Markov Chains and De-initialising Processes

by

Gareth O. Roberts* and Jeffrey S. Rosenthal**

(November 1998; last revised July 2000.)

Abstract. We define a notion of de-initialising Markov chains. We prove that to analyse convergence of Markov chains to stationarity, it suffices to analyse convergence of a de-initialising chain. Applications are given to Markov chain Monte Carlo algorithms and to convergence diagnostics.

1. Introduction.

Although Markov chains are routinely used in many probabilistic and algorithmic applications, the complexity of the state space can easily make the analysis of its convergence properties difficult. However, in some cases, it is possible to consider "simpler" processes which contain all the relevant convergence information for the chain of interest, and such that the analysis of the derived process is much more straightforward. Loosely speaking, we shall call such a process *de-initialising* for the chain of interest (although we shall find that there are a number of different natural de-initialising notions).

This paper therefore investigates this notion of de-initialising. A major motivation for this comes from Markov chains induced by various types of Markov chain Monte Carlo (MCMC) algorithms, including the Gibbs sampler and the slice sampler. We shall prove results bounding the total variation distance to stationarity of a Markov chain, in terms of

* Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England Internet: `g.o.roberts@lancaster.ac.uk`. Supported in part by EPSRC of the U.K.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: `jeff@math.toronto.edu`. Supported in part by NSERC of Canada.

the distance to stationarity of an appropriate de-initialising process. We shall also prove a general result which sheds some light on the use of convergence diagnostics for MCMC algorithms.

Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain on a state space $\mathcal{X}$. Let $\{Y_n\}_{n=1}^{\infty}$ be a second chain (not necessarily Markovian), on a state space $\mathcal{Y}$. We shall say that $\{Y_n\}$ is *de-initialising* for $\{X_n\}$ if for each $n \in \mathbf{N}$, conditional on Y_n, we have that X_n is conditionally independent of X_0.

More formally, $\{Y_n\}$ is de-initialising for $\{X_n\}$ if for each $n \geq 1$ (but *not* necessarily for $n = 0$),

$$\mathcal{L}\,(X_n \,|\, X_0, Y_n) = \mathcal{L}\,(X_n \,|\, Y_n) \tag{1}$$

i.e. the conditional distribution of X_n given X_0 and Y_n is required to be a function of Y_n only.

Equation (1) also explains the terminology “de-initialising”. Indeed, in the presence of Y_n, the distribution of X_n no longer depends on its initial value X_0. It is this lack of dependence on initial value which makes X_n be “de-initialised”; and it is the agent Y_n which is performing this “de-initialising”. Such reasoning also makes clear that de-initialising is closely related to Markov chain convergence issues. Indeed, a chain can be said to converge when it completely forgets its initial value.

Remark. An expression like $\mathcal{L}\,(X_n \,|\, X_0, Y_n)$ in (1) is really short-hand for the *regular conditional distribution* of X_n given the sigma-algebra generated by X_0 and Y_n. These conditional distributions are in fact only defined up to a set of probability 0, so all equations such as (1) should be taken as holding with probability 1 only. For a formal definition of conditional probability see e.g. Billingsley (1995, p. 439). We assume throughout that $\mathcal{X}$ is a *standard Borel space* so that these regular conditional distributions always exist; see e.g. Durrett, 1991, pp. 27 and 199).

Intuitively, the distribution of X_n is “completely determined” by the value of Y_n, i.e. once we know Y_n then we know all the history we need to make predictions about X_n. (Note that X_n may not be a deterministic function of Y_n, but it is a “random function” of Y_n in some sense.)

Our notion of de-initialising is somewhat related to the classical statistical notion of *sufficiency* introduced by Fisher (1920), see e.g. Cox and Hinkley (1974) and Lauritzen (1988). Indeed, if we adopt the point of view that X_0 is a statistical parameter of interest, then Y_n being de-initialising for X_n is formally equivalent to Y_n being a sufficient statistic for the parameter X_0, given the datum X_n.

In the case where $Y_n = t_n(X_0, \ldots, X_n)$ is a deterministic function of $X_0, \ldots, X_n$, and where we are working relative to a family of probability distributions $\mathcal{P}$, there are additional notions of sufficiency more closely related to the present paper. For example, Barndorff-Nielsen and Skibinsky (1963) and Skibinsky (1967) introduced the notion of *adequacy* of a sufficient statistic to another random variable; and Lauritzen (1972; 1974; 1988) introduced the notion of *total sufficiency* for sequences of random variables. Our notion of "forward de-initialising" presented in Section 2 herein, together with classical sufficiency relative to some family $\mathcal{P}$, is equivalent to total sufficiency in the sense of Lauritzen. Furthermore, Lauritzen's total sufficiency is equivalent to adequacy of Y_n for the collections $X_{n+1}, \ldots, X_{n+k}$, for all $k \in \mathbf{N}$. In addition, $\{Y_n\}$ is *transitive* for $\{X_n\}$ (Bahadur, 1954; Lauritzen, 1988, p. 29) if Y_n is conditionally independent of the history $X_0, \ldots, X_{n-1}$, given Y_{n-1}. This is quite similar to our definition of de-initialising, except with the conditional independence being for Y_n instead of for X_n.

Whilst these classical notions of sufficiency are clearly related to the concepts considered here, our motivation is very different. Our interest is specifically in summarising the convergence of Markov chains to stationarity in terms of simpler processes. We are not concerned here with families of distributions $\mathcal{P}$ for dependent data. Instead, we concentrate on summarising the distribution of a single Markov chain sequence.

In Section 2, we demonstrate that the total variation distance between the distributions of the de-initialising process at fixed time started at two different initial values bounds the corresponding quantity for the original Markov chain. Thus, the de-initialising process can be used to bound the convergence rate of the chain of interest.

Section 3 gives a number of examples of de-initialising processes, and some applications of the results of Section 2. In Section 4, we clarify the logical relationships between the various notions of de-initialising that we have introduced. Section 5 develops an application of de-initialising processes to diagnosing convergence of Markov chain Monte Carlo algorithms. Section 6 introduces *partial de-initialising*, a notion of de-initialising after a particular stopping time, with an application to the independence sampler.

2. Implications of de-initialising.

The first result of this paper states that the total variation distance to stationary for $\{X_n\}$ is bounded above by that for a de-initialising chain $\{Y_n\}$. This result is analogous to the result of Rosenthal (1992, Proposition 1 (4)) for pseudo-finite chains, but is much more general. To state it cleanly, we shall use the short-hand notation

$$\mathcal{L}(X_n \,|\, X_0 \sim \mu) \;\equiv\; \int \mathcal{L}(X_n \,|\, X_0 = x)\, \mu(dx)\,,$$

i.e. with probabilities given by

$$\mathbf{P}(X_n \in S \,|\, X_0 \sim \mu) \;\equiv\; \int \mathbf{P}(X_n \in S \,|\, X_0 = x)\, \mu(dx)$$

and expectations given by

$$\int f(y)\, \mathbf{P}(X_n \in dy \,|\, X_0 \sim \mu) \;\equiv\; \int\!\!\int f(y)\, \mathbf{P}(X_n \in dy \,|\, X_0 = x)\, \mu(dx)\,. \qquad (2)$$

We then have

Theorem 1. *Let $\{Y_n\}$ be de-initialising for $\{X_n\}$. Then for any two initial distributions μ and μ',*

$$\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) - \mathcal{L}(X_n \,|\, X_0 \sim \mu')\| \;\leq\; \|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - \mathcal{L}(Y_n \,|\, X_0 \sim \mu')\|\,,$$

where $\|\cdots\|$ denotes total variation distance. In particular, if μ and μ' are point masses, then

$$\|\mathcal{L}(X_n \,|\, X_0 = x) - \mathcal{L}(X_n \,|\, X_0 = x')\| \;\leq\; \|\mathcal{L}(Y_n \,|\, X_0 = x) - \mathcal{L}(Y_n \,|\, X_0 = x')\|\,.$$

Proof. Recall that

$$\|\nu - \nu'\| = \sup_S |\nu(S) - \nu'(S)| \tag{3}$$

and also

$$\|\nu - \nu'\| = \sup_{0 \le f \le 1} \left| \int f \, d\nu - \int f \, d\nu' \right| \tag{4}$$

Now, for any measurable set S, we have

$$\begin{aligned}
&|\mathbf{P}(X_n \in S \,|\, X_0 \sim \mu) - \mathbf{P}(X_n \in S \,|\, X_0 \sim \mu')| \\
&= \left| \int \mathbf{P}(X_n \in S \,|\, X_0 = x) \, \mu(dx) - \int \mathbf{P}(X_n \in S \,|\, X_0 = x) \, \mu'(dx) \right| \\
&= \left| \int \int \mathbf{P}(X_n \in S \,|\, X_0 = x, Y_n = y) \, \mathbf{P}(Y_n \in dy \,|\, X_0 = x) \mu(dx) \right. \\
&\quad \left. - \int \int \mathbf{P}(X_n \in S \,|\, X_0 = x, Y_n = y) \, \mathbf{P}(Y_n \in dy \,|\, X_0 = x) \mu'(dx) \right| \\
&= \left| \int \int \mathbf{P}(X_n \in S \,|\, Y_n = y) \, \mathbf{P}(Y_n \in dy \,|\, X_0 = x) \mu(dx) \right. \\
&\quad \left. - \int \int \mathbf{P}(X_n \in S \,|\, Y_n = y) \, \mathbf{P}(Y_n \in dy \,|\, X_0 = x) \mu'(dx) \right| \\
&= \left| \int \int f(y) \, \mathbf{P}(Y_n \in dy \,|\, X_0 = x) \mu(dx) - \int \int f(y) \, \mathbf{P}(Y_n \in dy \,|\, X_0 = x) \mu'(dx) \right|
\end{aligned}$$

where $f(y) = \mathbf{P}(X_n \in S \,|\, Y_n = y)$, so that $0 \le f(y) \le 1$. By (2), this is a difference of expectations. Hence, from (4), we have

$$\left| \mathbf{P}(X_n \in S \,|\, X_0 \sim \mu) - \mathbf{P}(X_n \in S \,|\, X_0 \sim \mu') \right| \le \|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - \mathcal{L}(Y_n \,|\, X_0 \sim \mu')\| \,.$$

Since this is true for any S, the result now follows from (3). ■

For example, we have the following.

Lemma 2. *If there are deterministic measurable functions $f_1, f_2, \ldots$ such that $X_n = f_n(Y_n)$, then $\{Y_n\}$ is de-initialising for $\{X_n\}$.*

Proof. Indeed, in this case

$$\mathcal{L}\,(X_n \,|\, X_0, \ldots, X_{n-1}, Y_n) \;=\; \delta_{f_n(Y_n)}(\cdot)\,,$$

which gives the result. ∎

We shall call Markov chains $\{X_n\}$ and $\{Y_n\}$ *co-de-initialising* if $\{Y_n\}$ is de-initialising for $\{X_n\}$, and also $\{X_n\}$ is de-initialising for $\{Y_n\}$. We immediately have

Corollary 3. *If $\{X_n\}$ and $\{Y_n\}$ are co-de-initialising Markov chains, then for $n \geq 1$,*

$$\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) - \mathcal{L}(X_n \,|\, X_0 \sim \mu')\| \;=\; \|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - \mathcal{L}(Y_n \,|\, X_0 \sim \mu')\|\,.$$

Say that $\{Y_n\}$ is *functionally de-initialising* for $\{X_n\}$ if it is de-initialising and Markovian, and also $Y_n = h_n(X_n)$ for some deterministic measurable functions h_n. Say that $\{Y_n\}$ is *homogeneously functionally de-initialising* for $\{X_n\}$ if in addition we can choose the same function h_n for each n, i.e. $\{Y_n\}$ is Markovian and de-initialising for $\{X_n\}$, and also $Y_n = f(X_n)$ for each n.

Now, if the Markov chain $\{X_n\}$ has a stationary distribution $\pi(\cdot)$, and if $\{Y_n\}$ is homogeneously functionally de-initialising for $\{X_n\}$, then $\{Y_n\}$ will have stationary distribution $f_*\pi$ defined by $(f_*\pi)(S) = \pi\left(f^{-1}(S)\right)$. Furthermore, by stationarity we will have $\mathcal{L}(X_n \,|\, X_0 \sim \pi) = \pi(\cdot)$ and $\mathcal{L}(Y_n \,|\, X_0 \sim \pi) = f_*\pi(\cdot)$. It follows from Lemma 2 that $\{X_n\}$ and $\{Y_n\}$ are co-de-initialising. Hence, setting $\mu' = \pi$ in Corollary 3, we obtain

Corollary 4. *Let $\{X_n\}$ be a Markov chain with stationary distribution $\pi(\cdot)$. Let $Y_n = f(X_n)$ for some measurable function $f : \mathcal{X} \to \mathcal{Y}$, and suppose that $\{Y_n\}$ is Markovian and is de-initialising for $\{X_n\}$. Then*

$$\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) - \pi(\cdot)\| \;=\; \|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - (f_*\pi)(\cdot)\|\,.$$

That is, we can obtain bounds on convergence rate of a chain to its stationary distribution, in terms of corresponding bounds for a homogeneously functionally de-initialising chain.

Remark. Even if $\{Y_n\} = \{f(X_n)\}$ is not Markovian, it will still be a *stationary process* (in the sense of e.g. p. 129 of Bhattacharya and Waymire, 1990), provided that $X_0 \sim \pi(\cdot)$, and one can still consider bounds on $\|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - (f_* \pi)(\cdot)\|$. However, in this case the terminology and notation becomes more cumbersome, and the results become less interesting, so we do not pursue that here.

We shall also consider certain other notions of de-initialising. Say that $\{Y_n\}$ is *backward de-initialising* for $\{X_n\}$ if for $n \geq 1$,

$$\mathcal{L}\left(X_n \,|\, X_0, X_1, \ldots, X_{n-1}, Y_n\right) = \mathcal{L}\left(X_n \,|\, Y_n\right) \,, \tag{5}$$

i.e. this distribution conditional on the entire history of $\{X_n\}$ is also a function of Y_n only. (Many, but not all, of our examples of de-initialising Markov chains are also backward de-initialising.) Say that $\{Y_n\}$ is *forward de-initialising* for $\{X_n\}$ if

$$\mathcal{L}(X_{n+1}, X_{n+2}, \ldots \,|\, X_0, \ldots, X_n, Y_n) = \mathcal{L}(X_{n+1}, X_{n+2}, \ldots \,|\, Y_n)\,.$$

Say that $\{Y_n\}$ is *totally de-initialising* for $\{X_n\}$ if

$$\mathcal{L}(X_n \,|\, X_0, \ldots, X_{n-1}, Y_n, X_{n+1}, X_{n+2}, \ldots) = \mathcal{L}(X_n \,|\, Y_n)\,.$$

(Obviously, total de-initialising implies both backward and forward de-initialising.)

The logical implications of these various notions of de-initialising are explored in Section 4 herein.

We shall also use the following.

Proposition 5. *Let $\{X_n\}$ be a Markov chain with transition probabilities $P(x, \cdot)$. If we can write $P(x, \cdot) = R(h(x), \cdot)$ for some measurable function $h : \mathcal{X} \to \mathcal{Y}$ and some probability distributions $R(y, \cdot)$ on $\mathcal{X}$, then $\{h(X_{n-1})\}$ is backward de-initialising for $\{X_n\}$. Furthermore $\{h(X_n)\}$ is forward de-initialising for $\{X_n\}$.*

Proof. We see that

$$\mathcal{L}(X_n \,|\, X_0, \ldots, X_{n-1}, h(X_{n-1})) = R(h(X_{n-1}), \cdot)$$

and hence equals $\mathcal{L}(X_n \,|\, h(X_{n-1}))$. For the second statement, we similarly see that

$$\mathcal{L}(X_{n+1}, X_{n+2}, \ldots \,|\, X_0, \ldots, X_n, h(X_n)) = \mathcal{L}(X_{n+1}, X_{n+2}, \ldots \,|\, h(X_n)) \,. \qquad \blacksquare$$

The following result illustrates an interesting difference, in general, between the notion of de-initialising and the classical notion of sufficiency.

Proposition 6. *Even if $\{Y_n\}$ is de-initialising for $\{X_n\}$, it may be that there is some sequence $\{Z_n\}$ of random variables such that $\{(Y_n, Z_n)\}$ is not de-initialising for $\{X_n\}$.*

Proof. Let X_0, X_n, and Z_n be any three random variables which are pairwise independent but are not independent. Let Y_n be identically zero (say). Then by pairwise independence, we have $\mathcal{L}(X_n \,|\, X_0, Y_n) = \mathcal{L}(X_n) = \mathcal{L}(X_n \,|\, Y_n)$, so that $\{Y_n\}$ is indeed de-initialising for $\{X_n\}$. Now, since X_n is independent of Z_n, $\mathcal{L}(X_n \,|\, Y_n, Z_n) = \mathcal{L}(X_n)$. On the other hand, since X_n is *not* independent of the pair (X_0, Z_n), therefore $\mathcal{L}(X_n \,|\, X_0, Y_n, Z_n) = \mathcal{L}(X_n \,|\, X_0, Z_n) \neq \mathcal{L}(X_n)$. It follows from these two observations that $\mathcal{L}(X_n \,|\, X_0, Y_n, Z_n) \neq \mathcal{L}(X_n \,|\, Y_n, Z_n)$. Hence, $\{(Y_n, Z_n)\}$ is not de-initialising for $\{X_n\}$. $\blacksquare$

On the other hand, if Z_n is required to be a function of $X_0, \ldots, X_{n-1}$, and if $\{Y_n\}$ is backward de-initialising for $\{X_n\}$, then clearly so is $\{(Y_n, Z_n)\}$. This corresponds closely to the situation for classical sufficiency, since there the "statistics" are required to be functions of the data.

3. Examples.

Examples of de-initialising Markov chains include:

1. *Deterministic functions.* If there are deterministic functions $f_1, f_2, \ldots$ such that $X_n = f_n(Y_n)$, then by Lemma 2, $\{Y_n\}$ is de-initialising for $\{X_n\}$. Thus, Theorem 1 states that to study convergence of $\{f_n(Y_n)\}$, it suffices to study convergence of $\{Y_n\}$. In fact, here $\{Y_n\}$ is totally de-initialising for $\{X_n\}$.

2. *Delayed chain.* If $Y_n = X_{\max(0,\, n-k)}$ for some fixed $k \in \mathbf{N}$, then again $\{Y_n\}$ is de-initialising for $\{X_n\}$. In fact, for $k \geq n$, we have $\mathcal{L}(X_n \,|\, X_0, Y_n) = P^k(Y_n, \cdot)$. Indeed, in this case Theorem 1 corresponds to the well-known statement that

$$\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) - \mathcal{L}(X_n \,|\, X_0 \sim \mu')\| \;\leq\; \|\mathcal{L}(X_{n-k} \,|\, X_0 \sim \mu) - \mathcal{L}(X_{n-k} \,|\, X_0 \sim \mu')\|\,,$$

i.e. that total variation distance cannot increase with time. (Note that this example is *not* forward de-initialising. Furthermore, it is not backward de-initialising either unless $k = 1$.)

3. *Two-variable data augmentation* (Tanner and Wong, 1984; Gelfand and Smith, 1990; Rosenthal, 1993). Here there is some target distribution $\pi(\cdot)$ on a product space $\mathcal{X} \times \mathcal{Y}$, with regular conditional distributions $\pi_{X|Y}(dx|y)$ and $\pi_{Y|X}(dy|x)$. A Markov chain $\{(X_n, Y_n)\}$ on $\mathcal{X} \times \mathcal{Y}$ is defined by alternately choosing $Y_{n+1} \sim \pi_{Y|X}(dy \,|\, X_n)$ and $X_{n+1} \sim \pi_{X|Y}(dx \,|\, Y_{n+1})$, for $n = 0, 1, 2, \ldots$. In this case we clearly have

$$\mathcal{L}\left(X_n \,|\, X_0, \ldots, X_{n-1}, Y_n\right) \;=\; \pi_{X|Y}(dx \,|\, Y_n)\,,$$

so that again $\{Y_n\}$ is de-initialising (and backward de-initialising) for $\{X_n\}$ and also for $\{(X_n, Y_n)\}$. Hence, to study convergence of $\{(X_n, Y_n)\}$ it suffices to study convergence of $\{Y_n\}$, a fact used in Rosenthal (1993). In this example, it is also true that $\{X_n\}$ is forward de-initialising for $\{(X_n, Y_n)\}$.

In fact, we can write

$$\mathcal{L}\left((X_n, Y_n) \,|\, (X_{n-1}, Y_{n-1})\right) = R(X_{n-1}, \cdot)$$

for appropriate choice of $R(\cdot,\cdot)$. Hence, setting $h\left((X_n, Y_n)\right) = X_n$, we can apply Proposition 5. We conclude that $\{X_n\}$ is forward de-initialising for $\{(X_n, Y_n)\}$, and that $\{X_{n-1}\}$ is backward de-initialising for $\{(X_n, Y_n)\}$.

One example of the use of this property is in the Bayesian analysis of finite mixtures (see for example Diebolt and Robert, 1994). For these models, the space of missing data is finite, and therefore the Markov chain consisting of just the missing data is uniformly ergodic. Furthermore, the missing data is co-de-initialising for the entire chain. Consequently, by Corollary 3, the data augmentation algorithm is also uniformly ergodic. This observation was termed the *duality principle* by Diebolt and Robert (1994).

4. *Pseudo-finite Markov chains*, or *Markov chains of finite rank* (Hoekstra and Steutel, 1984; Runnenburg and Steutel, 1962; Rosenthal, 1992). Here

$$\mathbf{P}\left(X_{n+1} \in \cdot \,|\, X_n = x\right) \;=\; \sum_{j=1}^{m} f_j(x)\, Q_j(\cdot)$$

for some finite number $m \in \mathbf{N}$, where $f_i : \mathcal{X} \to [0,1]$ are deterministic functions summing to 1, and $Q_j(\cdot)$ are fixed probability measures on $\mathcal{X}$. In this case, we can define a second Markov chain $\{Y_n\}$ on $\mathcal{Y} = \{1, 2, \ldots, m\}$ by $\mathbf{P}(Y_1 = j) = \mathbf{E}\left(f_j(X_0)\right)$, and

$$\mathbf{P}\left(Y_{n+1} = j \,|\, Y_n = i\right) \;=\; \mathbf{E}_{Q_i}(f_j)\,.$$

Then $\{Y_{n-1}\}$ is de-initialising (in fact, backward de-initialising) for $\{X_n\}$. Indeed, here

$$\mathcal{L}\left(X_n \,|\, X_0, \ldots, X_{n-1}, Y_{n-1}\right) \;=\; Q_{Y_{n-1}}(\cdot)\,.$$

Intuitively, Y_n keeps track of "which of the Q_i distributions the variable X_n is currently in". The result of Theorem 1, for the special case of pseudo-finite chains, was presented in Rosenthal (1992, Proposition 1 (4)). Furthermore, here $\{Y_n\}$ is forward de-initialising for $\{X_n\}$. Thus, interestingly, $\{Y_n\}$ satisfies the conclusions, but not the hypotheses, of Proposition 5.

5. *Slice samplers* (Swendsen and Wang, 1987; Besag and Green, 1993; Higdon, 1996; Damien et al., 1997; Mira and Tierney, 1997; Fishman, 1996; Neal, 1997; Roberts

and Rosenthal, 1997b, 1999). Let $f : \mathcal{X} \to [0, \infty)$ be a non-negative L^1 function, where $\mathcal{X} \subseteq \mathbf{R}^d$ and $\int f \, d\lambda_d > 0$. The slice sampler is defined as follows. Given X_n, we first choose $Z_{n+1} \in \mathbf{R}$ uniformly from the interval $[0, f(X_n)]$. We then choose X_{n+1} uniformly from the set $\{x \in \mathcal{X}; \ f(x) \geq Z_{n+1}\}$. Then the law of X_n converges (as $n \to \infty$) to the distribution on $\mathcal{X}$ having density proportional to f. (In fact, it is easily checked that the marginal chain $\{X_n\}$ is *reversible* with respect to this distribution.) Such samplers are a common way of approximately sampling from a high-dimensional density. Corresponding to a slice sampler is a second Markov chain $\{Y_n\}$ on $\mathcal{Y} = [0, \infty)$, defined by $Y_n = f(X_n)$. Note that X_n is *not* in general a deterministic function of Y_n since (for $d \geq 2$, say) the function f will not be invertible. However, it is still true that for $n \geq 1$ we have

$$\mathcal{L}\left(X_n \,|\, X_0, \ldots, X_{n-1}, Y_n\right) \ = \ \mathbf{Unif}\left(L(Y_n)\right) \ ,$$

where $L(y) = \{x \in \mathcal{X}; \ f(x) \geq y\}$ and $\mathbf{Unif}(R)$ is the uniform distribution (i.e., normalised Lebesgue measure) on the region R. Hence, again, $\{Y_n\}$ is de-initialising for $\{X_n\}$. In fact, here $\{Y_n\}$ is itself Markovian, so that $\{Y_n\}$ is also backward de-initialising, forward de-initialising, totally de-initialising, and functionally de-initialising; and $\{X_n\}$ and $\{Y_n\}$ are co-de-initialising. These facts were used implicitly in the detailed study of slice samplers by Roberts and Rosenthal (1997b, 1999).

6. *The Gibbs sampler* (Geman and Geman, 1984; Gelfand and Smith, 1990). Let $\pi(\cdot)$ be a probability distribution on $\mathbf{R}^k$. The Gibbs sampler for π is a Markov chain with transition probabilities given by

$$P(\mathbf{x}, d\mathbf{y}) = \prod_{i=1}^{k} \pi(dy^i | \mathbf{y}^{j<i}, \mathbf{x}^{j>i}) \,.$$

In this case, we see that x^1 does not appear in the formula for $P(\mathbf{x}, d\mathbf{y})$. Let $\mathbf{X}_{n-1}^{-i} \equiv (X_n^1, \ldots, X_n^{i-1}, X_n^{i+1}, \ldots, X_n^k)$. Then by Proposition 5, we see that $\{\mathbf{X}_{n-1}^{-k}\}$ is backward de-initialising for $\{\mathbf{X}_n\}$ (a fact used in Rosenthal, 1995, Lemma 7), and $\{\mathbf{X}_n^{-1}\}$ is forward de-initialising for $\mathbf{X}_n$. The special case $k = 2$ corresponds to data augmentation as in Example 3 above.

Remark. Ideas related to de-initialising also arise in the study of quasi-stationarity. For example, let $\{X_s\}_{s\geq 0}$ be a continuous-time Markov process. If $t > 0$ is fixed, and τ is a stopping time, then to prove weak convergence of $\mathcal{L}\left(\{X_s\}_{0\leq s\leq t} \,|\, \tau > T\right)$ to a limiting distribution, as $T \to \infty$, it suffices to prove convergence of $\mathcal{L}\left(X_t \,|\, \tau > T\right)$; see for example Jacka and Roberts (1997). In the present context, this translates as saying that X_t is de-initialising for $\{X_s\}_{0\leq s\leq t}$ with regards to the event $\{\tau > T\}$, whenever $T \geq t$.

Finally, we note that good examples of the conditional independencies implicit in our notions of de-initialising can be written in terms of directed graphical models (see for example Lauritzen, 1996; Whittaker, 1990). We give three examples to illustrate this.

Backward de-initialising is implied by the following graphical model, which describes the conditional independence structure in e.g. Example 3 (data augmentation) above:

Figure 1. A graphical example of backward de-initialising.

Backward de-initialising is implied by the following graphical model, which describes the conditional independence structure in e.g. Example 4 (pseudo-finite chain) above, or Example 6 (the Gibbs sampler) above with $Y_n = X_n^{-1}$.

Figure 2. A graphical example of forward de-initialising.

Finally, total de-initialising would be implied by the following graph, which describes the dependencies in Example 5 (the slice sampler) above, and which also appears naturally in the study of hidden Markov models (see e.g. Elliot et al., 1995):

Figure 3. A graphical example of total de-initialising.

Note that the conditional independencies described in the above graphical models are not *required* by our various notions of de-initialising.

4. Relationships between different notions of de-initialising.

We note that if $\{Y_n\}$ is backward de-initialising for $\{X_n\}$, then $\{Y_{n-1}\}$ is automatically *one-step forward de-initialising* for $\{X_n\}$, meaning that

$$\mathcal{L}(X_{n+1}\,|\,X_0,\ldots,X_n,Y_n) = \mathcal{L}(X_{n+1}\,|\,Y_n)$$

for $n \geq 1$. Indeed, this follows immediately by substituting Y_{n-1} for Y_n in equation (5).

However, this one-step forward de-initialising does *not* imply forward de-initialising as we have defined it. For example, let X_0, X_2, Y_0 be three random variables which are pairwise independent but not three-way independent, and let $X_1, X_3, \ldots, Y_1, Y_2, \ldots$ all be independent of everything. Then $\{X_n\}$ is Markovian (in fact, an independent sequence), but is *not* Markovian if we first condition on Y_0. We then have that $\mathcal{L}(X_1\,|\,X_0, Y_0) = \mathcal{L}(X_1\,|\,Y_0)$, so that $\{Y_n\}$ is indeed one-step forward de-initialising for $\{X_n\}$. On the other hand, $\mathcal{L}(X_2\,|\,Y_0) = \mathcal{L}(X_2)$, but $\mathcal{L}(X_2\,|\,X_0, Y_0) \neq \mathcal{L}(X_2)$. Hence, $\{Y_n\}$ is not forward de-initialising for $\{X_n\}$.

We also note that if $\{Y_n\}$ is functionally de-initialising for $\{X_n\}$, then by Lemma 2 above, $\{X_n\}$ and $\{Y_n\}$ are *co-de-initialising*, so that Corollary 3 applies. It also follows that $\{Y_n\}$ is automatically forward de-initialising for $\{X_n\}$ as well; this is seen by a direct application of the Markov property for $\{X_n\}$.

We summarise the logical relationships between our various notions of de-initialising in Figure 4.

Figure 4. Logical relationships between different notions of de-initialising.

5. Application to diagnosis of convergence.

In many Markov chain Monte Carlo (MCMC) applications, the user attempts to diagnose convergence of a Markov chain $\{X_n\}_{n=0}^{N}$ to its stationary distribution $\pi(\cdot)$, by monitoring a low-dimensional functional. (See for example Gelfand and Smith, 1990; Cowles and Carlin, 1996; Brooks and Roberts, 1998.)

For example, perhaps the user monitors the values $h(X_n)$ for some function $h : \mathcal{X} \to \mathbf{R}$. A quantity often computed is the *empirical lag-k autocorrelation*, defined for $k \in \mathbf{N}$ by

$$EAC_{h,k} \;=\; \frac{\sum_{i=1}^{N-k}(h(X_i) - m_h)(h(X_{i+k}) - m_h)}{(N-k+1)\; v_h}\,,$$

where $m_h = \frac{1}{N}\sum_{i=1}^{N} h(X_i)$ and $v_h = \frac{1}{N-1}\sum_{i=1}^{N}\left(h(X_i) - m_h\right)^2$ are the empirical mean and variance. $EAC_{h,k}$ is thus an estimator of the true stationary autocorrelation

$$AC_{h,k} \;=\; \operatorname{corr}\bigl(h(X_0), h(X_k)\bigr)\,.$$

under the assumption of stationarity (i.e., with $X_0 \sim \pi$).

Now, if $EAC_{h,k}$ is, say, rather large for $k < 40$ and very small for $k \geq 40$, then one is tempted to conclude that the Markov chain converges to stationarity after approximately 40 iterations.

One difficulty with such an approach is that it is far from clear how to choose the function h for which to compute autocorrelations. It is often the case that for certain choices of h the autocorrelations will be small, while for other choices of h they will be large (see for example Roberts and Rosenthal, 1999). In such cases, one is really interested in the *maximal lag-k autocorrelation* defined by

$$\gamma_k \;=\; \sup_h \big|AC_{h,k}\big|\,, \tag{6}$$

where the supremum is taken over all choices of nonconstant functions $h : \mathcal{X} \to \mathbf{R}$ having finite variance under π. However, in practice one is typically forced to approximate γ_k by the maximum of $AC_{h,k}$ (or, even, of $EAC_{h,k}$) for a certain finite number of choices of h, and it is not clear how such choices of h are to be made.

The notion of de-initialising can assist with such choices. In particular, we have the following.

Theorem 7. *Let $\{X_n\}$ be a Markov chain which is reversible with respect to a stationary distribution π. Suppose that for some nonconstant function $f : \mathcal{X} \to \mathcal{Y}$, setting $Y_n = f(X_n)$, we have that $\{Y_n\}$ is Markovian and is a de-initialising chain for $\{X_n\}$. Then with γ_k as in (6), we have that*

$$\gamma_k \;=\; \sup_g \big|AC_{g\circ f,k}\big|\,.$$

In words, this theorem says that the supremum in (6) is achieved somewhere on the set of functions of the form $h = g \circ f$, i.e. on a choice of h which depends on x only through $f(x)$ (alternatively, on a choice of h which is $\sigma(f)$-measurable).

In practice, this means that, when choosing functions h to compute autocorrelations, it suffices to restrict attention to those functions which depend only on the de-initialising chain $\{Y_n\}$. For example, if Y_n consists of the first few coordinates of X_n, then the functions h need depend only on those same first few coordinates of X_n.

To prove Theorem 7, we require the following two well-known propositions. To state them, let $L_b(\pi)$ be the set of all probability measures μ such that $\frac{d\mu}{d\pi}$ is an essentially-bounded function. Also let $\langle\cdot,\cdot\rangle$ be the usual $L^2(\pi)$ inner product, i.e. $\langle f, g\rangle = \int \pi(dx) f(x) g(x)$ for $f, g : \mathcal{X} \to \mathbf{R}$, and $\|f\| = \langle f, f\rangle^{1/2}$. Finally, let $\|P_0\| = \sup\limits_{f\in L_0^2(\pi),\ \|f\|=1} \|P_0 f\|$ be the $L^2(\pi)$ operator norm of the operator P_0 defined by

$$(P_0 h)(x) \;=\; \mathbf{E}\big(h(X_1)\,|\,X_0 = x\big)\,, \qquad h \in L_0^2(\pi)\,,$$

where

$$L_0^2(\pi) = \left\{ h : \mathcal{X} \to \mathbf{R} ; \int \pi(dy)h(y) = 0, \int \pi(dy)h^2(y) < \infty \right\} .$$

Proposition 8. *Let $\{X_n\}$ be a Markov chain on a state space $\mathcal{X}$, which is reversible with respect to a stationary distribution μ. Let γ_k be as in (6). Then*

$$\gamma_k \ = \ \|P_0\|^k .$$

Proof. By shifting and rescaling as necessary, it suffices in the definition of γ_k to restrict attention to functions h in the collection

$$S \ = \ \left\{ h : \mathcal{X} \to \mathbf{R}; \int \pi(dx)h(x) = 0, \int \pi(dx)h^2(x) = 1 \right\} .$$

For $h \in S$, with $X_0 \sim \pi$, we compute that

$$\begin{aligned} \operatorname{corr}\Big(h(X_0),\ h(X_k)\Big) \ &= \ \mathbf{E}\Big(h(X_0)\,h(X_k)\Big) \\ &= \ \int\int \pi(dx_0)h(x_0)P(x_0, dx_k)h(x_k) \\ &= \ \int \pi(dx_0)h(x_0)(P_0^k h)(x_0) \\ &= \ \langle h,\ P_0^k h \rangle . \end{aligned}$$

(For similar reasoning see e.g. Amit, 1991.)

Hence, using self-adjointness and Lemma A1 from the Appendix, we conclude that

$$\gamma_k \ = \ \sup_h \left| \operatorname{corr}\Big(h(X_0),\ h(X_k)\Big) \right| \ = \ \sup_{h \in S} \left| \operatorname{corr}\Big(h(X_0),\ h(X_k)\Big) \right|$$

$$= \ \sup_{h \in S} |\langle h, P_0^k h \rangle| \ = \ \sup_{h \in S} \|P_0^k h\| \ = \ \|P_0^k\| \ = \ \|P_0\|^k ,$$

as claimed. ∎

Remarks.

1. This proof makes use of the technical result Lemma A1. However, if k is even, then we can instead write
$$\operatorname{corr}\Big(h(X_0),\ h(X_k)\Big) \;=\; \langle P_0^{k/2}h,\ P_0^{k/2}h\rangle \;=\; \|P_0^{k/2}h\|^2\,,$$
and the result then follows easily without requiring Lemma A1. (This also shows that if k is even then $\operatorname{corr}(h(X_0), h(X_k)) \geq 0$.)
2. Similarly, if the spectrum of P_0 consists only of pure eigenvalues $\{\lambda_i\}$ with corresponding orthonormal eigenvectors $\{e_i\}$, then we can decompose $h = \sum_i a_i e_i$, so that $P_0^k h = \sum_i a_i \lambda_i^k e_i$. In this case $\langle h, h\rangle = \sum_i a_i^2$ and $\big|\langle h, P_0^k h\rangle\big| = \big|\sum_i a_i^2 \lambda_i^k\big|$, while $\|P_0\| = \sup_i |\lambda_i|$, so that Proposition 8 follows easily. However, in general P_0 may have continuous spectrum, so that the Spectral Theorem is required to make this approach rigorous.
3. For a related but different result, see Lemma 2.3 of Liu, Wong and Kong (1994).

Proposition 9. *Let $\{X_n\}$ be a Markov chain on a state space $\mathcal{X}$, which is reversible with respect to a stationary distribution μ. Let γ_k be as in (6). Then*
$$\frac{1}{k}\log\gamma_k \;=\; \sup_{\mu\in L_b(\pi)}\ \lim_{n\to\infty}\frac{1}{n}\log\|\mathcal{L}(X_n\,|\,X_0\sim\mu)\;-\;\pi(\cdot)\|\,.$$
(We allow for the special case when both sides equal $-\infty$.)

Proof. We have (see e.g. Theorem 2 of Roberts and Rosenthal, 1997a) that
$$\sup_{\mu\in L_b(\pi)}\ \lim_{n\to\infty}\frac{1}{n}\log\|\mathcal{L}(X_n\,|\,X_0\sim\mu)\;-\;\pi(\cdot)\| \;=\; \log\|P_0\|\,,$$
with P_0 as above. But from Proposition 8, we have that $\gamma_k = \|P_0\|^k$, or that $(\gamma_k)^{1/k} = \|P_0\|$. The result follows by taking logs. ∎

Proof of Theorem 7. We note that $\sup_g \big|AC_{g\circ f,k}\big|$ is the maximal lag-k autocorrelation for the chain $\{Y_n\}$. Hence, from Proposition 9,
$$\frac{1}{k}\log\left(\sup_g\big|AC_{g\circ f,k}\big|\right) \;=\; \sup_{\nu\in L_b(f_*\pi)}\ \lim_{n\to\infty}\frac{1}{n}\log\|\mathcal{L}(Y_n\,|\,Y_0\sim\nu)\;-\;(f_*\pi)(\cdot)\|\,.$$

Now, if $\mu = \mathcal{L}(X_0)$ and $\nu = \mathcal{L}(Y_0 \,|\, X_0 \sim \mu) = f_*\mu$, then for $(f_*\pi)$-a.e. y,

$$\frac{d\nu}{d(f_*\pi)}(y) \;=\; \int_{f^{-1}(y)} \frac{d\mu}{d\pi}(x)\,\rho(dx)\,,$$

where $\rho(\cdot) = \mathcal{L}(X_0 \,|\, Y_0 = y)$ is the conditional distribution of X_0 conditional on being in the set $f^{-1}(y)$. Informally, $\frac{d\nu}{d(f_*\pi)}(y)$ is a "weighted average" of $\frac{d\mu}{d\pi}(x)$ over $x \in f^{-1}(y)$. Hence, if $\frac{d\mu}{d\pi} \le M$ then $\frac{d\nu}{d(f_*\pi)} \le M$. Therefore, $\mathcal{L}(Y_0 \,|\, X_0 \sim \mu) \in L_b(f_*\pi)$ whenever $\mathcal{L}(X_0) \in L_b(\pi)$.

We conclude that

$$\frac{1}{k}\log\left(\sup_g \left|AC_{g\circ f,k}\right|\right) \;\ge\; \sup_{\mu\in L_b(\pi)} \lim_{n\to\infty} \frac{1}{n}\log\|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) \;-\; (f_*\pi)(\cdot)\|\,.$$

But then from Corollary 4, it follows that

$$\frac{1}{k}\log\left(\sup_g \left|AC_{g\circ f,k}\right|\right) \;\ge\; \sup_{\mu\in L_b(\pi)} \lim_{n\to\infty} \frac{1}{n}\log\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) \;-\; \pi(\cdot)\|\,.$$

Hence, from Proposition 9, we have

$$\frac{1}{k}\log\left(\sup_g \left|AC_{g\circ f,k}\right|\right) \;\ge\; \frac{1}{k}\log\left(\sup_h \left|AC_{h,k}\right|\right)\,.$$

We conclude that

$$\sup_g \left|AC_{g\circ f,k}\right| \;\ge\; \sup_h \left|AC_{h,k}\right|,$$

On the other hand, we clearly have

$$\sup_g \left|AC_{g\circ f,k}\right| \;\le\; \sup_h \left|AC_{h,k}\right|,$$

so it follows that

$$\sup_g \left|AC_{g\circ f,k}\right| \;=\; \sup_h \left|AC_{h,k}\right|,$$

as claimed. ∎

As a specific application of Theorem 7, consider the slice sampler of Example 5 above. In Roberts and Rosenthal (1999), the slice sampler was examined for specific choices of the function f, and autocorrelations for a number of different functions h were analysed. In light of Theorem 7, the only autocorrelation functions that needed to be analysed were those of the form $h = g \circ f$, i.e. those which depended only on the values $f(X_n)$.

6. Partial de-initialising.

Say that $\{Y_n\}$ is *partially de-initialising for* $\{X_n\}$ *on* $\{C_n\}$, if there are events $\{C_n\}$ such that

$$\mathbf{P}\left(X_n \in A \,|\, X_0, Y_n\right) = \mathbf{P}\left(X_n \in A \,|\, Y_n\right) \qquad \text{on } C_n \text{ (w.p. 1)}\,.$$

Theorem 10. *Let $\{Y_n\}$ be partially de-initialising for $\{X_n\}$ on $\{C_n\}$. Then for any initial distributions μ and μ',*

$$\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) - \mathcal{L}(X_n \,|\, X_0 \sim \mu')\| \;\leq\; \|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - \mathcal{L}(Y_n \,|\, X_0 \sim \mu')\| + \mathbf{P}(C_n^C)\,,$$

Proof. We have that

$$|\mathbf{P}(X_n \in S \,|\, X_0 \sim \mu) - \mathbf{P}(X_n \in S \,|\, X_0 \sim \mu')|$$

$$\left|\int \mathbf{P}(X_n \in S \,|\, X_0 = x)\mu(dx) - \int \mathbf{P}(X_n \in S \,|\, X_0 = x)\mu'(dx)\right|$$

$$= \left|\int\int_{C_n \cup C_n^C} \mathbf{P}(X_n \in S \,|\, X_0 = x, Y_n = y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu(dx)\right.$$

$$\left. - \int\int_{C_n \cup C_n^C} \mathbf{P}(X_n \in S \,|\, X_0 = x, Y_n = y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu'(dx)\right|$$

$$\leq \left|\int\int_{C_n} \mathbf{P}(X_n \in S \,|\, Y_n = y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu(dx)\right.$$

$$\left. - \int\int_{C_n} \mathbf{P}(X_n \in S \,|\, Y_n = y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu'(dx)\right|$$

$$+ \left|\int\int_{C_n^C} \mathbf{P}(X_n \in S \,|\, Y_n = y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu(dx)\right.$$

$$\left. - \int\int_{C_n^C} \mathbf{P}(X_n \in S \,|\, Y_n = y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu'(dx)\right|$$

$$\leq \left|\int\int f(y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu(dx) - \int\int f(y)\, \mathbf{P}(Y_n \in dy \,|\, X_0 = x)\mu'(dx)\right| + \mathbf{P}(C_n^C)$$

where $f(y) = \mathbf{1}_{C_n}(y)\,\mathbf{P}(X_n \in S \,|\, Y_n = y)$, so that $0 \le f(y) \le 1$. The result now follows just like for Theorem 1. ∎

One special case is $C_n = \{\tau \le n\}$, where τ is a stopping time for $\{X_n\}$. [In fact, usual de-initialising corresponds to $C_0 = \emptyset$ and $C_n = \Omega$ for $n \ge 1$, i.e. $C_n = \{\tau \le n\}$ where $\tau \equiv 1$.]

For example, consider the *independence sampler*, which is defined in terms of a target density π and a proposal density q. Given X_n, it chooses (conditionally independently) $Z_{n+1} \sim q(z)\,dz$. It then either "accepts" Z_{n+1} (i.e., sets $X_{n+1} = Z_{n+1}$) with probability $\min\big(1,\ \pi(Z_{n+1})q(X_n) \,/\, \pi(X_n)q(Z_{n+1})\big)$, or else "rejects" Z_{n+1} (i.e. sets $X_{n+1} = X_n$) with the remaining probability.

Given an independence sampler $\{X_n\}$, let τ be the first time that the sampler accepts a proposed move, and let $Y_n = q(X_n)/\pi(X_n)$. Then it is straightforward to see that $\{Y_n\}$ is partially de-initialising for $\{X_n\}$ on $\{\tau \le n\}$. Furthermore, clearly $\mathbf{P}(\tau \le n) = (1-\alpha)^n$ where α is the probability that the sampler accepts Z_1. We thus obtain from Theorem 10:

Corollary 11. *Let $\{X_n\}$ be an independence sampler relative to a target density π and a proposal density q. Let τ be the first time the sampler accepts a proposed move, and let $\alpha = \mathbf{P}(\tau = 1)$ be the probability that the first proposed move is accepted. Let $Y_n = q(X_n)/\pi(X_n)$, and let $\nu(\cdot)$ be the corresponding stationary distribution of $\{Y_n\}$. Then*

$$\|\mathcal{L}(X_n \,|\, X_0 \sim \mu) - \pi(\cdot)\| \;\le\; \|\mathcal{L}(Y_n \,|\, X_0 \sim \mu) - \nu\| + (1-\alpha)^n \,.$$

Now, for large n the correction term $(1-\alpha)^n$ will typically be quite small. Hence, the total variation distance bounds on $\{Y_n\}$ are very close to corresponding bounds on $\{X_n\}$.

Remark. It would also be possible to consider partial future de-initialising, partial functional de-initialising (e.g. the independence sampler), partial backward de-initialising (again e.g. the independence sampler), etc., but we do not pursue those notions here.

Appendix: An Operator Theory Lemma.

The proof of Proposition 8 requires the following well-known technical property of self-adjoint operators.

Lemma A1. *Let H be any self-adjoint operator on any real or complex Hilbert space $\mathcal{H}$. Then*

$$\sup_{\substack{f\in\mathcal{H}\\ \|f\|=1}} \big|\langle f, Hf\rangle\big| \;=\; \|H\|\,.$$

Proof. Clearly $\sup\limits_{\|f\|=1} \big|\langle f, Hf\rangle\big| \le \|H\|$, so it suffices to find a sequence $\{g_n\}$ of vectors with $\|g_n\| = 1$ and $|\langle g_n, Hg_n\rangle| \to \|H\|$.

By the definition of $\|H\|$, we can find a sequence of vectors $\{f_n\}$ with $\|f_n\| = 1$ and $\|Hf_n\| \to \|H\|$. But then we compute that

$$\begin{aligned}
\big\|H^2 f_n - \|H\|^2 f_n\big\|^2 \;&=\; \Big\langle H^2 f_n - \|H\|^2 f_n,\; H^2 f_n - \|H\|^2 f_n\Big\rangle \\
&=\; \|H^2 f_n\|^2 - 2\,\|H\|^2\,\langle H^2 f_n,\; f_n\rangle + \|H\|^4 \\
&=\; \|H^2 f_n\|^2 - 2\,\|H\|^2\,\langle Hf_n,\; Hf_n\rangle + \|H\|^4 \\
&=\; \|H^2 f_n\|^2 - 2\,\|H\|^2\,\|Hf_n\|^2 + \|H\|^4 \\
&\le\; \|H\|^2\,\|Hf_n\|^2 - 2\,\|H\|^2\|Hf_n\|^2 + \|H\|^4 \\
&\to\; \|H\|^2\,\|H\|^2 - 2\,\|H\|^2\,\|H\|^2 + \|H\|^4 \\
&=\; 0\,.
\end{aligned}$$

That is, $H^2 f_n - \|H\|^2 f_n \to 0$.

Now,

$$H^2 f_n - \|H\|^2 f_n = \big(H + \|H\|I\big)\big(H - \|H\|I\big) f_n$$

(where I is the identity operator). To make use of the fact that this approaches 0, we note that we must have either (a) $\big(H - \|H\|I\big) f_n \to 0$, or (b) $\|\big(H - \|H\|I\big) f_n\| \ge \epsilon$ for infinitely many n and some fixed $\epsilon > 0$.

In case (a), we set $g_n = f_n$ to obtain that $\langle g_n,\; (H - \|H\|I) g_n\rangle \to 0$, so that $\langle g_n,\; Hg_n\rangle \to \|H\|$, as desired.

In case (b), we restrict to those n with $\|\big(H - \|H\|I\big) f_n\| \ge \epsilon$, and set $g_n = (H - \|H\|I) f_n \,/\, \|(H - \|H\|I) f_n\|$. We then obtain that $\|g_n\| = 1$, and that $\langle g_n, (H + \|H\|I) g_n\rangle \to$

0, so that $\langle g_n,\ Hg_n \rangle \to -\|H\|$. The result then follows by taking absolute values. ■

Remarks.

1. In operator theory, the quantity $\sup\limits_{\|f\|=1} \big|\langle f, Hf \rangle\big|$ is referred to as the *numerical radius* of H.
2. Lemma A1 is false if H is not required to be self-adjoint. For example, let H be the operator on $\mathbf{R}^2$ which rotates each vector clockwise by 90 degrees. Then $\sup_{\|f\|=1} \big|\langle f, Hf \rangle\big| = 0$ even though $\|H\| = 1$.
3. Lemma A1 is often stated over a complex Hilbert space. However, for our purposes we need the result over a real Hilbert space. In the context of the present paper, a complex Hilbert space corresponds to the supremum (6) including correlations of complex-valued functions with their complex conjugates. Indeed, on a complex Hilbert space, Lemma A1 is true within a factor of 2 even if H is not self-adjoint; see e.g. Halmos (1951), page 33. However, note the explicit use of $i = \sqrt{-1}$ in Halmos's proof, which is why his result does not hold on a real Hilbert space (cf. the previous remark).

Acknowledgements. We are very grateful to Peter Rosenthal for assistance with Lemma A1, and to Steffen Lauritzen for providing numerous references about classical sufficiency. We thank the two anonymous referees for very careful readings and many useful comments.

REFERENCES

Amit, Y. (1991), On the rates of convergence of stochastic relaxation for Gaussian and Non-Gaussian distributions. *J. Multivariate Analysis* **38**, 89–99.

Bahadur, R.R. (1954), Sufficiency and statistical decision functions. *Ann. Math. Stat.* **25**, 423–462.

Barndorff-Nielsen, O. and Skibinsky, M. (1963), Adequate subfields and almost sufficiency. Appl. Math. Publ. No. **329**, Brookhaven National Laboratory.

Besag, J.E. and Green, P.J. (1993), Spatial statistics and Bayesian computation (with discussion). *J. Royal Stat. Soc. Ser. B* **55**, 25–38.

R.N. Bhattacharya and E.C. Waymire (1990), Stochastic processes with applications. Wiley & Sons, New York.

Billingsley, P. (1995), Probability and Measure, 3rd ed. John Wiley & Sons, New York.

Brooks, S.P. and Roberts, G.O. (1998), Convergence assessment techniques for Markov chain Monte Carlo. Stat. and Comput. **8**, 319–335.

Cowles, M.K. and Carlin, B.P. (1996), Markov chain Monte Carlo convergence diagnostics: A comparative Review. J. Amer. Stat. Assoc. **91**, 883–904.

Cox, D.R. and Hinkley, D.V. (1974), Theoretical statistics. Chapman and Hall, London.

Damien, P., Wakefield, J.C. and Walker, S. (1997), Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. *J. Royal Stat. Soc., Series B*, to appear.

Diebolt, J. and Robert, C.P. (1994), Estimation of finite mixture distributions through Bayesian sampling. *J. Royal Stat. Soc. Ser. B* **56**, 363–375.

Durrett, R. (1991), Probability: theory and examples. Wadsworth, Pacific Grove, California.

Elliott, R.J., Aggoun, L., and Moore, J.B. (1995), Hidden Markov models: estimation and control. New York : Springer-Verlag.

Fisher, R.A. (1920), A mathematical examination of the models of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices R. Astronomical Soc.* **80**, 758–770. Reprinted (1950) in Fisher's *Contributions to Mathematical Statistics*. Wiley, New York.

Fishman, G. (1996), An analysis of Swendsen-Wang and Related Sampling Methods. Preprint, University of North Carolina, Department of Operations Research.

Gelfand, A.E. and Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities. J. Amer. Stat. Soc. **85**, 398–409.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence

6, 721-741.

Halmos, P.R. (1951), Introduction to Hilbert space and the theory of spectral multiplicity. Chelsea Publishing Company, New York.

Higdon, D.M. (1996), Auxiliary variable methods for Markov chain Monte Carlo with applications. Preprint, Institute of Statistics and Decision Sciences, Duke University.

Hoekstra, A.H. and Steutel, F.W. (1984), Limit theorems for Markov chains of finite rank. *Linear Alg. Appl.* **60**, 65–77.

Jacka, S.D. and Roberts, G.O. (1997), Strong forms of weak convergence. *Stoch. Proc. Appl.* **67**, 41–53.

Lauritzen, S.L. (1972), Sufficiency and time series analysis. Preprint No. 11, Inst. of Math. Statistics, University of Copenhagen.

Lauritzen, S.L. (1974), Sufficiency, prediction and extreme models. *Scand. J. Stat.* **1**, 128–134.

Lauritzen, S.L. (1988), Extremal families and systems of sufficient statistics. Springer Lecture Notes in Statistics, Vol. **49**. Springer, Berlin, New York.

Lauritzen, S.L. (1996), Graphical models. Oxford: Clarendon Press.

Liu, J.S., Wong, W., and Kong, A. (1994), Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81**, 27-40.

Mira, A. and Tierney, L. (1997), On the use of auxiliary variables in Markov chain Monte Carlo sampling. Preprint, School of Statistics, University of Minnesota.

Neal, R. (1997) Markov chain Monte Carlo methods based on 'slicing' the density function. Preprint.

Roberts, G.O. and Rosenthal, J.S. (1997a), Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2**, paper 2.

Roberts, G.O. and Rosenthal, J.S. (1997b), Convergence of slice sampler Markov chains. *J. Royal Stat. Soc. Ser. B*, to appear.

Roberts, G.O. and Rosenthal, J.S. (1999), The polar slice sampler. Preprint.

Rosenthal, J.S. (1992), Convergence of pseudo-finite Markov chains. Unpublished manuscript.

Rosenthal, J.S. (1993), Rates of convergence for data augmentation on finite sample spaces. Ann. Appl. Prob., Vol. **3**, No. **3**, 819–839.

Rosenthal, J.S. (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558-566.

Runnenburg, J.Th. and Steutel, F.W. (1962), On Markov chains the transition function of which is a finite sum of products of functions of one variable (summary). *Ann. Math. Stat.* **33**, 1483–1484.

Skibinsky, M. (1967), Adequate subfields and sufficiency. *Ann. Math. Stat.* **38**, 155-161.

Swendsen, R.H. and Wang, J.S. (1987), Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86–88.

Tanner, M. and Wong, W. (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Soc. **81**, 528–550.

Whittaker, J. (1990), Graphical models in applied multivariate statistics. New York: Wiley.