Quantitative bounds on convergence of time-inhomogeneous Markov Chains

R. Douc[*], E. Moulines[†], and Jeffrey S. Rosenthal[‡]

October 4, 2006

Abstract

Convergence rates of Markov chains have been widely studied in recent years. In particular, quantitative bounds on convergence rates have been studied in various forms by Meyn and Tweedie (1994), Rosenthal (1995), Roberts and Tweedie (1999), Jones and Hobert (2001), Fort (2001) and others. In this paper, we first extend a result of Rosenthal (1995) concerning quantitative convergence rates for time homogeneous Markov chains. Our extension allows us to consider f-total variation distance (instead of total variation) and time inhomogeneous Markov chains. We apply our results to simulated annealing.

Key words and phrases: Convergence rate, coupling, Markov Chain Monte Carlo, simulated annealing, f-total variation.

AMS 1991 subject classification. 60J27, 60J22.

[*]CMAP, École Polytechnique, route de Saclay, 91128 Palaiseau cedex, France. douc@cmapx.polytechnique.fr.

[†]Département TSI / CNRS URA 820, Ecole Nationale Supérieure des Télécommunications (ENST), 46 Rue Barrault, 75634 Paris Cedex 13, France. moulines@tsi.enst.fr.

[‡]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. jeff@math.toronto.edu. Supported in part by NSERC of Canada.

1 Time-homogeneous case

1.1 Introduction

Let P be a Markov transition kernel defined on a general state-space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Denote by P^k the corresponding k-step transition kernel. For ξ a probability measure on $\mathcal{B}(\mathcal{X})$ and f a Borel function, define $\xi P(A) = \int \xi(dy) P(y, A)$ and $Pf(x) = \int P(x, dy) f(y)$.

For $f : \mathcal{X} \to [1, \infty)$, the *$f$-total variation* or *$f$-norm* of a signed measure μ on $\mathcal{B}(\mathcal{X})$ is defined as

$$\|\mu\|_f := \sup_{|\phi| \le f} |\mu(\phi)|.$$

When $f \equiv 1$, the f-norm is the total variation norm, which is denoted $\|\mu\|_{\mathrm{TV}}$. Our goal is to find explicit bounds on rates of convergence of $\xi P^n - \xi' P^n$ to zero. In the special case in which P has a stationary distribution π, this corresponds to bounding the convergence of ξP^n to π. Our results extend and sharpen the non-quantitative results developed in, for example, (Meyn and Tweedie, 1993, Chapter 15, 16), where one typically finds conditions under which there exists some *rate function* $r(n)$ such that $r(n) \|P^n(x, \cdot) - \pi\|_f \to 0$ as $n \to \infty$.

The problem of getting explicit bounds on $\|P^n(x, \cdot) - \pi\|_f$ has received much attention in recent years, motivated by control of convergence for Markov Chain Monte Carlo and operation research problems (see e.g. Jones and Hobert (2001)). Most of the results available only cover total variation bound (see Rosenthal (1995) and Roberts and Tweedie (1999)). To the best of our knowledge, the only explicit bound in f-total variation distance is (Meyn and Tweedie, 1994, Theorem 2.3). This bound is based upon the Nummelin's splitting construction and depends in a very intricate way on the constants of the kernel. In this section, we use a different approach, based on *coupling*. We obtain a bound (Theorem 2) which is simple, very generally applicable, and although not tight does improve upon (Meyn and Tweedie, 1994, Theorem 2.3).

1.2 Assumptions and Lemma

Let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. To use the coupling construction, we first need a set where coupling may occur. We assume:

(**A1**) There exist a set $\bar{C} \subset \mathcal{X} \times \mathcal{X}$, a constant $\varepsilon > 0$, and a family of probability measures $\{\nu_{x,x'}, (x,x') \in \bar{C}\}$ on $\mathcal{X}$, with

$$P(x,A) \wedge P(x',A) \geq \varepsilon\, \nu_{x,x'}(A), \qquad \forall A \in \mathcal{B}(\mathcal{X}),\ (x,x') \in \bar{C}. \tag{1}$$

Following Bickel and Ritov (2001), we call $\bar{C}$ a $(1,\varepsilon)$-*coupling set*. For simplicity, only one-step minorisation is considered in this paper. Adaptations to m-step minorisation can be carried out as in Rosenthal (1995). We note that condition (1) is in many cases satisfied by setting $\bar{C} = C \times C$, where C is a so-called *pseudo-small* set. Recall that a subset $C \subset \mathcal{X}$ is $(1,\varepsilon)$-*pseudo-small* if there exist a constant $\varepsilon > 0$ and a family of probability measure $\{\nu_{x,x'}, (x,x') \in C \times C\}$ with $P(x,\cdot) \wedge P(x',\cdot) \geq \varepsilon \nu_{x,x'}(\cdot)$ for all $(x,x') \in C \times C$ (see Roberts and Rosenthal (2001)). We stress that C is a subset of $\mathcal{X}$ and that, despite the obvious similarity, a $(1,\varepsilon)$-pseudo small set is not a $(1,\varepsilon)$-coupling set. Recall finally that a set C is $(1,\varepsilon)$-small if it is $(1,\varepsilon)$-pseudo-small with the same minorizing probability measure $\nu = \nu_{x,x'}$ for all $(x,x') \in C \times C$. The primary motivation for using $(1,\varepsilon)$-coupling set is that the usual pairwise coupling argument can be used without change and that, in some cases detailed below, $(1,\varepsilon)$-coupling sets can be significantly larger than the product of $(1,\varepsilon)$-pseudo-small sets.

To introduce the coupling construction, some additional definitions are required. Let $\bar{R}$ be a Markov transition kernel satisfying, for all $(x,x') \in \bar{C}$ and all $A \in \mathcal{B}(\mathcal{X})$,

$$\bar{R}(x,x';A \times \mathcal{X}) = (1-\varepsilon)^{-1}(P(x,A) - \varepsilon\nu_{x,x'}(A)) \quad \text{and}$$
$$\bar{R}(x,x';\mathcal{X} \times A) = (1-\varepsilon)^{-1}(P(x',A) - \varepsilon\nu_{x,x'}(A)). \tag{2}$$

For example, one may set, for $(x,x') \in \bar{C}$,

$$\bar{R}(x,x';A \times A') = \left((1-\varepsilon)^{-1}(P(x,A) - \varepsilon\nu_{x,x'}(A))\right)\ \left((1-\varepsilon)^{-1}(P(x',A') - \varepsilon\nu_{x,x'}(A'))\right),$$

but other more tricky constructions may also be considered. Similarly, let $\bar{P}$ be a Markov transition kernel on $\mathcal{X} \times \mathcal{X}$ such that, for $(x,x') \in \bar{C}$ and all $A, A' \in \mathcal{B}(\mathcal{X})$,

$$\bar{P}(x,x';A \times A') = (1-\varepsilon)\bar{R}(x,x';A \times A') + \varepsilon\nu_{x,x'}(A \cap A'), \tag{3}$$

and satisfying, for $(x,x') \notin \bar{C}$ and all $A \in \mathcal{B}(\mathcal{X})$,

$$\bar{P}(x,x';A \times \mathcal{X}) = P(x,A) \quad \text{and} \quad \bar{P}(x,x';\mathcal{X} \times A) = P(x',A). \tag{4}$$

For example, one may one again set, for $(x,x') \notin \bar{C}$, $\bar{P}(x,x';A \times A') = P(x,A)P(x',A')$, to get that $\bar{P}$ satisfies (4) for all $(x,x') \in \mathcal{X} \times \mathcal{X}$.

Define the product space $\mathsf{Z} = \mathcal{X} \times \mathcal{X} \times \{0,1\}$, and the associated product sigma-algebra $\mathcal{B}(\mathsf{Z})$. We shall define on the space $(\mathsf{Z}^{\mathbb{N}}, \mathcal{B}(\mathsf{Z})^{\otimes \mathbb{N}})$ a Markov chain $(Z_n := (X_n, X'_n, d_n), n \geq 0)$. Indeed, given Z_n, we construct Z_{n+1} as follows. If $d_n = 1$, then draw $X_{n+1} \sim P(X_n, \cdot)$, and set $X'_{n+1} = X_{n+1}$ and $d_{n+1} = 1$. If $d_n = 0$ and $(X_n, X'_n) \in \bar{C}$, flip a coin with probability of heads ε. If the coin comes up heads, then draw X from $\nu_{X_n, X'_n}(\cdot)$, and set $X_{n+1} = X'_{n+1} = X$, and $d_{n+1} = 1$. If the coin comes up tails, then draw (X_{n+1}, X'_{n+1}) from the residual kernel $\bar{R}(X_n, X'_n; \cdot)$ and set $d_{n+1} = 0$. If $d_n = 0$ and $(X_n, X'_n) \notin \bar{C}$, then draw (X_{n+1}, X'_{n+1}) according to the kernel $\bar{P}(X_n, X'_n; \cdot)$ and set $d_{n+1} = 0$. Here d_n is called a *bell variable*; it indicates whether the chains have coupled ($d_n = 1$) or not ($d_n = 0$) by time n.

For μ a probability measure on $\mathcal{B}(\mathsf{Z})$, denote by $\mathbb{P}_\mu$ the probability measure induced on $(\mathsf{Z}^{\mathbb{N}}, \mathcal{B}(\mathsf{Z})^{\otimes \mathbb{N}})$ by the Markov chain $(Z_n, n \geq 0)$ with initial distribution μ. The corresponding expectation operator will be denoted by $\mathbb{E}_\mu$. It is then easily checked that $(X_n, n \geq 0)$ and $(X'_n, n \geq 0)$ are each marginally updated according to the transition kernel P, *i.e.* for any n, for any initial distributions ξ and ξ', and for any $A, A' \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{P}_{\xi \otimes \xi' \otimes \delta_0}(Z_n \in A \times \mathcal{X} \times \{0,1\}) = \xi P^n(A) \quad \text{and} \quad \mathbb{P}_{\xi \otimes \xi' \otimes \delta_0}(Z_n \in \mathcal{X} \times A' \times \{0,1\}) = \xi' P^n(A'), \tag{5}$$

where δ_x is the Dirac measure centered on x and $\otimes$ is the tensor product of measures. Define the *coupling time* $T = \inf\{k \geq 1;\ d_k = 1\}$ (with the convention $\inf \emptyset = \infty$). Let P^* be the Markov kernel defined, for all $(x,x') \in \mathcal{X} \times \mathcal{X}$ and all $A \in \mathcal{B}(\mathcal{X} \times \mathcal{X})$, by

$$P^*(x,x';A) = \begin{cases} \bar{P}(x,x';A) & \text{if } (x,x') \notin \bar{C}, \\ \bar{R}(x,x';A) & \text{if } (x,x') \in \bar{C}. \end{cases} \tag{6}$$

For μ a probability measure on $\mathcal{X} \times \mathcal{X}$, denote by $\mathbb{P}^*_\mu$ and $\mathbb{E}^*_\mu$ the probability and the expectation induced by the Markov chain on $\mathcal{X} \times \mathcal{X}$ with initial distribution μ and transition kernel P^*.

Lemma 1. *Assume (A1). Then, for any $n \geq 0$ and any non-negative Borel function $\phi : (\mathcal{X} \times \mathcal{X})^{n+1} \to \mathbb{R}^+$, we have*

$$\mathbb{E}_{\xi \otimes \xi' \otimes \delta_0} \left\{ \phi(\bar{X}_0, \ldots, \bar{X}_n)\ \mathbf{1}(d_n = 0) \right\} = \mathbb{E}^*_{\xi \otimes \xi'} \left\{ \phi(\bar{X}_0, \ldots, \bar{X}_n)\ (1-\varepsilon)^{N_{n-1}} \right\},$$

where $\bar{X}_i := (X_i, X_i')$, $N_i := \sum_{j=0}^{i} \mathbf{1}_{\bar{C}}(\bar{X}_j)$ *and* $N_{-1} := 0$.

Proof. We first verify that the result holds for all functions $\phi(\bar{x}_0, \ldots, \bar{x}_n) = \prod_{i=0}^{n} \psi_i(\bar{x}_i)$ where $\bar{x}_i := (x_i, x_i')$ and $(\psi_i, i \geq 0)$ are non-negative Borel functions on $\mathcal{B}(\mathcal{X} \times \mathcal{X})$. The proof is by induction. For $n = 0$, the result is obvious. Assume that the result holds up to order $n-1$ for some $n \geq 1$. We have

$$\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\phi(\bar{X}_0, \ldots, \bar{X}_n)\mathbf{1}(d_n = 0)\right\} = \mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\prod_{i=0}^{n-1} \psi_i(\bar{X}_i)\mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})\ \psi_n(\bar{X}_n)\mathbf{1}(d_n = 0)\right\}$$
$$+ \mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\prod_{i=0}^{n-1} \psi_i(\bar{X}_i)\mathbf{1}_{\bar{C}}(\bar{X}_{n-1})\psi_n(\bar{X}_n)\mathbf{1}(d_n = 0)\right\},$$

where $\bar{C}^c := \mathcal{X}\backslash\bar{C}$. Define $\mathcal{G}_k = \sigma(Z_i = (\bar{X}_i, d_i), 0 \leq i \leq k)$. Note that, for $n \geq 1$,

$$\mathbb{E}\left\{\psi_n(\bar{X}_n)\mathbf{1}(d_n = 0) \mid \mathcal{G}_{n-1}\right\}\mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})\mathbf{1}(d_{n-1} = 0) = \bar{P}\psi_n(\bar{X}_{n-1})\mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})\mathbf{1}(d_{n-1} = 0).$$

Since $N_{n-2}\ \mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1}) = N_{n-1}\ \mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})$ and $\bar{P}(x, x'; \cdot) = P^*(x, x'; \cdot)$ for $(x, x') \notin \bar{C}$ we have, under the induction assumption,

$$\begin{aligned}
&\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\prod_{i=0}^{n-1} \psi_i(\bar{X}_i)\mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})\ \psi_n(\bar{X}_n)\mathbf{1}(d_n = 0)\right\} \qquad (7)\\
&= \mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\prod_{i=0}^{n-1} \psi_i(\bar{X}_i)\mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})\bar{P}\psi_n(\bar{X}_{n-1})\mathbf{1}(d_{n-1} = 0)\right\}\\
&= \mathbb{E}^*_{\xi\otimes\xi'}\left\{\prod_{i=0}^{n-1} \psi_i(\bar{X}_i)\ \mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})P^*\psi_n(\bar{X}_{n-1})\ (1-\varepsilon)^{N_{n-1}}\right\}\\
&= \mathbb{E}^*_{\xi\otimes\xi'}\left\{\prod_{i=0}^{n} \psi_i(\bar{X}_i)\ \mathbf{1}_{\bar{C}^c}(\bar{X}_{n-1})\ (1-\varepsilon)^{N_{n-1}}\right\}.
\end{aligned}$$

Similarly, note that

$$\mathbb{E}\left\{\mathbf{1}(d_n = 0)\psi_n(\bar{X}_n)|\mathcal{G}_{n-1}\right\}\mathbf{1}_{\bar{C}}(\bar{X}_{n-1})\mathbf{1}(d_{n-1} = 0) = (1-\varepsilon)\bar{R}\psi_n(\bar{X}_{n-1})\mathbf{1}(d_{n-1} = 0).$$

Since $(N_{n-2}+1)\mathbf{1}_{\bar{C}}(\bar{X}_{n-1}) = N_{n-1}\mathbf{1}_{\bar{C}}(\bar{X}_{n-1})$ and $\bar{R}(x, x'; \cdot) = P^*(x, x'; \cdot)$ for all $(x, x') \in \bar{C}$, the induction assumption implies

$$\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\prod_{i=0}^{n-1}\psi_i(\bar{X}_i)\mathbf{1}_{\bar{C}}(\bar{X}_{n-1})\ \psi_n(\bar{X}_n)\mathbf{1}(d_n=0)\right\} \tag{8}$$
$$= (1-\varepsilon)\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\prod_{i=0}^{n-1}\psi_i(\bar{X}_i)\ \mathbf{1}_{\bar{C}}(\bar{X}_{n-1})\ \bar{R}\psi_n(\bar{X}_{n-1})\mathbf{1}(d_{n-1}=0)\right\}$$
$$= \mathbb{E}^*_{\xi\otimes\xi'}\left\{\prod_{i=0}^{n-1}\psi_i(\bar{X}_i)\ \mathbf{1}_{\bar{C}}(\bar{X}_{n-1})P^*\psi_n(\bar{X}_{n-1})\ (1-\varepsilon)^{N_{n-1}}\right\}$$
$$= \mathbb{E}^*_{\xi\otimes\xi'}\left\{\prod_{i=0}^{n}\psi_i(\bar{X}_i)\mathbf{1}_{\bar{C}}(\bar{X}_{n-1})(1-\varepsilon)^{N_{n-1}}\right\}.$$

Thus, the two measures on $\mathcal{B}(\mathcal{X}\times\mathcal{X})^{\otimes(n+1)}$ defined respectively by

$$A \mapsto \mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\mathbf{1}_A(\bar{X}_0,\ldots,\bar{X}_n)\mathbf{1}(d_n=0)\right\} \quad \text{and} \quad A \mapsto \mathbb{E}^*_{\xi\otimes\xi'}\left\{\mathbf{1}_A(\bar{X}_0,\ldots,\bar{X}_n)(1-\varepsilon)^{N_{n-1}}\right\}$$

are equal on the monotone class $\mathcal{C} = \{A:\ A = A_0\times\ldots\times A_n, A_i\in\mathcal{B}(\mathcal{X}\times\mathcal{X})\ \}$, and thus these two measures coincide on the product sigma-algebra, which concludes the proof. □

1.3 Main Time-Homogeneous Result

Let $f:\mathcal{X}\to[1,\infty]$ and let $\phi:\mathcal{X}\to\mathbb{R}$ be any Borel function such that $\sup_{x\in\mathcal{X}}|\phi(x)|/f(x)<\infty$. Using (5), the classical coupling inequality (see e.g. (Thorisson, 2000, Chapter 2, section 3)) implies that

$$\left|\xi P^n\phi - \xi' P^n\phi\right| = \left|\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{\phi(X_n)-\phi(X'_n)\right\}\right|$$
$$= \left|\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{(\phi(X_n)-\phi(X'_n))\mathbf{1}(d_n=0)\right\}\right|$$
$$\le \left(\sup_{x\in\mathcal{X}}|\phi(x)|/f(x)\right)\ \mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{(f(X_n)+f(X'_n))\mathbf{1}(d_n=0)\right\}.$$

By Lemma 1,

$$\mathbb{E}_{\xi\otimes\xi'\otimes\delta_0}\left\{(f(X_n)+f(X'_n))\mathbf{1}(d_n=0)\right\} = \mathbb{E}^*_{\xi\otimes\xi'}\left\{(f(X_n)+f(X'_n))(1-\varepsilon)^{N_{n-1}}\right\}.$$

Thus, the following key coupling inequality holds:

$$\left|\xi P^n\phi - \xi' P^n\phi\right| \le \left(\sup_{x\in\mathcal{X}}|\phi(x)|/f(x)\right)\ \mathbb{E}^*_{\xi\otimes\xi'}\left\{(f(X_n)+f(X'_n))(1-\varepsilon)^{N_{n-1}}\right\}. \tag{9}$$

To bound the term in the RHS of (9), we need a *drift condition* outside $\bar{C}$ for the kernel P^*:

(**A2**) There exist a function $\bar{V} : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ and constants b and λ, $0 < \lambda < 1$, such that

$$P^* \bar{V} \leq \lambda \bar{V} + b \mathbf{1}_{\bar{C}}. \tag{10}$$

Theorem 2. *Assume (A1)-(A2). Let $f : \mathcal{X} \to [1, \infty)$ be a function which satisfies $f(x) + f(x') \leq 2\bar{V}(x, x')$, for all $(x, x') \in \mathcal{X} \times \mathcal{X}$. Then, for all $j \in \{1, \dots, n+1\}$ and for all initial probability measures ξ and ξ' on $\mathcal{X}$,*

$$\|\xi P^n - \xi' P^n\|_{\mathrm{TV}} \leq 2(1-\varepsilon)^j \mathbf{1}(j \leq n) + 2\lambda^n B^{j-1} (\xi \otimes \xi')(\bar{V}), \tag{11}$$

$$\|\xi P^n - \xi' P^n\|_f \leq 2(1-\varepsilon)^j \left(b(1-\lambda)^{-1} + \lambda^n (\xi \otimes \xi')(\bar{V}) \right) \mathbf{1}(j \leq n) + 2\lambda^n B^{j-1} \, (\xi \otimes \xi')(\bar{V}), \tag{12}$$

where

$$B = 1 \vee \left((1-\varepsilon)\lambda^{-1} \sup_{(x,x') \in \bar{C}} \bar{R}\bar{V}(x, x') \right).$$

Proof. For any $j \in \{1, \dots, n+1\}$, we have

$$\begin{aligned} \mathbb{E}^*_{\xi \otimes \xi'} \left\{ (f(X_n) + f(X'_n))(1-\varepsilon)^{N_{n-1}} \right\} &\leq \mathbb{E}^*_{\xi \otimes \xi'} \left\{ (f(X_n) + f(X'_n))(1-\varepsilon)^{N_{n-1}} \mathbf{1}(N_{n-1} \geq j) \right\} \\ &\quad + 2\mathbb{E}^*_{\xi \otimes \xi'} \left\{ \bar{V}(\bar{X}_n)(1-\varepsilon)^{N_{n-1}} \mathbf{1}(N_{n-1} < j) \right\}. \end{aligned} \tag{13}$$

Consider the first term of the RHS of (13). We have

$$\mathbb{E}^*_{\xi \otimes \xi'} \left\{ (f(X_n) + f(X'_n))(1-\varepsilon)^{N_{n-1}} \mathbf{1}(N_{n-1} \geq j) \right\} \leq (1-\varepsilon)^j \mathbb{E}^*_{\xi \otimes \xi'} \left\{ (f(X_n) + f(X'_n)) \right\}. \tag{14}$$

If $f \equiv 1$, then $\mathbb{E}^*_{\xi \otimes \xi'} \{ (f(X_n) + f(X'_n)) \} = 2$. Otherwise, by repeated application of the drift condition (A2), we have :

$$(P^*)^n \bar{V} \leq \lambda (P^*)^{n-1} \bar{V} + b \leq \lambda^n \bar{V} + b \sum_{k=0}^{n-1} \lambda^k \leq \lambda^n \bar{V} + b/(1-\lambda).$$

Since $f(x) + f(x') \leq 2\bar{V}(x, x')$, we get :

$$\mathbb{E}^*_{\xi \otimes \xi'} \left\{ (f(X_n) + f(X'_n)) \right\} \leq 2 \, \mathbb{E}^*_{\xi \otimes \xi'} \left\{ \bar{V}(\bar{X}_n) \right\} \leq 2\lambda^n (\xi \otimes \xi')(\bar{V}) + 2b/(1-\lambda).$$

Consider the second term of the RHS of (13). Denote for $s \geq 0$,

$$M_s := \lambda^{-s} B^{-N_{s-1}} \bar{V}(\bar{X}_s)(1-\varepsilon)^{N_{s-1}}.$$

We show that $(M_s, s \geq 0)$ is a $(\mathcal{F}, \mathbb{P}^*_{\xi\otimes\xi'})$-supermartingale, where $\mathcal{F} := (\mathcal{F}_s := \sigma(\bar{X}_i, i \leq s), s \geq 0)$. The definition of N_s and the drift condition (A2) imply

$$\mathbf{1}_{\bar{C}^c}(\bar{X}_s) N_s = \mathbf{1}_{\bar{C}^c}(\bar{X}_s) N_{s-1} \quad \text{and} \quad \mathbf{1}_{\bar{C}^c}(\bar{X}_s) P^* \bar{V}(\bar{X}_s) \leq \mathbf{1}_{\bar{C}^c}(\bar{X}_s)\, \lambda \bar{V}(\bar{X}_s).$$

Thus, we have

$$\begin{aligned}\mathbb{E}^* \left\{ M_{s+1} | \mathcal{F}_s \right\} \mathbf{1}_{\bar{C}^c}(\bar{X}_s) &= \lambda^{-(s+1)} B^{-N_s} P^* \bar{V}(\bar{X}_s)(1-\varepsilon)^{N_s} \mathbf{1}_{\bar{C}^c}(\bar{X}_s) \\ &= \lambda^{-(s+1)} B^{-N_{s-1}} P^* \bar{V}(\bar{X}_s)(1-\varepsilon)^{N_{s-1}} \mathbf{1}_{\bar{C}^c}(\bar{X}_s) \leq M_s \mathbf{1}_{\bar{C}^c}(\bar{X}_s). \quad (15)\end{aligned}$$

By definition, $\sup_{(x,x') \in \bar{C}} \bar{R}\bar{V}(x,x') \leq \lambda(1-\varepsilon)^{-1} B$. Since by construction $\mathbf{1}_{\bar{C}} P^* \bar{V} = \mathbf{1}_{\bar{C}} \bar{R} \bar{V}$, we have

$$\mathbb{E}^* \{ \bar{V}(\bar{X}_{s+1}) | \mathcal{F}_s \} \mathbf{1}_{\bar{C}}(\bar{X}_s) = \bar{R}\bar{V}(\bar{X}_s)\; \mathbf{1}_{\bar{C}}(\bar{X}_s) \leq \lambda(1-\varepsilon)^{-1} B\; \mathbf{1}_{\bar{C}}(\bar{X}_s).$$

Since $\mathbf{1}_{\bar{C}}(\bar{X}_s) N_s = \mathbf{1}_{\bar{C}}(\bar{X}_s)(N_{s-1}+1)$, we have

$$\begin{aligned}&\mathbb{E}^* \{ M_{s+1} | \mathcal{F}_s \} \mathbf{1}_{\bar{C}}(\bar{X}_s) \leq \\ &\qquad \lambda^{-(s+1)} B^{-1} B^{-N_{s-1}} (1-\varepsilon)^{N_{s-1}+1}\; \lambda(1-\varepsilon)^{-1} B \mathbf{1}_{\bar{C}}(\bar{X}_s) \leq M_s \mathbf{1}_{\bar{C}}(\bar{X}_s). \quad (16)\end{aligned}$$

Eqs. (15) and (16) show that $(M_s, s \geq 0)$ is a $(\mathcal{F}, \mathbb{P}^*_{\xi\otimes\xi'})$-supermartingale. By the optional stopping theorem $\mathbb{E}^*_{\xi\otimes\xi'}\{M_n\} \leq \mathbb{E}^*_{\xi\otimes\xi'}\{M_0\}$. Since $B \geq 1$, we have $\mathbf{1}(N_{n-1} < j) \leq B^{j-1} B^{-N_{n-1}}$ which implies

$$\begin{aligned}\mathbb{E}^*_{\xi\otimes\xi'} \left\{ \bar{V}(\bar{X}_n)(1-\varepsilon)^{N_{n-1}} \mathbf{1}(N_{n-1} < j) \right\} &\leq \lambda^n B^{j-1} \mathbb{E}^*_{\xi\otimes\xi'} \left\{ \lambda^{-n} B^{-N_{n-1}} \bar{V}(\bar{X}_n)(1-\varepsilon)^{N_{n-1}} \right\} \\ &\leq \lambda^n B^{j-1} \mathbb{E}^*_{\xi\otimes\xi'} \left\{ M_n \right\} \leq \lambda^n B^{j-1} \xi \otimes \xi'(\bar{V}). \quad (17)\end{aligned}$$

By combining (14) and (17) for $f \equiv 1$, we have

$$\mathbb{E}^*_{\xi\otimes\xi'} \left\{ (f(X_n) + f(X'_n))(1-\varepsilon)^{N_{n-1}} \right\} \leq 2(1-\varepsilon)^j\; \mathbf{1}(j \leq n) + 2\lambda^n B^{j-1} \xi \otimes \xi'(\bar{V}),$$

and (11) follows from (9). Similarly, for f such that $f(x) + f(x') \leq 2\bar{V}(x,x')$, we have

$$\mathbb{E}^*_{\xi\otimes\xi'} \left\{ (f(X_n) + f(X'_n))(1-\varepsilon)^{N_{n-1}} \right\} \;\leq\; 2(1-\varepsilon)^j \left(\lambda^n (\xi \otimes \xi')(\bar{V}) + b/(1-\lambda) \right) + 2\lambda^n B^{j-1} \xi \otimes \xi'(\bar{V}),$$

and (12) follows from (9). □

1.4 Application to convergence to stationarity

If P has a stationary distribution π, *i.e.* if $\pi P = \pi$, then we can choose $\xi' = \pi$. Then $\pi P^n = \pi$ for all n, and hence the results (11) and (12) allow to bound $\|\xi P^n - \pi\|_{\mathrm{TV}}$ and $\|\xi P^n - \pi\|_f$, respectively.

To compare our result with Meyn and Tweedie (1994), Rosenthal (1995) and Roberts and Tweedie (1999), we will now derive from the explicit expressions of the bounds provided in Theorem 2 the rate of convergence for the total variation distance or the f-norm, *i.e.* we find a bound for $\limsup_{n\to\infty} n^{-1} \log \|P^n(x,\cdot) - \pi\|_f$. We follow the approach originally taken in Rosenthal (1995) but we adapt the results to the expression of the bound given in Theorem 2.

Proposition 3. *Assume (A1)-(A2), and that $\pi P = \pi$. Let $f : \mathcal{X} \to [1,\infty)$ be a function satisfying $f(x) + f(x') \le 2\bar{V}(x,x')$, for all $(x,x') \in \mathcal{X} \times \mathcal{X}$. Then, for all $x \in \mathcal{X}$,*

$$\limsup_{n\to\infty} n^{-1} \log \|P^n(x,\cdot) - \pi\|_f \le \begin{cases} \frac{-\log(\lambda)\log(1-\varepsilon)}{\log((M-\varepsilon)/\lambda) - \log(1-\varepsilon)} & \text{if} \quad \frac{M-\varepsilon}{\lambda} \ge 1, \\ \log(\lambda) & \text{if} \quad \frac{M-\varepsilon}{\lambda} < 1. \end{cases} \tag{18}$$

where $M := \sup_{(x,x')\in\bar{C}} \bar{P}\,\bar{V}(x,x')$.

Proof. By definition of $\bar{P}$ (see (3)), for all $(x,x') \in \bar{C}$ we have

$$(1-\varepsilon)\bar{R}\,\bar{V}(x,x') + \varepsilon \int \nu_{x,x'}(dy)\bar{V}(y,y) = \bar{P}\,\bar{V}(x,x') \ge (1-\varepsilon)\bar{R}\,\bar{V}(x,x') + \varepsilon,$$

where we have used that $\bar{V} \ge 1$. Thus,

$$\sup_{(x,x')\in\bar{C}} \bar{R}\,\bar{V}(x,x') \le \frac{M-\varepsilon}{1-\varepsilon},$$

which implies

$$(1-\varepsilon) \sup_{(x,x')\in\bar{C}} \bar{R}\,\bar{V}(x,x')\lambda^{-1} \le (M-\varepsilon)\lambda^{-1}.$$

Assume first $(M-\varepsilon)\lambda^{-1} \ge 1$, the bounds for total variation and f-norm may be expressed for $j \in \{1,\dots,n\}$,

$$\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \le 2(1-\varepsilon)^j + 2\lambda^{n-j+1}(M-\varepsilon)^{j-1} \int \bar{V}(x,x')\pi(dx'),$$

$$\|P^n(x,\cdot) - \pi\|_f \le \frac{2b(1-\varepsilon)^j}{1-\lambda} + 2\lambda^n \left((1-\varepsilon)^j + \lambda^{-j+1}(M-\varepsilon)^{j-1}\right) \int \bar{V}(x,x')\pi(dx')$$

The result follows by choosing j as:

$$j = \left\lfloor \frac{-\log(\lambda) n}{\log((M-\varepsilon)/\lambda) - \log(1-\varepsilon)} \right\rfloor .$$

When $(M-\varepsilon)\lambda^{-1} < 1$, we put $j = n+1$ in (11) and (12) showing that

$$\|\xi P^n - \xi' P^n\|_{\mathrm{TV}} \leq 2\lambda^n \int \bar{V}(x,x')\pi(dx') \quad \text{and} \quad \|\xi P^n - \xi' P^n\|_f \leq 2\lambda^n \int \bar{V}(x,x')\pi(dx').$$

The result follows. □

Remark 1. The bounds we find in this paper for the f-total variation distance are the same than the ones found for the total variation distance by (Roberts and Tweedie, 1999, Theorem 2.3).

In some applications, the minorization and drift conditions (A1)-(A2) are more naturally expressed in terms of the kernel P and it is thus required to derive the bivariate drift and minorisation conditions from the corresponding single variate conditions ((Rosenthal, 1995, Theorem 12), (Roberts and Tweedie, 1999, Section 5)). The crucial point here is to relate the bivariate drift condition (A2) to single variate drift condition. We follow essentially the argument leading to (Rosenthal, 1995, Theorem 12), which allows us to construct such a drift function $\bar{V}$ from univariate test functions (see (Roberts and Tweedie, 1999, Theorem 5.2) for a refinement of this result).

Consider the following assumption:

(S) There exist a function V and a constant c such that

- the level set $C = \{x \in \mathcal{X} \ : V(x) \leq c\}$ is $(1,\varepsilon)$-small, *i.e.* $P(x,\cdot) \geq \varepsilon\nu(\cdot)$ for all $x \in C$ for some $\varepsilon > 0$ and some probability measure ν.
- there exist $\lambda_c < 1$ and $b_c < \infty$ such that $PV \leq \lambda_c V + b_c \mathbf{1}_C$ and $\lambda_c + b_c/(1+c) < 1$.

Under **(S)**, $\bar{C} = \{(x,x'); V(x) \leq c, V(x') \leq c\}$ is a $(1,\varepsilon)$-coupling set, *i.e.* for all $(x,x') \in \bar{C}$ and all $A \in \mathcal{B}(\mathcal{X})$, $P(x,A) \wedge P(x',A) \geq \varepsilon\nu(A)$. Define the univariate residual kernel R as

$$R(x,A) = (1-\varepsilon)^{-1}(P(x,A) - \varepsilon\nu(A)), \quad \forall x \in C, \quad \forall A \in \mathcal{B}(\mathcal{X}). \tag{19}$$

To apply Theorem 1, we need to define the kernels $\bar{R}$, $\bar{P}$ and P^*. Because the drift condition is expressed on the univariate kernel P, we define both $\bar{R}$ and $\bar{P}$ from the corresponding univariate kernels R and P. More precisely, for all $A, A' \in \mathcal{B}(\mathcal{X})$, define

$$\bar{R}(x, x'; A \times A') := R(x, A)R(x', A') \quad \text{if } (x, x') \in \bar{C}, \tag{20}$$

$$\bar{P}(x, x'; A \times A') := P(x, A)P(x', A') \quad \text{if } (x, x') \notin \bar{C}. \tag{21}$$

These kernels satisfy (2) and (4).

Proposition 4. *Assume **(S)**. Then **(A1)** is satisfied with $\bar{C} = C \times C$ and $\nu_{x,x'} = \nu$ for all $(x, x') \in C \times C$. Define P^* as in (6) with $\bar{R}$ and $\bar{P}$ given in (20) and (21). Then (A2) is satisfied with $\bar{V}(x, x') = (1/2)(V(x) + V(x'))$ for all $(x, x') \in \mathcal{X} \times \mathcal{X}$ with*

$$\lambda = \lambda_c + b_c/(1 + c) \quad \textit{and} \quad b = \left\{ \frac{c\varepsilon\lambda_c}{1 - \varepsilon} - \frac{cb_c}{1 + c} \right\} \vee 0 + \frac{b_c - \varepsilon}{1 - \varepsilon}$$

Proof. The proof follows from (Roberts and Tweedie, 1999, Theorem 5.2). Since, for $(x, x') \notin \bar{C}$, $(1 + c)/2 \leq \bar{V}(x, x')$, we have

$$P^*\bar{V}(x, x') \leq \lambda_c \bar{V}(x, x') + \frac{b_c}{2} \leq \left(\lambda_c + \frac{b_c}{1 + c}\right) \bar{V}(x, x'), \quad \forall (x, x') \notin C \times C,$$

and for $(x, x') \in C \times C$,

$$P^*\bar{V}(x, x') = \frac{1}{2}(RV(x) + RV(x')) = \frac{1}{2(1 - \varepsilon)} \big(PV(x) + PV(x') - 2\varepsilon\, \nu(V)\big) \leq \\ \frac{\lambda_c}{(1 - \varepsilon)} \bar{V}(x, x') + \frac{b_c - \varepsilon}{1 - \varepsilon} \leq \lambda \bar{V}(x, x') + b$$

where we have used that, for $(x, x') \in \bar{C}$, $\bar{V}(x, x') \leq c$. The proof follows. □

Under **(S)**, we may thus apply Theorem 2 with $f = V$ which yields explicit bounds for the total-variation and the V-norm, under the assumptions used by Rosenthal (1995) and Roberts and Tweedie (1999) to obtain bounds for the total variation distance (see also Rosenthal (2002)). It is worthwhile to note that (see the discussion above) the rate of convergence in V-norm is the *same* as the rate of convergence in total variation.

Remark 2. It may be checked that, if the sets $\{V \leq d\}$ are 1-small for all $d \geq c$, then assumption **(S)** is always satisfied for large enough d (see (Roberts and Tweedie, 1999, discussion following Theorem 5.2)).

We summarize the discussion above in the following Theorem.

Theorem 5. *Assume* $\boldsymbol{S}$*. Then, for all* $j \in \{1, \dots, n+1\}$ *and for all initial probability measures* ξ *and* ξ' *on* $\mathcal{X}$*,*

$$\|\xi P^n - \xi' P^n\|_{\mathrm{TV}} \le 2(1-\varepsilon)^j \mathbf{1}(j \le n) + \lambda^n B^{j-1} \left(\xi(V) + \xi'(V)\right),$$
$$\|\xi P^n - \xi' P^n\|_V \le 2(1-\varepsilon)^j \left(b(1-\lambda)^{-1} + \lambda^n(\xi(V) + \xi'(V))/2\right) \mathbf{1}(j \le n) + \lambda^n B^{j-1} \ \left(\xi(V) + \xi'(V)\right),$$

where $\lambda = \lambda_c + b_c/(1+c)$ *and*

$$B = 1 \vee \left((1-\varepsilon)\lambda^{-1} \sup_{x \in C} RV(x) \right).$$

1.5 Example

We conclude this section by a simple example showing a situation where we can exploit the additional degree of flexibility brought by $(1,\varepsilon)$-coupling sets. Consider the Markov chain on $\mathbb{R}^d$ defined for $k \in \mathbb{Z}^+$ by:

$$X_{k+1} = g(X_k) + Z_k,$$

where

1. g is a Lipshitz function over $\mathbb{R}^d$ for some norm $\|\cdot\|$ with Lipshitz constant
$$\|g\|_{\mathrm{Lip}} = \sup_{\substack{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d \\ x \ne y}} \frac{\|g(x) - g(y)\|}{\|x - y\|} < 1,$$
2. $(Z_k, k \ge 0)$ is a sequence of independent and identically distributed random vectors, with density q w.r.t. Lebesgue measure on $\mathbb{R}^d$. In addition, q is positive and continuous.

It is known (see *e.g.* Doukhan and Ghindes (1980)) that under these assumptions the Markov chain is positive recurrent and thus has a unique invariant distribution. Define for $\delta > 0$,

$$\bar{C}(\delta) := \left\{ (x, x') \in \mathbb{R}^d \times \mathbb{R}^d : \|x - x'\| \le \delta \right\}. \tag{22}$$

Using $a \wedge b = (1/2)\left((a+b) - |a-b|\right)$, it is easily shown that, for all $(x,x') \in \bar{C}(\delta)$ and all $A \in \mathcal{B}(\mathbb{R}^d)$,

$$P(x,A) \wedge P(x',A) \ge \frac{1}{2} \int_A \left(q(z - g(x)) + q(z - g(x')) - |q(z - g(x)) - q(z - g(x'))| \right) dz,$$

and thus $P(x,A) \wedge P(x',A) \geq \varepsilon(\delta)\nu_{x,x'}(A)$ with

$$\nu_{x,x'}(A) = \frac{\int_A \left(q(z-g(x)) + q(z-g(x')) - |q(z-g(x)) - q(z-g(x'))|\right) dz}{2 - \int |q(z-g(x)) - q(z-g(x'))| dz},$$
$$\varepsilon(\delta) = 1 - \frac{1}{2} \sup_{(x,x')\in\bar{C}(\delta)} \int |q(z-(g(x)-g(x'))) - q(z)| dz. \tag{23}$$

Note that for all $(x,x') \in \bar{C}(\delta)$, $\|g(x)-g(x')\| \leq \|g\|_{\mathrm{Lip}}\|x-x'\| \leq \|g\|_{\mathrm{Lip}}\delta$. Since the function $u \to \int |q(z-u) - q(z)|dz$ is continuous, and q is everywhere positive, for all $\delta > 0$, the set $\bar{C}(\delta)$ is a $(1, \varepsilon(\delta))$-coupling set.

Let $\delta > 0$. For all $(x,x') \in \mathbb{R}^d \times \mathbb{R}^d$ and all $A, A' \in \mathcal{B}(\mathbb{R}^d)$, define $\bar{P}$ by

$$\bar{P}(x,x'; A \times A') = \int \mathbf{1}_A(f(x)+z)\mathbf{1}_{A'}(f(x')+z)q(z)dz,$$

and let, for $(x,x') \in \bar{C}(\delta)$,

$$\bar{R}_\delta(x,x'; A \times A') = (1-\varepsilon(\delta))^{-1}(\bar{P}(x,x'; A\times A') - \varepsilon(\delta)\nu_{x,x'}(A \cap A')).$$

It is easily checked that $\bar{R}_\delta$ and $\bar{P}$ satisfy (2) and (4), respectively. Finally, define P^*_δ as in (6).

We now determine an explicit bound for the total variation distance. Put $\bar{V}(x,x') = 1 + \|x-x'\|$. Note that, for all $(x,x') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\bar{P}\bar{V}(x,x') = 1 + \|g(x)-g(x')\| \leq 1 + \|g\|_{\mathrm{Lip}}\|x-x'\|.$$

Choose λ such that $\|g\|_{\mathrm{Lip}} < \lambda < 1$. By construction, for all $(x,x') \notin \bar{C}(\delta)$, we have $\|x-x'\| \geq \delta$. Hence, for any $\delta > (1-\lambda)/(\lambda - \|g\|_{\mathrm{Lip}})$ and all $(x,x') \notin \bar{C}(0,\delta)$ we have

$$\begin{aligned}1 + \|g\|_{\mathrm{Lip}}\|x-x'\| &= \lambda(1+\|x-x'\|) + (1-\lambda-(\lambda-\|g\|_{\mathrm{Lip}})\|x-x'\|)\\ &\leq \lambda(1+\|x-x'\|) + (1-\lambda-(\lambda-\|g\|_{\mathrm{Lip}})\delta) < \lambda(1+\|x-x'\|).\end{aligned}$$

It remains to prove that $\sup_{(x,x')\in\bar{C}(\delta)} \bar{R}\bar{V}(x,x') < \infty$. Note that

$$\sup_{(x,x')\in\bar{C}(\delta)} \bar{R}\bar{V}(x,x') \leq \frac{\sup_{(x,x')\in\bar{C}(\delta)} \bar{P}\bar{V}(x,x') - \varepsilon(\delta)}{1-\varepsilon(\delta)} \leq \frac{1 + \|g\|_{\mathrm{Lip}}\delta - \varepsilon(\delta)}{1-\varepsilon(\delta)}.$$

Summarizing our findings, for any λ with $\|g\|_{\mathrm{Lip}} < \lambda < 1$, and any $\delta > (1-\lambda)/(\lambda - \|g\|_{\mathrm{Lip}})$, (A1) is satisfied with $\varepsilon := \varepsilon(\delta)$, and (A2) is satisfied with $\bar{V}(x,x') = 1 + \|x-x'\|$. We may thus apply Theorem 2 to obtain a total variation distance bound, as follows. (Note that with this choice of bivariate drift function $\bar{V}$ we may only compute total variation bound; the condition $f(x) + f(x') \leq 2(1 + \|x - x'\|)$ indeed implies that $f \leq 1$.)

Proposition 6. *For all λ such that $\|g\|_{\mathrm{Lip}} < \lambda < 1$, for all $\delta > (1-\lambda)/(\lambda - \|g\|_{\mathrm{Lip}})$, for all $j \in \{1, \ldots, n+1\}$ and for all initial probability measures ξ and ξ' on $\mathcal{X}$,*

$$\|\xi P^n - \xi' P^n\|_{\mathrm{TV}} \leq 2(1 - \varepsilon(\delta))^j \, \mathbf{1}(j \leq n) + 2\lambda^n B^{j-1} \left(1 + \int\int \xi(dx)\xi'(dx')\|x - x'\|\right),$$

where $\varepsilon(\delta)$ is defined in Eq. (23) and

$$B = 1 \vee \left\{\lambda^{-1}(1 + \|g\|_{\mathrm{Lip}}\delta - \varepsilon(\delta))\right\}.$$

2 Time-inhomogeneous case

We now proceed with extending Theorem 2 to time inhomogeneous chains. Specifically, we consider a family $(P_k, k \geq 1)$ of Markov transition kernels. That is, we allow $P_k(x, A)$ to depend not only on the starting point x and the target subset A, but also on the time parameter k. For example, this would be the case for simulated annealing and hidden Markov models; a specific example is discussed in Section 3.

2.1 Assumptions and Lemma

The assumptions and notations parallel those from the time-homogeneous case. We first assume the following minorisation condition.

(NS1) There exist a sequence $(\bar{C}_k, k \geq 1)$ of subsets of $\mathcal{X} \times \mathcal{X}$, $\bar{C}_k \subset \mathcal{X} \times \mathcal{X}$, a sequence $(\varepsilon_k, k \geq 1)$, $\varepsilon_k \geq 0$, and a family of probability measures $(\nu_{k,x,x'}, (x, x') \in \bar{C}_k, k \geq 1)$ such that

$$P_k(x, \cdot) \wedge P_k(x', \cdot) \geq \varepsilon_k \nu_{k,x,x'}(\cdot).$$

Let $(\bar{P}_k, k \geq 1)$ a family of transitions kernels satisfying, for all k, the analog of (4) with $P = P_k$, and let $(\bar{R}_k, k \geq 0)$ be a family of transition kernels verifying, for all k, the analog of (2) with $P = P_k$, $\nu_{x,x'} = \nu_{k,x,x'}$, $\varepsilon = \varepsilon_k$ and $\bar{C} = \bar{C}_k$. The proof is based on straightforward adaptation of the coupling construction used in the homogeneous case. For $n \geq 0$, if $(X_n, X'_n) \in \bar{C}_{n+1}$ and $d_n = 0$ flip a coin with probability of success ε_{n+1}. If the coin comes up heads, then draw X_{n+1} from ν_{n+1,X_n,X'_n}, and set $X_{n+1} = X'_{n+1}$ and $d_{n+1} = 1$. Otherwise, draw (X_{n+1}, X'_{n+1}) from $\bar{R}_{n+1}(X_n, X'_n; \cdot)$ and set $d_{n+1} = 0$. If $(X_n, X'_n) \notin \bar{C}_{n+1}$ and $d_n = 0$, then draw (X_{n+1}, X'_{n+1}) from $\bar{P}_{n+1}(X_n, X'_n; \cdot)$ and set $d_{n+1} = 0$. Finally, define $(P^*_k, k \geq 0)$ be the family of transitions kernels defined as the analog of (6). For μ a probability measure on $\mathcal{X} \times \mathcal{X}$, denote $\mathbb{P}^*_\mu$ and $\mathbb{E}^*_\mu$ the probability and the expectation induced by the Markov chain with initial distribution μ and transition kernels $(P^*_k, k \geq 0)$.

Lemma 7. *Assume (NS1) and let $f : \mathcal{X} \to [1, +\infty)$. For any probability measures ξ, ξ' on $\mathcal{B}(\mathcal{X})$, for any $n \geq 1$,*

$$\|\xi P_1 \dots P_n - \xi' P_1 \dots P_n\|_f \leq \mathbb{E}^*_{\xi \otimes \xi'} \left\{ (f(X_n) + f(X'_n)) \prod_{i=1}^{n} \left(1 - \varepsilon_i \mathbf{1}_{\bar{C}_i}(\bar{X}_{i-1})\right) \right\}, \tag{24}$$

where $\bar{X}_i = (X_i, X'_i)$.

The proof can be adapted from Lemma 1 and Eq. (9). We also assume the following drift condition:

(**NS2**) There exist a family of functions $\{\bar{V}_k\}_{k \geq 0}$, $\bar{V}_k : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ and two sequences $(\lambda_k, k \geq 0)$, $0 \leq \lambda_k \leq 1$ for all $k \geq 1$ and $(b_k, k \geq 0)$ such that:

$$P^*_{k+1} \bar{V}_{k+1} \leq \lambda_k \bar{V}_k + b_k \mathbf{1}_{\bar{C}_{k+1}}, \quad \forall k \geq 0. \tag{25}$$

Define for $j \in \{1, \dots, k\}$,

$$(1 - \varepsilon)_{j,k} := \max_{1 \leq k_1 < \dots < k_j \leq k} \prod_{l=1}^{j} (1 - \varepsilon_{k_l}) \qquad \text{and} \quad B_{j,k} := \max_{1 \leq k_1 < \dots < k_j \leq k} \prod_{l=1}^{j} B_{k_l}$$

where, for any integer k,

$$B_k := 1 \vee \left((1 - \varepsilon_k) \left(\sup_{(x,x') \in \bar{C}_k} \bar{R}_k \bar{V}_k(x, x') \right) \lambda_{k-1}^{-1} \right). \tag{26}$$

By convention, we set $B_{0,k} = 1$ for all k.

2.2 Main time-inhomogeneous result

We can now state our main result, as follows.

Theorem 8. *Assume (NS1) and (NS2). Let $(f_k, k \geq 0)$ be a family of functions such that, for all $k \geq 0$, $f_k(x) + f_k(x') \leq 2\bar{V}_k(x, x')$, for all $(x, x') \in \mathcal{X} \times \mathcal{X}$. Then, for all $j \in \{1, \dots, n+1\}$ and for all initial probability measures ξ and ξ',*

$$\|\xi P_1 \dots P_n - \xi' P_1 \dots P_n\|_{\mathrm{TV}} \leq 2(1-\varepsilon)_{j,n} \mathbf{1}(j \leq n) + 2\left(\prod_{s=0}^{n-1} \lambda_s\right) B_{j-1,n}\, (\xi \otimes \xi')(\bar{V}_0), \tag{27}$$

$$\|\xi P_1 \dots P_n - \xi' P_1 \dots P_n\|_{f_n} \leq 2(1-\varepsilon)_{j,n} D_n \mathbf{1}(j \leq n) + 2\left(\prod_{s=0}^{n-1} \lambda_s\right) B_{j-1,n}\, (\xi \otimes \xi')(\bar{V}_0), \tag{28}$$

where $D_n := \left(\prod_{l=0}^{n-1} \lambda_l\right) \xi \otimes \xi'(V_0) + \sum_{j=0}^{n-1} \left(\prod_{l=j+1}^{n-1} \lambda_l\right) b_j$ with the convention $\prod_{l=i}^{j} \lambda_l = 1$ when $i > j$.

Proof. The proof is along the same lines as for the time-homogeneous case. Denote: $N_k = \sum_{j=0}^{k} \mathbf{1}_{\bar{C}_{j+1}}(X_j, X'_j)$. For any $j \in \{1, \dots, n+1\}$, we have:

$$\mathbb{E}^*_{\xi \otimes \xi'}\left\{(f_n(X_n) + f_n(X'_n)) \prod_{i=1}^{n} (1 - \varepsilon_i \mathbf{1}_{\bar{C}_i}(\bar{X}_{i-1}))\right\} \leq$$

$$(1-\varepsilon)_{j,n} \mathbb{E}^*_{\xi \otimes \xi'}\left\{(f_n(X_n) + f_n(X'_n))\right\} \mathbf{1}(j \leq n) + 2\mathbb{E}^*_{\xi \otimes \xi'}\left\{\bar{V}_n(\bar{X}_n) \prod_{i=1}^{n} (1 - \varepsilon_i \mathbf{1}_{\bar{C}_i}(\bar{X}_{i-1})) \mathbf{1}(N_{n-1} < j)\right\}.$$

where we have used that $\prod_{i=1}^{n}(1 - \varepsilon_i \mathbf{1}_{\bar{C}_i}(\bar{X}_{i-1}))\mathbf{1}(N_{n-1} \geq j) \leq (1-\varepsilon)_{j,n}$. When $f_n \equiv 1$,

$$\mathbb{E}^*_{\xi \otimes \xi'}\left\{(f_n(X_n) + f_n(X'_n))\right\} = 2.$$

Otherwise,

$$\mathbb{E}^*_{\xi \otimes \xi'}\left\{(f_n(X_n) + f_n(X'_n))\right\} \leq 2\mathbb{E}^*_{\xi \otimes \xi'}\left\{\bar{V}_n(\bar{X}_n)\right\} \leq 2D_n.$$

Now, since by definition $B_j \geq 1$ (see (26)), we have $B_{j,n} \leq B_{j',n}$ for all $0 \leq j \leq j' \leq n$ and

$$\mathbf{1}(N_{n-1} \leq j-1)(B_{j-1,n})^{-1} \leq (B_{N_{n-1},n})^{-1}$$

which implies that:

$$\mathbb{E}^*_{\xi \otimes \xi'}\left\{\bar{V}_n(\bar{X}_n) \prod_{i=1}^{n} (1 - \varepsilon_i \mathbf{1}_{\bar{C}_i}(\bar{X}_{i-1})) \mathbf{1}(N_{n-1} < j)\right\} \leq \left(\prod_{j=0}^{n-1} \lambda_j\right) B_{j-1,n}\, \mathbb{E}^*_{\xi \otimes \xi'}\left\{M_n\right\} \tag{29}$$

where, for $s \geq 0$:

$$M_s := \left(\prod_{j=0}^{s-1} \lambda_j \right)^{-1} \left(B_{1,N_{s-1},s} \right)^{-1} \prod_{j=1}^{s} (1 - \varepsilon_j \mathbf{1}_{\bar{C}_j}(\bar{X}_{j-1})) \, \bar{V}_s(X_s, X'_s). \tag{30}$$

As above, $(M_s, s \geq 0)$ is a $(\mathcal{F}, \mathbb{P}^*_{\xi \otimes \xi'})$-supermartingale w.r.t. where $\mathcal{F} := \{\mathcal{F}_s := \sigma(\bar{X}_j, 0 \leq j \leq s), s \geq 0)\}$ which concludes the proof. □

3 Application to Simulated Annealing

In this section, we apply the results above to study the convergence of the Simulated Annealing (SA) algorithm for continuous global optimization (see Locatelli (2002), Locatelli (2001), Fouskakis and Draper (2001), Andrieu et al. (2001) and the references therein).

3.1 Assumptions

Let f be a function defined on $\mathbb{R}$, and let $\mathcal{M}$ be the set of global minima of f (to keep the discussion simple, multi-dimensional versions are not considered here). We assume that:

- **(SA0)** the function f is twice continuously differentiable and there exist $\alpha > 0$, $x_1 \in \mathbb{R}$, such that for all $y \geq x \geq x_1$,

$$f(y) - f(x) \geq \alpha(y - x), \tag{31}$$

  and similarly for all $y \leq x \leq -x_1$,

$$f(y) - f(x) \geq \alpha(x - y). \tag{32}$$

- **(SA1)** For each $x \in \mathcal{M}$, we have $f''(x) > 0$.

Under **(SA0)**, $\mathcal{M} \subseteq [-x_1, x_1]$, i.e. the set of global minima of f are contained in the interval $[-x_1, x_1]$. Assumption **(SA1)** implies that the global minima are isolated and thus, that the set $\mathcal{M}$ is finite. **(SA0)** implies that for all $\gamma \geq 0$, $\int \exp(-\gamma f(y)) \, \mu^{\text{Leb}}(dy) < \infty$, where μ^{Leb} is the Lebesgue measure over $\mathbb{R}$.

Consider a *candidate transition kernel*, $Q(x, A)$, $x \in \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$, which generates potential transitions for a discrete time Markov chain evolving on $\mathbb{R}$. We focus on the case where the candidate points are proposed from a random walk with increment distribution having a density q with respect to μ^{Leb}: $Q(x, A) = \int_A q(y - x)\mu^{\text{Leb}}(dy)$, $A \in \mathcal{B}(\mathbb{R})$. In addition, it is assumed that:

- **(SA2)** The proposal density q is continuous and strictly positive and symmetric: $q(y) > 0$ and $q(y) = q(-y)$.

3.2 The Random Walk Metropolis Hastings (RWMH) Algorithm

The Random Walk Metropolis Hastings (RWMH) algorithm corresponds to the Hastings-Metropolis algorithm introduced in Metropolis et al. (1953) and Hastings (1970). It proceeds as follows, to sample from the (unnormalized) distribution $\exp(-\gamma f(x))\mu^{\text{Leb}}(dx)$ for $\gamma > 0$. (For RWMH, the "inverse temperature" parameter γ is held constant. We shall see later that with Simulated Annealing, by contrast, γ is modified at each iteration of the algorithm.)

Given the current state x, a candidate new state y is chosen according to the law $Q(x, \cdot)$. This candidate y is then accepted with probability $\alpha_\gamma(x, y)$, where:

$$\alpha_\gamma(x, y) = 1 \wedge (\exp(-\gamma(f(y) - f(x))) \, .$$

The RWMH kernel is thus given by:

$$K_\gamma(x, A) = \int_A \alpha_\gamma(x, y) q(y - x)\mu^{\text{Leb}}(dy) + \delta_x(A) \int (1 - \alpha_\gamma(x, y)) q(y - x)\mu^{\text{Leb}}(dy), \quad A \in \mathcal{B}(\mathbb{R}). \tag{33}$$

It then follows that $\pi_\gamma(\cdot)$ is a stationary distribution for K_γ, where:

$$\pi_\gamma(A) = \frac{\int_A \exp(-\gamma f(x))\mu^{\text{Leb}}(dx)}{\int_{\mathbb{R}} \exp(-\gamma f(x))\mu^{\text{Leb}}(dx)}, \quad \forall\, A \in \mathcal{B}(\mathbb{R}).$$

The RWMH algorithm on $\mathbb{R}$ has been extensively studied by Mengersen and Tweedie (1996), where they show that the transition kernels K_γ are π_γ-irreducible (Lemma 1.1) and that all the compact sets are small (Lemma 1.2).

Lemma 9. *Assume* ***(SA0)-(SA2)****. Then, for every compact subset* C *of* $\mathbb{R}$ *such that* $\mu^{\mathrm{Leb}}(C) > 0$*, we have for all* $x \in C$*,* $K_\gamma(x, A) \geq \varepsilon_\gamma \nu_\gamma(A)$*, with:*

$$\varepsilon_\gamma := \varepsilon e^{-\gamma d} \lambda^{\mathrm{Leb}}(C) \quad \text{and} \quad \nu(A) := \frac{\lambda^{\mathrm{Leb}}(A \cap C)}{\lambda^{\mathrm{Leb}}(C)}, \tag{34}$$

where

$$d := \sup_{x \in C} f(x) - \inf_{x \in C} f(x) \quad \text{and} \quad \varepsilon := \inf_{(x,y) \in C \times C} q(y - x) > 0. \tag{35}$$

Proof. For all $x \in C$,

$$K_\gamma(x, A) \geq \int_{A \cap C} \left(e^{-\gamma(f(y) - f(x))} \wedge 1\right) q(y - x) \mu^{\mathrm{Leb}}(dy) \geq \varepsilon e^{-\gamma d} \lambda^{\mathrm{Leb}}(A \cap C).$$

□

To apply Theorem 8, we need to find drift functions satisfying drift conditions outside the compact sets of $\mathbb{R}$. The existence of drift functions for the RWMH algorithm has been shown by (Mengersen and Tweedie, 1996, Theorem 3.2). The proposition below relaxes some of the assumptions required in their result, and shows that the same drift function can be taken for all the Markov kernels K_γ for large enough γ. For $0 < s \leq \gamma$, let $V_s(x) := e^{s f(x)}$, and

$$r(\gamma, s) := 1 - \left(\frac{\gamma - s}{\gamma}\right)^{\gamma/s} + \left(\frac{\gamma - s}{\gamma}\right)^{(\gamma - s)/s}. \tag{36}$$

Proposition 10. *Assume* ***(SA0)-(SA2)****. Then, for all* β *such that* $1/2 < \beta < 1$*, there exist* $\underline{x} < \infty$*,* $\underline{\gamma} > 0$ *and* $s > 0$ *such that*

i) $\frac{K_\gamma V_s(x)}{V_s(x)} \leq r(\gamma, s)$ *for all* $x \in \mathbb{R}$ *and* $\gamma \geq 0$*,*

ii) $\frac{K_\gamma V_s(x)}{V_s(x)} \leq \beta$ *for all* $|x| \geq \underline{x}$ *and* $\gamma \geq \underline{\gamma}$*.*

Proof. By Eq (33), and using that $V_s(y) = e^{s f(x)}$, we have for $\gamma > s > 0$,

$$\frac{K_\gamma V_s(x)}{V_s(x)} = \int \varphi_{\gamma,s}\left(e^{-(f(y) - f(x))}\right) q(y - x) \mu^{\mathrm{Leb}}(dy), \tag{37}$$

where $\varphi_{\gamma,s}(u) := u^{-s}(u^\gamma \wedge 1) + 1 - (u^\gamma \wedge 1)$. It is easily checked that for all $u \geq 0$,

$$\varphi_{\gamma,s}(u) \leq \varphi_{\gamma,s}\left[\left(\frac{\gamma - s}{\gamma}\right)^{1/s}\right] = r(\gamma, s), \tag{38}$$

which proves the first assertion of the proposition. Now, for any $\varepsilon > 0$, we will prove that there exists some $\underline{x}$ such that

$$\lim_{\gamma\to\infty} \sup_{x\geq\underline{x}} \frac{K_\gamma V_s(x)}{V_s(x)} \leq \varepsilon + 1/2.$$

The proof of the corresponding inequality where $x \geq \underline{x}$ is replaced by $x \leq -\underline{x}$ follows the same lines. Choose $M > 0$ such that:

$$\int_{-\infty}^{-M} q(z)\mu^{\mathrm{Leb}}(dz) \leq \varepsilon/2.$$

Inserting this inequality in (37) where $z = y - x$ and using (38) yields

$$\frac{K_\gamma V_s(x)}{V_s(x)} \leq \int_{-M}^{0} \varphi_{\gamma,s}(e^{-(f(x+z)-f(x))})q(z)\mu^{\mathrm{Leb}}(dz) + r(\gamma, s)\left(\int_0^\infty q(z)\mu^{\mathrm{Leb}}(dz) + \varepsilon/2\right)$$

For all $x \geq \underline{x} := x_1 + M$ and all $-M \leq z \leq 0$, we have by assumption **(SA0)**, $e^{-(f(x+z)-f(x))} \geq \exp(-\alpha z) \geq 1$ and since $\varphi_{\gamma,s}(u) = u^{-s}$ for $u \geq 1$,

$$\frac{K_\gamma V_s(x)}{V_s(x)} \leq \int_{-M}^{0} e^{\alpha s z} q(z)\mu^{\mathrm{Leb}}(dz) + r(\gamma, s)(1 + \varepsilon)/2$$

Now, choose s sufficiently large so that the first term of the right hand side is less than $\varepsilon/2$. Once s is chosen, it is easily checked that $\lim_{\gamma\to\infty} r(\gamma, s) = 1$. This proves the second assertion. □

Define $\bar{K}_\gamma(x, x'; A \times A') = K_\gamma(x, A)K_\gamma(x', A')$ and for $s \geq 0$, $\bar{V}_s(x, x') = (1/2)\left(V_s(x) + V_s(x')\right)$.

Proposition 11. *Assume **(SA0)-(SA2)**. For all $s \geq 0$ and for all $c \geq 0$, $\{V_s \leq c\}$ is a compact 1-small set for K_γ. Moreover, there exist $0 \leq \lambda_0 < \lambda < 1$, $s > 0$, $c_0 \leq c$, b and $\underline{\gamma}$ such that, for all $\gamma \geq \underline{\gamma}$,*

$$K_\gamma V_s \leq \lambda_0 V_s + b\mathbf{1}_{\{V_s \leq c_0\}}, \tag{39}$$

$$\bar{K}_\gamma \bar{V}_s \leq \lambda \bar{V}_s + b\mathbf{1}_{\{V_s \leq c\}\times\{V_s \leq c\}}. \tag{40}$$

Proof. The compactness of $\{V_s \leq c\}$ is straightforward from **(SA0)**. Then, by Lemma 9, it is a 1-small set for K_γ. Eq. (39) follows from Proposition 10. To prove (40), write for $c \geq c_0$,

$$\bar{K}_\gamma \bar{V}_s \leq \lambda_0 \bar{V}_s + b\mathbf{1}_{\{V_s \leq c\}\times\{V_s \leq c\}} + (b/2)\left(\mathbf{1}_{\{V_s \leq c\}\times\{V_s > c\}} + \mathbf{1}_{\{V_s > c\}\times\{V_s \leq c\}}\right).$$

Set $0 \leq \lambda_0 < \lambda < 1$ and $c = (b/(\lambda - \lambda_0) - 1) \vee c_0$. We have for all $(x, x') \in \{V_s \leq c\} \times \{V_s > c\}$

$$b/2 \leq (\lambda - \lambda_0)(1 + c)/2 \leq (\lambda - \lambda_0)\bar{V}_s(x, x'),$$

which implies

$$\left(\lambda_0 \bar{V}_s + (b/2)\right) \mathbf{1}_{\{V_s \leq c\} \times \{V_s > c\}} \leq \lambda \bar{V}_s \, \mathbf{1}_{\{V_s \leq c\} \times \{V_s > c\}}.$$

This concludes the proof. □

The key point in the above result (also outlined in Andrieu et al. (2001)) is that, for large enough γ ($\gamma \geq \underline{\gamma}$), all the transition kernels $\bar{K}_\gamma$ satisfy a drift condition outside the *same small set* $\{V_s \leq c\} \times \{V_s \leq c\}$, with the *same* drift function $\bar{V}_s$ and the *same* constants λ and b.

3.3 The Simulated Annealing Algorithm

We now consider the simulated annealing case. Here $\gamma = \gamma_i$ depends on the iteration, and for the i^{th} iteration, the kernel $P_i = K_{\gamma_i}$ is used. Define similarly $\bar{P}_i = \bar{K}_{\gamma_i}$, and $\pi_i = \pi_{\gamma_i}$. Denote $\bar{C} = \{V_s \leq c\} \times \{V_s \leq c\}$, with the constants s and c chosen to satisfy (40). For $(x, x') \in \bar{C}$, set $\bar{R}_i(x, x'; A \times A') = R_i(x, A)R_i(x', A')$, with:

$$R_i(x, A) = (1 - \varepsilon_i)^{-1}(P_i(x, A) - \varepsilon_i \nu_i(A)), \quad \varepsilon_i = \varepsilon_{\gamma_i} \quad \text{and} \quad \nu_i = \nu_{\gamma_i} \tag{41}$$

where ε_γ and ν_γ are defined in (34). We may now state the main result of this section.

Theorem 12. *Assume* ***(SA0)-(SA2)****. For $\xi \geq 0$, set:*

$$\gamma_i = \frac{\log(i + 1)}{d(1 + \xi)} + \underline{\gamma} \tag{42}$$

where d is defined in (35). Then for any initial probability measure μ, we have

$$\lim_{n \to \infty} \|\mu P_1 \dots P_n - \pi_n\|_{\text{TV}} = 0. \tag{43}$$

Proof. For any $1 \leq m \leq n$, we have

$$\|\mu P_1 \dots P_n - \pi_n\|_{\text{TV}} \leq \|(\mu P_1 \dots P_m)P_{m+1} \dots P_n - \pi_m P_{m+1} \dots P_n\|_{\text{TV}} + \sum_{l=m}^{n-1} \|\pi_l P_{l+1} P_{l+2} \dots P_n - \pi_{l+1} P_{l+1} P_{l+2} \dots P_n\|_{\text{TV}}. \tag{44}$$

Let $(a_n, n \geq 0)$ be a sequence of integers such that $\limsup_{n\to\infty}(a_n^{-1} + a_n/n) = 0$. Note that for sufficiently large n,

$$(\lambda^{\mathrm{Leb}}(C))^{-1} \sum_{i=n-a_n}^{n} \varepsilon_i = \varepsilon \sum_{i=n-a_n}^{n} e^{-\gamma_i d} = e^{-\underline{\gamma} d}\, \varepsilon \sum_{i=n-a_n}^{n} (1+i)^{-1/(1+\xi)}.$$

Hence $\lim_{n\to\infty} \sum_{i=n-a_n}^{n} \varepsilon_i = \infty$.

From Proposition 11, we have $\sup_i \sup_{(x,x')\in\bar{C}} \bar{R}_i \bar{V}_s(x,x') < \infty$, and thus there exists an integer l such that $\lambda^l \sup_i \sup_{(x,x')\in C} \bar{R}_i \bar{V}_s(x,x') \leq \lambda$, with $\lambda < 1$ satisfying (40). Since

$$\lambda^l \sup_i \sup_{(x,x')\in C} \bar{R}_i \bar{V}_s(x,x')\lambda^{-1} \leq 1,$$

Theorem 8 implies that, for all $n \geq (l+1)a_n$, and any initial distributions ξ and ξ',

$$\begin{aligned}\|\xi P_{n-(l+1)a_n} \dots P_n - \xi' P_{n-(l+1)a_n} \dots P_n\|_{\mathrm{TV}} &\leq \Big[\prod_{i=n-a_n}^{n} (1-\varepsilon_i)\Big] + \lambda^{a_n} \xi \otimes \xi'(\bar{V}_s) \\ &\leq \exp\left(- \sum_{i=n-a_n}^{n} \varepsilon_i \right) + \lambda^{a_n} \xi \otimes \xi'(\bar{V}_s).\end{aligned}$$

To bound the first term in the RHS of (44), we use the expression above with $\xi = \mu P_1 \dots P_m$ and $\xi' = \pi_m$ with $m = n-(l+1)a_n - 1$. Eq. (39) implies that for any initial distribution μ and any integer m,

$$\mu P_1 \dots P_m V_s \leq \lambda_0^m \mu V_s + \frac{b}{1-\lambda_0},$$

Since $\pi_m P_m = \pi_m$,

$$\pi_m V_s \leq \lambda_0 \pi_m V_s + b \quad \Rightarrow \quad \pi_m V_s \leq \frac{b}{1-\lambda_0}.$$

Hence $\mu P_1 \dots P_m \otimes \pi_m(\bar{V}_s) \leq \lambda_0^m \mu V_s/2 + b/(1-\lambda_0) < \infty$ which implies:

$$\lim_{n\to\infty} \|(\mu P_1 \dots P_{n-(l+1)a_n-1}) P_{n-(l+1)a_n} \dots P_n - \pi_{n-(l+1)a_n-1} P_{n-(l+1)a_n} \dots P_n\|_{\mathrm{TV}} = 0. \qquad (45)$$

We now bound the second term in the RHS of (44). For any $l \in \{1,\dots,n\}$, $\|\pi_l P_{l+1} \dots P_n - \pi_{l+1} P_{l+1} \dots P_n\|_{\mathrm{TV}} \leq \|\pi_l - \pi_{l+1}\|_{\mathrm{TV}}$ and thus:

$$\sum_{l=m}^{n-1} \|\pi_l P_{l+1} \dots P_n - \pi_{l+1} P_{l+1} \dots P_n\|_{\mathrm{TV}} \leq \sum_{l=m}^{n-1} \|\pi_l - \pi_{l+1}\|_{\mathrm{TV}}.$$

To bound this difference we use Lemma 13 in the Appendix, which simplifies the argument in Haario et al. (2001). This Lemma shows that:

$$\sum_{l=m}^{n-1} \|\pi_l - \pi_{l+1}\|_{\mathrm{TV}} \leq 2\log(Z(\gamma_m)/Z(\gamma_n)), \tag{46}$$

where $Z(\gamma) = \int_{\mathbb{R}} e^{-\gamma f(x)} \mu^{\mathrm{Leb}}(dx) / \sup_{x\in\mathbb{R}} e^{-\gamma f(x)}$. Using the Laplace formula (see e.g. Barndorff-Nielsen and Cox (1989)), it may be shown that:

$$Z(\gamma) = (2\pi\gamma^{-1})^{1/2} \left(\sum_{x\in\mathcal{M}} (f''(x))^{-1/2} \right) (1 + o(1)) \quad \text{as} \quad \gamma \to \infty. \tag{47}$$

where $\mathcal{M}$ is the set of global minima of $f(x)$ (recall that, under the stated assumptions, these minima are isolated and there are only a finite number of them). For any integer j, (46) and (47) show that:

$$\begin{aligned} \lim_{n\to\infty} \sum_{l=n-ja_n}^{n-1} \|\pi_l - \pi_{l+1}\|_{\mathrm{TV}} &\leq 2 \lim_{n\to\infty} \log \left(\frac{Z(\gamma_{n-ja_n})}{Z(\gamma_n)} \right) \\ &\leq \lim_{n\to\infty} \log\left(\gamma_n/\gamma_{n-ja_n}\right) = 0. \end{aligned} \tag{48}$$

Together with (45), this concludes the proof. □

A Appendix: Technical Lemmas

Lemma 13. *Let h be a non negative function on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$. Assume that $0 < \int h^\gamma d\mu < \infty$ for all $\gamma > \gamma_0 > 0$, and $\|h\|_\infty = \mathrm{esssup}_{\mathcal{X}} h(x) := \inf\{M : \mu\{x : h(x) > M\} = 0\} < \infty$. For $\gamma \geq \gamma_0$, denote by μ_γ the measure over $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with probability density function $h^\gamma / \int h^\gamma d\mu$ w.r.t μ. Then, for $\gamma' \geq \gamma \geq \gamma_0$,*

$$\|\mu_\gamma - \mu_{\gamma'}\|_{\mathrm{TV}} \leq 2 \log\left(\frac{Z(\gamma)}{Z(\gamma')}\right), \qquad Z(\gamma) = \frac{\int h^\gamma d\mu}{\|h\|_\infty^\gamma}.$$

Proof. Sheffé's identity shows that:

$$\|\mu_\gamma - \mu_{\gamma'}\|_{\mathrm{TV}} = \int |f - g| d\mu,$$

where $f = h^\gamma / \int h^\gamma d\mu$ and $g = h^{\gamma'} / \int h^{\gamma'} d\mu$. Note that $f/\|f\|_\infty = (h/\|h\|_\infty)^\gamma \geq g/\|g\|_\infty = (h/\|h\|_\infty)^{\gamma'}$ μ-a.e. and $\|g\|_\infty/\|f\|_\infty = Z(\gamma)/Z(\gamma')$. The proof follows from Lemma 14 below, which may be of independent interest. □

Lemma 14. *Let f and g be two probability density functions w.r.t to a common dominating measure μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Assume that $\|f\|_\infty < \infty$ and $\|g\|_\infty < \infty$, and $f(x)/\|f\|_\infty \geq g(x)/\|g\|_\infty$ μ-a.s. Then,*

$$\int |f - g| d\mu \leq 2 \log(\|g\|_\infty / \|f\|_\infty).$$

Proof. Using the inequality $(\|f\|_\infty/\|g\|_\infty) g \leq f$ and $|f - g| = f + g - 2(f \wedge g)$, we have:

$$\int |f - g| d\mu = 2\left(1 - \int f \wedge g d\mu\right) \leq 2\left(1 - \int \frac{\|f\|_\infty g}{\|g\|_\infty} \wedge g d\mu\right) = 2\left(1 - \frac{\|f\|_\infty}{\|g\|_\infty}\right).$$

and the proof follows from the inequality $1 - x \leq \log(1/x)$, for $x > 0$. □

Acknowledgements. We sincerely thank the referee and AE for their very careful reading of the manuscript, which has led to many improvements.

References

Andrieu, C., Breyer, L. and Doucet, A. (2001). Convergence of simulated annealing using Foster-Lyapunov criteria. *Journal of Applied Probability* **38** 975–994.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics.* Chapman-Hall, London.

Bickel, P. and Ritov, Y. (2001). Ergodicity of the conditional chain of general state space HMM. Work in progress.

Doukhan, P. and Ghindes, M. (1980). Etude du processus $x_{n+1} = f(X_n) + \varepsilon_n$. *C.R.A.S* **Série A** 921–923.

Fort, G. (2001). *Contrôle explicite d'ergodicité de chaînes de Markov: applications à l'analyse de convergence de l'algorithme Monte-Carlo EM.* Ph.D. thesis, Université de Paris VI.

Fouskakis, D. and Draper, D. (2001). Stochastic optimization: A review. Preprint, University of California, Santa Cruz.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–334.

Locatelli, M. (2001). Convergence and first hitting time of simulated annealing algorithms for continuous global optimization. *Mathematical Methods of Operations Research* **54** 171–199.

Locatelli, M. (2002). Simulated annealing algorithms for continuous global optimization. In *Handbook of Global Optimization: Volume* (P. P. Romeijn and H.E., eds.). 179–229.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics* **24** 101–121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of state calculations by fast computing machines,. *J. Chemical Physics* **21** 1087–1091.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Communication and Control Engineering series, Springer-Verlag, London.

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for convergence rates of Markov chains. *Annals of Applied Probability* **4** 981–1011.

Roberts, G. and Tweedie, R. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and Their Applications* **80** 211–229.

Roberts, G. O. and Rosenthal, J. S. (2001). Small and pseudo-small sets for Markov chains. *Stochastic Models* **17** 121–145.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal American Statistical Association* **90** 558–566.

ROSENTHAL, J. S. (2002). Quantitative convergence rates of markov chains: a simple account. *Electronic Communications in Probability* 123–138.

THORISSON, H. (2000). *Coupling, Stationarity and Regeneration.* Probability and its Applications, Springer-Verlag, New-York.