Submitted to Bernoulli

An adaptive multiple-try Metropolis algorithm

SIMON FONTAINE[1,2,*] and MYLÈNE BÉDARD[1]

[1] *Université de Montréal, Département de mathématiques et de statistique, 2920 chemin de la Tour, Montréal, QC, Canada, H3T 1J4.*

[2] *University of Michigan, Department of Statistics. West Hall, 1085 South University, Ann Arbor, MI, U.S.A., 48109. E-mail:* [*] simfont@umich.edu

Markov chain Monte Carlo (MCMC) methods, specifically samplers based on random walks, often have difficulty handling target distributions with complex geometry such as multi-modality. We propose an adaptive multiple-try Metropolis algorithm designed to tackle such problems by combining the flexibility of multiple-proposal samplers with the user-friendliness and optimality of adaptive algorithms. We prove the ergodicity of the resulting Markov chain with respect to the target distribution using common techniques in the adaptive MCMC literature. In a Bayesian model for loss of heterozygosity in cancer cells, we find that our method outperforms traditional adaptive samplers, non-adaptive multiple-try Metropolis samplers, and various more sophisticated competing methods.

Keywords: adaptive scaling, ergodicity, limit theorem, loss of heterozygosity, multiple candidates, random walk sampler, robust adaptation.

MSC 2010 subject classifications: Primary 60J22, Secondary 60J10, 60J20, 65C40, 62F15.

1. Introduction

Suppose we wish to find an expectation of the form $\pi(f) = \mathbb{E}_{X\sim\pi}\{f(X)\}$ for some π-integrable function $f : \mathcal{X} \to \mathbb{R}^q$ on some state space $\mathcal{X} \subseteq \mathbb{R}^d$. Monte Carlo (MC) methods make use of a law of large numbers, i.e.

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \stackrel{\mathcal{C}}{\longrightarrow} \pi(f) , \qquad n \to \infty , \tag{1.1}$$

for some mode of convergence $\mathcal{C} \in \{\text{in probability}, \text{ almost surely}\}$, to estimate $\pi(f)$ using the sample average. The conditions under which (1.1) holds rely on the joint distribution of $(X_n)_{n=1}^N$. For example, an i.i.d. assumption ($X_n \stackrel{\text{i.i.d.}}{\sim} \pi$, $n = 1, \dots, N$) is often sufficient to verify a law of large numbers. When π is even moderately complex however, it is generally impossible to sample directly from that distribution. Markov chain Monte Carlo (MCMC) methods provide a way to produce a sample $(X_n)_{n=1}^N$ that verifies a law of large numbers for some class of functions f while not requiring direct sampling from the target distribution π.

imsart-bj ver. 2014/10/16 file: article.tex date: December 23, 2020

One of the most common MCMC methods is the Metropolis-Hastings (MH) algorithm (Hastings, 1970), which produces the sample $(X_n)_{n=1}^N$ sequentially from time $n = 1$ through $n = N$. At time n, instead of sampling directly from π, we sample from a proposal distribution $q(\cdot|x)$ which may or may not depend on the previous sample point $X_{n-1} = x$. Since the candidate $Y \sim q(\cdot|x)$ is sampled from q and not from π, we need to proceed to an accept/reject step to adjust for that bias. The next point X_n to be included in the sample is chosen to be equal to y with probability

$$\alpha_{\mathrm{MH}}(y|x) = \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\} , \tag{1.2}$$

known as the *MH acceptance probability*, and equal to the previous observation x with probability $1 - \alpha_{\mathrm{MH}}(y|x)$. Now, since the distribution of X_n only depends on the previous time point $X_{n-1} = x$, then $\mathbf{X} = (X_n)_{n\geqslant 1}$ actually forms a Markov chain.

A homogeneous Markov chain $\mathbf{X}$ on a state space $\mathcal{X}$ with Markov transition P is ergodic with respect to some distribution Π for some initial state $x \in \mathcal{X}$ if

$$\lim_{n\to\infty} \|P^n(\cdot|x) - \Pi(\cdot)\|_{\mathrm{TV}} = 0 , \tag{1.3}$$

where $\|\mu\|_{\mathrm{TV}} = \sup_{B\in\mathcal{B}(\mathcal{X})} |\mu(B)|$ denotes the total variation norm of the signed measure μ, and where

$$P^m(B|x) = \int_{\mathcal{X}} P^{m-1}(B|y)\, P(\,\mathrm{d}y|x) , \qquad m > 1$$

is the *iterated* Markov transition with base case $P^1(B|x) = P(B|x)$. Typically, the ergodicity of a homogeneous Markov chain with respect to a density π is established through results such as in Tierney (1994, Theorem 1) where it is required that (1) π be the stationary density of the Markov transition P with density p, (2) the chain be π-irreducible, and (3) the chain be aperiodic. While ergodicity and laws of large numbers are two different concepts, the conditions used to verify the former are sufficient to verify the latter for all π-integrable functions (Meyn and Tweedie, 2009).

A sufficient condition for $\mathbf{X}$ to admit π as its stationary distribution is the *detailed balance* condition on the densities (Robert and Casella, 2004),

$$p(y|x)\,\pi(x) = p(x|y)\,\pi(y) , \qquad \forall\, x, y \in \mathcal{X} . \tag{1.4}$$

By construction, the chain $\mathbf{X}$ generated using a MH algorithm satisfies the first ergodicity condition since the expression of the MH acceptance probability (1.2) is specifically chosen to satisfy (1.4). A sufficient condition for aperiodicity and π-irreducibility of MH chains is the local positivity of the proposal density q,

$$\|x - y\|_2 < \delta \quad \Rightarrow \quad q(y|x) > \varepsilon , \tag{1.5}$$

for some $\delta, \varepsilon > 0$ ($\|\cdot\|_2$ is the Euclidean norm), together with the assumption that π is bounded above and away from 0 on any compact subset of the state space $\mathcal{X}$ (Robert and Casella, 2004). This type of condition can easily be verified for the Metropolis algorithm

(Metropolis et al., 1953), a special case of the MH algorithm that uses a symmetric random-walk proposal density,

$$q\,(y|x) = q\,(y - x) = q\,(x - y)\ \ .$$

Since its initial development, the MH algorithm—and most notably the Metropolis sampler—has seen multiple proposed improvements, of which we consider two here. The *Multiple-try Metropolis* (MTM) algorithm defines a variant of the Metropolis sampler where several candidates are generated in a given iteration; this technique produces a transition that is better adapted to the specific geometry of the target density, leading to an improved state space exploration. The *Adaptive Metropolis* (AM) algorithm, on the other hand, uses a random-walk proposal density at each iteration but adapts it through time to match the covariance of the target, therefore producing higher-quality candidates.

Both algorithms improve on the vanilla Metropolis sampler, but each suffers from the exact problem that the other algorithm aims at solving. The MTM sampler requires a large amount of hand-tuning that adaptive algorithms perform automatically; the AM algorithm typically uses simple proposal densities that may not be well-suited to target densities featuring complex geometries, such a multi-modal densities. In this article, we propose a novel adaptive MCMC algorithm that unites the advantages of these two samplers and therefore fixes some of their respective flaws.

Related work We review some of the recent attempts at integrating adaptation within the multiple-try framework. Martino et al. (2018) propose the *adaptive independant sticky multiple-try Metropolis*, which uses a non-parametric independent proposal density. Multiple candidates are sampled from this non-parametric density adapted using rejected points. Being non-parametric, this method does not extend well beyond a few dimensions for full-dimensional samplers. Casarin, Craiu and Leisen (2013) propose the *interacting multiple-try Metropolis* in which multiple parallel MTM chains interact with each other and MTM selection weights are adapted using all chains. Yang et al. (2019) propose an *adaptive component-wise multiple-try Metropolis* algorithm that consists of a multiple-try generalization of the Metroplis-within-Gibbs where one-dimensional proposals are adapted using MTM selection proportions of the chain's past. Tran, Pitt and Kohn (2016) briefly mention that their proposed *adaptive correlated Metropolis-Hastings* algorithm could be extended to include multiple candidates.

Principal contributions The main contribution of this research is the proposed aMTM algorithm, which consists of an adaptive MCMC sampler with full-dimensional adapted multiple-try Metropolis proposal densities. Introducing adaptation in MTM algorithms is a natural extension to the current literature; it is surprising that nothing has yet been published on the subject. We derive an ergodicity result under the assumption that both the sample and parameter spaces are bounded. Intermediary theoretical results are readily extendable to other adaptive MCMC algorithms and provide slightly more general knowledge about MTM transitions. We provide an implementation of our proposed algorithm in a R package called aMTM available at https://github.com/fontaine618/aMTM, which mostly consists of a wrapper for the main sampling function written in C++.

Paper organization In Section 2, we introduce some background on MTM and adaptive algorithms. Section 3 contains a general description of our proposed algorithm along with some variants. The validity of our sampler is discussed in Section 4, where ergodicity is proven. We conduct simulation experiments in Section 5 to assess the performance of our approach. Supplement A contains additional details on particular variants of the algorithms, results and proofs omitted from the main text.

2. Background

2.1. Multiple-try Metropolis

A natural extension to the MH algorithm is to consider K candidate points per iteration instead of a single one (Liu, Liang and Wong, 2000). The resulting *multiple-try Metropolis* sampler must therefore include an additional step that randomly selects a proposal among the set of K candidates $Y^{(1:K)} \sim q(\cdot|x)$ according to some positive sampling weight function $w^{(k)}(\cdot|x)$, which may depend on the index k of the candidate and on the previous state x of the chain. Throughout the text, exponents in parentheses $^{(k)}$ index candidates with the convention that $^{(1:K)}$ selects all candidates while $^{(-k)}$ omits the k-th candidate. Standardizing these weights, we obtain the probability of choosing $y^{(k)} = y$ as the official candidate:

$$\bar{w}^{(k)}(y, y^{(-k)}|x) = \frac{w^{(k)}(y^{(k)}|x)}{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)} \ .$$

Once an official candidate $k \in \{1, \ldots, K\}$ is selected, the proposed value $y^{(k)} = y$ must go through an accept/reject step in order to become the next state of the chain. As before, the acceptance probability is chosen such as to satisfy the detailed balance condition (1.4), which basically requires that the trajectory produced by a stationary Markov chain be equally probable when run forward or backward in time. To satisfy this condition, we thus need to generate a shadow sample $x_*^{(j)}$, $j = 1, \ldots, K$, that mimics the generation of a candidate set, and then select the official candidate x for going from y to x (instead of from x to y). That is, we let $X_*^{(k)} = x$ and sample $X_*^{(-k)} \sim q^{(-k)}(\cdot|y)$, where $q^{(-k)}$ is the conditional proposal distribution given the k-th component $X_*^{(k)} = x$. Using the following *MTM acceptance probability*

$$\alpha_{\text{MTM}}\left(y, y^{(-k)}|x, x_*^{(-k)}\right) = \min\left\{1, \frac{\pi(y)q^{(k)}(x|y)\bar{w}^{(k)}(x, x_*^{(-k)}|y)}{\pi(x)q^{(k)}(y|x)\bar{w}^{(k)}(y, y^{(-k)}|x)}\right\} \tag{2.1}$$

is then sufficient to verify (1.4), assuming that the marginal density of the k-th candidate, $q^{(k)}(\cdot|\cdot)$, satisfies

$$q^{(k)}(x|y) > 0 \qquad \Leftrightarrow \qquad q^{(k)}(y|x) > 0 \ , \tag{2.2}$$

for all $k = 1, \dots, K$ (see Supplement A, Proposition B.1).

The MTM design and the detailed balance condition's verification do not impose any restriction on the *joint* distribution of candidates, only on their *marginal* distributions. The set of candidates can thus be generated in any way and choosing appropriate correlation structures within the candidate set can greatly improve the algorithm's performance. The simplest choice is to generate candidates independently, but this does not make use of the sampler's full potential. Indeed, nothing prevents two candidates from being very close to one another, which does not improve the state space exploration.

Extremely antithetic (EA) candidates are generated so that their pairwise Euclidean distances be maximized. Craiu and Lemieux (2007, Section 3.1) achieve this by introducing a correlation of $\rho = -1/(K-1)$ between the K candidates. For example, if marginal Gaussian proposal distributions with unit spherical covariances I_d are used, this yields the joint covariance matrix

$$\mathrm{Var}\left(Y^{(1:K)}\right) = \begin{pmatrix} I_d & \cdots & \rho I_d \\ \vdots & \ddots & \vdots \\ \rho I_d & \cdots & I_d \end{pmatrix} \in \mathbb{R}^{dK \times dK} . \tag{2.3}$$

To produce candidates using different covariance matrices, we can simply generate K d-dimensional Gaussian observations using (2.3) and transform them using a Cholesky decomposition. To produce the shadow sample, we need to compute the conditional distribution of $Y^{(-k)}$ given $Y^{(k)} = y^{(k)}$. In the unit spherical covariance case, we can show (Fontaine, 2019, Section 5.3.4.2) that this corresponds to a Gaussian distribution with some specific mean and the following joint covariance,

$$\mathrm{Var}\left(Y^{(-k)}|y^{(k)}\right) = (1-\rho) \begin{pmatrix} (1+\rho) I_d & \cdots & \rho I_d \\ \vdots & \ddots & \vdots \\ \rho I_d & \cdots & (1+\rho) I_d \end{pmatrix} \in \mathbb{R}^{d(K-1) \times d(K-1)} .$$

Randomized quasi-Monte Carlo (RQMC) methods are constructed using a (random) regularly-spaced grid on the unit hyper-cube before going through a probability integral transform. For example, Craiu and Lemieux (2007, Section 3.2) construct such a grid using a Korobov rule. The RQMC candidates were generalized by Bédard, Douc and Moulines (2012) to *common random number* candidates, where the regular grid assumption is removed. We refer the reader to Supplement A, Section A.1, for a summary on how to perform these various sampling schemes with a multivariate Gaussian random walk proposal density.

The detailed balance condition (1.4) only requires that weight functions $w^{(k)}$ be positive everywhere. Users therefore have the freedom to choose functions that favor some particular behaviour. To encourage large jumps for instance, $w^{(k)}$ could be chosen to contain the factor $\|y - x\|_2$ (see, e.g., Yang et al., 2019). We refer the reader to Martino and Read (2013) for an extensive study of different weight functions. The two most common choices—and the ones that seem to perform best empirically—are importance weights,

$$w^{(k)}_{\mathrm{imp}}(y|x) = \frac{\pi(y)}{q^{(k)}(y|x)} , \tag{2.4}$$

and weights proportional to the target density,

$$w^{(k)}_{\text{prop}}\,(y|x) = \pi(y)\ . \tag{2.5}$$

The ergodicity of MTM chains can be verified using the same conditions as for the Metropolis algorithm. The assumption that (1.5) holds for each proposal density $q^{(k)}$, $k = 1, \ldots, K$, is sufficient to ensure π-irreducibility and aperiodicicity as long as both π and $w^{(k)}(\cdot|x)$ are bounded above and below on any compact subset of $\mathcal{X}$ for each k and for each fixed x (see Supplement A, Proposition B.2). The same conditions are also sufficient to establish a strong law of large numbers for all π-integrable functions.

Algorithm 1 summarizes the MTM sampler in its most general form—the joint proposal density and weight functions are left completely free.

Algorithm 1 Multiple-try Metropolis (MTM)

Input Target density π with support $\mathcal{X} \subseteq \mathbb{R}^d$, MC sample size N, joint proposal distribution q with marginals $q^{(k)}$ and conditionals $q^{(-k)}$, $k = 1, \ldots, K$, weight functions w^k, $k = 1, \ldots, K$.

Procedure

1. *Initialization.* Initialize the state to $x_0 \in \mathcal{X}$.
2. *MCMC iteration.* For $n = 0, \ldots, N-1$, do:
   (a) *Candidates generation.* Sample $y^{(1:K)} \sim q(\cdot|x_n)$;
   (b) *Weights.* Compute $w^{(k)}(y^{(k)}|x_n)$, $k = 1, \ldots, K$;
   (c) *Proposal selection.* Sample $k \in \{1, \ldots, K\}$ with probability proportional to weights $w^{(k)}$, $k = 1, \ldots, K$, and set $y = y^{(k)}$;
   (d) *Shadow sample.* Sample $x_*^{(-k)} \sim q^{(-k)}(\cdot|y^{(k)}, x_n)$ and set $x_*^{(k)} = x_n$;
   (e) *Reverse weights.* Compute $w^{(k)}(x_*^{(k)}|y)$, $k = 1, \ldots, K$;
   (f) *Acceptance probability.* Compute $\alpha_{\text{MTM}}(y, y^{(-k)}|x, x_*^{(-k)})$ using (2.1);
   (g) *Acceptance.* Accept the proposal ($x_{n+1} = y$) with probability α_{MTM}; otherwise reject the proposal ($x_{n+1} = x_n$).

Output The MC sample $\{x_n\}_{n=1}^N$.

2.2. Optimal scaling of MCMC

While a law of large numbers (1.1) guarantees that the sample average converges toward the desired expected value, it does not provide any insight about the estimation error for finite samples. The advantage of *central limit theorems* (CLTs) is that they provide information about the asymptotic distribution of Monte Carlo estimates. A Markov chain $\mathbf{X} \subseteq \mathcal{X}$ satisfies a CLT for a function f if there exists a constant $\sigma_f^2 < \infty$, known as the *asymptotic variance*, such that

$$\sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N} f\left(X_n\right) - \pi(f)\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_f^2\right)\ , \qquad N \to \infty\ , \tag{2.6}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. In MCMC contexts, CLTs are often used to produce Monte Carlo standard errors by applying the asymptotic result to a finite sample. Replacing convergence with approximation, (2.6) suggests, for large N,

$$\frac{1}{N}\sum_{n=1}^{N} f\left(X_n\right) \approx \mathcal{N}\left(\pi(f), \sigma_f^2/N\right) ,$$

so we may use an estimate of $\sigma_f/\sqrt{N}$ as the Monte Carlo standard error.

Now, if we fix the function of interest f and the target density π, we find that different Markov transitions yield different values for the asymptotic variance σ_f^2, hence different Monte Carlo estimation precisions. Therefore, choosing a transition producing a small—ideally the smallest—asymptotic variance is an important aspect of MCMC theory.

We note that the transition of Metropolis-Hastings samplers is entirely defined by the proposal density q. In the simple case of i.i.d. targets and Metropolis proposal densities with spherical Gaussian steps, Roberts, Gelman and Gilks (1997) showed that, asymptotically as the dimension $d \to \infty$, the optimal scaling of the proposal variance $\sigma_d^2 I_d$ is given by $\sigma_d^2 = (2.38)^2/d$ and is associated to an optimal acceptance probability of 0.234. Similar results were eventually obtained for more general target densities (Roberts and Rosenthal, 2001; Bédard, 2007, 2008a; Bédard, 2008b; Bédard and Rosenthal, 2008; Breyer and Roberts, 2000; Beskos, Roberts and Stuart, 2009; Sherlock and Roberts, 2009), for other algorithms (Roberts and Rosenthal, 1998; Breyer, Piccioni and Scarlatti, 2004; Pillai, Stuart and Thiéry, 2012; Bédard, Douc and Moulines, 2014; Beskos et al., 2013), and also for finite-dimensional targets (Gelman, Roberts and Gilks, 1996; Sherlock, Fearnhead and Roberts, 2010). These results provide guidelines for MCMC users to choose near-optimal proposal densities.

Bédard, Douc and Moulines (2012) studied MTM samplers and obtained asymptotically optimal scaling results ($d \to \infty$) for each fixed number of candidates $K = 1, \ldots, 5$. They considered various sampling schemes, including independent and EA candidates; their results, partially summarized in Table 1, apply to i.i.d. targets with spherical multivariate Gaussian candidates and weights proportional to the target. As K increases, we notice a growth in the optimal acceptance probability, which indicates that the chain has access to higher-quality selected proposals. Extremely antithetic candidates lead to acceptance rates that are significantly larger than those of independent candidates, meaning that an adequate correlation structure may substantially improves the quality of the selected proposal.

2.3. Adaptive MCMC

Optimal scaling results such as those presented in Section 2.2 implicitly require that the covariance of the target distribution be known. Indeed, the Gaussian random walk proposal uses a covariance that should be a multiple of the true covariance. In practice, the true covariance is never known—recall that we wish to compute some expectation

MTM optimal acceptance probability

Sampling scheme	Number of candidates (K) 1	2	3	4	5
Independent	0.23	0.32	0.37	0.39	0.41
EA	0.23	0.46	0.52	0.54	0.55

Table 1. Optimal acceptance probability of the MTM with spherical multivariate Gaussian candidates and weights proportional to π, a target with i.i.d. components (Bédard, Douc and Moulines, 2012).

$\pi(f)$ and that variance is the special case $f = (I - \pi(I))^2$, $I(x) = x$—so it is a dubious assumption to make.

To use optimal scaling results without knowledge of the true covariance, Haario, Saksman and Tamminen (2001) propose the *adaptive Metropolis* sampler, which learns the true covariance over time using the growing sample. Suppose that the time-n estimate of the covariance is Σ_n. A Metropolis iteration is performed using $s_d\Sigma_n$ as the proposal variance of the Gaussian random walk, where $s_d = (2.38)^2/d$ (from optimal scaling results). Once x_{n+1} is selected as either the proposal y or the previous state x_n, we define Σ_{n+1} as the empirical covariance of the sample $(x_i)_{i=1}^{n+1}$ (to which a small multiple of the identity matrix is added to ensure non-singularity). Simple recursions allow the computation of Σ_{n+1} from Σ_n without much work so this extra step is computationally cheap.

Since transitions change at every iteration, the chain is no longer homogeneous. The ergodicity property (1.3) is not defined properly for inhomogeneous chains so we require different definitions for studying the convergence of adaptive MCMC: we will use the setting in Roberts and Rosenthal (2007, Section 2). Furthermore, since the proposal variance depends on all past samples, the Markovian property of the chain is destroyed. Still, under some mild conditions, Haario, Saksman and Tamminen (2001) verified that AM chains satisfy a strong law of large numbers for bounded functions. The same conditions also imply convergence of the adaptive proposal covariance to the true target covariance (up to the small identity matrix added). Andrieu and Moulines (2006, see also Saksman and Vihola, 2010) later weakened the sufficient conditions for a strong law of large numbers, for an expanded class of functions, and provided a central limit theorem.

Following the work of Atchadé and Rosenthal (2005), Roberts and Rosenthal (2007) developed a simple yet powerful framework to study the ergodicity of adaptive MCMC in a more general setting. They showed that two main conditions are sufficient to verify both the ergodicity of adaptive chains and a weak law of large numbers. To state these sufficient conditions, let $\mathcal{Y}$ be some indexing set for the family of possible transitions $\{P_\gamma : \gamma \in \mathcal{Y}\}$ and let Γ_n denote the (random) index of the chain's transition at time $n \geqslant 0$. When the transitions are all within the same parametric family, the indexing set corresponds to the parameter space where the Γ_n's lie. For example, the indexing of the AM transition may be performed using the adapted covariance Σ_n.

The first condition, termed *diminishing adaptation* (DA), requires that subsequent transitions change less, in probability, as the chain progresses:

$$\mathrm{dist}(P_{\Gamma_n}, P_{\Gamma_{n+1}}) \xrightarrow{\mathbb{P}} 0 \;, \qquad n \to \infty \;, \tag{2.7}$$

where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability and where

$$\mathrm{dist}(P_{\gamma_n}, P_{\gamma_{n+1}}) := \sup_{x \in \mathcal{X}} \left\| P_{\gamma_n}(\cdot|x) - P_{\gamma_{n+1}}(\cdot|x) \right\|_{\mathrm{TV}}$$

is the distance between two consecutive transitions. In particular, convergence of transitions is not a necessary condition, even though it is often satisfied and desired.

The second condition, coined *bounded convergence* (BC, also called *containment*), states that all transitions are individually ergodic with respect to the target density and that their convergence rates do not degenerate, at least in probability. For all $\varepsilon > 0$, the process $\{M_\varepsilon(X_n, \Gamma_n)\}_{n \geqslant 0}$ is bounded in probability conditionally on the initial state $X_0 = x_*$ and initial transition index $\Gamma_0 = \gamma_*$, where

$$M_\varepsilon(x, \gamma) := \inf_m \left\{ m \geqslant 1 : \left\| P_\gamma^m(\cdot|x) - \pi(\cdot) \right\|_{\mathrm{TV}} \right\} \tag{2.8}$$

is the ε-convergence time of the homogeneous chain using transition P_γ with parameter $\gamma \in \mathcal{Y}$ and starting at $x \in \mathcal{X}$. Following the work of Roberts and Rosenthal (2007), most adaptive MCMC algorithms have their ergodicity verified using DA and BC, or some derivatives of these conditions.

3. Description of the algorithm

3.1. Motivation

Before introducing our proposed algorithm, we consider a toy example that exhibits some of the shortcomings of the AM and MTM samplers taken separately. Let π be a two-dimensional mixture of two Gaussian densities with weights $w_1 = 0.3$ and $w_2 = 0.7$, with means $\mu_1 = (20, 0)^\top$ and $\mu_2 = (0, 8)^\top$, and with covariance matrices $\Sigma_1 = \mathrm{diag}(9, 1)$ and $\Sigma_2 = \mathrm{diag}(1, 9)$. An i.i.d. sample of size $N = 10,000$ from that density may be found in Figure 1(a).

Multimodal densities are notoriously hard to sample using simple algorithms. Indeed, two types of moves are required to adequately explore the whole support of such distributions: local moves to explore a given mode and global ones to jump between modes. A single proposal density generally cannot do both efficiently because of the different scales on which they lie (see, e.g., a Metropolis sampler in Figure 1(b) featuring a very low acceptance rate). Furthermore, using adaptation to find an optimal proposal covariance matrix leads to the optimization of a single type of moves. Depending on the initialization, the AM algorithm will either converge to the covariance of one mode (Figure 1(c)) or to the global covariance of the target (Figure 1(d)). While the latter yields decent results, the acceptance rate is still small for a two-dimensional target density.

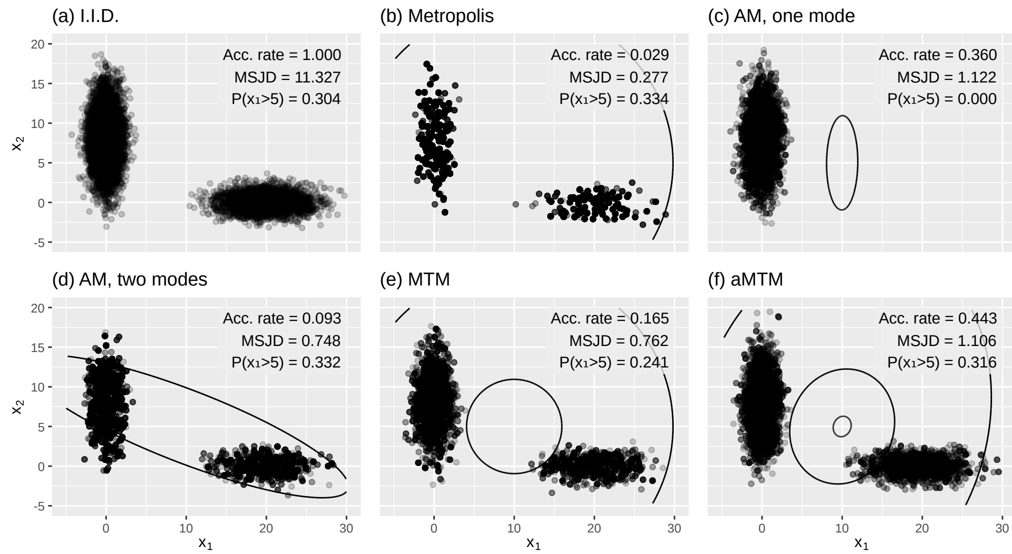

Figure 1. Samples for the bimodal density of Section 3.1 obtained using different samplers and parameters. Ellipses show the final proposal covariance(s); $P(x_1 > 5)$ is the proportion of points with $x_1 > 5$; MSJD is the mean squared jumping distance.

MTM samplers are better suited for multimodal densities because the different proposal densities can model various types of jumps. When proposal densities are well adjusted to the target, the resulting chain may offer good performances (e.g., Figure 1(e)). The tuning of proposal densities must however be done by hand, which rapidly becomes impractical with increasing dimensions.

Hence, adaptation of the MTM's proposal densities could bring the best of both worlds together: automatic tuning of the proposal distributions and better fit to the target's distinct characteristics. Our proposed method, whose description follows, can achieve improved performance with minimal tuning (see Figure 1(f)).

3.2. General algorithm

Adaptive MCMC algorithms are essentially defined by two components: a family of transitions and a way to move from one transition to another. The results of Roberts and Rosenthal (2007) suggest a way to construct ergodic adaptive MCMC algorithms: we define both components so that the DA and BC conditions be easy to verify. The BC condition is mostly related to the family of transitions while the DA condition is related to the agreement between successive transitions.

With that in mind, we propose to equip the MTM sampler with adaptation in the following way. The family of transitions is chosen to be MTM transitions in which candidates are generated using Gaussian random walks with a fixed, common correlation structure between the candidates and a fixed weight function. Keeping the correlation structure and weight function fixed will help verifying the DA condition since switching between correlation structures or weight functions would create discontinuous jumps between transitions. We will consider the four correlation structures described in Section 2.1 (independent, EA, RQMC, common random variable) as well as the weight functions (2.4) and (2.5). Algorithm 2 contains an abstract description of our proposed algorithm with arbitrary adaptation.

Algorithm 2 Adaptive multiple-try Metropolis (aMTM)

Input Target density π, MC sample size N, adaptation procedure $q_n \mapsto q_{n+1}$ inside some family of proposal densities $\mathcal{Q}$, weight functions w^k, $k = 1, \dots, K$.

Procedure

1. *Initialization.* Initialize the state to $x_0 \in \mathbb{R}^d$ and the joint proposal density to $q_0 \in \mathcal{Q}$.
2. *MCMC iteration.* For $n = 0, \dots, N-1$, do:
   (a) *MTM sampling.* Generate x_{n+1} from x_n using the current joint proposal distribution q_n and the weight functions w^k, $k = 1, \dots, K$, and according to one MTM sampling iteration (Algorithm 1, Step 2).
   (b) *Adaptation.* Update q_n to q_{n+1} according to the specified adaptation procedure.

Output The MC sample $\{x_n\}_{n=1}^N$.

3.3. Adaptation variants

When the family of proposal densities is chosen to be multivariate Gaussian random walks, then the joint proposal density q also is a (possibly singular) multivariate Gaussian density. In particular, each marginal density is uniquely determined by the covariance matrix $\Sigma^{(k)}$, $k = 1, \dots, K$. With a fixed correlation structure, the correlations between the candidates are all known given the marginal distributions. Hence, adaptation of q is reduced to adaptation of $\Sigma^{(k)}$, $k = 1, \dots, K$. The adaptation between two consecutive marginal covariances $\Sigma^{(k)}$ is inspired from existing schemes used to improve upon the Metropolis algorithm. Additional adaptation variants are also discussed in Supplement A, Section A.2.

AM updates A first update rule is given by the AM updates of Haario, Saksman and Tamminen (2001). At time n, the k-th proposal covariance is $s_d\Sigma_n^{(k)}$ for some scale $s_d > 0$. The recursion for the update $\Sigma_n^{(k)}$ satisfies

$$\mu_{n+1}^{(k)} = \mu_n^{(k)} + \gamma_{n+1}\left(x_{n+1} - \mu_n^{(k)}\right) , \tag{3.1}$$

$$\Sigma_{n+1}^{(k)} = \Sigma_{n+1}^{(k)} + \gamma_{n+1}\left[\left(x_{n+1} - \mu_n^{(k)}\right)\left(x_{n+1} - \mu_n^{(k)}\right)^\top - \Sigma_{n+1}^{(k)}\right] , \tag{3.2}$$

where $\mu_n^{(k)}$ is the running mean of the k-th component and $\gamma_{n+1} > 0$ is the adaptation step.

ASWAM updates Optimal scaling results provide guidelines about the choice of s_d given the dimension d of the target density. The AM algorithm uses that information to directly scale the proposal covariance. Now, these results also provide an optimal acceptance rate which can be used, instead of the scale itself, to tune the marginal covariances. Empirical evidence shows that the optimal acceptance rate is much less sensitive to a change of target density than the optimal scale. Thus, aiming at an optimal acceptance rate rather than an optimal scaling is a more robust adaptation principle.

Based on this argument, a second update rule is provided by the *adaptive scaling within adaptive Metropolis* (ASWAM) updates of Andrieu and Thoms (2008). The idea is to compute the running mean and covariance as in the AM updates (3.1) and (3.2), but to also adapt the scale s_d toward a value that yields an acceptance rate approaching some target rate $\alpha_* \in [0,1]$. The marginal covariance for candidate k at time n is $\lambda_n^{(k)} \Sigma_n^{(k)}$, where the scale $\lambda_n^{(k)}$ is updated using

$$\log\left(\lambda_{n+1}^{(k)}\right) = \log\left(\lambda_n^{(k)}\right) + \gamma_{n+1}\left[\alpha_{\mathrm{MTM}}\left(y, y^{(-k)} | x, x_*^{(-k)}\right) - \alpha_*\right] .$$

RAM updates An alternative to ASWAM updates is the *robust adaptive Metropolis* (RAM) of Vihola (2012), whose updates are better suited to target densities with no finite second moment. In a single step, the marginal covariance is updated to approach both the (pseudo-)covariance of the target and a target acceptance rate. Given a square root decomposition $\Sigma_n^{(k)} = S_n^{(k)} S_n^{(k)\top}$, the next marginal covariance is given by

$$\Sigma_{n+1}^{(k)} = S_n^{(k)} \left\{ I_d + \gamma_{n+1}\left[\alpha_{\mathrm{MTM}}\left(y, y^{(-k)} | x, x_*^{(-k)}\right) - \alpha_*\right] \frac{z_n^{(k)} z_n^{(k)\top}}{\left\| z_n^{(k)} \right\|_2^2} \right\} S_n^{(k)\top} ,$$

where $z_n^{(k)} = (S_n^{(k)})^{-1}[y - x_n]$ is the standardized proposed step.

4. Validity

The aMTM sampler is first and foremost an adaptive algorithm. The regularity assumptions that are imposed to verify the theoretical properties of the aMTM (ergodicity, LLN) are therefore very similar to other adaptive methods, such as the AM of Haario, Saksman and Tamminen (2001).

Although Section 3 describes several variants of the aMTM sampler, it is not possible to simultaneously consider all these variants when proving the ergodicity of this algorithm. In what follows, we consider some general results that can be used in verifying the ergodicity, but we relegate the details applicable to specific instances of the algorithm to the supplementary material. For example, the simple case of independent candidates and weights proportional to the target density requires no further assumption; other variants may require stronger assumptions, which are discussed in Supplement A, Section B.3.

Before pursuing, let us introduce some more notation. The target distribution has a density π with respect to Lebesgue measure, with support $\mathcal{X} = \{x \in \mathbb{R}^d \,|\pi(x) > 0\} \subseteq \mathbb{R}^d$. We are interested in the expectation of a function $f : \mathbb{R}^d \to \mathbb{R}^p$ such that $\pi(|f|) < \infty$. At time $n \in \mathbb{N}$, the K proposal covariance matrices of the MTM kernel are $\Sigma = (\Sigma^{(1)}, \dots, \Sigma^{(K)})$ with $\Sigma^{(k)} \in \mathcal{C}_d^+$, $k = 1, \dots, K$, where $\mathcal{C}_d^+$ denotes the cone of symmetric positive-definite $d \times d$ matrices. The parameter space therefore is $\Theta = (\mathcal{C}_d^+)^K$ and the marginal proposal densities are $q_\theta^{(k)}(y|x) = \varphi(y|x, \Sigma^{(k)})$, where $\varphi(\cdot|\mu, \Sigma)$ denotes a d-dimensional normal density with mean μ and covariance Σ.

We assume that $\mathcal{X}$ and Θ both are compact; these assumptions are similar to those in Haario, Saksman and Tamminen (2001) and greatly simplify proofs in Section 4.1, where the algorithm's ergodicity is studied. Generalizations to unbounded cases are discussed in Section 4.2, while limit theorems are considered in Section 4.3.

4.1. Ergodicity

First let us recall a result from Roberts and Rosenthal (2007), which will be the basis of the aMTM algorithm ergodicity's analysis.

Theorem 4.1 (Roberts and Rosenthal, 2007, Theorem 2)**.** *Consider an adaptive MCMC algorithm using a family of Markov transitions $\{P_\theta\}_{\theta\in\Theta}$ and let $x_0 \in \mathcal{X}$ and $\theta_0 \in \Theta$ be the initial state and transition index, respectively. Suppose each transition P_θ admits the target density π as its stationary distribution and suppose the algorithm satisfies the diminishing adaptation* (2.7) *and the bounded convergence* (2.8) *conditions for these initial values. Then, the algorithm is ergodic to the target density for these initial values.*

In this result, there are three main conditions to verify: stationarity, diminishing adaptation and bounded convergence. In our context, the family of Markov transitions $\{P_\theta\}_{\theta\in\Theta}$ consists in MTM transitions with some fixed correlation structure and fixed weight function. The index θ corresponds to the collection of marginal covariances Σ.

4.1.1. Stationary distribution

As mentioned in Section 2.1, a sufficient condition for P_θ to admit π as its stationary distribution is to satisfy (2.2). Conveniently, any symmetrical proposal density meets this requirement; in particular, a multivariate Gaussian random walk assumption is sufficient to establish the stationarity of π. The proof of (2.2)'s sufficiency (in Supplement A, Proposition B.1) is valid for any choice of correlation structure and any weight function.

4.1.2. Bounded convergence

Our study of the aMTM's bounded convergence relies on Craiu et al. (2015, Proposition 23):

Proposition 4.1 (Craiu et al., 2015, Proposition 23). *Consider an adaptive MCMC algorithm using a family of Markov transitions $\{P_\theta\}_{\theta\in\Theta}$ such that Θ is compact, and such that each P_θ admits the target density π as stationary distribution and is Harris-ergodic to π. If, for all $n \geqslant 1$, the application $(x, \theta) \mapsto \Delta_n(x, \theta)$ with*

$$\Delta_n(x, \theta) := \|P_\theta^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}}$$

is jointly continuous in (x, θ) for all $(x, \theta) \in \mathcal{X} \times \Theta$ and if $\{X_n\}_{n\geqslant 1}$ is bounded in probability, then the algorithm satisfies the bounded convergence condition (2.8).

Based on the previous result, the verification of the bounded convergence further requires (1) that each MTM transition be Harris-ergodic with respect to π, (2) that the parameter space Θ be compact, (3) that $\Delta_n(x, \theta)$ be continuous for each n, and (4) that $\{X_n\}_{n\geqslant 1}$ be bounded in probability. The intuition behind this result is that each transition is ergodic with some rate of convergence. We then suppose that this rate of convergence varies continuously on the compact set $\mathcal{X} \times \Theta$ (at least in probability) so that it remains bounded, whence "bounded convergence".

We can show that Harris erogodicity of MTM transitions only requires the verification of π-irreducibility and aperiodicity. Indeed, we have the following result.

Proposition 4.2. *Let P be a MTM transition for a target density π. If P is π-irreducible, then P is Harris-recurrent.*

Proof. The complete argument may be found in Supplement A, Proposition B.3, but is almost identical to the proof by Tierney (1994, Corollary 2) in the Metropolis case. □

We note that a similar implication exists for Metropolis-Hastings transitions. It will then be convenient to observe that π-irreducibility and aperiodicity follow from the condition stated in the following result, which is reminiscent of similar results for Metropolis-Hastings algorithms with extra assumptions on the weight functions.

Proposition 4.3. *Let P be a MTM transition for a target density π with connected support $\mathcal{X}$. Suppose that π is bounded above on $\mathcal{X}$ and below on any compact subset of $\mathcal{X}$. Suppose that, for each $k = 1, \ldots, K$, there exist δ, $\varepsilon > 0$ such that the marginal proposal densities are symmetric and satisfy*

$$q^{(k)}(y|x) > \varepsilon \ , \qquad \forall\, x,\, y \in \mathcal{X} : \|y - x\|_2 < \delta \ .$$

Suppose that, for all fixed $x \in \mathcal{X}$, the weights $w^{(k)}(\cdot|x)$ are bounded above and below on any compact set. Then, the kernel P is π-irreducible and aperiodic.

Proof. The complete argument may be found in Supplement A, Proposition B.2, but is completely analogous to the proof by Robert and Casella (2004, Lemma 7.6) in the MH case. □

The verification of Δ_n's continuity is inspired from a proof by Roberts and Rosenthal (2007, Corollary 11) in the case of a Metropolis-Hastings sampler. We refer the reader to Supplement A, Section B.3.4, for a complete proof, which requires no further assumption other than those already mentioned.

Under the assumption that $\mathcal{X}$ is compact, we directly have that $\{X_n\}_{n\geqslant 1}$ is bounded in probability. The more general case where $\mathcal{X}$ is unbounded requires more care and will be discussed in Section 4.2.

4.1.3. Diminishing adaptation

The diminishing adaptation condition (2.7) is easier to verify in the context of stochastic approximations. In particular, we recognize the covariance updates described in Section 3.3 as those of a Robbins-Monro algorithm (Robbins and Monro, 1951), which consists of updates taking the following form:

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, (k, y^{(1:K)}, x_*^{(1:K)})) \ , \tag{4.1}$$

for some function $H : \Theta \times \{1, \dots, K\} \times \mathcal{X}^{2K} \to \mathbb{R}^{d_\theta}$ with $\Theta \subseteq \mathbb{R}^{d_\theta}$. In the case where running means $\mu_n^{(k)}$ or scales $\lambda_n^{(k)}$ are used, we add them to θ and augment Θ accordingly.

Following the work of Andrieu and Moulines (2006) and Saksman and Vihola (2010), we can prove the following result for adaptive MCMC algorithms.

Proposition 4.4. *Suppose that the transition update function $H_\theta(\cdot) = H(\theta, \cdot)$ is 1-Lipschitz in θ, that is, there exists $C < \infty$ such that for every pair $(\theta, \theta') \in \Theta \times \Theta$ and for every bounded function f we have*

$$\|P_\theta f - P_{\theta'} f\|_1 \leqslant C \|f\|_1 \|\theta - \theta'\|_2 \ , \tag{4.2}$$

where $\|\cdot\|_1$ is defined for a function $f : \mathcal{X} \to \mathbb{R}^p$ by $\|f\|_1 = \sup_{x\in\mathcal{X}} \|f(x)\|_2$ and where

$$P_\theta f(z) = \int f(x) P_\theta(x|z)\, \mathrm{d}x.$$

Suppose that $\{\theta_n\}_{n\geqslant 1}$ in bounded in probability; if

$$\sup_{\theta\in\Theta} \|H_\theta\|_1 < \infty \ , \qquad \forall\, x \in \mathcal{X} \ , \theta \in \Theta \ , \tag{4.3}$$

and if the sequence of adaptation steps converges to 0, $\gamma_n \to 0$, then the adaptive MCMC algorithm satisfies the diminishing adaptation condition (2.7).

Proof. See Supplement A, Proposition B.4, for a stronger statement of which Proposition 4.4 is a special case when Θ is compact. $\square$

The previous result reduces the verification of the diminishing adaptation to the verification of the Lipschitz transitions condition (4.2) and the bounded updates condition (4.3). Verifying the Lipschitz transitions condition (4.2) does not require any additional assumption under the simple case of independent proposals and weights proportional to the target density. For other aMTM variants, the verification of the Lipschitz transition must be made on a case by case basis; see Supplement A, Section B.3.2, for a discussion. Verifying the bounded updates condition (4.3) is trivial under the assumption that $\mathcal{X}$ is compact because any update rule described in Section 3.3 will only involve bounded quantities (Supplement A, Section B.3.3).

4.2. Generalizations to unbounded spaces

The major assumptions made in Section 4.1 were the compactness of the state space $\mathcal{X}$ and parameter space Θ. These conditions greatly simplify the verification of the algorithm's ergodicity, but also substantially restrict the theoretical applicability of the proposed sampler. In this section, we discuss different approaches that could be used to relax or even remove these assumptions.

Assuming $\mathcal{X}$ to be compact may seem a major impediment to the practical use of the algorithm since target densities often have unbounded supports. One might then worry about the fact that the theoretical results of the previous section only apply to a very restricted class of target densities. Now, a simple workaround is to consider the target $\tilde{\pi} = \pi|_{\widetilde{\mathcal{X}}}$, a version of the initial π restricted to a compact set $\widetilde{\mathcal{X}} \subset \mathcal{X}$, which can be chosen arbitrarily large. In that case, the expectation $\tilde{\pi}(f)$ of the resulting MC estimate will be virtually indistinguishable from the original expectation $\pi(f)$ provided that $\widetilde{\mathcal{X}}$ is chosen large enough. In practice, this approach corresponds to rejecting any proposal that lies outside of $\widetilde{\mathcal{X}}$.

In contrast, the compactness of Θ does not reduce the scope of theoretical results; in reality, it only restricts the family of MTM transitions on which adaptation can be performed. In the aMTM algorithm, the space Θ lies within the product of K convex cones of symmetric positive definite matrices. We can therefore simply choose Θ compact by bounding the eigenvalues of the covariance matrices inside some interval. Since users typically have some idea of their problem's scaling, it is easy to find reasonable bounds so that Θ contains the most efficient MTM transitions. In any case, practical implementations are subject to program and machine limitations so any symmetric positive definite matrix will lie in some definitive compact set when stored.

4.2.1. Compact coverages

It is important to note that, because of the extensive similarities between MH and MTM algorithms, results applicable to the AM sampler are expected to have aMTM counterparts holding under fairly similar conditions. Here is an example of a construction used to prove the ergodicity of the AM algorithm on unbounded domains.

Dating back to the work of Chen, Lei and Gao (1988), *compact coverages* or *troncation at randomly varying bounds* or *sequentially constrained adaptive MCMC algorithms* is the idea of performing a Robbins-Monro stochastic approximation—which covers most adaptive MCMC algorithms as a special case—within some compact set and expanding that set when necessary. More explicitly, a compact coverage of Θ is a sequence of compact sets $\{\mathcal{K}_r\}_{r\geqslant 0}$ increasing to Θ, i.e. such that $\cup_{r\geqslant 0}\mathcal{K}_r = \Theta$ and $\mathcal{K}_r \subset \operatorname{int}(\mathcal{K}_{r+1})$. Then, the adaptation step of the sampler is modified so that the parameter θ_{n+1} is updated only if the new value lies in $\mathcal{K}_{n+1}$.

Subject to some regularity conditions on the target density π, Saksman and Vihola (2010, Section 5) show that the sequentially constrained AM algorithm is ergodic with respect to π for $\mathcal{X}$ and Θ unbounded (a generalization of Andrieu and Moulines, 2006, Theorem 2). Furthermore, Vihola (2011, Section 5) extends these results to the ASWAM sampler and Vihola (2012, Theorem 6) uses results from Vihola (2011) to verify the ergodicity of the RAM algorithm on unbounded domains.

Unfortunately, the *compact coverages* method of proof relies heavily on the geometry of the acceptation and rejection regions around the current state. In the case of MTM transitions, these regions become incredibly complicated because of the multiple candidates and shadow points; it is therefore far from easy to extend these proofs to the aMTM case.

4.2.2. Bounded adaptation and combocontinuity

Another setting under which the ergodicity of adaptive MCMC algorithms with unbounded state space can be studied is that of *bounded adaptation*, introduced by Craiu et al. (2015). Consider $\widetilde{\mathcal{X}}$, a compact subset of $\mathcal{X}$, and let us modify the adaptive MCMC as follows: whenever the current state is outside of $\widetilde{\mathcal{X}}$, a fixed transition is used and the parameter θ is not updated. We also assume *bounded jumps*; this means that there exists $D < \infty$ such that the probability of moving from $x \in \mathcal{X}$ to a point that is at most D away is 1 uniformly in $\theta \in \Theta$. This can be enforced by construction, by using proposal densities truncated beyond D. Craiu et al. (2015, Theorem 21) then show that the AM algorithm verifies the bounded convergence condition, provided that the sampler features a continuous transition (or a continuous proposal in the case of MH algorithms). Note that we still require the compactness of the parameter space Θ in that case. Rosenthal and Yang (2018) extend this result to more general adaptive MCMC algorithms that verify a *combocontinuity* condition, i.e. samplers using a transition density that can be written as a finite combination of continuous densities. In particular, MTM transitions fall under the scope of combocontinuity assuming that all proposals are continuous densities.

4.3. Limit theorems

Two other interesting characteristics of MC estimates are satisfying a law of large numbers and a central limit theorem. Indeed, ergodicity guarantees that the marginal distribution of the chain converges to the target distribution, but does not directly inform us on the properties of the estimate itself.

Under ergodicity, it is not hard to verify a weak law of large numbers for any bounded function: Roberts and Rosenthal (2007, Theorem 23) show that Bounded Convergence and Diminishing Adaptation are sufficient conditions in that case. However, extending the result to a strong LLN or broadening the class of functions over which it applies generally are trickier tasks.

Typically, strong LLN for adaptive MCMC algorithms require some sort of V*-ergodicity* condition, where V is some test function that ultimately controls the convergence rate. The obtained results thus apply to any V^α-bounded function f for some $\alpha \in [0, 1)$, i.e. such that $\sup_{x\in\mathcal{X}} |f(x)|/V^\alpha(x) < \infty$ (Andrieu and Moulines, 2006, Theorem 8). In our context, this method of proof however requires $V \equiv 1$, which then only allows bounded functions. Now, the context of compact coverages described in Section 4.2.1 could potentially enable verifying a strong LLN for the aMTM algorithm as was done for the AM algorithm (Saksman and Vihola, 2010, Theorem 10), the ASWAM algorithm (Atchadé and Fort, 2010, Proposition 5), and the RAM algorithm (Vihola, 2012, Theorem 6), all of which allow π^{-1}-bounded functions.

The story is very similar when it comes to obtaining a central limit theorem for the aMTM sampler. Andrieu and Moulines (2006, Theorem 9) provide a CLT for adaptive MCMC using compact coverages assuming $\pi^{-\alpha}$-ergodicity, which holds for any $\pi^{-\alpha/2}$-bounded function with $\alpha \in [0, 1)$. Saksman and Vihola (2010, Theorem 18) derive a similar result for the specific case of the AM algorithm with $\alpha = 1$.

5. Numerical experiments

5.1. Simulation experiments

Fontaine (2019) contains multiple simulation experiments investigating the many variants of the aMTM algorithm. For the MTM sampling component of the algorithm, the user can specify a correlation structure, a weight function and the number of candidates; for the adaptation component, the user can specify the update scheme, the target acceptance rate and the step-size sequence. We refer the reader to Fontaine (2019) for the details, but report here our main observations.

Extremely antithetic and randomized quasi-Monte Carlo candidates tend to perform slightly better than independent proposals and significantly better than common random variable proposals. These two correlation structures encourage a better spread of candidates over the sample space and it is therefore not surprising to record better mixing and space exploration. These results seem consistent both with theory—EA candidates have larger optimal acceptance rates (Bédard, Douc and Moulines, 2012)—and with practice (Craiu and Lemieux, 2007). Experimenting with importance weights and weights proportional to the target did not yield a clear favorite; this seems to agree with similar experiments (see, e.g., the extensive analysis of Martino and Read (2013).)

It is important to keep in mind that the complexity of the aMTM scales with $2K-1$ as the target evaluation is generally the computational bottleneck of each iteration. Striking a balance between more efficient and costlier iterations is thus a crucial problem. Gladly, we find that a small number of candidates—between 2 and 5, depending on the target—is generally enough to obtain the largest improvements in performance compared to single-candidate samplers.

In terms of adaptation, we find that updates using a target acceptance rate (RAM, ASWAM) outperform the simple AM and that RAM updates seem to improve marginally on the ASWAM in some cases. As for the target acceptance rate, we observe that rates relatively smaller than the optimal ones (Bédard, Douc and Moulines, 2012) perform best: these optimal rates are obtained for well-behaved targets, so it not surprising to find that smaller rates are preferable. Generally, we find that rates in the range $[0.2, 0.5]$ produce results with somewhat uniform performances. Finally, for step sizes of the form $n^{-\gamma}$, we observe that AM and ASWAM updates tend not to be significantly affected by the value of $\gamma \in [0.5, 1]$, while RAM updates seem to benefit from values closer to the lower bound.

5.2. Loss of heterozygosity in esophageal cancer cells

Problem description. During a cancer's progression, affected cells undergo genetic changes such as loss of chromosomes sections. This abnormality, called *loss of heterozygosity* (LOH), can be detected in laboratory: the Seattle Barrett's Esophagus research project (Barrett et al., 1996) collected LOH rates for 40 different regions of the genome. For each region $i = 1, \dots, 40$, we denote by X_i the number of cells with detected LOH and by N_i the total number of cells analyzed; Figure 2 (top) contains an histogram of the proportions of LOH within each region.

It is hypothesized that there are two causes for LOH in cancer cells: a "background" LOH, possibly caused by the cancer's progression, and a "systematic" LOH, due to the presence of tumor suppressor genes (TSGs). Regions with higher rates of LOH are therefore suspected to contain TSGs: modeling LOH rates is thus of interest for cancer researchers in order to identify regions with such TSGs. The existence of these two regimes suggests modeling LOH rates using mixture models: component membership probabilities may provide insight on the presence of TSGs. For more information on localization of TSGs and on LOH, we refer to Desai (2000) and references therein.

Desai (2000) suggests several two-components mixture models for LOH in cancer cells, where each component is either a binomial or a beta-binomial distribution. Following the analysis of this dataset by Warnes (2001), we consider a mixture of a binomial component and a beta-binomial component. The likelihood is given by

$$X_i \mid N_i, \eta, \pi_1, \pi_2, \gamma \sim \eta \mathrm{Binomial}\left(N_i, \pi_1\right) + (1-\eta)\mathrm{BetaBinomial}\left(N_i, \pi_2, \gamma\right) \ ,$$

where $\eta \in [0, 1]$ controls the mixture weights, $\pi_1, \pi_2 \in [0, 1]$ control the center of each component, and $\gamma \in \mathbb{R}$ controls the (logit) spread of the Beta-binomial component ($\gamma \to -\infty$ corresponds to a binomial distribution and $\gamma \to \infty$ to a discrete uniform distribution). The prior on the four model parameters is taken to be uniform over a set of plausible values:

$$(\eta, \pi_1, \pi_2, \gamma) \sim \mathrm{Uniform}\left([0,1]^3 \times [-30, 30]\right) \ .$$

Depending on which component is used to model each regime, the mixture model features some clear multi-modality. Indeed, the posterior distribution exhibits a first mode (Figure 2, top left) where the binomial component models the lower (background)

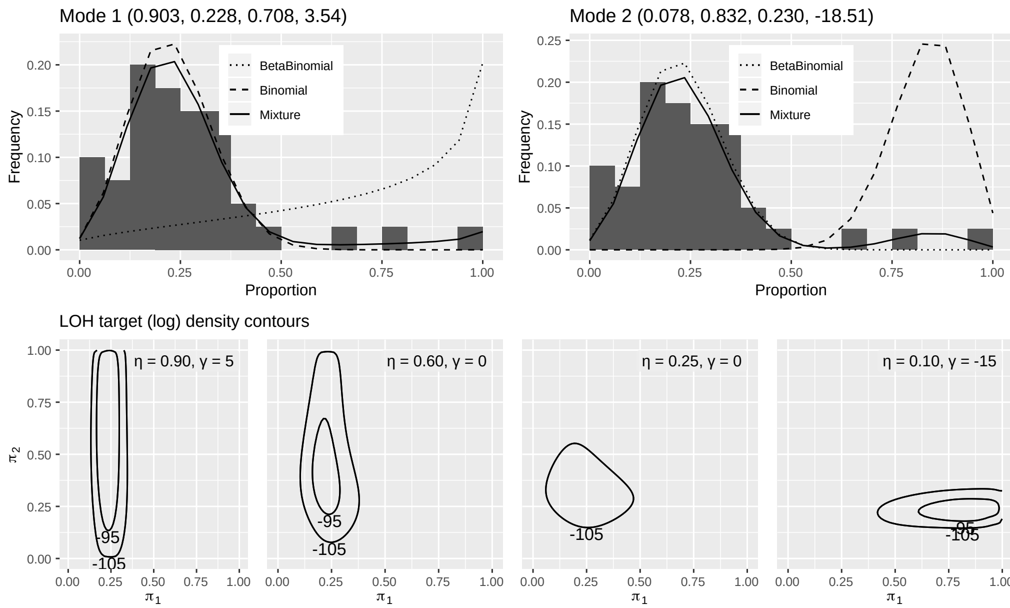

Figure 2. (Top) Histogram of observed proportions of LOH (Barrett et al., 1996) together with the distribution of each mixture component (dotted: Beta-Binomial; dashed: Binomial) and of the resulting mixture (solid) for the two posterior modes $(\eta, \pi_1, \pi_2, \gamma)$. (Bottom) Contours of the posterior density plots in (π_1, π_2) for some fixed values of (η, γ).

LOH rates and the beta-binomial component models the higher (systematic) LOH rates; the second mode (Figure 2, top right) exchanges this assignment. A few cross-sections of the posterior distribution are depicted in Figure 2 (bottom).

Experiment. Obtaining samples from a multi-modal posterior is a notoriously hard task for standard MCMC algorithms and multiple samplers have been proposed to approach such problems. We detail some proposals that were applied to this very LOH mixture model problem. First, Warnes (2001) proposes the *normal kernel coupler* (NKC), which uses interacting chains forming a normal kernel density estimate of the target. Conveniently, in their analysis of the LOH data, they provide global and per-mode posterior means computed using numerical integration (*adaptive quadrature*, AQ) against which we can compare our results. Second, a variety of MCMC algorithms using regional adaptation have been proposed: Mixed RAPT (Craiu, Rosenthal and Yang, 2009), RAPTOR (Bai, Craiu and Di Narzo, 2011), and OPRA (Grenon-Godbout and Bédard, 2020+). Third, VanDerwerken and Schmidler (2013) propose to sample from multimodal distributions using *parallel sampling* (PS). Fourth, Casarin, Craiu and Leisen (2013) propose an *interacting multiple-try Metropolis* algorithm and apply it to the LOH model, but no numerical results are reported.

We compare our proposed aMTM algorithm to these methods both in terms of estimate quality and in terms of MCMC chain metrics. Across the different analyses of the LOH model posterior in the literature, we find multiple reportings such as means, standard deviations and quantiles, either calculated globally or restricted to each of the two modes; we will provide all those estimates for comparison. Inspecting slices of the posterior density (Figure 2, bottom) and various scatter plots emerging from other samplers (Warnes, 2001; Craiu, Rosenthal and Yang, 2009; Bai, Craiu and Di Narzo, 2011), we define the two regions of interest. Specifically, by dividing the space using the boundary $\pi_1 = 0.4$ we find that each region contains one mode, with Mode 1 being assigned to the region $\pi_1 < 0.4$. In addition to parameter estimates, we will be interested in estimating the weight of each mode: numerical integration (Warnes, 2001) indicates that Mode 2 carries 3.0% of the target weight. Furthermore, we provide some chain statistics: the *mean squared jumping distance* (MSJD), the acceptance rate and the average marginal autocorrelation ($\overline{|\mathrm{ACF}|}$) as described in Bai, Craiu and Di Narzo (2011). [1]

For comparative purposes, we include AM, ASWAM, and RAM as single-candidate samplers as well as a non-adaptive MTM sampler with $K = 3$ candidates. The MTM samplers' proposal covariances are chosen to be (1) all equal to an estimate of the target's covariance (*Common*), (2) a downscaled version of that same global covariance (factors of 1, 0.1 and 0.01, *Scaled*), and (3) adjusted to each mode (a global component, plus one component for each mode, *Oracle*). We produce chains for different tuning parameters (target acceptance rate, adaptation parameter, etc.) and report the best instances. For our aMTM algorithm, we use RAM updates with a target acceptance rate of $\alpha = 0.2$. In all cases, we produce chains of length $N = 10,000$ with a burn-in of $1,000$ iterations; we proceed to 100 replications with random initializations of the chain uniform on the support and we report means and standard errors across those replications.

Results. Table 2 contains chain statistics obtained from all of our methods, along with results from other sources. One of the hardest elements to get right while sampling multimodal distributions using MCMC samplers is the respective weight of each mode. We observe that only MTM methods achieve weight estimates that are close to the truth (0.030), while single candidate samplers often find a single mode. This phenomenon also explains why AM and ASWAM exhibit higher MSJD and acceptance rates as sampling from a unimodal distribution leads to better mixing. Inspecting the non-adaptive MTM samplers, we find that relatively well-adjusted proposals can lead to decent chain properties: using proposals on varying scales yields large MSJD and acceptance rates while these samplers still spend an appropriate amount of time in each mode. Our aMTM sampler, which does not require such fine tuning, achieves similar if not better chain statistics and maintains accurate estimates of the modes' respective weights.

Turning to the global parameter estimates presented in Table 3, we compare our method to other proposed algorithms. We find that estimating correctly the modes' weights greatly improves the accuracy of the estimates. Indeed, the NKC slightly oversamples from the smaller mode which introduces a fairly large bias, especially for quantile

[1] The authors mention averaging the first 40 absolute lag-correlations while their code averages the first 1600; we use the latter here to obtain comparable results.

LOH Binomial-BetaBinomial mixture model: Chain statistics				
Algorithm	Details	MSJD	Acc. rate	$\mathbb{P}$(mode 2)
AM		1.338 (0.043)	0.191 (0.006)	0.280 (0.045)
ASWAM	$\alpha = 0.4$	1.674 (0.043)	0.312 (0.006)	0.262 (0.044)
RAM	$\alpha = 0.3$	0.257 (0.006)	0.034 (0.001)	0.086 (0.025)
MTM($K = 3$)	Common proposals	0.104 (0.002)	0.011 (0.000)	0.032 (0.008)
	Scaled proposals	0.858 (0.002)	0.259 (0.000)	0.047 (0.008)
	Oracle proposals	0.381 (0.003)	0.041 (0.000)	0.028 (0.009)
aMTM($K = 3$)	RAM, $\alpha = 0.2$	0.864 (0.005)	0.265 (0.001)	0.038 (0.010)
NKC				0.047
RAPTOR			0.194	

Table 2. MCMC chain statistics for the LOH mixture model. MSJD is the mean squared jumping distance; Acc. rate is the acceptance rate of the chain; $\mathbb{P}$(mode 2) is the proportion of points in Region 2 (defined by $\pi_1 > 0.4$). Statistics are shown as mean (standard error) over 100 random initializations.

estimates with probabilities that are close to the smaller mode's weight. Our estimates agree with those obtained from numerical integration and those from RAPTOR, and seem to improve on the estimates obtained from NKC, Mixed RAPT, OPRA, and PS. Furthermore, the mixing of the marginal chains, evaluated through $\overline{|\mathrm{ACF}|}$, seems to be slightly better in aMTM chains than in the RAPTOR chain, which is not surprising given the larger acceptance rate of aMTM (26.5 % vs. 19.4 %).

When restricting the estimates to either of the two regions (Table 4), we find that our method yields estimates that agree more closely with numerical integration than NKC and Mixed RAPT, especially for the smaller mode.

6. Discussion

The proposed adaptive multiple-try Metropolis algorithm is a natural extension of both the adaptive Metropolis (AM) sampler (Haario, Saksman and Tamminen, 2001) and the multiple-try Metropolis (MTM) algorithm (Liu, Liang and Wong, 2000). It combines the flexibility of the MTM and the ease of use of adaptive samplers. Indeed, in our multimodal LOH example, the aMTM sampler produces accurate samples with limited tuning: in terms of mixing, space exploration and estimates, our method outperforms non-adaptive MTM and single candidate adaptive samplers, and is at least on par with more sophisticated methods such as RAPTOR (Bai, Craiu and Di Narzo, 2011).

Being a full-dimensional sampler, the aMTM algorithm is better suited to relatively low-dimensional target distributions. The flexibility induced by the multiple proposals and the adaptation makes it an interesting option for MCMC users dealing with target exhibiting multimodality or, more generally, complex geometry. The computational overhead of this method mostly emerges from the multiple target evaluations, but performance improvement can be observed with just a few additional proposals.

LOH Binomial-BetaBinomial mixture model: Global estimates

Parameter (AQ mean)	Method	Mean	Std dev.	$Q_{0.025}$	$Q_{0.975}$	$\overline{\lvert\text{ACF}\rvert}$
η (0.832)	aMTM	0.823 (0.008)	0.124 (0.008)	0.507 (0.025)	0.963 (0.000)	0.105 (0.015)
	NKC	0.82		0.075	0.965	
	RAPTOR	0.828	0.155			0.15
	Mixed RAPT	0.838				
	OPRA	0.901				
	PS	0.816 (0.001)				
π_1 (0.246)	aMTM	0.252 (0.006)	0.062 (0.008)	0.193 (0.000)	0.408 (0.025)	0.092 (0.017)
	NKC	0.257		0.193	0.829	
	RAPTOR	0.248	0.106			0.19
	Mixed RAPT	0.275				
	OPRA	0.230				
	PS	0.299 (0.001)				
π_2 (0.617)	aMTM	0.613 (0.005)	0.170 (0.002)	0.293 (0.004)	0.906 (0.001)	0.104 (0.009)
	NKC	0.612		0.230	0.912	
	RAPTOR	0.614	0.174			0.05
	Mixed RAPT	0.679				
	OPRA	0.729				
	PS	0.678 (0.002)				
γ (12.82)	aMTM	12.542 (0.317)	11.321 (0.205)	−13.479 (0.963)	29.119 (0.030)	0.106 (0.010)
	NKC	12.3		−21.2	29.3	
	RAPTOR	12.732	11.561			0.09
	Mixed RAPT	13.435				
	OPRA	12.401				
	PS	9.49 (0.51)				

Table 3. Global estimates for the LOH mixture model. Q_p denotes the p-th quantile; $\overline{\lvert\text{ACF}\rvert}$ is the average marginal autocorrelation as defined in Bai, Craiu and Di Narzo (2011). Statistics are shown as mean (standard error) over 100 random initializations for the aMTM sampler. See text for a description of the methods.

Acknowledgements

This work has been supported by the Natural Sciences and Engineering Research Council of Canada.

Supplementary Material

Supplement A: Supplement to "An adaptive multiple-try Metropolis algorithm"
(doi: TBD; .pdf). Additional details on aMTM variants. Intermediary results and proofs.

Supplement B: Package `aMTM`: Adaptive multiple-try Metropolis algorithm
(; R package). The main sampling routine is implemented in `C++`; the `R` package consists of a wrapper for the sampler as well as some utility functions for chain statistics and plotting. The output is compatible with the `R` package `coda`.

LOH Binomial-BetaBinomial mixture model: Mode estimates

Parameter (AQ mean)	Method	Mean	Std dev.	Q 0.025	Q 0.975
Mode 1 estimates					
η (0.854)	aMTM	0.852 (0.008)	0.082 (0.008)	0.649 (0.025)	0.964 (0.000)
	NKC	0.856		0.656	0.966
	Mixed RAPT	0.897			
π_1 (0.229)	aMTM	0.229 (0.000)	0.019 (0.000)	0.193 (0.000)	0.267 (0.000)
	NKC	0.229		0.192	0.266
	Mixed RAPT	0.229			
π_2 (0.629)	aMTM	0.628 (0.002)	0.162 (0.001)	0.317 (0.002)	0.907 (0.001)
	NKC	0.631		0.319	0.913
	Mixed RAPT	0.714			
γ (13.73)	aMTM	13.709 (0.105)	10.401 (0.105)	-9.519 (0.825)	29.175 (0.021)
	NKC	13.7		-4.97	29.3
	Mixed RAPT	15.661			
Mode 2 estimates					
η (0.091)	aMTM	0.089 (0.001)	0.044 (0.001)	0.021 (0.001)	0.188 (0.003)
	NKC	0.084		0.017	0.219
	Mixed RAPT	0.079			
π_1 (0.825)	aMTM	0.814 (0.006)	0.062 (0.008)	0.694 (0.000)	0.930 (0.025)
	NKC	0.832		0.741	0.914
	Mixed RAPT	0.863			
π_2 (0.232)	aMTM	0.231 (0.000)	0.018 (0.000)	0.198 (0.000)	0.267 (0.001)
	NKC	0.23		0.199	0.261
	Mixed RAPT	0.237			
γ (-16.28)	aMTM	-16.559 (0.296)	7.422 (0.124)	-28.743 (0.162)	-4.303 (0.328)
	NKC	-17.5		-29.5	-4.11
	Mixed RAPT	-14.796			

Table 4. Per-mode estimates for the LOH mixture model. Q_p denotes the p-th quantile. Statistics are shown as mean (standard error) over 100 random initializations for the aMTM sampler. See text for a description of the methods.

Supplement C: Code producing the results
(; R code). Contains the `R` code defining and running the experiments, processing the results and generating the output.

References

Andrieu, C. and Moulines, É. (2006). On the Ergodicity Properties of Some Adaptive MCMC Algorithms. *Ann. Appl. Probab.* **16** 1462–1505. MR2260070

Andrieu, C. and Thoms, J. (2008). A Tutorial on Adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882

Atchadé, Y. and Fort, G. (2010). Limit Theorems for Some Adaptive MCMC Algorithms with Subgeometric Kernels. *Bernoulli* **16** 116–154. MR2648752

Atchadé, Y. F. and Rosenthal, J. S. (2005). On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli* **11** 815–828. MR2172842

Bai, Y., Craiu, R. V. and Di Narzo, A. F. (2011). Divide and Conquer: A Mixture-based Approach to Regional Adaptation for MCMC. *J. Comput. Graph. Statist.* **20** 63–79. Supplementary material available online. MR2816538

Barrett, M. T., Galipeau, P. C., Sanchez, C. A., Emond, M. J. and Reid, B. J. (1996). Determination of the Frequency of Loss of Heterozygosity in Esophageal Adenocarcinoma by Cell Sorting, Whole Genome Amplification and Microsatellite Polymorphisms. *Oncogene* **12** 1873–1878.

Bédard, M. (2007). Weak Convergence of Metropolis Algorithms for Non-*IID* Target Distributions. *The Annals of Applied Probability* **17** 1222–1244. MR2344305

Bédard, M. (2008a). Efficient Sampling Using Metropolis Algorithms: Applications of Optimal Scaling Results. *Journal of Computational and Graphical Statistics* **17** 312–332. MR2439962

Bédard, M. (2008b). Optimal Acceptance Rates for Metropolis Algorithms: Moving beyond 0.234. *Stochastic Process. Appl.* **118** 2198–2222. MR2474348

Bédard, M., Douc, R. and Moulines, E. (2012). Scaling Analysis of Multiple-try MCMC Methods. *Stochastic Process. Appl.* **122** 758–786. MR2891436

Bédard, M., Douc, R. and Moulines, É. (2014). Scaling Analysis of Delayed Rejection MCMC Methods. *Methodol. Comput. Appl. Probab.* **16** 811–838. MR3270597

Bédard, M. and Rosenthal, J. S. (2008). Optimal Scaling of Metropolis Algorithms: Heading toward General Target Ddistributions. *Canadian Journal of Statistics* **36** 483–503. MR2532248

Beskos, A., Roberts, G. and Stuart, A. (2009). Optimal Scalings for Local Metropolis-Hastings Chains on Nonproduct Targets in High Dimensions. *Ann. Appl. Probab.* **19** 863–898. MR2537193

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M. and Stuart, A. (2013). Optimal Tuning of the Hybrid Monte Carlo Algorithm. *Bernoulli* **19** 1501–1534. MR3129023

Breyer, L. A., Piccioni, M. and Scarlatti, S. (2004). Optimal Scaling of MaLa for Nonlinear Regression. *Ann. Appl. Probab.* **14** 1479–1505. MR2071431

Breyer, L. A. and Roberts, G. O. (2000). From Metropolis to Diffusions: Gibbs States and Optimal Scaling. *Stochastic Process. Appl.* **90** 181–206. MR1794535

Casarin, R., Craiu, R. and Leisen, F. (2013). Interacting Multiple Try Algorithms with Different Proposal Distributions. *Stat. Comput.* **23** 185–200. MR3016937

Chen, H. F., Lei, G. and Gao, A. J. (1988). Convergence and Robustness of the Robbins-Monro Algorithm Truncated at Randomly Varying Bounds. *Stochastic Process. Appl.* **27** 217–231. MR931029

Craiu, R. V. and Lemieux, C. (2007). Acceleration of the Multiple-try Metropolis Algorithm Using Antithetic and Stratified Sampling. *Stat. Comput.* **17** 109–120. MR2380640

Craiu, R. V., Rosenthal, J. and Yang, C. (2009). Learn from Thy Neighbor:

Parallel-chain and Regional Adaptive MCMC. *J. Amer. Statist. Assoc.* **104** 1454–1466. MR2750572

Craiu, R. V., Gray, L., Ł atuszyński, K., Madras, N., Roberts, G. O. and Rosenthal, J. S. (2015). Stability of Adversarial Markov Chains, with an Application to Adaptive MCMC Algorithms. *Ann. Appl. Probab.* **25** 3592–3623. MR3404645

Desai, M. (2000). Mixture Models for Genetic Changes in Cancer Cells, PhD thesis, University of Washington.

Fontaine, S. (2019). MCMC adaptatifs à essais multiples, Master's thesis, Department of Mathematics and Statistics, University of Montreal.

Gelman, A., Roberts, G. O. and Gilks, W. R. (1996). Efficient Metropolis Jumping Rules. 599–607. MR1425429

Grenon-Godbout, N. and Bédard, M. (2020+). On-line Partitioning of the Sample Space in the Regional Adaptation Algorithm. *Can J. Statistics.*

Haario, H., Saksman, E. and Tamminen, J. (2001). An Adaptive Metropolis Algorithm. *Bernoulli* **7** 223–242. MR1828504

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57** 97–109. MR3363437

Liu, J. S., Liang, F. and Wong, W. H. (2000). The Multiple-try Method and Local Optimization in Metropolis Sampling. *J. Amer. Statist. Assoc.* **95** 121–134. MR1803145

Martino, L. and Read, J. (2013). On the Flexibility of the Design of Multiple Try Metropolis Schemes. *Comput. Statist.* **28** 2797–2823. MR3141364

Martino, L., Casarin, R., Leisen, F. and Luengo, D. (2018). Adaptive Independent Sticky MCMC Algorithms. *EURASIP Journal on Advances in Signal Processing* **2018** 5.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21** 1087–1092.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, Second ed. Cambridge University Press, Cambridge With a prologue by Peter W. Glynn. MR2509253

Nummelin, E. (1984). *General Irreducible Markov Chains and Nonnegative Operators. Cambridge Tracts in Mathematics* **83**. Cambridge University Press, Cambridge. MR776608

Petersen, K. B. and Pedersen, M. S. (2008). The Matrix Cookbook. *Technical University of Denmark* **7** 510.

Pillai, N. S., Stuart, A. M. and Thiéry, A. H. (2012). Optimal Scaling and Diffusion Limits for the Langevin Algorithm in High Dimensions. *Ann. Appl. Probab.* **22** 2320–2356. MR3024970

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Ann. Math. Statistics* **22** 400–407. MR42668

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. MR2080278

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *Ann. Appl. Probab.* **7** 110–

120. MR1428751

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal Scaling of Discrete Approximations to Langevin Diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 255–268. MR1625691

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statist. Sci.* **16** 351–367. MR1888450

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms. *J. Appl. Probab.* **44** 458–475. MR2340211

Rosenthal, J. S. and Yang, J. (2018). Ergodicity of Combocontinuous Adaptive MCMC Algorithms. *Methodol. Comput. Appl. Probab.* **20** 535–551. MR3803934

Saksman, E. and Vihola, M. (2010). On the Ergodicity of the Adaptive Metropolis Algorithm on Unbounded Domains. *Ann. Appl. Probab.* **20** 2178–2203. MR2759732

Sherlock, C., Fearnhead, P. and Roberts, G. O. (2010). The Random Walk Metropolis: Linking Theory and Practice through a Case Study. *Statist. Sci.* **25** 172–190. MR2789988

Sherlock, C. and Roberts, G. (2009). Optimal Scaling of the Random Walk Metropolis on Elliptically Symmetric Unimodal Targets. *Bernoulli* **15** 774–798. MR2555199

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Ann. Statist.* **22** 1701–1762. With discussion and a rejoinder by the author. MR1329166

Tran, M.-N., Pitt, M. K. and Kohn, R. (2016). Adaptive Metropolis-Hastings Sampling Using Reversible Dependent Mixture Proposals. *Stat. Comput.* **26** 361–381. MR3439379

VanDerwerken, D. N. and Schmidler, S. C. (2013). Parallel Markov Chain Monte Carlo. *arXiv preprint arXiv:1312.7479.*

Vihola, M. (2011). On the Stability and Ergodicity of Adaptive Scaling Metropolis Algorithms. *Stochastic Process. Appl.* **121** 2839–2860. MR2844543

Vihola, M. (2012). Robust Adaptive Metropolis Algorithm with Coerced Acceptance Rate. *Stat. Comput.* **22** 997–1008. MR2950080

Warnes, G. R. (2001). The Normal Kernel Coupler: An Adaptive Markov Chain Monte Carlo Method for Efficiently Sampling from Multi-modal Distributions.

Yang, J., Levi, E., Craiu, R. V. and Rosenthal, J. S. (2019). Adaptive Component-Wise Multiple-Try Metropolis Sampling. *J. Comput. Graph. Statist.* **28** 276–289. MR3974879

Appendix A: Additional details on aMTM variants

A.1. Sampling schemes

Table 5 contains an algorithmic description of different types of candidates used in the aMTM algorithm. The derivations for the EA candidates can be found in Fontaine (2019, Section 5.3.4.2); those for the RQMC candidates can be found in Fontaine (2019, Example 4.3). Independent and common random variable candidates have trivial formulations.

A.2. Update variations

Local updates A notable difference between (ASW)AM and RAM updates is the use of the running mean $\mu_n^{(k)}$. While the (ASW)AM update uses the difference between the new point and the current estimate for the mean $(x_{n+1} - \mu_n^{(k)})$ to update the covariance, the RAM update rather uses the proposed step $(y - x_n)$. The latter seems more appropriate to locally adjust proposal densities to the target distribution. Indeed, using a running mean will potentially produce marginal covariances that are all similar to one another; using the proposed step may prevent this uniformisation. We thus propose to modify the (ASW)AM updates in (3.2) by making *local updates* in which $x_{n+1} - \mu_n^{(k)}$ is replaced by $y - x_n$. In that case, the running mean update (3.1) is no longer required.

Up to this point, the only marginal covariance updated in a given iteration is that of the selected candidate. We now propose two adaptation schemes imposing some conditions on the other marginal covariances.

Global proposal We propose to consider the first proposal density ($k = 1$) as a *global* one. Its marginal covariance is thus adapted at each iteration—using any of the three update rules—no matter if the candidate was generated from this proposal or not. Then, we expect that marginal covariance to approach the target's global covariance, while the covariance matrices of the other densities should explore more local properties of the target density. This approach is particularly well-suited to multimodal densities as the global density provides a way to jump between modes, while other densities propose jumps within specific modes. Computationaly, adapting a second covariance at every iteration doubles the adaptation cost.

Scale adaptation In the ASWAM case, we propose to adapt the scale parameter $\lambda_n^{(k)}$ of densities that are not selected very often; indeed, these densities are rarely adapted and may therefore never recover from a bad initialization. Given a target floor selection rate $s_* \in [0, 1]$, we decrease the scale parameter whenever a proposal density's selection rate drops below s_*. Indeed, for importance weights (2.4) or weights proportional to the target (2.5), the fact of being selected too rarely is generally related to the scale being to large.

Type	**Candidates (Step (a))**	**Shadow points (Step (d))**
Independent	Sample $$Y^{(j)} \sim q_\theta^{(j)}(\cdot\|x_n)$$ independently for $j = 1, \dots, K$.	Sample $$Y^{(j)} \sim q_\theta^{(j)}(\cdot\|y)$$ independently for $j \neq k$.
EA $\rho = \frac{-1}{K-1}$	Beforehand, compute the singular value decomposition (SVD) $$\Psi_K = (X\Lambda^{1/2})(X\Lambda^{1/2})^\top ,$$ where $$\Psi_K = \rho I_K \otimes I_d + (1-\rho) I_{dK} .$$ For $k = 1, \dots, K$: – Sample $Z^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mathbf{0}_d, I_d)$, – Compute $u^{(k)} = X\Lambda^{1/2} z^{(k)}$, – Compute $y^{(k)} = x_n + S^{(k)} u^{(k)}$.	Beforehand, compute the SVD $$\Phi_{K-1} = (X'\Lambda'^{1/2})(X'\Lambda'^{1/2})^\top,$$ where $$\Phi_{K-1} = \rho I_{K-1} \otimes I_d + I_{d(K-1)} .$$ For $j \neq k$: – Sample $Z^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mathbf{0}_d, I_d)$, – Compute $u_*^{(j)} = X'\Lambda'^{1/2} z^{(j)}$, – Compute $$x_*^{(j)} = y + S^{(j)} \left(u_*^{(j)} - \rho u^{(k)} \right) .$$
RQMC Koborov rule with $1 \leqslant a < K$	Sample $$U \sim \text{Uniform}[0,1)^d .$$ For $k = 1, \dots, K$: – Compute $$u^{(k)} \equiv_1 \frac{k-1}{K} \left(1, a, \dots, a^{d-1}\right) + u ,$$ – Compute $z^{(k)} = F^{-1}(u^{(k)})$, – Compute $y^{(k)} = x_n + S^{(k)} z^{(k)}$.	Compute $$u_* = F\left((S^{(k)})^{-1}(y - x_n) \right) .$$ For $j \neq k$: – Compute $$u_*^{(j)} \equiv_1 \frac{j-1}{K} \left(1, a, \dots, a^{d-1}\right) + u_* ,$$ – Compute $z_*^{(j)} = F^{-1}(u_*^{(j)})$, – Compute $x_*^{(j)} = y + S^{(j)} z_*^{(j)}$.
Common RV	Sample $$Z \sim \mathcal{N}_d(\mathbf{0}_d, I_d) .$$ For $k = 1, \dots, K$, compute $$y^{(k)} = x_n + S^{(k)} z .$$	Compute $$z_* = (S^{(k)})^{-1}(y - x_n) .$$ For $j \neq k$, compute $$x_*^{(j)} = y + S^{(j)} z_* .$$

Table 5. Summary of the different types of candidates used in the aMTM algorithm at the MTM sampling step; the rest of the MTM sampling remains unchanged. The sampling of candidates and shadow points is described algorithmically. Notation: x_n is the current state of the chain, F is the standard normal CDF, $y = y^{(k)}$ is the selected proposal, $\Sigma^{(k)} = S^{(k)} S^{(k)\top}$ is the square root decomposition and $\otimes$ denotes the usual Kronecker product.

Appendix B: Proofs

B.1. Results on MTM transitions

Consider a MTM transition P with joint proposal density $q(\cdot|x)$. For $k = 1, \ldots, K$, the conditional density of $y^{(-k)}$ given $y^{(k)}$ is $q^{(-k)}(\cdot|y^{(k)}, x)$ and the marginal density of $y^{(k)}$ is $q^{(k)}(\cdot|x)$. The generalized MTM acceptance probability is given by

$$\alpha_{\mathrm{MTM}}\left(y, y^{(-k)}|x, x_*^{(-k)}\right) = \min\left\{1, \frac{\pi(y)q^{(k)}(x|y)\bar{w}^{(k)}(x, x_*^{(-k)}|y)}{\pi(x)q^{(k)}(y|x)\bar{w}^{(k)}(y, y^{(-k)}|x)}\right\} ,$$

where $x_*^{(-k)} \sim q^{(-k)}(\cdot|x, y)$ are the shadow points, and where

$$\bar{w}^{(k)}(y, y^{(-k)}|x) = \frac{w^{(k)}(y|x)}{\sum_{j=1}^{K} w^{(j)}(y^{(j)}|x)}$$

is the probability of choosing the k-th candidate.

We define some notation in order to simplify the MTM transition. The transition admits the following (pseudo-)density:

$$p(y|x) = a(y|x) + R(x)\delta_x(y) ,$$

where $a(y|x)$ is the density for transitioning from x to y using any of the K candidates and any shadow sample, and where $R(x) = 1 - \int_{\mathcal{X}} a(y|x)\,\mathrm{d}y$ is the integrated rejection probability. Since moving from x to y can be achieved through any of the K candidates, we therefore decompose

$$a(y|x) = \sum_{k=1}^{K} A^{(k)}(y|x)q^{(k)}(y|x) ,$$

where $A^{(k)}(y|x)$ is the density for accepting a move from x to y through the k-th candidate. We have

$$\begin{aligned} A^{(k)}(y|x) = \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} & q^{(-k)}(y^{(-k)}|y, x)\bar{w}(y; y^{(-k)}|x)\alpha_{\mathrm{MTM}}(y, y^{(-k)}|x, x_*^{(-k)}) \\ & \times q^{(-k)}(x_*^{(-k)}|y, x)\,\mathrm{d}y^{(-k)}\,\mathrm{d}x_*^{(-k)} . \end{aligned}$$

Finally, we define the integrated probability of accepting the k-th candidate as the new point,

$$\overline{A}^{(k)}(x) = \int_{\mathcal{X}} A^{(k)}(y|x)q^{(k)}(y|x)\,\mathrm{d}y ,$$

so that we may write

$$R(x) = 1 - \sum_{k=1}^{K} \overline{A}^{(k)}(x) .$$

Naturally, when we want to make the dependence of a MTM transition on its set of parameters θ explicit, we simply index position each of the above definitions by θ.

Proposition B.1. *Suppose that the marginal proposal densities satisfy*

$$q^{(k)}(y|x) > 0 \quad \Leftrightarrow \quad q^{(k)}(x|y) > 0 \ , \qquad k = 1\,, \ldots, K \ .$$

Then, the MTM transition satisfies the detailed balance condition,

$$p(x|y)\pi(y) = p(y|x)\pi(x) \ , \qquad \forall\, x, y \in \mathcal{X} \ , \tag{B.1}$$

for any weight function $w^{(k)}(y|x)$ that is positive whenever $x, y \in \mathcal{X}$.

Proof. If $x = y$, then (B.1) trivially holds. Thus, we may assume that $y \neq x$, in which case $\delta_X(y) = 0$ and

$$\begin{aligned} p(y|x) &= a(y|x) \\ &= \sum_{k=1}^{K} \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} q(y, y^{(-k)}|x)\bar{w}(y; y^{(-k)}|x)\alpha_{\mathrm{MTM}}(y, y^{(-k)}|x, x_*^{(-k)}) \\ &\qquad\qquad \times q^{(-k)}(x_*^{(-k)}|y, x)\, \mathrm{d}y^{(-k)}\, \mathrm{d}x_*^{(-k)} \ . \end{aligned}$$

Then, we decompose the joint proposal density as

$$q(y, y^{(-k)}|x) = q^{(-k)}(y^{(-k)}|y, x)q^{(k)}(y|x) \ .$$

We also rewrite the MTM acceptance probability in a more symmetric form,

$$\begin{aligned} \alpha_{\mathrm{MTM}}(y, y^{(-k)}|x, x_*^{(-k)}) &= \pi(y)q^{(k)}(x|y)\bar{w}^{(k)}(x; x_*^{(-k)}|y) \\ &\quad \times \min\left\{ \frac{1}{\pi(y)q^{(k)}(x|y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)}, \frac{1}{\pi(x)q^{(k)}(y|x)\bar{w}^{(k)}(y; y^{(-k)}|x)} \right\} \ . \end{aligned}$$

We can now write

$$\begin{aligned} a(y|x)\pi(x) = \sum_{k=1}^{K} \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} & \pi(x)q^{(-k)}(y^{(-k)}|y, x)q^{(k)}(y|x)\bar{w}(y; y^{(-k)}|x) \\ & \times \pi(y)q^{(k)}(x|y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)q^{(-k)}(x_*^{(-k)}|y, x) \\ & \times \min\left\{ \frac{1}{\pi(y)q^{(k)}(x|y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)}, \frac{1}{\pi(x)q^{(k)}(y|x)\bar{w}^{(k)}(y; y^{(-k)}|x)} \right\} \\ & \mathrm{d}y^{(-k)}\, \mathrm{d}x_*^{(-k)} \ . \end{aligned}$$

By direct inspection, we see that the expression is completely symmetric under the swap $(y, y^{(-k)}) \leftrightarrow (x, x_*^{(-k)})$. Hence,

$$a(y|x)\pi(x) = a(x|y)\pi(y) \ ,$$

and the detailed balance condition (B.1) is satisfied. □

***Remark* B.1.** Note that a similar result was obtained by Casarin, Craiu and Leisen (2013, in the appendix) for a slightly less general form of MTM acceptance probability.

Proposition B.2. *Let π be a target density with connected support $\mathcal{X}$. Suppose that π and the weight function $w^{(k)}(\cdot|x)$ are bounded above on $\mathcal{X}$ and below on any compact subset of $\mathcal{X}$, for any fixed $x \in \mathcal{X}$. Further suppose that there exists δ, $\varepsilon > 0$ such that the marginal proposal densities are locally positive, that is,*

$$\|x-y\|_2 < \delta \quad \Rightarrow \quad q^{(k)}(y|x) > \varepsilon \;, \qquad k = 1\,, \dots, K \;.$$

Then, the MTM transition is π-irreducible and aperiodic.

Proof. The proof of π-irreducilibity appeals to Meyn and Tweedie (2009, Proposition 4.2.1) which states that a transition P is ϕ-irreducible if and only if, for all $x \in \mathcal{X}$ and for all measurable B such that $\phi(B) > 0$, there exists $m \in \mathbb{N}$ with $P^m(B|x) > 0$. Thus, let us consider $x \in \mathcal{X}$ as well as a measurable set $B \subseteq \mathcal{X}$ with positive probability $\pi(B) > 0$. By connectedness of $\mathcal{X}$, we can find a path between x and any point in B. In particular, we can always find a path of length $m \in \mathbb{N}$ from x to some $x_m \in B$ such that each step is at most of size δ, i.e. $\|x_i - x_{i-1}\|_2 < \delta$ $(i = 1, \dots, m)$ and each x_i has positive density $\pi(x_i) > 0$. Around each x_i, we consider the ball of radius δ, denoted

$$B_\delta(x_i) = \{x \in \mathbb{R}^d \mid \|x_i - x\|_2 \leqslant \delta\} \;.$$

Since x_i is in the support of π, then $\pi\left(B_\delta(x_i)\right) > 0$ by the definition of a support. Now, we show that the transition from one ball to the next happens with positive probability. Consider $i \in \{0\,, \dots, m-1\}$ and $x \in B_\delta(x_i)$. Then, the probability of landing in the next ball is bounded below by the probability of landing in the next ball through an accepted proposal, i.e.

$$P\left(B_\delta(x_{i+1})|x\right) \geqslant \sum_{k=1}^{K} \int_{B_\delta(x_{i+1})} A^{(k)}(y|x) q^{(k)}(y|x)\,\mathrm{d}y \;.$$

Now, $A^{(k)}$ is positive for any $y \in B_\delta(x_{i+1})$ since it is the expectation of a positive function ($\bar{w} > 0$ and $\alpha_{\mathrm{MTM}} > 0$ both follow from the assumptions). Then, since the marginal density $q^{(k)}$ is also positive on $B_\delta(x_{i+1})$ and since $B_\delta(x_{i+1})$ has positive probability, we find $P\left(B_\delta(x_{i+1})|x\right) > 0$. By induction, we can show that the i-step transition $P^i(B_\delta(x_i)|x)$ is positive for $i = 1\,, \dots, m$. In particular, it holds for m so that $P^m(B_\delta(x_m)|x) > 0$ from which we find $P^m(B|x) > 0$ because $x_m \in B \cap \mathcal{X}$. By Meyn and Tweedie (2009, Proposition 4.2.1), P is π-irreducible.

To prove aperiodicity, we show that P is strongly aperiodic, meaning that there exists a $(\nu, 1)$-small measurable set B with $\nu(B) > 0$. A $(\nu, 1)$-small set is such that, for all $x \in B$ and for all measurable sets C,

$$P(C|x) \geqslant \nu(C) \;. \tag{B.2}$$

We now consider $B = B_{\delta/2}(x)$ and construct a measure ν concentrated on B satisfying the minorization condition (B.2). Thus, let us consider $x \in B$ and C measurable. Then, we can bound $P(C|x) \geqslant P(C \cap B|x)$; we can also bound the latter using accepted proposals, leading to

$$P(C|x) \geqslant \sum_{k=1}^{K} \int_{C\cap B} A^{(k)}(y|x) q^{(k)}(y|x)\,\mathrm{d}y \ .$$

We now define K partitions of the support, one for each candidate. Given the current state x, the candidates $y^{(-k)}$, and the shadow sample $x_*^{(-k)}$, this partition groups together all the proposals y that are automatically accepted, given that the k-th candidate was selected:

$$D^{(k)}(x) = \left\{ y \in \mathcal{X} \;\middle|\; \frac{\pi(y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)}{\pi(x)\bar{w}^{(k)}(y; y^{(-k)}|x)} \leqslant 1 \right\} \ .$$

Note that, contrarily to what is suggested by the notation, $D^{(k)}(x)$ also is a function of $y^{(-k)}$ and $x_*^{(-k)}$. For $y \in D^{(k)}(x)$, we have

$$\alpha_{\mathrm{MTM}}(y, y^{(-k)}|x, x_*^{(-k)}) = \frac{\pi(y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)}{\pi(x)\bar{w}^{(k)}(y; y^{(-k)}|x)} \ ,$$

while for $y \notin D^{(k)}$, we have $\alpha_{\mathrm{MTM}}(y, y^{(-k)}|x, x_*^{(-k)}) = 1$. We can now split the integral over $B \cap C$ into two parts over which the form of α_{MTM} is known.

For $y \in D^{(k)}$, the integrand takes the form

$$\begin{aligned}
&\bar{w}^{(k)}(y; y^{(-k)}|x)\frac{\pi(y)\bar{w}^{(k)}(x; x_*^{(-k)}|y)}{\pi(x)\bar{w}^{(k)}(y; y^{(-k)}|x)} q^{(-k)}(y^{(-k)}|y, x) q^{(-k)}(x_*^{(-k)}|x, y) q^{(k)}(y|x)\\
&= \bar{w}^{(k)}(x; x_*^{(-k)}|y)\frac{\pi(y)}{\pi(x)} q^{(-k)}(y^{(-k)}|y, x) q^{(-k)}(x_*^{(-k)}|x, y) q^{(k)}(y|x) \ .
\end{aligned}$$

Since we will integrate over $y^{(-k)}$ and $x_*^{(-k)}$, we try to bound all terms that are not the densities of these variables. In particular, we search a lower bound for

$$\bar{w}^{(k)}(x; x_*^{(-k)}|y)\frac{\pi(y)}{\pi(x)} q^{(k)}(y|x) \ .$$

When $x \in B$ and $y \in C \cap B \subseteq B$, we can bound each term by making use of the assumptions. Indeed, we have $\|y - x\|_2 \leqslant \delta$ so that $q^{(k)}(y|x) \geqslant \varepsilon$. Furthermore, from the conditions on the weight functions, there exists $0 < a < A < \infty$ such that $w^{(k)}(x|y) > a$ and $w^{(j)}(x_*^{(j)}|y) \leqslant A$ so that $\bar{w}^{(k)}(x; x_*^{(-k)}|y) \geqslant a/KA$ for all $\|y - x\|_2 \leqslant \delta$ and all $x_*^{(-k)}$. Hence, we find

$$\bar{w}^{(k)}(x; x_*^{(-k)}|y)\frac{\pi(y)}{\pi(x)} q^{(k)}(y|x) \geqslant \frac{a\varepsilon}{KA}\frac{\pi(y)}{\pi(x)} \geqslant \frac{a\varepsilon}{KA}\frac{\inf_{y\in B}\pi(y)}{\sup_{y\in B}\pi(y)} \ ,$$

which is positive because all quantities are positive (π is bounded below and above on $B = B_{\delta/2}(x)$ compact).

On $y \notin D^{(k)}$, the integrand takes the form

$$\bar{w}^{(k)}(y; y^{(-k)}|x) q^{(-k)}(y^{(-k)}|y, x) q^{(-k)}(x_*^{(-k)}|x, y) q^{(k)}(y|x) \ ,$$

which means we aim to bound

$$\bar{w}^{(k)}(y; y^{(-k)}|x) q^{(k)}(y|x) \ .$$

For the same reasons as before, we have $\bar{w}^{(k)}(x; x_*^{(-k)}|y) \geqslant a/KA$ and $q^{(k)}(y|x) \geqslant \varepsilon$. Then, we note that $\inf_B \pi / \sup_B \pi$ is always less than 1 so we find the same bound as in the case $y \in D^{(k)}$, i.e.

$$\bar{w}^{(k)}(y; y^{(-k)}|x) q^{(k)}(y|x) \geqslant \frac{a\varepsilon}{KA} \geqslant \frac{a\varepsilon}{KA} \frac{\inf_{y \in B} \pi(y)}{\sup_{y \in B} \pi(y)} \ .$$

We therefore find the following bound on $P(C|x)$:

$$\begin{aligned}
P(C|x) &\geqslant \sum_{k=1}^{K} \int_{C \cap B} A^{(k)}(y|x) q^{(k)}(y|x) \, \mathrm{d}y \\
&= \sum_{k=1}^{K} \left(\int_{C \cap B \cap D^{(k)}} + \int_{C \cap B \cap (D^{(k)})^c} \right) A^{(k)}(y|x) q^{(k)}(y|x) \, \mathrm{d}y \\
&\geqslant \sum_{k=1}^{K} \left(\int_{C \cap B \cap D^{(k)}} + \int_{C \cap B \cap (D^{(k)})^c} \right) \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} \\
&\qquad q^{(-k)}(y^{(-k)}|y, x) q^{(-k)}(x_*^{(-k)}|x, y) \frac{a\varepsilon}{KA} \frac{\inf_B \pi}{\sup_B \pi} \, \mathrm{d}x_*^{(-k)} \, \mathrm{d}y^{(-k)} \, \mathrm{d}y \\
&= \sum_{k=1}^{K} \int_{C \cap B} \frac{a\varepsilon}{KA} \frac{\inf_B \pi}{\sup_B \pi} \int_{\mathcal{X}^{K-1}} \int_{\mathcal{X}^{K-1}} q^{(-k)}(y^{(-k)}|y, x) q^{(-k)}(x_*^{(-k)}|x, y) \, \mathrm{d}x_*^{(-k)} \, \mathrm{d}y^{(-k)} \, \mathrm{d}y \\
&= \sum_{k=1}^{K} \int_{C \cap B} \frac{a\varepsilon}{KA} \frac{\inf_B \pi}{\sup_B \pi} \, \mathrm{d}y \\
&= \frac{a\varepsilon}{A} \frac{\inf_B \pi}{\sup_B \pi} \lambda^{\mathrm{Leb}} (C \cap B) \ ,
\end{aligned}$$

where λ^{Leb} is the Lebesgue measure on $\mathbb{R}^d$. Since

$$\frac{a\varepsilon}{A} \frac{\inf_B \pi}{\sup_B \pi} =: c_0 > 0 \ ,$$

we have that

$$P(C|x) \geqslant \nu(C) \ ,$$

where $\nu(C) = c_0 \lambda^{\mathrm{Leb}} (C \cap B)$ is a non-trivial measure concentrated on B, as required. Finally, we note that $\nu(B) = c_0 \lambda^{\mathrm{Leb}}(B) > 0$ since $c_0 > 0$ and $\lambda^{\mathrm{Leb}}(B) > 0$, where B is a ball with positive radius $\delta/2 > 0$. $\square$

In the context of Markov transitions, a function $h : \mathbb{R}^d \to [0, \infty]$ is said to be *harmonic* for a transition P if $h = Ph$ everywhere, that is,

$$h(x) = \int_{\mathcal{X}} h(y) P(\,\mathrm{d}y|x) \;, \qquad x \in \mathbb{R}^d \;.$$

From Tierney (1994, Theorem 2), we know that a recurrent Markov transition P is Harris-recurrent if and only if every bounded harmonic function is a constant function. We use this result to show that recurrence and Harris-recurrence happen simultaneously for MTM transitions.

Proposition B.3. *Let P be a MTM transition for a given target density π. If P is π-irreducible, then P is Harris-recurrent.*

Proof. By Proposition B.1, the MTM transition satisfies the detailed balance condition. By Robert and Casella (2004, Theorem 6.46), the MTM transition admits π as its invariant distribution. By Tierney (1994, Theorem 1), the MTM transition is positive recurrent. From Nummelin (1984, Proposition 3.13), we know that a recurrent π-irreducible Markov transition P is such that every bounded harmonic function h is constant at least π-almost everywhere. Hence, we only require to extend that result to every $x \in \mathbb{R}^d$.

We define the set H as containing the points over which a function h is not constant, i.e.,

$$H = \{x \in \mathcal{X} \mid h(x) \neq \pi h\} \,.$$

By the above argument, we find $\pi(H) = 0$. Then, since the measure of H is null, the probability of transitioning from $x \in \mathcal{X}$ to H must also be 0:

$$a\,(H|x) = \int_H \sum_{k=1}^{K} A^{(k)}(y|x) q^{(k)}(y|x)\,\mathrm{d}y = 0 \;,$$

since $q^{(k)}$ is assumed to be a density and H has zero measure.

Now, since h is harmonic with respect to P, we can decompose

$$h(x) = \int_{\mathbb{R}^d} h(y) P(\,\mathrm{d}y|x) = \int_H h(y) P(\,\mathrm{d}y|x) + \int_{H^c} h(y) P(\,\mathrm{d}y|x) \;.$$

The former term satisfies

$$\begin{aligned}\int_H h(y) P(\,\mathrm{d}y|x) &= \int_H h(y) \sum_{k=1}^{K} \left[A^{(k)}(y|x) q^{(k)}(\,\mathrm{d}y|x)\right] + \int_H h(y) R(x) \delta_x(\,\mathrm{d}y) \\ &= 0 + h(x) R(x)\, \mathbb{I}\left(h(x) \neq \pi h\right) \;.\end{aligned}$$

For the latter term, we obtain

$$\begin{aligned}\int_{H^c} h(y) P(\,\mathrm{d}y|x) &= \int_{H^c} \pi h \sum_{k=1}^{K} \left[A^{(k)}(y|x) q^{(k)}(\,\mathrm{d}y|x)\right] + \int_{H^c} \pi h R(x) \delta_x(\,\mathrm{d}y) \\ &= \pi h (1 - R(x)) + \pi h R(x)\, \mathbb{I}\left(h(x) = \pi h\right) \;.\end{aligned}$$

Combining both expressions, we find that h must satisfy

$$h(x) = \pi h + R(x)(h(x) - \pi h)\,\mathbb{I}(h(x) \neq \pi h)\ .$$

Factorizing yield

$$0 = (h(x) - \pi h)(R(x)\,\mathbb{I}(h(x) \neq \pi h) - 1)\ . \tag{B.3}$$

Thus, if $h(x) \neq \pi h$, we must have $R(x) = 1$, but this would contradict the π-irreducibility of the transition whenever $x \in \mathcal{X}$ because this means that the probability of staying at x is 1. Hence, the only points where we can have $h(x) \neq \pi h$ are $x \notin \mathcal{X}$. Then, by construction of MTM transitions, $0 < R(x) < 1$ so that $h(x) = \pi h$ must hold to satisfy (B.3). This shows $h \equiv \pi h$. □

B.2. Results on adaptive MCMC

B.2.1. Additional background on adaptive algorithms

We say that a family of MCMC transitions $\{P_\theta\}_{\theta\in\Theta}$ satisfies the *uniform geometric ergodicity on compact sets* condition (Andrieu and Moulines, 2006, Assumption A1) if there exists a test function $V : \mathcal{X} \to [1, \infty)$ with $\sup_{\mathcal{X}} V < \infty$ such that, for any compact $\mathcal{K} \subseteq \Theta$, the following two conditions hold :

(i) *Minorisation.* There exists $C \in \mathcal{B}(\mathcal{X})$, $\delta > 0$ and a probability measure ν with $\nu(C) > 0$ such that

$$P_\theta(A|x) \geqslant \delta\nu(A)\ , \qquad \forall\ A \in \mathcal{B}(\mathcal{X}),\ \theta \in \mathcal{K},\ x \in C\ .$$

(ii) *Geometric drift.* There exists $\lambda \in [0, 1)$ and $b \in (0, \infty)$ such that

$$P_\theta V(x) \leqslant \begin{cases} \lambda V(x), & x \notin C, \\ b, & x \in C, \end{cases} \qquad \forall\ \theta \in \mathcal{K}\ ,$$

where $P_\theta V(x) = \int V(z)P_\theta(z|x)\,\mathrm{d}z$.

We say that a family of update functions $\{H_\theta\}_{\theta\in\Theta}$ is *V-Lipschitz* for some test function V (typically the same as in the geometric drift condition) if, for any compact $\mathcal{K} \subseteq \Theta$, we have

$$\sup_{\theta\in\mathcal{K}} \|H_\theta\|_V < \infty \qquad \text{and} \qquad \sup_{\theta\neq\theta'\in\mathcal{K}\times\mathcal{K}} \|\theta - \theta'\|_2^{-1} \|H_\theta - H_{\theta'}\|_V < \infty,$$

where $\|\mu\|_V$ defines the V-norm of the function f for some test function V, that is,

$$\|f\|_V = \sup_{x\in\mathcal{X}} \frac{\|f(x)\|_2}{V(x)}.$$

We say that a family of transitions is *V-Lipschitz on $\mathcal{K}$* if there exists $C < \infty$ such that, for all functions $f : \mathcal{X} \to \mathbb{R}$, with $\|f\|_V < \infty$, and all $r \in [0, 1]$,

$$\|P_\theta f - P_{\theta'} f\|_{V^r} \leqslant C\|f\|_{V^r}\|\theta - \theta'\|_2, \qquad \forall\ \theta, \theta' \in \mathcal{K}\,.$$

B.2.2. Diminishing adaptation

Define the V-norm of a (possibly signed) measure μ as

$$|||\mu|||_V = \sup_{g:|g|\leqslant V} |\mu(g)|,$$

where we use the triple bar notation to differentiate with the V-norm of a function defined earlier. Note that $|||\cdot|||_1$ is equivalent to $\|\cdot\|_{\mathrm{TV}}$:

$$\|\mu\|_{\mathrm{TV}} = \sup_{B\in\mathcal{B}(\mathcal{X})} |\mu(B)| = \frac{1}{2} \sup_{g:|g|\leqslant 1} |\mu(g)| = \frac{1}{2}|||\mu|||_1.$$

Proposition B.4. *Suppose $\{P_\theta\}_{\theta\in\Theta}$ satisfies the uniform geometric ergodicity on compact sets, $\{H_\theta\}_{\theta\in\Theta}$ is V-Lipschitz and $\{P_\theta\}_{\theta\in\Theta}$ is V-Lipschitz on any compact subset of Θ for the same test function V. If $\sup_{\theta\in\mathcal{K}} \|H_\theta\|_V < \infty$ for any $\mathcal{K} \subseteq \Theta$ compact and $\{\theta_n\}_{n\geqslant 0}$ is bounded in probability, then the adaptive MCMC algorithm is such that*

$$\left|\left|\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\right|\right|_V \xrightarrow{\mathbb{P}} 0, \qquad n\to\infty.$$

In particular, if $V \equiv 1$, then the algorithm satisfies the Diminishing Adaptation condition.

Proof. From the condition on H_θ, for any compact $\mathcal{K}\subseteq\Theta$, we have $\widetilde{H}(\mathcal{K}) := \sup_{\theta\in\mathcal{K}} \|H_\theta\|_V < \infty$. In particular, we have

$$\|H(\theta,x)\|_2 \leqslant \widetilde{H}(\mathcal{K})V(x) , \qquad \forall\, x\in\mathcal{X},\, \theta\in\mathcal{K} \ .$$

The uniform geometric ergodicity on compact sets ensures that $\{V(X_n)\}_{n\geqslant 0}$ is bounded in probability (Fontaine, 2019, Proposition 3.5). Hence, for all $\varepsilon > 0$, there exists $\widetilde{V} = \widetilde{V}(\varepsilon) < \infty$ such that $\mathbb{P}_{x_0,\theta_0}\left(V(X_n) \leqslant \widetilde{V}\right) \geqslant 1 - \frac{\varepsilon}{4}$ for all $n \geqslant 1$, where (x_0,θ_0) are the initial values of the joint chain $\{(X_n,\theta_n)\}_{n\geqslant 0}$. Then,

$$\mathbb{P}_{x_0,\theta_0}\left(\|H(\theta,X_n)\|_2 \leqslant \widetilde{H}(\mathcal{K})\widetilde{V} \mid \theta\in\mathcal{K}\right) \geqslant 1 - \frac{\varepsilon}{4} \ .$$

Then, for $\theta_{n+1} - \theta_n = \gamma_{n+1}H(\theta_n, X_n)$, we find

$$\mathbb{P}_{x_0,\theta_0}\left(\|\theta_{n+1}-\theta_n\|_2 \leqslant \gamma_{n+1}\widetilde{H}(\mathcal{K})\widetilde{V} \mid \theta_n\in\mathcal{K}\right) \geqslant 1 - \frac{\varepsilon}{4} \ , \qquad \forall\, \theta_n\in\mathcal{K} \ .$$

Since $\{\theta_n\}_{n\geqslant 0}$ is bounded in probability, there exists a compact $\mathcal{K}\subset\Theta$ such that

$$\mathbb{P}_{x_0,\theta_0}\left(\theta_n\in\mathcal{K}\right) \geqslant 1 - \frac{\varepsilon}{4} \ ,$$

from which we find

$$\mathbb{P}_{x_0,\theta_0}\left(\|\theta_{n+1}-\theta_n\|_2 \leqslant \gamma_{n+1}\widetilde{H}(\mathcal{K})\widetilde{V}\right) \geqslant \left(1-\frac{\varepsilon}{4}\right)^2 \ .$$

The Lipschitz transition condition implies that there exists $C < \infty$ with

$$\left\|P_{\theta_{n+1}} f - P_{\theta_n} f\right\|_V \leqslant C\|f\|_V \|\theta_{n+1} - \theta_n\|_2, \qquad \forall\, x \in \mathcal{X},\, \theta_{n+1},\, \theta_n \in \mathcal{K}\,.$$

Thus, we can bound

$$\begin{aligned}
\left|\!\left|\!\left|P_{\theta_{n+1}}(\cdot \mid x) - P_{\theta_n}(\cdot \mid x)\right|\!\right|\!\right|_V &= \sup_{f:|f|\leqslant V} \left|(P_{\theta_{n+1}} f - P_{\theta_n} f)(x)\right| \\
&\leqslant \sup_{f:|f|\leqslant V} \sup_{y\in\mathcal{X}} \left|(P_{\theta_{n+1}} f - P_{\theta_n} f)(y)\right| \\
&= \sup_{g:|g|\leqslant 1} \sup_{y\in\mathcal{X}} \frac{1}{V(y)} \left|(P_{\theta_{n+1}} g - P_{\theta_n} g)(y)\right| \\
&= \sup_{g:|g|\leqslant 1} \left\|P_{\theta_{n+1}} g - P_{\theta_n} g\right\|_V \\
&\leqslant \sup_{g:|g|\leqslant 1} C\|g\|_V \|\theta_{n+1} - \theta_n\|_2 \\
&= C\|\theta_{n+1} - \theta_n\|_2 .
\end{aligned}$$

We then find

$$\mathbb{P}_{x_0,\theta_0}\left(\left|\!\left|\!\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\!\right|\!\right|_V \leqslant \gamma_{n+1} C \widetilde{H}(\mathcal{K}) \widetilde{V} \mid \theta_{n+1} \in \mathcal{K}\right) \geqslant \left(1 - \frac{\varepsilon}{4}\right)^2 \,.$$

Using the boundedness in probability of $\theta_{n+1} \in \mathcal{K}$:

$$\mathbb{P}_{x_0,\theta_0}\left(\left|\!\left|\!\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\!\right|\!\right|_V \leqslant \gamma_{n+1} C \widetilde{H}(\mathcal{K}) \widetilde{V}\right) \geqslant \left(1 - \frac{\varepsilon}{4}\right)^3 \,.$$

Finally, $\gamma_n \to 0$ implies that, for any $\varepsilon' > 0$, there exists $M = M(\varepsilon') \in \mathbb{N}$ such that $\gamma_{n+1} C \widetilde{H}(\mathcal{K}) \widetilde{V} \leqslant \varepsilon'$ whenever $n \geqslant M$. Hence, for all $n \geqslant M$, we have

$$\mathbb{P}_{x_0,\theta_0}\left(\left|\!\left|\!\left|P_{\theta_{n+1}} - P_{\theta_n}\right|\!\right|\!\right|_V \leqslant \varepsilon'\right) \geqslant \left(1 - \frac{\varepsilon}{4}\right)^3 \geqslant 1 - \varepsilon \,.$$

□

B.3. Results on the aMTM algorithm

B.3.1. Diminishing adaptation

Theorem B.1. *Let π be a target density with compact support $\mathcal{X} \subseteq \mathbb{R}^d$. Consider a family of MTM transitions $\{P_\theta\}_{\theta\in\Theta}$ with compact parameter space Θ and satisfying the V-Lipschitz condition on Θ. An adaptive MTM algorithm on $\{P_\theta\}_{\theta\in\Theta}$ using the V-Lipschitz update family H_θ satisfying*

$$\sup_{\theta\in\mathcal{K}} \|H_\theta\|_V < \infty\ ,$$

with $V \equiv 1$ satisfies the diminishing adaptation condition.

Proof. We verify the conditions of Proposition B.4. When Θ and $\mathcal{X}$ are assumed to be compact, we can simply choose $V \equiv 1$ as the test function in the uniform geometric ergodicity condition. Since Θ is assumed to be compact, we directly have $\{\theta_n\}_{n\geqslant 0}$ bounded and therefore bounded in probability. The conditions on the family of updates are verified by hypothesis. □

B.3.2. Lipschitz transitions

Proposition B.5. *Let $\{\varphi_\Sigma\}_{\Sigma\in\mathcal{S}}$ be a collection of d-dimensional Gaussian densities with mean $\mathbf{0}_d$ and covariance $\Sigma \in \mathcal{S} \subseteq \mathcal{C}_d^+$. If $\mathcal{S}$ is compact, then*

$$\int_{\mathbb{R}^d} |\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)|\lambda(\,\mathrm{d}z) \leqslant \frac{d}{\lambda_{\min}} \|\Sigma - \Sigma'\|_F \ ,$$

where $\lambda_{\min} > 0$ is the smallest possible eigenvalue of a covariance $\Sigma \in \mathcal{S}$ and where $\|\cdot\|_F$ denotes the usual Frobenius norm.

Proof. Since $\mathcal{S}$ is compact, we can find $0 < \lambda_{\min} \leqslant \lambda_{\max} < \infty$ such that all eigenvalues of any $\Sigma \in \mathcal{S}$ are contained in $[\lambda_{\min}, \lambda_{\max}]$. Inspired by a step in the proof of Haario, Saksman and Tamminen (2001, Theorem 1), we consider the convex combination of $\Sigma, \Sigma' \in \mathcal{S}$, i.e.

$$\Sigma_t = (1-t)\Sigma + t\Sigma' = \Sigma + t(\Sigma' - \Sigma) \ .$$

While we do not require $\mathcal{S}$ to be convex, we know that $\mathcal{C}_+^d$ is indeed convex so that $\Sigma_t \in \mathcal{C}_+^d$ for any $t \in [0,1]$. In particular, φ_{Σ_t} is a well-defined d-dimensional Gaussian distribution for any $t \in [0,1]$. The purpose of this convex combination is the following identity, resulting from the fundamental theorem of calculus:

$$\int_0^1 \left(\frac{\partial}{\partial t}\varphi_{\Sigma_t}(z)\right) \mathrm{d}t = \varphi_{\Sigma_t}(z)\Big|_{t=0}^{t=1} = \varphi_\Sigma(z) - \varphi_{\Sigma'}(z) \ .$$

This identity holds as long as $\varphi_{\Sigma_t}(z)$ is differentiable w.r.t. t, but this will be verified implicitly in the following calculations. We then proceed to relate the previous identity to $\|\Sigma - \Sigma'\|_F$.

Logarithmic differentiation gives us

$$\begin{aligned}\frac{\partial}{\partial t}\varphi_{\Sigma_t}(z) &= \varphi_{\Sigma_t}(z)\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\\ &= -\frac{1}{2}\varphi_{\Sigma_t}(z)\frac{\partial}{\partial t}\left[d\log(2\pi) + \log\det(\Sigma_t) + z^\top\Sigma_t^{-1}z\ .\right]\end{aligned}$$

Then, using matrix derivative identities (Petersen and Pedersen, 2008), we find

$$\frac{\partial}{\partial t}z^\top\Sigma_t^{-1}z = \mathrm{tr}\left(-\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right) \ ,$$

which yields

$$\frac{\partial}{\partial t}\varphi_{\Sigma_t}(z) = -\frac{1}{2}\varphi_{\Sigma_t}(z)\,\mathrm{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma) - \Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right) \ .$$

Therefore, we find

$$\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z) = -\frac{1}{2}\,\mathrm{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma) - \Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right) \ ,$$

which can be bounded, using the triangle inequality, by

$$\left|\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\right| \leqslant \left|\mathrm{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| + \left|\mathrm{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| \ .$$

Now, we may use the general matrix norm inequality $|\mathrm{tr}(AB)| \leqslant \|A\|_F\|B\|_F$ to bound

$$\left|\mathrm{tr}\left(\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| \leqslant \left\|\Sigma_t^{-1}\right\|_F\|\Sigma' - \Sigma\|_F \ ,$$

as well as

$$\begin{aligned}\left|\mathrm{tr}\left(\Sigma_t^{-1}zz^\top\Sigma_t^{-1}(\Sigma' - \Sigma)\right)\right| &\leqslant \left\|\Sigma_t^{-1}zz^\top\Sigma_t^{-1}\right\|_F\|\Sigma' - \Sigma\|_F \\ &\leqslant z^\top\Sigma_t^{-2}z\|\Sigma' - \Sigma\|_F \ .\end{aligned}$$

Hence,

$$\left|\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\right| \leqslant \left(\left\|\Sigma_t^{-1}\right\|_F + z^\top\Sigma_t^{-2}z\right)\|\Sigma' - \Sigma\|_F \ .$$

Now, from the theory of Gaussian quadratic forms, we have

$$\int \left(z^\top\Sigma_t^{-2}z\right)\varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}z) = \mathrm{tr}\left(\Sigma_t^{-2}\Sigma_t\right) = \mathrm{tr}\left(\Sigma_t^{-1}\right) \ ,$$

which allows us to compute

$$\int \left(\left\|\Sigma_t^{-1}\right\|_F + z^\top\Sigma_t^{-2}z\right)\varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}z) = \left\|\Sigma_t^{-1}\right\|_F + \mathrm{tr}\left(\Sigma_t^{-1}\right) \ .$$

Finally, the bounded eigenvalues yield the following bounds,

$$\begin{aligned}\left\|\Sigma_t^{-1}\right\|_F^2 &= \sum_{i=1}^d \lambda_i^2(\Sigma_t^{-1}) = \sum_{i=1}^d \lambda_i^{-2}(\Sigma_t) \leqslant d\lambda_{\min}^{-2} \ , \\ \mathrm{tr}\left(\Sigma_t^{-1}\right) &= \sum_{i=1}^d \lambda_i(\Sigma_t^{-1}) \leqslant d\lambda_{\min}^{-1} \ ,\end{aligned}$$

which, in turn, give

$$\int \left(\left\|\Sigma_t^{-1}\right\|_F + z^\top \Sigma_t^{-2} z\right) \varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}z) \leqslant \sqrt{d}\lambda_{\min}^{-1} + d\lambda_{\min}^{-1} \leqslant 2d\lambda_{\min}^{-1} \,.$$

We conclude that

$$\begin{aligned}
\int |\varphi_\Sigma(z) - \varphi_{\Sigma'}(z)|\lambda(\,\mathrm{d}z) &= \int \left|\int_0^1 \frac{\partial}{\partial t}\varphi_{\Sigma_t}(z)\,\mathrm{d}t\right|\lambda(\,\mathrm{d}z) \\
&\leqslant \int \int_0^1 \frac{1}{2}\varphi_{\Sigma_t}(z)\left|\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\right|\mathrm{d}t\lambda(\,\mathrm{d}z) \\
&= \frac{1}{2}\int_0^1 \int \left|\frac{\partial}{\partial t}\log\varphi_{\Sigma_t}(z)\right|\varphi_{\Sigma_t}(z)\lambda(\,\mathrm{d}z)\,\mathrm{d}t \\
&\leqslant \frac{1}{2}\int_0^1 2d\lambda_{\min}^{-1}\|\Sigma' - \Sigma\|_F\,\mathrm{d}t \\
&= \frac{d}{\lambda_{\min}}\|\Sigma' - \Sigma\|_F \,.
\end{aligned}$$

□

Proposition B.6. *Consider a family of MTM transitions $\{P_\theta\}_{\theta\in\Theta}$ with Gaussian random walk marginal proposal densities whose covariances are contained in a compact subset of $\mathcal{C}_d^+$, the cone of symmetric positive-definite matrices. Suppose that the following Lipschitz condition holds: there exists $L < \infty$ such that, for all $x, y \in \mathcal{X}$*

$$\left|A_\theta^{(k)}(y|x) - A_{\theta'}^{(k)}(y|x)\right| \leqslant L\|\theta - \theta'\|_2 \,. \tag{B.4}$$

Then, there exists $C < \infty$ such that, for all functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_1 < \infty$,

$$\|P_\theta f - P_{\theta'} f\|_1 \leqslant C\|f\|_1\|\theta - \theta'\|_2 \,.$$

In particular, $\{P_\theta\}_{\theta\in\Theta}$ is V-Lipschitz for $V \equiv 1$.

Proof. By definition, we have

$$\|P_\theta f - P_{\theta'} f\|_1 = \sup_{x\in\mathcal{X}} |P_\theta f(x) - P_{\theta'} f(x)| \,.$$

For $\|f\|_1 < \infty$, we have

$$\frac{|f(x)|}{\|f\|_1} \leqslant 1 \,, \qquad \forall\, x \in \mathcal{X} \,. \tag{B.5}$$

We first consider the following development of $P_\theta f(x) - P_{\theta'} f(x)$:

$$\begin{aligned} P_\theta f(x) - P_{\theta'} f(x) &= \int_{\mathcal{X}} f(y) P_\theta(y|x) \lambda(\mathrm{d}y) - \int_{\mathcal{X}} f(y) P_{\theta'}(y|x) \lambda(\mathrm{d}y) \\ &= \int_{\mathcal{X}} f(y) \left[P_\theta(y|x) - P_{\theta'}(y|x) \right] \lambda(\mathrm{d}y) \\ &= \int_{\mathcal{X}} f(y) \left[R_\theta(x)\delta_x(y) + p_\theta(y|x) - R_{\theta'}(x)\delta_x(y) - p_{\theta'}(y|x) \right] \lambda(\mathrm{d}y) \\ &= \int_{\mathcal{X}} f(y) \left[(R_\theta(x) - R_{\theta'}(x))\, \delta_x(y) + (p_\theta(y|x) - p_{\theta'}(y|x)) \right] \lambda(\mathrm{d}y) \ . \end{aligned}$$

Then, using properties of integrals and the inequality (B.5), we find

$$\begin{aligned} \frac{|P_\theta f(x) - P_{\theta'} f(x)|}{\|f\|_1} &\leqslant \int_{\mathcal{X}} \frac{f(y)}{\|f\|_1} \left[|R_\theta(x) - R_{\theta'}(x)| \delta_x(y) + |p_\theta(y|x) - p_{\theta'}(y|x)| \right] \lambda(\mathrm{d}y) \\ &\leqslant \int_{\mathcal{X}} \left[|R_\theta(x) - R_{\theta'}(x)| \delta_x(y) + |p_\theta(y|x) - p_{\theta'}(y|x)| \right] \lambda(\mathrm{d}y) \\ &= |R_\theta(x) - R_{\theta'}(x)| + \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\mathrm{d}y) \ . \end{aligned}$$

Now, we note that

$$\begin{aligned} |R_\theta(x) - R_{\theta'}(x)| &= \left| 1 - \int_{\mathcal{X}} p_\theta(y|x)\lambda(\mathrm{d}y) - 1 + \int_{\mathcal{X}} p_{\theta'}(y|x)\lambda(\mathrm{d}y) \right| \\ &= \left| \int_{\mathcal{X}} \left[p_{\theta'}(y|x) - p_\theta(y|x) \right] \lambda(\mathrm{d}y) \right| \\ &\leqslant \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\mathrm{d}y) \ , \end{aligned}$$

which allows us to bound

$$|P_\theta f(x) - P_{\theta'} f(x)| \leqslant 2\|f\|_1 \int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\mathrm{d}y) \ . \tag{B.6}$$

Rearranging terms, we can write

$$\begin{aligned} p_\theta(y|x) - p_{\theta'}(y|x) &= \sum_{k=1}^{K} A_\theta^{(k)}(y|x) q_\theta^{(k)}(y|x) - \sum_{k=1}^{K} A_{\theta'}^{(k)}(y|x) q_{\theta'}^{(k)}(y|x) \\ &= \sum_{k=1}^{K} \left[A_\theta^{(k)} q_\theta^{(k)} - A_{\theta'}^{(k)} q_{\theta'}^{(k)} \right] (y|x) \\ &= \sum_{k=1}^{K} \left[A_\theta^{(k)} q_\theta^{(k)} - A_\theta^{(k)} q_{\theta'}^{(k)} + A_\theta^{(k)} q_{\theta'}^{(k)} - A_{\theta'}^{(k)} q_{\theta'}^{(k)} \right] (y|x) \\ &= \sum_{k=1}^{K} \left[A_\theta^{(k)} \left(q_\theta^{(k)} - q_{\theta'}^{(k)} \right) + \left(A_\theta^{(k)} - A_{\theta'}^{(k)} \right) q_{\theta'}^{(k)} \right] (y|x) \ . \end{aligned} \tag{B.7}$$

In that last expression, it is possible to directly bound the first term. Indeed, $A_\theta^{(k)} \leqslant 1$ and $q_\theta^{(k)} - q_{\theta'}^{(k)}$ may be bounded by Proposition B.5:

$$\begin{aligned}
\int_{\mathcal{X}} \left| \sum_{k=1}^{K} A_\theta^{(k)} \left(q_\theta^{(k)} - q_{\theta'}^{(k)} \right) (y|x) \right| \lambda(\,\mathrm{d}y) &\leqslant \sum_{k=1}^{K} \int_{\mathcal{X}} 1 \cdot \left| q_\theta^{(k)}(y|x) - q_{\theta'}^{(k)}(y|x) \right| \lambda(\,\mathrm{d}y) \\
&\leqslant \sum_{k=1}^{K} \int_{\mathcal{X}} |\varphi_{\Sigma^{(k)}}(z) - \varphi_{\Sigma'^{(k)}}(z)| \lambda(\,\mathrm{d}z) \\
&\leqslant \frac{d}{\lambda_{\min}} \sum_{k=1}^{K} \left\| \Sigma^{(k)} - \Sigma'^{(k)} \right\|_F \\
&\leqslant \frac{d}{\lambda_{\min}} \sum_{k=1}^{K} \|\theta - \theta'\|_2 \,, \\
&= \frac{Kd}{\lambda_{\min}} \|\theta - \theta'\|_2 \,, \qquad \text{(B.8)}
\end{aligned}$$

where $\lambda_{\min} > 0$ is the smallest eigenvalue over covariances in $\mathcal{K}$ compact. The second term of (B.7) can be bounded using the Lipschitz condition on $A_\theta^{(k)}$:

$$\int_{\mathcal{X}} \left| A_\theta^{(k)} - A_{\theta'}^{(k)} \right| q_{\theta'}^{(k)}(y|x) \lambda(\,\mathrm{d}y) \leqslant \int_{\mathcal{X}} L \|\theta - \theta'\|_2 q_{\theta'}^{(k)}(y|x) \lambda(\,\mathrm{d}y) = L\|\theta - \theta'\|_2 \,. \qquad \text{(B.9)}$$

Combining (B.8) and (B.9), we can finally bound the integral in (B.6). Indeed, we find

$$\begin{aligned}
\int_{\mathcal{X}} |p_\theta(y|x) - p_{\theta'}(y|x)| \lambda(\,\mathrm{d}y) &\leqslant \sum_{k=1}^{K} \int_{\mathcal{X}} A_\theta^{(k)} \left| q_\theta^{(k)} - q_{\theta'}^{(k)} \right| (y|x) \lambda(\,\mathrm{d}y) \\
&\quad + \sum_{k=1}^{K} \int_{\mathcal{X}} \left| A_\theta^{(k)} - A_{\theta'}^{(k)} \right| q_{\theta'}^{(k)}(y|x) \lambda(\,\mathrm{d}y) \\
&\leqslant \frac{Kd}{\lambda_{\min}} \|\theta - \theta'\|_2 + KL\|\theta - \theta'\|_2 \\
&\leqslant K \left(\frac{d}{\lambda_{\min}} + L \right) \|\theta - \theta'\|_2 \,,
\end{aligned}$$

which concludes the proof. $\square$

The Lipschitz condition on the acceptance probability (B.4) highly depends on the specific instance of the aMTM algorithm implemented. In particular, the expression $A_\theta^{(k)}$ involves the weight function $w_\theta^{(k)}$, the acceptance probability $\alpha_\theta^{(k)}$ and the conditional densities $q_\theta^{(k)}$. Hence, the choices of weights and covariance structure among candidates influence how we can verify (B.4) so we must resort to a case-by-case approach. Fontaine (2019, Section 5.5.2) contains all the details so we only report the general ideas here.

First, when the weight function is independent of θ (e.g. proportional to the target density) and candidates are chosen to be independent, then we do not require any additional assumption.

When the weight function depends on θ, then we can extract some general sufficient conditions. We require that the weight function and acceptance probability be Lipschitz in θ uniformly over their arguments; for any $\theta \neq \theta' \in \Theta^2$,

$$\sup_{y, y^{(-k)}, x, \theta \neq \theta'} \frac{\left|\bar{w}_\theta(y, y^{(-k)}|x) - \bar{w}_{\theta'}(y, y^{(-k)}|x)\right|}{\|\theta - \theta'\|_2} < \infty \ ,$$

$$\sup_{y, y^{(-k)}, x, x_*^{(k)}, \theta \neq \theta'} \frac{\left|\alpha_\theta(y, y^{(-k)}|x, x_*^{(k)}) - \alpha_{\theta'}(y, y^{(-k)}|x, x_*^{(k)})\right|}{\|\theta - \theta'\|_2} < \infty \ .$$

In the independent case, such conditions are easily verified by choosing functions that have continuous gradients and by supposing Θ compact and convex. In the extremely antithetic case, we can use similar arguments, but the details are more tedious since the conditional densities $q^{(k)}$ lie in some strict subspace of $\mathcal{X}^{K-1}$. When candidates are deterministic (e.g. RQMC or common random variable), these conditions become simpler as the conditional densities become degenerate: we can then drop the dependence on $y^{(-k)}$ and on $x_*^{(-k)}$.

B.3.3. Bounded updates

The set of parameters is given by

$$\theta = \left(\theta^{(1)}, \ldots, \theta^{(K)}\right)$$

where, in general, each component consists of a moving average, a covariance and a scale factor:

$$\theta^{(k)} = \left(\mu^{(k)}, \Sigma^{(k)}, l^{(k)}\right), \qquad k = 1, \ldots, K,$$

where $l^{(K)} = \log \lambda^{(K)}$. We denote by $\|\cdot\|_2$ the $\mathcal{L}_2$-norm; for elements of θ that are matrices, the contribution to $\|\theta\|$ will thus be the Frobenius norm $\|\cdot\|_F$ which corresponds to the $\mathcal{L}_2$-norm of the vectorized matrix. At iteration n, the available information to be used by the adaptation function is given by

$$\Xi_n = \left(k_n, y^{(1:K)}, x_*^{(1:K)}\right) \ .$$

Hence, we can describe the update function as

$$H_\theta\left(\Xi_n\right) = \begin{pmatrix} H_\theta^{(1)}\left(\Xi_n\right) \\ \vdots \\ H_\theta^{(K)}\left(\Xi_n\right) \end{pmatrix} \ ,$$

where $H_\theta^{(k)}$ corresponds to the update of $\theta^{(k)}$ as introduced in (4.1). Then

$$H_\theta^{(k)}(\Xi_n) = \begin{pmatrix} H_{\mu,\Sigma}^{(k)}(\Xi_n) \\ H_{l,\alpha}^{(k)}(\Xi_n) \end{pmatrix} ,$$

where $H_{\mu,\Sigma}^{(k)}$ corresponds to the (joint) update of $\mu^{(k)}$ and $\Sigma^{(k)}$, and where $H_{l,\alpha}^{(k)}$ corresponds to the update of $l^{(k)}$ using the acceptance probability.

Bounding $H_{l,\alpha}^{(k)}$ is easily achieved. In the ASWAM case, we have

$$H_{l,\alpha}^{(k)}(\Xi_n) = \mathbb{I}(\{k_n = k\}) \left[\alpha_\theta\left(y; y^{(-k)} | x_n; x_*^{(-k)}\right) - \alpha_*\right] ,$$

which can be bounded by

$$\left|H_{l,\alpha}^{(k)}(\Xi_n)\right| \leqslant \mathbb{I}(\{k_n = k\}) \left|\alpha_\theta\left(y; y^{(-k)} | x_n; x_*^{(-k)}\right) - \alpha_*\right| \leqslant 1 .$$

For AM or RAM updates, $H_{l,\alpha}^{(k)} = 0$.

Lemma B.1. *Let $H_{\mu,\Sigma}^{(k)}$ denote the AM or ASWAM update of $(\mu^{(k)}, \Sigma^{(k)})$. Then, if the sample space $\mathcal{X}$ and parameter space Θ are both compact,*

$$\sup_{\theta \in \mathcal{K}} \left\| H_{\mu,\Sigma}^{(k)} \right\|_1 < \infty .$$

Proof. The AM and ASWAM updates are such that

$$H_{\mu,\Sigma}^{(k)}(\Xi_n) = \mathbb{I}(\{k_n = k\}) \begin{pmatrix} x_{n+1} - \mu^{(k)} \\ (x_{n+1} - \mu^{(k)})(x_{n+1} - \mu^{(k)})^\top - \Sigma^{(k)} \end{pmatrix} .$$

Thus, $H_{\mu,\Sigma}^{(k)}$ only depends on θ, k_n, and x_{n+1}.

By definition, we have

$$\left\| H_{\mu,\Sigma}^{(k)} \right\|_1 = \sup_{(x_{n+1}, k_n) \in \mathcal{X} \times \{1, \dots, K\}} \left\| H_{\mu,\Sigma}^{(k)}(\Xi_n) \right\|_2 .$$

Obviously, the supremum over $k_n \in \{1, \dots, K\}$ is attained for $k_n = k$ because of the term $\mathbb{I}(\{k_n = k\})$. Hence, we find the following bound

$$\begin{aligned}
\left\| H_{\mu,\Sigma}^{(k)} \right\|_2 &\leqslant \left\| x_{n+1} - \mu^{(k)} \right\|_2 + \left\| (x_{n+1} - \mu^{(k)})(x_{n+1} - \mu^{(k)})^\top - \Sigma^{(k)} \right\|_F \\
&\leqslant \|x_{n+1}\|_2 + \left\|\mu^{(k)}\right\|_2 + \left\|x_{n+1} x_{n+1}^\top\right\|_F \\
&\quad + 2\left\|\mu^{(k)} x_{n+1}^\top\right\|_F + \left\|\mu^{(k)} \mu^{(k)\top}\right\|_2 + \left\|\Sigma^{(k)}\right\|_F .
\end{aligned}$$

Assuming $\mathcal{X}$ and Θ compact, then x_{n+1}, $\mu^{(k)}$, and $\Sigma^{(k)}$ are all bounded so that $\left\| H_{\mu,\Sigma}^{(k)} \right\|_2$ is uniformly bounded for $(x, k) \in \mathcal{X} \times \{1, \dots, K\}$, as well as for $\theta \in \mathcal{K}$. □

For the RAM update, we rewrite it as a Robbins-Monro recursion:

$$\begin{aligned}\Sigma_{n+1}^{(k)} &= S_n^{(k)}\left(I_d + \gamma_{n+1}\,\mathbb{I}\left(\{k_n = k\}\right)\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S_n^{(k)\top}\\ &= S_n^{(k)}S_n^{(k)\top} + \gamma_{n+1}\,\mathbb{I}\left(\{k_n = k\}\right)S_n^{(k)}\left(\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S_n^{(k)\top}\\ &=: \Sigma_n^{(k)} + \gamma_{n+1}H_{\Sigma_n}^{(k)}\left(\Xi_n\right)\ ,\end{aligned}$$

where

$$H_{\Sigma_n}^{(k)}\left(\Xi_n\right) = \mathbb{I}\left(\{k_n = k\}\right)S_n^{(k)}\left(\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S_n^{(k)\top}\ ,$$

with $u_{n+1} = y - x_n$ and $\Sigma^{(k)} = S^{(k)}S^{(k)\top}$.

Lemma B.2. *Let $H_\Sigma^{(k)}$ denote the RAM update function of $\Sigma^{(k)}$. Then, if the sample space $\mathcal{X}$ and parameter space Θ are both compact,*

$$\sup_{\theta\in\mathcal{K}}\left\|H_\Sigma^{(k)}\right\|_1 < \infty\ .$$

Proof. The norm of $H_\Sigma^{(k)}$ can be bounded as follows:

$$\begin{aligned}\left\|H_\Sigma^{(k)}\left(\Xi_n\right)\right\|_2 &\leqslant \left\|S^{(k)}\left(\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right)S^{(k)\top}\right\|_2\\ &\leqslant \left\|S^{(k)}\right\|_2\left\|\left[\alpha_\theta(y; y^{(-k)}|x_n; x_*^{(-k)}) - \alpha_*\right]\frac{u_{n+1}u_{n+1}^\top}{\|u_{n+1}\|_2^2}\right\|_2\left\|S^{(k)\top}\right\|_2\\ &\leqslant \left\|S^{(k)}\right\|_2\frac{\left\|u_{n+1}u_{n+1}^\top\right\|_2}{\|u_{n+1}\|_2^2}\left\|S^{(k)}\right\|_2\\ &\leqslant \left\|S^{(k)}\right\|_2\frac{\|u_{n+1}\|_2^2}{\|u_{n+1}\|_2^2}\left\|S^{(k)}\right\|_2 = \left\|S^{(k)}\right\|_2^2\ .\end{aligned}$$

For Θ compact, we find $\left\|S^{(k)}\right\|_2$ to be uniformly bounded, which implies that $\left\|H_\Sigma^{(k)}\right\|_1$ is uniformly bounded for $\theta \in \Theta$ as well as for all Ξ_n. That is, $\sup_{\theta\in\mathcal{K}}\left\|H_\Sigma^{(k)}\right\|_1 < \infty$. □

B.3.4. Continuity of the convergence metric

Recall the metric used to compare the iterated transition to the target density:

$$\Delta_n(x, \theta) = \|P_\theta^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}}\ .$$

Lemma B.3. *Let $F : \mathcal{W} \to \mathbb{R}$ be a function defined by*

$$F(w) = \int_{\mathcal{T}} f(w,t)\lambda(\,\mathrm{d}t) \ ,$$

where $f : \mathcal{W} \times \mathcal{T} \to \mathbb{R}$ is continuous w.r.t. (w,t) and where λ denotes the Lebesgue measure on $\mathbb{R}$. Suppose there exists a function $g : \mathcal{T} \to \mathbb{R}$ such that $|f(w,t)| \leqslant |g(t)|$ for all $(w,t) \in \mathcal{W} \times \mathcal{T}$ with $\lambda(|g|) < \infty$. Then, F is a continuous function of w on the whole of $\mathcal{W}$.

Proof. The function F is continuous on the whole of $\mathcal{W}$ if and only if $\lim_{n\to\infty} F(w_n) = F(w)$ for any sequence $w_n \to w$. Then, let $\{w_n\}_{n\in\mathbb{N}} \subseteq \mathcal{W}$ be an arbitrary sequence with $w_n \to w \in \mathcal{W}$ and define, for all $n \in \mathbb{N}$, $f_n : \mathcal{T} \to \mathbb{R}$ by $f_n(t) = f(w_n, t)$. By the continuity of f w.r.t. w, we know that $f_n(t) \to f(w,t)$ point-wise. By hypothesis, we have

$$|f_n(t)| = |f(w_n,t)| \leqslant |g(t)| \ , \qquad n \in \mathbb{N} \ .$$

Now, write

$$\lim_{n\to\infty} F(w_n) = \lim_{n\to\infty} \int_{\mathcal{T}} f(w_n,t)\lambda(\,\mathrm{d}t) = \lim_{n\to\infty} \int_{\mathcal{T}} f_n(t)\lambda(\,\mathrm{d}t) = \lim_{n\to\infty} \lambda(f_n(\cdot)) \ .$$

By the Monotone Convergence Theorem, we find

$$\lim_{n\to\infty} F(w_n) = \lambda\left(\lim_{n\to\infty} f_n(\cdot)\right) = \lambda\left(f(w,\cdot)\right) = F(w) \ ,$$

which concludes the proof. □

Lemma B.4. *Let P_θ be a MTM transition using a given set parameters $\theta \in \Theta$. Then, the acceptance probability through candidate k from the current state x to some other state y, $A_\theta^{(k)}(y|x)$, is a continuous function of (x, y, θ) assuming that each of $q_\theta^{(-k)}$, $\bar{w}^{(k)}$ and α_{MTM} are continuous functions of their arguments and parameters, and that the conditional densities $q_\theta^{(-k)}$ are uniformly bounded above by some integrable function $q^+ : \mathcal{X}^{K-1} \to \mathbb{R}_{\geqslant 0}$.*

Proof. This result is a direct consequence of Lemma B.3. The complete argument may be found in Fontaine (2019, Lemma 5.11). □

Lemma B.5. *Let P_θ be a MTM transition using a given set parameters $\theta \in \Theta$. Then, the integrated acceptance probability through candidate k from the current state x, $\overline{A}_\theta^{(k)}(x)$, is a continuous function of (x,θ) assuming that $A_\theta^{(k)}(y|x)$ is a continuous function of (x, y, θ) and assuming that each $q_\theta^{(k)}(y|x)$ is a density, with respect to the Lebesgue measure on $\mathbb{R}^d$, such that there exists an integrable function $q^+ : \mathcal{X} \to \mathbb{R}_{\geqslant 0}$ with $q_\theta^{(k)}(y|x) \leqslant q^+(y)$ uniformly for (x, θ, k). Furthermore, the rejection probability $R_\theta(x)$ is also a continuous function of (x,θ).*

Proof. This result is a direct consequence of Lemma B.3. The complete argument may be found in Fontaine (2019, Lemma 5.12). □

For a Markov transition taking the form of a MH density, i.e.

$$P(\,\mathrm{d}y|x) = p(y|x)\lambda(\,\mathrm{d}y) + R(x)\delta_x(\,\mathrm{d}y)\ , \tag{B.10}$$

we can write the iterated transition using the following recursion,

$$P^n(\,\mathrm{d}y|x) = p^n(y|x)\lambda(\,\mathrm{d}y) + R^n(x)\delta_x(\,\mathrm{d}y),$$

where

$$p^n(y|x) = \int_{\mathcal{X}} p^{n-1}(y|z)p(z|x)\lambda(\,\mathrm{d}z)\ ,$$

with the convention $p^0(y|x) = \delta_x(y)$.

Corollary B.1. *Under the setup and conditions of Lemma B.5, the iterated MTM transition, $p^n_\theta(y|x)$, is a continuous function of (x, y, θ) for all $n \in \mathbb{N}$.*

Proof. We proceed by induction over $n \geqslant 1$ to show that $p^n_\theta(y|x)$ is continuous with respect to (x, y, θ) and is uniformly bounded by $K^n\bar{q}^{n-1}q^+(y)$, where $\bar{q} = \sup_{\mathcal{X}} q^+ < \infty$.

For $n = 1$, we have

$$p^1_\theta(y|x) = \int_{\mathcal{X}} \delta_z(y)p_\theta(z|x)\lambda(\,\mathrm{d}z) = p_\theta(y|x) = \sum_{k=1}^{K} A^{(k)}_\theta(y|x)q^{(k)}_\theta(y|x)\ ,$$

which is a sum of products of continuous functions and is therefore continuous. The uniform bound is direct:

$$\left|p^1_\theta(y|x)\right| \leqslant \sum_{k=1}^{K}\left|A^{(k)}_\theta(y|x)q^{(k)}_\theta(y|x)\right| \leqslant \sum_{k=1}^{K} 1\cdot q^+(y) = K\cdot q^+(y) = K^1\bar{q}^{1-1}\cdot q^+(y)\ .$$

For $n > 1$, we suppose that $p^{n-1}_\theta(y|x)$ is continuous and uniformly bounded by $K^{n-1}\bar{q}^{n-2}q^+(y)$. We use Lemma B.3; to this end, we let

$$F(w) = p^n_\theta(y|x) = \int_{\mathcal{X}} p^{n-1}_\theta(y|z)p_\theta(z|x)\lambda(\,\mathrm{d}z)\ ,$$

the variables $(x, y, \theta) = w \in \mathcal{W}$ with $\mathcal{W} = \mathcal{X}^2 \times \Theta$, the integrated variable $z = t \in \mathcal{T}$ with $\mathcal{T} = \mathcal{X}$, and the integrand

$$f(w, t) = p^{n-1}_\theta(y|z)p_\theta(z|x)\ .$$

Since $p_\theta^{n-1}(y|z)$ is continuous w.r.t. (y,z,θ) by induction hypothesis and since $p_\theta(z|x)$ is continuous w.r.t. (x,z,θ) by assumption (see the case $n=1$), we find that f is a continuous function of all its arguments. Also, the induction hypothesis implies the following uniform bound

$$\left|p_\theta^{n-1}(y|z)\right| \leqslant K^{n-1}\bar{q}^{n-2}q^+(y) \leqslant K^{n-1}\bar{q}^{n-2}\sup_{\mathcal{X}} q^+ = K^{n-1}\bar{q}^{n-1} \ .$$

Hence, we find

$$|f(w,t)| \leqslant \left|p_\theta^{n-1}(y|z)\right||p_\theta(z|x)| \leqslant K^n\bar{q}^{n-1}q^+(z) =: g(t) \ .$$

Since q^+ is integrable and $K, \bar{q} < \infty$, we find that g is integrable for each fixed n. Finally, Lemma B.3 implies that F is continuous w.r.t. w, that is, $p_\theta^n(y|x)$ is continuous w.r.t. $(x,y,\theta) \in \mathcal{X}^2 \times \Theta$ for each fixed n. □

Theorem B.2. *Let $\{P_\theta\}_{\theta\in\Theta}$ be a family of MTM transitions indexed by its set of parameters θ, and suppose that the target density π and each P_θ satisfy the conditions of Proposition B.2. Further suppose that Θ is compact and that the resulting chain $\{X_n\}_{n\in\mathbb{N}}$ is bounded in probability. Then, if the conditions of Lemma B.5 hold, the adaptive chain satisfies the bounded convergence condition.*

Proof. We use a result by Craiu et al. (2015, Proposition 23) restated as Proposition 4.1 in the main text.

All conditions of the result are verified except the continuity of $(x,\theta) \mapsto \Delta_n(x,\theta)$. Indeed, the MTM transitions all admit π as their invariant distribution because of the detailed balance condition (Proposition B.1). They are also ergodic with respect to π by Proposition B.2 and then Harris-ergodic with respect to π by Proposition B.3.

To verify the continuity of Δ_n, we proceed in a similar fashion to Roberts and Rosenthal (2007, Corollary 11) in the MH case. We develop Δ_n using the decomposition of the iterated transition (B.10):

$$\begin{aligned}
\Delta_n(x,\theta) &= \|P_\theta^n(\cdot|x) - \pi(\cdot)\|_{\mathrm{TV}} \\
&= \sup_{B\in\mathcal{B}(\mathcal{X})} |P_\theta^n(B|x) - \pi(B)| \\
&= \sup_{B\in\mathcal{B}(\mathcal{X})} \left|\int_B P_\theta^n(\,\mathrm{d}y|x) - \int_B \pi(\,\mathrm{d}y)\right| \\
&= \sup_{B\in\mathcal{B}(\mathcal{X})} \left|R_\theta^n(x)\delta_x(B) + \int_B p_\theta^n(\,\mathrm{d}y|x) - \int_B \pi(\,\mathrm{d}y)\right| \\
&= R_\theta^n(x) + \frac{1}{2}\int_{\mathcal{X}} |p_\theta^n(y|x) - \pi(y)|\lambda(\,\mathrm{d}y) \ .
\end{aligned}$$

By inspection of the last expression, we can show that Δ_n is indeed a continuous function of its arguments.

By Lemma B.5, we know that $R_\theta^n(x)$ is a continuous function of (x,θ).

We then use Lemma B.3 to show that the integral is indeed continuous w.r.t. (x, θ). By Corollary B.1, we know that $p_\theta^n(y|x)$ is continuous w.r.t. (x, y, θ). Since π is assumed to be a density w.r.t. the Lebesgue measure, we have that $|p_\theta^n(y|x) - \pi(y)|$ is continuous w.r.t. (x, y, θ). Thus, we only need a (x, θ)-uniform and integrable bound on $|p_\theta^n(y|x) - \pi(y)|$. By the triangle inequality, we have

$$|p_\theta^n(y|x) - \pi(y)| \leqslant p_\theta^n(y|x) + \pi(y) \ .$$

By Corollary B.1, we find a uniform and integrable bound on the first term; the target density is independent of (x, θ) and integrable:

$$|p_\theta^n(y|x) - \pi(y)| \leqslant K^n \bar{q}^{n-1} q^+(y) + \pi(y) \in L_1(\lambda) \ .$$

Hence, all conditions of Lemma B.3 are verified, which implies that $\int_{\mathcal{X}} |p_\theta^n(y|x) - \pi(y)|\lambda(\,\mathrm{d}y)$ is continuous w.r.t. (x, θ). We conclude that Δ_n is continuous w.r.t. (x, θ) since it corresponds to a linear combination of continuous functions. □

Received December 2020