Adaptive Gibbs samplers and related MCMC methods

Krzysztof Łatuszyński, Gareth O. Roberts and Jeffrey S. Rosenthal

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

G. O. Roberts
Department of Statistics
University of Warwick
CV4 7AL, Coventry, UK
e-mail: Gareth.O.Roberts@warwick.ac.uk

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

(January 27, 2011; revised June 29, 2011)

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. [30, 40]. Metropolis algorithms and Gibbs samplers (to be described in Section 2) are among the most common MCMC algorithms, c.f. [33, 26, 40].

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 [34, 38, 9, 10, 4, 39, 40, 43] which provide "prescriptions" of how to do this, though they typically depend on unknown characteristics of π.

For random scan Gibbs and Metropolis within 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 [27, 23, 24, 15, 45, 11] 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 provide virtually no help in exploring the state space, 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. (Of course, as with any Gibbs sampler, repeatedly updating one coordinate does not help with *distributional* convergence, it only helps with sampling the entire state space to produce good estimates.) Thus, an efficient Gibbs sampler for this example 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. [11]), 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).

Such adaptive algorithms are only valid if their ergodicity can be established. Unfortunately 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. [17, 18, 5, 1, 19, 12, 41, 42, 22, 47, 48, 13, 8, 6, 7, 44, 46, 2, 3, 14]) which includes counterintuitive examples where X_n fails to converge to the desired distribution π (c.f. [5, 41, 8, 22]), 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 [25], 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 6) and provides a counter-example to Theorem 2.1 of [25]. Next, we establish various positive results for ergodicity of adaptive Gibbs samplers. We consider adaptive random scan Gibbs samplers (`AdapRSG`) which update coordinate selection probabilities as the simulation progresses; adaptive random scan Metropolis-within-Gibbs samplers (`AdapRSMwG`) which update coordinate selection probabilities as the simulation progresses; and adaptive random scan adaptive Metropolis-within-Gibbs samplers (`AdapRSadapMwG`) that update coordinate selection probabilities as well as proposal distributions for the Metropolis steps. Positive results in the uniform setting are discussed in Section 4, whereas Section 5 deals with the nonuniform setting. 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.

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* $i \in \{1, \dots, d\}$ *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* $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$.

Whereas the standard approach is to choose the coordinate i at the first step uniformly at random, which corresponds to $\alpha = (1/d, \dots, 1/d)$, 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 [27] and also observed empirically e.g. in Bayesian variable selection for linear models in statistical genetics [45, 11].

Throughout the paper we denote the transition kernel of a random scan Gibbs sampler with selection probabilities α as P_α and the transition kernel of a single Gibbs update of coordinate i is denoted as P_i, hence $P_\alpha = \sum_{i=1}^d \alpha_i P_i$.

We consider a class of adaptive random scan Gibbs samplers where selection probabilities $\alpha = (\alpha_1, \dots, \alpha_d)$ 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 $X_{n-1}, \dots, X_0$, and current selection probabilities $\alpha_{n-1} = (\alpha_{n-1,1}, \dots, \alpha_{n-1,d})$ amounts to the following steps, where $R_n(\cdot)$ is some update rule for α_n.

Algorithm 2.2 (`AdapRSG`).

1. *Set* $\alpha_n := R_n(\alpha_0, \dots, \alpha_{n-1}, X_{n-1}, \dots, X_0) \in \mathcal{Y}$
2. *Choose coordinate* $i \in \{1, \dots, d\}$ *according to selection probabilities* α_n
3. *Draw* $Y \sim \pi(\cdot|X_{n-1,-i})$
4. *Set* $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$

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. [41, 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 or Metropolis-Hastings step [29, 20]. 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 [29] 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. [19, 42], 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* $\alpha_n := R_n(\alpha_0,\ldots,\alpha_{n-1},X_{n-1},\ldots,X_0)\in\mathcal{Y}$
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 Sections 4.2 and 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 [45, 11]. Versions of such algorithms with fixed selection probabilities are considered by e.g. [19] and [42]. 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 $\gamma_n = (\gamma_{n,1},\ldots,\gamma_{n,d})$ 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* $\alpha_n := R_n(\alpha_0,\ldots,\alpha_{n-1},X_{n-1},\ldots,X_0,\gamma_{n-1},\ldots,\gamma_0)\in\mathcal{Y}$

2. *Set* $\gamma_n := R'_n(\alpha_0, \ldots, \alpha_{n-1}, X_{n-1}, \ldots, X_0, \gamma_{n-1}, \ldots, \gamma_0) \in \Gamma_1 \times \ldots \times \Gamma_n$
3. *Choose coordinate* $i \in \{1, \ldots, d\}$ *according to selection probabilities* α*, i.e. with* $\mathbb{P}(i = j) = \alpha_j$
4. *Draw* $Y \sim Q_{X_{n-1,-i}, \gamma_{n-1,i}}(X_{n-1,i}, \cdot)$
5. *With probability given by (3),*

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

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 Sections 4.3 and 5 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 [25] 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 [25], 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 $\mathbb{N} = \{1, 2, \ldots\}$, 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:

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

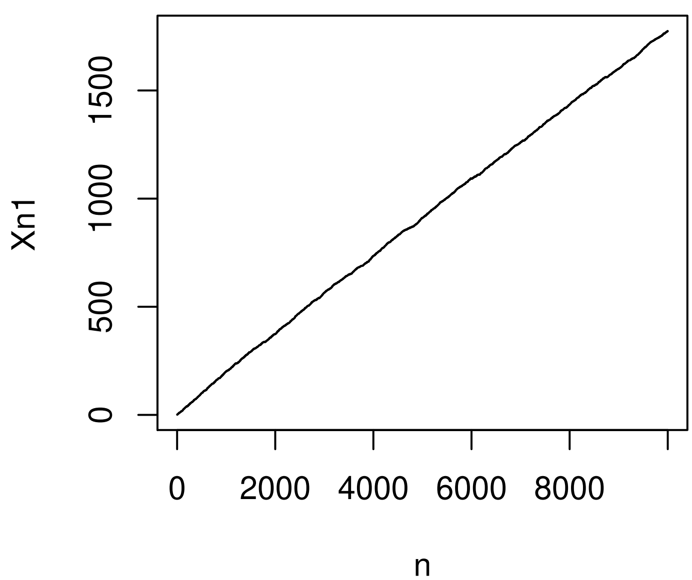
for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$.

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. [30, 40]). 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,* $||\pi_n - \pi||_{TV} \nrightarrow 0$.

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 6. We also simulated Example 3.1 on a computer (with the (a_n) as defined in Section 6), 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 - the uniform case

We now present positive results about ergodicity of adaptive Gibbs samplers under various assumptions. Results of this section 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. [30, 40] 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 the results of this section may be sufficient for a pragmatical user. The nonuniform case is considered in the following Section 5.

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 assume that all our selection probabilities are in this set $\mathcal{Y}$.

Remark 4.1. The above assumption may seem constraining, it is however irrelevant in practice. The additional computational effort on top of the unknown optimal strategy α^* (that may be in $\Delta_{d-1} - \mathcal{Y}$) is easily controlled by setting $\varepsilon := (Kd)^{-1}$ that effectively upperbounds it by $1/K$. The argument can be easily made rigorous e.g. in terms of the total variation distance or the asymptotic variance.

4.1. Adaptive random scan Gibbs samplers

The main result of this section is the following.

Theorem 4.2. *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 4.3.

1. Assumption *(b)* will typically be verified for $\beta = (1/d, \dots, 1/d)$; see also Proposition 4.8 below.
2. We expect that most adaptive random scan Gibbs samplers will be designed so that $|\alpha_n - \alpha_{n-1}| \le a_n$ for every $n \ge 1$, $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$, and $\omega \in \Omega$, for some deterministic sequence $a_n \to 0$ (which holds for e.g. the adaptations considered in [11]). 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 $\tilde{\mathcal{X}} = \mathcal{X} \cap (\{1, \dots, M\} \times \{1, \dots, M\})$ for some $1 < M < \infty$,, then the corresponding adaptive Gibbs sampler is ergodic, and (7) holds.

Before we proceed with the proof of Theorem 4.2, we need some preliminary lemmas, which may be of independent interest.

Lemma 4.4. *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.* $\sup_{x_0 \in \mathcal{X}, \alpha \in \mathcal{Y}} T(x_0, \alpha, n) \le M\rho^n \to 0$.

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 [30] 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 $x \in \mathcal{X}$,

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

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

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

Since $\beta \in \mathcal{Y}$, we have $1 \ge r \ge \frac{\varepsilon}{1-(d-1)\varepsilon} > 0$ 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

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

By Theorem 8 of [40] condition (9) implies

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

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

Lemma 4.5. *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*

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

Proof. Let $\delta := |\alpha - \alpha'|$. Then $r := \min_i \frac{\alpha'_i}{\alpha_i} \ge \frac{\varepsilon}{\varepsilon + \max_i |\alpha_i - \alpha'_i|} \ge \frac{\varepsilon}{\varepsilon+\delta}$ and reasoning as in the proof of Lemma 4.4 we can write $P_{\alpha'} = rP_\alpha + (1-r)P_q$ for some q and compute

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

as claimed. □

Corollary 4.6. *$P_\alpha(x,B)$ as a function of α on $\mathcal{Y}$ is Lipschitz with Lipschitz constant $1/\varepsilon$ for every fixed set $B \in \mathcal{B}(\mathcal{X})$.*

Corollary 4.7. *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 4.2. We conclude the result from Theorem 1 of [41] that requires simultaneous uniform ergodicity and diminishing adaptation. Simultaneous uniform ergodicity results from combining assumption (b) and Lemma 4.4. Diminishing adaptation results from assumption (a) with Corollary 4.7. Moreover note that Lemma 4.4 is uniform in x_0 and α_0 and (a') yields uniformly diminishing adaptation again by Corollary 4.7. A look into the proof of Theorem 1 [41] reveals that this suffices for the uniform part of Theorem 4.2. □

Finally, we note that verifying uniform ergodicity of a random scan Gibbs sampler, as required by assumption (b) of Theorem 4.2, may not be straightforward. Such issues have been investigated in e.g. [35] and more recently in relation to the parametrization of hierarchical models (see [32] 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). See also Theorem 2 of [31] for a related result.

Proposition 4.8. *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 4.4 it follows that `RSG(`α`)` is uniformly ergodic for any $\alpha \in \mathcal{Y}$. □

4.2. Adaptive random scan Metropolis-within-Gibbs

In this section we consider random scan Metropolis-within-Gibbs sampler algorithms (see also Section 5 for the nonuniform case). 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 4.9. *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.*

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

We have the following counterpart of Theorem 4.2.

Theorem 4.10. *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 4.9 holds.

Then `AdapRSMwG` is ergodic, i.e.

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

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.*

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

Remark 4.11. Remarks 4.3.1–4.3.3 still apply. Also, assumption 4.9 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. [28].
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. [30, 40].

To prove Theorem 4.10 we build on the approach of [37]. In particular recall the following notions of reversibility and of strong uniform ergodicity.

Definition 4.12. We say that a transition kernel P on $\mathcal{X}$ is reversible with respect to its stationary distribution π, if for any $A, B \in \mathcal{B}(\mathcal{X})$

$$\int_A P(x, B)\pi(\mathrm{d}x) = \int_B P(y, A)\pi(\mathrm{d}y).$$

Definition 4.13. 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

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

By Proposition 1 in [37], 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 4.14. *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*

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

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 [40] 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

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

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

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

Moreover by reversibility

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

and consequently

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

Now define

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

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

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

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

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

And with k defined above by (16) we have

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

This completes the proof. □

We will need the following generalization of Lemma 4.4.

Lemma 4.15. *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 [37] it is $(m,s_1)-$strongly uniformly ergodic for some m and s_1. Therefore, and arguing as in the proof of Lemma 4.4, c.f. (9), there exist $s_2\ge\big(\frac{\varepsilon}{1-(d-1)\varepsilon}\big)^m$, s.t. for every $\alpha\in\mathcal{Y}$ and every $x\in\mathcal{X}$

$$P_\alpha^m(x,\cdot)\ge s_2P_\beta^m(x,\cdot)\ge s_1s_2\pi(\cdot). \tag{17}$$

Set $m'=m$ and $s'=s_1s_2$. □

Proof of Theorem 4.10. We proceed as in the proof of Theorem 4.2, i.e. establish diminishing adaptation and simultaneous uniform ergodicity and conclude (12) and (13) from Theorem 1 of [41]. Observe that Lemma 4.5 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 4.7. To establish simultaneous uniform ergodicity, observe that by Assumption 4.9 and Lemma 4.14 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 4.15 the family `RSG(`α`)`, $\alpha \in \mathcal{Y}$ is $(m', s')-$strongly uniformly ergodic, therefore by Theorem 2 of [37] 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 [37]. □

We close this section with the following alternative version of Theorem 4.10.

Theorem 4.16. *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.

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

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.

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

Proof. Diminishing adaptation results from assumption (a) and Corollary 4.7. Simultaneous uniform ergodicity can be established as in the proof of Lemma 4.4. The claim follows from Theorem 1 of [41]. □

Remark 4.17. Whereas the statement of Theorem 4.16 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 4.10.

4.3. Adaptive random scan adaptive Metropolis-within-Gibbs

In this section, and also later in Section 5, 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. [11] 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 [34, 38, 9, 10, 4, 39, 40, 42, 43].

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 4.9 must be reformulated accordingly, as follows.

Assumption 4.18. *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.*

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

We have the following counterpart of Theorems 4.2 and 4.10.

Theorem 4.19. *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 $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$ and $\gamma_0 \in \Gamma$.*
(b) there exists $\beta \in \mathcal{Y}$ s.t. `RSG(`β`)` *is uniformly ergodic.*
(c) Assumption 4.18 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

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

for fixed starting values $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$ and γ_0.

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

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

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 4.20. Remarks 4.3.1–4.3.3 still apply. And, Remark 4.11 applies for verifying Assumption 4.18. 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 [41].

To establish simultaneous uniform ergodicity we proceed as in the proof of Theorem 4.10. Observe that by Assumption 4.18 and Lemma 4.14 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 4.15 the family $\texttt{RSG}(\alpha)$, $\alpha \in \mathcal{Y}$ is (m', s')−strongly uniformly ergodic, therefore by Theorem 2 of [37] 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 [37].

For diminishing adaptation we write

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

The first term above converges to 0 in probability by Corollary 4.7 and assumption (a). The second term

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

converges to 0 in probability as a mixture of terms that converge to 0 in probability. $\square$

The following lemma can be used to verify assumption (d) of Theorem 4.19; see also Example 4.22 below.

Lemma 4.21. *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*

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

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 4.19.

Proof. The first statement con be concluded from Proposition 12.3 of [1], however to be self-contained we provide the argument. 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

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

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

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

For the second statement note that condition (ii) implies 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 $\alpha_i(x,y) = \min\left\{1, \frac{\pi(y)q_i(y,x)}{\pi(x)q_i(x,y)}\right\}$. Indeed for any $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ using (22) we have

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

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

We now provide an example to show that diminishing adaptation of proposals as in Lemma 4.21 does not necessarily imply condition (d) of Theorem 4.19, so some additional assumption is required, e.g. (i) or (ii) of Lemma 4.21.

Example 4.22. Consider a sequence of Metropolis algorithms with transition kernels $P_1, P_2, \ldots$ 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

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

5. Ergodicity - nonuniform case

In this section we consider the case where nonadaptive kernels are not necessary uniformly ergodic. We study adaptive random scan Gibbs adaptive Metropolis within Gibbs (`AdapRSadapMwG`) algorithms in the nonuniform setting, with parameters $\alpha \in \mathcal{Y}$ and $\gamma_i \in \Gamma_i, i = 1, \dots, d$, subject to adaptation. The conclusions we draw apply immediately to adaptive random scan Gibbs Metropolis within Gibbs (`AdapRSMwG`) algorithms by keeping the parameters γ_i fixed for the Metropolis within Gibbs steps.

We keep the assumption that selection probabilities are in $\mathcal{Y}$ defined in (5), whereas the uniform ergodicity assumption will be replaced by some natural regularity conditions on the target density.

Our strategy is to use the generic approach of [41] and to verify the diminishing adaptation and the containment conditions. The containment condition has been extensively studied in [8] and it is essentially necessary for ergodicity of adaptive chains, see Theorem 2 therein for the precise result. In particular containment is implied by simultaneous geometrical ergodicity for the adaptive kernels. More precisely, we shall use the following result of [8].

Theorem 5.1 (Corollary 2 of [8])**.** *Consider the family $\{P_\gamma : \gamma \in \Gamma\}$ of Markov chains on $\mathcal{X} \subseteq \mathbb{R}^d$, satisfying the following conditions*

(i) for any compact set $C \in \mathcal{B}(\mathcal{X})$, there exist some integer $m > 0$, and real $\rho > 0$, and a probability measure ν_γ on C s.t.

$$P_\gamma^m(x,\cdot) \ \geq \ \rho\nu_\gamma(\cdot) \quad \textit{for all} \ \ x \in C,$$

(ii) there exists a function $V : \mathcal{X} \to (1, \infty)$, s.t. for any compact set $C \in \mathcal{B}(\mathcal{X})$, we have $\sup_{x \in C} V(x) < \infty$, $\pi(V) < \infty$, and

$$\limsup_{|x| \to \infty} \sup_{\gamma \in \Gamma} \frac{P_\gamma V(x)}{V(x)} < 1,$$

then for any adaptive strategy using $\{P_\gamma : \gamma \in \Gamma\}$, containment holds.

Throughout this section we assume $\mathcal{X}_i = \mathbb{R}$ for $i = 1, ..., d$, and $\mathcal{X} = \mathbb{R}^d$ and let μ_k denote the Lebsque measure on $\mathbb{R}^k$. By $\{e_1, ..., e_d\}$ denote the coordinate unit vectors and let $|\cdot|$ be the Euclidean norm.

Our focus is on random walk Metropolis proposals with symmetric densities for updating $X_i | X_{-i}$ denoted as $q_{i,\gamma_i}(\cdot)$, $\gamma_i \in \Gamma_i$. We shall work in the following setting, extensively studied for nonadaptive Metropolis within Gibbs algorithms in [16], see also [37] for related work and [21] for analysis of the random walk Metropolis algorithm.

Assumption 5.2. *The target distribution π is absolutely continuous with respect to μ_d with strictly positive and continuous density $\pi(\cdot)$ on $\mathcal{X}$.*

Assumption 5.3. *The family $\{q_{i,\gamma_i}\}_{1 \le i \le d; \gamma_i \in \Gamma_i}$ of symmetric proposal densities with respect to μ_1 (one-dimensional Lebesgue measure) is such that there exist constants $\eta_i > 0$, $\delta_i > 0$, for $i = 1, ..., d$, s.t.*

$$\inf_{|x| \le \delta_i} q_{i,\gamma_i}(x) \ge \eta_i \qquad \text{for every} \quad 1 \le i \le d \quad \text{and} \quad \gamma_i \in \Gamma_i. \tag{23}$$

Assumption 5.4. *There exist $0 < \delta < \Delta \le \infty$, such that*

$$\xi := \inf_{1 \le i \le d, \gamma_i \in \Gamma_i} \int_\delta^\Delta q_{i,\gamma_i}(y) \mu_1(dy) \quad > \quad 0, \tag{24}$$

and, for any sequence $x = \{x^j\}$ with $\lim_{j \to \infty} |x^j| = +\infty$ there exists a subsequence $\tilde{x} = \{\tilde{x}^j\}$ s.t. for some $i \in \{1, \ldots, d\}$ and all $y \in [\delta, \Delta]$,

$$\lim_{j \to \infty} \frac{\pi(\tilde{x}^j)}{\pi(\tilde{x}^j - sign(\tilde{x}_i^j) y e_i)} = 0 \quad \text{and} \quad \lim_{j \to \infty} \frac{\pi(\tilde{x}^j + sign(\tilde{x}_i^j) y e_i)}{\pi(\tilde{x}^j)} = 0. \tag{25}$$

Discussion of the seemingly involved 5.4 and simple criterions for checking it are given in [16]. It was shown in [16] that under these assumptions nonadaptive random scan Metropolis withing Gibbs algorithms are geometrically ergodic for subexponential densities. We establish ergodicity of the doubly adaptive `AdapRSadapMwG` algorithm in the same setting.

Theorem 5.5. *Let π be a subexponential density and let the selection probabilities $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Moreover assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ in probability for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$, $\gamma_i \in \Gamma_i$, $i = 1, \ldots, d$;

(b) 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 $\alpha_0 \in \mathcal{Y}$, $\gamma_i \in \Gamma_i$, $i = 1, \dots, d$;

(c) Assumptions 5.2, 5.3, 5.4 hold.

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

$$T(x_0, \alpha_0, \gamma_0, n) \to 0 \qquad as \quad n \to \infty. \tag{26}$$

Before proving this result we state its counterpart for densities that are log-concave in the tails. This is another typical setting carefully studied in the context of geometric ergodicity of nonadaptive chains ([16, 37, 28]) where Assumption 5.4 is replaced by the following two conditions.

Assumption 5.6. *There exists an $\phi > 0$ and δ s.t. $1/\phi \leq \delta < \Delta \leq \infty$ and, for any sequence $x := \{x^j\}$ with $\lim_{j\to\infty} |x^j| = +\infty$, there exists a subsequence $\tilde{x} := \{\tilde{x}^j\}$ s.t. for some $i \in \{1, \dots, d\}$ and for all $y \in [\delta, \Delta]$,*

$$\lim_{j\to\infty} \frac{\pi(\tilde{x}^j)}{\pi(\tilde{x}^j - sign(\tilde{x}^j_i) y e_i)} \leq \exp\{-\phi y\} \qquad and$$
$$\lim_{j\to\infty} \frac{\pi(\tilde{x}^j + sign(\tilde{x}^j_i) y e_i)}{\pi(\tilde{x}^j)} \leq \exp\{-\phi y\}. \tag{27}$$

Assumption 5.7.

$$\inf_{1 \leq i \leq d, \gamma_i \in \Gamma_i} \int_\delta^\Delta y q_{i,\gamma_i}(y) \mu_1(dy) \geq \frac{8}{\varepsilon \phi (e-1)}.$$

Remark 5.8. As remarked in [16], Assumption 5.6 generalizes the one-dimensional definition of log-concavity in the tails and Assumption 5.7 is easy to ensure, at least if $\Delta = \infty$, by taking the proposal distribution to be a mixture of an adaptive component and an uniform on $[-U, U]$ for U large enough or a mean zero Gaussian with large enough variance.

Theorem 5.9. *Let the selection probabilities $\alpha_n \in \mathcal{Y}$ for all n, with $\mathcal{Y}$ as in (5). Moreover assume that*

(a) $|\alpha_n - \alpha_{n-1}| \to 0$ in probability for fixed starting values $x_0 \in \mathcal{X}$ and $\alpha_0 \in \mathcal{Y}$, $\gamma_i \in \Gamma_i$, $i = 1, \dots, d$;

(b) 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 $\alpha_0 \in \mathcal{Y}$, $\gamma_i \in \Gamma_i$, $i = 1, \dots, d$;

(c) Assumptions 5.2, 5.3, 5.6, 5.7 hold.

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

$$T(x_0, \alpha_0, \gamma_0, n) \to 0 \qquad as \quad n \to \infty. \tag{28}$$

We now proceed to proofs.

Proof of Theorem 5.5. Ergodicity will follow from Theorem 2 of [41] by establishing diminishing adaptation and containment condition. Diminishing adaptation can be verified as in the proof of Theorem 4.19. Containment will result from Theorem 5.1.

Recall that $P_{\alpha,\gamma}$ is the random scan Metropolis within Gibbs kernel with selection probabilities α and proposals indexed by $\{\gamma_i\}_{1\leq i\leq d}$. To verify the small set condition (i), observe that Assumptions 5.2 and 5.3 imply that for every compact set C and every vector $\gamma_i \in \Gamma_i$, $i \in 1,\ldots,d$, we can find m^* and ρ^* independent of $\{\gamma_i\}$, and such that $P^{m^*}_{1/d,\gamma}(x,\cdot) \geq \rho^*\nu(\cdot)$ for all $x \in C$. Hence, arguing as in the proof of Lemma 4.4, there exist m and ρ, independent of $\alpha \in \mathcal{Y}$ and $\{\gamma_i\}$, such that $P^m_{\alpha,\gamma}(x,\cdot) \geq \rho\nu(\cdot)$ for all $x \in C$.

To establish the drift condition (ii), let $V_s := \pi(x)^{-s}$ for some $s \in (0,1)$ to be specified later. Then by Proposition 3 of [37] for all $1 \leq i \leq d$, $\gamma_i \in \Gamma_i$, and $x \in \mathbb{R}^d$ we have

$$P_{i,\gamma_i}V_s(x) \leq r(s)V_s(x) \quad \text{where} \quad r(s) := 1 + s(1-s)^{1/s-1}. \tag{29}$$

Since $r(s) \to 1$ as $s \to 0$, we can choose s small enough, so that

$$r(s) \quad < \quad 1 + \frac{\varepsilon\xi}{1-2\varepsilon\xi}. \tag{30}$$

The rest of the argument follows the proof of Theorem 2 in [16]. We repeat most of it since we need to ensure it is independent of α and γ. Assume by contradiction that there exists am $\mathbb{R}^d-$valued sequence $\{x^j\}$ s.t.

$$\limsup_{j\to\infty} \sup_{\alpha\in\mathcal{Y},\gamma_i\in\Gamma_i,1\leq i\leq d} P_{\alpha,\gamma}V_s(x^j)/V_s(x^j) \geq 1.$$

Then there exists a subsequence $\{\hat{x}^j\}$ such that

$$\lim_{j\to\infty} \sup_{\alpha\in\mathcal{Y},\gamma_i\in\Gamma_i,1\leq i\leq d} P_{\alpha,\gamma}V_s(\hat{x}^j)/V_s(\hat{x}^j) \geq 1.$$

Moreover, as shown in [16] (proof of Theorem 2, p. 129), there exists an integer $k \in \{1,\ldots,d\}$ and a further subsequence $\{\tilde{x}^j\}$, such that

$$\lim_{j\to\infty} \sup_{\gamma_k\in\Gamma_k} P_{k,\gamma_k}V_s(\tilde{x}^j)/V_s(\tilde{x}^j) \quad \leq \quad r(s) - (2r(s)-1)\xi. \tag{31}$$

The contradiction follows from (29), (30) and (31), since

$$\begin{aligned}
\lim_{j\to\infty} \sup_{\alpha\in\mathcal{Y},\gamma_i\in\Gamma_i,1\le i\le d} \frac{P_{\alpha,\gamma}V_s(\tilde{x}^j)}{V_s(\tilde{x}^j)} &= \lim_{j\to\infty}\sup_{\alpha\in\mathcal{Y}}\sum_{i=1}^d \alpha_i \sup_{\gamma_i\in\Gamma_i}\frac{P_{i,\gamma_i}V_s(\tilde{x}^j)}{V_s(\tilde{x}^j)} \\
&= \lim_{j\to\infty}\sup_{\alpha\in\mathcal{Y}}\left(\alpha_k \sup_{\gamma_k\in\Gamma_k} P_{k,\gamma_k}V_s(\tilde{x}^j)/V_s(\tilde{x}^j) + \sum_{i\neq k}\alpha_i \sup_{\gamma_i\in\Gamma_i}\frac{P_{i,\gamma_i}V_s(\tilde{x}^j)}{V_s(\tilde{x}^j)}\right) \\
&\le \varepsilon\big(r(s)-(2r(s)-1)\xi\big)+(1-\varepsilon)r(s) \;<\; 1.
\end{aligned}$$

□

Proof of Theorem 5.9. The proof is identical to the proof of Theorem 5.5 with the only difference that now the drift condition (ii) of Theorem 5.1 will be established under Assumptions 5.6 and 5.7.

Establishing (ii) of Theorem 5.1 will follow closely the proof of Theorem 3 in [16]. Let again $V_s := \pi(x)^{-s}$ for some $s\in(0,1)$ to be specified later and recall that (29) holds for all $1\le i\le d$, $\gamma_i\in\Gamma_i$, and $x\in\mathbb{R}^d$. Assume by contradiction that there exists am $\mathbb{R}^d-$valued sequence $\{x^j\}$ s.t.

$$\limsup_{j\to\infty} \sup_{\alpha\in\mathcal{Y},\gamma_i\in\Gamma_i,1\le i\le d} P_{\alpha,\gamma}V_s(x^j)/V_s(x^j)\ge 1.$$

Then there exists a subsequence $\{\hat{x}^j\}$ such that

$$\lim_{j\to\infty} \sup_{\alpha\in\mathcal{Y},\gamma_i\in\Gamma_i,1\le i\le d} P_{\alpha,\gamma}V_s(\hat{x}^j)/V_s(\hat{x}^j)\ge 1.$$

Moreover, as shown in [16] (proof of Theorem 3, p. 137, equation (15)), there exists an integer $k\in\{1,\dots,d\}$ and a further subsequence $\{\tilde{x}^j\}$, such that

$$\begin{aligned}
\lim_{j\to\infty} P_{k,\gamma_k}V_s(\tilde{x}^j)/V_s(\tilde{x}^j) \;&\le\; r(s)-(2r(s)-1)\mathcal{J}_{\gamma_k}(0) \\
&\quad +\mathcal{J}_{\gamma_k}(\phi s)+\mathcal{J}_{\gamma_k}(\phi(1-s))-\mathcal{J}_{\gamma_k}(\phi),
\end{aligned} \tag{32}$$

where for $b>0$,

$$\mathcal{J}_{\gamma_k}(b) \;=\; \int_\delta^\Delta e^{-by} q_{k,\gamma_k}(y)\mu_1(dy).$$

Now from (29) and (32) compute

$$\begin{aligned}
\lim_{j\to\infty} \sup_{\alpha\in\mathcal{Y},\gamma_i\in\Gamma_i,1\le i\le d} \frac{P_{\alpha,\gamma}V_s(\tilde{x}^j)}{V_s(\tilde{x}^j)} &= \lim_{j\to\infty}\sup_{\alpha\in\mathcal{Y}}\sum_{i=1}^d \alpha_i \sup_{\gamma_i\in\Gamma_i}\frac{P_{i,\gamma_i}V_s(\tilde{x}^j)}{V_s(\tilde{x}^j)} \\
&= \lim_{j\to\infty}\sup_{\alpha\in\mathcal{Y}}\left(\alpha_k \sup_{\gamma_k\in\Gamma_k} P_{k,\gamma_k}V_s(\tilde{x}^j)/V_s(\tilde{x}^j) + \sum_{i\neq k}\alpha_i \sup_{\gamma_i\in\Gamma_i}\frac{P_{i,\gamma_i}V_s(\tilde{x}^j)}{V_s(\tilde{x}^j)}\right) \\
&\le r(s)-\varepsilon\inf_{\gamma_k\in\Gamma_k}\Big((2r(s)-1)\mathcal{J}_{\gamma_k}(0)+\mathcal{J}_{\gamma_k}(\phi s)+\mathcal{J}_{\gamma_k}(\phi(1-s))-\mathcal{J}_{\gamma_k}(\phi)\Big) \\
&= \sup_{\gamma_k\in\Gamma_k}\Big(r(s)-\varepsilon\big((2r(s)-1)\mathcal{J}_{\gamma_k}(0)+\mathcal{J}_{\gamma_k}(\phi s)+\mathcal{J}_{\gamma_k}(\phi(1-s))-\mathcal{J}_{\gamma_k}(\phi)\big)\Big) \\
&=: \sup_{\gamma_k\in\Gamma_k}\mathcal{H}(\gamma_k,\phi,s).
\end{aligned}$$

The result will follow if we can find such an s that $\sup_{\gamma_k\in\Gamma_k}\mathcal{H}(\gamma_k,\phi,s) < 1$. Note that $\mathcal{H}(\gamma_k,\phi,0)=1$ for every $\gamma_k\in\Gamma_k$ and the function is differentiable. Therefore it is enough to show that there exist $\kappa_1>0$ and $\kappa_2>0$ such that

$$\frac{\partial}{\partial s}\mathcal{H}(\gamma_k,\phi,s) \quad < \quad -\kappa_1 \qquad \text{for all} \quad \gamma_k\in\Gamma_k \quad \text{and} \quad s\in(0,\kappa_2)$$

and conclude (ii) with $V_s(x)=\pi^{-s}(x)$ and $s:=\kappa_2$. To this end compute

$$\begin{aligned}\frac{1}{\varepsilon}\frac{\partial}{\partial s}\mathcal{H}(\gamma_k,\phi,s) \quad &= \quad \Big(\frac{1}{\varepsilon}-2\mathcal{J}_{\gamma_k}(0)\Big)\frac{\partial}{\partial s}r(s)-\phi\int_\delta^\Delta ye^{-\phi sy}q_{\gamma_k}(y)\mu_1(dy)\\ &\qquad +\phi\int_\delta^\Delta ye^{-\phi(1-s)y}q_{\gamma_k}(y)\mu_1(dy)\\ &= \quad \frac{1}{\varepsilon}\frac{(1-s)^{1/s}\log(1-s)}{s(s-1)}-\phi I_1+\phi I_2 \quad =: \quad \clubsuit,\end{aligned}$$

and notice that by $1/\phi\le\delta$ and Assumption 5.7, for s small enough we have

$$\begin{aligned} I_1-I_2 \quad &\ge \quad \frac{e-1}{2e}\int_\delta^\Delta yq_{\gamma_k}(y)\mu_1(dy)\\ &\ge \quad \frac{e-1}{2e}\frac{8}{\varepsilon\phi(e-1)} \quad = \quad \frac{4}{\varepsilon\phi e}, \qquad \text{and}\\ \frac{(1-s)^{1/s}\log(1-s)}{s(s-1)} \quad &\le \quad \frac{2}{e}.\end{aligned}$$

Consequently there exists $\kappa_2>0$ s.t. for all $s\in(0,\kappa_2)$

$$\clubsuit \quad \le \quad \frac{2}{\varepsilon e}-\frac{4\phi}{\varepsilon\phi e} \quad = \quad -\frac{2}{\varepsilon e} \quad =: \quad \kappa_1 \quad < \quad 0.$$

$\square$

Example 5.10. We now give an example involving a simple generalised linear mixed model. Consider the model and prior given by

$$\begin{aligned} Y_i \quad &\sim \quad Pois\left(e^{\theta+X_i}\right) &(33)\\ X_i \quad &\sim \quad N(0,1) &(34)\\ \theta \quad &\sim \quad N(0,1) &(35)\end{aligned}$$

The model is chosen to be extremely simple so as to not detract from the argument used to demonstrate ergodicity of `adapRSadapMwG`, although this argument readily generalises to different exponential families, link functions and random effect distributions.

We consider simulating from the posterior distribution of $\theta,\mathbf{X}$ given observations $y_1,\dots y_n$ using `adapRSadapMwG`. More specifically we set

$$q_{x_{-i},\gamma}(x_i,y_i)=\frac{\exp\left\{-(y_i-x_i)^2/2\gamma\right\}}{\sqrt{2\pi\gamma}} \tag{36}$$

where the range of permissible scales γ is restricted to be in some range $\Re = [a, b]$ with $0 < a \le b < \infty$. We are in the subexponential tail case and specifically we have the following.

Proposition 5.11. *Consider `adapRSadapMwG` applied to model (33) using any adaptive scheme satisfying the conditions (a) and (b) of Theorem 5.5. Then the scheme is ergodic.*

For the proof, we require the following definition from [16]. We let

$$\Phi = \{\text{functions } \phi : \mathbb{R}^+ \to \mathbb{R}^+;\ \phi(x) \to \infty \text{ as } x \to \infty\}.$$

Proof of Proposition 5.11. According to Theorem 5.5, it remains to check conditions 5.2, 5.3, 5.4 hold. Conditions 5.2 and 5.3 hold by construction, while condition 5.4 consists of two separate conditions. One of these, given in (24), holds by construction from (36). Moreover, [16] shows that (25) can be replaced by the following condition: there exist functions $\{\phi_i \in \Phi,\ 1 \le i \le d\}$ such that $i \in \{1, \dots, d\}$ and all $y \in [\delta, \Delta]$,

$$\lim_{|x_i|\to\infty} \sup_{\{x_{-i};\ \phi_j(|x_j|)\le\phi_i(|x_i|),\ j\ne i\}} \frac{\pi(\tilde{x}^j)}{\pi(\tilde{x}^j - sign(\tilde{x}^j_i) y e_i)} = 0 \qquad (37)$$

$$\text{and} \quad \lim_{|x_i|\to\infty} \sup_{\{x_{-i};\ \phi_j(|x_j|)\le\phi_i(|x_i|),\ j\ne i\}} \frac{\pi(\tilde{x}^j + sign(\tilde{x}^j_i) y e_i)}{\pi(\tilde{x}^j)} = 0. \qquad (38)$$

Now take $\phi_i(x) = x$ for all $1 \le i \le d$ so that (37) can be rewritten as the two conditions

$$\lim_{|x_i|\to\infty} \sup_{\{x_{-i};\ |x_j|\le|x_i|,\ j\ne i\}} \exp\left\{\int_{-y}^{0} \nabla_i \log \pi(x + sign(x_i) z e_i) dz\right\} = 0 \qquad (39)$$

$$\lim_{|x_i|\to\infty} \sup_{\{x_{-i};\ |x_j|\le|x_i|,\ j\ne i\}} \exp\left\{\int_{0}^{y} \nabla_i \log \pi(x + sign(x_i) z e_i) dz\right\} = 0 \qquad (40)$$

for all $y \in [\delta, \Delta]$, where ∇_i denotes the derivative in the ith direction. We shall show that uniformly on the set $S_i(x_i)$ which is defined to be $\{x_{-i};\ |x_j| \le |x_i|,\ j \ne i\}$ the function $\nabla_i \log \pi(x)$ converges to $-\infty$ as $x_i \to +\infty$ and to $+\infty$ as x_i approaches $-\infty$.

Now we have $d = n+1$ and let i correspond to the component x_i for $1 \le i \le n$ with $n + 1$ denoting the component θ. Therefore for $1 \le i \le n$

$$\nabla_i \log \pi(x) = -e^{\theta + x_i} + y_i - x_i$$

and

$$\nabla_{n+1} \log \pi(x) = -\sum_{i=1}^{n} e^{\theta + x_i} - \sum_{i=1}^{n} y_i - \theta$$

Now for $x_i > 0$, $1 \le i \le n$

$$\nabla_i \log \pi(x) \ge y_i - x_i$$

which is diverging to $-\infty$ independently of x_{-i}. Similarly,

$$\nabla_{n+1} \log \pi(x) \geq \sum_{i=1}^{n} y_i - \theta$$

diverging to $-\infty$ independently of $\{x_i;\ 1 \leq i \leq n\}$.

For $x_i < 0$, $1 \leq i \leq n$ and $(x_{-i}, \theta) \in S_i(x_i)$,

$$\nabla_i \log \pi(x) \leq y_i - x_i + 1$$

again diverging to $+\infty$ uniformly. Finally for $\theta < 0$ and $x \in S_{n+1}(\theta)$

$$\nabla_{n+1} \log \pi(x) \geq -n + \sum_{i=1}^{n} y_i - \theta$$

again demonstrating the required uniform convergence. Thus ergodicity holds. □

Remark 5.12. The random effect distribution in Example 5.10 can be altered to give different results. For instance if the distribution is doubly exponential Theorem 4.2 can be applied using very similar arguments to those used above. Extensions to more complex hierarchical models are clearly possible though we don't pursue this here

Remark 5.13. An important problem that we have not focused on involves the construction of explicit adaptive strategies. Since little is known about the optimisation of the Random Scan Random Walk Metropolis even in the non-adaptive case, this is not a straightforward question. Further work we are engaged in is exploring adaptation to attept to maximise a given optimality criterion for the chosen class of samplers. Two possible strategies are

- to scale the proposal variance to approach 2.4 times the empirically observed conditional variance;
- to scale the proposal variance to achieve an algorithm with acceptance proportion approximately 0.44.

Both these methods are founded in theoretical arguments, see for instance [39].

6. 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

$$\nu_n \ := \ \left\{ \frac{1}{4} - \frac{1}{a_n} \ , \ \frac{1}{2} \ , \ \frac{1}{4} + \frac{1}{a_n} \right\}. \tag{41}$$

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

$$\Omega_{\tilde{X}\geq S} := \{\omega \in \Omega : \tilde{X}_n(\omega) \geq S_n(\omega) \ \text{ for every } \ n\} \tag{42}$$

and

$$\Omega_\infty := \{\omega \in \Omega : S_n(\omega) \to \infty\}. \tag{43}$$

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

$$\mathbb{P}(\Omega_{\tilde{X}\geq S} \cap \Omega_\infty) > 0. \tag{44}$$

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

$$\mathbb{P}(S_k^{k+n} - \mathbb{E}S_k^{k+n} \leq -nt) \leq \exp\{-\frac{1}{2}nt^2\}. \tag{45}$$

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

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

Remark 6.1. 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 than required.

Lemma 6.2. *Let Y_n and S_n be as defined above and let*

$$\Omega_1 \quad := \quad \Big\{\omega \in \Omega : S_k = k \quad \textit{for every} \quad 0 < k \le c_1\Big\}. \tag{46}$$

$$\Omega_n \quad := \quad \Big\{\omega \in \Omega : S_k \ge \frac{b_{n-1}}{2} \quad \textit{for every} \quad c_{n-1} < k \le c_n\Big\} \quad \textit{for} \ \ n \ge 2. \tag{47}$$

Then

$$\mathbb{P}\bigg(\bigcap_{n=1}^{\infty} \Omega_n\bigg) > 0. \tag{48}$$

Remark 6.3. 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 $S_{c_{n-1}}^{c_n}$ is a sum of b_n i.i.d. random variables with $\mathbb{E}S_{c_{n-1}}^{c_n} = \frac{2b_n}{10+\log(n)}$. Therefore for every $n \ge 1$ by Hoeffding's inequality with $t = 1/(10+\log(n))$, we can also write

$$\mathbb{P}\Big(S_{c_{n-1}}^{c_n} \le \frac{b_n}{10+\log(n)}\Big) \le \exp\bigg\{-\frac{1}{2}\frac{b_n}{(10+\log(n))^2}\bigg\} =: p_n.$$

Therefore using the above bound iteratively we obtain

$$\mathbb{P}(S_{c_1} = b_1,\ S_{c_n} \ge b_n \ \text{ for every } \ n \ge 2) \ \ge \ p_{1,S} \prod_{n=2}^{\infty} (1 - p_n). \tag{49}$$

Note that $\{S_{c_n} \ge b_n\} \subseteq \Omega_n$ by the choice of b_n, and hence equation (49) implies also

$$\mathbb{P}\bigg(\bigcap_{n=1}^{\infty} \Omega_n\bigg) \ge p_{1,S} \prod_{n=2}^{\infty} (1 - p_n). \tag{50}$$

Clearly in this case

$$p_{1,S} \prod_{n=2}^{\infty} (1 - p_n) > 0 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \log(1 - p_n) > -\infty \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} p_n < \infty. \tag{51}$$

We conclude (51) by comparing p_n with $1/n^2$. We show that there exists n_0 such that for $n \ge n_0$ the series p_n decreases quicker then the series $1/n^2$ and therefore p_n is summable. We check that

$$\log \frac{p_{n-1}}{p_n} > \log \frac{n^2}{(n-1)^2} \qquad \text{for} \quad n \ge n_0. \tag{52}$$

Indeed

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

Now recall that b_{n-1} is an increasing sequence. Moreover the numerator 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 $n\geq n_0$

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

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

$$\mathbb{P}\left(\left(\bigcap_{n=1}^{\infty}\Omega_n\right)\cap\Omega_{\tilde{X}\geq S}\right)\geq C\mathbb{P}\left(\bigcap_{n=1}^{\infty}\Omega_n\right)\qquad\text{for some universal}\quad C>0,\tag{53}$$

and therefore

$$\mathbb{P}\left(\Omega_{\tilde{X}\geq S}\cap\Omega_\infty\right)>0.\tag{54}$$

Thus nonergodicity of $(X_n)_{n\geq 0}$ will follow from Lemma 6.2. We start with the following observation.

Lemma 6.4. *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 $\tilde{X}_{n-1}$

$$\mathbb{P}(\tilde{X}_n-\tilde{X}_{n-1}=1,Y_n=1)\geq\mathbb{P}(\tilde{X}_n-\tilde{X}_{n-1}=1)\mathbb{P}(Y_n=1),\tag{55}$$

(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.

$$Y_n=1\quad\Rightarrow\quad\tilde{X}_n-\tilde{X}_{n-1}=1\tag{56}$$

$$\tilde{X}_n-\tilde{X}_{n-1}=-1\quad\Rightarrow\quad Y_n=-1.\tag{57}$$

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

$$\Big((\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2}\ ,\ 1-(\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2}-(\frac{1}{4}+\frac{2}{a_n})\ ,\ \frac{1}{4}+\frac{2}{a_n}\Big), \tag{58}$$

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

$$\Big(\frac{1}{4}-\frac{2}{a_n}\ ,\ 1-(\frac{1}{4}-\frac{2}{a_n})-(\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2}\ ,\ (\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2}\Big), \tag{59}$$

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 (58) 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 (59). Let $\leq_{\mathrm{st}}$ denote stochastic ordering. By simple algebra both measures defined in (58) and (59) are stochastically bigger than

$$\mu_n^i \quad = \quad (\mu_{n,1}^i, \mu_{n,2}^i, \mu_{n,3}^i), \tag{60}$$

where

$$\begin{aligned}
\mu_{n,1}^i \quad &= \quad (\frac{1}{4}-\frac{2}{a_n})(1+\frac{2}{i}) \quad = \quad \frac{1}{4}-\frac{1}{a_n}-\frac{2i+8-a_n}{2ia_n}, & (61)\\
\mu_{n,2}^i \quad &= \quad 1-(\frac{1}{4}-\frac{2}{a_n})(1+\frac{2}{i})-(\frac{1}{4}+\frac{2}{a_n})(1-\frac{2}{\max\{4,i\}}), &\\
\mu_{n,3}^i \quad &= \quad (\frac{1}{4}+\frac{2}{a_n})(1-\frac{2}{\max\{4,i\}}) \quad = \quad \frac{1}{4}+\frac{1}{a_n}+\frac{2\max\{4,i\}-8-a_n}{2a_n\max\{4,i\}}. & (62)
\end{aligned}$$

Recall ν_n, the distribution of Y_n defined in (41). Examine (61) and (62) to see that if $2i-8 \geq a_n$, then $\mu_n^i \geq_{\mathrm{st}} \nu_n$. Hence in this case also the distribution of $\tilde{X}_n - \tilde{X}_{n-1}$ is stochastically bigger than the distribution of Y_n. The joint probability distribution of $(\tilde{X}_n - \tilde{X}_{n-1}, Y_n)$ satisfying (56) and (57) follows. □

Proof of Proposition 3.2. Define

$$\Omega_{1,\tilde{X}} := \Big\{\omega \in \Omega : \tilde{X}_n - \tilde{X}_{n-1} = 1 \quad \text{for every} \quad 0 < n \leq c_1\Big\}. \tag{63}$$

Since the distribution of $\tilde{X}_n - \tilde{X}_{n-1}$ is stochastically bigger than μ_n^i defined in (60) 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 6.4 *(a)* we have

$$\mathbb{P}\big(\Omega_{1,\tilde{X}} \cap \Omega_1\big) \geq p_{1,S}\, p_{1,\tilde{X}} > 0. \tag{64}$$

Since $S_{c_1} = \tilde{X}_{c_1} = c_1 = b_1$, on $\Omega_{1,\tilde{X}} \cap \Omega_1$, the requirements for Lemma 6.4 *(b)* hold for $n-1 = c_1$. We shall use Lemma 6.4 *(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 6.4 *(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

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

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

Acknowledgements

While preparing this manuscript KŁ was a postdoctoral fellow at the Department of Statistics, University of Toronto. KŁ and JSR were partially funded by NSERC of Canada. KŁ and GOR were partially funded by CRiSM and other grants from EPSRC.

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 (2010): 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] L. Bottolo, S. Richardson, and J.S. Rosenthal (2010): Bayesian models for sparse regression analysis of high dimensional data. In preparation.

[12] 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.

[13] R.V. Craiu, J.S. Rosenthal, and C. Yang (2009): Learn From Thy Neighbor: Parallel-Chain and Regional Adaptive MCMC. *J. Amer. Stat. Assoc.* **488**, 1454–1466.

[14] R.V. Craiu and A.F. Di Narzo (2009): A Mixture-Based Approach to Regional Adaption for MCMC. *Preprint.*

[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] G. Fort, E. Moulines, G.O. Roberts, J.S. Rosenthal (2003): On the geometric ergodicity of hybrid samplers. *J. Appl. Prob.* 40, 123–146.

[17] 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.

[18] H. Haario, E. Saksman, and J. Tamminen (2001): An adaptive Metropolis algorithm. *Bernoulli* 7, 223–242.

[19] H. Haario, E. Saksman, and J. Tamminen (2005): Componentwise adaptation for high dimensional MCMC. *Computational Statistics* 20, 265–273.

[20] W.K. Hastings (1970): Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1), 97–109.

[21] S.F. Jarner, E. Hansen (2000): Geometric Ergodicity of Metropolis Algorithms. *Stoch. Process. Appl.* 85, 341–361.

[22] K. Łatuszyński (2008): Regeneration and Fixed-Width Analysis of Markov Chain Monte Carlo Algorithms. PhD Dissertation. Available at: arXiv:0907.4716v1

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

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

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

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

[27] 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.

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

[29] 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.

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

[31] R.C. Neath, G.L. Jones (2009): *Variable-at-a-time Implementations of Metropolis-Hastings*, arXiv:0903.0664v1

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

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

[34] 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.

[35] 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.

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

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

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

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

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

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

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

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

[44] E. Saksman and M. Vihola (2008): On the Ergodicity of the Adaptive Metropolis Algorithm on Unbounded Domains. *Preprint.*

[45] 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
[46] M. Vihola (2009): On the Stability and Ergodicity of an Adaptive Scaling Metropolis Algorithm. *Preprint.*
[47] C. Yang (2008): On The Weak Law Of Large Numbers For Unbounded Functionals For Adaptive MCMC. *Preprint.*
[48] C. Yang (2008): Recurrent and Ergodic Properties of Adaptive MCMC. *Preprint.*