb{E}}$) defined as follows:

$$\bar{P}((x_n, y_n), d(x_{n+1}, y_{n+1})) = \frac{\mathbb{P}(y_{n+1}|x_{n+1})\mathbb{P}(x_n, dx_{n+1}, y_n)}{\mathbb{P}(y_n|x_n)}$$

where

$$\mathbb{P}(y=0|x) = \begin{cases} 1 & x \in C_d^c \\ 1-\bar{\epsilon} & x \in C_d \end{cases}$$

$$\mathbb{P}(y=1|x) = \begin{cases} 0 & x \in C_d^c \\ \bar{\epsilon} & x \in C_d \end{cases}$$

and

$$\mathbb{P}(x, x' \in A, y'=0) = \begin{cases} K(x,A) & x \in C_d^c \\ K(x,A) - \bar{\epsilon}\nu(A) & x \in C_d \end{cases}$$

$$\mathbb{P}(x, x' \in A, y'=1) = \begin{cases} 0 & x \in C_d^c \\ \bar{\epsilon}\nu(A) & x \in C_d \end{cases}$$

We will first establish that $C = C_d \times \{0,1\}$ is a small set. Let $z \in C_d \times \{0\}$, then we have that:

$$\begin{aligned} \bar{P}((x,0), A \times B) &= \int_{A\times B} \frac{\mathbb{P}(y_{n+1}|x_{n+1})}{1-\bar{\epsilon}} K(x, dx_{n+1}) - \bar{\epsilon}\nu(dx_{n+1}) \\ &\geq \int_{A\times B} \frac{\mathbb{P}(y_{n+1}|x_{n+1})}{1-\bar{\epsilon}} (\epsilon - \bar{\epsilon})\nu(dx_{n+1}) \end{aligned}$$

with $A \in \mathcal{E}$ and $B \in \sigma(\{0,1\})$. Since $C_d \times \{1\}$ is an atom we can conclude that, for $z \in C$:

$$\bar{P}(z, \cdot) \geq \tilde{\epsilon}\bar{\nu}(\cdot)$$

where the definition of $\bar{\nu}$ clear.

For the Foster-Lyapunov condition, we have the following formulation. Let $z \in C_d^c \times \{0\}$ and assume that the new Lyapunov function is such that $\bar{V}(z) = V(x)$, then we have:

$$\bar{P}(\bar{V})(z) \leq \lambda \bar{V}(z).$$

If $z \in C_d \times \{0\}$ then

$$\begin{aligned}\bar{P}(\bar{V})(z) &\leq \frac{[\lambda \bar{V}(z) + b - \bar{\epsilon}\nu(V)]}{1-\bar{\epsilon}} \\ &\leq \frac{[\lambda d + b - \bar{\epsilon}\nu(V)]}{1-\bar{\epsilon}}.\end{aligned}$$

Finally, let $z \in C_d \times \{1\}$, then:

$$\bar{P}(\bar{V})(z) \leq \nu(V).$$

Putting this together yields:

$$\bar{P}(\bar{V})(z) \leq \lambda \bar{V}(z) + \bar{b}\mathbb{I}_{\bar{C}}(z)$$

with $\bar{b}$ as above.

To conclude, we will relate the bound (B.32) that we have proved, to the Poisson equation itself. Let $\alpha = C_d \times \{1\}$, then, since we are in the strongly aperiodic case:

$$\begin{aligned}\hat{f}(x) &= \bar{\mathbb{E}}_x[\sum_{k=1}^{\tau_\alpha - 1} \bar{f}(X_k)] \\ &\leq |\bar{f}|_V \left[\big(1+\lambda\big) V(x) + \frac{\bar{b}(1-\bar{\epsilon})}{(\epsilon - \bar{\epsilon})\nu(C_d)} \right]\end{aligned}$$

where we have applied (B.32) and used the small set and drift conditions for the split chain.

□

Proposition B.4. *Let K_μ, $\mu \in \mathscr{P}(E)$ be as in (NL1), then $\forall n \geq 0$, $x \in E$ measurable $f : E \to \mathbb{R}$ such that $K^n(|f|)(x) < \infty$ we have:*

$$K_\mu^n(f)(x) = (1-\epsilon)^n K^n(f)(x) + \epsilon \sum_{l=1}^{n} (1-\epsilon)^{l-1} \Phi(\mu) K^{l-1}(f).$$

Proof. The result is clearly true for $n = 1$, so assume for n and consider $K_\mu^{n+1}(f)(x)$:

$$\begin{aligned}K_\mu^{n+1}(f)(x) &= K_\mu \left[(1-\epsilon)^n K^n(f)(x) + \epsilon \sum_{l=1}^{n} (1-\epsilon)^{l-1} \Phi(\mu) K^{l-1}(f) \right](x) \\ &= (1-\epsilon)^{n+1} K^{n+1}(f)(x) + \epsilon(1-\epsilon)^n \Phi(\mu) K^n(f) + \epsilon \sum_{l=1}^{n} (1-\epsilon)^{l-1} \Phi(\mu) K^{l-1}(f)\end{aligned}$$

from which the proof clearly follows. □

Proposition B.5. *Consider (NL3). Assume (A2-i-ii-iii-iv (a or b)). Then, for $\xi, \mu \in \mathscr{P}_\infty(E)$ we have the following decomposition for the differences in the solution to the Poisson equation:*

$$\begin{aligned}\hat{f}_\xi(x) - \hat{f}_\mu(x) = \sum_{n \in \mathbb{N}_0} \Bigg\{ &\sum_{i=0}^{n-1} \big([K_\xi^i - \omega(\xi)](K_\xi - K_\mu)\{[K_\mu^{n-i-1} - \omega(\mu)](f)\}(x) \big) - \\ &[\omega(\xi) - \omega(\mu)]([K_\mu^n - \omega(\mu)](f)) \Bigg\}.\end{aligned}$$

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

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

since

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

□

Appendix C. A Coupling Argument for $U-$statistics of Markov chains

Define a probability space $(\Omega\times\Omega, \mathcal{F}\otimes\mathcal{F}, \widetilde{\mathbb{P}})$ and a polish space $(E,\mathcal{E})$, such that $\Omega = E^{\mathbb{N}}$, $\mathcal{F} = \mathcal{E}^{\otimes\mathbb{N}}$. Define Markov chains $(\Omega,\mathcal{F},\{X_n\}_{n\geq 0},\mathbb{P}_x)$ $(\Omega,\mathcal{F},\{Y_n\}_{n\geq 0},\mathbb{P}_\pi)$, such that $\widetilde{\mathbb{P}}(\Omega\times\cdot) = \mathbb{P}_x(\cdot)$, $\widetilde{\mathbb{P}}(\cdot\times\Omega) = \mathbb{P}_\pi(\cdot)$. That is, that $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 0}$ are Markov chains of the same transition and different initial distributions and $\widetilde{\mathbb{P}}$ admits $\mathbb{P}_x$ and $\mathbb{P}_\pi$ as marginals. Denote the $U-$statistic of X (resp. Y) as $S_{n,X}^{\odot q}(f)$ (resp. $S_{n,Y}^{\odot q}(f)$). We then have the following result.

Proposition C.1. *Consider the process $\{X_n, Y_n\}_{n\geq 0}$ on $(\Omega\times\Omega, \mathcal{F}\otimes\mathcal{F}, \widetilde{\mathbb{P}})$ and assume that $\mathbb{P}_x$ is induced by a geometrically ergodic Markov kernel. Then:*

$$\lim_{n\to\infty} |S_{n,X}^{\odot q}(f) - S_{n,Y}^{\odot q}(f)| = 0$$

$\widetilde{\mathbb{P}}-$a.s..

Proof. First, note that there exist a $\widetilde{\mathbb{P}}-$a.s finite coupling time τ. For example, by the following argument. Letting $\mathbb{P}_x^{(n)}$ (resp. $\mathbb{P}_\pi^{(n)}$) denote the law of $(X_n, X_{n+1}, \dots)$ (resp. $(Y_n, Y_{n+1}, \dots)$):

$$\lim_{n\to\infty} \|\mathbb{P}_x^{(n)} - \mathbb{P}_\pi^{(n)}\|_{TV} \quad = \quad 0$$

then by Theorem 2.1 of Goldstein (1979), there exists a $\widetilde{\mathbb{P}}-$a.s finite coupling time.

Second, consider the $U-$statistic:

$$S_{n,X}^{\odot q}(f) = \frac{1}{(n+1)_q} \sum_{\alpha \in <q,n+1>} f(X_{\alpha(1)}, \ldots, X_{\alpha(q)})$$

recalling that $(n)_q = n!/(n-q)!$. Now, any statements about $S_{n,Y}^{\odot q}(f)$ can be transferred to those of $S_{n,X}^{\odot q}(f)$ via our coupling construction. This is because the proportion of terms that contain variables before coupling is, for $n+1-\tau \geq q$:

$$\frac{\sum_{k=1}^{\tau \wedge q} (\tau)_k (n+1-\tau)_{q-k}}{(n+1)_q}$$

which, if $\tau < \infty$, goes to zero as $n \to \infty$; since the former can occur $\widetilde{\mathbb{P}}-$a.s. we have the desired result. □

REFERENCES

Aaronson, J., Burton, R., Dehling, H., Gilhat, D., Hill, T. & Weiss, B. (1996). Strong laws for $L-$ and $U-$ statistics. *Trans. Amer. Math. Soc.*, **348**, 2845–2866.

Andrieu, C. & Atchadé, Y. F. (2005). On the efficiency of adaptive MCMC algorithms. Technical Report, Department of Mathematics, University of Bristol.

Andrieu, C. & Moulines É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Prob.*, **16**, 1462–1505.

Andrieu, C. & Robert, C. P. (2001). Controlled MCMC for optimal sampling. Technical Report 0125, Cahiers du Cérémade.

Andrieu, C., Breyer, L, A. & Doucet, A. (2001). Convergence of simulated annealing using Foster-Lyapunov criteria. *J. Appl. Prob.*, **38**, 975–994.

Andrieu, C., Moulines E. & Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Contr. Optim.*, **44**, 283–312.

Andrieu, C., Jasra, A., Doucet, A. & Del Moral, P. (2007). On the convergence of the equi-energy sampler. *Stoch. Anal. Appl.* (in press).

Applebaum, D. (2004). *Lévy Processes and Stochastic Calculus.* CUP: Cambridge.

Atchadé, Y. F. (2006). Resampling from the past to improve MCMC algorithms. Technical Report, Department of Mathematics & Statistics, University of Ottawa.

Atchadé, Y. F. & Rosenthal J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, **11**, 815–828.

Borovkova, S., Burton, R. & Dehling, H. (1999). Consistency of the Takens estimator for the correlation dimension. *Ann. Appl. Prob.*, **9**, 376–390.

Breyer, L, A. & Roberts, G. O. (2001). Catalytic perfect simulation. *Methodol. Computat. Appl. Prob.*, **3**, 161–177.

Brockwell, A. E. (2005). Recursive kernel density estimation of the likelihood for generalized state-space models. Technical Report no. 816, Department of Statistics, Carnegie Mellon University.

Burkholder, D. L. (1973). Distribution function inequalities for Martingales. *Ann. Prob.*, **1**, 19–42.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

Del Moral, P. & Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. In *Séminaire de Probabilités XXXVII*, Ed. Azéma, J., Emery, M., Ledoux, M. and Yor, M., *Lecture Notes in Math.* **1832**, 415-446. Springer: Berlin.

Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.

Del Moral, P. & Miclo, L. (2004). On convergence of chains with occupational self-interactions. *Proc. R. Soc. Lond. A*, **460**, 325–46.

Doucet, A., De Freitas, J. F. G. & Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer: New York.

Gander, M. P. S. & Stephens, D. A. (2007). Stochastic volatility modelling with general marginal distributions: Inference, prediction and model selection. *J. Statist. Plan. Infer.* (in press).

Geyer, C. (1991), Markov chain maximum likelihood. In *Computing Science and Statistics: The 23rd symposium on the interface*, (E. Keramigas ed), 156-63 Fairfax: Interface Foundation.

Goldstein, S. (1979). Maximal Coupling. *Probab. Theory Relat. Fields*, **46**, 193–204.

Glynn, P. W. & Meyn S. P. (1996). A Lyapunov bound for solutions of the Poisson equation. *Ann. Prob.*, **24**, 916–931.

Grams, W. F. & Serfling R., J. (1973). Convergence rates for $U-$statistics and related statistics. *Ann. Statist.*, **1**, 153–160.

Haario, H., Saksman E. & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.

Hobert, J. P. & Robert C. P. (2004). A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. *Ann. Appl. Prob.*, **14**, 1295–1305.

HOEFFDING W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, **19**, 293–325.

JARNER, S. F. & ROBERTS G. O. (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Prob.*, **12**, 224–247.

JASRA, A., STEPHENS, D. A. & HOLMES, C. C. (2007). On population-based simulation for static inference. *Statist. Comp.*, (in press).

JASRA, A., STEPHENS, D. A. & HOLMES, C. C. (2005). Population-based reversible jump Markov chain Monte Carlo. Technical Report, Imperial College London.

KOU, S. C, ZHOU, Q., & WONG, W. H. (2006). Equi-energy sampler with applications to statistical inference and statistical mechanics (with discussion). *Ann. Statist.*, **34**, 1581–1619.

KUSHNER, H. & YIN, G. (1997). *Stochastic Approximation Algorithms and Applications.* Springer: New York.

MEYN, S. P. & TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer: New York.

MEYN, S. P. & TWEEDIE R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Prob.*, **4**, 981–1011.

NUMMELIN, E. (1984). *General Irreducible Markov chains and Non-Negative Operators.* CUP: Cambridge.

ROBERT, C. P., & CASELLA, G. (2004). *Monte Carlo Statistical Methods.* Springer: New York.

ROBERTS, G. O. & ROSENTHAL J. S. (1998). Two convergence properties of hybrid samplers. *Ann. Appl. Prob.*, **8**, 397–407.

ROBERTS, G. O. & ROSENTHAL J. S. (2006). Coupling and ergodicity of adaptive MCMC. Technical Report, University of Lancaster.

ROBERTS, G.O., ROSENTHAL, J, S. & SCHWARTZ, P. O. (1998). Convergence properties of peturbed Markov chains. *J. Appl. Prob.*, **35**, 1–11.

ROBERTS, G. O., PAPASPILIOPOULOS, O. & DELLAPORTAS, P. (2004). Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *J. R. Statist. Soc. B*, **66**, 369-393.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BRISTOL
BRISTOL

DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE LONDON
LONDON

England

E-Mail: c.andrieu@bris.ac.uk

England

E-mail:a.jasra@ic.ac.uk

Department of Statistics

University of British Columbia

Vancouver

Canada

E-Mail: arnaud@stat.ubc.ca

Institut de Mathématiques

University of Bordeaux

Bordeaux

France

E-mail:Pierre.Del-Moral@math.u-bordeaux1.fr