

Convergence of Simulated Annealing using Foster-Lyapunov Criteria

Christophe Andrieu¹, Laird A. Breyer² and Arnaud Doucet³

¹Department of Mathematics, Statistics Group,
University of Bristol,

University Walk, Bristol BS8 1TW, U.K.

²Department of Mathematics and Statistics,
Fylde College, Lancaster University, Lancaster, LA1 4YF, U.K.

³Department of Electrical and Electronic Engineering,
University of Melbourne,
Parkville, Victoria 3010, Australia.

Email: c.andrieu@bris.maths.ac.uk - laird@lbreyer.com -
doucet@ee.mu.oz.au

September 7, 2001

¹Corresponding author.

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, non-homogeneous Markov chain

Contents

1	Introduction	2
2	Convergence under general conditions	4
2.1	Bounding the estimation error	4
2.1.1	Bounding the term $\mathbb{P}(N_{i-k_0+1,m} < j)$	6
2.1.2	Bounding the term $\prod_{l=1}^j (1 - \varepsilon_{i-lk_0})$	8
2.2	Bounding the approximation bias	9
2.3	The convergence result	10
3	Examples of applications	11
3.1	On the condition on λ, d and b	11
3.2	Metropolis simulated annealing on \mathbb{R}^k	12
3.2.1	Drift condition	13
3.2.2	Minorization condition	13
3.2.3	Regularity of maxima and finite entropy conditions	14
3.3	Random scan hybrid Metropolis simulated annealing on \mathbb{R}^k	14
3.3.1	Drift condition	15
3.3.2	Minorization condition	15
3.3.3	Regularity of maxima and finite entropy conditions	16
3.4	Examples	16
4	Acknowledgements	16

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 [Van Laarhoven and Arts (1987)], [Locatelli (2000)]. Technically speaking, a uniform minorization condition (or Doeblin condition) is assumed for both strong and weak convergence of the process [Van Laarhoven and Arts (1987)]. 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 non compact 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 [Roberts and Tweedie (1996)], [Roberts and Tweedie (1996)], [Rosenthal (1995)], [Stramer and Tweedie (1999a)], [Stramer and Tweedie (1999b)]. 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. The simulated annealing is one of them and has found many applications in statistical image processing [Geman and Geman (1984)], statistical signal processing [Andrieu and Doucet (2000)], physics and computer science [Van Laarhoven and Arts (1987)]. The method exploits the fact that extrema are preserved under monotone transformations. If γ is any positive real number, then the probability density

$$\bar{\pi}^\gamma(x) \triangleq \frac{\pi(x)^\gamma}{\int_{\mathcal{X}} \pi(y)^\gamma dy}$$

also has the same global maxima. Here dx 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 $\bar{\pi}^\gamma$ 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 $\bar{\pi}^\gamma$ 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

$$P_{\gamma_i}(dy|x) \triangleq \mathbb{P}\left(X^{(i+1)} \in dy \mid X^{(i)} = x\right),$$

are such that

$$\int_{\mathcal{X}} \bar{\pi}^{\gamma_i}(x) P_{\gamma_i}(dy|x) dx = \bar{\pi}^{\gamma_i}(y) dy. \quad (1)$$

The aim of this paper is to give conditions under which $\|\mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^{\gamma_i}\| \rightarrow 0$ as $i \rightarrow +\infty$, 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 hold, the distributions $(\bar{\pi}^{\gamma_i})_{i \in \mathbb{N}}$ converge weakly to a mixture of point masses concentrated on the set

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

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

$$\lim_{i \rightarrow +\infty} \mathbb{P}\left(\hat{\pi}\left(X^{(i)}\right) \geq 1 - \epsilon\right) = 1.$$

where $\hat{\pi}(x) \triangleq \pi(x) / \|\pi\|_\infty$. 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 $\lim_{i \rightarrow +\infty} \|\mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^\infty\| = 0$, when the set \mathcal{X}_{\max} is of null measure, see [Haario and Sacksman (1991), 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 [Geman and Geman (1984)], [Mitra *et al.* (1986)], [Van Laarhoven and Arts (1987)].

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

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

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

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

Most cases treated in the literature assume that the transition kernels P_{γ_i} each satisfy a uniform ergodicity assumption *i.e.* there exists an integer k_0 and constants $\varepsilon_i > 0$ such that $\mathbb{P}(X^{(i+k_0)} \in dy | X^{(i)} = x) \geq \varepsilon_i \mu_i(y) dy$ 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*

$$\mathbb{P}(X^{(i+k_0)} \in dy | X^{(i)} = x) \geq \varepsilon_i \mu_i(y) dy, \text{ if } x \in C.$$

Here the constants ε_i are assumed non increasing, 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 [Fort and Moulines (2000)], [Jarner and Roberts (2001)] 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 $(\bar{\pi}^{\gamma_i})_{i \in \mathbb{N}}$ 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 $\left[-\frac{\partial^2 \log \pi(x)}{\partial x_i \partial x_j} \Big|_{\tilde{x}}\right]$ is a $k \times k$ matrix whose determinant is non-zero.*

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.**

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

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

$$\begin{aligned} \left\| \mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^{\gamma_i} \right\| &\leq \left\| \mathbb{P}(X^{(i)} \in \cdot) - \mathbb{P}(X^{(i)} \in \cdot | X^{(m)} \sim \bar{\pi}^{\gamma_m}) \right\| \\ &\quad + \sum_{k=m}^{i-1} \|\bar{\pi}^{\gamma_{k+1}} - \bar{\pi}^{\gamma_k}\|, \end{aligned} \tag{2}$$

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 $i \rightarrow +\infty$. 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 $i \rightarrow +\infty$.

The main result is given in Theorem 10 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 $\|\mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^{\gamma_i}\|$ 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 10 in Subsection 2.3.

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

$$\begin{aligned} \left\| \mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^{\gamma_i} \right\| &\leq \left\| \mathbb{P}(X^{(i)} \in \cdot) - \mathbb{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \bar{\pi}^{\gamma_m}) \right\| \\ &\quad + \sum_{k=m}^{i-1} \|\bar{\pi}^{\gamma_{k+1}} - \bar{\pi}^{\gamma_k}\|. \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}(X^{(i)} \in dx) - \bar{\pi}^{\gamma_i}(dx) &= \mathbb{P}(X^{(i)} \in dx) - \mathbb{P}(X^{(i)} \in dx \mid X^{(m)} \sim \bar{\pi}^{\gamma_m}) \\ &\quad + \sum_{k=m}^{i-1} \mathbb{P}(X^{(i)} \in dx \mid X^{(k)} \sim \bar{\pi}^{\gamma_k}) - \mathbb{P}(X^{(i)} \in dx \mid X^{(k+1)} \sim \bar{\pi}^{\gamma_{k+1}}). \end{aligned} \quad (3)$$

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

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

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

$$\left\| \mathbb{P}(X^{(i)} \in \cdot) - \mathbb{P}(X^{(i)} \in \cdot \mid X^{(m)} \sim \bar{\pi}^{\gamma_m}) \right\|$$

by a standard coupling construction [Lindvall (1992)], and using a method pioneered by [Rosenthal (1995)] 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 $F_i : \mathcal{X} \rightarrow \mathcal{X}$ such that $X^{(i+1)} = F_i(X^{(i)})$. In this case we also have

$$P_{\gamma_i}(dy \mid x) = \mathbb{P}(F_i(x) \in dy).$$

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 $i \geq m$.

- Initialisation $\phi_{1,m}^{(m)} \sim \mathbb{P}(X^{(m)} \in \cdot)$, $\phi_{2,m}^{(m)} \sim \bar{\pi}^{\gamma_m}$, $d = 0$ 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 ε_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 $\phi_{j,m}^{(i+k_0)} \sim \frac{\mathbb{P}(X^{(i+k_0)} \in \cdot | \phi_{j,m}^{(i)}) - \varepsilon_i \mu_i(X^{(i+k_0)} \in \cdot)}{1 - \varepsilon_i}$, for $j = 1, 2$.
3. $\phi_{j,m}^{(i+1)}, \dots, \phi_{j,m}^{(i+k_0-1)} \sim \mathbb{P}(X^{(i+1)} \in \cdot, \dots, X^{(i+k_0-1)} \in \cdot | X^{(i)} = \phi_{j,m}^{(i)}, X^{(i+k_0)} = \phi_{j,m}^{(i+k_0)})$, 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

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

and for $k \geq 2$

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

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

$$\left\| \mathbb{P}(X^{(i)} \in \cdot) - \mathbb{P}(X^{(i)} \in \cdot | X^{(m)} \sim \bar{\pi}^{\gamma_m}) \right\| \leq \prod_{l=1}^j (1 - \varepsilon_{i-lk_0}) + \mathbb{P}(N_{i-k_0+1,m} < j). \quad (4)$$

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

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

as the event on the second line is included in that of the third line. One has the following identity

$$\mathbb{P}(\tau_{C,m} > i) = \mathbb{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} \geq j) + \mathbb{P}(\tau_{C,m} > i \text{ and } N_{i-k_0+1,m} < j). \quad (5)$$

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

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

since the series ε_i is non increasing. Then using identity (5) we obtain

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

as

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

■

2.1.1 Bounding the term $\mathbb{P}(N_{i-k_0+1,m} < j)$

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 10. It is stated in Proposition 3 (compare with [Rosenthal (1995)] Lemma 3) and its proof is a direct combination of Lemmas 5 and 6 below.

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

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

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 $\lambda_* \triangleq \lambda + b/(2(1+d))$ and $b_* \triangleq b(1+2d)/(2(1+d))$ then we have

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

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

Corollary 4 *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 $0 < r < \min \left\{ 1/k_0, \left| \frac{-\log(\lambda_*)}{\log(A)} \right| \right\}$ then*

$$\lim_{i \rightarrow +\infty} \mathbb{P}(N_{i-k_0+1,m_i} < j_i) = 0. \quad (6)$$

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

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

holds, allowing us to use Proposition 3. If $A \leq 1$ then (6) holds for any $r > 0$. If $A > 1$ then by noting that for $r < \frac{-\log(\lambda_*)}{\log(A)}$ we have $\lambda_* A^r < 1$ the result of (6) still holds. ■

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

Lemma 5 *Fix m, i and j such that $i > k_0 + m$ and define $r_1 = r_{1,m} \triangleq \tau_{1,m} - m$ and $r_i = r_{i,m} \triangleq \tau_{i+1,m} - \tau_{i,m}$ for $i > 1$. We have*

$$\mathbb{P}(N_{i-k_0+1,m} < j) \leq \lambda_*^{i-k_0+1} \mathbb{E}(\lambda_*^{-r_1} \mathbb{E}(\lambda_*^{-r_2} \mathbb{E}(\lambda_*^{-r_3} \dots \mathbb{E}(\lambda_*^{-r_j} | r_1, \dots, r_{j-1}) \dots | r_1, r_2) | r_1)) \quad (7)$$

Proof. Let us write

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

by Markov's inequality. Then the result follows from the identity $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y|X]]$. ■

This is the second technical lemma we need.

Lemma 6 *Define $V^*(x_1, x_2) \triangleq \frac{1}{2}[V(x_1) + V(x_2)]$. From the drift and minorization conditions (Assumptions 1 and 2), assuming $\mathbb{E}[V(X^{(0)})] < +\infty$, we have for any m :*

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

$$\mathbb{E}(\lambda_*^{-r_{1,m}}) \leq \sup_{i > m} \mathbb{E}\left(V^*\left(\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)}\right)\right) \leq \frac{b}{1-\lambda} + \frac{1}{2} \mathbb{E}[V(X^{(0)})] < +\infty. \quad (8)$$

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

$$\mathbb{E} \left(\lambda_*^{-r_{i,m}} \mid r_{1,m}, \dots, r_{i-1,m} \right) \leq \lambda_*^{k_0} \sup_{x \in C} V(x) + b_* \sum_{k=0}^{k_0-1} \lambda_*^k. \quad (9)$$

Proof. Following [Rosenthal (1995)] and by Theorem 5.2 of [Roberts and Tweedie (1999)] observe that by the drift condition (Assumption 1)

$$\mathbb{E} \left[V^* \left(\phi_{1,m}^{(i)}, \phi_{2,m}^{(i)} \right) \mid \phi_{1,m}^{(i-1)}, \phi_{2,m}^{(i-1)} \right] \leq \lambda_* V^* \left(\phi_{1,m}^{(i-1)}, \phi_{2,m}^{(i-1)} \right) + b_* \mathbb{I}_C \left(\phi_{1,m}^{(i-1)} \right) \mathbb{I}_C \left(\phi_{2,m}^{(i-1)} \right). \quad (10)$$

Then, following [Rosenthal (1995)], we introduce the functions

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

whose series has non increasing (in k) expectation from (10) (at least for $\tau_{i-1} + k_0 \leq k$ when $i > 1$ as before τ_i one does not go into C , and for $k \geq m$ when $i = 1$). Thus, and because $V^*(\cdot, \cdot) \geq 1$, one obtains

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

From Assumption 1 one has for appropriate j and i

$$P_{\gamma_i} V \left(\phi_{j,m}^{(i)} \right) \leq \lambda V \left(\phi_{j,m}^{(i)} \right) + b \mathbb{I}_C \left(\phi_{j,m}^{(i)} \right),$$

from which we deduce that:

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

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

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

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

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

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

which by recursion leads for any $i \geq 0$

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

that is if $\mathbb{E} \left[V \left(X^{(0)} \right) \right] < +\infty$

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

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

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

which is the first result we wanted to prove.

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

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

Contrary to [Rosenthal (1995)] we cannot yet conclude as the Markov chain is non homogeneous. We therefore find an upper bound for this term. For $k_0 > 1$

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

where we have used (10) to obtain the inequality.

Successively applying (k_0 times) this procedure we obtain

$$\mathbb{E} \left[V^* \left(\phi_{1,m}^{(\tau_{i-1,m} + k_0)}, \phi_{2,m}^{(\tau_{i-1,m} + k_0)} \right) \mid \phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})} \right] \leq \lambda_*^{k_0} V^* \left(\phi_{1,m}^{(\tau_{i-1,m})}, \phi_{2,m}^{(\tau_{i-1,m})} \right) + b_* \sum_{k=0}^{k_0-1} \lambda_*^k,$$

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 $j \rightarrow +\infty$ if and only if $\sum_{l=1}^j \varepsilon_{i-lk_0}$ diverges as $j \rightarrow +\infty$. 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 $\varepsilon_i = \varepsilon \alpha \sum_{j=0}^{k_0-1} \gamma_{i+j}$ with $\alpha \in (0, 1)$ and $\varepsilon > 0$ as it is a case often met in practice. Such examples will be met in Section 3 for our two applications with the Metropolis algorithm, but can also be found in [Geman and Geman (1984)], [Van Laarhoven and Arts (1987)]. 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 $j \rightarrow +\infty$ (j being a function in i). We have the following proposition:

Proposition 7 *For all i we assume that $\varepsilon_i = \varepsilon \alpha \sum_{j=0}^{k_0-1} \gamma_{i+j}$ and if we choose $\gamma_i = \frac{\log(i+\varrho-k_0+1)}{-k_0 \log(\alpha)(1+\zeta)}$ (where $\alpha \in (0, 1)$, $\varepsilon > 0$ and $\varrho > 0$). Then for $j_i = \lfloor ri^{1-\zeta/2} \rfloor$ (where $\zeta \in (0, 1)$) with $0 < r < \min \left\{ 1/k_0, \left| \frac{-\log(\lambda_*)}{\log(A)} \right| \right\}$ we obtain*

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

Proof. We follow [Haario and Sacksman (1991)] for the proof. We denote $a \triangleq 1/(1+\zeta)$ 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:

$$\begin{aligned}
\varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2} \rfloor} \alpha \sum_{j=0}^{k_0-1} \gamma_{i-lk_0+j} &> \varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2} \rfloor} \alpha^{k_0} (\gamma_{i-lk_0+k_0-1}) \\
&> \varepsilon \sum_{l=1}^{\lfloor ri^{1-\zeta/2} \rfloor} (i-lk_0+\varrho)^{-a}
\end{aligned}$$

$$\begin{aligned}
&> \frac{\varepsilon}{k_0 a \zeta} \int_{i-k_0 \lfloor r i^{1-\zeta/2} \rfloor}^i (x + \varrho)^{-a} dx \\
&> \frac{\varepsilon}{k_0 a \zeta} \left[(i + \varrho)^{a\zeta} - \left(i - k_0 \lfloor r i^{1-\zeta/2} \rfloor + \varrho \right)^{a\zeta} \right] \\
&> \frac{\varepsilon}{k_0 a \zeta} (i + \varrho)^{a\zeta} \left[1 - \left(1 - \frac{k_0 r i^{1-\zeta/2}}{i + \varrho} \right)^{a\zeta} \right] \\
&\stackrel{i \rightarrow +\infty}{>} \frac{\varepsilon}{k_0 a \zeta} (i + \varrho)^{a\zeta} \left[1 - \left(1 - a\zeta \frac{k_0 r i^{1-\zeta/2}}{i + \varrho} \right) \right] \\
&\stackrel{i \rightarrow +\infty}{\rightarrow} +\infty.
\end{aligned}$$

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

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

■

2.2 Bounding the approximation bias

In this subsection the main result is Proposition 9, 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 π^{γ_i} converge weakly to a limit π^∞ which is a mixture of point masses concentrated on \mathcal{X}_{\max} . This is a standard result which relies on Assumption 3 and [Hwang (1980)]:

$$\pi^\infty(dx) = \lim_{\gamma \rightarrow +\infty} \frac{\sum_{\tilde{x} \in \mathcal{X}_{\max}} (2\pi\gamma^{-1})^{\dim(\tilde{x})/2} \alpha(\tilde{x}) \delta_{\tilde{x}}(dx)}{\sum_{\tilde{x} \in \mathcal{X}_{\max}} (2\pi\gamma^{-1})^{\dim(\tilde{x})/2} \alpha(\tilde{x})}, \quad \alpha(\tilde{x}) \triangleq \left[\left\| -\frac{\partial^2 \log \pi(x)}{\partial x_m \partial x_n} \right\|_{x=\tilde{x}} \right]^{-1/2}, \quad (11)$$

which reduces to

$$\pi^\infty(dx) = \lim_{\gamma \rightarrow +\infty} \frac{\sum_{\tilde{x} \in \mathcal{X}_{\max}} \alpha(\tilde{x}) \delta_{\tilde{x}}(dx)}{\sum_{\tilde{x} \in \mathcal{X}_{\max}} \alpha(\tilde{x})}$$

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 8 For each γ_i for which the integral exists, define $Z(\gamma_i) \triangleq \int_{\mathcal{X}} \hat{\pi}^{\gamma_i}(x) dx$. We have

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

whenever m is sufficiently large.

Proof. Due to Assumption 4 we extend the proof of Theorem 3.2. in [Haario and Sacksman (1991), pp. 871-872] as follows. Haario and Sacksman assume that $-\log(\pi)$ is bounded on \mathcal{X} which allows them to differentiate under the integral sign. More generally, note that for i sufficiently large and all $x \in \mathcal{X}$

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

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

$$-\frac{dZ(\gamma_i)}{d\gamma_i} = - \int_{\mathcal{X}} \log(\hat{\pi}(x)) \exp[\gamma_i \log(\hat{\pi}(x))] dx.$$

This is the starting point of the proof of Theorem 3.2. in [Haario and Sacksman (1991)], which is then the same. ■

We now prove the main result of this section.

Proposition 9 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 7, and $\varrho > 0$), then

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

under Assumption 4.

Proof. Under Assumption 4 we can apply the result of Lemma 8

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

It is shown in [Hwang (1980)] that

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

so by substituting the expression for γ_i we get

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

■

2.3 The convergence result

We now combine Proposition 1, Proposition 2, Corollary 4 of Proposition 3, Proposition 7, and Proposition 9 into the following theorem.

Theorem 10 Suppose Assumptions 1, 2, 3, 4, hold with $\varepsilon_i = \varepsilon \alpha^{\sum_{j=0}^{k_0-1} \gamma_{i+j}}$ for some $\alpha \in (0, 1)$ and $\varepsilon > 0$, $\mathbb{E}[V(X^{(0)})] < +\infty$ and $\gamma_i = \frac{\log(i+\varrho-k_0+1)}{-k_0 \log(\alpha)(1+\zeta)}$ (with $\varrho > 0$), then, as $i \rightarrow +\infty$

$$\left\| \mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^{\gamma_i} \right\| \leq \prod_{l=1}^{j_i} (1 - \varepsilon_{i-lk_0}) + \lambda_*^{i-k_0+1} A^{j_i-1} B + k \log \left[\frac{\log(i+\varrho-k_0+1)}{\log(m_i+\varrho-k_0+1)} \right], \quad (12)$$

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

$$\lim_{i \rightarrow +\infty} \left\| \mathbb{P}(X^{(i)} \in \cdot) - \bar{\pi}^{\gamma_i} \right\| = 0$$

for $j_i = \lfloor ri^{1-\zeta/2} \rfloor$ with $0 < r < \min \left\{ 1/k_0, \left| \frac{-\log(\lambda_*)}{\log(A)} \right| \right\}$ 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 [Haario and Sacksman (1991), pp. 880] that

$$\lim_{i \rightarrow +\infty} \mathbb{P}(\hat{\pi}(X^{(i)}) \geq 1 - \epsilon) = 1$$

for all $\epsilon > 0$.

Remark 1 It is possible to further simplify the bound in Eq. (12) 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 one usually starts with a family of time homogeneous chains with respective stationary distributions $\bar{\pi}^{\gamma_i}$, all of which satisfy some uniform minorization condition. The latter is equivalent to uniform ergodicity [Meyn and Tweedie (1993)]. If each chain with stationary distribution $\bar{\pi}^{\gamma_i}$ 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 10 applies, for a suitable cooling schedule. No attempt has been made to discover a fastest cooling schedule.

3.1 On the condition on λ, d and b

We first start proving a result that shows that the condition $d \geq b / (2(1 - \lambda)) - 1$ 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 11 *Under the three following assumptions:*

- *The following limit exists*

$$\lim_{|x| \rightarrow +\infty} PV(x) / V(x) = \lambda < 1, \quad (13)$$

- *For any $d > 0$ the set $C_d = \{x; V(x) \leq d\}$ is small, i.e. Assumption 2 is satisfied (see also [Meyn and Tweedie (1993)]).*
- *$PV(x)$ is bounded on any compact set.*
- *$V(x)$ goes to infinity when $|x|$ is large*

$$\lim_{|x| \rightarrow +\infty} V(x) = +\infty, \quad (14)$$

then there exists d such that

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

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

$$\begin{aligned} \lambda_d &= \sup_{x \in C_d^c} PV(x) / V(x), \\ b_d &= \sup_{x \in C_d} (PV(x) - \lambda_d V(x)), \end{aligned} \quad (15)$$

and

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

Proof. For x sufficiently large, we have

$$\begin{aligned} PV(x) - \lambda_d V(x) &= (PV(x) / V(x) - \lambda_d) V(x) \\ &= \varepsilon_d(x) V(x), \end{aligned} \quad (16)$$

and we can write

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

where from (13) and (15)

$$\begin{aligned} \lim_{|x| \rightarrow +\infty} PV(x)/V(x) - \lambda &= 0, \\ \lim_{d \rightarrow +\infty} \lambda - \lambda_d &= 0^-. \end{aligned}$$

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 $x \in S_\alpha^c$, where $S_\beta \triangleq \{x; |x| \leq \beta\}$ 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 $C_d \cap S_\alpha^c \neq \emptyset$

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

and thus

$$d \geq \sup_{x \in C_d \cap S_\alpha^c} \frac{(PV(x) - \lambda_d V(x))}{2(1-\lambda_d)} - 1. \quad (17)$$

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 (14) there exists $d \geq d_{\varepsilon,0}$ sufficiently large such that for $\alpha \geq \alpha_{0,\varepsilon}$ as defined above

$$d \geq \sup_{x \in C_d} V(x) \geq \frac{\sup_{x \in C_d \cap S_\alpha} (PV(x) - \lambda_d V(x))}{2(1-\lambda_d)} - 1 \quad (18)$$

and together with (17) we conclude. ■

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

$$F_i(x) = \begin{cases} x + w & (w \sim q) \text{ if } (\pi(x+w)/\pi(x))^{\gamma_i} > \xi \sim \mathcal{U}_{[0,1]} \\ x & \text{otherwise,} \end{cases} \quad (19)$$

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

$$|w| < \delta \Rightarrow q(w) > \varepsilon. \quad (20)$$

Note that if we were to choose a sequence $(\gamma_i)_{i \in \mathbb{N}}$ to be constant and equal to one, 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

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

so that (1) p. 2 holds. The quantity $r_{\gamma_i}(x)$ is chosen such that $P_{\gamma_i}(\mathbb{R}^k | x) = 1$. We shall make the following assumptions on π [Jarner and Hansen (1998)].

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

$$\lim_{|x| \rightarrow +\infty} \frac{x}{|x|} \cdot \nabla \log \pi(x) = -\infty, \quad (21)$$

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

$$\limsup_{|x| \rightarrow +\infty} \frac{x}{|x|} \cdot \frac{\nabla \pi(x)}{|\nabla \pi(x)|} < 0. \quad (22)$$

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

Proposition 12 *Let $X_1^{(i+1)} = F_i(X_1^{(i)})$ where F_i is defined by (19). Assume that Assumption 3, (21)-(22) and $\mathbb{E}[V(X_1^{(0)})] < +\infty$ (with $V(x) = \hat{\pi}^{-1/2}(x)$) are satisfied and that we choose the series $\gamma_i = \frac{\log(i+q-k_0+1)}{-\log(\alpha)(1+\zeta)}$ (where α is obtained from (23) to correspond to the definition in Proposition 7. The quantities ζ, q are also defined in Proposition 7). Then Theorem 10 holds.*

We now discuss the four assumptions necessary for Theorem 10 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 13 *If (21) and (22) hold then there exist d, C, λ and b such that for any $\gamma \geq 1$*

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

Proof. [Jarner and Hansen (1998)] 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.* $I(x, dy) \triangleq \delta_x(dy)$). Then

$$\begin{aligned} P_\gamma V(x) &= V(x) + \int_{\mathcal{X}} q(x, dy) \min \left\{ 1, \left(\frac{\pi(y)}{\pi(x)} \right)^\gamma \left\{ \hat{\pi}^{-1/2}(y) - \hat{\pi}^{-1/2}(x) \right\} \right\} \\ &\leq V(x) + \int_{\mathcal{X}} q(x, dy) \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \left\{ \hat{\pi}^{-1/2}(y) - \hat{\pi}^{-1/2}(x) \right\} \right\} = PV(x), \end{aligned}$$

and the inequality holds because whenever $\pi^\gamma(y) \leq \pi^\gamma(x)$ we also have $\hat{\pi}^{-1/2}(y) \geq \hat{\pi}^{-1/2}(x)$. ■

3.2.2 Minorization condition

Using (20), we obtain the minorization condition (Assumption 2).

Lemma 14 *For all $i \geq 0$ the compact subsets of \mathbb{R}^k are small for P_{γ_i} .*

Proof. See Lemma 4 in [Roberts and Rosenthal (1998)] and its proof. One obtains for all x

$$\begin{aligned} P_\gamma(dy|x) &\geq \varepsilon \eta^\gamma(x, \delta/2) dy, \\ \eta(x, \delta/2) &\triangleq \frac{\inf_{w \in B(x, \delta/2)} \pi(w)}{\sup_{w \in B(x, \delta/2)} \pi(w)}, \end{aligned} \quad (23)$$

where $B(x, L)$ is the ball of radius L centered at x , and ε, δ are as in (20). Then, by convolving the transition kernels, one sees that every compact set is a small set (See Proposition 5.5.5 (ii) and Theorem 5.5.7 in [Meyn and Tweedie (1993)]). ■

3.2.3 Regularity of maxima and finite entropy conditions

The regularity Assumption 3 is assumed and Assumption 4 (finite entropy) follows by (21).

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 $X_2^{(0)} \sim v$, with v some probability distribution, and $X_2^{(i+1)} = F_i(X_2^{(i)})$, where

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

e_j denotes the j^{th} coordinate vector and q_j denotes a symmetric proposal distribution that updates x in direction e_j . We shall assume that $(q_j)_{j=1, \dots, k}$ 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 $j = 1, \dots, k$

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

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” [Tierney (1994)].

For each $\gamma > 1$, let $P_{\gamma_i, j}(dy|x) \triangleq \mathbb{P}(F_i(x) \in dy | j, x)$ be the kernel

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

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

$$P_{\gamma_i, RS}(dy|x) \triangleq \mathbb{P}(F_i(x) \in dy) = \sum_{j=1}^k \frac{1}{k} P_{\gamma_i, j}(dy|x).$$

Conditions for the geometric convergence of the homogeneous Markov chain when $\gamma_i = 1$ for all i have been proposed in [Roberts and Rosenthal (1998)], but further simplified in [Fort, Moulines and Roberts (2001)]. We state here two assumptions. Assumption A3 is the weakest, and probably the most difficult to check in practice, but is implied by A4 which is generally easier to check.

- (A3 [Fort, Moulines and Roberts (2001)]) For any sequence $x^{(j)} \in \mathcal{X}$ such that $\lim_{j \rightarrow +\infty} |x^{(j)}| = +\infty$, one may extract a subsequence $\tilde{x}^{(j)}$ such that for some $i \in \{1, \dots, k\}$ and all $y \in \mathbb{R}^+$,

$$\lim_{j \rightarrow +\infty} \frac{p(\tilde{x}^{(j)})}{p(\tilde{x}^{(j)} \pm y e_i)} = 0 \text{ and } \lim_{j \rightarrow +\infty} \frac{p(\tilde{x}^{(j)} \pm y e_i)}{p(\tilde{x}^{(j)})} = 0. \quad (24)$$

- (A4 [Fort, Moulines and Roberts (2001)]) The target distribution π is continuously differentiable for large $|x|$ and for all $i \in \{1, \dots, k\}$

$$\lim_{x_i \rightarrow +\infty} \sup_{\{x_{-i} \in [-x_i, x_i]^{k-1}\}} \nabla_i \log \pi(x) = -\infty \text{ and } \lim_{x_i \rightarrow +\infty} \inf_{\{x_{-i} \in [-x_i, x_i]^{k-1}\}} \nabla_i \log \pi(x) = +\infty. \quad (25)$$

where x_i is the i th 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 15 Define $(X_2^{(i)})_{i \in \mathbb{N}}$ as above, and suppose that (24) or (25), $\mathbb{E} [V (X_2^{(0)})] < +\infty$ (with $V(x) = \hat{\pi}^{-c}(x)$ and some $c \in (0, 1)$) hold and that the series $\gamma_i = \frac{\log(i+q-k_0+1)}{-k_0 \log(\alpha)(1+\zeta)}$ (where α is obtained from (28) to correspond to the definition in Proposition 7). The quantities ζ , q are also defined in Proposition 7 and k_0 depends on the choice of the small set C defined in Lemma 17. Then Theorem 10 holds for the chain $(X_2^{(i)})_{i \in \mathbb{N}}$.

The assumptions made can probably be simplified in the spirit of [Jarner and Hansen (1998)] with adaptations from [Roberts and Rosenthal (1998)], but we shall not investigate this here. We now discuss the four assumptions for Theorem 10 to apply.

3.3.1 Drift condition

The required drift condition is obtained through the following lemma:

Lemma 16 If (24)-(25) hold, then for any compact set C and any temperature $\gamma \geq 1$

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

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

Proof. In Theorem 1 of [Fort, Moulines and Roberts (2001)] 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 $V_0(x) = \pi^{-c}(x)$ then

$$\limsup_{|x| \rightarrow +\infty} \frac{P_{1,RS}V_0(x)}{V_0(x)} < 1. \quad (26)$$

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

$$\limsup_{|x| \rightarrow +\infty} \frac{P_{\gamma,RS}V(x)}{V(x)} < 1 \quad (27)$$

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

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

but then, with an argument similar to that used in Lemma 13,

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

for $\gamma > 1$ necessarily

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

which contradicts (26). Therefore (27) holds uniformly for γ . ■

3.3.2 Minorization condition

To prove the minorization condition, we will need the following lemma:

Lemma 17 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 [Roberts and Rosenthal (1998)], see also Lemma 14 in the present paper. Calling $s_{\gamma_i}(x, y)$ the continuous component of the transition kernel $P_{\gamma_i,RS}$ one has for suitable constants a_1, a_2 and δ

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

Then by composition for any N and $x \in [-N\delta/2, N\delta/2]$

$$s_{\gamma_i} \dots s_{\gamma_{i+kN-1}}(x, y) \geq \frac{1}{k^{kN}} a_1^{kN} \prod_{j=i}^{i+kN-1} a_2^{\gamma_j} \text{ whenever } |w_i| < N\delta/2 \text{ for } i = 1, \dots, k \quad (28)$$

which shows that $[-N\delta/2, N\delta/2]$ is a small set for $P_{\gamma_i, RS}^{kN}$. The end of the proof is the same as for Lemma 4 in [Roberts and Rosenthal (1998)]. ■

3.3.3 Regularity of maxima and finite entropy conditions

The regularity 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 [Jarner and Hansen (1998)] and [Fort, Moulines and Roberts (2001)] 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 Eq. (1.33) in [Locatelli (2000)] for example). However the examples given here are interesting as they demonstrate that simulated annealing is sound for global optimization on a non compact set, as long as some tail properties are satisfied.

Example 1 *The conditions of Theorem 10 are satisfied for a Random walk Metropolis annealing algorithm in order to maximize the density (defined on \mathbb{R}^k)*

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

Example 2 *The conditions of Theorem 10 are **not** satisfied for a Random walk Metropolis annealing algorithm in order to maximize the density*

$$\pi(x_1, x_2) \propto \exp(-.01(x_1^2 + x_1^2 x_2^2 + x_2^2)).$$

See Fig. 1 for a contour plot of π .

Example 3 *However, the conditions of Theorem 10 are satisfied for the **Random Scan** Random walk Metropolis annealing algorithm in order to optimize*

$$\pi(x_1, x_2) \propto \exp(-.01(x_1^2 + x_1^2 x_2^2 + x_2^2)),$$

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

4 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.

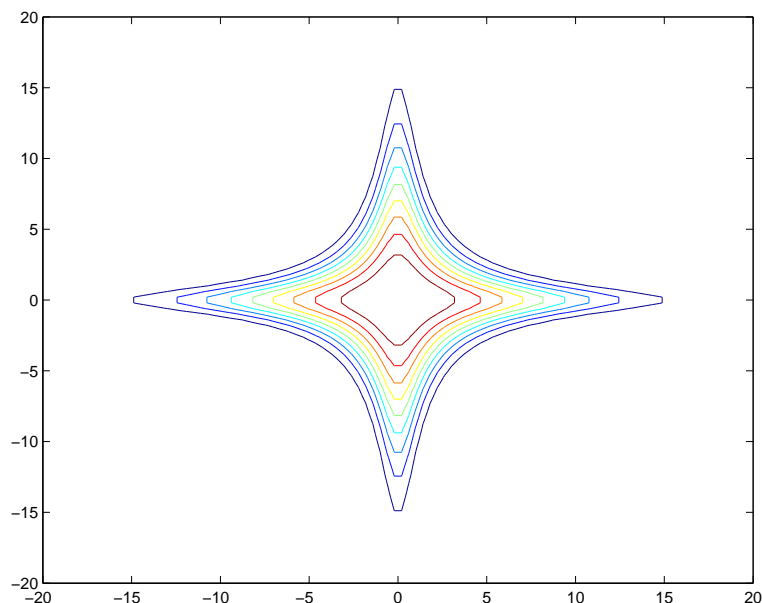


Figure 1: Contour plot of $\exp(-.01(x_1^2 + x_1^2x_2^2 + x_2^2))$.

References

- [Andrieu and Doucet (2000)] Andrieu, C. and Doucet, A. (2000), Simulated annealing for maximum a posteriori parameter estimation of hidden Markov models, *IEEE Trans. Information Theory*, vol. 46, no. 3, 994-1004.
- [Billingsley, (1985)] Billingsley, P. (1985), *Probability and Measure*, 2nd ed, Wiley.
- [Fort and Moulines (2000)] Fort, G. and Moulines, É. (2000), V-subgeometric ergodicity for a Hastings-Metropolis algorithm, *Statistics and Probability Letters*, **49**(4), 401-410.
- [Fort, Moulines and Roberts (2001)] Fort G., Moulines É. and Roberts, G.O. (2001), A Geometrically Ergodic Hybrid Sampler for Sub-Exponential Densities, *submitted*.
- [Gelfand and Mitter (1993)] Gelfand, S.B. and Mitter, S.K. (1993), Metropolis-type annealing algorithms for global optimization in \mathbb{R}^d , *SIAM J. Control Optimization*, **31**(1), 111-131.
- [Geman and Geman (1984)] Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, *IEEE Trans. Patt. Ana. Mac. Int.*, **6**, 721-741.
- [Haario and Sacksman (1991)] Haario, A and Sacksman, E. (1991), Simulated annealing in general state space, *Adv. Appl. Prob.*, **23**, 866-893.
- [Hwang (1980)] Hwang, C., (1980) Laplace's Method Revisited: weak convergence of probability measures, *Ann. Prob.*, **8**(6), 1177-1182.
- [Jarner and Hansen (1998)] Jarner, S. F. and Hansen, E. (1998), Geometric ergodicity of Metropolis algorithms, *Stoch. Proc. Appl.*, **85**, 341-161.
- [Jarner and Roberts (2001)] Jarner, S. F. and Roberts, G. (2001), Polynomial convergence rates of Markov chains, *to appear in Ann. Appl. Prob.*
- [Lindvall (1992)] Lindvall, T. (1992), *Lectures on the Coupling Method*, Wiley & Sons, New York.

- [Locatelli (2000)] Locatelli, M. (2000), Simulated annealing algorithms for continuous global optimization: convergence conditions, *Journal of Optimization Theory and Applications*, **104**(1),121-133.
- [Mengersen (1994)] Mengersen, K.L. and Tweedie, R.L. (1994), Rates of convergence of the Hastings and Metropolis algorithm, *Annals of Statistics*, **24**, 101-121.
- [Meyn and Tweedie (1993)] Meyn, S.P. and Tweedie, R.L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag.
- [Mitra *et al.* (1986)] Mitra, D., Romeo, F., and Sangiovanni-Vincentelli, A. (1986), Convergence and finite-time behavior of simulated annealing, *Adv. App. Prob.*, **18**, 747-771.
- [Roberts and Tweedie (1996)] Roberts, G.O. and Tweedie, R.L. (1996), Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms, *Biometrika*, **83**, 96-110.
- [Roberts and Tweedie (1999)] Roberts, G.O. and Tweedie, R.L. (1999), Bounds on regeneration times and convergence rates for Markov chains, *Stoch. Proc. Applic.*, **80**, 211-229.
- [Roberts and Rosenthal (1998)] Roberts, G.O. and Rosenthal, J.S. (1998), Two convergence properties of hybrid samplers, *Ann. Appl. Prob.*, **8**, 397-407.
- [Roberts and Tweedie (1996)] Roberts, G.O. and Tweedie, R.L. (1996), Exponential convergence of Langevin diffusions and their discrete approximations, *Bernoulli*, **2**, 341-364.
- [Rosenthal (1995)] Rosenthal, J.S. (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association*, **90**, 558-566.
- [Stramer and Tweedie (1999a)] Tweedie, R. and Stramer, O. (1999), Langevin-type models I: diffusions with given stationary distributions, and their discretizations, *Methodology and Computing in Applied Probability*, **1**, 283-306.
- [Stramer and Tweedie (1999b)] Tweedie, R. and Stramer, O. (1999), Langevin-type models II: self-targeting candidates for MCMC algorithms, *Methodology and Computing in Applied Probability*, **1**, 307-328.
- [Tierney (1994)] L. Tierney (1994), Markov chains for exploring posterior distributions, (with discussion) *Annals of Statistics*, 1701-1762.
- [Van Laarhoven and Arts (1987)] Van Laarhoven, P.J. and Arts, E.H.L. (1987), *Simulated Annealing: Theory and Applications*, Reidel Pub., Amsterdam.