ADAPTIVE CHAINS

LARS HOLDEN
NORWEGIAN COMPUTING CENTER

ABSTRACT Adaptive chains are chains that are able to learn from all previous elements in the chain. It is an extension of Markov chains. It is proved convergence of adaptive chains that satisfies a strong Doeblin condition (i.e., the transition density r from x_i to y satisfies $r(y, x_1, x_2, \dots, x_i) \geq a_i \pi(y)$ for all $x_1, \dots, x_i, y$ in the state space.) By using the previous iterations of the adaptive chain, it is possible to increase a_i which will improve convergence compared with Markov chains. It is also proved a decreasion rate in the covariance between element x_i and x_{i+j} as j increases.

The results may also be applied on regeneration chains where only the history before the last regeneration is used. Particularly interesting is the adaptive Metropolis-Hastings algorithm. Adaptive simulated annealing is also described and convergence is proved when the temperature decreases proportional with $M/\log i$. The convergence is due to contraction properties of integral operators with a stationary distribution and that satisfies a strong Doeblin condition. The algorithm is particularly useful when it is necessary with many samples from the same distribution like in Bayesian estimation, and in applications where it is very expensive to calculate the limiting density like inverse problems and optimisation.

1991 MATHEMATICS SUBJECT CLASSIFICATION 65C05, 65U05

KEY WORDS Adaptive chains, adaptive Metropolis-Hastings, regeneration, Markov chain, simulated annealing, inverse problems and optimisation.

1. INTRODUCTION

Markov chains are widely used as models and computational devices in areas ranging from statistics to physics. The theory and applications of Markov chains are very active fields of research: see, for example, Meyn and Tweedie (1993), Gilks et al. (1996) and Geyer (1992). In many Markov chains, one gains knowledge about the limiting distribution as the number of iterations increases. This knowledge may be used to adapt the transition densities in order to improve the convergence of the chain. Since the chain is adaptive, it is not a Markov chain. Part of Markov chain theory may be adapted to adaptive chains. There has been a large number of papers on adaptive chains the last years. Some paper focus on regeneration; Roberts and Tweedie (1998) and Gilks et al. (1998). At the regeneration time the present state is independent of the previous states. States before the last regeneration may then be used in the transition density. The chain is a Markov chain between each regeneration time. In Haario et al A (1998) and

Research supported by Research Council of Norway.

Haario et al B (1998) the full history is used in a Metropolis algorithm. It is made some assumptions on using Gaussian distributions. See also Gilks et al. (1994) and Roberts and Gilks (1994) for more specialised adaptive chains.

This paper proves convergence of regeneration chains using part of the history and adaptive chains using the full history. The only assumptions on the transition function are that it satisfies a strong Doeblin condition and a stability criterion. The results may also be applied on simulated annealing and simulated tempering. The improvements of the presented technique compared with standard Markov chains, are most significant if many samples from the same distribution are drawn. This is the case in most Bayesian applications. By using all previous samples of the chain, one may use methods from optimisation theory e.g. gradients in order to identify the most likely part of the sate space. Methods used in optimisation theory usually identify the likely areas much faster than traditionally MCMC techniques. It is equal important to identity these areas independent on whether the objective is to find the most likely state in the state space or the objective is to sample from a distribution.

Convergence is proved in both relative supremum norm $\|p\|_{\pi,\infty} = \sup_x\{|p(x) - \pi(x)|/\pi(x)\}$ and the total variation norm as $\|p - \pi\|_{T.V.} = \sup_{C\subset\Omega} \left|\int_C (p(x) - \pi(x))\, d\mu(x)\right|$. These norms are used in this paper since the total variation norm is the most used norm and the relative supremum norm is needed in order to bound a function of the correlation between different states in the chain. The relative supremum norm is stronger than most other norms.

2. A motivating example

Assume we want to sample from a function $\pi(x) = cf(x)$ for $x \in [0,1]$ where c is an unknown constant with the adaptive and regenerate Metropolis-Hastings algorithm. Let $\tilde{y}^j = (y_1, ..., y_j)$ denote the first j values of y.

1. Generate an initial state $x_1 \in \Omega$ from the density p_1.
2. For $i = 1, \ldots, n$:
   (a) Generate a state y_{i+1} from the density $q(y_{i+1}, x_i, \tilde{y}^{t(i)})$.
   (b) Calculate $\alpha_i(y_{i+1}, x_i, \tilde{y}^{t(i)}) = \min\left\{1, \frac{\pi(y_{i+1})q(x_i, y_{i+1}, \tilde{y}^{t(i)})}{\pi(x_i)q(y_{i+1}, x_i, \tilde{y}^{t(i)})}\right\}$.
   (c) Set $x_{i+1} = \begin{cases} y_{i+1} & \text{with probability } \alpha_i(y_{i+1}, x_i, \tilde{y}^{t(i)}) \\ x_i & \text{with probability } 1 - \alpha_i(y_{i+1}, x_i, \tilde{y}^{t(i)}). \end{cases}$

In an adaptive chain $t(i+1) \equiv i$. In a regenerate chain $t(1) = 0$ and $t(i+1) = i$ with probability

$$H(y_{i+1}, x_i, \tilde{y}^{t(i)}) = \begin{cases} max\{\kappa/w(x_i, \tilde{y}^{t(i)}), \kappa/w(y_{i+1}, \tilde{y}^{t(i)})\} & \text{if } w(x_i, \tilde{y}^{t(i)}) > \kappa, \text{ and } w(y_{i+1}, \tilde{y}^{t(i)}) > \kappa \\ max\{w(x_i, \tilde{y}^{t(i)})/\kappa, w(y_{i+1}, \tilde{y}^{t(i)})/\kappa\} & \text{if } w(x_i, \tilde{y}^{t(i)}) < \kappa, \text{ and } w(y_{i+1}, \tilde{y}^{t(i)}) < \kappa \\ 1 & \text{otherwise} \end{cases}$$

and otherwise $t(i+1) = t(i)$. In the expression $\kappa > 0$ is a constant and $w(z, \tilde{y}^{t(i)}) = \pi(z)/q(z, \tilde{y}^{t(i)})$. The algorithm is most efficient if κ is a central value of $w(z, \tilde{y}^{t(i)})$. This formula for the regenerate chain is only valid for

proposal function independent of the present position i.e. x_i. See Mykland et al (1995).

We will use four different proposal functions:

1. Markov chain: $q_1(y) \equiv 1$
2. Markov chain: $q_2(y, x_i) \sim \text{Uniform}(\max\{0, x_i - .05\}, \min\{1, x_i + .05\})$
3. Adaptive chain: $q_3(y, \tilde{y}^{i-1}) = \theta_i \max\{f(y_{k(y)}), 0.1\}$ where θ_i is a normalising constant and $k(y) = \text{argmin}_{1 \le j \le i-1}\{|y - y_j|\}$. We assume that $\pi(x_1) > 0$.
4. Regenerate chain: $q_4(y, \tilde{y}^{t(i)}) = \theta_i \max\{f(y_{k(y)}), 0.1\}$ where θ_i is a normalising constant and $k(y) = \text{argmin}_{1 \le j \le t(i)}\{|y - y_j|\}$. We assume that $\pi(x_1) > 0$.

The proposal function q_2 has some similarities with random walk while q_3 and q_4 are step functions that in each point uses the nearest evaluation of f. The proposal functions q_3 and q_4 do not give Markov chains since they use previous states/evaluations $y_1, ..., y_{t(i)}$. In a later section it is proved that the adaptive chain generated by using q_3 converges to a chain ϕ_i which again converges to π. ϕ_i is defined by the same algorithm as above but with p_1 replaced by π. The regenerate chain using q_4 converges directly to π.

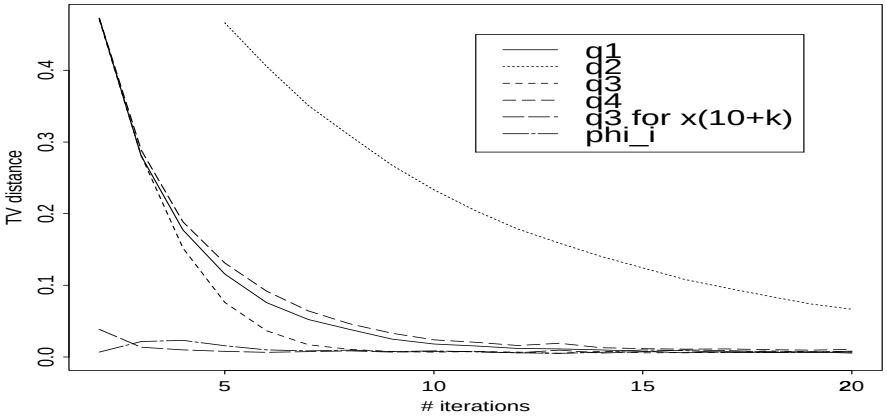

FIGURE 1. The total variance distance between the target and the actual distribution at each of the first 20 iterations. The convergence rate is in the following order: q_3 (adaptive, conditioned on $x_{10} = .5$ and using the first 10 evaluations of f), q_3 (adaptive), $q_1 \equiv 1$, q_4 (regenerate) and the poorest one q_2 (Random walk). The curve starting in origo which starts converging after 4 iterations, is the ϕ_i curve that the adaptive chain approaches. The results are based on statistics from 100.000 simulated chains.

These methods are tested on the function $f(x) = x^3$ with $x_1 = 0.5$. Simulations have been performed for each of the three methods. The total variance distance between the target distribution and the actual distribution is shown in figure 1, while the estimated auto correlations for lags up to 20 are shown in figure 2. The total variance distance shows how fast the chain

converges to the limiting distribution. The auto correlation shows how fast the dependency between x_n and x_{n+k} vanishes. In figure 1 it is also shown the distribution for x_{10+k} assuming that element $x_{10} = 0.5$ and using the adaptive q_3. For q_1 and q_2 the distributions for x_{10+k} and x_k are identical assuming $x_{10} = .5$ and $x_1 = .5$ respectively. The proposal functions q_3 and q_4 learn from more samples. In all cases the convergence depends critically on how close q is π. The random walk proposal function, q_2, is shown to be largely inferior of the other methods. The probability for regeneration is so small in the start of the chain that the chain using q_4 is not better than the chain using q_1. The adaptive q_3 is better and considerably better if it may use some previous evaluations of f. The regenerate q_4 has the same decrease in auto correlation as the plotted adaptive q_3 but needs a longer burn-in period.

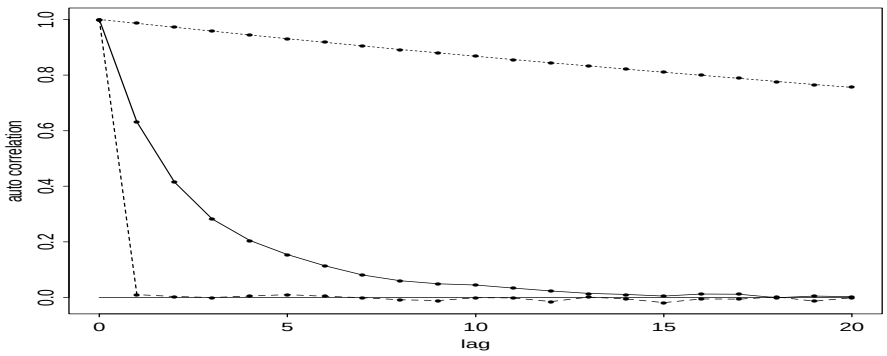

FIGURE 2. The auto correlation of three chains having different proposal functions. The dotted line correspond to q_1, while the solid line correspond to q_2 and the dashed line to q_3. The regenerate q_4 chain gave almost identical curve as the adaptive q_3. The results are based on a long chain of length 10000, with a burn-in of 100 iterations.

The two last methods are not Markov chains since they depend on all previous elements in the chain. In other examples the history may be represented as estimates of some parameters β. Then the chain could in some cases be represented as a Markov chain in the state space (x, β). In the example above this is not natural since all previous states are used. Hence, it is not natural to consider it as a Markov chain. This proposal function is chosen since it is easy to generalise to several dimensions. If the number of samples is large, it is probably necessary to estimate some parameters in the proposal function.

The convergence depends critically on how close q is π. If the transition density of a Markov chain satisfies $r(y|x) \geq a\pi(y)$ for all x, y in the state space, then it converges geometrically with rate $(1-a)^i$. For adaptive and regenerate chains the product $\prod_i (1-a_i)$ where a_i satisfies $r_i(y, x_i, \tilde{y}^{t(i)}) \geq a_i\pi(y)$, is important for the convergence and the decreasion of the covariance between element x_j and x_k. This is proved in the following sections.

3. Definition of adaptive and regenerate chains

Assume that we want to sample from a function $\pi(x) = cf(x)$ for $x \in \Omega$ where c is an unknown constant with an iterative method. An adaptive chain is defined as follows.

Adaptive chain.

1. Generate an initial state $x_1 \in \Omega$ from the density p_1
2. For $i = 1, \ldots, n$:
   Generate a new state $x_{i+1} = y$ from the density $r_i(y, x_i, \tilde{x}^{i-1})$.

Note that r_i has variable number of arguments. $r_i(y, x_i, \tilde{x}^{i-1})$ is interpreted as a transition density from x_i to y that depends on the history $\tilde{x}^i = (x_1, \ldots, x_i)$. The last element in the chain x_i may also be used in the history.

We will also define the regenerate chain. Let the function h_i and the density ν_i satisfy

$$r_i(y, x, \tilde{x}^j) \geq h_i(x, \tilde{x}^j)\nu_i(y) \qquad \text{for all } (y, x, \tilde{x}^j) \in \Omega^{j+2} \ . \tag{1}$$

Regenerate chain.

1. Generate an initial state $x_1 \in \Omega$ from the density p_1
2. Set $t(1) = 0$
3. For $i = 1, \ldots, n$:
   (a) Generate a new state $x_{i+1} = y$ from the density $r_i(y, x_i, \tilde{x}^{t(i)})$.
   (b) Set $t(i+1) = i$ with probability $H(y, x, \tilde{x}^{t(i)}) = h_i(x_i, \tilde{x}^{t(i)})\nu_i(y)/r_i(y, x_i, \tilde{x}^{t(i)})$ and else $t(i+1) = t(i)$.

The regenerate chain only uses the first $t(i)$ elements in the chain in the transition density, not the last elements $x_{t(i)+1}, \cdots, x_i$. The iterations where $t(i+1) = i$ is called a regeneration. The regeneration chain is a Markov chain between each regeneration. In each regeneration we may consider x_{i+1} as drawn from ν_i instead of r_i hence it becomes independent of $\tilde{x}^{t(i)}$. It may be complicated to calculate H. See the example in section 2 and the references Mykland et al (1995) and Gilks et al. (1998).

Convergence both for the adaptive chain and the regenerate chain is proved in this paper. The author believe that normally will the adaptive chain approach π faster than both standard Markov chains and regenerate chains since it may use all the previous elements in the chain. The adaptive chain is also easier to use than regenerate chains since it is not necessary to consider the regeneration.

In order to prove convergence of the adaptive chain, it is necessary to introduce another chain ϕ_i which is a slight perturbation of π. ϕ_i is defined by the same transition density as p_i but using $\phi_1 = \pi$ instead of p_1. The convergence of p_i towards ϕ_i may be proved similarly as convergence of

Markov chains satisfying the strong Doeblin condition. In order to prove that ϕ_i approaches π, it is necessary with some new techniques.

If the objective is to draw n samples, then sample no. $m_1, m_2, \dots, m_n$ may be drawn. Let $m_i - m_{i-1}$ decrease when i increases since the increased history will improve convergence and decrease covariance.

4. Main assumptions and results

Let $\Omega^1 = \Omega \subset \mathbb{R}^n$ be a Borel measurable state space or, alternatively, let Ω be a discrete state space, and $\Omega^{i+1} = \Omega \times \Omega^i$, and $\tilde{x}^i = (x_1, ..., x_i) \in \Omega^i$. Let $\mu(\tilde{x}^i)$ be the product measure on Ω^i. Let further the limiting density π and the transition density r be integrable with respect to μ including point mass distributions. The results are also valid for more general state spaces Ω.

In some algorithms like simulated annealing and simulated tempering there is defined a sequence of densities $\{\pi_i\}_i$ where $\pi_i \to \pi$ when $i \to \infty$. In standard Markov chain applications we may set $\pi_i \equiv \pi$ for all $i > 0$. This sequence is included in the theorems since it makes the theorems more general and the proofs do not change. Define

$$c_i = \|\pi_{i+1} - \pi_i\|_{T.V}. \tag{2}$$

The transition density in iteration i is

$$s_i(y, x) = \int_{\Omega^{t(i)}} r_i(y, x, \tilde{x}^{t(i)}) P(\tilde{x}^{t(i)}, x) d\mu(\tilde{x}^{t(i)}) \tag{3}$$

where $P(\tilde{x}^{t(i)}, x)$ is the conditional distribution for $\tilde{x}^{t(i)}$ given $x_i = x$ and the initial distribution is p_1. If another initial distribution is used, P and s_i are still well defined but P has not the same interpretation. Note that when applied on adaptive chains $t(i+1) = i$. If Metropolis-Hastings is used and it is prefered to use the proposed history $\tilde{y}^j$ instead for the history of the chain $\tilde{x}^j$, then $P(\tilde{x}^{t(i)}, x)$ should be replaced by $Q(\tilde{y}^{t(i)}, x)$, the conditional distribution for $\tilde{y}^{t(i)}$ given x. This only leads to changes in the definitions. The theorems and proofs are identical.

The density after i iterations, p_i, is defined by

$$p_{i+1}(y) = \int_\Omega s_i(y, x) p_i(x) d\mu(x) \tag{4}$$

where p_1 is defined by the user. In each iteration p_i approaches the density ϕ_i defined by

$$\phi_{i+1}(y) = \int_\Omega s_i(y, x) \phi_i(x) d\mu(x) \tag{5}$$

and $\phi_1 \equiv \pi_1$. Note that s_i and ϕ_i depend on p_1 but not on the chain x_i.

It is possible to construct the transition function r_i such that it satisfies

$$\pi_i(y) = \int_\Omega r_i(y, x, \tilde{x}^{t(i)}) \pi_i(x) d\mu(x) \qquad \text{for all } \tilde{x}^{t(i)} \in \Omega^{t(i)} \text{ and all integers } i > 0. \tag{6}$$

This assumption may be verified for the Metropolis-Hastings algorithm. The transition kernel for the Metropolis-Hastings algorithm is

$$r_i(y,x,\tilde{x}^{t(i)}) = q(y,x,\tilde{x}^{t(i)})\alpha_i(y,x,\tilde{x}^{t(i)}) + I(y=x)(1-\int q(z,x,\tilde{x}^{t(i)})\alpha_i(z,x,\tilde{x}^{t(i)})d\mu(z))$$

where

$$\alpha_i(y,x,\tilde{x}^{t(i)}) = \min(1, \frac{\pi_i(y)q(x,y,\tilde{x}^{t(i)})}{\pi_i(x)q(y,x,\tilde{x}^{t(i)})}). \tag{7}$$

Using the fact that

$$\pi_i(x)q(y,x,\tilde{x}^{t(i)})\alpha_i(y,x,\tilde{x}^{t(i)}) = \pi_i(y)q(x,y,\tilde{x}^{t(i)})\alpha_i(x,y,\tilde{x}^{t(i)})$$

which follows from (7), we obtain the detailed balance equation:

$$\pi_i(x)r_i(y,x,\tilde{x}^{t(i)}) = \pi_i(y)r_i(x,y,\tilde{x}^{t(i)})$$

Integrating both sides with respect to x gives (6).

Define the transition density $\hat{r}_i$ by

$$\hat{r}_i(y,x) = \int_{\Omega^{t(i)}} r_i(y,x,\tilde{x}^{t(i)})P(\tilde{x}^{t(i)})d\mu(\tilde{x}^{t(i)}) \tag{8}$$

where $P(\tilde{x}^{t(i)})$ is the distribution for $\tilde{x}^{t(i)}$ conditioned on p_1 as initial distribution. If π_i satisfies the equation (6) then it also satisfies

$$\pi_i(y) = \int_\Omega \hat{r}_i(y,x)\pi_i(x)d\mu(x) \qquad \text{for all integers } i>0 \tag{9}$$

since

$$\begin{aligned}\pi_i(y) &= \int_{\Omega^{t(i)}} \pi_i(y)P(\tilde{x}^{t(i)})d\mu(\tilde{x}^{t(i)}) \\ &= \int_{\Omega^{t(i)+1}} r_i(y,x,\tilde{x}^{t(i)})\pi_i(x)P(\tilde{x}^{t(i)})d\mu(x,\tilde{x}^{t(i)}) \\ &= \int_\Omega \hat{r}_i(y,x)\pi_i(x)d\mu(x).\end{aligned}$$

Notice that $s_1 = \hat{r}_1$.

If regeneration chain is used, then at the regeneration times when $t(i+1)=i$, x_{i+1} is independent of the history $\tilde{x}^{t(i)}$, hence $P(\tilde{x}^{t(i)},x) = P(\tilde{x}^{t(i)})$. Then $s_i \equiv \hat{r}_i$ for $i>0$ implying $\phi_i \equiv \pi_i$ for $i>0$.

For adaptive chains the difference $\phi_i - \pi_i$ depends on how close s approaches $\hat{r}$. Define b_i and b_i^S by

$$b_i = \| \int_\Omega (s_i(.,x) - \hat{r}_i(.,x))\pi_i(x)d\mu(x)\|_{\text{T.V.}} \tag{10}$$

and

$$b_i^S = \sup_{y\in\Omega,\pi_i(y)>0} | \int_\Omega (s_i(y,x) - \hat{r}_i(y,x))\pi_i(x)d\mu(x)|/\pi_i(y). \tag{11}$$

Note that

$$s_i(y,x) - \hat{r}_i(y,x) = \int_{\Omega^{t(i)}} r_i(y,x,\tilde{x}^{t(i)})(P(\tilde{x}^{t(i)},x) - P(\tilde{x}^{t(i)}))d\mu(\tilde{x}^{t(i)}).$$

One may argue that b_i and b_i^S are small based on $r_i(y, x, \tilde{x}^{t(i)})$ is not too sensitive to the history e.g. due to that $r_i(y, x, \tilde{x}^{t(i)})$ approaches a function $\tilde{r}(y, x)$ when i increases. Another argument is that the most of the history is independent of the present state. Hence $P(\tilde{x}^{t(i)}, x) \approx P(\tilde{x}^{t(i)})$ in most of $\Omega^{\tilde{x}^{t(i)}}$.

The difference $\phi_i - p_i$ depends on the constant a_i in the strong Doeblin condition: Let $a_i \in [0, 1]$, satisfy

$$s_i(y, x) \geq a_i \phi_i(y) \qquad \text{for all } x, y \in \Omega \text{ and all } i > 0. \tag{12}$$

and

$$\hat{r}_i(y, x) \geq a_i \pi_i(y) \qquad \text{for all } x, y \in \Omega \text{ and all } i > 0. \tag{13}$$

The strong Doeblin condition may be described as the chain defined by s_i and $\hat{r}_i$ respectively have probability a_i to move to the distribution ϕ_i and π_i respectively in iteration i. See Doob (1953), p. 197, and Meyn and Tweedie (1993) p. 391 for references to the Doeblin condition and Holden (1997) for reference to the strong Doeblin condition. The strong Doeblin condition is discussed in separate section.

Then it is possible to formulate the convergence in total variation norm. The proof is inspired by the proof of Theorem 16.2.1 in (Meyn and Tweedie 1993). Note that the the $\phi_i - \pi_i$ difference is independent of the initial density p_1.

THEOREM 1 Assume the adaptive/regenerate chain satisfies (9) and use the definitions in (2), (3), (4), (5), (8), (10), (12) and (13). For adaptive chains $t(i + 1) = i$ and for regenerate chains $\phi_i \equiv \pi_i$. Then the following is satisfied

$$\|p_{i+1} - \phi_{i+1}\|_{T.V.} \leq \prod_{j=1}^{i} (1 - a_j). \tag{14}$$

$$\|\phi_{i+1} - \pi_{i+1}\|_{T.V.} \leq \sum_{j=1}^{i} (c_j + b_j) \prod_{k=j+1}^{i} (1 - a_k). \tag{15}$$

and

$$\|p_{i+1} - \pi\|_{T.V.} \leq \|\pi_{i+1} - \pi\|_{T.V.} + \prod_{j=1}^{i} (1 - a_j) + \sum_{j=1}^{i} (c_j + b_j) \prod_{k=j+1}^{i} (1 - a_k) \tag{16}$$

If $a_j \geq a > 0$ for all $j > 0$ and $c_j + b_j \to 0$ when $j \to \infty$, then $\|\phi_{i+1} - \pi_{i+1}\|_{T.V.} \to 0$ when $i \to \infty$. If $a_j \geq a > 0$ and $c_j + b_j \leq d^j < 1$ for all $j > 0$, then there is the following geometric convergence $\|\phi_{i+1} - \pi_{i+1}\|_{T.V.} \leq \frac{d(1-a)^i}{1-a-d}$.

It is possible to give a similar theorem for convergence in the relative supremum norm. Note that $\|p_1\|_{\pi,\infty}$ is not well-defined if π or p_1 has point mass distribution. The convergence in this norm may be generalised to the case where there is point mass in p_1. This is necessary in order to apply the theorem on the Metropolis–Hastings algorithm. Let $A_i \subset \Omega$ be the

points where there is point mass distribution in p_i. A_i is countable since $\int p_i(x)d\mu(x) = 1$. Define

$$S_i = \int_{A_i} p_i(x)d\mu(x),$$

the pointwise relative error

$$R^i(x) = (p_i(x) - \phi_i(x))/\phi_i(x) = p_i(x)/\phi_i(x) - 1,$$

the maximum relative error

$$R^i_M = \sup_{x\in\Omega\setminus A_i} \left|R^i(x)\right|,$$

and

$$D_i = \sup_{y\in\Omega\setminus A_{i+1}, x\in A_i} \{(s_i(y,x)/\phi_i(y) - a_i)/(1-a_i)\}.$$

If there is not point mass in p_i and ϕ, then $\|p_i\|_{\phi_i,\infty} = R^i_M$.

Convergence in the relative supremum norm is proved using techniques from Holden (1997).

THEOREM 2 Assume the adaptive/regenerate chain satisfies (9) and use the definitions in (2), (3), (4), (5), (8), (11), (12) and (13). For adaptive chains $t(i+1) = i$ and for regenerate chains $\phi_i \equiv \pi_i$. If $\pi(x) > 0$ for $x \in \Omega$, and there is neither point mass distribution in p_i, ϕ_i, π_i, s_i nor $\hat{r}_i$ then

$$\|p_{i+1}\|_{\phi_{i+1},\infty} \le \|p_1\|_{\pi_1,\infty} \prod_{j=1}^{i}(1-a_j). \tag{17}$$

(18)

$$\|\phi_{i+1}\|_{\pi_i,\infty} \le \sum_{j=2}^{i}(\|\pi_{i+1}\|_{\pi_i,\infty} + b^S_j)(\|\pi_j\|_{\pi_{j+1},\infty} + 1) \prod_{k=j+1}^{i}(1-a_k)(\|\pi_k\|_{\pi_{k+1},\infty} + 1).$$

and

(19)

$$\|p_i\|_{\pi,\infty} \le \|p_i\|_{\phi_i,\infty}(1 + \|\phi_i\|_{\pi_i,\infty})(1 + \|\pi_i\|_{\pi,\infty}) + \|\phi_i\|_{\pi_i,\infty}(1 + \|\pi_i\|_{\pi,\infty}) + \|\pi_i\|_{\pi,\infty}$$

If $a_j \ge a > 0$ for all $j > 0$ and $\|\pi_{i+1}\|_{\pi,\infty} + b^S_j \to 0$ when $j \to \infty$, then

$$\|\phi_{i+1}\|_{\pi_i,\infty} \to 0 \qquad \text{when } i \to \infty\ . \tag{20}$$

If $a_j \ge a > 0$ and $\|\pi_{i+1}\|_{\pi,\infty} + b^S_j \le d^j < 1$ for all $j > 0$, then there is the following geometric convergence

$$\|\phi_{i+1}\|_{\pi_i,\infty} \le \frac{d(1-a)^i}{1-a-d}. \tag{21}$$

If there is point mass distribution in p_i and $s_i(x_{i+1}, x_i)$ but only for $x_i \in A_i$ and $x_{i+1} \in A_{i+1}$, then (18) and (19) are still valid when $\|p_{i+1}\|_{.,\infty}$ is interpreted as the part without point mass. In addition we have

$$S_{i+1} \le S_1 \prod_{j=1}^{i}(1-a_j).$$

and

$$R_M^{i+1} \leq (R_M^1 + S_1 \sum_{j=1}^{i} D_j) \prod_{j=1}^{i} (1 - a_j).$$

Similar results are valid for most other norms. It is well known and easy to prove that if $H = L_q, L_\infty$ or the total variation norm, then

$$\|f\|_H \leq \|f\|_{\pi,\infty} \|\pi\|_H. \tag{22}$$

Hence, the bound on the relative supremum norm may be used to prove convergence in other norms, see (Holden 1997).

5. Estimation results

Adaptive and regenerate chains may be used for estimation similarly as Markov chains are used. The error in expectation and the correlation between different terms may be bounded by the norms used in Theorem 1 and 2. Define $p_{j|i}$ as the the distribution of x_j given x_i, and $\sigma_i^2 = \int (f(x) - f_i)^2 p_i(x) d\mu(x)$, $f_i = \int f(x) p_i(x) d\mu(x)$. σ_π and f_π are defined correspondingly with p_i replaced by π. In addition, define $\sigma_{A_{j|i}}^2 = \sup_{y \in A_{j|i}} (f(y) - f_{j|i})^2$.

Theorem 3
The expected value satisfies

$$|f_i - f_\pi| \leq f_\pi \|p_i\|_{\pi,\infty}.$$

The covariance between two terms of the chain satisfies

$$|\mathrm{Cov}(f(x_i), f(x_j))| \leq 2\sigma_i (\sigma_\pi^2 + (f_{j|i} - f_\pi)^2)^{1/2} \|p_{j|i} - \pi\|_{T.V.}^{1/2} \|p_{j|i}\|_{\pi,\infty}^{1/2}$$

If there is point mass distribution in $p_{j|i}$, then

$$|\mathrm{Cov}(f(x_i), f(x_j))| \leq 2\sigma_i \|p_{j|i} - \pi\|_{T.V.}^{1/2} ((\sigma_\pi^2 + (f_{j|i} - f_\pi)^2) R_M^{j|i} + \sigma_{A_{j|i}}^2 S_{j|i})^{1/2}$$

Combining Theorem 1 and 2 imply that the covariance for regenerate chains decreases with the rate $\prod_{k=i+1}^{j} (1 - a_k)$. It is easy to apply these bounds on estimators like $1/n \sum_i g(x_{m_i})$.

6. Further discussion on the Doeblin condition

The strong Doeblin condition assumption is crucial in this paper. If equation

$$r_i(y, x, \tilde{x}^{t(i)}) \geq a_i \phi_i(y) \qquad \text{for all } (y, x, \tilde{x}^{t(i)}) \in \Omega^{i+1} \text{ and all } i > 0. \tag{23}$$

is satisfied then the strong Doeblin condition (12) is satisfied since

$$\begin{aligned} s_i(y, x) &= \int_{\Omega^{t(i)}} r_i(y, x, \tilde{x}^{t(i)}) P(\tilde{x}^{t(i)}, x) d\mu(\tilde{x}^{t(i)}) \\ &\geq a_i \phi_i(y) \int_{\Omega^{t(i)}} P(\tilde{x}^{t(i)}, x) d\mu(\tilde{x}^{t(i)}) = a_i \phi_i(y) \end{aligned}$$

The calculation is similar for $\hat{r}_i$ using (13).

The adaptive Metropolis–Hastings algorithm, as described in the previous section, is one method to make a transition function that satisfies (23). The Metropolis–Hastings algorithm satisfies the strong Doeblin condition (12) with the same a_i if $q(y,x,\tilde{x}^{t(i)}) \geq a_i\pi(y)$. Let $(y,x,\tilde{x}^{t(i)}) \in \Omega^{t(i)+2}$, then

$$\begin{aligned} r_i(y,x,\tilde{x}^{t(i)}) &\geq \alpha_i(y,x,\tilde{x}^{t(i)})q(y,x,\tilde{x}^{t(i)}) \\ &= \min\left\{q(y,x,\tilde{x}^{t(i)}), \frac{\pi(y)}{\pi(x)}q(x,y,\tilde{x}^{t(i)})\right\} \geq a_i\pi(y). \end{aligned}$$

The strong Doeblin condition may seem restrictive. In chains like random walk where $|x_i - x_{i-1}|$ is small the strong Doeblin condition is not satisfied. For these kinds of chains it is possible to define a new chain which consists of every n'th element in the original chain i.e. $z_k = x_{nk}$. This new chain may satisfy the strong Doeblin condition. Formalise this by for every $x, y \in \Omega$ define a sequence $\{D_{i,x,y}\}_{i=0}^{n}$ where $D_{i,x,y} \subset \Omega$. Let $D_{0,x,y} = \{x\}$, $D_{n,x,y} = \{y\}$, and for every $u \in D_{i,x,y}$ and $v \in D_{i+1,x,y}$, $r_i(v,u,\tilde{x}^{t(i)}) \geq a_i\phi(v)$. Then the transition density $r^n(x_{n(k+1)}, x_{nk}, \tilde{x}^{t(n(k-1))})$ for the new chain satisfies assuming $a_i = a$ for all $i > 0$

$$\begin{aligned} &r^n(x_{n(k+1)}, x_{nk}, \tilde{x}^{t(n(k-1))}) = \tilde{r}^n(x_{n(k+1)}, x_{nk}, \tilde{x}^{t(nk)}) \\ &= \int_{\Omega^{n-1}} \prod_{j=0}^{n-1} r(x_{nk+j+1}, x_{nk+j}, \tilde{x}^{t(nk+j)}) d\mu(x_{nk+1}, \ldots, x_{n(k+1)-1}) \\ &\geq \int_{\prod_{j=1}^{n-1} D_{j,x,y}} \prod_{j=0}^{n-1} r(x_{nk+j+1}, x_{nk+j}, \tilde{x}^{t(nk+j)}) d\mu(x_{nk+1}, \ldots, x_{n(k+1)-1}) \\ &\geq a\phi_{n(k+1)}(x_{n(k+1)}) \int_{\prod_{j=1}^{n-1} D_{j,x,y}} \prod_{j=0}^{n-2} a\phi_{nk+j+1}(x_{nk+j+1}) d\mu(x_{nk+1}, \ldots, x_{n(k+1)-1}) \\ &= a^n \phi_{n(k+1)}(x_{n(k+1)}) \prod_{j=1}^{n-1} \int_{D_{j,x,y}} \phi_{nk+j}(x_{nk+j}) d\mu(x_{nk+j}) \end{aligned}$$

The transition density r^n for z_i should only depend on z_k and $\tilde{z}^{t(k)}$. The history $\tilde{x}^{n(k-1)}$ used by $\tilde{r}$ above may be found from r and the z history $\tilde{z}^{t(k)}$. Hence, this chain satisfies the strong Doeblin condition with

$$\tilde{a}_k^n = a^n \inf_{x,y\in\Omega} \{\prod_{i=1}^{n-1} [\int_{D_{i,x,y}} \phi_{nk+i}(x) d\mu(x)]\}. \tag{24}$$

If it is not possible to move between any two states in a finite number of steps, often the following weaker assumption is satisfied: Let $\{B_i\}$ be a sequence of sets with $\cup_i B_i = \Omega$ where

$$s_i(y,x) \geq a_i\phi_i(y) \qquad \text{for all } i \geq 0 \text{ and } x, y \in B_i. \tag{25}$$

where $a_i > 0$. Similar theorems as in this paper may be proved using the above sequence of Doeblin condition. This is used in (Holden 1997)

7. Practical advice

It is natural to let q approximate π with the knowledge of the previous iterations. Assume $\Omega \subset \mathbf{R}$. Let $q_5(y, \tilde{x}^{t(i)})$ denote a piecewise kth order polynomial approximation to π using $\pi(y_i)$, $i = 1, \cdots, t(i)$. Then

$|q_5(y,x,\tilde{x}^{t(i)}) - \pi(y)| \leq c_k/t^{k+1}(i)$ for bounded and sufficient smooth π. This implies that $1-a_i, b_i, b_i^S \leq c_k/t^{k+1}(i)$ which gives convergence according to Theorem 1 and 2. The convergence p_i towards ϕ_i for adaptive chains will be with the rate $c_k^i/(i!)^{k+1}$ which is considerably faster than standard MCMC. It is possible to generalise this to $\Omega \subset \mathbf{R}^n$ using Taylor expansion in this dimension.

There is a danger of adapting too fast to the knowledge based on few samples. In the proof of Theorem 1 it is proved that the relative error $R^i(y) = (p_i(y) - \phi_i(y))/\phi_i(y)$ satisfies

$$R^{i+1}(y) = \int_\Omega \frac{s_i(y,x)}{\phi_i(y)} R^i(x)\phi_i(x)d\mu(x).$$

Notice that

$$\int_\Omega R^i(x)\phi_i(x)d\mu(x) = 0,$$

hence the relative error decreases faster the better $s_i(y,x)$ approximates $\phi_i(y)$. This approximation may be used in evaluating how close the transition density should adapt the limiting density based on a small number of iterations. Some more robust proposal functions are discussed in the following paragraphs.

The simplest way to reduce the adaption on the proposal function is to truncate it i.e. $q_6(y,x,\tilde{x}^{t(i)}) = \max\{q_5(y,x,\tilde{x}^{t(i)}), \rho_i\}$. If $\rho_i = \rho > 0$ for all i, then convergence is ensured. But if $\pi(x) < \rho$ in part of Ω, then the convergence will not be faster than geometric since $a_i \leq a < 1$ for all i. If $\rho_i \to 0$, convergence may be faster than geometric. But, the easily proved convergence $\prod_i (1 - \rho_i/\sup_x\{\pi(x)\})$ is slower than geometric.

The above proposal function may not be robust when $\pi(x)$ has many modes. A more robust proposal function than q_3 is the following

$$q_7(y,x,\tilde{x}^{t(i)}) = \begin{cases} \min(\eta_i, \max\left\{q_5(y,x,\tilde{x}^{t(i)}), \tau_i\right\}) & \text{if } |y - x_{k(y)}| < \psi_i \\ 1 & \text{otherwise} \end{cases}$$

where ψ_i depends on the expected smoothness of f. η_i should be chosen close to 1 until all modes where identified. This reduces the probability of resampling in modes that already are identified. τ_i should be close to 0 and decrease slowly in order to reduce the probability for sampling in areas where it is known to be no modes.

One may or may not use the last state x in the interpolator q_5. If it is used it is more difficult to calculate the ratio α due to an integrational constant. Preliminary numerical tests are not conclusive whether it is best to use x or not in q_5. The difference is illustrated below. Let z be a local optima for $\pi(y)$ which is not reflected in the initial proposal function. In a Markov chain the density for proposing z is small, but if it is proposed then the probability for accepting the proposal is large and the probably for leaving the state small. An adaptive chain behaves like the Markov chain until the local optima is identified. After it is identified, will the density for proposing z be large, and when it is proposed both the probabilities for accepting and leaving the proposal are large. If the present state is not included in the interpolator, will the probability for leaving the local optima for the first time be small.

The period where it remains in the optima partly compensate for the period before the optima was recognised and therefore sampled with too small probability. A regenerate chain behaves as a Markov chain until the first regeneration after the local optima were identified. After the regeneration it behaves as an adaptive chain. The probability for a regeneration is small until there has been a regeneration after all local optima are identified.

In some applications it is natural to vary between different transition functions $r_i = \sum_j \gamma_{i,j}\tilde{r}_j$ where $\gamma_{i,\tau(i)} = 1$ for an integer function τ and $\gamma_{i,j} = 0$ for $j \neq \tau(i)$. The choice parameter τ may either be systematic or stochastic. The chain converges as long as it satisfies (9) and $1 - a_i$ and b_i are sufficient small. It is also sufficient that the sub chain $x_{m_1}, x_{m_2}, \ldots$ satisfies the strong Doeblin condition using (24). If τ is stochastic and it is possible to move to the same position for different transition functions it may be complicated to calculate the acceptance ratio in the Metropolis Hastings algorithm. If there are common bounds on a_i and b_i for all functions τ, we may assume τ is fixed. The convergence is then satisfied for all τ.

Adaptive chains may also be used on inverse problems i.e. find $x_0 \in \Omega$ where $g(x_0) = 0$ and in optimalization i.e. find $x_0 \in \Omega$ where $g(x_0) \geq g(x)$ for all $x \in \Omega$ where there is some noise in the evaluations of g. Bayesian statisticians often consider the measurement error and include prior knowledge. Hence they prefer to study formulas like $\pi(x) = f(x)\exp(-\beta g^2(x))$ or $\pi(x) = f(x)\exp(\beta g(x))$ where β is a constant and $f(x)$ represents prior knowledge of x_0. Traditional numerical methods for the inverse and optimalization problem neglect f and use gradients of g calculated from the previous evaluations of g. These methods are usually faster on the pure inverse/optimisation problem ($''g''$ problem) than traditional MCMC methods are on the Bayesian problem ($''\pi''$ problem). Using adaptive chains it is possible to use some of the same numerical techniques, but in a Bayesian framework and hopefully keep some of the efficiency of the numerical methods.

8. An abstract mathematical formulation

The adaptive chain is described in an abstract mathematical formulation in this section. Let K be the set of measurable densities on Ω. Let further V be the set of integral operators, $T : K \to K$ with π as invariant distribution i.e. $T(\pi) = \pi$. It is possible to construct elements in T from any one-parametric family $q(\cdot, x) \in K$ for every $x \in \Omega$ using the Metropolis-Hastings construction. If $q(y, x) \geq a\pi(y)$ for every $x, y \in \Omega$, then the corresponding kernel of the operator satisfies $r(y, x) \geq a\pi(y)$ for every $y \in \Omega$.

Let $\{T_i\}$ be a sequence of elements in V with corresponding a_i in the strong Doeblin condition and densities p_i after applying the operator T_i. Then the theorem in the following section states that the sequence p_i converges towards π with rate $\prod_i(1 - a_i)$. Hence, there may be convergence even if $a_i = 0$ for some elements in the chain as long as a sub-chain has $a_{m_i} > 0$.

For the adaptive chain we have $T_i(\phi_i) = \phi_i$. Then it is also necessary to define a chain of operator U_i satisfying $U_i(\pi) = \pi$ and bound the difference $T_i - U_i$.

There may be different kernel in each iteration and the kernel may change deterministically or stochastically. The convergence is due to contraction properties from the strong Doeblin condition in the space V. The adaptiveness is not critical for convergence, it is only an opportunity to use knowledge in order to improve convergence and reduce correlation.

9. Simulated annealing

In simulated annealing see Geman (1984), it is simulated from the density

$$\pi_i(x) = \beta_i f(x) \exp(-g(x)/T(i)) \tag{26}$$

where it is assumed that $T(i) \to 0$ when $i \to \infty$. Then $\pi_i \to \pi$ pointwise where

$$\pi(x) = \beta f(x) \text{ for } x \in \Omega_{g=0}. \tag{27}$$

Assume that the chain satisfies

$$s_i(y,x) \geq a\phi_i(y) \exp(-M/T(i)) \quad \text{for all } i \geq 0 \text{ and } x, y \in \Omega. \tag{28}$$

This follows from (24) if for some values of $x, y \in \Omega$ it is necessary to pass domains $D_{x,y} \subset \Omega$, where $g(x) > 0$ in order to move from x to y. We may then formulate the following corollary to Theorem 1 which may be used to bound the convergence rate for simulated annealing combined with ordinary MCMC, regenerate chains and adaptive chains. The corollary gives an expression for the convergence rate and has the same bounds on the decrease of T as in Geman (1984).

Corollary 1 The simulated annealing chain satisfies

$$\|p_i - \pi\|_{T.V.} \leq \|\pi_i - \pi\|_{T.V.} + \|\phi_i - \pi_i\|_{T.V.} + \|p_i - \phi_i\|_{T.V.} \tag{29}$$

If $\lim_{n\to\infty} \int_{\Omega_{0<g(x)<1/n}} f(x) d\mu(x) = 0$, then $\|\pi_i - \pi\|_{T.V.} \to 0$.

If the chain satisfies (28) and $T(k) > M/\log(k)$ and $\sum_{j=1}^{\infty}(c_j + b_j) < L$, then for $i \geq 2^n j$

$$\|p_i - \pi_i\|_{T.V.} \leq (L+1)\exp[-an/2] + \sum_{k=j+1}^{i} (c_j + b_j) \to 0$$

when $i, j, n \to \infty$.

Proof The inequality in (29) follows from the triangle inequality. The second part of the corollary follows from

$$\begin{aligned}
\|\pi_i - \pi\|_{T.V.} &= \beta_i \int_{\Omega_{g>0}} f(x) \exp(-g(x)/T(i)) d\mu(x) \\
&\leq \beta(\int_{\Omega_{g>1/n}} f(x) \exp(-g(x)/T(i)) d\mu(x) \\
&\quad + \int_{\Omega_{1/n \geq g>0}} f(x) \exp(-1/(nT(i))) d\mu(x)) \to 0
\end{aligned}$$

when $n \to \infty$.

Set $c_0 = 1$ and $b_0 = 0$. Then (28) implies

$$\begin{aligned}\|p_i - \pi_i\| &\leq \sum_{j=0}^{i-1}(c_j + b_j) \prod_{k=j+1}^{i-1} (1 - a\exp(-M/T(k))) \\ &= \sum_{j=0}^{i-1}(c_j + b_j)\exp[\sum_{k=j+1}^{i-1} \log(1 - a\exp(-M/T(k)))] \\ &\leq \sum_{j=0}^{i-1}(c_j + b_j)\exp[-a\sum_{k=j+1}^{i-1} \exp(-M/T(k))].\end{aligned}$$

If $T(k) > M/\log(k)$

$$\sum_{k=j}^{i}\exp(-M/T(k)) \geq \sum_{k=j}^{i}\exp(-\log(k)) \geq \sum_{k=j}^{i} k^{-1} > n/2$$

for $i \geq 2^n j$. Then

$$\|p_i - \pi_i\| \leq (L+1)\exp[-an/2] + \sum_{k=j}^{i-1}(c_j + b_j) \to 0$$

when $i, j, n \to \infty$. □

10. Some examples

This section gives two small examples. In the first example, it is not possible to use the Metropolis-Hastings algorithm and the second example use simulated annealing.

Example 1 Let $\Omega = (0,1)$, and let f be an unknown continuous monotone increasing function which satisfies $f(y) = 0$ for one value of $y \in (0.1, 0.9)$. The function f may be evaluated in a point x at a high cost. We want to simulate from the distribution $\pi(x) = \text{Uniform}(y - .1, y + .1)$. Hence it is not possible to evaluate the limiting function π and $\pi(x) = 0$ in the largest part of Ω. It is simulated from the adaptive chain algorithm by defining:

$$y_i^L = \max\{0.1, \text{all points } x \text{ where } f \text{ is evaluated and } f(x) < 0\}$$

$$y_i^R = \min\{.9, \text{all points } x \text{ where } f \text{ is evaluated and } f(x) > 0\}.$$

$$r_i(y, x, \tilde{x}^i) = \text{Uniform}(y^L - .1, y^R + .1)$$

We may apply Theorem 1. With probability 1, $a_i \to, 1$ and $b_i, c_i \to 0$, when $i \to \infty$ hence also $\|p_i - \pi\|_{T.V.} \to 0$ with probability 1 when $i \to \infty$.

Example 2 Let $\Omega = [-2, 2]$, $f \equiv 1$, and $g = \max\{1 - |x|, 0\}$. Let π_i be defined by (26) and π be defined by (27) i.e. $\pi(x) = 1/2$ for $1 \leq |x| \leq 2$. Simulate by Metropolis Hastings simulated annealing i.e. oordinary Metropolis Hastings with π_i instead of π in iteration i. If q approximate π in Ω, simulated annealing is not needed. If q only approximate $c\pi$ for a constant c in an interval $(x - 1, x + 1)$ close to the present value x, then simulated annealing is a reasonable algorithm. But the chain x_i does not

satisfy the strong Doeblin condition since it is not possible to jump between any to states in one step. However, the chain $z_i = x_{8i}$ satisfies (28). Hence, z_i and then also x_i converge with $T(k) = M/\log k$ for M sufficient small.

11. Proof of Theorems

Proof of Theorem 1

The strong Doeblin condition (12) implies that s_i may be written as $s_i(y,x) = a_i\phi_i(y) + (1-a_i)v_i(y,x)$ for a transition density v_i. Then

$$\begin{aligned}
\|p_{i+1} - \phi_{i+1}\|_{\text{T.V.}} &= \| \int_\Omega s_i(.,x)(p_i(x) - \phi_i(x))d\mu(x)\|_{\text{T.V.}} \\
&= \| \int_\Omega (a_i\phi_i(.) + (1-a_i)v_i(.,x))(p_i(x) - \phi_i(x))d\mu(x)\|_{\text{T.V.}} \\
&= (1-a_i)\| \int_\Omega v_i(.,x)(p_i(x) - \phi_i(x))d\mu(x)\|_{\text{T.V.}} \\
&\leq (1-a_i)\|p_i - \phi_i\|_{\text{T.V.}}
\end{aligned}$$

Equation (14) is then proved using induction and $\|p_1 - \phi_1\|_{\text{T.V.}} \leq 1$. In order to prove (15) it is necessary with the intermediate result

$$\begin{aligned}
\|\phi_{i+1} - \pi_{i+1}\|_{\text{T.V.}} &\leq \|\pi_{i+1} - \pi_i\|_{\text{T.V.}} + \|\phi_{i+1} - \pi_i\|_{\text{T.V.}} \\
&= c_i + \| \int_\Omega s_i(.,x)\phi_i(x) - \hat{r}_i(.,x)\pi_i(x)d\mu(x)\|_{\text{T.V.}} \\
&= c_i + \| \int_\Omega (s_i(.,x)(\phi_i(x) - \pi_i(x)) + (s_i(.,x) - \hat{r}_i(.,x))\pi_i(x))d\mu(x)\|_{\text{T.V.}} \\
&\leq c_i + \| \int_\Omega s_i(.,x)(\phi_i(x) - \pi_i(x))d\mu(x)\|_{\text{T.V.}} \\
&\qquad + \| \int_\Omega (s_i(.,x) - \hat{r}_i(.,x))\pi_i(x)d\mu(x)\|_{\text{T.V.}} \\
&\leq c_i + b_i + (1-a_i)\|\phi_i - \pi_i\|_{\text{T.V.}}
\end{aligned}$$

Then (15) follows by induction using that $\phi_1 \equiv \pi_1$. Equation (16) is proved by combining the results above and the triangle inequality

$$\|p_{i+1} - \pi\|_{T.V.} \leq \|\pi_{i+1} - \pi\|_{T.V.} + \|\pi_{i+1} - \phi_{i+1}\|_{T.V.} + \|\phi_{i+1} - p_{i+1}\|_{T.V.}$$

If $a_j \geq a > 0$ and $c_j + b_j \to 0$ when $j \to \infty$, then for any $\varepsilon > 0$ there exists k such that $c_j + b_j < \varepsilon$ for $j > k$. This gives for $i > k$

$$\begin{aligned}
\|\phi_{i+1} - \pi_{i+1}\|_{T.V.} &\leq \sum_{j=1}^{i}(c_j + b_j) \prod_{k=j+1}^{i}(1-a_k) \\
&\leq \sum_{j=1}^{k}(c_j + b_j)(1-a)^{i-j} + \varepsilon \sum_{j=0}^{i-k-1}(1-a)^j \\
&< (1-a)^{i-k}\sum_{j=1}^{k}(c_j + b_j) + \varepsilon\frac{1}{a}
\end{aligned}$$

which may be made arbitrary small by choosing ε small and i large.

If $a_j \geq a > 0$ and $c_j + b_j \leq d^j < 1$ for all $j > 0$, then

$$\begin{aligned}
\|\phi_{i+1} - \pi_{i+1}\|_{T.V.} &\leq \sum_{j=1}^{i}(c_j + b_j)\prod_{k=j+1}^{i}(1-a_k) \leq \sum_{j=1}^{i} d^j(1-a)^{i-j} \\
&= (1-a)^i \sum_{j=1}^{i}(d/(1-a))^j \\
&\leq d(1-a)^{i-1}\frac{1}{1-d/(1-a)} \\
&= \frac{d}{1-a-d}(1-a)^i.
\end{aligned}$$

□

Proof of Theorem 2 Convergence in the relative supremum norm is proved by combining (4) and (5). This gives

$$p_{i+1}(y) - \phi_{i+1}(y) = \int_\Omega s_i(y,x)(p_i(x) - \phi_i(x))d\mu(x)$$

which implies

$$\frac{p_{i+1}(y)}{\phi_{i+1}(y)} - 1 = \int_\Omega \frac{s(y,x)}{\phi_{i+1}(y)}(p_i(x) - \phi_i(x))d\mu(x).$$

This is only well-defined for $\phi_{i+1}(y) > 0$. Further calculation gives

$$\begin{aligned}
R^{i+1}(y) &= \int_\Omega \frac{s_i(y,x)}{\phi_{i+1}(y)} R^i(x)\phi_i(x)d\mu(x) \\
&= R^i_M \int_\Omega \frac{s_i(y,x)}{\phi_{i+1}(y)}\phi_i(x)d\mu(x) \\
&\qquad - \int_\Omega \frac{s_i(y,x)}{\phi_{i+1}(y)}(R^i_M - R^i(x))\phi(x)d\mu(x) \\
&\leq R^i_M - a_i \int_\Omega (R^i_M - R^i(x))\pi(x)d\mu(x) \\
&= R^i_M(1-a_i) + a_i \int_\Omega R^i(x)\phi_i(x)d\mu(x) \\
&= R^i_M(1-a_i).
\end{aligned}$$

The strong Doeblin condition (12) is used in the inequality. Define $\tilde{p}$ such that the corresponding $\tilde{R}^i = -R^i$. Note that $\tilde{p}$ may be negative and thus not a density. Perform the same calculation as above with $\tilde{R}$ replacing R. This gives

$$-R^{i+1}(y) \leq R^i_M(1-a_i)$$

which combined with the inequality above gives $R^{i+1}_M \leq R^i_M(1-a_i)$. Induction gives (17).

In order to prove (18) it is necessary with the intermediate result

$$
\begin{aligned}
|\phi_{i+1}(y) - \pi_{i+1}(y)| &\leq |\pi_{i+1}(y) - \pi_i(y)| + |\phi_{i+1}(y) - \pi_i(y)| \\
&\leq \|\pi_{i+1}\|_{\pi_i,\infty}\pi_i(y) + |\int_\Omega s_i(y,x)\phi_i(x) - \hat{r}_i(y,x)\pi_i(x)d\mu(x)| \\
&= \|\pi_{i+1}\|_{\pi_i,\infty}\pi_i(y) + \\
&\quad |\int_\Omega (s_i(y,x)(\phi_i(x) - \pi_i(x)) + (s(.,x) - \hat{r}_i(y,x))\pi_i(x))d\mu(x)| \\
&\leq \|\pi_{i+1}\|_{\pi_i,\infty}\pi_i(y) + |\int_\Omega s_i(y,x)(\phi_i(x) - \pi_i(x))d\mu(x)| \\
&\quad + |\int_\Omega (s_i(y,x) - \hat{r}_i(y,x))\pi_i(x)d\mu(x)| \\
&\leq \|\pi_{i+1}\|_{\pi_i,\infty}\pi_i(y) + (1-a_i)\|\phi_i(y) - \pi_i(y)\|_{\pi_i,\infty}\pi_i(y) + b_i^S\pi_i(y) \\
&= (\|\pi_{i+1}\|_{\pi_i,\infty} + (1-a_i)\|\phi_i\|_{\pi_i,\infty} + b_i^S)\pi_i(y) \\
&= (\|\pi_{i+1}\|_{\pi_i,\infty} + (1-a_i)\|\phi_i\|_{\pi_i,\infty} + b_i^S)\pi_{i+1}(y)(\frac{\pi_i(y) - \pi_{i+1}(y)}{\pi_{i+1}(y)} + 1) \\
&\leq (\|\pi_{i+1}\|_{\pi_i,\infty} + (1-a_i)\|\phi_i\|_{\pi_i,\infty} + b_i^S)\pi_{i+1}(y)(\|\pi_i\|_{\pi_{i+1},\infty} + 1)
\end{aligned}
$$

where the bound on R^i from the previous calculation is used Hence

$$\|\phi_{i+1}\|_{\pi_{i+1},\infty} \leq ((1-a_i)\|\phi_i\|_{\pi_i,\infty} + b_i^S + \|\pi_{i+1}\|_{\pi_i,\infty})(\|\pi_i\|_{\pi_{i+1},\infty} + 1).$$

Then (18) follows by induction using $\phi_1 \equiv \pi_1$. Equation (19) follows from the calculation below

$$
\begin{aligned}
\|p_i\|_{\pi,\infty} &= \sup_{y\in\Omega}\left|\frac{p_i(y)}{\pi(y)} - 1\right| \\
&= \sup_{y\in\Omega}\left|(\frac{p_i(y)}{\phi_i(y)} - 1)\frac{\phi_i(y)}{\pi(y)} + \frac{\phi_i(y)}{\pi(y)} - 1\right| \\
&= \sup_{y\in\Omega}\left|(\frac{p_i(y)}{\phi_i(y)} - 1)\frac{\phi_i(y)}{\pi_i(y)}\frac{\pi_i(y)}{\pi(y)} + (\frac{\phi_i(y)}{\pi_i(y)} - 1)\frac{\pi_i(y)}{\pi(y)} + \frac{\pi_i(y)}{\pi(y)} - 1\right| \\
&\leq \|p_i\|_{\phi_i,\infty}(1 + \|\phi_i\|_{\pi_i,\infty})(1 + \|\pi_i\|_{\pi,\infty}) + \|\phi_i\|_{\pi_i,\infty}(1 + \|\pi_i\|_{\pi,\infty}) + \|\pi_i\|_{\pi,\infty}.
\end{aligned}
$$

Equation (20) and (21) are proved similarly as in the proof for Theorem 1.

It is left to prove the theorem when there is point mass distribution in p_i. The bound on S_i follows directly from $\int p_i(x)d\mu(x) = 1$ and the strong Doeblin condition (12). The bound on R^i is found similarly as the first

calculation in the proof.

$$\begin{aligned}
R^{i+1}(y) &= R^i_M - \int_{\Omega\setminus A_i} \frac{s_i(y,x)}{\phi_i(y)}(R^i_M - R^i(x))\phi_i(x)d\mu(x) \\
&\qquad + \int_{A_i} \frac{s_i(y,x)}{\phi_i(y)}p_i(x)d\mu(x) \\
&\le R^i_M - a_i \int_{\Omega\setminus A_i} (R^i_M - R^i(x))\phi_i(x)d\mu(x) \\
&\qquad + \int_{A_i} \frac{s_i(y,x)}{\phi_i(y)}p_i(x)d\mu(x) \\
&= R^i_M(1-a_i) + a_i \int_{\Omega\setminus A_i} R^i(x)\phi(x)d\mu(x) \\
&\qquad + \int_{A_i} \frac{s_i(y,x)}{\phi_i(y)}p_i(x)d\mu(x) \\
&= R^i_M(1-a_i) - a_i \int_{A_i} p_i(x)\pi(x)d\mu(x) \\
&\qquad + \int_{A_i} \frac{s_i(y,x)}{\pi(y)}p_i(x)d\mu(x) \\
&= R^i_M(1-a_i) + \int_{A_i} (\frac{s_i(y,x)}{\phi_i(y)} - a_i)p_i(x)d\mu(x) \\
&= (R^i_M + D_i S_i)(1-a_i)
\end{aligned}$$

Induction gives

$$R^{i+1}_M \le (R^1_M + S_1 \sum_{j=1}^{i} D_j) \prod_{j=1}^{i} (1-a_j)$$

□

Proof of Theorem 3 The bound on the expected value is found from

$$\begin{aligned}
|f_i - f_\pi| &= \int_\Omega f(x)(p_i(x) - \pi(x))d\mu(x) \\
&\le \int_\Omega f(x)\pi(x)|(p_i(x) - \pi(x))/\pi(x)|d\mu(x) \\
&\le f_\pi \|p_i\|_{\pi,\infty}.
\end{aligned}$$

The bound on the covariance follows from

$$
\begin{aligned}
|\mathrm{Cov}(f(x_i),f(x_j))| &= |\int_{\Omega^2}(f(x_i)-f_i)(f(x_j)-f_j)p_i(x_i)p_{j|i}(x_j|x_i)d\mu(x_i,x_j)| \\
&= |\int_{\Omega^2}(f(x_i)-f_i)(f(x_j)-f_j)p_i(x_i)(p_{j|i}(x_j|x_i)-\pi(x_j))d\mu(x_i,x_j)| \\
&\leq (\int_{\Omega^2}(f(x_i)-f_i)^2 p_i(x)|p_{j|i}(x_j|x_i)-\pi(x_j)|d\mu(x_i,x_j))^{1/2} \\
&\quad (\int_{\Omega^2}(f(x_j)-f_j)^2 p_i(x)|p_{j|i}(x_j|x_i)-\pi(x_j)|d\mu(x_i,x_j))^{1/2} \\
&\leq 2\sigma_i \|p_{j|i}-\pi\|_{T.V.}^{1/2} \\
&\quad (\int_{\Omega^2}(f(x_j)-f_j)^2 p_i(x)\pi(x_j)|\frac{p_{j|i}(x_j|x_i)-\pi(x_j)}{\pi(x_j)}|d\mu(x_i,x_j))^{1/2} \\
&\leq 2\sigma_i \|p_{j|i}-\pi\|_{T.V.}^{1/2}\|p_{j|i}\|_{\pi,\infty}^{1/2}(\int_{\Omega}(f(x_j)-f_j)^2\pi(x_j)d\mu(x_j))^{1/2} \\
&= 2\sigma_i \|p_{j|i}-\pi\|_{T.V.}^{1/2}\|p_{j|i}\|_{\pi,\infty}^{1/2} \\
&\quad (\int_{\Omega}((f(x_j)-f_\pi)^2+(f_\pi-f_j))^2)\pi(x_j)d\mu(x_j))^{1/2} \\
&= 2\sigma_i(\sigma_\pi^2+(f_\pi-f_j))^2)^{1/2}\|p_{j|i}-\pi\|_{T.V.}^{1/2}\|p_{j|i}\|_{\pi,\infty}^{1/2}.
\end{aligned}
$$

When there is point mass in p_i the calculation is similar. □

Acknowledgement. The author thanks Fred Espen Benth, Arnoldo Frigessi, Jon Gjerde, Peter Green, Radford Neal, Øivind Skare and Anne Randi Syversveen for valuable comments. The author is grateful for financial support from Norwegian Research Council.

References

Doob, J. L. (1953) *Stochastic Processes.* Wiley, New York.

S. Geman and D. Geman, (1984) Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images, IEEE Trans. on Pattern Analysis and Machine Intelligence, **6**, 721–741

Geyer, C.J. and Møller, J. (1993) Simulation procedures and likelihood inference for spatial point processes, Technical Report 260, Dept. of Theor. Stat., Institute of Mathematics, University of Aarhus.

Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statistical Science* **7**, 473–483.

Gilks, W.R., Roberts, G.O. and George, E. I.(1994) Adaptive direction sampling, The Statistican, 43, 179-189

Gilks, W.R., S. Richardson and D.J. Spiegelhalter, (1996) Markov Chain Monte Carlo in Practice, Chapman and Hall, London

Gilks, W.R., G.O. Roberts and S. K. Sahu, (1996) Adaptive Markov Chain Monte Carlo through Regeneration, Preprint.

Harrio, H., Saksman, E. and Tamminen, T.(1998) An Adaptive Metropolis algorithm, Preprint

Harrio, H., Saksman, E. and Tamminen, T. (1998) Adaptive proposal distribution for random walk Metropolis algortihm, Preprint

Holden, L. (1997) Geometric convergence of Markov Chains, Preprint, University of Oslo.

Mengersen, K. L. and Tweedie, R. L. (1996) Rates of Convergence of the Hastings and Metropolis Algorithms. *The Annals of Statistics* **24**, 101–121.

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

Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers, J. Amer. Statist. Assoc., 90, 233-241.

Ripley, B. D., (1987) Stochastic Simulation, John Wiley & Sons, New York.
Roberts, G.O. and Gilks, W.R., (1994) Convergence of adaptive direction sampling, J. Multiv., Anal. 49, 287-298.
Roberts, G.O. and Tweedie R. L.,(1998) Bounds on regeneration times and convergence rates for Markov chains, Preprint

Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway
E-mail address: Lars.Holden@nr.no