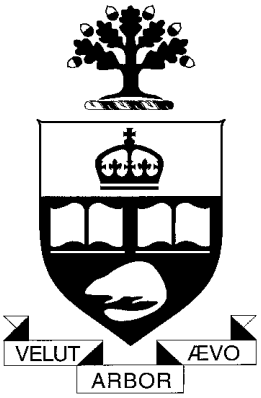


Adaptive Gibbs samplers

by

K. Łatuszyński
Department of Statistics
University of Warwick

and

J. S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 1001 January 14, 2010

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Adaptive Gibbs samplers

Krzysztof Łatuszyński and Jeffrey S. Rosenthal

K. Łatuszyński
Department of Statistics
University of Warwick
CV4 7AL, Coventry, UK
e-mail: latuch@gmail.com

J. S. Rosenthal
Department of Statistics
University of Toronto
Toronto, Ontario, Canada, M5S 3G3
e-mail: jeff@math.toronto.edu

(January, 2010)

Abstract: We consider various versions of adaptive Gibbs and Metropolis-within-Gibbs samplers, which update their selection probabilities (and perhaps also their proposal distributions) on the fly during a run, by learning as they go in an attempt to optimise the algorithm. We present a cautionary example of how even a simple-seeming adaptive Gibbs sampler may fail to converge. We then present various positive results guaranteeing convergence of adaptive Gibbs samplers under certain conditions.

AMS 2000 subject classifications: Primary 60J05, 65C05; secondary 62F15.
Keywords and phrases: MCMC estimation, adaptive MCMC, Gibbs sampling.

1. Introduction

Markov chain Monte Carlo is a commonly used approach to evaluating expectations of the form $\theta := \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x)$, where π is an intractable probability measure, e.g. known up to a normalising constant. One simulates $(X_n)_{n\geq 0}$, an ergodic Markov chain on $\mathcal{X}$, evolving according to a transition kernel P with stationary limiting distribution π and, typically, takes ergodic average as an estimate of θ. The approach is justified by asymptotic Markov chain theory, see e.g. [29, 38]. Metropolis algorithms and Gibbs samplers (to be described in Section 2) are among the most common MCMC algorithms, c.f. [31, 25, 38].

The quality of an estimate produced by an MCMC algorithm depends on probabilistic properties of the underlying Markov chain. Designing an appropriate transition kernel P that guarantees rapid convergence to stationarity and efficient simulation is often a challenging task, especially in high dimensions. For Metropolis algorithms there are various optimal scaling results [32, 36, 9, 10, 4, 37, 38, 41] which provide "prescriptions" of how to do this, though they typically depend on unknown characteristics of π.

For random scan Gibbs samplers, a further design decision is choosing the selection probabilities (i.e., coordinate weightings) which will be used to select which coordinate to update next. These are usually chosen to be uniform, but some recent work [26, 22, 23, 15, 43, 12] has suggested that non-uniform weightings may sometimes be preferable.

For a very simple toy example to illustrate this issue, suppose $\mathcal{X} = [0,1] \times [-100, 100]$, with $\pi(x_1, x_2) \propto x_1^{100}(1 + \sin(x_2))$. Then with respect to x_1, this π puts almost all of the mass right up against the line $x_1 = 1$. Thus, repeated Gibbs sampler updates of the coordinate x_1 make virtually no difference, and do not need to be done often at all (unless the functional f of interest is *extremely* sensitive to tiny changes in x_1). By contrast, with respect to x_2, this π is a highly multi-modal density with wide support and many peaks and valleys, requiring many updates to the coordinate x_2 in order to explore the state space appropriately. Thus, an efficient Gibbs sampler would not update each of x_1 and x_2 equally often; rather, it would update x_2 very often and x_1 hardly at all. Of course, in this simple example, it is easy to see directly that x_1 should be updated less than x_2, and furthermore such efficiencies would only improve the sampler by approximately a factor of 2. However, in a high-dimensional example (c.f. [12]), such issues could be much more significant and also much more difficult to detect manually.

One promising avenue to address this challenge is *adaptive MCMC algorithms*. As an MCMC simulation progresses, more and more information about the target distribution π is learned. Adaptive MCMC attempts to use this new information to redesign the transition kernel P on the fly, based on the current simulation output. That is, the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$. So, in the above toy example, a good adaptive Gibbs sampler would somehow automatically "learn" to update x_1 less often, without requiring the user to determine this manually (which could be difficult or impossible in a very high-dimensional problem).

Unfortunately, such adaptive algorithms are only valid if their ergodicity can be established. The stochastic process $(X_n)_{n\geq 0}$ for an adaptive algorithm is no longer a Markov chain; the potential benefit of adaptive MCMC comes at the price of requiring more sophisticated theoretical analysis. There is substantial and rapidly growing literature on both theory and practice of adaptive MCMC (see e.g. [16, 17, 5, 1, 18, 13, 39, 40, 21, 45, 46, 14, 8, 6, 7, 42, 44, 2, 3]) which includes counterintuitive examples where X_n fails to converge to the desired distribution π (c.f. [5, 39, 8, 21]), as well as many results guaranteeing ergodicity under various assumptions. Most of the previous work on ergodicity of adaptive MCMC has concentrated on adapting Metropolis and related algorithms, with less attention paid to ergodicity when adapting the selection probabilities for random scan Gibbs samplers.

Motivated by such considerations, in the present paper we study the ergodicity of various types of adaptive Gibbs samplers. To our knowledge, proofs of ergodicity for adaptively-weighted Gibbs samplers have previously been considered only by [24], and we shall provide a counter-example below (Example 3.1) to demonstrate that their main result is not correct. In view of this, we are not

aware of any valid ergodicity results in the literature that consider adapting selection probabilities of random scan Gibbs samplers, and we attempt to fill that gap herein.

This paper is organised as follows. We begin in Section 2 with basic definitions. In Section 3 we present a cautionary Example 3.1, where a seemingly ergodic adaptive Gibbs sampler is in fact transient (as we prove formally later in Section 8) and provides a counter-example to Theorem 2.1 of [24]. Next, we establish various positive results for ergodicity of adaptive Gibbs samplers. In Section 4, we consider adaptive random scan Gibbs samplers (`AdapRSG`) which update coordinate selection probabilities as the simulation progresses; in Section 5, we consider adaptive random scan Metropolis-within-Gibbs samplers (`AdapRSMwG`) which update coordinate selection probabilities as the simulation progresses; and in Section 6, we consider adaptive random scan adaptive Metropolis-within-Gibbs samplers (`AdapRSadapMwG`) that update coordinate selection probabilities as well as proposal distributions for the Metropolis steps – the case that corresponds most closely to the adaptations performed in the statistical genetics work of [12]. In each case, we prove that under reasonably mild conditions, the adaptive Gibbs samplers are guaranteed to be ergodic, although our cautionary example does show that it is important to verify some conditions before applying such algorithms. Finally, in Section 7 we consider particular methods of simultaneously adapting the selection probabilities and proposal distributions, and prove that in addition to being ergodic, such algorithms are approximately optimal under certain strong assumptions.

2. Preliminaries

Gibbs samplers are commonly used MCMC algorithms for sampling from complicated high-dimensional probability distributions π in cases where the full conditional distributions of π are easy to sample from. To define them, let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be an $d-$dimensional state space where $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and write $X_n \in \mathcal{X}$ as $X_n = (X_{n,1}, \ldots, X_{n,d})$. We shall use the shorthand notation

$$X_{n,-i} \;:=\; (X_{n,1}, \ldots, X_{n,i-1}, X_{n,i+1}, \ldots, X_{n,d})\,,$$

and similarly $\mathcal{X}_{-i} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_{i-1} \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_d$.

Let $\pi(\cdot|x_{-i})$ denote the conditional distribution of $Z_i \,|\, Z_{-i} = x_{-i}$ where $Z \sim \pi$. The random scan Gibbs sampler draws X_n given X_{n-1} (iteratively for $n = 1, 2, 3, \ldots$) by first choosing one coordinate at random according to some selection probabilities $\alpha = (\alpha_1, \ldots, \alpha_d)$ (e.g. uniformly), and then updating that coordinate by a draw from its conditional distribution. More precisely, the Gibbs sampler transition kernel $P = P_\alpha$ is the result of performing the following three steps.

Algorithm 2.1 (`RSG`(α))**.**

1. *Choose coordinate $i \in \{1, \ldots, d\}$ according to selection probabilities α, i.e. with $\mathbb{P}(i = j) = \alpha_j$*

2. *Draw* $Y \sim \pi(\cdot|X_{n-1,-i})$
3. *Set* $X_n := (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})$.

Whereas the standard approach is to choose the coordinate i at the first step uniformly at random, which corresponds to $\alpha = (1/d, \ldots, 1/d)$, this may be a substantial waste of simulation effort if d is large and variability of coordinates differs significantly. This has been discussed theoretically in [26] and also observed empirically e.g. in Bayesian variable selection for linear models in statistical genetics [43, 12]. We consider a class of adaptive random scan Gibbs samplers where selection probabilities $\alpha = (\alpha_1, \ldots, \alpha_d)$ are subject to optimization within some subset $\mathcal{Y} \subseteq [0,1]^d$ of possible choices. Therefore a single step of our generic adaptive algorithm for drawing X_n given the trajectory $X_{n-1}, \ldots, X_0$, and current selection probabilities $\alpha_{n-1} = (\alpha_{n-1,1}, \ldots, \alpha_{n-1,d})$ amounts to the following steps, where $R_n(\cdot)$ is some update rule for α_n.

Algorithm 2.2 (`AdapRSG`)**.**

1. *Set* $\alpha_n := R_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0) \in \mathcal{Y}$
2. *Choose coordinate* $i \in \{1, \ldots, d\}$ *according to selection probabilities* α_n
3. *Draw* $Y \sim \pi(\cdot|X_{n-1,-i})$
4. *Set* $X_n := (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})$

Algorithm 2.2 defines P_n, the transition kernel used at time n, and α_n plays here the role of Γ_n in the more general adaptive setting of e.g. [39, 8]. Let $\pi_n = \pi_n(x_0, \alpha_0)$ denote the distribution of X_n induced by Algorithm 2.1 or 2.2, given starting values x_0 and α_0, i.e. for $B \in \mathcal{B}(\mathcal{X})$,

$$\pi_n(B) = \pi_n\big((x_0, \alpha_0), B\big) := \mathbb{P}(X_n \in B | X_0 = x_0, \alpha_0). \tag{1}$$

Clearly if one uses Algorithm 2.1 then $\alpha_0 = \alpha$ remains fixed and $\pi_n(x_0, \alpha)(B) = P_\alpha^n(x_0, B)$. By $\|\nu - \mu\|_{TV}$ denote the total variation distance between probability measures ν and μ. Let

$$T(x_0, \alpha_0, n) := \|\pi_n(x_0, \alpha_0) - \pi\|_{TV}. \tag{2}$$

We call the adaptive Algorithm 2.2 *ergodic* if $T(x_0, \alpha_0, n) \to 0$ for π-almost every starting state x_0 and all $\alpha_0 \in \mathcal{Y}$.

We shall also consider random scan Metropolis-within-Gibbs samplers that instead of sampling from the full conditional at step (2) of Algorithm 2.1 (respectively at step (3) of Algorithm 2.2), perform a single Metropolis step. More precisely, given $X_{n-1,-i}$ the i-th coordinate $X_{n-1,i}$ is updated by a draw Y from the proposal distribution $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ with the usual Metropolis acceptance probability for the marginal stationary distribution $\pi(\cdot|X_{n-1,-i})$. Such Metropolis-within-Gibbs algorithms were originally proposed by [28] and have been very widely used. Versions of this algorithm which adapt the proposal distributions $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ were considered by e.g. [18, 40], but always with fixed (usually uniform) coordinate selection probabilities. If instead the proposal distributions $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ remain fixed, but the selection probabilities α_i are adapted on the fly, we obtain the following algorithm (where $q_{x,-i}(x, y)$ is the density function for $Q_{x,-i}(x, \cdot)$).

Algorithm 2.3 (`AdapRSMwG`)**.**

1. *Set* $\alpha_n := R_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0) \in \mathcal{Y}$
2. *Choose coordinate* $i \in \{1, \ldots, d\}$ *according to selection probabilities* α_n
3. *Draw* $Y \sim Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$
4. *With probability*

$$\min\left(1, \quad \frac{\pi(Y|X_{n-1,-i})\ q_{X_{n-1,-i}}(Y, X_{n-1,i})}{\pi(X_{n-1}|X_{n-1,-i})\ q_{X_{n-1,-i}}(X_{n-1,i}, Y)}\right), \tag{3}$$

accept the proposal and set

$$X_n = (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})\,;$$

otherwise, reject the proposal and set $X_n = X_{n-1}$.

Ergodicity of `AdapRSMwG` is considered in Section 5 below. Of course, if the proposal distribution $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ is symmetric about X_{n-1}, then the q factors in the acceptance probability (3) cancel out, and (3) reduces to the simpler probability $\min\big(1,\ \pi(Y|X_{n-1,-i})/\pi(X_{n-1}|X_{n-1,-i})\big)$.

We shall also consider versions of the algorithm in which the proposal distributions $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ are also chosen adaptively, from some family $\{Q_{x_{-i},\gamma}\}_{\gamma\in\Gamma_i}$ with corresponding density functions $q_{x_{-i},\gamma}$, as in e.g. the statistical genetics application [43, 12]. Versions of such algorithms with fixed selection probabilities are considered by e.g. [18] and [40]. They require additional adaptation parameters $\gamma_{n,i}$ that are updated on the fly and are allowed to depend on the past trajectories. More precisely, if $\gamma_n = (\gamma_{1,n}, \ldots, \gamma_{d,n})$ and $\mathcal{G}_n = \sigma\{X_0, \ldots, X_n, \alpha_0, \ldots, \alpha_n, \gamma_0, \ldots, \gamma_n\}$, then the conditional distribution of γ_n given $\mathcal{G}_{n-1}$ can be specified by the particular algorithm used, via a second update function R'_n. If we combine such proposal distribution adaptions with coordinate selection probability adaptions, this results in a doubly-adaptive algorithm, as follows.

Algorithm 2.4 (`AdapRSadapMwG`)**.**

1. *Set* $\alpha_n := R_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0, \gamma_{n-1}, \ldots, \gamma_0) \in \mathcal{Y}$
2. *Set* $\gamma_n := R'_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0, \gamma_{n-1}, \ldots, \gamma_0) \in \Gamma_1 \times \ldots \times \Gamma_n$
3. *Choose coordinate* $i \in \{1, \ldots, d\}$ *according to selection probabilities* α*, i.e. with* $\mathbb{P}(i = j) = \alpha_j$
4. *Draw* $Y \sim Q_{X_{n-1,-i},\gamma_{n-1}}(X_{n-1,i}, \cdot)$
5. *With probability (3),*

$$\min\left(1, \quad \frac{\pi(Y|X_{n-1,-i})\ q_{X_{n-1,-i},\gamma_{n-1}}(Y, X_{n-1,i})}{\pi(X_{n-1}|X_{n-1,-i})\ q_{X_{n-1,-i},\gamma_{n-1}}(X_{n-1,i}, Y)}\right),$$

accept the proposal and set

$$X_n = (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})\,;$$

otherwise, reject the proposal and set $X_n = X_{n-1}$.

Ergodicity of `AdapRSadapMwG` is considered in Section 6 below.

3. A counter-example

Adaptive algorithms destroy the Markovian nature of $(X_n)_{n\geq 0}$, and are thus notoriously difficult to analyse theoretically. In particular, it is easy to be tricked into thinking that a simple adaptive algorithm "must" be ergodic when in fact it is not.

For example, Theorem 2.1 of [24] states that ergodicity of adaptive Gibbs samplers follows from the following two simple conditions:

(i) $\alpha_n \to \alpha$ a.s. for some fixed $\alpha \in (0,1)^d$; and
(ii) The random scan Gibbs sampler with fixed selection probabilities α induces an ergodic Markov chain with stationary distribution π.

Unfortunately, this claim is false, i.e. (i) and (ii) alone do not guarantee ergodicity, as the following example and proposition demonstrate. (It seems that in the proof of Theorem 2.1 in [24], the same measure is used to represent trajectories of the adaptive process and of a corresponding non-adaptive process, which is not correct and thus leads to the error.)

Example 3.1. Let $\mathbb{N} = \{1, 2, \dots\}$, and let the state space $\mathcal{X} = \{(i,j) \in \mathbb{N}\times\mathbb{N} : i = j \text{ or } i = j+1\}$, with target distribution given by $\pi(i,j) \propto j^{-2}$. On $\mathcal{X}$, consider a class of adaptive random scan Gibbs samplers for π, as defined by Algorithm 2.2, with update rule given by:

$$R_n\Big(\alpha_{n-1}, X_{n-1} = (i,j)\Big) = \begin{cases} \left\{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\right\} & \text{if} \quad i = j, \\ \\ \left\{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\right\} & \text{if} \quad i = j+1, \end{cases} \tag{4}$$

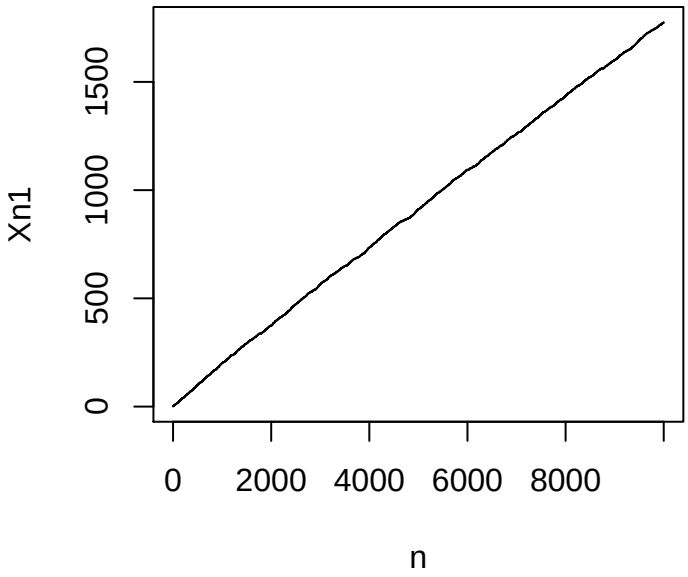
for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$.

Example 3.1 satisfies assumptions (i) and (ii) above. Indeed, (i) clearly holds since $\alpha_n \to \alpha := (\frac{1}{2}, \frac{1}{2})$, and (ii) follows immediately from the standard Markov chain properties of irreducibility and aperiodicity (c.f. [29, 38]). However, if a_n increases to ∞ slowly enough, then the example exhibits transient behaviour and is not ergodic. More precisely, we shall prove the following:

Proposition 3.2. *There exists a choice of the (a_n) for which the process $(X_n)_{n\geq 0}$ defined in Example 3.1 is not ergodic. Specifically, starting at $X_0 = (1,1)$, we have $\mathbb{P}(X_{n,1} \to \infty) > 0$, i.e. the process exhibits transient behaviour with positive probability, so it does not converge in distribution to any probability measure on $\mathcal{X}$. In particular, $||\pi_n - \pi||_{TV} \not\to 0$.*

Remark 3.3. In fact, we believe that in Proposition 3.2, $\mathbb{P}(X_{n,1} \to \infty) = 1$, though to reduce technicalities we only prove that $\mathbb{P}(X_{n,1} \to \infty) > 0$, which is sufficient to establish non-ergodicity.

A detailed proof of Proposition 3.2 is presented in Section 8. We also simulated Example 3.1 on a computer (with the (a_n) as defined in Section 8), resulting in the following trace plot of $X_{n,1}$ which illustrates the transient behaviour since $X_{n,1}$ increases quickly and steadily as a function of n:



4. Ergodicity of adaptive random scan Gibbs samplers

We now present various positive results about ergodicity of adaptive Gibbs samplers under various assumptions. Most of our results are specific to *uniformly ergodic* chains. (Recall that a Markov chain with transition kernel P is uniformly ergodic if there exist $M < \infty$ and $\rho < 1$ s.t. $\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq M\rho^n$ for every $x \in \mathcal{X}$; see e.g. [29, 38] for this and other notions related to general state space Markov chains.) In some sense this is a severe restriction, since most MCMC algorithms arising in statistical applications are not uniformly ergodic. However, truncating the variables involved at some (very large) value is usually sufficient to ensure uniform ergodicity without affecting the statistical conclusions in any practical sense, so this is not an insurmountable practical problem. We do plan to separately consider adaptive Gibbs samplers in the non-uniformly ergodic case, but that case appears to be considerably more technical so we do not pursue it further here.

To continue, recall that `RSG(`α`)` stands for random scan Gibbs sampler with selection probabilities α as defined by Algorithm 2.1, and `AdapRSG` is the adaptive version as defined by Algorithm 2.2. For notation, let $\Delta_{d-1} := \{(p_1, \ldots, p_d) \in \mathbb{R}^d : p_i \geq 0,\ \sum_{i=1}^d p_i = 1\}$ be the $(d-1)-$dimensional probability simplex, and

let

$$\mathcal{Y} := [\varepsilon, 1]^d \cap \Delta_{d-1} \tag{5}$$

for some $0 < \varepsilon \leq 1/d$. We shall generally assume that all our selection probabilities are in this set $\mathcal{Y}$, to avoid difficulties arising when one or more of the selection probabilities approach zero so certain coordinates are virtually never updated and thus get "stuck".

The main result of this section is the following.

Theorem 4.1. *Let the selection probabilities $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ *in probability for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$.*
(b) there exists $\beta \in \mathcal{Y}$ s.t. `RSG(`β`)` *is uniformly ergodic.*

Then `AdapRSG` *is ergodic, i.e.*

$$T(x_0, \alpha_0, n) \to 0 \qquad as \quad n \to \infty. \tag{6}$$

Moreover, if

(a') $\sup_{x_0,\alpha_0} |\alpha_n - \alpha_{n-1}| \to 0$ *in probability,*

then convergence of `AdapRSG` *is also uniform over all x_0, α_0, i.e.*

$$\sup_{x_0,\alpha_0} T(x_0, \alpha_0, n) \to 0 \qquad as \quad n \to \infty. \tag{7}$$

Remark 4.2.
1. Assumption *(b)* will typically be verified for $\beta = (1/d, \ldots, 1/d)$; see also Proposition 4.7 below.
2. We expect that most adaptive random scan Gibbs samplers will be designed so that $|\alpha_n - \alpha_{n-1}| \leq a_n$ for every $n \geq 1$, $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$, and $\omega \in \Omega$, for some deterministic sequence $a_n \to 0$ (which holds for e.g. the adaptations considered in [12]). In such cases, *(a')* is automatically satisfied.
3. The sequence α_n is not required to converge, and in particular the amount of adaptation, i.e. $\sum_{n=1}^{\infty} |\alpha_n - \alpha_{n-1}|$, is allowed to be infinite.
4. In Example 3.1, condition (a') is satisfied but condition (b) is not.
5. If we modify Example 3.1 by truncating the state space to say $\tilde{\mathcal{X}} = \mathcal{X} \cap (\{1, \ldots, M\} \times \{1, \ldots, M\})$ for some $1 < M < \infty$,, then the corresponding adaptive Gibbs sampler is ergodic, and (7) holds.

Before we proceed with the proof of Theorem 4.1, we need some preliminary lemmas, which may be of independent interest.

Lemma 4.3. *Let $\beta \in \mathcal{Y}$ with $\mathcal{Y}$ as in (5). If* `RSG(`β`)` *is uniformly ergodic, then also* `RSG(`α`)` *is uniformly ergodic for every $\alpha \in \mathcal{Y}$. Moreover there exist $M < \infty$ and $\rho < 1$ s.t. $\sup_{x_0 \in \mathcal{X}, \alpha \in \mathcal{Y}} T(x_0, \alpha, n) \leq M\rho^n \to 0$.*

Proof. Let P_β be the transition kernel of `RSG(`β`)`. It is well known that for uniformly ergodic Markov chains the whole state space $\mathcal{X}$ is small (c.f. Theorem

5.2.1 and 5.2.4 in [29] with their $\psi = \pi$). Thus there exists $s > 0$, a probability measure μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and a positive integer m, s.t. for every $x \in \mathcal{X}$,

$$P_\beta^m(x, \cdot) \geq s\mu(\cdot). \tag{8}$$

Fix $\alpha \in \mathcal{Y}$ and let

$$r := \min_i \frac{\alpha_i}{\beta_i}.$$

Since $\beta \in \mathcal{Y}$, we have $1 \geq r \geq \frac{\varepsilon}{1-(d-1)\varepsilon} > 0$ and P_α can be written as a mixture of transition kernels of two random scan Gibbs samplers, namely

$$P_\alpha = rP_\beta + (1-r)P_q, \qquad \text{where} \quad q = \frac{\alpha - r\beta}{1-r}.$$

This combined with (8) implies

$$\begin{aligned} P_\alpha^m(x, \cdot) &\geq r^m P_\beta^m(x, \cdot) \geq r^m s\mu(\cdot) \\ &\geq \Big(\frac{\varepsilon}{1-(d-1)\varepsilon}\Big)^m s\mu(\cdot) \qquad \text{for every} \quad x \in \mathcal{X}. \end{aligned} \tag{9}$$

By Theorem 8 of [38] condition (9) implies

$$\|P_\alpha^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq \left(1 - \Big(\frac{\varepsilon}{1-(d-1)\varepsilon}\Big)^m s\right)^{\lfloor n/m \rfloor} \qquad \text{for all} \quad x \in \mathcal{X}. \tag{10}$$

Since the right hand side of (10) does not depend on α, the claim follows. □

Lemma 4.4. *Let P_α and $P_{\alpha'}$ be random scan Gibbs samplers using selection probabilities $\alpha, \alpha' \in \mathcal{Y} := [\varepsilon, 1-(d-1)\varepsilon]^d$ for some $\varepsilon > 0$. Then*

$$\|P_\alpha(x, \cdot) - P_{\alpha'}(x, \cdot)\|_{TV} \leq \frac{|\alpha - \alpha'|}{\varepsilon + |\alpha - \alpha'|} \leq \frac{|\alpha - \alpha'|}{\varepsilon}. \tag{11}$$

Proof. Let $\delta := |\alpha - \alpha'|$. Then $r := \min_i \frac{\alpha_i'}{\alpha_i} \geq \frac{\varepsilon}{\varepsilon + \max_i |\alpha_i - \alpha_i'|} \geq \frac{\varepsilon}{\varepsilon+\delta}$ and reasoning as in the proof of Lemma 4.3 we can write $P_{\alpha'} = rP_\alpha + (1-r)P_q$ for some q and compute

$$\begin{aligned} \|P_\alpha(x, \cdot) - P_{\alpha'}(x, \cdot)\|_{TV} &= \|(rP_\alpha + (1-r)P_\alpha) - (rP_\alpha + (1-r)P_q)\|_{TV} \\ &= (1-r)\|P_\alpha - P_q\|_{TV} \leq \frac{\delta}{\varepsilon + \delta}, \end{aligned}$$

as claimed. □

Corollary 4.5. *$P_\alpha(x, B)$ as a function of α on $\mathcal{Y}$ is Lipshitz with Lipshitz constant $1/\varepsilon$ for every fixed set $B \in \mathcal{B}(\mathcal{X})$.*

Corollary 4.6. *If $|\alpha_n - \alpha_{n-1}| \to 0$ in probability, then also $\sup_{x \in \mathcal{X}} \|P_{\alpha_n}(x, \cdot) - P_{\alpha_{n-1}}(x, \cdot)\|_{TV} \to 0$ in probability.*

Proof of Theorem 4.1. We conclude the result from Theorem 1 of [39] that requires simultaneous uniform ergodicity and diminishing adaptation. Simultaneous uniform ergodicity results from combining assumption (b) and Lemma 4.3. Diminishing adaptation results from assumption (a) with Corollary 4.6. Moreover note that Lemma 4.3 is uniform in x_0 and α_0 and (a') yields uniformly diminishing adaptation again by Corollary 4.6. A look into the proof of Theorem 1 [39] reveals that this suffices for the uniform part of Theorem 4.1. □

Finally, we note that verifying uniform ergodicity of a random scan Gibbs sampler, as required by assumption (b) of Theorem 4.1, may not be straightforward. Such issues have been investigated in e.g. [33] and more recently in relation to the parametrization of hierarchical models (see [30] and references therein). In the following proposition, we show that to verify uniform ergodicity of any random scan Gibbs sampler, it suffices to verify uniform ergodicity of the corresponding systematic scan Gibbs sampler (which updates the coordinates $1, 2, \ldots, d$ in sequence rather than select coordinates randomly).

Proposition 4.7. *Let $\alpha \in \mathcal{Y}$ with $\mathcal{Y}$ as in (5). If the systematic scan Gibbs sampler is uniformly ergodic, then so is* `RSG(`α`)`.

Proof. Let

$$P = P_1 P_2 \cdots P_d$$

be the transition kernel of the uniformly ergodic systematic scan Gibbs sampler, where P_i stands for the step that updates coordinate i. By the minorisation condition characterisation, there exist $s > 0$, a probability measure μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and a positive integer m, s.t. for every $x \in \mathcal{X}$,

$$P^m(x, \cdot) \geq s\mu(\cdot).$$

However, the probability that the random scan Gibbs sampler $P_{1/d}$ in its md subsequent steps will update the coordinates in exactly the same order is $(1/d)^{md} > 0$. Therefore the following minorisation condition holds for the random scan Gibbs sampler.

$$P_{1/d}^{md}(x, \cdot) \geq (1/d)^{md} s\mu(\cdot).$$

We conclude that `RSG(`$1/d$`)` is uniformly ergodic, and then by Lemma 4.3 it follows that `RSG(`α`)` is uniformly ergodic for any $\alpha \in \mathcal{Y}$. □

5. Adaptive random scan Metropolis-within-Gibbs

In this section we consider random scan Metropolis-within-Gibbs sampler algorithms. Thus, given $X_{n-1,-i}$, the i-th coordinate $X_{n-1,i}$ is updated by a draw Y from the proposal distribution $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ with the usual Metropolis acceptance probability for the marginal stationary distribution $\pi(\cdot|X_{n-1,-i})$. Here, we consider Algorithm `AdapRSMwG`, where the proposal distributions $Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$ remain fixed, but the selection probabilities α_i are adapted on the fly. We shall prove ergodicity of such algorithms under some circumstances. (The more general algorithm `AdapRSadapMwG` is then considered in the following section.)

To continue, let $P_{x_{-i}}$ denote the resulting Metropolis transition kernel for obtaining $X_{n,i}|X_{n-1,i}$ given $X_{n-1,-i} = x_{-i}$. We shall require the following assumption.

Assumption 5.1. *For every $i \in \{1, \ldots, d\}$ the transition kernel $P_{x_{-i}}$ is uniformly ergodic for every $x_{-i} \in \mathcal{X}_{-i}$. Moreover there exist $s_i > 0$ and an integer m_i s.t. for every $x_{-i} \in \mathcal{X}_{-i}$ there exists a probability measure $\nu_{x_{-i}}$ on $(\mathcal{X}_i, \mathcal{B}(\mathcal{X}_i))$, s.t.*

$$P^{m_i}_{x_{-i}}(x_i, \cdot) \geq s_i \nu_{x_{-i}}(\cdot) \qquad \text{for every} \quad x_i \in \mathcal{X}_i.$$

We have the following counterpart of Theorem 4.1.

Theorem 5.2. *Let $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ in probability for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$.
(b) there exists $\beta \in \mathcal{Y}$ s.t. `RSG(β)` is uniformly ergodic.
(c) Assumption 5.1 holds.

Then `AdapRSMwG` is ergodic, i.e.

$$T(x_0, \alpha_0, n) \to 0 \qquad \text{as} \quad n \to \infty. \tag{12}$$

Moreover, if

(a') $\sup_{x_0, \alpha_0} |\alpha_n - \alpha_{n-1}| \to 0$ in probability,

then convergence of `AdapRSMwG` is also uniform over all x_0, α_0, i.e.

$$\sup_{x_0, \alpha_0} T(x_0, \alpha_0, n) \to 0 \qquad \text{as} \quad n \to \infty. \tag{13}$$

Remark 5.3. Remarks 4.2.1–4.2.3 still apply. Also, assumption 5.1 can easily be verified in some cases of interest, e.g.

1. Independence samplers are essentially uniformly ergodic if and only if the candidate density is bounded below by a multiple of the stationary density, i.e. $q(\mathrm{d}x) \geq s\pi(\mathrm{d}x)$ for some $s > 0$, c.f. [27].
2. The Metropolis-Hastings algorithm with continuous and positive proposal density $q(\cdot, \cdot)$ and bounded target density π is uniformly ergodic if the state space is compact, c.f. [29, 38].

To prove Theorem 5.2 we build on the approach of [35]. In particular recall the following notion of strong uniform ergodicity.

Definition 5.4. We say that a transition kernel P on $\mathcal{X}$ with stationary distribution π is $(m, s)-$*strongly uniformly ergodic*, if for some $s > 0$ and positive integer m

$$P^m(x, \cdot) \geq s\pi(\cdot) \qquad \text{for every} \quad x \in \mathcal{X}.$$

Moreover, we will say that a family of Markov chains $\{P_\gamma\}_{\gamma \in \Gamma}$ on $\mathcal{X}$ with stationary distribution π is $(m, s)-$*simultaneously strongly uniformly ergodic*, if for some $s > 0$ and positive integer m

$$P^m_\gamma(x, \cdot) \geq s\pi(\cdot) \qquad \text{for every} \quad x \in \mathcal{X} \quad \text{and} \quad \gamma \in \Gamma.$$

By Proposition 1 in [35], if a Markov chain is both uniformly ergodic and reversible, then it is strongly uniformly ergodic. The following lemma improves over this result by controlling both involved parameters.

Lemma 5.5. *Let μ be a probability measure on $\mathcal{X}$, let m be a positive integer and let $s > 0$. If a reversible transition kernel P satisfies the condition*

$$P^m(x, \cdot) \geq s\mu(\cdot) \qquad \text{for every} \quad x \in \mathcal{X},$$

then it is $\left(\left(\left\lfloor \frac{\log(s/4)}{\log(1-s)} \right\rfloor + 2\right) m, \frac{s^2}{8}\right)$ −strongly uniformly ergodic.

Proof. By Theorem 8 of [38] for every $A \in \mathcal{B}(\mathcal{X})$ we have

$$\|P^n(x, A) - \pi(A)\|_{TV} \leq (1-s)^{\lfloor n/m \rfloor},$$

And in particular

$$\|P^{km}(x, A) - \pi(A)\|_{TV} \leq s/4 \qquad \text{for} \quad k \geq \frac{\log(s/4)}{\log(1-s)}. \tag{14}$$

Since π is stationary for P, we have $\pi(\cdot) \geq s\mu(\cdot)$ and thus an upper bound for the Radon-Nikodym derivative

$$\mathrm{d}\mu/\mathrm{d}\pi \leq 1/s. \tag{15}$$

Moreover by reversibility

$$\pi(\mathrm{d}x)P^m(x, \mathrm{d}y) = \pi(\mathrm{d}y)P^m(y, \mathrm{d}x) \geq \pi(\mathrm{d}y)s\mu(\mathrm{d}x)$$

and consequently

$$P^m(x, \mathrm{d}y) \geq s\big(\mu(\mathrm{d}x)/\pi(\mathrm{d}x)\big)\pi(\mathrm{d}y). \tag{16}$$

Now define

$$A := \{x \in \mathcal{X} : \mu(\mathrm{d}x)/\pi(\mathrm{d}x) \geq 1/2\}$$

Clearly $\mu(A^c) \leq 1/2$. Therefore by (15) we have

$$1/2 \leq \mu(A) \leq (1/s)\pi(A)$$

and hence $\pi(A) \geq s/2$. Moreover (14) yields

$$P^{km}(x, A) \geq s/4 \qquad \text{for} \quad k := \left\lfloor \frac{\log(s/4)}{\log(1-s)} \right\rfloor + 1.$$

And with k defined above by (16) we have

$$\begin{aligned} P^{km+m}(x, \cdot) &= \int_{\mathcal{X}} P^{km}(x, \mathrm{d}z)P^m(z, \cdot) \geq \int_A P^{km}(x, \mathrm{d}z)P^m(z, \cdot) \\ &\geq \int_A P^{km}(x, \mathrm{d}z)(s/2)\pi(\cdot) \geq (s^2/8)\pi(\cdot). \end{aligned}$$

This completes the proof. □

We will need the following generalization of Lemma 4.3.

Lemma 5.6. *Let $\beta \in \mathcal{Y}$ with $\mathcal{Y}$ as in (5). If* `RSG(β)` *is uniformly ergodic then there exist $s' > 0$ and a positive integer m' s.t. the family $\{$`RSG(α)`$\}_{\alpha\in\mathcal{Y}}$ is $(m', s')-$simultaneously strongly uniformly ergodic.*

Proof. $P_\beta(x,\cdot)$ is uniformly ergodic and reversible, therefore by Proposition 1 in [35] it is $(m, s_1)-$strongly uniformly ergodic for some m and s_1. Therefore, and arguing as in the proof of Lemma 4.3, c.f. (9), there exist $s_2 \geq \left(\frac{\varepsilon}{1-(d-1)\varepsilon}\right)^m$, s.t. for every $\alpha \in \mathcal{Y}$ and every $x \in \mathcal{X}$

$$P_\alpha^m(x,\cdot) \geq s_2 P_\beta^m(x,\cdot) \geq s_1 s_2 \pi(\cdot). \tag{17}$$

Set $m' = m$ and $s' = s_1 s_2$. □

Proof of Theorem 5.2. We proceed as in the proof of Theorem 4.1, i.e. establish diminishing adaptation and simultaneous uniform ergodicity and conclude (12) and (13) from Theorem 1 of [39]. Observe that Lemma 4.4 applies for random scan Metropolis-within-Gibbs algorithms exactly the same way as for random scan Gibbs samplers. Thus diminishing adaptation results from assumption (a) and Corollary 4.6. To establish simultaneous uniform ergodicity, observe that by Assumption 5.1 and Lemma 5.5 the Metropolis transition kernel for ith coordinate i.e. $P_{x_{-i}}$ has stationary distribution $\pi(\cdot|x_{-i})$ and is $\left(\left(\left\lfloor \frac{\log(s_i/4)}{\log(1-s_i)} \right\rfloor + 2\right) m_i, \frac{s_i^2}{8}\right)$ −strongly uniformly ergodic. Moreover by Lemma 5.6 the family `RSG(α)`, $\alpha \in \mathcal{Y}$ is $(m', s')-$strongly uniformly ergodic, therefore by Theorem 2 of [35] the family of random scan Metropolis-within-Gibbs samplers with selection probabilities $\alpha \in \mathcal{Y}$, `RSMwG(α)`, is $(m_*, s_*)-$simultaneously strongly uniformly ergodic with m_* and s_* given as in [35]. □

We close this section with the following alternative version of Theorem 5.2.

Theorem 5.7. *Let $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ *in probability for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$.*
(b) there exists $\beta \in \mathcal{Y}$ s.t. `RSMwG(β)` *is uniformly ergodic.*

Then `AdapRSMwG` *is ergodic, i.e.*

$$T(x_0, \alpha_0, n) \to 0 \qquad as \quad n \to \infty. \tag{18}$$

Moreover, if

(a') $\sup_{x_0,\alpha_0} |\alpha_n - \alpha_{n-1}| \to 0$ *in probability,*

then convergence of `AdapRSMwG` *is also uniform over all x_0, α_0, i.e.*

$$\sup_{x_0,\alpha_0} T(x_0, \alpha_0, n) \to 0 \qquad as \quad n \to \infty. \tag{19}$$

Proof. Diminishing adaptation results from assumption (a) and Corollary 4.6. Simultaneous uniform ergodicity can be established as in the proof of Lemma 4.3. The claim follows from Theorem 1 of [39]. □

Remark 5.8. Whereas the statement of Theorem 5.7 may be useful in specific examples, typically condition (b), the uniform ergodicity of a random scan Metropolis-within-Gibbs sampler, will be not available and establishing it will involve conditions required by Theorem 5.2.

6. Adaptive random scan adaptive Metropolis-within-Gibbs

In this section, we consider the adaptive random scan adaptive Metropolis-within-Gibbs algorithm `AdapRSadapMwG`, that updates both selection probabilities of the Gibbs kernel and proposal distributions of the Metropolis step. Thus, given $X_{n-1,-i}$, the i-th coordinate $X_{n-1,i}$ is updated by a draw Y from a proposal distribution $Q_{X_{n-1,-i},\,\gamma_{n,i}}(X_{n-1,i},\cdot)$ with the usual acceptance probability. This doubly-adaptive algorithm has been used by e.g. [12] for an application in statistical genetics. As with adaptive Metropolis algorithms, the adaption of the proposal distributions in this setting is motivated by optimal scaling results for random walk Metropolis algorithms [32, 36, 9, 10, 4, 37, 38, 40, 41].

Let $P_{x_{-i},\,\gamma_{n,i}}$ denote the resulting Metropolis transition kernel for obtaining $X_{n,i}|X_{n-1,i}$ given $X_{n-1,-i} = x_{-i}$. We will prove ergodicity of this generalised algorithm using tools from the previous section. Assumption 5.1 must be reformulated accordingly, as follows.

Assumption 6.1. *For every $i \in \{1,\dots,d\}$, $x_{-i} \in \mathcal{X}_{-i}$ and $\gamma_i \in \Gamma_i$, the transition kernel $P_{x_{-i},\,\gamma_i}$ is uniformly ergodic. Moreover there exist $s_i > 0$ and an integer m_i s.t. for every $x_{-i} \in \mathcal{X}_{-i}$ and $\gamma_i \in \Gamma_i$ there exists a probability measure $\nu_{x_{-i},\,\gamma_i}$ on $(\mathcal{X}_i, \mathcal{B}(\mathcal{X}_i))$, s.t.*

$$P^{m_i}_{x_{-i},\,\gamma_i}(x_i,\cdot) \geq s_i \nu_{x_{-i},\,\gamma_i}(\cdot) \qquad \text{for every} \quad x_i \in \mathcal{X}_i.$$

We have the following counterpart of Theorems 4.1 and 5.2.

Theorem 6.2. *Let $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ *in probability for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$.*
(b) there exists $\beta \in \mathcal{Y}$ s.t. `RSG(`β`)` is uniformly ergodic.
(c) Assumption 6.1 holds.
(d) The Metropolis-within-Gibbs kernels exhibit diminishing adaptation, i.e. for every $i \in \{1,\dots,d\}$ the $\mathcal{G}_{n+1}$ measurable random variable

$$\sup_{x\in\mathcal{X}} \|P_{x_{-i},\,\gamma_{n+1,i}}(x_i,\cdot) - P_{x_{-i},\,\gamma_{n,i}}(x_i,\cdot)\|_{TV} \to 0 \text{ in probability, as } n \to \infty,$$

for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$.

Then `AdapRSadapMwG` is ergodic, i.e.

$$T(x_0, \alpha_0, n) \to 0 \qquad \text{as} \quad n \to \infty. \tag{20}$$

Moreover, if

(a') $\sup_{x_0,\alpha_0} |\alpha_n - \alpha_{n-1}| \to 0$ *in probability,*

(d') $\sup_{x_0,\alpha_0}\sup_{x\in\mathcal{X}}\|P_{x_{-i},\,\gamma_{n+1,i}}(x_i,\cdot)-P_{x_{-i},\,\gamma_{n,i}}(x_i,\cdot)\|_{TV}\to 0$ *in probability, then convergence of* `AdapRSadapMwG` *is also uniform over all* x_0,α_0, *i.e.*

$$\sup_{x_0,\alpha_0} T(x_0,\alpha_0,n)\to 0 \qquad as \quad n\to\infty. \tag{21}$$

Remark 6.3. Remarks 4.2.1–4.2.3 still apply. And, Remark 5.3 applies for verifying Assumption 6.1. Verifying condition (d) is discussed after the proof.

Proof. We again proceed by establishing diminishing adaptation and simultaneous uniform ergodicity and concluding the result from Theorem 1 of [39]. To establish simultaneous uniform ergodicity we proceed as in the proof of Theorem 5.2. Observe that by Assumption 6.1 and Lemma 5.5 every adaptive Metropolis transition kernel for ith coordinate i.e. $P_{x_{-i},\,\gamma_i}$ has stationary distribution $\pi(\cdot|x_{-i})$ and is $\left(\left(\left\lfloor\frac{\log(s_i/4)}{\log(1-s_i)}\right\rfloor+2\right)m_i,\,\frac{s_i^2}{8}\right)$−strongly uniformly ergodic. Moreover, by Lemma 5.6 the family $\texttt{RSG}(\alpha)$, $\alpha\in\mathcal{Y}$ is (m',s')−strongly uniformly ergodic, therefore by Theorem 2 of [35] the family of random scan Metropolis-within-Gibbs samplers with selection probabilities $\alpha\in\mathcal{Y}$ and proposals indexed by $\gamma\in\Gamma$, is (m_*,s_*)−simultaneously strongly uniformly ergodic with m_* and s_* given as in [35].

For diminishing adaptation we write

$$\begin{aligned}\sup_{x\in\mathcal{X}}\|P_{\alpha_n,\,\gamma_n}(x,\cdot)-P_{\alpha_{n-1},\,\gamma_{n-1}}(x,\cdot)\|_{TV} \;\le\; & \\ \sup_{x\in\mathcal{X}}\|P_{\alpha_n,\,\gamma_n}(x,\cdot)-P_{\alpha_{n-1},\,\gamma_n}(x,\cdot)\|_{TV} & \\ +\;\sup_{x\in\mathcal{X}}\|P_{\alpha_{n-1},\,\gamma_n}(x,\cdot)-P_{\alpha_{n-1},\,\gamma_{n-1}}(x,\cdot)\|_{TV} &\end{aligned}$$

The first term above converges to 0 in probability by Corollary 4.6 and assumption (a). The second term

$$\begin{aligned}\sup_{x\in\mathcal{X}}\|P_{\alpha_{n-1},\,\gamma_n}(x,\cdot)-P_{\alpha_{n-1},\,\gamma_{n-1}}(x,\cdot)\|_{TV} \;\le\; & \\ \sum_{i=1}^{d}\alpha_{n-1,i}\sup_{x\in\mathcal{X}}\|P_{x_{-i},\,\gamma_{n+1,i}}(x_i,\cdot)-P_{x_{-i},\,\gamma_{n,i}}(x_i,\cdot)\|_{TV} &\end{aligned}$$

converges to 0 in probability as a mixture of terms that converge to 0 in probability. □

The following lemma can be used to verify assumption (d) of Theorem 6.2; see also Example 6.5 below.

Lemma 6.4. *Assume that the adaptive proposals exhibit diminishing adaptation i.e. for every* $i\in\{1,\dots,d\}$ *the* $\mathcal{G}_{n+1}$ *measurable random variable*

$$\sup_{x\in\mathcal{X}}\|Q_{x_{-i},\,\gamma_{n+1,i}}(x_i,\cdot)-Q_{x_{-i},\,\gamma_{n,i}}(x_i,\cdot)\|_{TV}\to 0 \text{ in probability, as } n\to\infty,$$

for fixed starting values $x_0\in\mathcal{X}$ *and* $\alpha_0\in\mathcal{Y}$.

Then any of the following conditions

(i) The Metropolis proposals have symmetric densities, i.e.

$$q_{x_{-i},\,\gamma_{n,i}}(x_i, y_i) \;=\; q_{x_{-i},\,\gamma_{n,i}}(y_i, x_i),$$

(ii) $\mathcal{X}_i$ is compact for every i, π is continuous, everywhere positive and bounded, implies condition (d) of Theorem 6.2.

Proof. Let P_1, P_2 denote transition kernels and Q_1, Q_2 proposal kernels of two generic Metropolis algorithms for sampling from π on arbitrary state space $\mathcal{X}$. To see that (i) implies (d) we check that

$$\|P_1(x,\cdot) - P_2(x,\cdot)\|_{TV} \;\leq\; 2\|Q_1(x,\cdot) - Q_2(x,\cdot)\|_{TV}.$$

Indeed, the acceptance probability

$$\alpha(x,y) = \min\Big\{1, \frac{\pi(y)}{\pi(x)}\Big\} \;\in [0,1]$$

does not depend on the proposal, and for any $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ we compute

$$\begin{aligned} |P_1(x,A) - P_2(x,A)| \;&\leq\; \Big|\int_A \alpha(x,y)\big(q_1(y) - q_2(y)\big)\mathrm{d}y\Big| \\ &\qquad + \mathbb{I}_{\{x\in A\}}\Big|\int_{\mathcal{X}} \big(1-\alpha(x,y)\big)\big(q_1(y) - q_2(y)\big)\mathrm{d}y\Big| \\ &\leq\; 2\|Q_1(x,\cdot) - Q_2(x,\cdot)\|_{TV}. \end{aligned}$$

Condition (ii) implies that there exists $K < \infty$, s.t. $\pi(y)/\pi(x) \leq K$ for every $x, y \in \mathcal{X}$. To conclude that (d) results from (ii) note that

$$|\min\{a,b\} - \min\{c,d\}| \;<\; |a-c| + |b-d| \tag{22}$$

and recall acceptance probabilities $\alpha_i(x,y) = \min\Big\{1, \frac{\pi(y)q_i(y,x)}{\pi(x)q_i(x,y)}\Big\}$. Indeed for any $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ using (22) we have

$$\begin{aligned} |P_1(x,A) - P_2(x,A)| \;&\leq\; \Big|\int_A \Big(\min\Big\{q_1(x,y), \frac{\pi(y)}{\pi(x)}q_1(y,x)\Big\} \\ &\qquad\qquad - \min\Big\{q_2(x,y), \frac{\pi(y)}{\pi(x)}q_2(y,x)\Big\}\Big)\mathrm{d}y\Big| \\ &\qquad + \mathbb{I}_{\{x\in A\}}\Big|\int_{\mathcal{X}} \Big(\big(1-\alpha_1(x,y)\big)q_1(x,y) \\ &\qquad\qquad -\big(1-\alpha_2(x,y)\big)q_2(x,y)\Big)\mathrm{d}y\Big| \\ &\leq\; 4(K+1)\|Q_1(x,\cdot) - Q_2(x,\cdot)\|_{TV} \end{aligned}$$

And the claim follows since a random scan Metropolis-within-Gibbs sampler is a mixture of Metropolis samplers. □

We now provide an example to show that diminishing adaptation of proposals as in Lemma 6.4 does not necessarily imply condition (d) of Theorem 6.2, so some additional assumption is required, e.g. (i) or (ii) of Lemma 6.4.

Example 6.5. Consider a sequence of Metropolis algorithms with transition kernels $P_1, P_2, \dots$ designed for sampling from $\pi(k) = p^k(1-p)$ on $\mathcal{X} = \{0, 1, \dots\}$. The transition kernel P_n results from using proposal kernel Q_n and the standard acceptance rule, where

$$Q_n(j,k) \quad = \quad q_n(k) \quad := \quad \begin{cases} p^k\left(\frac{1}{1-p} - p^n + p^{2n}\right)^{-1} & \text{for} \quad k \neq n, \\ p^{2n}\left(\frac{1}{1-p} - p^n + p^{2n}\right)^{-1} & \text{for} \quad k = n. \end{cases}$$

Clearly

$$\sup_{j \in \mathcal{X}} \|Q_{n+1}(j,\cdot) - Q_n(j,\cdot)\|_{TV} \quad = \quad q_{n+1}(n) - q_n(n) \to 0.$$

However

$$\begin{aligned} \sup_{j \in \mathcal{X}} \|P_{n+1}(j,\cdot) - P_n(j,\cdot)\|_{TV} \quad &\geq \quad P_{n+1}(n,0) - P_n(n,0) \\ &= \quad \min\left\{q_{n+1}(0), \frac{\pi(0)}{\pi(n)} q_{n+1}(n)\right\} \\ &\qquad - \quad \min\left\{q_n(0), \frac{\pi(0)}{\pi(n)} q_n(n)\right\} \\ &= \quad q_{n+1}(0) - q_n(0) p^n \to 1 - p \neq 0. \end{aligned}$$

7. A specific Metropolis-within-Gibbs adaptive choice

As an application of the previous section, we discuss a particular method of adapting the α_i selection probabilities for the doubly-adaptive Metropolis-within-Gibbs algorithms. We are motivated by two closely-related componentwise adaptation algorithms, from [18] and from Section 3 of [40]. Briefly, these algorithms use a deterministic scan Metropolis-within-Gibbs sampler and perform a random walk Metropolis step for updating coordinate i by proposing a normal increment to $X_{n-1,i}$, i.e. the proposal $Y_{n,i} \sim N(X_{n-1,i}, \sigma^2_{n,i})$. The proposal variance $\sigma^2_{n,i}$ is subject to adaptation. Haario et al. in [18] use

$$\sigma^{2,\mathrm{HST}}_{n,i} \quad = \quad (2.4)^2(s^2_{n,i} + 0.05), \tag{23}$$

where $s^2_{n,i}$ is the sample variance of $X_{0,i}, \dots, X_{n-1,i}$, whereas Roberts and Rosenthal in [40] take

$$\sigma^{2,\mathrm{RR}}_{n,i} \quad = \quad e^{ls_i}, \tag{24}$$

and ls_i is updated every batch of 50 iterations by adding or subtracting $\delta(n) = O(n^{-1/2})$. Specifically, ls_i is increased by $\delta(n)$ if the fraction of acceptances of variable i was more then 0.44 on the last batch and decreased if it was less.

Both rules have theoretical motivation, c.f. [37]; $\sigma^{2,\mathrm{HST}}_{n,i}$ comes from diffusion limit considerations in infinite dimensions and $\sigma^{2,\mathrm{RR}}_{n,i}$ is motivated by one dimensional Gaussian target densities. Conclusions drawn in this very special situations are observed empirically to be robust in a wide range of examples that are neither high-dimensional nor Gaussian [37, 40].

In this section, we use a random scan Gibbs sampler instead of a deterministic scan, and optimise the coordinate selection probabilities α_i simultaneously with proposal variances. We aim at minimizing the asymptotic variance. Under certain strong conditions (Assumption 7.1) that allow for illustrative analysis and explicit calculations, we shall provide approximately optimal adaptions for the α_i in equations (43) and (44) below, and shall prove ergodicity of the corresponding algorithms in Theorem 7.3. More general adaptation algorithms for random scan Gibbs samplers have been investigated by others (e.g. [26, 24, 22, 23]).

Assumption 7.1. *The following conditions hold.*

(i) The stationary distribution on $\mathcal{X} = \mathbb{R}^d$ is of the product form

$$\pi(x) \quad = \quad \prod_{i=1}^{d} C_i \, g(C_i x_i), \tag{25}$$

where g is a one dimensional density and C_i, $i = 1, \ldots, d$, are unknown, strictly positive constants.

(ii) The second moment of g exists, i.e. $\sigma^2 := Var_g Z < \infty$.

(iii) The one-dimensional random walk Metropolis algorithm with $N(x, 1)$ proposal distributions and target density g is uniformly ergodic.

We consider an adaptive random scan adaptive random walk Metropolis-within-Gibbs algorithm `AdapRSadapMwG`, with Gaussian proposals, for estimating expectation of a linear target function

$$f(x) \quad = \quad a_0 + \sum_{i=1}^{d} a_i x_i. \tag{26}$$

A random scan Gibbs sampler for a target density of product form (25) is uniformly ergodic, therefore arguing as in the proof of Theorem 6.2, under Assumption 7.1 a random scan Metropolis-within-Gibbs with $N(x, 1)$ proposals is uniformly ergodic. Moreover, by (ii), function f defined in (26) is square integrable and the Markov chain CLT holds, i.e. for any initial distribution of X_0

$$n^{-1/2}\left(\sum_{k=0}^{n-1} f(X_i) - n\mathbb{E}_\pi f(X) \right) \quad \to \quad N(0, \sigma^2_{\mathrm{as}}), \qquad \text{as} \quad n \to \infty, \tag{27}$$

where the asymptotic variance $\sigma^2_{\mathrm{as}} < \infty$ can be written as

$$\sigma^2_{\mathrm{as}} \quad = \quad \tau_f Var_\pi f(X), \qquad \text{and} \tag{28}$$

$$\tau_f \quad = \quad 1 + 2\sum_{k=1}^{\infty} Cor_\pi(f(X_0), f(X_k)), \tag{29}$$

is the stationary integrated autocorrelation time. Markov chain CLTs and asymptotic variance formulae are discussed e.g. in [38, 19, 11]. Note that under Assumption 7.1 the asymptotic variance decomposes and some explicit computations are possible.

$$\sigma^2_{\text{as}} = \sum_{i=1}^{d} \sigma^2_{\text{as},i} = \sum_{i=1}^{d} \tau_{f,i} Var_{\pi,i} f(X), \qquad \text{where} \tag{30}$$

$$\tau_{f,i} = 1 + 2\sum_{k=1}^{\infty} Cor_{\pi}(X_{0,i}, X_{k,i}), \qquad \text{and} \tag{31}$$

$$Var_{\pi,i} f(X) := Var_{\pi}(a_i X_{0,i}) = \frac{a_i^2}{C_i^2}\sigma^2. \tag{32}$$

To compute $\tau_{f,i}$ for a random scan Metropolis-within-Gibbs sampler in the present setting, we focus solely on coordinate i, i.e. the Markov chain $X_{n,i}$, $n = 0, 1, \ldots$ Due to the product form of π, the distribution of $X_{n,i}|X_n$ does not depend on $X_{n,-i}$. Let P_i be the transition kernel that describes the dynamics of $X_{n,i}$, $n = 0, 1, \ldots$ and let $\alpha = (\alpha_1, \ldots, \alpha_d)$ denote the (fixed) selection probabilities. We write P_i as a mixture

$$P_i(x_i, \cdot) = (1 - \alpha_i)\text{Id} + \alpha_i P_i^{\text{Metrop}}(x_i, \cdot), \tag{33}$$

where Id denotes the identity kernel and P_i^{Metrop} performs a single Metropolis step for the target distribution $C_i g(C_i x)$. Thus P_i is a lazy version of P_i^{Metrop}, since it performs a P_i^{Metrop} step if coordinate i is selected with probability α_i and an identity step otherwise. We will use Lemma 7.2 below, which is a general result about asymptotic variance of lazy reversible Markov chains. Suppose

$$h \in L_0^2(\pi) := \{h \in L^2(\pi) : \pi h = 0\},$$

and denote

$$\sigma^2_{h,H} := \lim_{n\to\infty} \frac{1}{n} Var\Big(\sum_{i=0}^{n-1} h(Z_i)\Big),$$

where $Z_0, Z_2, \ldots$ is a Markov chain with transition kernel H and initial distribution π that is stationary for H.

Lemma 7.2. *Let P be a reversible transition kernel with stationary measure π. Let $\delta \in (0, 1)$ and by P_δ denote its lazy version*

$$P_\delta = (1 - \delta)Id + \delta P.$$

Then

$$\sigma^2_{h,P_\delta} = \frac{1}{\delta}\sigma^2_{h,P} + \frac{1-\delta}{\delta}\pi h^2. \tag{34}$$

Proof. The proof is based on the functional analytic approach (see e.g. [20, 34]). A reversible transition kernel P with invariant distribution π is a self-adjoint operator on $L_0^2(\pi)$ with spectral radius bounded by 1. By the spectral decomposition theorem for self adjoint operators, for each $h \in L_0^2(\pi)$ there exists a finite positive measure $E_{h,P}$ on $[-1,1]$, such that

$$\langle h, P^n h\rangle \quad = \quad \int_{[-1,1]} x^n E_{h,P}(dx),$$

for all integers $n \geq 0$. Thus in particular

$$\pi h^2 \quad = \quad \int_{[-1,1]} 1 E_{h,P}(dx), \tag{35}$$

$$\sigma^2_{h,P} \quad = \quad \int_{[-1,1]} \frac{1+x}{1-x} E_{h,P}(dx). \tag{36}$$

Since

$$P_\delta^n \quad = \quad \big((1-\delta)\mathrm{Id} + \delta P\big)^n = \sum_{k=0}^{n} \binom{n}{k} (1-\delta)^k \delta^{n-k} P^{n-k},$$

we have

$$\begin{aligned}
\langle h, P_\delta^n h\rangle &= \int_{[-1,1]} ((1-\delta)+\delta x)^n E_{h,P}(dx), \qquad \text{and consequently} \\
\sigma^2_{h,P_\delta} &= \int_{[-1,1]} \frac{1+1-\delta+\delta x}{1-1+\delta-\delta x} E_{h,P}(dx) \\
&= \int_{[-1,1]} \frac{1}{\delta}\left(\frac{1+x}{1-x} + 1 - \delta\right) E_{h,P}(dx) \\
&= \frac{1}{\delta}\int_{[-1,1]} \frac{1+x}{1-x} E_{h,P}(dx) + \frac{1-\delta}{\delta}\int_{[-1,1]} 1 E_{h,P}(dx),
\end{aligned}$$

as claimed. □

Let

$$\tilde{\sigma}^2_{\mathrm{as},i} \quad = \quad \tilde{\tau}_{f,i} Var_\pi(a_i X_{0,i}) \quad = \quad \tilde{\tau}_{f,i} \frac{a_i^2}{C_i^2}\sigma^2, \tag{37}$$

be the asymptotic variance of the Metropolis kernel P_i^{Metrop} defined in (33). Here $\tilde{\tau}_{f,i}$ is its stationary integrated autocorrelation time. From Lemma 7.2 we have the following formula for $\sigma^2_{as,i}$ of (30).

$$\begin{aligned}
\sigma^2_{\mathrm{as},i} &= \frac{1}{\alpha_i}\tilde{\sigma}^2_{\mathrm{as},i} + \frac{1-\alpha_i}{\alpha_i}\frac{a_i^2}{C_i^2}\sigma^2, \qquad \text{hence} \\
\tau_{f,i} &= \frac{1}{\alpha_i}\tilde{\tau}_{\mathrm{as},i} + \frac{1-\alpha_i}{\alpha_i}.
\end{aligned} \tag{38}$$

Now we take advantage of the fact that f is linear and of the actual adaptation of the proposal variances performed by both versions, i.e. HST and RR. Namely, they aim at minimizing their integrated autocorrelation time $\tilde{\tau}_{\mathrm{as},i}$. Under Assumption 7.1 the conditional distributions are equally shaped up to the scaling constant C_i. However the adaptive algorithm will learn C_i and adjust the proposal variance accordingly. We conclude that after an initial learning period the following proportionality relation will hold approximately

$$\sigma_{n,i}^{2,\mathrm{HST}} \;\propto\; \sigma_{n,i}^{2,\mathrm{RR}} \;\propto\; 1/C_i^2,$$

and also the stationary integrated autocorrelation times for the adapted P_i^{Metrop} will be close to the (unknown) optimal value, say T, i.e.

$$\tilde{\tau}_{\mathrm{as},i} \;\approx\; T. \tag{39}$$

Typically $T \gg 1$, hence we can approximately write (using (38), (30), (31), (32) and (33))

$$\begin{aligned} \tau_{f,i} &\approx T/\alpha_i, \\ \sigma_{\mathrm{as},i}^2 &\approx \frac{T a_i^2}{C_i^2 \alpha_i}\sigma^2, \qquad \text{and finally} \end{aligned} \tag{40}$$

$$\sigma_{\mathrm{as}}^2 \;\approx\; T\sigma^2 \sum_{i=1}^{d} \frac{a_i^2}{C_i^2\alpha_i} \;\propto\; \sum_{i=1}^{d} \frac{\sigma_{n,i}^{2,\mathrm{HST}} a_i^2}{\alpha_i} \;\propto\; \sum_{i=1}^{d} \frac{\sigma_{n,i}^{2,\mathrm{RR}} a_i^2}{\alpha_i}. \tag{41}$$

The last expression is minimised for

$$\alpha_i \;\propto\; \left(\sigma_{n,i}^{2,\mathrm{HST}} a_i^2\right)^{1/2} \;\propto\; \left(\sigma_{n,i}^{2,\mathrm{RR}} a_i^2\right)^{1/2}, \tag{42}$$

which yields a very intuitive prescription for adapting selection probabilities, namely by setting

$$\alpha_{n,i}^{\mathrm{HST}} := \frac{\left(\sigma_{n,i}^{2,\mathrm{HST}} a_i^2\right)^{1/2}}{\sum_{k=1}^{d}\left(\sigma_{n,k}^{2,\mathrm{HST}} a_k^2\right)^{1/2}} \qquad \text{for the HST version of [18], and} \tag{43}$$

$$\alpha_{n,i}^{\mathrm{RR}} := \frac{\left(\sigma_{n,i}^{2,\mathrm{RR}} a_i^2\right)^{1/2}}{\sum_{k=1}^{d}\left(\sigma_{n,k}^{2,\mathrm{RR}} a_k^2\right)^{1/2}} \qquad \text{for the RR version of [40].} \tag{44}$$

The above argument shows: *(43) and (44) are approximately optimal choices of adaptive selection probabilities for these algorithms, at least for target densities of the form (25).*

We next prove ergodicity of these algorithms. Let HST-algorithm denote an `AdapRSadapMwG` that uses (23) for updating proposal variances and (43) for updating selection probabilities. Similarly let RR-algorithm follow (24) and (44) with additional restriction for ls_i to stay in $[-M, M]$ for some fixed, large $M < \infty$ (which technically plays the role of 0.05 in (23) for the HST-algorithm).

Theorem 7.3. *Under Assumption 7.1 the HST- and RR-algorithms are ergodic.*

Proof. It is enough to check that the assumptions of Theorem 6.2 are satisfied. We do this for the HST-algorithm; the proof for the RR-algorithm follows in the same way. Condition (b) is immediately implied by Assumption 7.1 (i), since (b) requires only that the full Gibbs sampler is uniformly ergodic, which is obvious for a product target density of the form (25). Next, observe that Assumption 7.1 (iii) implies that the support of $C_i g(C_i x)$, say $S_{g,i}$, is bounded, therefore the sample variance estimate in (23) is bounded from above and for the HST-algorithm, for every $i \in \{1, \ldots, d\}$,

$$\sigma_{n,i}^{2,\mathrm{HST}} \quad \in \quad [(2.4)^2 0.05,\ K_i] \ =: \ S_{\sigma,i} \tag{45}$$

for some $K_i < \infty$. Thus (a) holds since the denominator in (43) is bounded from below and the change in sample variance

$$\sigma_{n,i}^{2,\mathrm{HST}} - \sigma_{n+1,i}^{2,\mathrm{HST}} \quad = \quad O(n^{-1}). \tag{46}$$

Condition (d) results from (46), (45) and Lemma 6.4 (i). We are left with (c). Let $\phi_\sigma(\cdot)$ denote the density function of $N(0, \sigma^2)$. Since

$$\sup_{i \in \{1,\ldots,d\}; x,y \in S_{g,i}; \sigma_1, \sigma_2 \in S_{\sigma,i}} \phi_{\sigma_1}(x-y)/\phi_{\sigma_2}(x-y) < \infty,$$

the Radon-Nikodym derivative of all pairs of proposals for every coordinate is bounded and hence Assumption 6.1 is implied again by Assumption 7.1 (iii). □

Remark 7.4.

1. Condition (i) of Assumption 7.1 is very restrictive, however it already proved extremely helpful in understanding high dimensional MCMC algorithms via diffusion limits [32, 37, 9], and conclusions drawn under (i) are empirically observed to be robust even if the condition is violated. It is essential to investigate its robustness also in the Gibbs sampler setting.
2. Minor generalisations to (i) are straightforward, e.g. our conclusions hold for $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$, where $\mathcal{X}_i = \mathbb{R}^k$.
3. Condition (iii) of Assumption 7.1 is required to ensure asymptotic validity of our algorithm by Theorem 6.2. We will report separately on ergodicity of adaptive random scan Gibbs samplers in the non-uniform case.

8. Proof of Proposition 3.2

The analysis of Example 3.1 is somewhat delicate since the process is both time and space inhomogeneous (as are most nontrivial adaptive MCMC algorithms). To establish Proposition 3.2, we will define a couple of auxiliary stochastic process. Consider the following one dimensional process $(\tilde{X}_n)_{n \geq 0}$ obtained from $(X_n)_{n \geq 0}$ by

$$\tilde{X}_n := X_{n,1} + X_{n,2} - 2.$$

Clearly $\tilde{X}_n - \tilde{X}_{n-1} \in \{-1, 0, 1\}$, moreover $X_{n,1} \to \infty$ and $X_{n,2} \to \infty$ if and only if $\tilde{X}_n \to \infty$. Note that the dynamics of $(\tilde{X}_n)_{n\geq 0}$ are also both time and space inhomogeneous.

We will also use an auxiliary random-walk-like space homogeneous process

$$S_0 = 0 \quad \text{and} \quad S_n := \sum_{i=1}^{n} Y_i, \quad \text{for } n \geq 1,$$

where $Y_1, Y_2, \ldots$ are independent random variables taking values in $\{-1, 0, 1\}$. Let the distribution of Y_n on $\{-1, 0, 1\}$ be

$$\nu_n \ := \ \left\{ \frac{1}{4} - \frac{1}{a_n} \ , \ \frac{1}{2} \ , \ \frac{1}{4} + \frac{1}{a_n} \right\}. \tag{47}$$

We shall couple $(\tilde{X}_n)_{n\geq 0}$ with $(S_n)_{n\geq 0}$, i.e. define them on the same probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$, by specifying the joint distribution of $(\tilde{X}_n, S_n)_{n\geq 0}$ so that the marginal distributions remain unchanged. We describe the details of the construction later. Now define

$$\Omega_{\tilde{X}\geq S} := \{\omega \in \Omega : \tilde{X}_n(\omega) \geq S_n(\omega) \ \text{ for every } \ n\} \tag{48}$$

and

$$\Omega_\infty := \{\omega \in \Omega : S_n(\omega) \to \infty\}. \tag{49}$$

Clearly, if $\omega \in \Omega_{\tilde{X}\geq S} \cap \Omega_\infty$, then $\tilde{X}_n(\omega) \to \infty$. In the sequel we show that for our coupling construction

$$\mathbb{P}(\Omega_{\tilde{X}\geq S} \cap \Omega_\infty) > 0. \tag{50}$$

We shall use the Hoeffding's inequality for $S_k^{k+n} := S_{k+n} - S_k$. Since $Y_n \in [-1, 1]$, it yields for every $t > 0$,

$$\mathbb{P}(S_k^{k+n} - \mathbb{E}S_k^{k+n} \leq -nt) \leq \exp\{-\frac{1}{2}nt^2\}. \tag{51}$$

Note that $\mathbb{E}Y_n = 2/a_n$ and thus $\mathbb{E}S_k^{k+n} = 2\sum_{i=k+1}^{k+n} 1/a_i$. The following choice for the sequence a_n will facilitate further calculations. Let

$$\begin{aligned}
b_0 &= 0, \\
b_1 &= 1000, \\
b_n &= b_{n-1}\Big(1 + \frac{1}{10 + \log(n)}\Big), \qquad \text{for} \quad n \geq 2 \\
c_n &= \sum_{i=0}^{n} b_n, \\
a_n &= 10 + \log(k), \qquad\qquad\qquad \text{for} \quad c_{k-1} < n \leq c_k.
\end{aligned}$$

Remark 8.1. To keep notation reasonable we ignore the fact that b_n will not be an integer. It should be clear that this does not affect proofs, as the constants we have defined, i.e. b_1 and a_1 are bigger then required.

Lemma 8.2. *Let Y_n and S_n be as defined above and let*

$$\Omega_1 := \Big\{\omega \in \Omega : S_k = k \quad \textit{for every} \quad 0 < k \leq c_1\Big\}. \tag{52}$$

$$\Omega_n := \Big\{\omega \in \Omega : S_k \geq \frac{b_{n-1}}{2} \quad \textit{for every} \quad c_{n-1} < k \leq c_n\Big\} \quad \textit{for} \;\; n \geq 2. \tag{53}$$

Then

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \Omega_n\Big) > 0. \tag{54}$$

Remark 8.3. Note that $b_n \nearrow \infty$ and therefore $\bigcap_{n=1}^{\infty} \Omega_n \subset \Omega_\infty$.

Proof. With positive probability, say $p_{1,S}$, we have $Y_1 = \cdots = Y_{1000} = 1$ which gives $S_{c_1} = 1000 = b_1$. Hence $\mathbb{P}(\Omega_1) = p_{1,S} > 0$. Moreover recall that $S^{c_n}_{c_{n-1}}$ is a sum of b_n i.i.d. random variables with $\mathbb{E}S^{c_n}_{c_{n-1}} = \frac{2b_n}{10+\log(n)}$. Therefore for every $n \geq 1$ by Hoeffding's inequality with $t = 1/(10 + \log(n))$, we can also write

$$\mathbb{P}\Big(S^{c_n}_{c_{n-1}} \leq \frac{b_n}{10 + \log(n)}\Big) \leq \exp\left\{ -\frac{1}{2}\frac{b_n}{(10 + \log(n))^2}\right\} =: p_n.$$

Therefore using the above bound iteratively we obtain

$$\mathbb{P}(S_{c_1} = b_1,\; S_{c_n} \geq b_n \;\text{ for every }\; n \geq 2) \;\geq\; p_{1,S} \prod_{n=2}^{\infty}(1 - p_n). \tag{55}$$

Now consider the minimum of S_k for $c_{n-1} < k \leq c_n$ and $n \geq 2$. The worst case is when the process S_k goes monotonically down and then monotonically up for $c_{n-1} < k \leq c_n$. By the choice of b_n, equation (55) implies also

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \Omega_n\Big) \geq p_{1,S} \prod_{n=2}^{\infty}(1 - p_n). \tag{56}$$

Clearly in this case

$$p_{1,S} \prod_{n=2}^{\infty}(1 - p_n) > 0 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \log(1 - p_n) > -\infty \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} p_n < \infty. \tag{57}$$

We conclude (57) by comparing p_n with $1/n^2$. We show that there exists n_0 such that for $n \geq n_0$ the series p_n decreases quicker then the series $1/n^2$ and therefore p_n is summable. We check that

$$\log \frac{p_{n-1}}{p_n} > \log \frac{n^2}{(n-1)^2} \qquad \text{for} \quad n \geq n_0. \tag{58}$$

Indeed

$$\begin{aligned}\log\frac{p_{n-1}}{p_n} &= -\frac{1}{2}\left(\frac{b_{n-1}}{(10+\log(n-1))^2}-\frac{b_n}{(10+\log(n))^2}\right)\\ &= \frac{b_{n-1}}{2}\left(\frac{11+\log(n)}{(10+\log(n))^3}-\frac{1}{(10+\log(n-1))^2}\right)\\ &= \frac{b_{n-1}}{2}\left(\frac{(11+\log(n))(10+\log(n-1))^2-(10+\log(n))^3}{(10+\log(n))^3(10+\log(n-1))^2}\right).\end{aligned}$$

Now recall that b_{n-1} is an increasing sequence. Moreover the enumerator can be rewritten as

$$(10+\log(n))\Big((10+\log(n-1))^2-(10+\log(n))^2\Big)+(10+\log(n-1))^2,$$

now use $a^2-b^2=(a+b)(a-b)$ to identify the leading term $(10+\log(n-1))^2$. Consequently there exists a constant C and $n_0\in\mathbb{N}$ s.t. for $n\geq n_0$

$$\log\frac{p_{n-1}}{p_n} \;\geq\; \frac{C}{(10+\log(n))^3} \;>\; \frac{2}{n-1} \;>\; \log\frac{n^2}{(n-1)^2}.$$

Hence $\sum_{n=1}^{\infty}p_n<\infty$ follows. □

Now we will describe the coupling construction of $(\tilde{X}_n)_{n\geq 0}$ and $(S_n)_{n\geq 0}$. We already remarked that $\bigcap_{n=1}^{\infty}\Omega_n\subset\Omega_\infty$. We will define a coupling that implies also

$$\mathbb{P}\left(\left(\bigcap_{n=1}^{\infty}\Omega_n\right)\cap\Omega_{\tilde{X}\geq S}\right)\geq C\mathbb{P}\left(\bigcap_{n=1}^{\infty}\Omega_n\right)\qquad\text{for some universal}\quad C>0, \tag{59}$$

and therefore

$$\mathbb{P}\left(\Omega_{\tilde{X}\geq S}\cap\Omega_\infty\right)>0. \tag{60}$$

Thus nonergodicity of $(X_n)_{n\geq 0}$ will follow from Lemma 8.2. We start with the following observation.

Lemma 8.4. *There exists a coupling of $\tilde{X}_n-\tilde{X}_{n-1}$ and Y_n, such that*

(a) For every $n\geq 1$ and every value of $\tilde{X}_{n-1}$

$$\mathbb{P}(\tilde{X}_n-\tilde{X}_{n-1}=1,Y_n=1)\geq\mathbb{P}(\tilde{X}_n-\tilde{X}_{n-1}=1)\mathbb{P}(Y_n=1), \tag{61}$$

(b) Write even or odd $\tilde{X}_{n-1}$ as $\tilde{X}_{n-1}=2i-2$ or $\tilde{X}_{n-1}=2i-3$ respectively. If $2i-8\geq a_n$ then the following implications hold a.s.

$$Y_n=1 \quad\Rightarrow\quad \tilde{X}_n-\tilde{X}_{n-1}=1 \tag{62}$$

$$\tilde{X}_n-\tilde{X}_{n-1}=-1 \quad\Rightarrow\quad Y_n=-1. \tag{63}$$

Proof. Property *(a)* is a simple fact for any two $\{-1,0,1\}$ valued random variables Z and Z' with distributions say $\{d_1,d_2,d_3\}$ and $\{d'_1,d'_2,d'_3\}$. Assign $\mathbb{P}(Z=Z'=1):=\min\{d_3,d'_3\}$ and *(a)* follows. To establish *(b)* we analyse the dynamics of $(X_n)_{n\geq 0}$ and consequently of $(\tilde{X}_n)_{n\geq 0}$. Recall Algorithm 2.2 and the update rule for α_n in (4). Given $X_{n-1}=(i,j)$, the algorithm will obtain the value of α_n in step 1, next draw a coordinate according to $(\alpha_{n,1},\alpha_{n,2})$ in step 2. In steps 3 and 4 it will move according to conditional distributions for updating the first or the second coordinate. These distributions are

$$(1/2,1/2) \qquad \text{and} \qquad \left(\frac{i^2}{i^2+(i-1)^2},\frac{(i-1)^2}{i^2+(i-1)^2}\right)$$

respectively. Hence given $X_{n-1}=(i,i)$ the distribution of $X_n\in\{(i,i-1),(i,i),(i+1,i)\}$ is

$$\left((\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2}\ ,\ 1-(\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2}-(\frac{1}{4}+\frac{2}{a_n})\ ,\ \frac{1}{4}+\frac{2}{a_n}\right), \quad (64)$$

whereas if $X_{n-1}=(i,i-1)$ then $X_n\in\{(i-1,i-1),(i,i-1),(i,i)\}$ with probabilities

$$\left(\frac{1}{4}-\frac{2}{a_n}\ ,\ 1-(\frac{1}{4}-\frac{2}{a_n})-(\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2}\ ,\ (\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2}\right), \quad (65)$$

respectively. We can conclude the evolution of $(\tilde{X}_n)_{n\geq 0}$. Namely, if $\tilde{X}_{n-1}=2i-2$ then the distribution of $\tilde{X}_n-\tilde{X}_{n-1}\in\{-1,0,1\}$ is given by (64) and if $\tilde{X}_{n-1}=2i-3$ then the distribution of $\tilde{X}_n-\tilde{X}_{n-1}\in\{-1,0,1\}$ is given by (65). Let $\leq_{\mathrm{st}}$ denote stochastic ordering. By simple algebra both measures defined in (64) and (65) are stochastically bigger then

$$\mu^i_n \quad = \quad (\mu^i_{n,1},\mu^i_{n,2},\mu^i_{n,3}), \quad (66)$$

where

$$\mu^i_{n,1} \quad = \quad (\frac{1}{4}-\frac{2}{a_n})(1+\frac{2}{i}) \quad = \quad \frac{1}{4}-\frac{1}{a_n}-\frac{2i+8-a_n}{2ia_n}, \quad (67)$$

$$\mu^i_{n,2} \quad = \quad 1-(\frac{1}{4}-\frac{2}{a_n})(1+\frac{2}{i})-(\frac{1}{4}+\frac{2}{a_n})(1-\frac{2}{\max\{4,i\}}),$$

$$\mu^i_{n,3} \quad = \quad (\frac{1}{4}+\frac{2}{a_n})(1-\frac{2}{\max\{4,i\}}) \quad = \quad \frac{1}{4}+\frac{1}{a_n}+\frac{2\max\{4,i\}-8-a_n}{2a_n\max\{4,i\}}. \quad (68)$$

Recall ν_n, the distribution of Y_n defined in (47). Examine (67) and (68) to see that if $2i-8\geq a_n$, then $\mu^i_n\geq_{\mathrm{st}}\nu_n$. Hence in this case also the distribution of $\tilde{X}_n-\tilde{X}_{n-1}$ is stochastically bigger then the distribution of Y_n. The joint probability distribution of $(\tilde{X}_n-\tilde{X}_{n-1},Y_n)$ satisfying (62) and (63) follows. □

Proof of Proposition 3.2. Define

$$\Omega_{1,\tilde{X}}:=\Big\{\omega\in\Omega:\tilde{X}_n-\tilde{X}_{n-1}=1 \quad \text{for every} \quad 0<n\leq c_1\Big\}. \quad (69)$$

Since the distribution of $\tilde{X}_n - \tilde{X}_{n-1}$ is stochastically bigger then μ_n^i defined in (66) and $\mu_n^i(1) > c > 0$ for every i and n,

$$\mathbb{P}\big(\Omega_{1,\tilde{X}}\big) =: p_{1,\tilde{X}} > 0.$$

By Lemma 8.4 *(a)* we have

$$\mathbb{P}\big(\Omega_{1,\tilde{X}} \cap \Omega_1\big) \geq p_{1,S}\, p_{1,\tilde{X}} > 0. \tag{70}$$

Since $S_{c_1} = \tilde{X}_{c_1} = c_1 = b_1$, on $\Omega_{1,\tilde{X}} \cap \Omega_1$, the requirements for Lemma 8.4 *(b)* hold for $n-1 = c_1$. We shall use Lemma 8.4 *(b)* iteratively to keep $\tilde{X}_n \geq S_n$ for every n. Recall that we write $\tilde{X}_{n-1}$ as $\tilde{X}_{n-1} = 2i-2$ or $\tilde{X}_{n-1} = 2i-3$. If $2i - 8 \geq a_n$ and $\tilde{X}_{n-1} \geq S_{n-1}$ then by Lemma 8.4 *(b)* also $\tilde{X}_n \geq S_n$. Clearly if $\tilde{X}_k \geq S_k$ and $S_k \geq \frac{b_{n-1}}{2}$ for $c_{n-1} < k \leq c_n$ then $\tilde{X}_k \geq \frac{b_{n-1}}{2}$ for $c_{n-1} < k \leq c_n$, hence

$$2i - 2 \geq \frac{b_{n-1}}{2} \quad \text{for} \quad c_{n-1} < k \leq c_n.$$

This in turn gives $2i - 8 \geq \frac{b_{n-1}}{2} - 6$ for $c_{n-1} < k \leq c_n$ and since $a_k = 10 + \log(n)$, for the iterative construction to hold, we need $b_n \geq 32 + 2\log(n+1)$. By the definition of b_n and standard algebra we have

$$b_n \geq 1000 \left(1 + \sum_{i=2}^{n} \frac{1}{10 + \log(n)}\right) \geq 32 + 2\log(n+1) \quad \text{for every} \;\; n \geq 1.$$

Summarising the above argument provides

$$\begin{aligned}
\mathbb{P}(X_{n,1} \to \infty) &\geq \mathbb{P}\left(\Omega_\infty \cap \Omega_{\tilde{X} \geq S}\right) \geq \mathbb{P}\left(\left(\bigcap_{n=1}^{\infty} \Omega_n\right) \cap \Omega_{\tilde{X} \geq S}\right) \\
&\geq \mathbb{P}\left(\Omega_{1,\tilde{X}} \cap \left(\bigcap_{n=1}^{\infty} \Omega_n\right) \cap \Omega_{\tilde{X} \geq S}\right) \\
&\geq p_{1,\tilde{X}} p_{1,S} \prod_{n=2}^{\infty} (1 - p_n) > 0.
\end{aligned}$$

Hence $(X_n)_{n \geq 0}$ is not ergodic, and in particular $\|\pi_n - \pi\|_{\mathrm{TV}} \nrightarrow 0$. □

Acknowledgements

This paper was written while the first author was a postdoctoral fellow at the Department of Statistics, University of Toronto. Both authors were partially funded by NSERC of Canada.

References

[1] C. Andrieu and E. Moulines (2006): On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. *Ann. Appl. Probab.* 16(3), 1462–1505.

[2] Y. Atchadé and G. Fort (2008): Limit Theorems for some adaptive MCMC algorithms with sub-geometric kernels. *Bernoulli, to appear.*
[3] Y. Atchadé, G. Fort, E. Moulines, and P. Priouret (2009): Adaptive Markov Chain Monte Carlo: Theory and Methods. *Preprint.*
[4] Y.F. Athadé, G.O. Roberts, and J.S. Rosenthal, (2009): Optimal Scaling of Metropolis-coupled Markov Chain Monte Carlo. *Preprint.*
[5] Y.F. Atchadé and J.S. Rosenthal (2005): On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli* 11, 815–828.
[6] Y. Bai (2009): Simultaneous drift conditions for Adaptive Markov Chain Monte Carlo algorithms. *Preprint.*
[7] Y. Bai (2009): An Adaptive Directional Metropolis-within-Gibbs algorithm. *Preprint.*
[8] Y. Bai, G.O. Roberts, J.S. Rosenthal (2009): On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms. *Preprint.*
[9] M. Bédard (2007): Weak Convergence of Metropolis Algorithms for Non-iid Target Distributions. *Ann. Appl. Probab.* 17, 1222–44.
[10] M. Bédard (2008): Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and their Applications*, 118(12), 2198–2222.
[11] W. Bednorz, R. Latała and K. Łatuszyński (2008): A Regeneration Proof of the Central Limit Theorem for Uniformly Ergodic Markov Chains. *Elect. Comm. in Probab.* 13, 85–98.
[12] L. Bottolo, S. Richardson, and J.S. Rosenthal (2010): Bayesian models for sparse regression analysis of high dimensional data. In preparation.
[13] A.E. Brockwell and J.B. Kadane (2005): Identification of Regeneration Times in MCMC Simulation, with Application to Adaptive Schemes. *Journal of Computational and Graphical Statistics,* 14, 436–458.
[14] R.V. Craiu, J.S. Rosenthal, and C. Yang (2008): Learn From Thy Neighbor: Parallel-Chain and Regional Adaptive MCMC. *J. Amer. Stat. Assoc., to appear.*
[15] P. Diaconis, K. Khare, and L. Saloff-Coste (2008): Gibbs sampling, exponential families and orthogonal polynomials (with discussion and rejoinder). *Statistical Science* 23(2), 151–178.
[16] W.R. Gilks, G.O. Roberts, and S.K. Sahu (1998): Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* 93(443), 1045–1054.
[17] H. Haario, E. Saksman, and J. Tamminen (2001): An adaptive Metropolis algorithm. *Bernoulli* 7, 223–242.
[18] H. Haario, E. Saksman, and J. Tamminen (2005): Componentwise adaptation for high dimensional MCMC. *Computational Statistics* 20, 265–273.
[19] O. Häggström and J.S. Rosenthal (2007): On Variance Conditions for Markov Chain CLTs. *Elect. Comm. in Probab.* 12, 454-464.
[20] C. Kipnis, S.R.S. Varadhan, (1986): Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions. *Commun. Math. Phys.* 104, 1–19.
[21] K. Łatuszyński (2008): Regeneration and Fixed-Width Analysis of

Markov Chain Monte Carlo Algorithms. PhD Dissertation. Available at: arXiv:0907.4716v1

[22] R.A. Levine (2005): A note on Markov chain Monte Carlo sweep strategies. Journal of Statistical Computation and Simulation 75(4), 253–262.

[23] R.A. Levine, Z. Yu, W.G. Hanley, and J.A. Nitao (2005): Implementing Random Scan Gibbs Samplers. Computational Statistics 20, 177–196.

[24] R.A. Levine and G. Casella (2006): Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis* 97, 2071–2100.

[25] J.S. Liu (2001): *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

[26] J.S. Liu, W.H. Wong, and A. Kong (1995): Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *J. Roy. Stat. Soc. B* 57(1), 157–169.

[27] K.L. Mengersen and R.L. Tweedie (1996): Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* 24, 1, 101–121.

[28] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

[29] S.P. Meyn and R.L. Tweedie (1993): *Markov Chains and Stochastic Stability.* Springer-Verlag, London. Available at: probability.ca/MT

[30] O. Papaspiliopoulos and G.O. Roberts (2008): Stability of the Gibbs sampler for Bayesian hierarchical models. *Annals of Statistics* 36(1), 95–117.

[31] C.P. Robert and G. Casella (2004): *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

[32] G.O. Roberts, A. Gelman, and W.R. Gilks (1997): Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* 7, 110–120.

[33] G.O. Roberts and N.G. Polson (1994): On the geometric convergence of the Gibbs sampler. *J. R. Statist. Soc. B* 56(2), 377–384.

[34] G.O. Roberts and J.S. Rosenthal (1997): Geometric ergodicity and hybrid Markov chains. *Elec. Comm. Prob.* 2 (2).

[35] G.O. Roberts and J.S. Rosenthal (1998): Two convergence properties of hybrid samplers. *Ann. Appl. Prob.* 8(2), 397–407.

[36] G.O. Roberts and J.S. Rosenthal (1998): Optimal scaling of discrete approximations to Langevin diffusions. *J. Roy. Stat. Soc. B* 60, 255268.

[37] G.O. Roberts and J.S. Rosenthal (2001): Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.* 16, 351–367.

[38] G.O. Roberts and J.S. Rosenthal (2004): General state space Markov chains and MCMC algorithms. *Probability Surveys* 1, 20–71.

[39] G.O. Roberts and J.S. Rosenthal (2007): Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.,* 44, 458–475.

[40] G.O. Roberts and J.S. Rosenthal (2006): Examples of Adaptive MCMC *J. Comp. Graph. Stat.* 18(2), 349–367.

[41] J.S. Rosenthal (2008): Optimal Proposal Distributions and Adaptive MCMC. *Preprint.*

[42] E. Saksman and M. Vihola (2008): On the Ergodicity of the Adaptive

Metropolis Algorithm on Unbounded Domains. *Preprint.*

[43] E. Turro, N. Bochkina, A.M.K. Hein, and S. Richardson (2007): BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. BMC Bioinformatics 8, 439–448. Available at: http://www.biomedcentral.com/1471-2105/8/439

[44] M. Vihola (2009): On the Stability and Ergodicity of an Adaptive Scaling Metropolis Algorithm. *Preprint.*

[45] C. Yang (2008): On The Weak Law Of Large Numbers For Unbounded Functionals For Adaptive MCMC. *Preprint.*

[46] C. Yang (2008): Recurrent and Ergodic Properties of Adaptive MCMC. *Preprint.*