Dept. of Math. University of Oslo
Statistical Research Report No. 2
ISBN 82–553–1340–0 May 2002
ISSN 0806–3842

Markov Chain Monte Carlo algorithms with independent proposal distribution and their relationship to importance sampling and rejection sampling

J. Gåsemyr

May 24, 2002

Abstract

In this paper we define a class of MCMC algorithms, the generalized self regenerative chains (GSR), generalizing the SR chain of Sahu and Zhigljavski (2001), which contains rejection sampling as a special case. We show that this class contains members that are asymptotically more efficient and converge faster than the SR chains. We also consider generalizations of the Metropolis - Hastings independent chains or Metropolized independent sampling, and for some of these algorithms we are able to give the convergence rates and establish a lower bound for the asymptotic efficiency. All these MCMC algorithms use a proposal distribution that is independent of the current state. We discuss such algorithms generally. We are in particular interested in the number of times a given proposed value occurs consecutively as a state of the chain. We consider this number as a random integer weight that links these algorithms also to importance sampling. We show that for the generalizations of the SR and independent chains the expected values of these weights characterize the stationary distribution.

Key words: Adaptive algorithms, importance sampling, independent chain, Markov chain Monte Carlo algorithms, rejection sampling, self regenerative chain.

1 Introduction

Let $\mathcal{X}$ be a subset of a Euclidean space, on which a probability density π is defined. This density may only be known up to proportionality, and may be very intractable analytically due e. g. to a complex dependence structure or high dimensionality. Let p be another density that approximates π to some degree, from which samples can be generated more or less easily. Suppose we want to estimate $\mu = E_\pi h(X)$ for some real-valued function h defined on $\mathcal{X}$. There exist several methods for using an IID sequence $\{Y^n\}, n = 1, 2, \ldots, N$ of p-distributed variables to this end. Rejection sampling creates independent samples from π, but usually throws away a large proportion of the variables Y^n that are generated. Sampling importance resampling (SIR) generates a sample of approximately independent π-distributed variables. The density p can also be used as a proposal distribution for a Markov chain $\{X^t\}$, for which π is the stationary distribution (see e. g. example 1 and 2 below). After a suitable burn - in period, the chain consists of variables that are approximately or exactly π-distributed, but dependent. In all these methods, μ can be estimated by evaluating h at all members of the sample that is generated and computing the sample mean. On the other hand, importance sampling does not attempt to generate a sample from π. Instead, h is evaluated at each Y^n, but $h(y)$ is weighted by weight proportional to $w(y) = \pi(y)/p(y)$ in the sample mean estimate.

A unifying perspective linking importance sampling to the other algorithms that have been mentioned, arises from the observation that for these algorithms, each p-distributed proposal value Y^n is represented in the final sample a finite number W^n times. The numbers W^n are random, integer valued weights whose distribution depends on Y^n, and, for the Markov chain methods, also on the history of the chain. In the traditional MCMC algorithms, as e. g. the independent chain (see example 2 below), the weight W^n arises through a step-by-step process, very different from the computation of the deterministic importance sampling weight $w(Y^n)$. The value of X^t is determined by a separate iterative step for each t. However, example 1 describes an MCMC algorithm where W^n is determined by a single drawing from a history-independent integer valued distribution, a procedure much more analogous to importance sampling. The sample mean estimate for μ based on an MCMC sample can be written in the form

$$\hat{\mu} = [\sum_{n=1}^{N} h(Y^n) W^n] / [\sum_{n=1}^{N} W^n] \tag{1}$$

Interpreting W^n as $w(Y^n)$, this formula also covers importance sampling; except that the denominator may be replaced by N if w is known exactly, and not only up to proportionality.

The basic idea of this paper is to pursue this unifying perspective, with partcular emphasis on looking at the Markov chain algorithms from this angle. This leads to the construction of some new types of MCMC algorithms. We also study asymptotic efficiency and rate of convergence for several MCMC algorithms based on an independent proposal density p, and investigate the relationship between the random weight W^n and the deterministic (given Y^n) weight $w(Y^n)$.

We recall from Hastings (1970) that a Markov chain with π as stationary distribution can be constructed as follows: Let ψ be a density with the same support as π. Let $\beta(x,y)$ be a symmetric function such that $0 < \beta(x,y) \leq \pi(x)/\psi(x)$. Define

$$\alpha(x,y) = \beta(x,y)/(\pi(x)/\psi(x)). \tag{2}$$

At the $t+1$st iteration, variables X, U are generated independently from respectively ψ and the uniform density on $[0,1]$. Suppose that $X^t = x^t$ and $X = x$. Then X^{t+1} is defined to be x if $U \leq \alpha(x^t, x)$, and x^t otherwise. To fit in our framework, ψ and p must be related somehow through the generation of candidates from p, but are not necessarily identical (see e. g. example 1 below). We will refer to this as a chain of the Hastings type.

Example 1. Given that $Y^n = y$, let $W^n + 1$ be geometrically distributed with parameter $\alpha(y) = 1/(1+\kappa w(y))$, where κ is some constant > 0. Hence, $P(W^n = s|Y^n = y) = \alpha(y)(1-\alpha(y))^s$. The algorithm can be performed even if w is known only up to proportionality, due to the constant κ. This is the selfregenerative (SR) chain defined in Sahu and Zhigljavski (2001). The name is motivated by the fact that the chain regenerates according to the scheme defined by Gilks et al (1998) at each new value Y^n that is accepted. This chain is formally of the Hastings type. To see this, put

$$\phi(y) = p(y)q(y)/\int p(z)q(z)dz \tag{3}$$

where

$$q(y) = 1 - \alpha(y) = \kappa w(y)/(1+\kappa w(y)) \tag{4}$$

Then ϕ represents the distribution of the first accepted Y when successively drawing candidates Y from p, each being accepted with probability $q(Y)$. This may be intuitively obvious to some readers, but in any case it follows from the proof of theorem 10 in section 4. Then the chain can be understood probabilistically as follows: Generate a candidate y from ϕ, and move from x^t to y if $U \leq \alpha(x^t)$, otherwise stay in x^t. In the standard notation of the Hastings type chains (see (2)), we have $\alpha(x,y) = \alpha(x), \psi(y) = \phi(y)$. Hence, using (4) we have

$$\beta(x,y) = \alpha(x)\pi(x)/\phi(x) = \\ \int p(z)q(z)dz\alpha(x)\pi(x)/(p(x)(1-\alpha(x))) = \int p(z)q(z)dz/\kappa, \tag{5}$$

which is constant and in particular symmetric. In practice, the variable Y distributed according to ϕ is redundant and is not generated corresponding to t for which $X^{t-1} = X^t$, i.e. when $U > \alpha(x^{t-1})$. For the SR algorithm, the expected weights are given by

$$E(W^n|Y^n = y) = 1/\alpha(y) - 1 = \kappa w(y) \tag{6}$$

The SR algorithm is generalized in section 2. Conditions for π-stationarity are given, generalizing (5) and (6). We introduce a version, the OSR algorithm, which is theoretically optimal with respect to asymptotic variance. Convergence rates are also analysed.

Example 2. A Hastings type algorithm can be obtained by defining $\alpha(x, y) = \min(1, w(y)/w(x))$ and $\psi = p$. We then have $\beta(x, y) = \min(w(x), w(y))$. This is the Metropolis Hastings algorithm with independent proposal, also called independent chain (IC). This algorithm is studied extensively in Liu (1996). An adaptive version is studied in Gåsemyr (2000, 2002). It is proved in Gåsemyr (2000) that at stationarity

$$E(W^n|Y^n = y) = w(y) \tag{7}$$

This will be generalized in section 3 to a wider class of Markov chains. Section 3 also contains results on convergence rates and asymptotic efficiency for this class, subject to some additional conditions.

In section 4 the two types of chains studied in sections 2 and 3 are viewed from a common perspective. We define a probabilistic structure on the RIW sequences in such a way that the derived chains are Markov, and such that these two types of chains are special cases. Characterizations of π-stationarity are given, and a new algorithm using ingredients from both section 2 and section 3 is presented. In section 5 we give some final comments, primarily addressing the limitations in the applicability of simulation algorithms based on sampling from one single proposal density p, due to the difficulties of approximating very complex densities with such a proposal density.

We conclude this section by introducing some general terminology and notation that will be used throughout this paper. With a slight abuse of language, we will identify a probability density and the distribution it determines.

Definition 1 *Let $\{Y^n\}$ be a sequence of independent samples from p, and let $\{W^n\}$ be a corresponding sequence of associated, random, not necessarily independent, nonnegative integer weights. The sequence $\{(Y^n, W^n\}$ is called a random integer weight (RIW) p-sequence. Define $T_n = \sum_{i=1}^n W^i$ and $N_t = \min\{n : T_n \geq t\}$. Define the sequence $\{X^t\}$ by letting $X^t = Y^n$ if and only if $T_{n-1} < t \leq T_n$, or equivalently, if and only if $N_t = n$. Then we say that $\{X^t\}$ is the chain derived from the sequence $\{(Y^n, W^n)\}$.*

Adding or removing (Y^n, W^n) for which $W^n = 0$ does not affect the derived chain $\{X^t\}$, so the same chain may be derived from different RIW sequences. However, the chain $\{X^t\}$ determines uniquely a sequence $\{(Z^r, W_0^r)\}$ where $W_0^r > 0$ for all r. The relationship between the sequences $\{(Y^n, W^n)\}$ and $\{(Z^r, W_0^r)\}$ is given by the following equations:

$$\begin{aligned} R_n &= \textstyle\sum_{i=1}^n I(W^i > 0) \\ N_0(r) &= \min\{n : R_n = r\} \\ Z^r &= Y^{N_0(r)}, W_0^r = W^{N_0(r)} \end{aligned} \tag{8}$$

We have chosen to include all proofs in the main text, rather than in an appendix. This is because it is necessary to work through most proofs to develop a good understanding of the subject matter of the paper. However, the proofs of theorem 3, including the accompanying lemma 1, and theorem 9 may safely be left out on a first reading.

2 The generalized selfregenerative chain (GSR)

In this section we generalize example 1 of section 1, the selfregenerative (SR) chain of Sahu and Zhigljavski (2001).

2.1 Definition and basic properties of the GSR chain

For the GSR algorithm, the distribution of the weight W^n associated with the candidate Y^n generated from p is determined by functions q, α defined on $\mathcal{X}$, taking values in $[0, 1]$, and is given by the equation

$$P(W^n = s|Y^n = y) = P(VS = s) \tag{9}$$

where V, S are independent, V is Bernoulli with parameter $q(y)$ and S is geometric with parameter $\alpha(y)$. This means that

$$P(W^n = 0|Y^n = y) = 1 - q(y)$$

and

$$P(W^n = s|Y^n = y) = q(y)\alpha(y)(1 - \alpha(y))^{s-1}, s = 1, 2, \dots .$$

For the SR chain (see example 1) we have $q(y) = 1 - \alpha(y)$. Similar to (4) we define $\phi(y) = p(y)q(y)/\int p(z)q(z)dz$. Then clearly the chain $\{X^t\}$ derived from $\{(Y^n, W^n)\}$ according to definition 1 is Markov with transition kernel

$$k(x, y) = \alpha(x)\phi(y) \text{ if } x \neq y, \text{ and } P(X^{t+1} = X^t|X^t = x) = 1 - \alpha(x) \tag{10}$$

(For a formal proof in a more general context, see theorem 10 of section 4). We call this a GSR chain.

Theorem 1 *For a GSR chain the following conditions are equivalent:*

- *(i) The chain has π as stationary distribution.*
- *(ii) $\alpha(y) = q(y)/(\kappa w(y))$ for some constant $\kappa > 0$.*
- *(iii) $E(W^n|Y^n = y) = \kappa w(y)$*
- *(iv) $\alpha(y)\pi(y)/\phi(y)$ is constant.*

Proof: Using (10) it is seen that π is the stationary distribution for $\{X^t\}$ if and only if $\pi(y) = \int \pi(x)\alpha(x)\phi(y)dx + \pi(y)(1-\alpha(y))$. This is equivalent to $\alpha(y)\pi(y) = p(y)q(y)(\int \pi(x)\alpha(x)dx)/(\int p(z)q(z)dz)$. From this the equivalence of (i) and (ii) follows, noting that the stationarity condition is preserved when α is multiplied with some constant c such that $c\alpha(y) \leq 1$ for all y. Noting that $E(W^n|Y^n = y) = q(y)/\alpha(y)$, the equivalence of (ii), and (iii) is straightforward. The equivalence of (ii) and (iv) follows since $\alpha(y)\pi(y)/\phi(y) \propto \alpha(y)w(y)/q(y)$. •

Condition (iv) of the theorem expresses that a GSR chain with π as stationary distribution can be regarded formally as a Hastings type chain with the symmetric function $\beta(x, y)$ being constant, just as for the SR (see (5)).

In practice, the function $w(y)$ is often only known up to proportionality. Theorem 1 can nevertheless be used to construct a GSR chain with π as stationary distribution, since the unknown proportionality constant can be absorbed into the parameter κ.

Note that the accept - reject procedure for candidates Y^n may in fact be regarded as sampling from ϕ by means of rejection sampling with p as candidate density. To see this, introduce $v(y) = \phi(y)/p(y)$ and $v^* = \sup_y v(y) = q^*/(\int p(z)q(z)dz)$, where $q^* = \sup_y q(y)$. Hence, we have $q(y) = [p(y)q(y)/(p(y)\int p(z)q(z)dz)]\int p(z)q(z)dz = v(y)/(v^*/q^*)$. Since $q^* \leq 1$, this implies that $q(y)$ can be used as acceptance criterion for rejection sampling.

2.2 Asymptotic variance, introduction of the OSR algorithm

Consider a GSR for which π is the stationary distribution, and suppose we want to estimate $\mu = E_\pi h(X)$ for some function h for which $\sigma_\pi^2 = \text{var}_\pi h(X) < \infty$. The natural estimator based on an IID sample $Y^1, \ldots, Y^N$ from p is (cf. (1) and (2))

$$\hat{\mu}_N = [\sum_{t=1}^{T_N} h(X^t)]/T_N = [\sum_{n=1}^{N} h(Y^n)W^n]/[\sum_{n=1}^{N} W^n] = [\sum_{r=1}^{R_N} h(Z^r)W_0^r]/[\sum_{r=1}^{R_N} W_0^r] \quad (11)$$

Due to the underlying IID structure, it is easy to study the asymptotic properties of $\hat{\mu}_N$ by means of the central limit theorem. The choice of relevant time scale for evaluating the asymptotic performance is however not obvious. In most situations, the major contributor to computational time is the calculation of $w(y)$, since this involves the usually complicated calculation of a function proportional to $\pi(y)$. Part (ii) of theorem 1 shows that $w(y)$ is needed either for $\alpha(y)$ or for $q(y)$, but not necessarily for both. It is necessary to compute $q(Y^n)$ for each proposed value Y^n, while α is needed only for the accepted values Z^r. In any case, the Markov chain index t is irrelevant. In the SR algorithm, $w(y)$ enters into both $\alpha(y)$ and $q(y)$, and this is the case also for another family of GSR algorithms, the OSR algorithms, that will be suggested below. Hence, we have chosen to use the number N of candidates generated from p as a measure of elapsed time (cf. however the final comment of this subsection).

Theorem 2 *For a GSR chain with π as stationary distribution $\sqrt{N}(\hat{\mu}_N - \mu)$ converges in distrubution to $N(0, \sigma)$ as $N \to \infty$, where*

$$\sigma^2 = 2\int (h(y)-\mu)^2/q(y))w(y)\pi(y)dy - (1/\kappa)\sigma_\pi^2 \tag{12}$$

Proof: For every n we have

$$E(h(Y^n)W^n) = \int h(y)E(W^n|Y^n = y)p(y)dy = \int h(y)\kappa w(y)p(y)dy = \kappa\mu.$$

Replacing h by 1 we obtain $E(W^n) = \kappa$. From this we get $\lim_{N\to\infty} T_N/N = \kappa$, and also $E(h(Y^n)W^n - \mu W^n) = 0$. Hence, $\sqrt{N}(\hat{\mu}_N - \mu) = \sqrt{N}[\sum_{n=1}^N (h(Y^n)W^n - \mu W^n)/N](N/T_N)$ converges by the central limit theorem and Cramer's theorem in distribution to $N(0, \sigma)$, where using (9) and (ii) of theorem 1 we get

$$\begin{aligned}
\sigma^2 &= (1/\kappa)^2 \mathrm{var}(h(Y^n)W^n) = (1/\kappa)^2 E[(W^n)^2(h(Y^n)-\mu)^2] \\
&= (1/\kappa)^2 \int (h(y)-\mu)^2 E((W^n)^2|Y^n = y)p(y)dy \\
&= (1/\kappa)^2 \int (h(y)-\mu)^2 q(y)(2-\alpha(y))/(\alpha(y)^2)p(y)dy \\
&= 2\int [(h(y)-\mu)^2/q(y)]w(y)\pi(y)dy - (1/\kappa)\sigma_\pi^2. \qquad \bullet
\end{aligned}$$

Inserting $q(y) = \kappa w(y)/(1+\kappa w(y))$ in (12) we obtain the asymptotic variance for the SR algorithm as

$$\sigma_{SR}^2 = 2\int (h(y)-\mu)^2 w(y)\pi(y)dy + (1/\kappa)\sigma_\pi^2 = 2\sigma_{IS}^2 + (1/\kappa)\sigma_\pi^2, \tag{13}$$

where σ^2_{IS} is the asymptotic variance for importance sampling.

Equation (12) suggests that it would be efficient to choose $q(y)$ as large as possible, subject to the restrictions $q(y) \leq 1, \alpha(y) = q(y)/\kappa w(y) \leq 1$ Therefore, we propose to use

$$q(y) = \min(1, \kappa w(y)), \alpha(y) = \min(1/(\kappa w(y)), 1) \tag{14}$$

In view of (12) we choose to call the algorithm determined by (14) the optimal selfregenerative chain with parameter κ, abreviated OSR(κ). If $\sup_y w(y) = w^* < \infty$, and if $\kappa = \kappa_1 = 1/w^*$, then $q(y) = w(y)/w^*, \alpha(y) = 1$. Hence, the algorithm is equivalent to rejection sampling, and the resulting chain $\{X^t\}$ consists of independent samples from π. Another special case is obtained by the choice $\kappa = \kappa_2 = 1/w_*$ where $w_* = \inf_y w(y)$ (assuming $w_* > 0$). In this case $q(y) = 1$ and $P(W^n \geq 1) = 1$ for every n. For these two special cases, (12) yields respectively

$$\sigma^2_{\kappa_1} = w^* \sigma^2_\pi \tag{15}$$

and

$$\sigma^2_{\kappa_2} = 2\sigma^2_{IS} - w_* \sigma^2_\pi \tag{16}$$

The relative asymptotic efficiency of the OSR($1/w_*$)- algorithm compared to importance sampling is given by $re(OSR(1/w_*); IS) = \sigma^2_{IS}/\sigma^2_{\kappa_2} = 1/(2 - w_*\sigma^2_\pi/\sigma^2_{IS}) = 1/(2 - w_* re(IS))$, where $re(IS)$ denotes the asymptotic efficiency of importance sampling. This equality holds no matter which function h we want to estimate. Note that (16) is less than σ^2_{SR}, no matter which κ is used for the SR chain. In general, we have the following theorem:

Theorem 3 *Denote by σ^2_κ the asymptotic variance for the OSR(κ) chain. Define $A(\kappa) = \{y : \kappa w(y) \geq 1\}, B(\kappa) = A(\kappa)^c, \bar{B}(\kappa) = \{y : \kappa w(y) \leq 1\}$. Then*

$$\begin{aligned}\sigma^2_\kappa = 2 &\int_{A(\kappa)} (h(y) - \mu)^2 w(y)\pi(y)dy \\ &+ (2/\kappa) \int_{B(\kappa)} (h(y) - \mu)^2 w(y)p(y)dy - (1/\kappa)\sigma^2_\pi \end{aligned} \tag{17}$$

Furthermore, σ^2_κ is minimized by $\kappa = \kappa_0$, where κ_0 is determined by the inequalities

$$2\int_{B(\kappa_0)} (h(y) - \mu)^2 w(y)p(y)dy \leq \sigma^2_\pi \leq 2\int_{\bar{B}(\kappa_0)} (h(y) - \mu)^2 w(y)p(y)dy \tag{18}$$

The minimizing value κ_0 satisfies $1/w^* = \kappa_1 \leq \kappa_0 \leq \kappa_2 = 1/w_*$.

Proof: The verification of (17) is straightforward. Using lemma 1 below, differentiating σ_κ^2 yields right and left derivatives respectively

$$(d/d\kappa)^+\sigma_\kappa^2 = (-1/\kappa^2)[2\int_{B(\kappa)}(h(y)-\mu)^2w(y)p(y)dy - \sigma_\pi^2]$$

and

$$(d/d\kappa)^-\sigma_\kappa^2 = (-1/\kappa^2)[2\int_{\bar{B}(\kappa)}(h(y)-\mu)^2w(y)p(y)dy - \sigma_\pi^2].$$

Hence, the right derivative exceeds the left derivative at points of discontinuity, and the derivative increases in intervals of continuity. Since $\bar{B}(\kappa_1) = \mathcal{X}$, while $B(\kappa_2) = \emptyset$, the expressions in brackets range from σ_π^2 to $-\sigma_\pi^2$ as κ ranges from κ_1 to κ_2. Hence, there exists κ_0 such that $\kappa_1 \leq \kappa_0 \leq \kappa_2$ satisfying (18) and minimizing σ_κ^2. •

Lemma 1 *Let $A(\kappa), B(\kappa), \bar{B}(\kappa)$ be as in theorem 3. Then the function $f(\kappa) = \int_{A(\kappa)}(h(y)-\mu)^2w(y)\pi(y)dy + (1/\kappa)\int_{B(\kappa)}(h(y)-\mu)^2w(y)p(y)dy$ is continuous and piecewise differentiable with right and left derivatives respectively*

$$(d/d\kappa)^+f(\kappa) = (-1/\kappa)^2\int_{B(\kappa)}(h(y)-\mu)^2w(y)p(y)dy$$

and

$$(d/d\kappa)^-f(\kappa) = t(-1/\kappa)^2\int_{\bar{B}(\kappa)}(h(y)-\mu)^2w(y)p(y)dy$$

Proof: For $\delta > 0$, define $C(\delta)B(\kappa) - B(\kappa+\delta) = \{y : p(y)/\kappa > \pi(y) \geq p(y)/(\kappa+\delta)$. If $\delta_n \to 0^+$, then $\cap_{n=1}^\infty C(\delta_n) = \emptyset$ and $\cup_{n=1}^\infty B(\kappa+\delta_n) = B(\kappa) - \cap_{n=1}^\infty C(\delta_n) = B(\kappa)$. We have

$$\begin{aligned}\lim_{\delta\to0^+}(f(\kappa+\delta)-f(\kappa))/\delta &= \lim_{\delta\to0^+}(1/\delta)[\int_{C(\delta)}(h(y)-\mu)^2w(y)\pi(y)dy \\ &\quad + (1/(\kappa+\delta)-1/\kappa)\int_{B(\kappa+\delta)}(h(y)-\mu)^2w(y)p(y)dy \\ &\quad - (1/\kappa)\int_{C(\delta)}(h(y)-\mu)^2w(y)p(y)dy \\ &= \lim_{\delta\to0^+}[(1/\delta)\int_{C(\delta)}(h(y)-\mu)^2w(y)(\pi(y)-p(y)/\kappa)dy] \\ &\quad - (1/\kappa)^2\int_{B(\kappa)}(h(y)-\mu)^2w(y)p(y)dy. \qquad (19)\end{aligned}$$

On $C(\delta)$ we have $-\delta p(y)/\kappa(\kappa+\delta) \leq \pi(y) - p(y)/\kappa < 0$, and hence the expression in brackets in (19) tends to 0 as $\delta \to 0$. This takes care of the right derivative of f.

Now define $D(\delta) = B(\kappa - \delta) - B(\kappa) = \{y : p(y)/\kappa \leq \pi(y) < p(y)/(\kappa - \delta), \delta > 0.$ If $\delta_n \to 0^+$, then $\cap_{n=1}^{\infty} D(\delta_n) = D = \{y : \pi(y) = p(y)/\kappa\}$, and $\cap_{n=1}^{\infty} B(\kappa - \delta_n) = B(\kappa) \cup [\cap_{n=1}^{\infty} D(\delta_n)] = B(\kappa) \cup D = \bar{B}(\kappa)$. Hence,

$$\lim_{\delta \to 0^+} (f(\kappa - \delta) - f(\kappa))/(-\delta) = \lim_{\delta \to 0^+} (-1/\delta)[- \int_{D(\delta)} (h(y) - \mu)^2 w(y)\pi(y)dy$$

$$+ (1/(\kappa - \delta) - 1/\kappa) \int_{B(\kappa - \delta} (h(y) - \mu)^2 w(y)p(y)dy$$

$$+ (1/\kappa) \int_{D(\delta)} (h(y) - \mu)^2 w(y)p(y)dy]$$

$$= \lim_{\delta \to 0^+} [(-1/\delta) \int_{D(\delta)} (h(y) - \mu)^2 w(y)(-\pi(y) + p(y)/\kappa)dy]$$

$$- (1/\kappa)^2 \int_{\bar{B}(\kappa)} (h(y) - \mu)^2 w(y)p(y)dy.$$

The expression in brackets tends to $\lim_{\delta \to 0^+}[(-1/\delta) \int_D (h(y) - \mu)^2 w(y)(-\pi(y) + p(y)/\kappa)dy -$ $(1/\delta) \int_{D(\delta) - D} (h(y) - \mu)^2 w(y)(-\pi(y) + p(y)/\kappa)dy] = 0$, since on D, we have $\pi(y) = p(y)/\kappa$, while on $D(\delta) - D$, we have $0 > -\pi(y) + p(y)/\kappa > -\delta p(y)/\kappa(\kappa - \delta)$. •

Finally, note that if $w(y)$ can be bounded reasonably closely from below by a function $v(y)$ which is easily computed, then an alternative to the OSR algorthm is to define $q(y) = \min(1, \kappa v(y))$. In this case, the computational cost involved in the computation of $\hat{\mu}_N$ can be measured roughly in terms of the number R_N of accepted values, each of which involving the computation of $\alpha(Z^r) = q(Z^r)/(\kappa w(Z^r))$. The performance should then be measured in terms of the asymptotic variance $\sigma^2 \int p(y)q(y)dy$ for the variable $\sqrt{R}(\hat{\mu}_{N_0(R)} - \mu)$. Here σ^2 is as in theorem 2, and $\int p(y)q(y)dy = \lim_{R \to \infty} R/N_0(R)$ represents the proportion of proposed values that are accepted.

2.3 Speed of convergence

The underlying IID structure of the GSR algorithm indicates that when estimating $E_\pi h(X)$ forr some function h, it is unnecessary to exclude an initial set of sample values corresponding to a burn - in period. Therefore, from an estimation point of view the convergence rate would seem to be an irrelevant quantity. Nevertheless, we think that speed of convergence is theoretically interesting and a quantity that is important for the comparison with other algorithms, and hence deserves a thorough examination.

Based on the same reasoning as in the previous subsection (see the discussion following (12)), we will measure the rate of convergence in terms of the number of proposed p-distributed values Y^n rather than on the Markov chain time scale t. Noting from theorem 1 (iii) that $E(W^n) = \kappa$, we have $T_n \approx \kappa n$. Using this approximate equality a crude analysis of convergence rates can be performed through a conventional Markov chain convergence rate analysis for the chain $\{X^t\}$. If such an analysis yields the rate r, the relevant adjusted rate after conversion to the time scale of n is $\rho_0(\kappa) = r^\kappa$. This has been done in Sahu and Zhigljavski (2001) for the SR chain, yielding the rate

$$\rho_0^{SR}(\kappa) = [\kappa w^*/(1+\kappa w^*)]^\kappa = [1 + 1/(\kappa w^*)]^{-\kappa}, \tag{20}$$

assuming w^* is finite. We will generalize this by means of a coupling argument. This argument is also a part of the more accurate analysis of speed of convergence summarized in theorem 4 below.

Consider a GSR chain $\{X^t\}$ with π as stationary distribution, determined by $q(y)$ and $\alpha(y) = q(y)/(\kappa w(y))$ (cf. (ii) of theorem 1). Let $\{X_0^t\}$ be another chain with the same transition kernel, but started at stationarity, i.e. X_0^0 is π-distributed. We denote by p^t, p_0^t the densities of X^t and X_0^t respectively. the two chains are linked in the following way: For each t, let U^t be uniform on $[0, 1]$, and let $\tilde{Y}^t$ be a sample from the density $\phi(y) = p(y)q(y)/\int p(z)q(z)dz$. Define $L^t = I(U^t \leq \alpha(X^t)), L_0^t = I(U^t \leq \alpha(X_0^t))$. We put $X^{t+1} = X^t$ if $L^t = 0, X^{t+1} = \tilde{Y}^{t+1}$ if $L^t = 1$, and similarly $X_0^{t+1} = X_0^t$ if $L_0^t = 0, X_0^{t+1} = \tilde{Y}^{t+1}$ if $L_0^t = 1$. This realises the transition kernel (10) for both chains, and the two chains are coupled at $\tau = \min\{t : L^t = L_0^t = 1\}$. By the coupling inequality the total variation distance satisfies $(1/2)|p^t - \pi| \leq P(\tau \geq t$. Since we clearly have $P(L^t = L_0^t = 1) \geq \alpha^* = \inf_y \alpha(y) = \inf_y q(y)/(\kappa w(y))$ we obtain as an upper bound for the adjusted convergence rate

$$\rho_0(\kappa) = (1 - \alpha^*)^\kappa \tag{21}$$

This coincides with (20) in the case of an SR chain. For the OSR chain we obtain

$$\rho_0^{OSR}(\kappa) = [1 - 1/(\kappa w^*)]^\kappa \tag{22}$$

Both for the SR and the OSR we have that $\alpha^* > 0$ if and only if $w^* < \infty$. In general, $\alpha^* > 0 \Rightarrow w^* < \infty$. In most cases it is natural to choose α and q in such a way that $\alpha^* > 0$ if $w^* < \infty$.

It may be verified that if $a \in R - \{0\}$, then $(1 + a/\kappa)^\kappa$ is an increasing function of κ for $\kappa > \max(0, -a)$. Applying this to (20) and (22), it follows that the functions $\rho_0^{SR}(\kappa)$ and $\rho_0^{OSR}(\kappa)$ are respectively decreasing and increasing with κ, with e^{-1/w^*} as a common limit as $\kappa \to \infty$. Hence $\rho_0^{OSR}(\kappa') \leq e^{-1/w^*} \leq \rho_0^{SR}(\kappa'')$ for all κ', κ''.

The inadequacy of this analysis is clearly revealed when considering small values of κ for the OSR, with $\kappa = 1/w^*$, corresponding to rejection sampling, as an extreme case. Indeed, we have $\lim_{\kappa \to (1/w^*)^+} \rho_0^{OSR}(\kappa) = 0$, a rate that clearly does not reflect the true properties of rejection sampling. Hence, a more accurate analysis is needed.

Our ambition is to express the speed of convergence more directly in terms of the number n of p-distributed proposal values. It is by no means obvious how this should be done. Our approach is to measure the speed of convergence in terms of the number of proposals that have to be generated before the coupling time τ described above is reached.

Recall that according to our standard framework, the sequence $\{Y^n\}$ refers to the proposals needed to generate new states of the chain $\{X^t\}$, to be accepted whenever $V^n \leq q(Y^n)$, where V^n are IID variables, uniform on $[0, 1]$. This means that even though in principle proposals Y from p are needed to obtain $\tilde{Y}^t$ for every t, these are only needed in an abstract sense, and do not contribute to the computation time, unless $t = T_n + 1$ for some n. This is equivalent to $L^{t-1} = I(U^{t-1} \leq \alpha(X^{t-1})) = 1$. If also $L_0^{t-1} = I(U^{t-1} \leq \alpha(X_0^{t-1})) = 1$, then $X^t = X_0^t$, and the two chains stay identical from t onwards. If this occurs for the first time at $n = \nu$, we have $\tau = T_\nu + 1$.

Theorem 4 *Consider a GSR chain determined by α and q with π as stationary distribution, and suppose $\alpha^* = \min_y \alpha(y) > 0$. Define $\nu = \min\{n : W^n \geq 1, L_0^{T_n} = 1\}$, where $L_0^t = I(U^t \leq \alpha(X_0^t))$. Then $P(\nu > n) \leq (1 - \kappa\alpha^*)^n$.*

Proof: $P(\nu = n|\nu > n - 1) =$ the probability that Y^n is accepted, and when at $t = T_n$ it is discarded, X_0^t is also discarded, given that $\nu > n - 1$. This equals

$$P((V^n \leq q(Y^n)) \cap (L_0^{T_n} = 1)|\nu > n - 1) =$$
$$\int P(V^n \leq q(y))P(L_0^{T_n} = 1|Y^n = y, V^n \leq q(y), \nu > n - 1)p(y)dy \qquad (23)$$

For the second factor of the integrand we have

$P(L_0^{T_n} = 1|Y^n = y, V^n \leq q(y), \nu > n - 1) = \sum_{t=T_{n-1}+1}^{\infty} P(L_0^t = 1, T_n = t|Y^n = y, V^n \leq q(y), \nu > n - 1) = \sum_{t=T_{n-1}+1}^{\infty} P(U^t \leq \min(\alpha(y), \alpha(X_0^t))|T_n \geq t, Y^n = y, V^n \leq q(y), \nu > n - 1)P(T_n \geq t|Y^n = y, V^n \leq q(y), \nu > n - 1) \geq \sum_{j=0}^{\infty}(1 - \alpha(y))^j\alpha^* = \alpha^*/\alpha(y)$,

where the inequality follows by replacing $\alpha(X_0^t)$ by α^*. Inserting this in (23) we obtain $P(\nu = n|\nu > n - 1) \geq \int q(y)[\alpha^*/\alpha(y)]p(y)dy = \int \kappa w(y)\alpha^* p(y)dy = \kappa\alpha^*$. Hence, $P(\nu > n) \leq (1 - \kappa\alpha^*)^n$. •

In view of theorem 6 it is natural to introduce

$$\rho(\kappa) = 1 - \kappa\alpha^* \qquad (24)$$

Note that for the SR and the OSR the rates are given respectively by

$$\rho^{SR}(\kappa) = 1 - 1/(\kappa^{-1} + w^*) \qquad (25)$$

and

$$\rho^{OSR}(\kappa) = 1 - 1/w^* \tag{26}$$

Hence, $\rho^{OSR}(\kappa)$ is independent of κ, and is always smaller than $\rho^{SR}(\kappa)$. Unlike $\rho_0^{OSR}(\kappa)$, the rate $\rho^{OSR} = 1 - 1/w^*$ reflects the performance of rejection sampling in a natural way.

Another possibility for measuring the speed of convergence is expressed by the following theorem. It is based on an exact sampling idea that is also used in theorem 8 (see also theorem 7 and corollaries 1 and 2) of section 3. In that context, i. e. generalizations of independent chains, it is shown to lead to the same convergence rate as the one found by Liu (1996) in the special case of independent chains.

Theorem 5 *Consider a GSR chain determined by α and q with π as stationary distribution, and suppose $\alpha^* = \inf_y \alpha(y) > 0$. Define $\nu_1 = min\{n : V^n \leq w(Y^n)\kappa\alpha^*\}$, where $V^n = I(W^n \geq 1)$ are IID uniform variables. Also, define $\tau_1 = T_{\nu_1 - 1} + 1$. Then*

- *(i) $P(\nu_1 > n) = (1 - \kappa\alpha^*)^n$.*
- *(ii) X^t is π-distributed given that $t \geq \tau_1$.*

Proof: Note that $\kappa\alpha^* \leq q(y)/w(y)$ for all y implies that $w(y)\kappa\alpha^* \leq 1$. Therefore, we have $P(V^n \leq w(Y^n)\kappa\alpha^*) = \int w(y)\kappa\alpha^* p(y)dy = \kappa\alpha^*$, proving (i). Note that $w(y)\kappa\alpha^* = w(y)\inf_z(\kappa\alpha(z)) = w(y)\inf_z q(z)/w(z) \leq q(y)$. Hence $V^n \leq w(Y^n)\kappa\alpha^*$ implies that Y^n is accepted as a new state, i. e. $W^n \geq 1$. It follows that $W^{\nu_1} \geq 1$ and $X^{\tau_1} = X^{T_{\nu_1 - 1}+1} = Y^{\nu_1}$. Hence, for any $A \subseteq \mathcal{X}$ we have

$P(X^{\tau_1} \in A) = P(Y^{\nu_1} \in A) = \sum_{n=1}^{\infty} P((Y^n \in A) \cap (\nu_1 = n)) = \sum_{n=1}^{\infty} P((Y^n \in A) \cap (\nu_1 = n) | \nu_1 > n-1) P(\nu_1 > n-1) = \sum_{n=1}^{\infty} P((Y^n \in A) \cap (V^n \leq w(Y^n)\kappa\alpha^*))(1-\kappa\alpha^*)^{n-1} = (1/\kappa\alpha^*) \int_A \kappa\alpha^* w(y)p(y)dy = \int_A \pi(y)dy$. Hence, X^{τ_1} is π-distributed. It follows from the stationarity of the chain with respect to π that X^t is π-distributed given $t \geq \tau_1$. •

Since there seems to be no canonical way of measuring the speed of convergence in terms of n, it is somewhat reassuring that theorems 4 and 5 give the same answer. Theorem 5 is stronger in the sense that it describes a geometric rate for obtaining exact samples from π. On the other hand, in theorem 4 the geometric rate only gives an upper bound for the coupling time, so the actual coupling may potentially take place much faster.

3 The generalized independent chain (GIC)

Let $\alpha(x, y)$ be any function such that $0 < \alpha(x, y) \leq 1$. Consider a Markov chain where the current state x^t is replaced by a proposed value y sampled from p with probability $\alpha(x^t, y)$. With probability $1 - \alpha(x^t, y)$ the chain remains at x^t. Hence, the structure is the same as for a chain of the Hastings type with $\psi = p$, such as the IC described in example 2, except that we do not require in general that $\alpha(x, y)\pi(x)/\psi(x) = \alpha(x, y)w(x)$ is symmetric. Such a chain will be called a generalized independent chain (GIC). The absence of the symmetry condition implies that stationarity with respect to π must be demonstrated directly in the general case.

For the GIC we have the transition kernel $k(x, y) = \alpha(x, y)p(y)$ if $x \neq y$,

$$P(X^{t+1} = X^t | X^t = x) = 1 - \alpha(x) \tag{27}$$

where $\alpha(x) = \int p(z)\alpha(x, z)dz$. The distribution of the weights are given by

$$P(W^{t+1} = s | X^t = x, Y^{t+1} = y) = P(VS = s) \tag{28}$$

where V, S are independent, V is Bernoulli with parameter $\alpha(x, y)$ and S is geometric with parameter $\alpha(y)$. Clearly, X^t is determined by $Y^1, W^1, \ldots, Y^t, W^t$. Hence, (28) defines a conditional probability distribution for W^{t+1} given $Y^1, W^1, \ldots, Y^t, W^t, Y^{t+1}$, and $X^1, X^2, \ldots$ is the chain derived from the RIW sequence $\{(Y^t, W^t)\}$ in the sense of definition 1. We now prove, as alluded to in section 1 (see example 2), that the expected weights under stationarity for a GIC are given by the importance weights w, if the stationary distribution is π. In fact, this implication can be strengthened to an equivalence:

Theorem 6 *A GIC with p as proposal distribution has π as stationary distribution if and only if $E(W^{t+1} | Y^{t+1} = y, X^t \sim \pi) = w(y)$.*

Proof: By (27), π is the stationary distribution if and only if $\pi(y) = \int \pi(x)\alpha(x, y) p(y)dx + \pi(y)(1 - \alpha(y))$. This is equivalent to $w(y) = (\int \pi(x)\alpha(x, y)dx)/\alpha(y)$. This latter expression equals $E(W^{t+1} | Y^{t+1} = y, X^t \sim \pi)$ by (28). •

From now on we restrict attention to GIC's of the Hastings type; i. e. chains for which $\alpha(x, y)w(y)$ is symmetric. The following proposition shows that the IC replaces the current value more often than any other GIC of the Hastings type with the same proposal p.

Proposition 1 *Define $\alpha_{IC}(x, y) = \min(1, w(y)/w(x))$, $\alpha_{IC}(x) = \int \alpha_{IC}(x, z)p(z)dz$, where p is some proposal distribution. Then for any GIC of the Hastings type with the same proposal distribution and π as stationary distribution we have $\alpha(x, y) \leq \alpha_{IC}(x, y)$ for all $x, y \in \mathcal{X}$, and hence $\alpha(x) \leq \alpha_{IC}(x)$*

Proof: Any GIC of the Hastings type with p, π as respectively proposal and stationary distribution is determined by a symmetric function β satisfying (see (2)) $\beta(x, y) = \alpha(x, y)w(x) \leq w(x)$. By symmetry, we also have $\beta(x, y) \leq w(y)$, and hence $\beta(x, y) \leq \min(w(x), w(y))$. This implies $\alpha(x, y) \leq \alpha_{IC}(x, y)$. •

Suppose now that $w^* = \sup_x w(x)$ is finite. Let (Y^{t+1}, U^{t+1}) be the pair generated from $p\times$ the uniform density on $[0,1]$ at iteration $t+1$ when running an IC with target distribution π. For any $A \subseteq \mathcal{X}$ we have

$P(X^{t+1} \in A | X^t = x, U^{t+1} \leq w(Y^{t+1})/w^*) = P(X^{t+1} \in A | X^t$
$= x, U^{t+1} \leq \alpha_{IC}(x, Y^{t+1}), U^{t+1} \leq w(Y^{t+1})/w^*) = P(Y^{t+1} \in A | X^t = x, U^{t+1} \leq$
$\alpha_{IC}(x, Y^{t+1}), U^{t+1} \leq w(Y^{t+1})/w^*) = P(Y^{t+1} \in A | U^{t+1} \leq w(y^{t+1})/w^*) = \int_A \pi(y)dy,$

where the last equality follows from the well known properties of the rejection sampler. This proves the following theorem 7, and essentially also the subsequent corollary 1 and the generalization given in corollary 2. A suitably modified version of theorem 7 is also valid for adaptive independent chains, as defined in Gåsemyr (2002); see theorem 1 of that paper.

Theorem 7 *Let $\{X^t\}$ be an IC with target distribution π. Then the state X^{t+1} is π-distributed and independent of X^t given that $U^{t+1} \leq w(Y^{t+1})/w^*$.*

The following corollary is a trivial consequence of theorem 7:

Corollary 1 *Let $\{X^t\}$ be an IC with target distribution π. Let $\tau_m, m = 1, 2, \ldots$ be the successive times t for which $U^t \leq w(Y^t)/w^*)$, i.e. $\tau_m = \min\{t : \sum_{s=1}^t I(U^s \leq w(Y^s)/w^*) = m\}$. Then the X^{τ_m} are independent samples from π, and the sets $\{X^{\tau_m}, X^{\tau_m+1}, \ldots, X^{\tau_{m+1}-1}\}, m = 1, 2, \ldots$ are independent. Moreover, X^s is π-distributed for $s \geq t$, given that $t = \tau_1$.*

Note that the set $\{X^{\tau_1}, \ldots, X^{\tau_m}\}$ can be viewed as the sample from π obtained from a sequence of IID p-distributed variables $Y^1, \ldots, Y^T$ by means of rejection sampling, assuming $m = \max\{i : \tau_i \leq T\}$.

The proof of these results depends crucially on the inequality

$$\alpha(x, y) \geq w(y)/w^*, \tag{29}$$

which is not necessarily true for an arbitrary GIC. However, if $\alpha(x, y) \geq w(y)/(cw^*)$ for some $c > 1$, the above theorem and corollary still apply if we replace w^* with cw^* in the statements. Moreover, even if $\inf(\alpha(x, y)/w(y)) = 0$, we may construct a modified version for which these results apply:

Corollary 2 *For an arbitrary GIC of the Hastings type, having π as stationary distribution, define a modified symmetric function $\beta_M(x, y) = \max(\beta(x, y), w(x)w(y)/w^*)$ and correspondingly $\alpha_M(x, y) = \max(\alpha(x, y), w(y)/w^*)$. Then the correspondingly modified GIC has π as stationary distribution, and theorem 7 and corollary 1 apply.*

Proof: Note that β_M satisfies $\beta_M(x, y) \leq w(x)$, as required in (2). The stationarity follows by the symmetry of β_M. Noting that α_M satisfies (29), the proof is identical to the proof of theorem 7 and corollary 1. •

Note that if (29) is satisfied, then $\alpha_M(x,y) = \alpha(x,y), \beta_M(x,y) = \beta(x,y)$ for all x, y. For such GIC's, the convergence rate is easily derived. The following theorem shows that the convergence rate found by Liu (1996) for IC's can be generalized to such chains.

Theorem 8 *Suppose (29) is satisfied for a GIC of the Hastings type . Then the chain converges in total variation norm with a rate $r \leq 1 - 1/w^*$.*

Proof: Note that $P(U^t \leq w(Y^t)/w^*) = 1/w^*$, and hence $P(\tau_1 > t) = (1 - 1/w^*)^t$. By corollaries 1 and 2 $P(X^t \in A|\tau_1 \leq t) = \int_A \pi(x)dx$ for any $A \subseteq \mathcal{X}$. Conditioning on the events $(\tau_1 \leq t), (\tau_1 > t)$ shows that $|P(X^t \in A) - \int_A \pi(x)dx| \leq (1 - 1/w^*)^t$. •

We conclude this section by considering asymptotic efficiency. Let $\{Y^t\}$ be a sequence of independent samples from p. If w^* is finite, an estimate of $\mu = E_\pi(h(X))$ can be obtained by taking the mean of $h(X^{\tau_1}), \ldots, h(X^{\tau_m})$, where $\tau_i, i = 1, 2, \ldots, m$ are as in corollary 1. This is in fact equivalent to estimating μ by means of the rejection sampler (See the comment after corollary 1). An alternative estimate can be based on the entire output from running a GIC. In the next theorem, we compare the asymptotic efficiencies of these methods.

Theorem 9 *Consider a GIC $\{X^t\}$ of the Hastings type satisfying (29). Define $\hat{\mu}^R(m) = (1/m)\sum_{i=1}^m h(X^{\tau_i})$ and $\hat{\mu}^M(m) = (1/\tau_m)\sum_{t=1}^{\tau_m} h(X^t)$. Define $\rho(m) = var(\hat{\mu}^M(m))/var(\hat{\mu}^R(m))$. Then $\lim_{m\to\infty} \rho(m) \leq 2 - 1/w^*$.*

Proof: Define $\tau_0 = 0$ and $R_i = \tau_i - \tau_{i-1}, i = 1, \ldots, m$. Using corollaries 1 and 2 The variables $X^1, \ldots, X^{\tau_m}$ may be grouped into independent segments $\{X^1, \ldots, X^{\tau_1 - 1}\}$, $\{X^{\tau_1}, \ldots, X^{\tau_2 - 1}\}, \ldots, \{X^{\tau_m}\}$ with respectively $R_1 - 1, R_2, R_3, \ldots, R_m, 1$ variables. The variables $R_i, i = 1, \ldots, m$ are independent and geometrically distributed with parameter $1/w^*$. We have

$$\tau_m = \sum_{i=1}^m R_i \tag{30}$$

Define $\hat{\mu}_i = (1/R_i)\sum_{t=\tau_{i-1}}^{\tau_i - 1} h(X^t), i = 2, 3, \ldots, m$. For $i = 1$ the expression is for convenience slightly modified by replacing $h(X^0)$ by $h(X^{\tau_m})$. This gives

$$\hat{\mu}^M(m) = (1/\tau_m)\sum_{i=1}^m R_i\hat{\mu}_i \tag{31}$$

Clearly, the $\hat{\mu}_i$'s are independent. On the other hand, we make no assumptions on the covariance structure within each segment $\{X^{\tau_i}, X^{\tau_i+1}, \ldots, X^{\tau_{i+1}-1}\}$ of the Markov chain. This means that we may have $\text{cov}(h(X^t), h(X^{t+s})) = \sigma^2 = \text{var}_\pi(h(X))$ given that $\tau_i \leq t < t+s < \tau_{i+1}$ for some i, indicating a deterministic dependence between samples from the same segment. Hence, we only base our comparison on the very conservative bound

$$\sigma_M^2 = \mathrm{var}(\hat{\mu}_i) \leq \sigma^2 \tag{32}$$

Using (30), (31) and (32), the independence of the $\hat{\mu}_i$'s and the symmetry of the $R_i, i = 1, \ldots, m$ this gives

$$\begin{aligned}
&\rho(m) = m\,\mathrm{var}(\hat{\mu}^M(m))/\sigma^2 = (m/\sigma^2)[E(\mathrm{var}(\hat{\mu}^M(m)|\tau_1, \ldots, \tau_m))\\
&+\mathrm{var}(E(\hat{\mu}^M(m)|\tau_1, \ldots, \tau_m))] = (m/\sigma^2)[E((1/\tau_m^2)(\textstyle\sum_{i=1}^m R_i^2 \mathrm{var}(\hat{\mu}_i|\tau_1, \ldots, \tau_m))\\
&+\mathrm{var}(\mu|\tau_1, \ldots, \tau_m)] \leq mE((1/\tau_m^2)(\textstyle\sum_{i=1}^m R_i^2))\\
&= m^2 E((1/\tau_m^2)R_1^2) \leq m^2 E((R_1^2/(\textstyle\sum_{i=2}^m R_i)^2) = E(R_1^2)E(((1/m)\textstyle\sum_{i=2}^m R_i)^{-2}).
\end{aligned}$$

Since R_1 is geometrically distributed with parameter $1/w^*$, the first factor is $2(w^*)^2 - w^*$, while the second factor tends to $(1/w^*)^2$ by the strong law of large numbers and the bounded convergence theorem. This completes the proof. •

In fact, this matches exactly the result obtained by a completely different method in Liu (1996) in the case of an IC with finite state spaces. The result shows that rejection sampling may potentially be twice as efficient as the GIC. However, our result is based on assuming $\mathrm{cov}(h(X^t), h(X^{t+s})) = \sigma^2$ given that $\tau_i \leq t < t+s < \tau_{i+1}$ for some i. By a reasonable decay of autocovariances, the GIC will be much more efficient. Furthermore, if w^* has to be replaced by an upper bound c^*, the efficiency of the GIC remains unchanged, whereas the effeciency of the rejection method will be reduced.

If we modify the estimate for μ based on the GIC to $\hat{\mu}_1^M(m) = (1/m)\sum_{i=1}^m \hat{\mu}_i$, we obtain a corresponding ratio of variances $\rho_1(m)$ satisfying $\rho_1(m) \leq 1$. The fact that we may have $\rho(m) > 1$ for the standard estimate $\hat{\mu}^M(m)$ is accounted for by the extra variability due to the random weights $R_i/(\sum_{j=1}^m R_j)$ allotted to the $\hat{\mu}_i$. This does not necessarily mean that the estimate $\hat{\mu}_1^M(m)$ using fixed weights $1/m$ is better in practice.

4 A common framework for GSR and GIC

The GSR chains and the GIC's have many common features. Both types of chains are Markov, based on independent proposals from a fixed density p and are derived from an RIW p- sequence $\{(Y^n, W^n)\}$ as in definition 1, and the expected weight allotted to a proposed y is proportional to the importance weight $w(y)$ (unconditionally for the GSR; under stationarity for the GIC). However, the structures of the two types of chains, as described in the previous sections, are somewhat different. In this section we present a common framework within which these two types of chains can be viewed as special cases.

Let $\alpha(x)$ and $q(x,y)$ be functions on $\mathcal{X}$ and $\mathcal{X}^2$ respectively, taking values in $(0,1]$. Define an RIW p-sequence with corresponding derived chain $\{X^t\}$ by the equation

$$\begin{aligned} P(W^{n+1} = s|Y^1,&W^1,\dots,Y^n,W^n,Y^{n+1}) \\ &= P(W^{n+1}=s|X^{T_n},Y^{n+1}) = P(VS=s) \end{aligned} \tag{33}$$

where V,S are independent, V is Bernoulli with parameter $q(x,y)$ given that $X^{T_n}=x$, $Y^{n+1}=y$, and S is geometric with parameter $\alpha(y)$ given that $Y^{n+1}=y$. Note that this implies in particular that

$$P(W^{n+1} \geq s+1|W^{n+1} \geq s, Y^{n+1}=y) = 1-\alpha(y) \tag{34}$$

for any $s \geq 1$.

Theorem 10 *Let $\alpha(x)$ and $q(x,y)$ be functions on $\mathcal{X}$ and $\mathcal{X}^2$ respectively, taking values in $(0,1]$. Let $\{(Y^n,W^n)\}$ be the RIW p-sequence defined by α,q through (33). Then the chain $\{X^t\}$ derived from $\{(Y^n,W^n)\}$ is a Markov chain with $P(X^{t+1}=X^t|X^t=x)=1-\alpha(x)$, and transition kernel*

$$k(x,y) = \alpha(x)\phi(y|x) \textit{ for } y \neq x \tag{35}$$

where ϕ is defined by

$$\phi(y|x) = p(y)q(x,y)/\int p(z)q(x,z)dz \tag{36}$$

The chain $\{X^t\}$ is called the random integer weight (RIW) chain determined by α and q.

Proof: Let t be an arbitrary integer, and suppose $X^t=x^t$. If $N_t=n$, then $s=t-T_{n-1}\geq 1$ and $Y^n=X^t$. This implies $W^n \geq s$ and by (34) $P(X^{t+1}=X^t|X^1=x^1,\dots,X^t=x^t,N_t=n)=P(W^n\geq s+1|W^n\geq s,Y^n=x^t)=1-\alpha(x^t)$. By summing over the distribution of N_t it follows that $P(X^{t+1}=X^t|X^1=x^1,\dots,X^t=x^t)=1-\alpha(x^t)$. Furthermore, if $y\neq x^t$, the probability density for X^{t+1} at y given the history is given by $P(X^{t+1}\neq X^t|X^1=x^1,\dots,X^t=x^t)\times [p(y)q(x^t,y)+(1-\int p(z)q(x^t,z)dz)p(y)q(x^t,y)+(1-\int p(z)q(x^t,z)dz)^2p(y)q(x^t,y)+\dots]=\alpha(x^t)\phi(y|x^t)$. •

Corollary 3 *Let α,q be as in theorem 10, and let c be a real- valued function on $\mathcal{X}$. Suppose $q'(x,y)=c(x)q(x,y)\leq 1$ for all $x,y\in\mathcal{X}$. Then α,q and α,q' determine probabilistically identical RIW Markov chains, and the variables $N_t, t=1,2,\dots$ are stochastically minimized by choosing $q'(x,y)=q(x,y)/\sup_y q(x,y)$.*

Proof: The first assertion follows from (35) and (36), the second by the observation (implicit in the proof of theorem 10) that $\phi(y|x)$ represents the density for the first accepted value from a sequence of candidates y drawn from p, each being accepted with probability $q(x, y)$. Therefore, by maximizing q by choosing $q = q'$ as indicated, the number of candidates drawn before this first acceptance is stochastically minimized, thereby also minimizing N_t. •

If q and q' are related as in corollary 3, we say that α, q' determine a version of the RIW chain determined by α, q.

By setting $q(x, y) = q(y)$ independent of x in theorem 10, we obtain the GSR chains as a special case. If

$$\int p(z)q(x, z)dz = \alpha(x) \text{ for all } x \tag{37}$$

the transition kernel coincides with that of a GIC with $\alpha(x, y) = q(x, y)$, and the derived RIW chain is probabilistically equivalent to a GIC. Also, we may obtain a probabilistically equivalent GIC with $\alpha(x, y) = \alpha(x)q(x, y)/ \int p(z)q(x, z)dz$ if this latter quantity is bounded by 1. But even though the GIC is also derived from an RIW p-sequence, the probabilistic relationship between the sequence Y^n and the chain X^t is different. For the GIC, the samples y from p are not only used to obtain a $\phi(y|x)$-distributed successor to the current state x, but also to obtain a realisation of a geometric variable with parameter $\alpha(x) = \int p(z)\alpha(x, z)dz$, the weight associated with x. The first y that defeats x, thereby fixing the value of the geometric variable, is also accepted as the new state of the chain. This is usually a better way to implement the chain than an implementation using (33). This latter implementation requires knowing $\alpha(x)$, which is not easily computed from (37) in general.

So far, we have made no mention of any stationary distribution for an RIW chain. In the rest of this section, we will investigate the relationship between a stationary distribution π and other aspects of the framework.

Recall that we use the notation $\{Z^r, W_0^r\}$ for the subsequence of an RIW sequence $\{Y^n, W^n\}$ obtained by removing all (Y^n, W^n) with $W^n = 0$.

Theorem 11 *Let α, q be as in theorem 10, and let $\{(Y^n, W^n)\}$ be an RIW p-sequence with distribution defined by (33). Then $\{Z^r\}$ is a Markov chain with transition kernel $k(z^r, z) = \phi(z|z^r)$, and the derived RIW chain $\{X^t\}$ has π as stationary distribution if and only if $\{Z^r\}$ has a stationary distribution with density $\chi(x) = \pi(x)\alpha(x)/ \int \pi(z)\alpha(z)dz$.*

Proof: In view of the proof of theorem 10 we consider the first part as obvious. The chain $\{X^t\}$ has π as stationary distribution if and only if $\int \pi(x)\alpha(x)\phi(y|x)dx+$

$\pi(y)(1-\alpha(y) = \pi(y)$, which is equivalent to

$$\pi(y)\alpha(y) = \int \pi(x)\alpha(x)\phi(y|x)dx \tag{38}$$

The result follows by normalizing, i.e. dividing both sides by $\int \pi(z)\alpha(z)dz$. •

Theorem 11 generalizes the equivalence between (i) and (ii) in theorem 1. Indeed, in the case of a GSR chain, the corresponding chain $\{Z^r\}$ consists of IID samples from ϕ, and the transition kernel $\phi(y|x)$ and the stationary distribution χ both coincide with ϕ. Part (ii) of theorem 1 says that $\alpha(y)\pi(y) \propto p(y)q(y) \propto \phi(y) = \chi(y)$, consistent with the conclusion of theorem 11.

Corollary 4 *Suppose α, q determine an RIW chain with π as stationary distribution. Let c be a constant > 0 such that $\alpha'(x) = c\alpha(x) \leq 1$ for all x. Then the chain determined by α', q also has π as stationary distribution.*

Proof: Replacing α by α' leaves χ unchanged, and hence preserves the stationarity condition $\int \chi(x)\phi(y|x)dx = \chi(y)$. •

If α and α' are related as in corollary 4 and q and q' are related as in corollary 3, we say that α', q' determine a scaled version of the chain determined by α, q. The associated $\{Z^r\}$-chains have the same transition kernel and are probabilistically identical, but the geometric parameters determining the distribution of the positive weights W_0^r are scaled by a common factor.

The following proposition illustrates the concept of a scaled version of an RIW chain.

Proposition 2 *Consider a GSR chain determined by $\alpha(x)$ and $q(y)$ for which π is the stationary distribution. Then the GIC with $\alpha(x,y) = \alpha(x)q(y)$ is a scaled version which is of the Hastings type.*

Proof: We obtain a scaled version by defining $q'(x,y) = \alpha(x,y) = \alpha(x)q(y), \alpha'(x) = \alpha(x)\int p(z)q(z)dz$. This is a GIC, since $\int p(z)q'(x,z)dz = \alpha'(x)$ (see (37)). We have $\alpha(x,y)w(x) = \alpha(x)w(x)q(y) \propto q(x)q(y)$ by theorem 1. Hence, $\alpha(x,y)w(x)$ is symmetric, so that the GIC is of the Hastings type. •

As demonstrated by corollary 3 (see also corollary 4), there is a great flexibility in the choice of probabilistic structure for an RIW p-sequence from which an RIW chain with a particular probabilistic structure (or a particular stationary distribution) can be derived. Therefore, one can not expect a similar strict relationship between the stationary distribution π and the expected weights $E(W^n|Y^n = y)$ as indicated for GSR chains and GIC's in theorems 1 and 6 respectively. However, we have the following theorem.

Theorem 12 *Suppose π is the stationary distribution for the RIW chain determined by α and q, and let χ be as in theorem 11. Suppose that $\gamma(x) = \int p(z)q(x,z)dz$ satifies* $\inf_x \gamma(x) > 0$.

(i) Assume the initial state X^0 is χ- distributed. There exists a version of the chain for which $E(W^n|Y^n = y) = \kappa w(y)$ for some constant $\kappa > 0 (n = 1, 2, \dots)$.

(ii) Assume that the initial state X^0 is π- distributed. Then there exists a scaled version of the chain which is a GIC and has $E(W^n|Y^n = y) = w(y)$.

Proof: By theorem 11, if $X^0 = Z^0$ is χ-distributed, then Z^r is χ-distributed for all r. Under stationarity for $\{Z^r\}$ we have by (33) $P(W^n \geq 1|Y^n = y) = \int \chi(x)q(x,y)dx$, and $E(W^n|W^n \geq 1, Y^n = y) = 1/\alpha(y)$, and hence

$$E(W^n|Y^n = y) = (\int \chi(x)q(x,y)dx)/\alpha(y) \tag{39}$$

On the other hand, since π is the stationary distribution for the RIW chain, we obtain by using (38) and then (36) and theorem 11

$$\pi(y)\alpha(y) = \int \pi(x)\alpha(x)\phi(y|x)dx =$$

$$p(y)(\int \pi(z)\alpha(z)dz) \int \chi(x)[q(x,y)/\gamma(x)]dx. \tag{40}$$

Define $c = \sup_{x,y}[q(x,y)/\gamma(x)]$ and $c(x) = 1/(c\gamma(x))$. Define a new version of the chain by replacing $q(x,y)$ by $q'(x,y) = c(x)q(x,y)$. Dividing by $\pi(y)$ on both sides, (40) can then be written as $\alpha(y) = [c \int \pi(z)\alpha(z)dz/w(y)] \int \chi(x)q'(x,y)dx$. For this version we have by (39) $E(W^n|Y^n = y) = (\int \chi(x)q'(x,y)dx)/\alpha(y) = w(y)/(c \int \pi(z)\alpha(z)dz)$. This proves (i).

To prove (ii), define instead $c = \sup_{x,y}[\alpha(x)q(x,y)/\gamma(x)]$. Note that for any $x, \sup_y q(x,y) \geq \int p(z)q(x,z)dz = \gamma(x)$, so that $c \geq \alpha(x)$ for every x, and hence $\alpha'(x) = \alpha(x)/c \leq 1$. Also define $c(x) = \alpha(x)/(c\gamma(x))$ and $q'(x,y) = c(x)q(x,y) = (\alpha(x)q(x,y))/(c\gamma(x))$. Then $q'(x,y) \leq 1$ by the definition of c, and $\int p(z)q'(x,z)dz = \alpha(x)/c = \alpha'(x)$, so that the pair (α', q') satisfies (37) and consequently determines a scaled version of the original chain which is (probabilistically equivalent to) a GIC. By corollaries 3 and 4, this GIC still has π as stationary distribution, and by theorem 6 we obtain $E(W^n|Y^n = y) = w(y)$. •

We conclude this section by constructing an algorithm resulting in an RIW chain which may be convenient in some situations. As a motivation, recall the observation made at the end of section 2.1 that one step in the GSR algorithm consists of sampling y from ϕ by means of rejection sampling with p as proposal density. The other step consists of generating a weight which is geometrically distributed with parameter $\alpha(y)$. As described in section 2, the density ϕ is a secondary entity, determined by the choice of an acceptance probability q. Let us now change perspective and suppose that we choose ϕ as a presumably good approximation to π, which is analytically much more tractable than π, but that there exists no standard procedure for sampling directly from ϕ. Letting $v(y) = \phi(y)/p(y), v^* = \sup_y v(y)$, define $q(y) = v(y)/v^*, \alpha(y) = c\phi(y)/\pi(y)$ for some $c > 0$. Then q and α determine a GSR chain with π as stationary distribution by theorem 1, part (ii). Since the computation of $q(y)$ does not involve computation of $w(y) = \pi(y)/p(y)$, the generation of ϕ-distributed samples y in this way may work at an acceptable speed, even if v^* is quite large. We may regard this procedure as a way of subjecting candidates from p to a relatively cheap "prescreening" by ϕ, before putting them to the more expensive test by π.

The algorithm constructed in the following example is intended to handle the same kind of situation. It replaces the rejection sampling step in the procedure deskribed above by a GIC step, and may be an alternative if the rejection sampling step is too inefficient.

Example 3. Let ϕ be a density approximating π. Define $\alpha_1(y) = c\phi(y)/\pi(y)$ for some $c > 0$. Let $q(x, y)$ be an acceptance probability for accepting a candidate y from p, when the current state is x, for a GIC chain with ϕ as stationary distribution. For instance, we may have $q(x, y) = \min(1, \phi(y)p(x)/\phi(x)p(y))$. Now the idea is to replace the rejection sampling step in the algorithm described in the preceding paragraph by a GIC step determined by q. This means that the successor y of a state x for the chain, rather than being sampled from ϕ, is generated from the distribution $p(y)q(x, y) + (1 - \int p(z)q(x, z)dz)\delta_x(y)$, where δ_x denotes the Dirac measure at x. In particular, x may with positive probability be its own successor. In any case, the number of repica of the successor y is geometrically distributed with parameter $\alpha_1(y)$. Below we prove that this algorithm creates a Markov chain, derived from an RIW p-sequence, with π as stationary distribution.

Formally we construct an RIW p-sequence $\{(Y^n, W^n)\}$ representing the algorithm in the following way: Let $Y^1, Y^2, \ldots$ be an IID sequence of p-distributed variables. Given that $Y^n = y^n$, let $W_1^n, W_2^n, \ldots$ be independent and geometrically distributed with parameter $\alpha_1(y^n)$. Given that $X^{T_{n-1}} = x, Y^n = y^n, Y^{n+1} = y^{n+1}, Y^{n+2} = y^{n+2}, \ldots$, let $V_1^n, V_2^n, V_3^n, \ldots$ be mutually independent Bernoulli variables with parameters $q(x, y^n), 1 - q(y^n, y^{n+1}), 1 - q(y^n, y^{n+2}), \ldots$. Define $V^n = \sum_{i=1}^{\infty} \prod_{j=1}^{i} V_j^n$ and $W^n = \sum_{i=1}^{V^n} W_i^n$. The sequences $\{V_i^n, i = 1, 2, \ldots\}$ are linked for different values of n by requiring inductively that given

$$V_1^n = V_2^n = \cdots = V_i^n = 1 \tag{41}$$

we have $1 - V_{i+1}^n = V_1^{n+i}$, representing an indicator variable for replacing y^n by

y^{n+i} as a new state of the derived chain. This is consistent, since the Bernoulli parameters for V^n_{i+1} and V_1^{n+i} are respectively $1 - q(y^n, y^{n+i})$ and $q(y^n, y^{n+i})$ under the condition (41).

We want to prove that the chain $\{X^t\}$ derived from $\{(Y^n, W^n\}$ has π as stationary distribution. We denote by $\{X_1^t\}$ the chain derived from $\{(Y^n, V^n\}$. Note that for $i \geq 1$ we have

$P(V^n \geq i+1 | V^n \geq i, Y^n = y^n) = P(V^n_{i+1} = 1 | V^n_1 = \ldots = V^n_i = 1, Y^n = y^n) = 1 - \int p(y)q(y^n, y)dy$

Defining $\alpha_2(y) = \int p(z)q(y,z)dz$ we have that $\{X_1^t\}$ is an RIW chain determined by q and α_2, and in particular it is a GIC (cf. (33), (34) and (37)). Recall that we have chosen q in such a way that this chain has ϕ as stationary distribution.

We claim that $\{X^t\}$ is an RIW chain with transition kernel
$k(x,y) = \alpha_1(x)p(y)q(x,y)$ if $y \neq x$ and

$$P(X^{t+1} = X^t | X^t = x) = 1 - \alpha_1(x)\alpha_2(x) \tag{42}$$

To prove this, we argue as in the proof of theorem 10. Let t be an arbitrary integer. Suppose that

$$\sum_{k=1}^{n-1} W^k + \sum_{j=1}^{i-1} W_j^n < t \leq \sum_{k=1}^{n-1} W^k + \sum_{j=1}^{i} W_j^n \leq \sum_{k=1}^{n} W^k \tag{43}$$

In particular, (43) implies that $X^t = Y^n$. Denote the event described by (43) by E, and put
$s = t - \sum_{k=1}^{n-1} W^k$, $r = t - [\sum_{k=1}^{n-1} W^k + \sum_{j=1}^{i-1} W_j^n]$.

Then $s \geq r > 0$ and $P(X^{t+1} = X^t | X^1 = x^1, \ldots, X^t = x^t, E) = P(W^n \geq s+1 | Y^n = x^t, E) = P(W_i^n \geq r+1 | W_i^n \geq r, Y^n = x^t) + P(W_i^n = r | W_i^n \geq r, Y^n = x^t) P(V^n \geq i+1 | V^n \geq i, Y^n = x^t) = (1 - \alpha_1(x^t)) + \alpha_1(x^t)(1 - \alpha_2(x^t)) = 1 - \alpha_1(x^t)\alpha_2(x^t)$. If $y \neq x^t$, the conditional density at y for X^{t+1} given $X^1 = x^1, \ldots, X^t = x^t, E$ is given by $P(W_i^n = r | W_i^n \geq r, Y^n = x^t)p(y)P(V^n = i | V^n \geq i, Y^n = x^t, Y^{n+i} = y) = \alpha_1(x^t)p(y)q(x^t, y) = \alpha_1(x^t)\alpha_2(x^t)\phi(y|x^t)$, where by definition $\phi(y|x) = p(y)q(x,y)/(\int p(z)q(x,z)dz)$. Hence, (42) is confirmed, and the transition kernel is of the form given in theorem 10. It corresponds to the kernel of an RIW chain determined by q and $\alpha = \alpha_1\alpha_2$.

Note that the $\{Z^r\}$-chains corresponding to the RIW chains $\{X^t\}$ and $\{X_1^t\}$ are identical. Since $\{X_1^t\}$ has ϕ as stationary distribution, the stationary distribution for $\{Z^r\}$ has by theorem 11 density $\chi(y)$ proportional to $\alpha_2(y)\phi(y)$. Since $\alpha_1(y) = c\phi(y)/\pi(y)$, we obtain $\chi(y) \propto \alpha_2(y)\phi(y) \propto \alpha_1(y)\alpha_2(y)\pi(y)$. Hence, using theorem 11 again it follows that π is the stationary distribution for $\{X^t\}$. it is also easy to see that if $\inf_x q(x,y) > 0$ for all y, then the chain is Harris recurrent and hence ergodic.

5 Concluding remarks

In this paper we have treated different simulation algorithms, aimed at exploring a specific target distribution π, based on sampling from a single proposal distribution p. The applicability of such algorithms is limited, due to the difficulty of finding a suitable independent proposal p that approximates π closely enough. There are however some techniques that can be used in order to overcome these difficulties to a certain extent. One possibility is the algorithm described in example 3. In this algorithm, the processing of candidates from p is speeded up by putting them through a "pre-screening" accept - reject step based on another density ϕ, chosen to be less complex than π. Another possibility is to use an adaptive scheme to update the proposal density on the basis of previous iterations of the algorithm.

The construction of adaptive MCMC methods has become an active field of research. Methods based on updating an independent proposal distributioon are discussed in Gilks et al (1998), Gåsemyr, Natvig and Sørensen (2001), Sahu and Zhigljavski (2001) and Gåsemyr (2002).

A natural way of updating p is to choose p from a specific parametric class of distributions and update parameters of p, e. g. the covariance matrix for a multinormal p, on the basis of the history of the chain. Alternatively, Sahu and Zhigljavski (2001) suggest an updating scheme creating an increasingly complex mixture distribution, adding a new convex component whenever the algorithm discovers an area where the old p is unacceptably small compared to π.

Equally important is the question of when to update p. Gilks et al. construct regeneration times for the chain, and show that when p is updated at these times, ergodicity of the chain is preserved under certain conditions. These results are also the basis for the scheme of Sahu and Zhigljavski (2001), and work equally well for the generalization of their algorithm treated in section 2 of the present paper. Alternatively, p may be updated at fixed intervals (Gåsemyr, Natvig and Sørensen (2001), Gåsemyr (2002)). Another possibility is to update p at the times $\tau_1, \tau_2, \ldots$ of corollary 1 (and corollary 2) of the present paper. It may then be proved along the lines of the proof of theorem 1 of Gåsemyr (2002) that the chain will stay π-distributed from time τ_1 onwords.

The adaption methodology is not limited to the MCMC based algorithms. Any adaptive scheme constructed for the OSR chains introduced in section 2, can be applied to rejection sampling, which is a special case, namely OSR$(1/w^*)$. Note also that the OSR$(1/w_*)$ is very similar to importance sampling. It accepts each proposed value at least once, and equation (16) indicates that any adaptive scheme constructed for this algorithm should have comparable effects when transferred to importance sampling. In connection with the OSR algorithm we would also like to mention the possibility to change κ adaptively, in order to approach the optimal value κ_0 of theorem 3.

One task performed by the MCMC algorithms that can not be taken over by importance sampling, is to replace sampling from conditional distributions within a Gibbs sampling framework, when it is very difficult to sample from these conditionals directly. Breaking down a high-dimensional vector into single components or blocks containing a small number of components makes it more realistic to approximate the conditional distributions by members of some parametric class, even if the entire joint distribution is far too complex to allow such an approximation. Also for this kind of application, the adaption methodology can be useful. One possibility, the componentwise adaptive independent chain, (CAIC), is described in Gåsemyr (2002), see also Gåsemyr, Natvig and Nygård (2002).

Acknowledgement

We are very greatful to Professor Bent Natvig for his careful reading of the manuscript and helpful suggestions.

References

Gilks, W. R., Roberts, G. O. and Sahu, S. K.: "Adaptive Markov Chain Monte Carlo through regeneration". J. Amer. Statist. Ass. 93, 1045–1054 (1998).

Gåsemyr, J.: "On an adaptive version of the Metropolis - Hastings alogorithm with independent proposal distribution". Statistical research report, Department of mathematics, University of Oslo (2000).

Gåsemyr, J.: "On an adaptive version of the Metropolis - Hastings alogorithm with independent proposal distribution". To appear in Scand. J. Stat. (2002).

Gåsemyr, J., Natvig, B. and Sørensen, E.: "A comparison of two sequential Metropolis - Hastings algorithms with standard simulation techniques in Bayesian inference in reliability models involving a generalized gamma distribution". Methodol. Comput. Appl. Probab. 3, 51–73 (2001).

Gåsemyr, J., Natvig, B. and Nygård, C. S.: "An application of adaptive independent chain Metropolis - Hastings algorithms in Bayesian hazard rate estimation". Statistical reasearch report, Department of mathematics, University of Oslo (2002).

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

Liu, J. S.: "Metropolized independent sampling with comparison to rejection sampling and importance sampling". Statist. Comput. 6, 113–119 (1996).

Sahu, S. K. and Zhigljavski, A. A.: "Self regenerative Markov Chain Monte Carlo with adaptation". Technical report, University of Southampton (2001).