Convergence of Simulated Annealing Using Foster-Lyapunov Criteria
Author(s): Christophe Andrieu, Laird A. Breyer, Arnaud Doucet
Source: *Journal of Applied Probability*, Vol. 38, No. 4 (Dec., 2001), pp. 975-994
Published by: Applied Probability Trust
Stable URL: http://www.jstor.org/stable/3215778
Accessed: 01/06/2010 15:54

J. Appl. Prob. **38**, 975–994 (2001)
Printed in Israel

CONVERGENCE OF SIMULATED ANNEALING USING FOSTER–LYAPUNOV CRITERIA

CHRISTOPHE ANDRIEU,* *University of Bristol*
LAIRD A. BREYER,** *Lancaster University*
ARNAUD DOUCET,*** *University of Melbourne*

Abstract

Simulated annealing is a popular and much studied method for maximizing functions on finite or compact spaces. For noncompact state spaces, the method is still sound, but convergence results are scarce. We show here how to prove convergence in such cases, for Markov chains satisfying suitable drift and minorization conditions.

Keywords: Simulated annealing; Markov chain Monte Carlo; optimization; Foster–Lyapunov; nonhomogeneous Markov chain

AMS 2000 Subject Classification: Primary 60J27
Secondary 65C40

1. Introduction

Simulated annealing is widely used for finding the global maxima of complicated functions. Theoretical results on the convergence of this method are usually stated (and proved) only for functions whose domain is finite, or at the most compact [12], [24]. Technically speaking, a uniform minorization condition (or Doeblin condition) is assumed for both strong and weak convergence of the process [24].

In this paper, we show how to prove convergence results for simulated annealing when no such uniform minorization exists. In practice, this means that we can study problems with noncompact state spaces. Moreover, our results are framed in terms of Foster–Lyapunov drift conditions, which have already proved very successful in the analysis of homogeneous Markov chains, especially those related to MCMC (Markov chain Monte Carlo) algorithms [17], [18], [20], [21], [22]. To the best of our knowledge, such drift conditions have not been used before in the formulation of simulated annealing problems.

Consider the problem of finding the set of global maxima of some density π whose domain is a set $\mathcal{X}$, usually taken to be $\mathbb{R}^k$. This problem arises commonly in Bayesian statistics, in the context of maximum *a posteriori* (MAP) estimation problems, where the complexity is such that numerical methods are the only option. Simulated annealing is one of them and has found many applications in statistical image processing [6], statistical signal processing [1], physics

Received 5 June 2000; revision received 12 September 2001.

* Postal address: Department of Mathematics, Statistics Group, University of Bristol, University Walk, Bristol BS8 1TW, UK. Email address: c.andrieu@bris.ac.uk
** Postal address: Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster LA1 4YF, UK.
*** Postal address: Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3010, Australia.

and computer science [24]. The method exploits the fact that extrema are preserved under monotone transformations. If γ is any positive real number, then the probability density

$$\overline{\pi}^{\gamma}(x) := \frac{\pi(x)^{\gamma}}{\int_{\mathcal{X}} \pi(y)^{\gamma}\,\mathrm{d}y}$$

also has the same global maxima. Here $\mathrm{d}x$ denotes some σ-finite measure on $\mathcal{X}$, typically Lebesgue measure on $\mathbb{R}^k$, and we assume that the integral in the denominator exists. As γ tends to infinity, the density $\overline{\pi}^{\gamma}$ tends to a mixture delta function concentrated on the set of global maxima, if π is regular enough. Any stationary, homogeneous Markov chain with equilibrium density $\overline{\pi}^{\gamma}$ will therefore spend more time in the vicinity of these maxima, the bigger the value of γ.

To perform simulated annealing, we choose an increasing sequence γ_i and construct a time inhomogeneous Markov chain $X^{(i)}$ whose transition kernels

$$P_{\gamma_i}(\mathrm{d}y \mid x) := \mathrm{P}(X^{(i+1)} \in \mathrm{d}y \mid X^{(i)} = x)$$

are such that

$$\int_{\mathcal{X}} \overline{\pi}^{\gamma_i}(x) P_{\gamma_i}(\mathrm{d}y \mid x)\,\mathrm{d}x = \overline{\pi}^{\gamma_i}(y)\,\mathrm{d}y. \tag{1}$$

The aim of this paper is to give conditions under which $\|\mathrm{P}(X^{(i)} \in \cdot) - \overline{\pi}^{\gamma_i}\| \to 0$ as $i \to +\infty$, where $\|\cdot\|$ is the total variation norm of distributions. This requires Assumptions 1 and 2 below on the transition kernels. Moreover, if Assumptions 3 and 4, below, hold, the distributions $(\overline{\pi}^{\gamma_i})_{i\in\mathbb{N}}$ converge weakly to a mixture of point masses concentrated on the set

$$\mathcal{X}_{\max} := \arg\max_{x\in\mathcal{X}} \pi(x)$$

of global maxima of π. In particular, these two results imply immediately that, for all $\epsilon > 0$,

$$\lim_{i\to+\infty} \mathrm{P}(\widehat{\pi}(X^{(i)}) \geq 1 - \epsilon) = 1,$$

where $\widehat{\pi}(x) := \pi(x)/\|\pi\|_{\infty}$. However, our results are stronger, and we obtain the rate of convergence of the algorithm.

We remark also that it is not possible to prove the even stronger result which would state that $\lim_{i\to+\infty} \|\mathrm{P}(X^{(i)} \in \cdot) - \overline{\pi}^{\infty}\| = 0$, when the set $\mathcal{X}_{\max}$ is of null measure, see [7, p. 873]. Of course, when $\mathcal{X}$ is discrete with counting reference measure, this remark does not apply, and we can then obtain strong convergence under appropriate assumptions [6], [15], [24].

We now state and comment upon the four assumptions mentioned above.

Assumption 1. (Drift condition.) *There exist a function* $V : \mathcal{X} \to [1, +\infty)$ *and constants* $\lambda < 1$, $b < +\infty$ *such that for all* $i \in \mathbb{N}^*$ *and all* $x \in \mathcal{X}$,

$$P_{\gamma_i} V(x) \leq \lambda V(x) + b\mathbf{1}_C(x),$$

where $C := \{x \colon x \in \mathcal{X}$ *and* $V(x) \leq d\}$, *for some constant* $d \geq b/(2(1-\lambda)) - 1$, *and* $P_{\gamma_i} V(x) = \int_{\mathcal{X}} P_{\gamma_i}(\mathrm{d}y \mid x) V(y)$.

Most cases treated in the literature assume that the transition kernels P_{γ_i} each satisfy a uniform ergodicity assumption, i.e. there exist an integer k_0 and constants $\varepsilon_i > 0$ such that $\mathrm{P}(X^{(i+k_0)} \in \mathrm{d}y \mid X^{(i)} = x) \geq \varepsilon_i \mu_i(y)\,\mathrm{d}y$ for all $x \in \mathcal{X}$, where the μ_i are probability densities.

This certainly always holds when the state space $\mathcal{X}$ is finite, provided irreducibility is verified. Assumption 1 is not needed in this case. However, in more realistic situations where $\mathcal{X}$ is unbounded, uniform ergodicity is a very restrictive condition to impose on the transitions. By using a drift function V as above, it is possible to consider much more general situations, as commonly encountered in applications. Indeed, it is often easier to verify a series of local minorization conditions on the set C defined above, as in the following assumption.

Assumption 2. (Minorization condition.) *For some $k_0 \geq 1$ and each $i \in \mathbb{N}^*$, we have*

$$\mathrm{P}(X^{(i+k_0)} \in \mathrm{d}y \mid X^{(i)} = x) \geq \varepsilon_i \mu_i(y)\,\mathrm{d}y, \quad \textit{if } x \in C.$$

Here the constants ε_i are assumed nonincreasing and the μ_i are probability densities.

The uniformly ergodic case mentioned above corresponds to choosing $C = \mathcal{X}$. Of course, this implies that V in Assumption 1 is bounded. In the examples given at the end of the paper, choosing a bounded V will prove impossible. Assumption 1 is a geometric drift condition specially modified for the annealing setup. Recent progress on so-called polynomial rates of convergence [3], [10] parallels the older theory of geometric ergodicity for Markov chains. We expect that our methods will also generalize in this direction, thus weakening further the assumptions required, but we do not pursue this here.

We next list two conditions which ensure that the sequence of probabilities $(\overline{\pi}^{\gamma_i})_{i\in\mathbb{N}}$ converges to a mixture of point masses concentrated on $\mathcal{X}_{\max}$. The first condition precludes the possibility of saddle points, and ensures that there exist only a finite number of global maxima.

Assumption 3. (Regularity of maxima.) *The set $\mathcal{X}_{\max}$ of global maxima of π is finite and contained in the interior $\overset{\circ}{\mathcal{X}}$ of $\mathcal{X}$. At each point $\tilde{x} \in \mathcal{X}_{\max}$, the Hessian*

$$\left[-\frac{\partial^2 \log \pi(x)}{\partial x_i \partial x_j}\bigg|_{\tilde{x}}\right]$$

is a $k \times k$ matrix whose determinant is nonzero.

The final condition is usually required only when the state space is noncompact.

Assumption 4. (Finite entropy condition.) *The entropy of the probability distribution π is finite, i.e.*

$$-\int_{\mathcal{X}} \pi(x) \log \pi(x)\,\mathrm{d}x < +\infty.$$

The plan of this paper is as follows: in Section 2, we state and prove the following decomposition

$$\|\mathrm{P}(X^{(i)} \in \cdot) - \overline{\pi}^{\gamma_i}\| \leq \|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \overline{\pi}^{\gamma_m})\| + \sum_{k=m}^{i-1} \|\overline{\pi}^{\gamma_{k+1}} - \overline{\pi}^{\gamma_k}\|$$

of the error into an *estimation error* (the first term on the right) and an *approximation bias* (the second term on the right). The estimation error has an asymptotic behaviour which depends upon the properties of the transition kernel of $(X^{(i)})_{i\in\mathbb{N}}$, that is Assumptions 1 and 2. Subsection 2.1 is devoted to sufficient conditions for the vanishing of this term, as $i \to +\infty$. The approximation bias has an asymptotic behaviour dependent on regularity properties of π, that is Assumptions 3 and 4. In Subsection 2.2, we discuss sufficient conditions for this term to vanish, as $i \to +\infty$.

The main result is given in Theorem 1 in Subsection 2.3. Finally, in Section 3 we present two examples based on the commonly used random walk based Metropolis algorithm on $\mathbb{R}^k$, which is never uniformly ergodic (unless π has compact support), hence requiring all four hypotheses.

2. Convergence under general conditions

In this section we study the behaviour of $\|\mathrm{P}(X^{(i)} \in \cdot) - \pi^{\gamma_i}\|$ as i tends to infinity. As stated in the introduction, our results are based upon a decomposition into an estimation error and an approximation bias, which we now prove. Both these terms are then analysed separately, and finally combined into Theorem 1 in Subsection 2.3.

Proposition 1. *For all integers m, and i such that $m < i$, we have the estimate*

$$\|\mathrm{P}(X^{(i)} \in \cdot) - \pi^{\gamma_i}\| \leq \|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \pi^{\gamma_m})\| + \sum_{k=m}^{i-1} \|\pi^{\gamma_{k+1}} - \pi^{\gamma_k}\|.$$

Proof. We have for $m < i$ the telescoping sum

$$\begin{aligned}\mathrm{P}(X^{(i)} \in \mathrm{d}x) - \pi^{\gamma_i}(\mathrm{d}x) &= \mathrm{P}(X^{(i)} \in \mathrm{d}x) - \mathrm{P}(X^{(i)} \in \mathrm{d}x \mid X^{(m)} \sim \pi^{\gamma_m}) \\ &\quad + \sum_{k=m}^{i-1} \mathrm{P}(X^{(i)} \in \mathrm{d}x \mid X^{(k)} \sim \pi^{\gamma_k}) - \mathrm{P}(X^{(i)} \in \mathrm{d}x \mid X^{(k+1)} \sim \pi^{\gamma_{k+1}}).\end{aligned}$$

Upon using the fact that $\mathrm{P}(X^{(i)} \in \mathrm{d}x \mid X^{(k)} \sim \pi^{\gamma_k}) = \mathrm{P}(X^{(i)} \in \mathrm{d}x \mid X^{(k+1)} \sim \pi^{\gamma_k})$, which is a consequence of the stationarity relation $\pi^{\gamma_k} P_{\gamma_k} = \pi^{\gamma_k}$, we see by the triangle inequality and the definition $\|\nu\| := \sup_{A \in \mathcal{B}(\mathsf{X})} |\nu(A)|$, that

$$\begin{aligned}\|\mathrm{P}(X^{(i)} \in \cdot) - \pi^{\gamma_i}\| &\leq \|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \pi^{\gamma_m})\| \\ &\quad + \sum_{k=m}^{i-1} \left\| \int_{\mathsf{X}} \mathrm{P}(X^{(i)} \in \cdot \mid X^{(k)} = x)(\pi^{\gamma_k}(\mathrm{d}x) - \pi^{\gamma_{k+1}}(\mathrm{d}x)) \right\| \\ &\leq \|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \pi^{\gamma_m})\| + \sum_{k=m}^{i-1} \|\pi^{\gamma_k} - \pi^{\gamma_{k+1}}\|,\end{aligned}$$

where the last inequality is due to the contracting property of the transition kernel.

We now study the asymptotic behaviour of each of these terms.

2.1. Bounding the estimation error

In this section, we bound the estimation error

$$\|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \pi^{\gamma_m})\|$$

by a standard coupling construction [11], and using a method pioneered by Rosenthal [20] for time homogeneous Markov chains. In our time inhomogeneous setup, most of his arguments can be salvaged, but some subtle changes are required (see Subsections 2.1.1 and 2.1.2). The main result here is Proposition 2, where a bound on the estimation error is obtained that consists of a term due to the minorization condition, and another term due to the drift condition.

From the simulation point of view, the chain $(X^{(i)})_{i\in\mathbb{N}}$ is usually constructed by a sequence of random maps $F_i : \mathcal{X} \to \mathcal{X}$ such that $X^{(i+1)} = F_i(X^{(i)})$. In this case we also have

$$P_{\gamma_i}(\mathrm{d}y \mid x) = \mathrm{P}(F_i(x) \in \mathrm{d}y).$$

We shall need a pair of auxiliary processes $(\phi_{1,m}^{(i)})_{i\in\mathbb{N}}$ and $(\phi_{2,m}^{(i)})_{i\in\mathbb{N}}$, defined for $i \geq m$.

```
• Initialization $\phi_{1,m}^{(m)} \sim \mathrm{P}(X^{(m)} \in \cdot)$, $\phi_{2,m}^{(m)} \sim \bar{\pi}^{\gamma_m}$, $d = 0$ and $i = m$.
• Iteration $i + 1$
  If $d = 0$,
      If $\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)} \notin C \times C$ then $\phi_{j,m}^{(i+1)} = F_i(\phi_{j,m}^{(i)})$, for $j = 1, 2$, and $i := i + 1$.
      Else
        1. With probability $\varepsilon_i$, $\phi_{1,m}^{(i+k_0)} = \phi_{2,m}^{(i+k_0)} \sim \mu_i$, for $j = 1, 2$, and $d = 1$.
        2. Otherwise $\phi_{j,m}^{(i+k_0)} \sim (\mathrm{P}(X^{(i+k_0)} \in \cdot \mid \phi_{j,m}^{(i)}) - \varepsilon_i\mu_i(X^{(i+k_0)} \in \cdot))/(1 - \varepsilon_i)$, for $j = 1, 2$.
        3. $\phi_{j,m}^{(i+1)}, \ldots, \phi_{j,m}^{(i+k_0-1)} \sim \mathrm{P}(X^{(i+1)} \in \cdot, \ldots, X^{(i+k_0-1)} \in \cdot \mid X^{(i)} = \phi_{j,m}^{(i)}, X^{(i+k_0)} = \phi_{j,m}^{(i+k_0)})$, for $j = 1, 2$, and $i := i + k_0$.
  Else $\phi_{j,m}^{(i+1)} = F_i(\phi_{j,m}^{(i)})$, for $j = 1, 2$, and $i := i + 1$.
```

We easily verify that the Markov chains constructed in this way are marginally updated at each iteration according to the transition kernel P_{γ_i}. Note that if $(\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)}) \in C \times C$, then with probability ε_i we have $\phi_{1,m}^{(i)} = \phi_{2,m}^{(i)}$, and then the chains are identical for all subsequent times. We cannot in general simulate these chains easily, but we shall only use them to get the theoretical bound.

Proposition 2. *Set*

$$\tau_{1,m} := \min\{i : i > m + k_0 - 1, (\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)}) \in C \times C\}$$

and for $k \geq 2$

$$\tau_{k,m} := \inf\{i : i \geq k_0 + \tau_{k-1,m}, (\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)}) \in C \times C\}.$$

Then with $N_{i,m} := \max\{k : \tau_{k,m} < i\}$ *we have for all* i, m *and* j *such that* $i - (j+1)k_0 > m$,

$$\|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \bar{\pi}^{\gamma_m})\| \leq \prod_{l=1}^{j}(1 - \varepsilon_{i-lk_0}) + \mathrm{P}(N_{i-k_0+1,m} < j).$$

Proof. Define $\tau_{C,m} := \inf\{i > m : \phi_{1,m}^{(i)} = \phi_{2,m}^{(i)}\}$ for $m < i$. Then

$$\begin{aligned}\|\mathrm{P}(X^{(i)} \in \cdot) - \mathrm{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \bar{\pi}^{\gamma_m})\| &\leq \|\mathrm{P}(\phi_{1,m}^{(i)} \in \cdot) - \mathrm{P}(\phi_{2,m}^{(i)} \in \cdot)\| \\ &\leq \mathrm{P}(\phi_{1,m}^{(i)} \neq \phi_{2,m}^{(i)}) \\ &\leq \mathrm{P}(\tau_{C,m} > i),\end{aligned}$$

as the event on the second line is included in that of the third line. We have the following identity

$$\mathrm{P}(\tau_{C,m} > i) = \mathrm{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} \geq j) + \mathrm{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} < j). \tag{2}$$

Now the $\tau_{k,m}$ are the times, separated by at least k_0 iterations, when the chains meet in C, and $N_{i,m}$ counts the number of times that the chains meet in C before instant i, separated by at least k_0 iterations. Since coupling occurs with probability ε_i whenever both chains are in C we obtain the bound

$$\mathrm{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} \geq j) \leq \prod_{l=1}^{j}(1 - \varepsilon_{i-lk_0}),$$

since the series ε_i is nonincreasing. Then using (2) we obtain

$$\begin{aligned}\mathrm{P}(\tau_{C,m} > i) &\leq \prod_{l=1}^{j}(1 - \varepsilon_{i-lk_0}) + \mathrm{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} < j)\\ &\leq \prod_{l=1}^{j}(1 - \varepsilon_{i-lk_0}) + \mathrm{P}(N_{i-k_0+1,m} < j),\end{aligned}$$

as

$$\mathrm{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} < j) + \mathrm{P}(\tau_{C,m} \leq i \text{ and } N_{i-k_0+1,m} < j) = \mathrm{P}(N_{i-k_0+1,m} < j).$$

2.1.1. *Bounding the term* $\mathrm{P}(N_{i-k_0+1,m} < j)$. The main result of this section is a bound on the distribution of $N_{i,m}$ that will be used in the proof of Theorem 1. It is stated in Proposition 3 (cf. [20, Lemma 3]) and its proof is a direct combination of Lemmas 1 and 2 below.

Proposition 3. *There exist constants A, B and $\lambda_* < 1$ such that*

$$\mathrm{P}(N_{i-k_0+1,m} < j) \leq \lambda_*^{i-k_0+1} A^{j-1} B$$

for any m and i, j such that $i - (j+1)k_0 > m$.

The constants appearing in the above proposition are the following. Set $\lambda_* := \lambda + b/(2(1+d))$ and $b_* := b(1+2d)/(2(1+d))$; then we have

$$A := \lambda_*^{k_0} \sup_{x \in C} V(x) + b_* \sum_{k=0}^{k_0-1} \lambda_*^k, \qquad B := \frac{b}{1-\lambda} + \frac{1}{2}\mathrm{E}[V(X^{(0)})].$$

Note that we have $\lambda_* < 1$ by the drift condition (Assumption 1). For later use, we propose the following sufficient condition for $\lim_{i\to+\infty} \mathrm{P}(N_{i-k_0+1,m_i} < j_i) = 0$. The meaning of the constants will become clear at the end.

Corollary 1. *For $m_i = \lfloor i - i^{1-\zeta/2} \rfloor$ where $\zeta \in (0,1)$ and $j_i = \lfloor r i^{1-\zeta/2} \rfloor$ where*

$$0 < r < \min\left\{\frac{1}{k_0}, \left|\frac{-\log(\lambda_*)}{\log(A)}\right|\right\}$$

we have

$$\lim_{i\to+\infty} \mathrm{P}(N_{i-k_0+1,m_i} < j_i) = 0. \tag{3}$$

Proof. We note that for i sufficiently large, then the condition

$$i - (\lfloor r i^{1-\zeta/2} \rfloor + 1)k_0 > \lfloor i - i^{1-\zeta/2} \rfloor$$

holds, allowing us to use Proposition 3. If $A \le 1$ then (3) holds for any $r > 0$. If $A > 1$, then by noting that for $r < -\log(\lambda_*)/\log(A)$ we have $\lambda_* A^r < 1$ the result (3) still holds.

We now prove the first technical lemma used in Proposition 3.

Lemma 1. *Fix m, i and j such that $i > k_0 + m$ and define $r_1 = r_{1,m} := \tau_{1,m} - m$ and $r_i = r_{i,m} := \tau_{i+1,m} - \tau_{i,m}$ for $i > 1$. We have*

$$\begin{aligned}&\mathrm{P}(N_{i-k_0+1,m} < j)\\ &\quad \le \lambda_*^{i-k_0+1}\,\mathrm{E}(\lambda_*^{-r_1}\,\mathrm{E}(\lambda_*^{-r_2}\,\mathrm{E}(\lambda_*^{-r_3}\cdots\mathrm{E}(\lambda_*^{-r_j}\mid r_1,\ldots,r_{j-1})\ldots\mid r_1,r_2)\mid r_1)).\end{aligned}$$

Proof. Let us write

$$\begin{aligned}\mathrm{P}(N_{i-k_0+1,m} < j) &= \mathrm{P}(\tau_{j+1,m} > i - k_0 + 1)\\ &= \mathrm{P}\Bigl(\sum_{t=1}^{j} r_t > i - k_0 + 1\Bigr)\\ &= \mathrm{P}\Bigl(\lambda_*^{-\sum_{t=1}^{j} r_t} > \lambda_*^{-(i-k_0+1)}\Bigr)\\ &\le \lambda_*^{i-k_0+1}\,\mathrm{E}\Bigl(\prod_{t=1}^{j}\lambda_*^{-r_t}\Bigr)\end{aligned}$$

by Markov's inequality. Then the result follows from the identity $\mathrm{E}[XY] = \mathrm{E}[X\,\mathrm{E}[Y \mid X]]$.

This is the second technical lemma we need.

Lemma 2. *Define $V^*(x_1, x_2) := \frac{1}{2}[V(x_1) + V(x_2)]$. From the drift and minorization conditions (Assumptions 1 and 2), assuming $\mathrm{E}[V(X^{(0)})] < +\infty$, we have for any m:*

1. *The expectation $\mathrm{E}(\lambda_*^{-r_{1,m}})$ is uniformly bounded in m:*

$$\mathrm{E}(\lambda_*^{-r_{1,m}}) \le \sup_{i>m}\mathrm{E}(V^*(\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)})) \le \frac{b}{1-\lambda} + \frac{1}{2}\mathrm{E}[V(X^{(0)})] < +\infty.$$

2. *For any m, $i > 1$ and all $r_{1,m}, \ldots, r_{i-1,m}$*

$$\mathrm{E}(\lambda_*^{-r_{i,m}} \mid r_{1,m}, \ldots, r_{i-1,m}) \le \lambda_*^{k_0}\sup_{x\in C} V(x) + b_* \sum_{k=0}^{k_0-1}\lambda_*^{k}.$$

Proof. Following [20] and by Theorem 5.2 of [19], observe that by the drift condition (Assumption 1)

$$\mathrm{E}[V^*(\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)}) \mid \phi_{1,m}^{(i-1)}, \phi_{2,m}^{(i-1)}] \le \lambda_* V^*(\phi_{1,m}^{(i-1)}, \phi_{2,m}^{(i-1)}) + b_* \mathbf{1}_C(\phi_{1,m}^{(i-1)})\mathbf{1}_C(\phi_{2,m}^{(i-1)}). \quad (4)$$

Then, following [20], we introduce the functions

$$g_{i,m}(k) := \begin{cases} \lambda_*^{-(k-m)} V^*(\phi_{1,m}^{(k)}, \phi_{2,m}^{(k)}) & \text{if } k \leq \tau_{i,m}, \\ 0 & \text{if } k > \tau_{i,m}, \end{cases}$$

whose series has nonincreasing (in k) expectation from (4) (at least for $\tau_{i-1} + k_0 \leq k$ when $i > 1$, as before τ_i we do not go into C, and for $k \geq m$ when $i = 1$). Thus, and because $V^*(\cdot, \cdot) \geq 1$, we obtain

$$\mathrm{E}[\lambda_*^{-r_{1,m}}] \leq \mathrm{E}[g_{1,m}(r_{1,m})] \leq \mathrm{E}[V^*(\phi_{1,m}^{(m)}, \phi_{2,m}^{(m)})] = \tfrac{1}{2}[\mathrm{E}[V(\phi_{1,m}^{(m)})] + \mathrm{E}[V(\phi_{2,m}^{(m)})]].$$

From Assumption 1 we have for appropriate j and i

$$P_{\gamma_i} V(\phi_{j,m}^{(i)}) \leq \lambda V(\phi_{j,m}^{(i)}) + b\mathbf{1}_C(\phi_{j,m}^{(i)}),$$

from which we deduce that:

1. Upon using the fact that $\overline{\pi}^{\gamma_i} P_{\gamma_i} = \overline{\pi}^{\gamma_i}$ for any $m \geq 0$ and $i \geq m$

$$\mathrm{E}[V(\phi_{2,m}^{(i)})] \leq \lambda\, \mathrm{E}[V(\phi_{2,m}^{(i)})] + b,$$

that is, again for any $m \geq 0$ and $i \geq m$,

$$\mathrm{E}[V(\phi_{2,m}^{(i)})] \leq \frac{b}{1-\lambda}.$$

2. From the definition of $(\phi_{1,m}^{(i)})_{i \in \mathbb{N}}$

$$\mathrm{E}[V(\phi_{1,m}^{(i+1)})] \leq \lambda\, \mathrm{E}[V(\phi_{1,m}^{(i)})] + b,$$

which by recursion implies that for any $i \geq 0$

$$\mathrm{E}[V(\phi_{1,m}^{(i)})] \leq \lambda^i\, \mathrm{E}[V(X^{(0)})] + b \sum_{k=0}^{i-1} \lambda^k,$$

that is, if $\mathrm{E}[V(X^{(0)})] < +\infty$, then

$$\sup_{i \in \mathbb{N}} \mathrm{E}[V(\phi_{1,m}^{(i)})] \leq \frac{b}{1-\lambda} + \mathrm{E}[V(X^{(0)})] < +\infty.$$

Summarizing, we have shown that for all $m \geq 0$

$$\sup_{i \geq m} \mathrm{E}[V^*(\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)})] < \frac{b}{1-\lambda} + \frac{1}{2}\mathrm{E}[V(X^{(0)})] < +\infty,$$

which is the first result we wanted to prove.

We now prove the second result. The fact that $g_{i,m}(k)$ is nonincreasing for $k \geq \tau_{i-1,m} + k_0$ and $V^*(\cdot, \cdot) \geq 1$ allows us to deduce that:

$$\begin{aligned} \mathrm{E}[\lambda_*^{-r_{i,m}} \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] &= \mathrm{E}[\lambda_*^{-r_{i,m}} \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\ &= \mathrm{E}[\lambda_*^{\tau_{i-1,m}-m+m-\tau_{i,m}} \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\ &\leq \mathrm{E}[\lambda_*^{\tau_{i-1,m}-m} g_{i,m}(\tau_{i,m}) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\ &\leq \mathrm{E}[\lambda_*^{\tau_{i-1,m}-m} g_{i,m}(\tau_{i-1,m} + k_0) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\ &\leq \lambda_*^{-k_0}\, \mathrm{E}[V^*(\phi_{1,m}^{(\tau_{i-1,m}+k_0)}, \phi_{2,m}^{(\tau_{i-1,m}+k_0)}) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}]. \end{aligned}$$

Unlike in [20], we cannot yet conclude the proof as the Markov chain is nonhomogeneous. We therefore find an upper bound for this term. For $k_0 > 1$

$$\begin{aligned}
&\mathrm{E}[V^*(\phi_{1,m}^{(\tau_{i-1,m}+k_0)}, \phi_{2,m}^{(\tau_{i-1,m}+k_0)}) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\
&\quad = \mathrm{E}[\mathrm{E}[V^*(\phi_{1,m}^{(\tau_{i-1,m}+k_0)}, \phi_{2,m}^{(\tau_{i-1,m}+k_0)}) \mid \phi_{1,m}^{(\tau_{i-1,m}+k_0-1)}, \phi_{2,m}^{(\tau_{i-1,m}+k_0-1)}] \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\
&\quad \leq \lambda_* \,\mathrm{E}[V^*(\phi_{1,m}^{(\tau_{i-1,m}+k_0-1)}, \phi_{2,m}^{(\tau_{i-1,m}+k_0-1)}) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] + b_*,
\end{aligned}$$

where we have used (4) to obtain the inequality.

Successively applying this procedure (k_0 times) we obtain

$$\begin{aligned}
&\mathrm{E}[V^*(\phi_{1,m}^{(\tau_{i-1,m}+k_0)}, \phi_{2,m}^{(\tau_{i-1,m}+k_0)}) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}] \\
&\quad \leq \lambda_*^{k_0} V^*(\phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})}) + b_* \sum_{k=0}^{k_0-1} \lambda_*^k,
\end{aligned}$$

which leads to the second result.

2.1.2. *Bounding the term* $\prod_{l=1}^{j}(1-\varepsilon_{i-lk_0})$. The term $\prod_{l=1}^{j}(1-\varepsilon_{i-lk_0})$ converges to 0 as $j \to +\infty$ if and only if $\sum_{l=1}^{j} \varepsilon_{i-lk_0}$ diverges as $j \to +\infty$. Clearly the rate at which ε_i in Assumption 2 goes to zero controls the decay of the estimation error (2). In turn, the vanishing of the ε_i in Assumption 2 depends on the transition probabilities of $(X^{(i)})_{i\in\mathbb{N}}$ and thereby on the $(\gamma_i)_{i\in\mathbb{N}}$. From now on we assume that

$$\varepsilon_i = \varepsilon\alpha^{\sum_{j=0}^{k_0-1}\gamma_{i+j}} \quad \text{with } \alpha \in (0,1) \text{ and } \varepsilon > 0, \tag{5}$$

as it is a case often met in practice. Such examples will be given in Section 3 for our two applications with the Metropolis algorithm, but can also be found in [6], [24]. Since the ε_i are now simple functions of the γ_i we seek a condition on the series $(\gamma_i)_{i\in\mathbb{N}}$ so that the term $\prod_{l=1}^{j}(1-\varepsilon_{i-lk_0})$ converges to 0 as $j \to +\infty$ (j being a function of i). We have the following proposition.

Proposition 4. *For all i we assume that (5) holds and we let*

$$\gamma_i = \frac{\log(i+\varrho-k_0+1)}{-k_0\log(\alpha)(1+\zeta)}$$

(where $\alpha \in (0,1)$, $\varepsilon > 0$ and $\varrho > 0$). Then for $j_i = \lfloor r i^{1-\zeta/2} \rfloor$ (where $\zeta \in (0,1)$) with

$$0 < r < \min\left\{\frac{1}{k_0}, \left|\frac{-\log(\lambda_*)}{\log(A)}\right|\right\}$$

we obtain

$$\lim_{i\to+\infty} \prod_{l=1}^{j_i}(1-\varepsilon_{i-lk_0}) = 0.$$

Proof. We follow [7] for the proof. We denote $a := 1/(1+\zeta)$ and we study the sum $\sum_{l=1}^{j_i} \varepsilon_{i-lk_0}$ using the fact that $1/(\varrho + x)^a$ is a decreasing function:

$$\begin{aligned}
\varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2}\rfloor} \alpha^{\sum_{j=0}^{k_0-1} \gamma_{i-lk_0+j}} &> \varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2}\rfloor} \alpha^{k_0(\gamma_{i-lk_0+k_0-1})} > \varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2}\rfloor} (i - lk_0 + \varrho)^{-a} \\
&> \frac{\varepsilon}{k_0 a\zeta} \int_{i-k_0\lfloor ri^{1-\zeta/2}\rfloor}^{i} (x+\varrho)^{-a}\,\mathrm{d}x \\
&> \frac{\varepsilon}{k_0 a\zeta} [(i+\varrho)^{a\zeta} - (i - k_0\lfloor ri^{1-\zeta/2}\rfloor + \varrho)^{a\zeta}] \\
&> \frac{\varepsilon}{k_0 a\zeta}(i+\varrho)^{a\zeta}\left[1 - \left(1 - \frac{k_0 r i^{1-\zeta/2}}{i+\varrho}\right)^{a\zeta}\right] \\
&> \frac{\varepsilon}{k_0 a\zeta}(i+\varrho)^{a\zeta}\left[1 - \left(1 - a\zeta\frac{k_0 r i^{1-\zeta/2}}{i+\varrho}\right)\right] \quad \text{as } i \to +\infty \\
&\to +\infty \quad \text{as } i \to +\infty.
\end{aligned}$$

Thus for all $0 < r < \min\{1/k_0, |{-\log(\lambda_*)}/\log(A)|\}$ we have

$$\lim_{i\to+\infty} \varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2}\rfloor} \alpha^{\sum_{j=0}^{k_0-1} \gamma_{i-lk_0+j}} = +\infty.$$

2.2. Bounding the approximation bias

In this subsection the main result is Proposition 5, which gives a sufficient condition for the approximation bias to vanish in the limit. To understand this result note first that, as γ_i tends to infinity, the probability distributions $\overline{\pi}^{\gamma_i}$ converge weakly to a limit $\overline{\pi}^{\infty}$ which is a mixture of point masses concentrated on $\mathcal{X}_{\max}$. This is a standard result which relies on Assumption 3 and [8]:

$$\overline{\pi}^{\infty}(\mathrm{d}x) = \lim_{\gamma\to+\infty} \frac{\sum_{\widetilde{x}\in\mathcal{X}_{\max}} (2\pi\gamma^{-1})^{\dim(\widetilde{x})/2}\alpha(\widetilde{x})\delta_{\widetilde{x}}(\mathrm{d}x)}{\sum_{\widetilde{x}\in\mathcal{X}_{\max}} (2\pi\gamma^{-1})^{\dim(\widetilde{x})/2}\alpha(\widetilde{x})},$$

where

$$\alpha(\widetilde{x}) := \left[\left|-\frac{\partial^2 \log \pi(x)}{\partial x_m \partial x_n}\right|\Bigg|_{x=\widetilde{x}}\right]^{-1/2},$$

which reduces to

$$\overline{\pi}^{\infty}(\mathrm{d}x) = \lim_{\gamma\to+\infty} \frac{\sum_{\widetilde{x}\in\mathcal{X}_{\max}} \alpha(\widetilde{x})\delta_{\widetilde{x}}(\mathrm{d}x)}{\sum_{\widetilde{x}\in\mathcal{X}_{\max}} \alpha(\widetilde{x})}$$

when $\mathcal{X} = \mathbb{R}^k$. Note that when $\mathcal{X} = \bigcup_{k=1}^{n} \mathbb{R}^k$ it is necessary to rescale the temperature according to the dimension of the current subspace to obtain the same result.

Lemma 3. *For each γ_i for which the integral exists, define $Z(\gamma_i) := \int_{\mathcal{X}} \widetilde{\pi}^{\gamma_i}(x)\,\mathrm{d}x$. We have*

$$\sum_{k=m}^{i-1} \|\overline{\pi}^{\gamma_k} - \overline{\pi}^{\gamma_{k+1}}\| \le 2\log\left(\frac{Z(\gamma_m)}{Z(\gamma_i)}\right)$$

whenever m is sufficiently large.

Proof. Due to Assumption 4 we extend the proof of Theorem 3.2 in [7, pp. 871–872] as follows. There it is assumed that $-\log(\pi)$ is bounded on $\mathcal{X}$ which allows differentiation under the integral sign. More generally, note that for i sufficiently large and all $x \in \mathcal{X}$

$$0 \geq \log(\widetilde{\pi}(x)) \exp[\gamma_i \log(\widetilde{\pi}(x))] \geq \log(\widetilde{\pi}(x))\widetilde{\pi}(x),$$

which under Assumption 4 and using the dominated convergence theorem allows us to write

$$-\frac{\mathrm{d}Z(\gamma_i)}{\mathrm{d}\gamma_i} = -\int_{\mathcal{X}} \log(\widetilde{\pi}(x)) \exp[\gamma_i \log(\widetilde{\pi}(x))]\,\mathrm{d}x.$$

This is the starting point of the proof of Theorem 3.2 in [7], which is then the same.

We now prove the main result of this section.

Proposition 5. *For all* $\zeta \in (0, 1)$ *and* $m_i = \lfloor i - i^{1-\zeta/2} \rfloor$, *if*

$$\gamma_i = \frac{\log(i + \varrho - k_0 + 1)}{-k_0 \log(\alpha)(1 + \zeta)}$$

(where $\alpha \in (0, 1)$ *is defined in Proposition 4, and* $\varrho > 0$*), then*

$$\lim_{i \to +\infty} \sum_{k=m_i}^{i-1} \|\pi^{\gamma_k} - \pi^{\gamma_{k+1}}\| = 0$$

under Assumption 4.

Proof. Under Assumption 4 we can apply the result of Lemma 3, so

$$\sum_{k=m_i}^{i-1} \|\pi^{\gamma_k} - \pi^{\gamma_{k+1}}\| \leq 2 \log\left(\frac{Z(\gamma_{m_i})}{Z(\gamma_i)}\right).$$

It is shown in [8] that

$$\lim_{\gamma \to +\infty} \left[Z(\gamma) - \sum_{\widetilde{x} \in \mathcal{X}_{\max}} (2\pi\gamma^{-1})^{\dim(\widetilde{x})/2} \alpha(\widetilde{x}) \right] = 0,$$

so by substituting the expression for γ_i we get

$$\lim_{i \to +\infty} 2 \log\left(\frac{Z(\gamma_{m_i})}{Z(\gamma_i)}\right) = \lim_{i \to +\infty} k \log\left[\frac{\log(i + \varrho - k_0 + 1)}{\log(m_i + \varrho - k_0 + 1)}\right] = 0.$$

2.3. The convergence result

We now combine Proposition 1, Proposition 2, Corollary 1, Proposition 4 and Proposition 5 into the following theorem.

Theorem 1. *Suppose Assumptions 1, 2, 3, 4 hold with (5),* $\mathrm{E}[V(X^{(0)})] < +\infty$ *and*

$$\gamma_i = \frac{\log(i + \varrho - k_0 + 1)}{-k_0 \log(\alpha)(1 + \zeta)} \quad (\textit{with } \varrho > 0),$$

then, as $i \to +\infty$,

$$\|\mathrm{P}(X^{(i)} \in \cdot) - \overline{\pi}^{\gamma_i}\| \le \prod_{l=1}^{j_i}(1 - \varepsilon_{i-lk_0}) + \lambda_*^{i-k_0+1} A^{j_i-1} B + k \log\left[\frac{\log(i+\varrho-k_0+1)}{\log(m_i+\varrho-k_0+1)}\right], \qquad (6)$$

where $A := \lambda_*^{k_0} \sup_{x \in C} V(x) + b_* \sum_{k=0}^{k_0-1} \lambda_*^k$, $B := b/(1-\lambda) + \frac{1}{2}\mathrm{E}[V(X^{(0)})]$, $\lambda_* := \lambda + b/(2(1+d))$ *and* $b_* := b(1+2d)/(2(1+d))$. *Consequently*

$$\lim_{i \to +\infty} \|\mathrm{P}(X^{(i)} \in \cdot) - \overline{\pi}^{\gamma_i}\| = 0$$

for $j_i = \lfloor r i^{1-\zeta/2} \rfloor$ *with* $0 < r < \min\{1/k_0, |-\log(\lambda_*)/\log(A)|\}$ *and* $m_i = \lfloor i - i^{1-\zeta/2} \rfloor$ *for some* $\zeta \in (0, 1)$.

The above theorem implies, using Corollary 5.4 of [7, p. 880], that

$$\lim_{i \to +\infty} \mathrm{P}(\widehat{\pi}(X^{(i)}) \ge 1 - \epsilon) = 1$$

for all $\epsilon > 0$.

Remark 1. It is possible to further simplify the bound in (6) in order to obtain a convergence rate, but we do not pursue this here.

3. Examples of applications

In this section we give a couple of examples of Markov chains to which the results of the previous section apply. Using the Metropolis–Hastings method, many homogeneous chains satisfying $\pi P = \pi$ can be designed, where P denotes the transition kernel and π is a positive and continuously differentiable arbitrary target distribution, and these chains can be adapted to the problem of interest. Note that the assumptions on the distribution can be weakened, at the expense of longer developments. Here we shall present two classes of chains that are among the simplest geometrically, but not uniformly, ergodic ones (Subsections 3.2 and 3.3). Specific examples are given in Subsection 3.4.

As explained in the introduction, it is usually assumed in the literature that the annealing chains (which are time inhomogeneous) satisfy a uniform minorization condition. In Assumption 2 this corresponds to taking $C = \mathcal{X}$. To construct such chains we usually start with a family of time homogeneous chains with respective stationary distributions $\overline{\pi}^{\gamma_i}$, all of which satisfy some uniform minorization condition. The latter is equivalent to uniform ergodicity [14]. If each chain with stationary distribution $\overline{\pi}^{\gamma_i}$ is only geometrically ergodic, a uniform minorization is impossible. This is when Assumptions 1 and 2 become useful. For each example below, our presentation focuses on checking the drift condition (Assumption 1) and the minorization condition (Assumption 2), after which Theorem 1 applies, for a suitable cooling schedule. No attempt has been made to discover a fastest cooling schedule.

3.1. On the condition involving λ, d and b

We start by proving a result that shows that the condition $d \ge b/(2(1-\lambda)) - 1$ necessary in Assumption 1 can be satisfied when the drift function $V(x)$ satisfies very general conditions. This result is implicitly used in the remaining sections of this paper.

Lemma 4. *Under the following assumptions:*

- *the following limit exists:*
$$\lim_{|x|\to+\infty} \frac{PV(x)}{V(x)} = \lambda < 1; \tag{7}$$
- *for any $d > 0$ the set $C_d = \{x;\ V(x) \leq d\}$ is small, i.e. Assumption 2 is satisfied (see also [14]);*
- *$PV(x)$ is bounded on any compact set;*
- *$V(x)$ goes to infinity when $|x|$ is large:*
$$\lim_{|x|\to+\infty} V(x) = +\infty; \tag{8}$$

there exists d such that

$$PV(x) \leq \lambda_d V(x) + b_d \mathbf{1}_{C_d}(x),$$

with $C_d = \{x;\ V(x) \leq d\}$,

$$\lambda_d = \sup_{x\in C_d^{\mathrm{c}}} \frac{PV(x)}{V(x)}, \tag{9}$$
$$b_d = \sup_{x\in C_d} (PV(x) - \lambda_d V(x))$$

and

$$d \geq \frac{b_d}{2(1-\lambda_d)} - 1.$$

Proof. For x sufficiently large, we have

$$\begin{aligned} PV(x) - \lambda_d V(x) &= (PV(x)/V(x) - \lambda_d)V(x) \\ &= \varepsilon_d(x)V(x), \end{aligned}$$

and we can write

$$\varepsilon_d(x) = (PV(x)/V(x) - \lambda + \lambda - \lambda_d),$$

where from (7) and (9)

$$\lim_{|x|\to+\infty} \frac{PV(x)}{V(x)} - \lambda = 0,$$
$$\lim_{d\to+\infty} \lambda - \lambda_d = 0^-.$$

Thus for any $\varepsilon > 0$ there exist $d_{0,\varepsilon}$ and $\alpha_{0,\varepsilon} > 0$ such that for $d \geq d_{0,\varepsilon}$, $\alpha > \alpha_{0,\varepsilon}$ and $x \in S_\alpha^{\mathrm{c}}$, where $S_\beta := \{x;\ |x| \leq \beta\}$ for $\beta > 0$, then

$$\frac{\varepsilon_d(x)}{2(1-\lambda_d)} < \varepsilon.$$

Choosing ε such that for any x

$$V(x) \geq \varepsilon V(x) - 1,$$

we deduce that for $d \geq d_{0,\varepsilon}$ and $x \in S_\alpha^c$ and when $C_d \cap S_\alpha^c \neq \varnothing$

$$d = \sup_{x \in C_d} V(x) \geq \sup_{x \in C_d \cap S_\alpha^c} V(x) \geq \varepsilon \sup_{x \in C_d \cap S_\alpha^c} V(x) - 1,$$

and thus

$$d \geq \sup_{x \in C_d \cap S_\alpha^c} \frac{PV(x) - \lambda_d V(x)}{2(1-\lambda_d)} - 1. \tag{10}$$

We now consider the case when $x \in C_d \cap S_\alpha$. We notice that

$$\begin{aligned}\sup_{x \in C_d \cap S_\alpha} (PV(x) - \lambda_d V(x)) &\leq \sup_{x \in C_d \cap S_\alpha} (PV(x) - \lambda V(x)) \\ &\leq \sup_{x \in S_\alpha} (PV(x) - \lambda) < +\infty\end{aligned}$$

from the assumptions on $PV(x)$ and $V(x)$. Note that the upper bound is independent of d. From (8) there exists $d \geq d_{\varepsilon,0}$ *sufficiently large* such that for $\alpha \geq \alpha_{0,\varepsilon}$ as defined above

$$d \geq \sup_{x \in C_d} V(x) \geq \frac{\sup_{x \in C_d \cap S_\alpha} (PV(x) - \lambda_d V(x))}{2(1-\lambda_d)} - 1$$

and together with (10) we conclude the proof.

3.2. Metropolis simulated annealing on $\mathbb{R}^k$

We begin with a simple and well known example of a Markov chain with specified target distribution. Suppose that we define an annealing algorithm as follows: set $X_1^{(0)} \sim v$, where v is some probability distribution, and $X_1^{(i+1)} = F_i(X_1^{(i)})$, where

$$F_i(x) = \begin{cases} x + w \ (w \sim q) & \text{if } (\pi(x+w)/\pi(x))^{\gamma_i} > \xi \sim \mathcal{U}_{[0,1]}, \\ x & \text{otherwise,} \end{cases} \tag{11}$$

and q is a symmetric density with respect to the Lebesgue measure on $\mathbb{R}^k$, bounded away from zero and $+\infty$ on a neighbourhood of the origin in $\mathbb{R}^k$ and such that there exist δ and ε so that

$$|w| < \delta \quad \Rightarrow \quad q(w) > \varepsilon. \tag{12}$$

Note that if we were to choose a sequence $(\gamma_i)_{i \in \mathbb{N}}$ to be constant and equal to 1, we would have an ordinary random walk based Metropolis algorithm with stationary density π on $\mathcal{X}$. Such a Markov chain is never uniformly ergodic (assuming π is not compactly supported), but can be geometrically ergodic. The maps F_i have transition probabilities

$$\mathrm{P}(F_i(x) \in \mathrm{d}y) := q(w) \min\left\{1, \left(\frac{\pi(y = x + w)}{\pi(x)}\right)^{\gamma_i}\right\} + \delta_x(\mathrm{d}y) r_{\gamma_i}(x)$$

so that (1) holds. The quantity $r_{\gamma_i}(x)$ is chosen such that $P_{\gamma_i}(\mathbb{R}^k \mid x) = 1$. We shall make the following assumptions on π [9].

- The density π is continuously differentiable and its tails are lighter than any exponential,

$$\lim_{|x| \to +\infty} \frac{x}{|x|} \cdot \nabla \log \pi(x) = -\infty. \tag{13}$$

- The contours $\partial A(x) = \{y : \pi(y) = \pi(x)\}$ are asymptotically regular, i.e.

$$\limsup_{|x|\to+\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0. \tag{14}$$

We can now state convergence of simulated annealing for this process.

Proposition 6. *Let $X_1^{(i+1)} = F_i(X_1^{(i)})$ where F_i is defined by (11). Assume that Assumption 3, (13), (14) and $\mathrm{E}[V(X_1^{(0)})] < +\infty$ (with $V(x) = \widehat{\pi}^{-1/2}(x)$) are satisfied and that we choose the series*

$$\gamma_i = \frac{\log(i + \varrho - k_0 + 1)}{-\log(\alpha)(1 + \zeta)}$$

(where α is obtained from (15), below, to correspond to the definition in Proposition 4; the quantities ζ, ϱ are also defined in Proposition 4). Then Theorem 1 holds.

We now discuss the four assumptions necessary for Theorem 1 to apply.

3.2.1. *Drift condition.* Assumption 1 is a direct consequence of the following lemma. Note that with our choice of drift function V, the constant λ is best possible for all kernels P_γ.

Lemma 5. *If (13) and (14) hold, then there exist d, C, λ and b such that for any $\gamma \geqslant 1$*

$$P_\gamma V(x) \le \lambda V(x) + b\mathbf{1}_C(x).$$

Proof. Jarner and Hansen [9] have shown the result for $\gamma = 1$. Here we notice simply that for $\gamma > 1$,

$$P_\gamma V(x) = V(x) + (P_\gamma - I)V(x)$$

(where I is the transition kernel identity, i.e. $I(x, \mathrm{d}y) := \delta_x(\mathrm{d}y)$). Then

$$\begin{aligned} P_\gamma V(x) &= V(x) + \int_{\mathsf{X}} q(x, \mathrm{d}y) \min\left\{1, \left(\frac{\pi(y)}{\pi(x)}\right)^\gamma \{\widehat{\pi}^{-1/2}(y) - \widehat{\pi}^{-1/2}(x)\}\right\} \\ &\le V(x) + \int_{\mathsf{X}} q(x, \mathrm{d}y) \min\left\{1, \frac{\pi(y)}{\pi(x)}\{\widehat{\pi}^{-1/2}(y) - \widehat{\pi}^{-1/2}(x)\}\right\} = PV(x), \end{aligned}$$

and the inequality holds because whenever $\pi^\gamma(y) \le \pi^\gamma(x)$ we also have $\widehat{\pi}^{-1/2}(y) \ge \widehat{\pi}^{-1/2}(x)$.

3.2.2. *Minorization condition.* Using (12), we obtain the minorization condition (Assumption 2).

Lemma 6. *For all $i \ge 0$ the compact subsets of $\mathbb{R}^k$ are small for P_{γ_i}.*

Proof. See Lemma 4 in [16] and its proof. We obtain for all x

$$P_\gamma(\mathrm{d}y \mid x) \ge \varepsilon\eta^\gamma(x, \tfrac{1}{2}\delta)\,\mathrm{d}y, \tag{15}$$

$$\eta(x, \tfrac{1}{2}\delta) := \frac{\inf_{w\in B(x,\delta/2)} \pi(w)}{\sup_{w\in B(x,\delta/2)} \pi(w)},$$

where $B(x, L)$ is the ball of radius L centred at x, and ε, δ are as in (12). Then, by convolving the transition kernels, we see that every compact set is a small set (See Proposition 5.5.5(ii) and Theorem 5.5.7 in [14]).

3.2.3. *Regularity of maxima and finite entropy conditions.* The regularity Assumption 3 is assumed and Assumption 4 (finite entropy) follows by (13).

3.3. Random scan hybrid Metropolis simulated annealing on $\mathbb{R}^k$

We wish to illustrate here how to handle the case of updates which occur one direction at a time only. Therefore, let us consider a variation on the previous example.

Let us consider the Markov chain $X_2^{(i)}$ that satisfies $X_2^{(0)} \sim \upsilon$, with υ some probability distribution, and $X_2^{(i+1)} = F_i(X_2^{(i)})$, where

$$F_i(x) = \begin{cases} j \sim \mathcal{U}_{\{1,\ldots,k\}} & \\ x + w_j e_j \ (w_j \sim q_j) & \text{if } (\pi(x + w_j e_j)/\pi(x))^{\gamma_i} > \xi \sim \mathcal{U}_{[0,1]}, \\ x & \text{otherwise.} \end{cases}$$

Here, e_j denotes the jth coordinate vector and q_j denotes a symmetric proposal distribution that updates x in direction e_j. We shall assume that $(q_j)_{j=1,\ldots,k}$ is a family of symmetric densities with respect to the Lebesgue measure such that there exist positive constants ε_j and δ_j such that for all $j = 1, \ldots, k$

$$q_j(w_j) \geq \varepsilon_j \quad \text{for } |w_j| < \delta_j.$$

Contrary to the Markov chain described in the previous subsection, here the k components are updated 'one variable at a time' at random. The homogeneous version of this algorithm ($\gamma_i = 1$) corresponds to the Metropolis algorithm 'one at a time' [23].

For each $\gamma > 1$, let $P_{\gamma_i,j}(\mathrm{d}y \mid x) := \mathrm{P}(F_i(x) \in \mathrm{d}y \mid j, x)$ be the kernel

$$P_{\gamma_i,j}(\mathrm{d}y \mid x) = q_j(w_j) \min\left\{1, \left(\frac{\pi(y = x + w_j e_j)}{\pi(x)}\right)^{\gamma_i}\right\} + \delta_x(\mathrm{d}y) r_{\gamma_i,j}(x),$$

where $r_{\gamma_i,j}(x)$ is such that $P_{\gamma_i,j}(\mathbb{R}^k \mid x) = 1$ and let $P_{\gamma_i,RS}$ be the kernel generating $X_2^{(i)}$,

$$P_{\gamma_i,RS}(\mathrm{d}y \mid x) := \mathrm{P}(F_i(x) \in \mathrm{d}y) = \sum_{j=1}^{k} \frac{1}{k} P_{\gamma_i,j}(\mathrm{d}y \mid x).$$

Conditions for the geometric convergence of the homogeneous Markov chain when $\gamma_i = 1$ for all i have been proposed [16], but are further simplified in [4]. We state here two assumptions. Assumption 5 is the weakest, and probably the most difficult to check in practice, but is implied by Assumption 6 which is generally easier to check.

Assumption 5. ([4].) *For any sequence $x^{(j)} \in \mathcal{X}$ such that $\lim_{j\to+\infty} |x^{(j)}| = +\infty$, a subsequence $\widetilde{x}^{(j)}$ can be extracted such that for some $i \in \{1, \ldots, k\}$ and all $y \in \mathbb{R}^+$,*

$$\lim_{j\to+\infty} \frac{p(\widetilde{x}^{(j)})}{p(\widetilde{x}^{(j)} \pm y e_i)} = 0 \quad \textit{and} \quad \lim_{j\to+\infty} \frac{p(\widetilde{x}^{(j)} \pm y e_i)}{p(\widetilde{x}^{(j)})} = 0. \tag{16}$$

Assumption 6. ([4].) *The target distribution π is continuously differentiable for large $|x|$ and for all $i \in \{1, \ldots, k\}$*

$$\begin{aligned} \lim_{x_i\to+\infty} \sup_{\{x_{-i} \in [-x_i, x_i]^{k-1}\}} \nabla_i \log \pi(x) &= -\infty, \\ \lim_{x_i\to+\infty} \inf_{\{x_{-i} \in [-x_i, x_i]^{k-1}\}} \nabla_i \log \pi(x) &= +\infty, \end{aligned} \tag{17}$$

where x_i is the ith coordinate of $x \in \mathcal{X}$, x_{-i} the remaining coordinates and $\nabla_i f$ the partial derivative of f in direction e_i.

We can now state convergence of simulated annealing for this process.

Proposition 7. *Define $(X_2^{(i)})_{i\in\mathbb{N}}$ as above, and suppose that (16) or (17),* $\mathrm{E}[V(X_2^{(0)})] < +\infty$ *(with $V(x) = \widehat{\pi}^{-c}(x)$ and some $c \in (0,1)$) hold and that*

$$\gamma_i = \frac{\log(i + \varrho - k_0 + 1)}{-k_0 \log(\alpha)(1+\zeta)}$$

(where α is obtained from (20), below, to correspond to the definition in Proposition 4). The quantities ζ, ϱ are also defined in Proposition 4 and k_0 depends on the choice of the small set C defined in Lemma 8, below. Then Theorem 1 holds for the chain $(X_2^{(i)})_{i\in\mathbb{N}}$.

The assumptions made can probably be simplified in the spirit of [9] with adaptations from [16], but we shall not investigate this here. We now discuss the four assumptions for Theorem 1 to apply.

3.3.1. *Drift condition.* The required drift condition is obtained through the following lemma.

Lemma 7. *If (16) and (17) hold, then for any compact set C and any temperature $\gamma \geqslant 1$*

$$P_{\gamma,RS}V(x) \leq \lambda V(x) + b\mathbf{1}_C(x),$$

where $V(x) = \widehat{\pi}(x)^{-c}$ for some $c \in (0,1)$.

Proof. In Theorem 1 of [4] the geometric convergence of the homogeneous chain corresponding to $\gamma_i = 1$ was proved for some $c \in (0,1)$. In particular it was proven that for the test function $V_0(x) = \pi^{-c}(x)$

$$\limsup_{|x|\to+\infty} \frac{P_{1,RS}V_0(x)}{V_0(x)} < 1. \tag{18}$$

We can easily check that the proof extends to the test function $V(x) = \widehat{\pi}^{-c}(x)$. In our case we would like to prove that

$$\limsup_{|x|\to+\infty} \frac{P_{\gamma,RS}V(x)}{V(x)} < 1 \tag{19}$$

uniformly for γ. Here for contradiction, suppose that we have a sequence of points $x^{(i)}$ such that

$$\limsup_{i\to+\infty} \frac{P_{\gamma,RS}V(x^{(i)})}{V(x^{(i)})} \geq 1,$$

but then, with an argument similar to that used in Lemma 5,

$$P_{\gamma,RS}V(x) \leq P_{1,RS}V(x)$$

for $\gamma > 1$ and necessarily

$$\limsup_{i\to+\infty} \frac{P_{1,RS}V_0(x^{(i)})}{V_0(x^{(i)})} \geq 1,$$

which contradicts (18). Therefore (19) holds uniformly for γ.

3.3.2. *Minorization condition.* To prove the minorization condition, we will need the following lemma.

Lemma 8. *For all* $i \geq 0$ *the compact subsets of* $\mathbb{R}^k$ *are small for* $P_{\gamma_i,RS}$.

Proof. The proof is straightforward, and follows the proof of Lemma 4 in [16], see also Lemma 6 in the present paper. Calling $s_{\gamma_i}(x, y)$ the continuous component of the transition kernel $P_{\gamma_i,RS}$, we have for suitable constants a_1, a_2 and δ

$$s_{\gamma_i} \ldots s_{\gamma_{i+k-1}}(x, y) \geq \frac{1}{k^k} a_1^k \prod_{j=i}^{i+k-1} a_2^{\gamma_j} \quad \text{whenever } |w_i| < \tfrac{1}{2}\delta \text{ for } i = 1, \ldots, k.$$

Then by composition for any N and $x \in [-\frac{1}{2}N\delta, \frac{1}{2}N\delta]$

$$s_{\gamma_i} \ldots s_{\gamma_{i+kN-1}}(x, y) \geq \frac{1}{k^{kN}} a_1^{kN} \prod_{j=i}^{i+kN-1} a_2^{\gamma_j} \quad \text{whenever } |w_i| < \tfrac{1}{2}N\delta \text{ for } i = 1, \ldots, k, \tag{20}$$

which shows that $[-\frac{1}{2}N\delta, \frac{1}{2}N\delta]$ is a small set for $P^{kN}_{\gamma_i,RS}$. The end of the proof is the same as for Lemma 4 in [16].

3.3.3. *Regularity of maxima and finite entropy conditions.* The regularity condition (Assumption 3) is assumed. The finite entropy condition (Assumption 4) follows by the assumptions on π.

3.4. Examples

Specific examples of distributions for which the Metropolis algorithm and the random scan Metropolis algorithm are geometrically ergodic are reported in [9] and [4] respectively. Note that the examples given here are mostly trivial from an optimization point of view. They could be made more complex by adding a function difficult to optimize to π (see Equation (1.33) in [12] for example). However the examples given here are interesting as they demonstrate that simulated annealing is sound for global optimization on a noncompact set, as long as some tail properties are satisfied.

Example 1. The conditions of Theorem 1 are satisfied for a random walk Metropolis annealing algorithm in order to maximize the density (defined on $\mathbb{R}^k$)

$$\pi(x) \propto \exp(-|x|^s) \quad \text{with } s > 1.$$

Example 2. The conditions of Theorem 1 are *not* satisfied for a random walk Metropolis annealing algorithm in order to maximize the density

$$\pi(x_1, x_2) \propto \exp(-0.01(x_1^2 + x_1^2 x_2^2 + x_2^2)).$$

See Figure 1 for a contour plot of π.

Example 3. However, the conditions of Theorem 1 are satisfied for the *random scan* random walk Metropolis annealing algorithm in order to optimize

$$\pi(x_1, x_2) \propto \exp(-0.01(x_1^2 + x_1^2 x_2^2 + x_2^2)),$$

demonstrating the interest of such a 'one variable at a time' scheme, often recommended in practice.

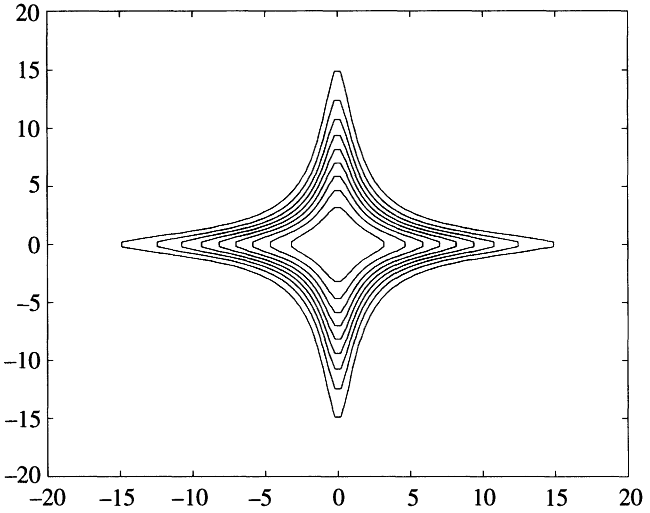

FIGURE 1: Contour plot of $\exp(-0.01(x_1^2 + x_1^2 x_2^2 + x_2^2))$.

Acknowledgements

We are grateful to Søren F. Jarner and Éric Moulines for useful discussions. Christophe Andrieu was financially supported by AT&T Research Laboratories, Cambridge, UK. Laird Breyer acknowledges the support of EU TMR network ERB-FMRX-CT96-0095 on Computational and Statistical Methods for the Analysis of Spatial Data. Arnaud Doucet was financially supported by EPSRC, UK.

References

[1] ANDRIEU, C. AND DOUCET, A. (2000). Simulated annealing for maximum a posteriori parameter estimation of hidden Markov models. *IEEE Trans. Inf. Theory* **46,** 994–1004.

[2] BILLINGSLEY, P. (1985). *Probability and Measure*, 2nd edn. John Wiley, Chichester.

[3] FORT, G. AND MOULINES, É. (2000). V-subgeometric ergodicity for a Hastings–Metropolis algorithm. *Statist. Prob. Lett.* **49,** 401–410.

[4] FORT, G., MOULINES, É. AND ROBERTS, G. O. (2001). A geometrically ergodic hybrid sampler for sub-exponential densities. Submitted.

[5] GELFAND, S. B. AND MITTER, S. K. (1993). Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$. *SIAM J. Control Optimization* **31,** 111–131.

[6] GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intellig.* **6,** 721–741.

[7] HAARIO, A. AND SAKSMAN, E. (1991). Simulated annealing in general state space. *Adv. Appl. Prob.* **23,** 866–893.

[8] HWANG, C. (1980). Laplace's method revisited: weak convergence of probability measures. *Ann. Prob.* **8,** 1177–1182.

[9] JARNER, S. F. AND HANSEN, E. (1998). Geometric ergodicity of Metropolis algorithms. *Stoch. Process. Appl.* **85,** 341–361.

[10] JARNER, S. F. AND ROBERTS, G. (2001). Polynomial convergence rates of Markov chains. To appear in *Ann. Appl. Prob.*

[11] LINDVALL, T. (1992). *Lectures on the Coupling Method.* John Wiley, New York.

[12] LOCATELLI, M. (2000). Simulated annealing algorithms for continuous global optimization: convergence conditions. *J. Optimization Theory Appl.* **104,** 121–133.

[13] MENGERSEN, K. L. AND TWEEDIE, R. L. (1994). Rates of convergence of the Hastings and Metropolis algorithm. *Ann. Statist.* **24,** 101–121.

[14] MEYN, S. P. AND TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer, Berlin.

[15] MITRA, D., ROMEO, F. AND SANGIOVANNI-VINCENTELLI, A. (1986). Convergence and finite-time behavior of simulated annealing. *Adv. Appl. Prob.* **18,** 747–771.

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

[17] ROBERTS, G. O. AND TWEEDIE, R. L. (1996). Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli* **2,** 341–364.

[18] ROBERTS, G. O. AND TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83,** 96–110.

[19] ROBERTS, G. O. AND TWEEDIE, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stoch. Process. Appl.* **80,** 211–229.

[20] ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90,** 558–566.

[21] TWEEDIE, R. AND STRAMER, O. (1999). Langevin-type models I: diffusions with given stationary distributions, and their discretizations. *Methodology Comput. Appl. Prob.* **1,** 283–306.

[22] TWEEDIE, R. AND STRAMER, O. (1999). Langevin-type models II: self-targeting candidates for MCMC algorithms. *Methodology Comput. Appl. Prob.* **1,** 307–328.

[23] TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22,** 1701–1762.

[24] VAN LAARHOVEN, P. J. AND ARTS, E. H. L. (1987). *Simulated Annealing: Theory and Applications.* Reidel, Amsterdam.