ORIGINAL PAPER

Convergence rate of multiple-try Metropolis independent sampler

Xiaodong Yang[1] · Jun S. Liu[1]

Abstract

The multiple-try Metropolis method is an interesting extension of the classical Metropolis–Hastings algorithm. However, theoretical understanding about its usefulness and convergence behavior is still lacking. We here derive the exact convergence rate for the multiple-try Metropolis Independent sampler (MTM-IS) via an explicit eigen analysis. As a by-product, we prove that an naive application of the MTM-IS is less efficient than using the simpler approach of "thinned" independent Metropolis–Hastings method at the same computational cost. We further explore more variants and find it possible to design more efficient algorithms by applying MTM to part of the target distribution or creating correlated multiple trials.

Keywords Convergence rate · Eigenvalues · Markov chain · Monte Carlo · Transition function

1 Introduction

1.1 Fundamental Metropolis–Hastings method

Markov chain Monte Carlo (MCMC) methods have played important roles in statistical computing and Bayesian inference and have attracted much attention from both theoretical researchers and practitioners. In a nutshell, the set of methods provide general and practical recipes for generating random draws from any given target probability distribution known up to a normalizing constant. Specifically, such an algorithm generates a time-homogeneous Markov chain with its stationary distribution being the target one. Under mild assumptions, this chain converges to the target distribution geometrically (Roberts and Tweedie 1996; Liu et al. 1995). See Liu (2008) and Brooks et al. (2011) for more comprehensive reviews. The scheme first proposed by Metropolis et al. (1953) and then generalized by Hastings (1970) is arguably the most popular and fundamental construction among all MCMC methods. Let $\pi(\cdot)$ denote the target probability distribution/density function on the state space $\mathcal{X}$. The *Metropolis–Hastings* method constructs a Markov chain $x^{(1)}, x^{(2)}, \ldots,$ on $\mathcal{X}$ as follows. At step $t + 1$, it proposes a

new state y from a user-specified transition function $p(x, y)$, i.e., $y \sim p(x^{(t)}, \cdot)$. Then, the next state $x^{(t+1)}$ is equal to y with probability ρ and to $x^{(t)}$ with probability $1 - \rho$, where

$$\rho = \min \left\{ 1, \frac{\pi(y) p(y, x^{(t)})}{\pi(x^{(t)}) p(x^{(t)}, y)} \right\}.$$

This design ensures that the generated Markov chain satisfies *the detailed balance* with respect to π, which guarantees the chain's reversibility and convergence under mild conditions.

1.2 Geometric convergence

A Markov chain with transition function A is said to be geometrically ergodic if, for π-almost everywhere x, $\|A^n(x, \cdot) - \pi(\cdot)\| \le C(x) r^n$ holds true with constant $r \in (0, 1)$. Here $\| \cdot \|$ denotes a distance metric between two probability measures, usually taken as the total variation (TV) distance. Other modes of convergence, such as convergence in χ^2-distance (which implies the convergence in total variation), have also been investigated (Liu et al. 1995; Liu 2008). Establishing this inequality and deriving sharp bounds on the rate r are seen as central tasks in studying MCMC algorithms (Tierney 1994; Liu et al. 1995; Roberts and Tweedie 1996).

As a generalization of the standard Metropolis–Hastings algorithm, the multiple-try Metropolis (MTM) scheme as formalized in Liu et al. (2000) allows one to draw multiple trials at each step and select one according to a specially designed probability distribution. Although intuitively

✉ Jun S. Liu
    jliu@stat.harvard.edu

    Xiaodong Yang
    yangxiaodong0912@gmail.com

1   Department of Statistics, Harvard University, Oxford Street, Cambridge, MA 02138, USA

the MTM scheme enables one to escape from local optimums more easily, there is little theoretical understanding of the convergence rate of any form of the MTM algorithm, making it a challenging practical concern when deciding whether a MTM approach should be employed for a specific problem. Existing theoretical results on the Metropolis–Hastings algorithm clearly cannot be easily extended to the MTM algorithm. Indeed, getting sharp bounds on the convergence rate of any general-purpose Metropolis–Hastings algorithm can be extremely challenging, except for the Independent Metropolis–Hastings (IMH) algorithm (which is also called the Metropolised independence sampler by Liu (1996) and the independence Metropolis chain by Tierney (1994)). We are therefore tempted to consider whether the IMH's multiple-try version, which we call the *multiple-try Metropolis Independent sampler* (MTM-IS), can be tackled theoretically.

1.3 Convergence rate of independent Metropolis–Hastings algorithm

Geometrical ergodicity is not guaranteed for a general Metropolis–Hastings algorithm unless we impose suitable restrictions (Roberts and Tweedie 1996), and exact convergence rates for Metropolis–Hastings algorithms are rare to find (Diaconis and Saloff-Coste 1998). In practice, geometric ergodicity is often established under the 'drift-and-minorization' framework (Diaconis et al. 2008). But this technique usually results in a very conservative bound of the convergence rate, not quite practically useful. Because of the very special structure of the IMH algorithm, explicit eigen-analyses of its transition matrix for the finite-discrete state space case were obtained by Liu (1996), which results in the exact convergence rate of the IMH algorithm (also a very tight bound on the constant in front of the rate) and offers a comparison with classical rejection sampling and importance sampling. Atchadé and Perron (2007) studies the continuous case by determining the full spectrum of the transition operator of the IMH algorithm. A recent preprint of Wang (2020) combines previous results and provides a lower bound, hence determining the exact convergence rate. In this paper, we impose similar conditions on the MTM-IS and study its exact convergence rate.

1.4 Multiple-try Metropolis and its variants

The original idea of multiple-try Metropolis (MTM) comes from chemical physicists interested in molecular simulations (Frenkel et al. 1996). Its general formulation constructed in Liu et al. (2000) inspires the development of Ensemble MCMC methods by Neal (2011), connects with particle filtering (Martino et al. 2014), and stimulates ideas of parallelizing MCMC (Calderhead 2014; Yang et al. 2018). We

refer interested readers to the review of Martino (2018). Intuitively, the MTM approach enables one to explore the sample space more broadly, and thus potentially gains efficiency in avoiding being trapped in local modes. The method has been incorporated in some applications such as model selection (Pandolfi et al. 2010) and Bayes factor estimation (Dai and Liu 2020).

In the context of molecular simulations (Frenkel et al. 1996), the multiple-try strategy is often applied to a target distribution in which the state space can be partitioned into two parts: position and orientation, i.e., $\mathbf{x} = (\mathbf{x}^p, \mathbf{x}^o)$. For a given $\mathbf{x}^p$, evaluating multiple configurations corresponding to different orientations, $\pi(\mathbf{x}^p, \mathbf{x}^{o1}), \ldots, \pi(\mathbf{x}^p, \mathbf{x}^{om})$ is not much more expensive than evaluating a single $\pi(\mathbf{x}^p, \mathbf{x}^o)$. Thus, MTM can be quite useful in facilitating an efficient move: we can propose the new configuration by (a) first proposing the position $\mathbf{x}^p_{(new)}$; (b) associating with it multiple orientations $\mathbf{x}^{o1}_{(new)}, \ldots, \mathbf{x}^{om}_{(new)}$; (c) picking one from them properly, and (d) using the MTM rule to do acceptance/rejection. In addition to this case, MTM is also particularly useful when combined with directional sampling, as in (Liu et al. 2000; Dai and Liu 2020). Specifically, given a sampling direction $\mathbf{e}$ at position $\mathbf{x}$, multiple trials are drawn simultaneously as $r_1, \ldots, r_m \sim p(r)$ to construct $\mathbf{y}_j = \mathbf{x} + r_j \mathbf{e}$.

Several variants of the MTM are worth mentioning: Craiu and Lemieux (2007) propose to use correlated trials to accelerate MTM and introduces antithetic and stratified sampling to bring correlation; Casarin et al. (2013) argue that multiple independent trials from different distributions are worth considering, and connect to interactive sampling algorithms. Theoretically, Bédard et al. (2012) conducts a scaling analysis for MTM. However, to the best of our efforts, we can not find any existing result on the convergence rate of an MTM algorithm.

In this paper, we report the exact convergence rate of the MTM-IS for general target $\pi(\cdot)$ and proposal $p(\cdot)$. The result is somewhat surprising as it shows that the MTM-IS with k multiple tries is not as efficient as simply repeating the standard IMH algorithm k times, thus suggesting that the we may want to design the k multiple proposals to be "over-dispersed" (e.g., negatively correlated) in order to take advantage of the MTM structure. Another useful scenario, as discussed previously and detailed in Sect. 5.1, is to help proposing a better configuration for a general Metropolis–Hastings algorithm by orienting part of the proposal better via MTM.

The rest of the article is organized as follows. Section 2 carries out an eigenvalue analysis of MTM-IS; Sect. 3 specifies the exact convergence rate of MTM-IS under the total variation distance and offers an inequality to compare MTM-IS with its corresponding "thinned" IMH algorithm (i.e., taking one draw from every k iterations of the sampler); Sect. 4

provides some empirical results for multivariate Gaussian and Gaussian mixtures; Sect. 5 discusses several variants and extensions of MTM; and Sect. 6 concludes the article with a short remark.

2 Eigen-analysis of multiple-try Metropolis independent sampler

2.1 Notations

Throughout the article, we use $\mathcal{X}$ to denote the state space, which can be either discrete or continuous. Notations $\pi(x)$ and $p(x, y)$ represent the target and proposal distributions, respectively, with $x, y \in \mathcal{X}$. If proposal distribution is independent of the current state x, we write it as $p(y)$. The actual transition function/probability/density of the MCMC algorithm is denoted by $A(x, y)$. A collection of multiple trials of size k is written as $\mathbf{y} = (y_1, \ldots, y_k)$. We consider the total variation distance for any two (signed) measures P and Q, which is defined as $\|P - Q\|_{TV} = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$, where $\mathcal{F}$ denotes the σ-field common to P and Q (e.g., the Borel σ-field for most common uses). In Sect. 5, we slightly abuse the notation by letting $p(x, \mathbf{y})$ be the proposal distribution for $x \in \mathcal{X}$ and $\mathbf{y} = (y_1, \ldots, y_k) \in \mathcal{X}^k$, as we would consider multiple correlated trials in this section. Besides, we write $p(x, \mathbf{y}_{(-j)} \mid y_j)$ as the conditional distribution of $\mathbf{y}_{(-j)} \equiv (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_k)$ given y_j and x. Lastly, $p_j(x, y_j) = \int p(x, \mathbf{y}) d\mathbf{y}_{-j}$ denotes the conditional marginal distribution of y_j given x.

2.2 Description of the algorithms

The general framework of the MTM as formulated in Liu et al. (2000) is summarized in Algorithm 1. Let the current state be x, and let the number of multiple tries be k. With a proposal transition function $p(x, y)$ that defines the conditional distribution of y, we define the *generalized importance weight* as

$$w(y|x) = \frac{\pi(y)}{p(x, y)} \lambda(x, y) \tag{1}$$

where λ is a symmetric non-negative function (i.e., $\lambda(x, y) = \lambda(y, x) \geq 0$, $\forall x, y$). Thus, the acceptance/rejection ratio in a general MH algorithm is just the ratio of the generalized importance weights.

Here, $x_1^*, x_2^*, \ldots, x_{k-1}^*$ are called *balancing trials*, which are drawn to guarantee the detailed balance. Liu et al. (2000) also extend the MTM for generating non-independent multiple trials such as semi-deterministic ones along a direction. If we choose $p(x, y) = p(y)$, we can modify this algorithm to avoid drawing additional balancing trials as the algorithm

---

Algorithm 1 Multiple-Try Metropolis: the current state is x.

1: Generate multiple trials $y_1, \ldots, y_k$ independently from $p(x, \cdot)$; compute their respective weights $w(y_j \mid x)$ as defined in (1) for $j = 1, \ldots, k$.
2: Select index J with probability proportional to $w(y_j \mid x)$ and define $y = y_J$.
3: Draw $x_1^*, x_2^*, \ldots, x_{k-1}^*$ independently from $p(y, \cdot)$. And set $x_k^* = x$.
4: Accept y with the ratio $\rho = \min\left\{1, \frac{\sum_j w(y_j|x)}{\sum_j w(x_j^*|y)}\right\}$.

---

is still valid if we simply replace the x_j^* by y_j in computing ρ. This modified version is summarized in Algorithm 2 and named the MTM-IS(k). In this case, we further select $\lambda(x, y) \equiv 1$ then the generalized importance weight (1) turns out to be $w(y \mid x) = \pi(y)/p(y)$, coinciding with the standard notation of importance ratio. In order to simplify the notations, we could write

$$w(y) = \pi(y)/p(y). \tag{2}$$

Algorithm 2 MTM-IS: the current state is x.

1: Draw multiple trials $y_1, \ldots, y_k$ independently from $p(\cdot)$; compute $w(y_j)$ as defined in (2) for $j = 1, \ldots, k$.
2: Select index J with probability proportional to $w(y_j)$ and define $y = y_J$.
3: Compute $W = \sum_{j=1}^k w(y_j)$.
4: Accept y with the following probability $\rho = \min\left\{1, \frac{W}{W - w(y) + w(x)}\right\}$.

In theory, we assume that π is absolutely continuous with respect to p, so that this importance weight can be interpreted as the Radon-Nikodym derivative. In practice, one should always choose p so that its support covers that of π for the algorithm to work well. The main result of this section is stated in Theorem 2, which can be viewed as a generalization of the results in Liu (1996) and Atchadé and Perron (2007) and provides the exact convergence rate of MTM-IS.

2.3 Transition distribution decomposition

Theorem 1 *The transition distribution of MTM-IS can be decomposed as*

$$A(x, dy) = R(x)\delta_x(dy) + \min\{H_k[w(x)], H_k[w(y)]\}\pi(y)dy, \tag{3}$$

where H_k is defined as

$$H_k(z) = k \underbrace{\int \ldots \int}_{k-1} \frac{1}{z + \sum_{i=1}^{k-1} w(y_i)} \prod_{i=1}^{k-1} p(y_i)dy_i, \tag{4}$$

and $R(x) = 1 - \int_{\mathcal{X}} \min\{H_k[w(x)], H_k[w(y)]\}\pi(y)\mathrm{d}y \in [0,1]$ denotes the rejection probability when the current state is $x \in \mathcal{X}$. In particular, $H_k(z)$ is a strictly decreasing function in z. For $k = 1$, H_k degenerates to $H_1(z) = z^{-1}$.

Proof Let $x \notin B \subset \mathcal{X}$ be measurable, the probability of proposing an element in B and accepting it is

$$A(x, B) = \mathbb{P}\left[\bigcup_{j=1}^{k}\left\{(y_j \in B) \cap (J = j) \cap (y_J \text{ gets accepted})\right\}\right]$$
$$= k\mathbb{P}\left[\{(y_k \in B) \cap (J = k) \cap (y_k \text{ gets accepted})\}\right].$$

The last equality appears irrelevant to x, but the importance ratio $w(x) = \pi(x)/p(x)$ matters when deciding whether or not the chosen y_J is accepted. Furthermore,

$$A(x, B)$$
$$= k\underbrace{\int \cdots \int}_{k-1} \int_B \frac{w(y)}{w(y) + \sum_{j=1}^{k-1} w(y_j)}$$
$$\min\left[1, \frac{w(y) + \sum_{j=1}^{k-1} w(y_j)}{w(x) + \sum_{j=1}^{k-1} w(y_j)}\right] p(y)\mathrm{d}y \prod_{j=1}^{k-1} p(y_j)\mathrm{d}y_j$$
$$= k\underbrace{\int \cdots \int}_{k-1} \int_B$$
$$\min\left[\frac{w(y)}{w(y) + \sum_{j=1}^{k-1} w(y_j)}, \frac{w(y)}{w(x) + \sum_{j=1}^{k-1} w(y_j)}\right]$$
$$p(y)\mathrm{d}y \prod_{j=1}^{k-1} p(y_j)\mathrm{d}y_j$$
$$= \int_B \min\{H_k[w(x)], H_k[w(y)]\}\pi(y)\mathrm{d}y,$$

where H_k is as defined in (4). Thus, the overall rejection probability is

$$R(x) = 1 - \int_{\mathcal{X}} \min\{H_k[w(x)], H_k[w(y)]\}\pi(y)\mathrm{d}y, \quad (5)$$

and the prescribed decomposition (3) is thus proved. $\qquad\square$

Let $w^* \triangleq \inf\{u > 0 : \pi(x : w(x) \leq u) = 1\}$ be the *essential supremum* of $w(x)$ on $\mathcal{X}$ w.r.t. $\pi(\cdot)$ (i.e., w^* is the smallest value such that $w(x) \leq w^*$ with π-probability 1). Since $H_k(w)$ is a monotone decreasing function of w (Theorem 1), we have an upper bound $R(x) \leq 1 - H_k(w^*)$. Furthermore, since

$$A(x, \mathrm{d}y) = R(x)\delta_x(\mathrm{d}y) + \min\{H_k[w(x)],$$
$$H_k[w(y)]\}\pi(y)\mathrm{d}y \geq H_k(w^*)\pi(y)\mathrm{d}y,$$

we have the following mixture representation of the transition function, convenient for comparing with π:

$$A(x, \mathrm{d}y) = H(w^*)\pi(y)\mathrm{d}y + [1 - H(w^*)]q_{\mathrm{res}}(x, \mathrm{d}y), \quad (6)$$

where $q_{\mathrm{res}}(x, B) := \dfrac{A(x, B) - H(w^*)\pi(B)}{1 - H(w^*)}$. This representation can be used to facilitate a coupling argument to prove the geometric convergence of the Markov chain (more details in Sect. 3).

2.4 Spectrum of the transition operator

Now we provide a result to fully characterize the spectrum of the transition operator induced by the MTM-IS algorithm. A similar result was derived for the IMH algorithm by Liu (1996) for the discrete state-space case, and then by Atchadé and Perron (2007) in general. To be concrete, we introduce the following definitions.

Definition 1 Let $A(x, y)$ be the transition function of a
    Markov chain with π as its invariant distribution. We define its transition operator $K : L^2(\pi) \to L^2(\pi)$ as

$$Kf(x) = \int f(y)A(x, y)dy. \quad (7)$$

It computes the conditional mean and is called the forward operator in Liu et al. (1995).

Definition 2 Let K_0 be the restriction of K onto $L_0^2(\pi)$, the orthogonal complement of the constant function of $L^2(\pi)$. Then the *spectrum* of K_0 is

$$\sigma(K_0) \triangleq \{\lambda \in \mathbb{R} : K_0 - \lambda I \text{ is non-invertible}\}. \quad (8)$$

The *essential range* of a function R is

$$\text{ess-ran}(R) \triangleq \{\lambda \in \mathbb{R} : \pi(x :| R(x) - \lambda | < \epsilon) > 0, \forall \epsilon > 0\}.$$

Theorem 2 *Let K be the transition operator defined by the MTM-IS algorithm, and let K_0 be similarly defined as in Definition 1. Then, $\sigma(K_0) \subseteq \text{ess-ran}(R)$, where R is the rejection probability defined in (5). The equality holds if $\forall \alpha \in \text{ess-ran}(R)$, $\pi\{y : R(y) = \alpha\} = 0$.*

Since the proof is mostly technical, we defer it to the Appendix. From (5) and Theorem 1, it is obvious to see that an upper bound of $R(x)$ is $1 - H_k(w^*)$. This implies that there is a gap between 1 and the upper edge $1 - H(w^*)$ of the spectrum $\sigma(K_0)$, provided that $w^* < \infty$. For the finite discrete state-space case, $H(w^*) = 1/w^*$, and $1 - H(w^*)$ is the exact convergence rate of the chain.

3 Convergence rate and algorithmic comparison

3.1 Convergence in χ^2-distance

The χ^2-distance between two probability distributions π and p is defined as

$$d_\chi^2(\pi, p) = \text{var}_\pi[p(x)/\pi(x)]. \tag{9}$$

Let $p_n(x) = A_n(p_0, x)$ denote the distribution of X_n, the state of the Markov chain after n steps from initialization p_0. It was shown in Liu et al. (1995) that $d_\chi(\pi, p_n) \leq \|K_0^n\|_2 d_\chi(\pi, p_0)$, where $\|\cdot\|_2$ is L^2-norm of the operator K_0. It is easy to show that (Liu et al. 1995)

$$\rho = \lim_{n \to \infty} \|K_0^n\|_2^{1/n} \tag{10}$$

is the spectral radius of K_0 (Liu et al. 1995), which is equal to the maximum of $\sigma(K_0)$ in absolute value. As shown in Theorem 2, this is bounded by $1 - H(w^*)$. Thus, $d_\chi(\pi, p_n) \leq (1 - H(w^*))^n d_\chi(\pi, p_0)$. It also follows from the Cauchy-Schwartz inequality that

$$\begin{aligned}
\|p_n - \pi\|_{L_1} &= \int \frac{|p_n(x) - \pi(x)|}{\sqrt{\pi(x)}} \sqrt{\pi(x)} dx \\
&\leq \left[\int \frac{(p_n(x) - \pi(x))^2}{\pi(x)} dx \right]^{1/2} = d_\chi(\pi, p_n) \\
&\leq (1 - H(w^*))^n d_\chi(\pi, p_0). \tag{11}
\end{aligned}$$

Thus, the L_1 distance between p_n and the target π, also known as their *total variation distribution* and denoted as $\|p_n - \pi\|_{TV}$, decreases geometrically bounded by the same rate.

3.2 Maximal total variation distance

Definition 3 Let the transition function of a Markov chain be $A(\cdot, \cdot)$, with the corresponding stationary distribution $\pi(\cdot)$. The *maximal total variation distance* between the Markov chain's n-step distribution and π is

$$d(n) = \operatorname*{ess\,sup}_{x \in \mathcal{X}} \|A^n(x, \cdot) - \pi(\cdot)\|_{TV}. \tag{12}$$

Moreover, the quantity

$$r = \limsup_{n \to \infty} d(n)^{\frac{1}{n}} \tag{13}$$

is called the *exact convergence rate* of the Markov chain.

Since the total variation distance is equivalent to the L_1 distance $\|p - \pi\|_{TV} = 2\|p - \pi\|_{L^1}$ between two probability measures π and p, it is easy to see from definition of (10) and Eq. (11) that rate $r \leq \rho$. In the following, we use another a coupling argument to validate this upper bound r. Moreover, we will also show that for the transition kernel defined by Algorithm 2, inequality $r \geq \rho$ also holds. We need the following lemmas to prove our results.

Lemma 1 (Coupling) (Levin and Peres 2017) *Suppose* $(\Psi_t, \widetilde{\Psi}_t)_{t=0}^\infty$ *are a pair of Markov chains with the same transition rule satisfying: (i) If* $\Psi_i = \widetilde{\Psi}_i$ *for some* i, *then for any* $j \geq i$, $\Psi_j = \widetilde{\Psi}_j$; *and (ii)* $\widetilde{\Psi}_0 \sim \pi$. *Then, for* $\tau = \min\{n : \Psi_n = \widetilde{\Psi}_n\}$, *we have a bound*

$$\|A^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq \mathbb{P}(\tau \geq n).$$

Lemma 2 (Lower bound) (Wang 2020) *Let* $R(x)$ *denote the rejection probability* (5) *given current state* x. *That is,*

$$R(x) = 1 - \int \min\{H[w(x)], H[w(y)]\} \pi(y) dy.$$

Then, we have a lower bound

$$\|A^n(x, \cdot) - \pi(\cdot)\|_{TV} \geq [R(x)]^n.$$

Theorem 3 *Consider the MTM-IS defined in Algorithm 2 and let* $w^* < \infty$ *be the essential supremum of* $w(x) = \pi(x)/p(x)$. *Then, the maximal total variation distance of the algorithm to its target distribution* π *is*

$$d(n) = [1 - H_k(w^*)]^n.$$

Thus, the exact convergence rate of the MTM-IS is $1 - H_k(w^*)$.

Proof We will establish that upper and lower bounds of $d(n)$ are equal in the limit.

3.2.1 Upper bound

An upper bound can be obtained by using the coupling idea of Lemma 1. Consider two Markov chains $\{x_t\}$ and $\{\tilde{x}_t\}$ defined by MTM-IS. Because of the the decomposition (6), we can interpret the actual transition measure $A(x, \cdot)$ as a mixture of $\pi(\cdot)$ and $q_{\text{res}}(x, \cdot)$, and define the following coupling rule for the two chains. First, we let $x_0 = x$ (for some arbitrary $x \in \mathcal{X}$) and assume that $\tilde{x}_0 \sim \pi(\cdot)$ as the initialization of these two chains. Then, suppose that the two chains are at x_t and $\tilde{x}_t$, respectively, at time t. If $x_t = \tilde{x}_t$, then sample x_{t+1} from $A(x_t, \cdot)$ and set $\widetilde{x}_{t+1} = x_{t+1}$. Thus, their future paths coalesce into one. If $x_t \neq \tilde{x}_t$, we draw $z \sim \text{Bernoulli}(H(w^*))$ and sample $x \sim \pi(\cdot)$. We set $x_{t+1} = \widetilde{x}_{t+1} = x$ if $z = 1$. Otherwise, we sample $x_{t+1} \sim q_{\text{res}}(x_t, \cdot)$ and $\widetilde{x}_{t+1} \sim q_{\text{res}}(\tilde{x}_t, \cdot)$, independently.

Our constructions of $\{x_t\}$ and $\{\tilde{x}_t\}$ have the following properties: (i) marginally these two chains both evolve according to $A(\cdot, \cdot)$; (ii) the distribution of x_t is exactly $A^t(x, \cdot)$ and the distribution of $\tilde{x}_t$ is $\pi(\cdot)$; (iii) once $x_t = \tilde{x}_t$ for some t, the two chains coalesce into one afterwards. Applying Lemma 1, we have

$$\|A^n(x, \cdot) - \pi(\cdot)\|_{TV} \le \mathbb{P}(\tau \ge n) \le [1 - H(w^*)]^n. \quad (14)$$

Taking the supremum over $x \in \mathcal{X}$ we have $d(n) \le [1 - H(w^*)]^n$.

3.2.2 Lower bound

For the lower bound, we consider the worst case as demonstrated in the proof of Lemma 2 in Wang (2020). In particular, if we can find some x^* such that $w(x^*) = w^*$, then the proof is over; but sometimes this is not achievable, in which case we take advantage of the continuity and monotonicity of H_k. For any $\epsilon > 0$, there exists $\delta > 0$ such that $H(w) < H(w^*) + \epsilon$ once $w^* - \delta < w \le w^*$. By the definition of essential supremum, we can always find some $x_\delta \in \mathcal{X}$ such that $w^* - \delta < w(x_\delta) \le w^*$, thus

$$d(n) \ge \|A^n(x_\delta, \cdot) - \pi(\cdot)\|_{TV} \ge R(x_\delta)^n \ge [1 - H(w^*) - \epsilon]^n,$$

since we know from (5) that

$$R(x_\delta) \ge 1 - \int_{\mathcal{X}} H_k[w(x_\delta)]\pi(\mathrm{d}y) \ge 1 - H(w^*) - \epsilon.$$

Letting $\epsilon \to 0$, we derive the final result. $\qquad\square$

3.3 Comparison with the IMH sampler

Since one iteration of MTM-IS is computationally as expensive as k-iterations of the IMH algorithm, we are interested in knowing which one has a better convergence rate. We denote the MTM-IS algorithm with k trials as MTM-IS(k) to emphasize the role of k. Correspondingly, we denote the k-fold thinned IMH algorithm IMH(k) (i.e., collecting 1 draw after every k steps of the standard IMH). Note, however, that a clear advantage of MTM-IS(k) over IMH(k) is that the former is straightforward to parallelise as suggested in Calderhead (2014), which can considerably speed up the algorithm in practice.

Previously, we obtain the exact convergence rate of MTM-IS(k) as $1 - H_k(w^*)$. We rewrite (4) as an expectation form to gain some insights:

$$H_k(z) = k \underbrace{\int \dots \int}_{k-1} \frac{1}{z + \sum_{i=1}^{k-1} w(y_i)} \prod_{i=1}^{k-1} p(y_i) \mathrm{d}y_i$$

$$= \mathbb{E}_p \left[\frac{k}{z + \sum_{i=1}^{k-1} w(X_i)} \right],$$

where $X_1, \dots, X_{k-1}$ are independent samples from $p(\cdot)$. Setting $k = 1$, the formula reduces to $H_1(z) = z^{-1}$, which gives rise to the exact convergence rate $1 - 1/w^*$ of the IMH algorithm as shown in Liu (1996) and Atchadé and Perron (2007). The convergence rate of IMH(k) is then exactly $(1 - 1/w^*)^k$. We have the following main result, whose proof is deferred to the Appendix.

Theorem 4 *With the same notations as in Theorem 3, we have*

$$1 - H_k(w^*) = 1 - \mathbb{E}_p \left[\frac{k}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] \ge \left(1 - \frac{1}{w^*} \right)^k \quad (15)$$

for any $k \ge 1$, where all X_i's are taken independently from $p(\cdot)$. Thus, MTM-IS(k) is no more efficient than IMH(k) although the two algorithms are of similar computational cost.

This theorem provides the first theoretical guidance on the use of MTM methods. It implies that in this rather simple MTM-IS framework, multiple independent proposals are not helpful in improving the the mixing of the algorithm. It is not surprising that IMH is preferable when the target distribution is "easy"—after all, the IMH is perfect if the proposal matches the target exactly and having multiple trials is simply a waste. It is surprising to us, though, that such a preference holds universally.

We speculate that k independent multiple proposals in a general MTM framework are also not more efficient than the corresponding k-fold thinned MCMC algorithm. It therefore casts a doubt on the utility of MTM. Our numerical experiences in the past suggest that the MTM strategy is most helpful in jumping among multiple modes of the target distribution (Liu et al. 2000; Dai and Liu 2020). Also as demonstrated in the molecular simulation literature (Frenkel et al. 1996), a form of *partial* MTM is very useful in building part of the proposal and will be examined in more detail in Sect. 5.1. More general correlated multiple proposals may also help (Craiu and Lemieux 2007) and will be discussed in Sects. 5.2 and 5.4.

4 Numerical illustrations

We illustrate the discrepancy between convergence rates of MTM-IS(k) and IMH(k) numerically. As expected, if the proposal p is already very close to target π, IMH(k) is significantly better than MTM-IS(k). The performance difference
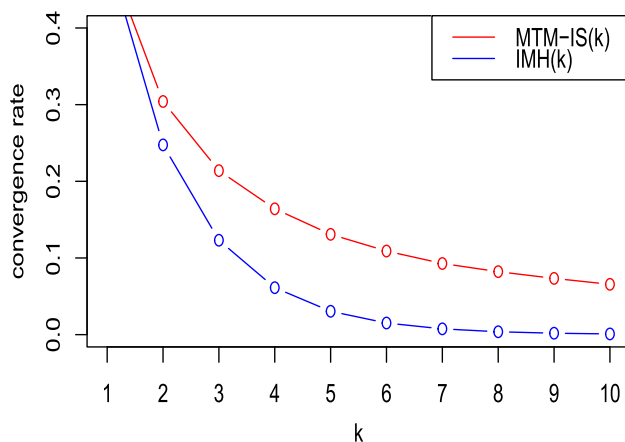
Fig. 1 Convergence rates for example 1 with a finite discrete target distribution

of the two algorithms becomes quite minimal if the proposal distribution differs from the target one considerably, i.e., when w^* is large. In these examples, the explicit convergence rate formula for MTM-IS(k) is still complicated, so we use Monte Carlo to approximate the expectation in (15).

4.1 Univariate examples

The first two examples were previously used in Liu (1996) to compare the IMH algorithm with importance sampling and rejection sampling and are reexamined here. The third example is a continuous case with an unbounded domain.

Example 1 Let the state space be $\mathcal{X} = \{1, \ldots, m\}$, $p(i) = 1/m$ and $\pi(i) = (2m + 1 - 2i)/m^2$, $p(i) = 1/m$. In this case, $w^* = 2 - 1/m$ is close to 2, leading to an approximate convergence rate 0.5 for the IMH algorithm. Figure 1 displays the convergence rates of MTM-IS(k) and IMH(k) with $m = 1000$ and k ranging from 1 to 10 computed from 50,000 independent uniform Monte Carlo samples.

Example 2 We consider the case where the target distribution is binomial Bin(m, θ), and $p(x) = 1/(m + 1)$ is uniform. Then

$$w(x) = (m + 1) \frac{m!}{x!(m - x)!} \theta^x (1 - \theta)^{m-x}.$$

Using the standard normal approximation, we find that

$$w^* \approx \sqrt{\frac{m}{2\pi\theta(1 - \theta)}}.$$

Figure 2 is computed from 50,000 independent uniform Monte Carlo samples with $m = 100$ for two θ values. We that in the latter case when the distribution is very skewed, the discrepancy between MTM-IS(k) and IMH(k) is much smaller.

Example 3 We investigate a one-dimensional continuous case with the target being $\mathcal{N}(0, 1)$, and the proposal distribution being a scaled t-distribution with 10 degrees of freedom, $p(x) = c t_{10}(cx)$ with $c \geq 1$. For practical uses of both importance sampling and IMH-type algorithms, we strongly recommend to choose a proposal distribution that has a heavier tail than but does not differ too much with the target. In our case, both t-distribution proposals satisfy the fat-tail requirement. But a larger c leads to a larger discrepancy between the target and the proposal. Figure 3 is computed based on 50,000 independent Monte Carlo samples with two choices of c, demonstrating that IMH(k) and MTM-IS(k) are nearly indistinguishable if the proposal does not align with the target well.

4.2 Multivariate Gaussian and Gaussian mixture

We first use multivariate Gaussian distributions as both the target and proposal to show some practical implications of our result. Let $\pi = \mathcal{N}(0, \mathbb{I}_d)$ and $T = \mathcal{N}(\vec{\mu}, \sigma^2 \mathbb{I}_d)$. Then we find that the importance weight can be expressed as:

$$w(\vec{x}) = \frac{\pi(\vec{x})}{p(\vec{x})} = \sigma^d$$
$$\exp\left[-\frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right)\|\vec{x}\|^2 - \frac{1}{\sigma^2}\langle\vec{x}, \vec{\mu}\rangle + \frac{1}{2\sigma^2}\|\vec{\mu}\|^2\right].$$

Therefore, $w^* = \sup w(\vec{x}) < \infty$ if either $\sigma > 1$ with an arbitrary $\vec{\mu}$ or $\sigma = 1$ with $\vec{\mu} = 0$. When $\sigma > 1$, the maximal importance weight $w^* \sim \sigma^d$ and thus the mixing time of IMH $\tau_{\mathrm{IMH}}(\delta) = \Omega(w^* \log(1/\delta))$ scales exponentially with the dimension d. In the same manner, the mixing time of MTM-IS also scales exponentially with d, and becomes worse as σ increases. Figure 4 supports that MTM-IS and consecutive IMH have almost the same mixing rates.

Next, we consider a Gaussian mixture distribution $\pi = \frac{1}{3}\mathcal{N}(0, \mathbb{I}_d) + \frac{2}{3}\mathcal{N}(\vec{1}, \mathbb{I}_d)$, where $\vec{1}$ is a d-dimensional vector filled with all 1's. Employing $T = \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$, we have the importance weight

$$w(\vec{x}) = \frac{1}{3}\sigma^d \exp\left[-\frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right)\|\vec{x}\|^2\right]$$
$$+ \frac{2}{3}\sigma^d \exp\left[-\frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right)\|\vec{x}\|^2 - \frac{1}{\sigma^2}\langle\vec{x}, \vec{1}\rangle + \frac{d}{2\sigma^2}\right].$$

It is easy to see that $w^* < \infty$ if and only if $\sigma > 1$. Figure 5 depicts theoretical convergence rates and log mixing times for varying dimension and proposal standard deviation σ. Again the mixing times scale exponentially with dimension d. Unlike the single Gaussian case, however, Fig. 5b shows that the slope of log mixing times is not a monotone function of σ.
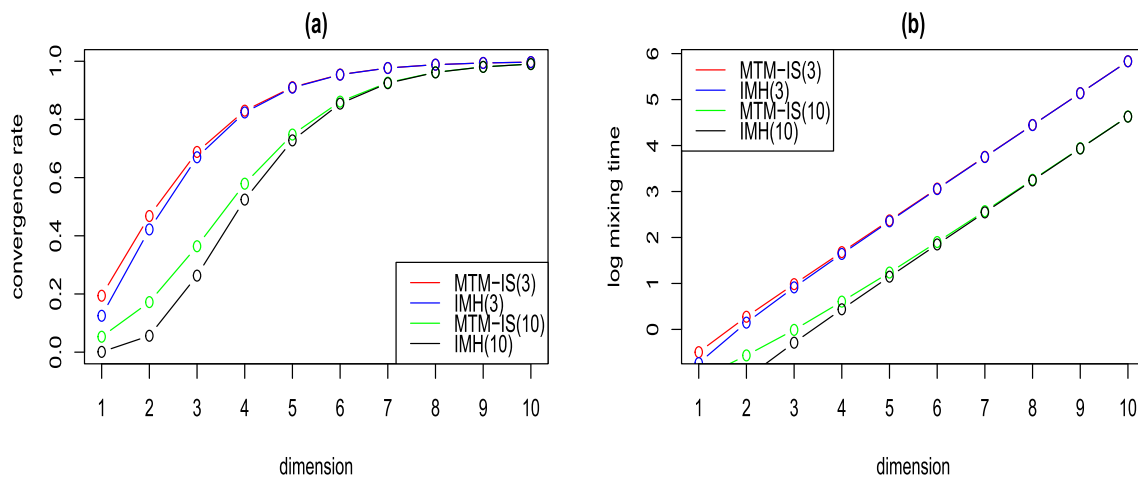
Figure 6 explores the optimization with σ. Specifically, Fig. 6a plots the convergence rates against varying σ when
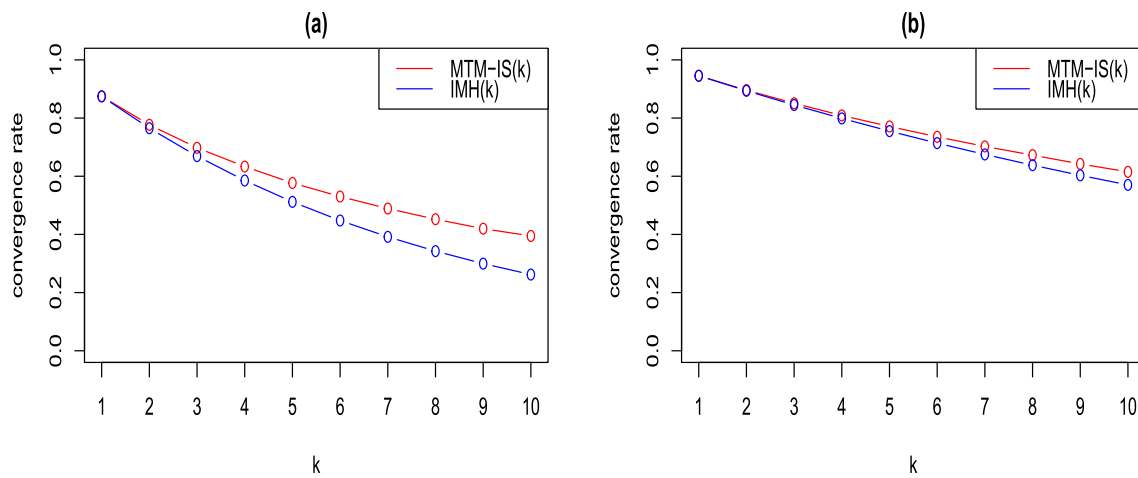
Fig. 2 Convergence rates for a binomial target with **a** $\theta=0.5$, and **b** $\theta=0.05$ (Example 2)
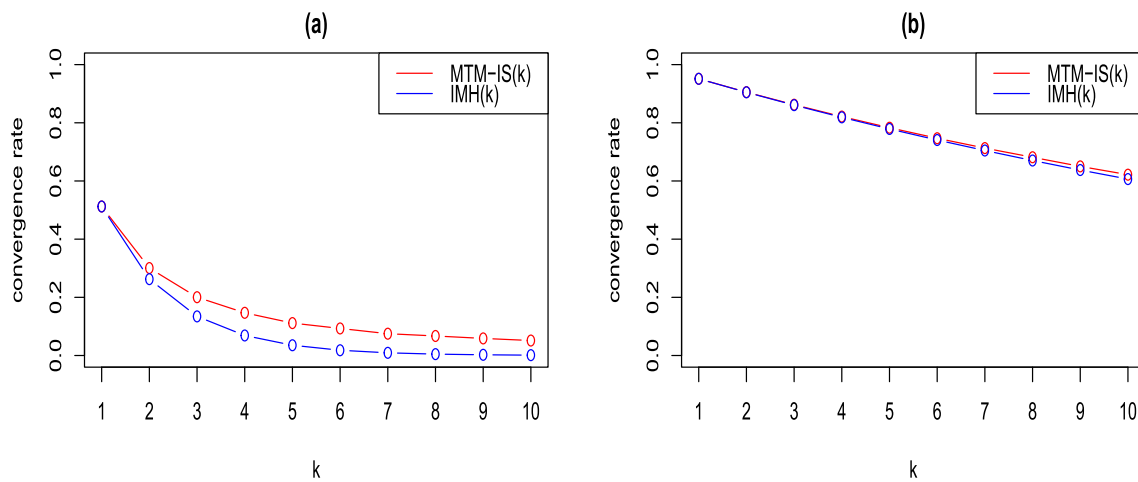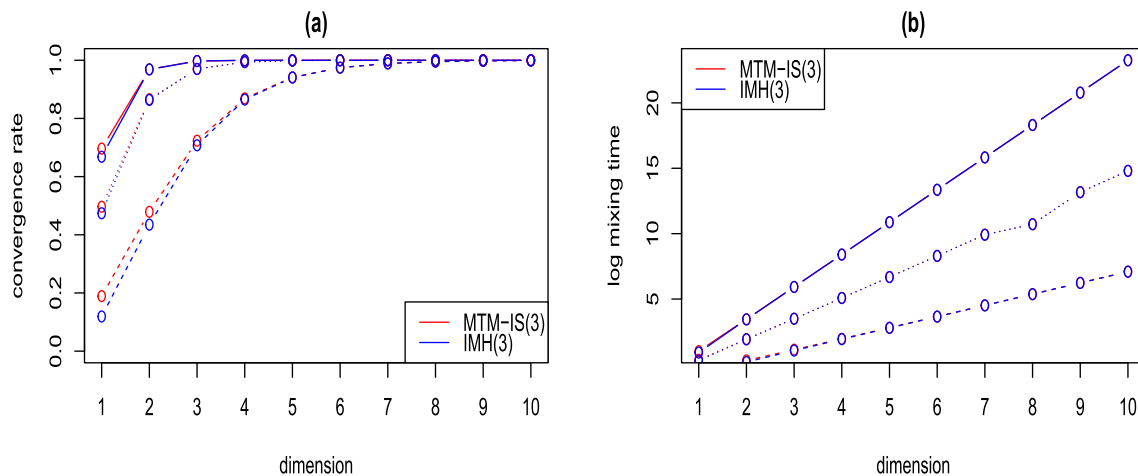

Fig. 3 Convergence rates for a standard normal target (Example 3) with the sampling distribution $p(x)$ being a scaled t-distribution $ct_{10}(cx)$ with **a** $c = 2$, **b** $c = 20$

Fig. 4 Convergence rates (left) and log-mixing times (right) for the standard multivariate Gaussian target $\pi = \mathcal{N}(0, \mathbb{I}_d)$ with proposal $p = \mathcal{N}(0, 4\mathbb{I}_d)$

Fig. 5 Convergence rates (left) and log mixing times (right) for a multivariate Gaussian mixture target $\pi = \frac{1}{3}\mathcal{N}(0, \mathbb{I}_d) + \frac{2}{3}\mathcal{N}(\vec{1}, \mathbb{I}_d)$ with proposal $p = \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$. Solid lines: $\sigma = 1.1$; dashed lines: $\sigma = 2$; dotted lines: $\sigma = 5$. MTM-IS and IMH are nearly indistinguishable

$d = 2$, showing that the optimal choice is $\sigma \approx 1.594949$. When d grows, the optimal σ remains approximately in the range of 1.55–1.62. Figure 6c indicates that the mixing time still scales exponentially with d even if σ is optimized.

5 Variants of multiple-try Metropolis

5.1 Partial MTM-IS: an efficient variant

To reflect how MTM has actually been used in molecular simulations (Frenkel et al. 1996), we assume a partition of the state-space, $\mathbf{x} = (\mathbf{x}^a, \mathbf{x}^b)$, and the corresponding partition of the target distribution $\pi(\mathbf{x}) \propto q(\mathbf{x}^a, \mathbf{x}^b) = q_a(\mathbf{x}^a) q_b(\mathbf{x}^b | \mathbf{x}^a)$, where q_b may not be normalized. We assume that $q_a(\mathbf{x}^a)$ is much more expensive to evaluate than $q_b(\mathbf{x}^b | \mathbf{x}^a)$. An important point to note is that we want to move $(\mathbf{x}^a, \mathbf{x}^b)$ jointly instead of iterating between conditional draws of $\mathbf{x}_a | \mathbf{x}_b$ and $\mathbf{x}_b | \mathbf{x}_a$ (for reasons such as the two components may be tightly coupled). We consider the independent proposal: $p(\mathbf{x}) = p_a(\mathbf{x}^a) p_b(\mathbf{x}^b | \mathbf{x}^a)$. A *Partial MTM-IS* algorithm is as follows:

Remark 1 (PMTM-IS versus MTM-IS) Note that, compared with the vanilla MTM-IS (Algorithm 2), PMTM-IS needs to draw extra balancing samples. Since we assume that sampling $\mathbf{x}^b$ and evaluating it are both very cheap, it is still worth doing. In this case, there are no standard IMH or MCMC variants for comparisons.

Typically, one iteration of IMH involves evaluating q_a / p_a twice (respectively on $\mathbf{x}^a$ and $\mathbf{y}^a$) and evaluating q_b / p_b twice (respectively on $\mathbf{x}^b | \mathbf{x}^a$ and $\mathbf{y}^b | \mathbf{y}^a$). In contrast, one iteration of Algorithm 3 consists of evaluating q_a / p_a twice (respectively on $\mathbf{x}^a$ and $\mathbf{y}^a$) and evaluating q_b / p_b for $2k$ times (respectively on $\mathbf{x}_j^b | \mathbf{x}^a$ and $\mathbf{y}_j^b | \mathbf{y}^a$ with $j = 1, \ldots, k$). When evaluating

---

Algorithm 3 PMTM-IS: the current state is $\mathbf{x} = (\mathbf{x}^a, \mathbf{x}^b)$.

1: Draw $\mathbf{y}^a$ from $p_a(\cdot)$; and draw multiple trials $\mathbf{y}_1^b, \ldots, \mathbf{y}_k^b$ independently from $p_b(\cdot \mid \mathbf{y}^a)$;
2: Draw i.i.d. "balancing trials" $\mathbf{x}_1^b, \ldots, \mathbf{x}_{k-1}^b$ from $p_b(\cdot | \mathbf{x}^a)$, and let $\mathbf{x}_k^b = \mathbf{x}^b$;
3: For $j = 1, \ldots, k$, compute

$$w_j \triangleq \frac{q_b(\mathbf{y}_j^b \mid \mathbf{y}^a) q_a(\mathbf{y}^a)}{p_b(\mathbf{y}_j^b \mid \mathbf{y}^a) p_a(\mathbf{y}^a)}, \quad w_j' = \frac{q_b(\mathbf{x}_j^b \mid \mathbf{x}^a) q_a(\mathbf{x}^a)}{p_b(\mathbf{x}_j^b \mid \mathbf{x}^a) p_a(\mathbf{x}^a)},$$

and set $W_y = \sum_{j=1}^k w_j$, $W_x = \sum_{j=1}^k w_j'$.
4: Select index J with probability proportional to w_j and define $\mathbf{y} = (\mathbf{y}^a, \mathbf{y}_J^b)$.
5: Accept y with probability $\rho = \min\left\{1, W_y / W_x\right\}$.

---

q_b is significantly computationally more expensive than q_a, Algorithm 3 nearly matches the computational cost of one-step IMH. Under certain reasonable regularity conditions, the following proposition shows that Algorithm 3 provably converges faster.

Proposition 1 *Let* $\mathbf{x} = (\mathbf{x}^a, \mathbf{x}^b)$, *and* $\pi(\mathbf{x}) = \pi_a(\mathbf{x}^a) \pi_b(\mathbf{x}^b | \mathbf{x}^a) \propto q_a(\mathbf{x}_a) q_b$
$(\mathbf{x}^b | \mathbf{x}^a)$, *where* π_a *and* π_b *are normalized marginal and conditional distributions. Under the following regularity conditions with proposal* p *(all parts normalized):*

$$ess \sup_{\mathbf{x}^a, \mathbf{x}^b} \frac{\pi(\mathbf{x}^a, \mathbf{x}^b)}{p(\mathbf{x}^a, \mathbf{x}^b)} = w^* < \infty, \tag{16}$$

IMH converges with rate $1 - 1/w^*$. *In contrast, the partial MTM-IS (Algorithm 3) has a convergence rate no slower than* $1 - 1/w^*$.
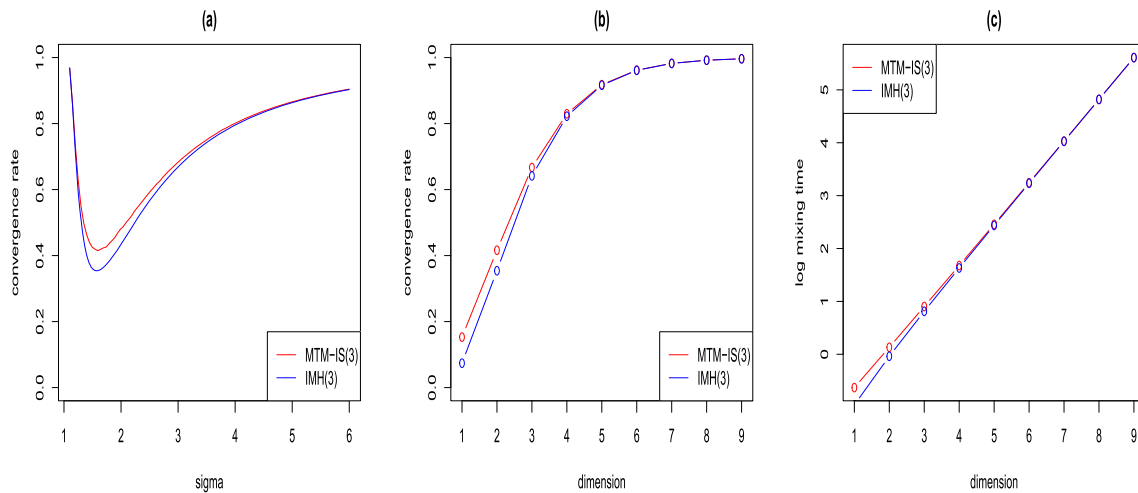
Fig. 6 Multidimensional mixture Gaussian: $\pi = \frac{1}{3}\mathcal{N}(0, \mathbb{I}_d) + \frac{2}{3}\mathcal{N}(\vec{1}, \mathbb{I}_d)$ and $T = \mathcal{N}(0, \sigma^2\mathbb{I}_d)$. **a** Convergence rates against σ with varying $1.1 \leq \sigma \leq 6$ when $d=2$; **b** and **c** Plot respectively the convergence rates and log mixing times against the varying dimensions under the optimized σ

Proof Noting that ess $\sup_\mathbf{x} \frac{\pi(\mathbf{x}^a, \mathbf{x}^b)}{p(\mathbf{x}^a, \mathbf{x}^b)} = w^*$, we obtain the convergence rate of IMH as $1 - 1/w^*$ by Theorem 3. As for Algorithm 3, we decompose the transition kernel as

$$
A((\mathbf{x}^a, \mathbf{x}^b), (\mathbf{y}^a, \mathbf{y}^b))
$$

$$
= k\mathbb{P}\Big[\big\{(\mathbf{y}^a \text{ gets proposed}) \cap (\mathbf{y}_k^b = \mathbf{y}^b)
$$

$$
\cap (J = k) \cap (\text{joint } (\mathbf{y}^a, \mathbf{y}^b) \text{ gets accepted})\big\}\Big]
$$

$$
= k \underbrace{\int_{\mathcal{X}^b} \cdots \int_{\mathcal{X}^b}}_{k-1} \underbrace{\int_{\mathcal{X}^b} \cdots \int_{\mathcal{X}^b}}_{k-1} \frac{p_a(\mathbf{y}^a) p_b(\mathbf{y}^b|\mathbf{y}^a) w_k}{\sum_{j=1}^k w_j}
$$

$$
\min\left\{1, \frac{\sum_{j=1}^k w_j}{\sum_{j=1}^k w_j'}\right\}
$$

$$
\prod_{j=1}^{k-1} p_b(\mathbf{y}_j^b|\mathbf{y}^a) p_b(\mathbf{x}_j^b|\mathbf{x}^a) \mathrm{d}\mathbf{y}_j^b \mathrm{d}\mathbf{x}_j^b.
$$

Suppose the normalizing constant of $q(\mathbf{x}^a, \mathbf{x}^b)$ is C, i.e., $\pi(\mathbf{x}^a, \mathbf{x}^b) = q(\mathbf{x}^a, \mathbf{x}^b)/C$. Then,

$$
\frac{p_a(\mathbf{y}^a) p_b(\mathbf{y}^b|\mathbf{y}^a) w_k}{\sum_{j=1}^k w_j} \min\left\{1, \frac{\sum_{j=1}^k w_j}{\sum_{j=1}^k w_j'}\right\}
$$

$$
= \frac{q(\mathbf{y}^a, \mathbf{y}^b)}{\max\left\{\sum_{j=1}^k w_j, \sum_{j=1}^k w_j'\right\}}
$$

$$
= \frac{q(\mathbf{y}^a, \mathbf{y}^b)/C}{\max\left\{\sum_{j=1}^k \frac{q(\mathbf{y}^a, \mathbf{y}_j^b)/C}{p(\mathbf{y}^a, \mathbf{y}_j^b)}, \sum_{j=1}^k \frac{q(\mathbf{x}^a, \mathbf{x}_j^b)/C}{p(\mathbf{x}^a, \mathbf{x}_j^b)}\right\}}
$$

$$
= \frac{\pi(\mathbf{y}^a, \mathbf{y}^b)}{\max\left\{\sum_{j=1}^k \frac{\pi(\mathbf{y}^a, \mathbf{y}_j^b)}{p(\mathbf{y}^a, \mathbf{y}_j^b)}, \sum_{j=1}^k \frac{\pi(\mathbf{x}^a, \mathbf{x}_j^b)}{p(\mathbf{x}^a, \mathbf{x}_j^b)}\right\}},
$$

in which $\mathbf{y}_k^b = \mathbf{y}^b$ and $\mathbf{x}_k^b = \mathbf{x}^b$. Therefore, it gives rise to

$$
A((\mathbf{x}^a, \mathbf{x}^b), (\mathbf{y}^a, \mathbf{y}^b))
$$

$$
= k\pi(\mathbf{y}^a, \mathbf{y}^b) \underbrace{\int_{\mathcal{X}^b} \cdots \int_{\mathcal{X}^b}}_{k-1} \underbrace{\int_{\mathcal{X}^b} \cdots \int_{\mathcal{X}^b}}_{k-1}
$$

$$
\frac{\prod_{j=1}^{k-1} p_b(\mathbf{y}_j^b|\mathbf{y}^a) p_b(\mathbf{x}_j^b|\mathbf{x}^a) \mathrm{d}\mathbf{y}_j^b \mathrm{d}\mathbf{x}_j^b}{\max\left\{W(\mathbf{y}^a; \mathbf{y}_{1:k-1}^b, \mathbf{y}^b), W(\mathbf{x}^a; \mathbf{x}_{1:k-1}^b, \mathbf{x}^b)\right\}},
$$

where $W(\mathbf{x}^a; \mathbf{x}_{1:k}^b) \triangleq \sum_{j=1}^k \frac{\pi_b(\mathbf{x}^a, \mathbf{x}_j^b)}{p_b(\mathbf{x}^a, \mathbf{x}_j^b)}$ for any $\mathbf{x}^a \in \mathcal{X}^a, \mathbf{x}_{1:k}^b = (\mathbf{x}_1^b, \ldots, \mathbf{x}_k^b) \in (\mathcal{X}^b)^k$. By definition, we find

$$
W(\mathbf{x}^a; \mathbf{x}_{1:k}^b) = \sum_{j=1}^k \frac{\pi_b(\mathbf{x}^a, \mathbf{x}_j^b)}{p_b(\mathbf{x}^a, \mathbf{x}_j^b)} \leq kw^*.
$$

The following inequality immediately follows:

$$
A((\mathbf{x}^a, \mathbf{x}^b), (\mathbf{y}^a, \mathbf{y}^b)) \geq \frac{\pi(\mathbf{y}^a, \mathbf{y}^b)}{w^*}. \tag{17}
$$

Surprisingly, (3) leads to a mixture decomposition like (17) and thus is sufficient to construct the upper bound in Theorem 3 by the coupling argument and Lemma 1. Therefore, the convergence rate of Algorithm 3 is no larger than $1 - 1/w^*$. However, the arguments for establishing matching lower bounds cannot directly apply due to the extra balancing trials $\mathbf{x}_j^b, 1 \leq j \leq k - 1$. So the exact convergence rate of Algorithm 3 remains unknown. $\qquad\square$

5.2 Correlated multiple trials

Compared with the original MTM, the partial MTM-IS differs in that its multiple trials $(\mathbf{y}^a, \mathbf{y}_1^b), \ldots, (\mathbf{y}^a, \mathbf{y}_k^b)$ are correlated due to the state space partitioning. As also demonstrated by Craiu and Lemieux (2007), we believe that generating correlated multiple trials is a key in designing efficient MTM algorithms. Although rigorous theoretical analysis for a general correlated MTM design is beyond our reach, we present some theoretical results for two special cases for finite state spaces, which may also be generalization to continuous state-spaces. Implications derived from the analysis apply more generally: good correlated multiple-tries can be obtained with the aid of a deterministic step.

5.2.1 Stratified sampling

Suppose $\mathcal{X}$ is a finite state space. We partition it into a few subgroups, $\mathcal{X}_1, \ldots, \mathcal{X}_B$ so that $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset, \forall i \neq j$ and $\cup_j \mathcal{X}_j = \mathcal{X}$. We begin with a block wise IMH step by sampling from $\{\mathcal{X}_1, \ldots, \mathcal{X}_B\}$ with weight $p(\mathcal{X}_j)$ and accept it with $w(\mathcal{X}_j) = \pi(\mathcal{X}_j)/p(\mathcal{X}_j)$ afterwards. Then, we draw y within the sampled block with probability proportional to $\pi(y)$. It is easy to see that the chain become stationary once it converges at the subgroup level. Thus, the convergence rate of this algorithm is

$$r_B = 1 - w(\mathcal{X}^*)^{-1},$$

where $\mathcal{X}^* = \arg\max_j w(\mathcal{X}_j)$. This is not generally better than the IMH(k), which has a convergence rate of $(1 - 1/w^*)^k > 1 - k/w^*$. But if the weights w's are very uneven and we can partition the states so that the weights $w(\mathcal{X}_j)$'s are more balanced, then the stratified IMH can improve upon IMH(k) significantly. We also note that the computation cost of this block-based MTM-IS(k) algorithm is no worse than IMH(2) (the first step of block sampling is no worse than 1-step IMH; and so is the second step of sampling within a block), much better than IMH(k) when k is large.

Example 4 (Example 1 continued) Let $\mathcal{X} = \{1, \ldots, N\}$, and suppose that the target $\pi(x) \propto x$ and $p(x) \propto 1$. Then, the original weights are $w(x) \propto x$ and $w^* = \frac{2N}{1+N} \approx 2$. Let $k = 2$, then IMH(2) has a rate of $(1 - 1/w^*)^2 \approx 0.25$, which is quite good. Assume that N is an even number and we partition the space as $\mathcal{X}_j = \{j, N - j + 1\}$ for $j = 1, \ldots, N/2$. Then $w(\mathcal{X}_j) \propto 1$, and the resulting MTM-IS(2) converges in one step. More generally, for an arbitrary distribution $\pi(x)$ and the uniform proposal $p(x) = (2N)^{-1}$, we have $w^* = \pi(x^*)$ with $x^* = \arg\max_x \pi(x)$. Thus, if we can partition the state space so that $\pi(\mathcal{X}_j)$ are approximately equal for $j = 1, \ldots, B$, the algorithm can be much improved.

5.2.2 Sampling without replacement

Another obvious way of introducing correlations for multiple proposals is to do sampling without replacement. Let $\mathcal{X} = \{1, \ldots, N\}$. To simplify the discussion, we here focus on the simple random sampling without replacement (SRSWOR, i.e., $p(\mathbf{y}) \propto 1$), although it is possible to extend the method to do sampling without replace with unequal probabilities using one of the schemes in Chen et al. (1994). The algorithm is as follows.

Algorithm 4 MTM-SRSWOR(k): Suppose that the current state is at x.

1: Draw $S = (y_1, \ldots, y_k) \subset (\mathcal{X} \backslash \{x\})$ jointly via SRSWOR.
2: Select index J with probability proportional to $w(y_j) = (N - 1)\pi(y_j)$ and define $y = y_J$.
3: Accept y with the ratio $\rho = \min\left\{1, \frac{w(y,x)+\sum_{i \neq J} w(y_i,x)}{w(x,y)+\sum_{i \neq J} w_i(y_i,y)}\right\} = \min\left\{1, \frac{\pi(y)+\sum_{j \neq J} \pi(y_j)}{\pi(x)+\sum_{j \neq J} \pi(y_j)}\right\}$.

The actual transition probability from x to $y \neq x$ for this scheme is

$$A(x, y) = \sum_{S_y^{(k-1)}} \frac{1}{\binom{N-1}{k}} \pi(y)$$

$$\min\left[\frac{1}{\pi(y) + \sum_{i<k} \pi(y_i)}, \frac{1}{\pi(x) + \sum_{i<k} \pi(y_i)}\right], \tag{18}$$

where $S_y^{(k-1)} \subset \mathcal{X} \backslash \{x, y\}$, $|S_y^{(k-1)}| = k - 1$, and $y_j \in S_y^{(k-1)}, \forall j < k$. Doing an exact eigenvalue decomposition of matrix A would have brought us a tight bound on the convergence rate. But A does not possess a nice low-rank property as that for the IMH sampler or the MTM-IS.

For $S \subset \mathcal{X}$, we define $\pi(S) = \sum_{x \in \mathcal{X}} \pi(x)$, $S^* = \arg\max_{\{S: |S|=k\}} \pi(S)$, and $x^* = \arg\max_x \pi(x)$. We find the following inequality to hold:

$$A(x, y) \geq \frac{k\pi(y)}{(N - 1)\pi(S^*)}, \quad x \neq y.$$

During each iteration, the chain stays at the current state if and only if the new proposal is rejected since in our construction of Algorithm 4, the proposal set is not allowed to contain the current state. We observe that $\rho \equiv 1$ whenever $x = x_* \triangleq \arg\min_x \pi(x)$, leading to $A(x_*, x_*) = 0$. This fact prevents us from using the previous coupling arguments directly. However, as we specify to some circumstances, we could still obtain satisfactory results.

Example 5 Choosing $k = 2$ and $\mathcal{X} = \{1, \ldots, N\}$, we set

$$\pi_1 = 1 - p, \quad \pi_2 = \cdots = \pi_N = \frac{p}{N-1}, \tag{19}$$

where $0 \le p \le (N-1)/N$, which guarantees that $x^* = 1$ and $\{2, \ldots, N\} \in \arg\min_x \pi(x)$. As a result, we know that $A(2, 2) = \cdots = A(N, N) = 0$. Furthermore, matrix A can be completely determined by the following four quantities:

$$a_1 = A(1, 2) = \frac{2\pi_2}{(N-1)(\pi_1 + \pi_2)},$$
$$a_2 = A(1, 1) = \frac{\pi_1 - \pi_2}{\pi_1 + \pi_2},$$
$$a_3 = A(2, 1) = \frac{2\pi_1}{(N-1)(\pi_1 + \pi_2)},$$
$$a_4 = A(2, 3) = \frac{(N-3)}{(N-1)(N-2)}$$
$$+ \frac{2\pi_2}{(N-1)(N-2)(\pi_1 + \pi_2)}.$$

We can then write out A as follows:

$$A = \begin{bmatrix} a_2 & a_1 & a_1 & a_1 & \ldots & a_1 \\ a_3 & 0 & a_4 & a_4 & \ldots & a_4 \\ a_3 & a_4 & 0 & a_4 & \ldots & a_4 \\ a_3 & a_4 & a_4 & 0 & \ldots & a_4 \\ \ldots & & & & & \ldots \\ a_3 & a_4 & a_4 & a_4 & \ldots & 0 \end{bmatrix}. \tag{20}$$

Now this matrix admits a useful low-rank decoupling: $A = G + ep^T$, where $e = [1, \ldots, 1]^T$, $p = [a_3, a_4, \ldots, a_4]^T$ and

$$G = \begin{bmatrix} a_2 - a_3 & a_1 - a_4 & a_1 - a_4 & a_1 - a_4 & \ldots & a_1 - a_4 \\ 0 & -a_4 & 0 & 0 & \ldots & 0 \\ 0 & 0 & -a_4 & 0 & \ldots & 0 \\ 0 & 0 & 0 & -a_4 & \ldots & 0 \\ \ldots & & & & & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -a_4 \end{bmatrix}. \tag{21}$$

Note that e is a common right eigenvector for both A and $A - G$, corresponding to the largest eigenvalue 1. Since $A - G$ is of rank 1, the remaining eigenvalues of A and G have to be the same. Hence the eigenvalues for A are $1, a_2 - a_3, -a_4, \ldots, -a_4$. This decoupling trick has also been used in Liu (1996) for the IMH algorithm. Given the convergence rate $(1 - 1/(N\pi_1))^2$ of IMH(2), it suffices to show

$$|a_2 - a_3| \le (1 - 1/(N\pi_1))^2, \quad a_4 \le (1 - 1/(N\pi_1))^2, \tag{22}$$

to prove that MTM-SRSWOR(2) is faster than IMH(2). Clearly, this holds true for $p = \dfrac{N-1}{2N}$, which leads to

$\pi_1 = 1/2 + 1/(2N)$, $\pi_2 = 1/(2N)$. In this case,

$$a_2 - a_3 = 1 - \frac{4N}{(N+2)(N-1)} < 1 - \frac{4}{N+1}$$
$$< \left(1 - \frac{2}{N+1}\right)^2 = \left(1 - \frac{1}{N\pi_1}\right)^2,$$
$$a_4 = \frac{(N-3)}{(N-1)(N-2)} + \frac{2}{(N-1)(N-2)(N+2)}$$
$$< \left(1 - \frac{1}{N\pi_1}\right)^2.$$

We note that designing a suitable parallel construction to do SRSWOR can speed up the algorithm considerably. Furthermore, when proposing multiple trials, we may also choose not to exclude x from the proposal set. In this case, we need to modify Algorithm 4 slightly to become Algorithm 5.

Algorithm 5 MTM-SRSWOR-II(k): the current state is x.

1: Draw a subset $S \subset \mathcal{X}$ of size k at random, denoted as $S = (y_1, \ldots, y_k)$.
2: Select index J with probability proportional to $w(y_j) = N\pi(y_j)$ and define $y = y_J$.
3: If $x \notin S$, accept y with probability $\rho = \min\left\{1, \frac{w(y) + \sum_{j \ne J} w(y_j)}{w(x) + \sum_{j \ne J} w(y_j)}\right\}$. If $x \in S$, accept y with probability 1.

5.3 Independent non-identical proposals

Besides introducing correlations between multiple trials, Craiu and Lemieux (2007) also suggests to use different proposals for generating multiple trials in each MTM iteration and provides some supportive empirical evidences. Here we consider a special case of MTM-IS(k) in which the multiple trials are generated from different proposals, i.e., $y_j \sim p_j(\cdot)$ independently for $j = 1, \ldots, k$. In this case, we also do not have to draw balancing trials. Defining $w_j(x) := \pi(x)/p_j(x)$, we summarize the procedure in Algorithm 6.

Algorithm 6 MTM with independent non-identical proposals with current state x.

1: Draw multiple trials $y_j \sim p_j(y_j)$, $j = 1, \ldots, k$ independently. Then compute $w_j(y_j) = \pi(y_j)/p_j(y_j)$.
2: Select index J with probability proportional to $w_j(y_j, x)$ and define $y = y_J$.
3: Accept y with the ratio $\rho = \min\left\{1, \frac{w_J(y) + \sum_{i \ne J} w_i(y_i)}{w_J(x) + \sum_{i \ne J} w_i(y_i)}\right\}$.

To demonstrate the effect of the *multiple-try* design employed in Algorithm 6, it should be compared with a *sequential k-step IMH* sampler. During one iteration, this

sampler runs an interior loop of length k, within which the j-th step proposes an independent proposal from p_j and then accepts/rejects it based on the MH rule as in the ordinary IMH sampler. This sequential IMH sampler has the same computational cost as Algorithm 6. The following theorem provides tight upper bounds for the convergence rates of the two algorithms, and its proof is deferred to appendixes.

Theorem 5 *Suppose target π is absolutely continuous with respect to every proposal p_j. Algorithm 6 and its corresponding sequential IMH sampler are geometrically convergent, with their corresponding respective convergent rates upper bounded by $1 - \sum_{j=1}^{k} \mathbb{E}_p \left[\frac{1}{w_j^* + \sum_{1 \le i \le k, i \ne j} w_i(X_i)} \right]$ and $\prod_{j=1}^{k} \left(1 - \frac{1}{w_j^*} \right)$, respectively, where $w_j^* := \sup_{x \in \mathcal{X}} w_j(x)$. Furthermore, the following inequality holds,*

$$1 - \sum_{j=1}^{k} \mathbb{E}_p \left[\frac{1}{w_j^* + \sum_{1 \le i \le k, i \ne j} w_i(X_i)} \right] \ge \prod_{i=1}^{k} \left(1 - \frac{1}{w_i^*} \right), \tag{23}$$

implying that the upper bound for Algorithm 6 is worse than that for the corresponding sequential IMH.

Remark 2 (Tightness of the lower bounds) Suppose $\exists x^*$ such that

$$w_j(x^*) = w_j^* := \sup_{x \in \mathcal{X}} w_j(x) = \sup_{x \in \mathcal{X}} \pi(x)/p_j(x) < \infty, \text{ for all } j, \tag{24}$$

i.e., different proposals have their importance weight functions w_j to attain their respective supremums at a same point x^*. Then, the convergence rates for both aforementioned algorithms attain their respective upper bounds. When $p_1 = \cdots = p_k$, condition (24) automatically holds, recovering the convergence rate result of Theorem 3. However, when there is no such a x^* as required by (24), the quantities claimed in Theorem 5 are only upper bounds. It remains unknown under what other conditions one algorithm can be provably better than the other. Our empirical study shows that their computational efficiencies are almost indistinguishable when the target distribution is "hard" relative to the proposals.

Example 6 We conducted a few simulations to examine convergence behaviors of Algorithm 6 and the corresponding sequential IMH sampler at the same computational cost.

As shown in Fig. 7, we considered target densities of the form of a mixture of two standard distributions with various dimensions. Top plots in Fig. 7 correspond to Gaussian mixture targets, $\pi = \frac{1}{2} \mathcal{N}(0, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(3, \mathbb{I}_d)$, with d=3, 4, and 5, respectively. Two different proposal distributions are

employed: $p_1 = \mathcal{N}(0, \mathbb{I}_d)$ and $p_2 = \mathcal{N}(0, 9\mathbb{I}_d)$. During one iteration of the MTM-IS(k) algorithm, $k/2$ trials are independently drawn from of p_1, and another $k/2$ trials from p_2. The bottom plots correspond to t-mixture distributions, $\pi = \frac{1}{2} t_3(0) + \frac{1}{2} t_3(\mathbf{4})$, for d=1, 2, and 3. Two different proposal distributions are: $p_1 = t_3(0)$ and $p_2 = t_5(0)$, and the same implementation of MTM-IS(k) as the previous case is employed. These plots show that Algorithm 6 and its corresponding *sequential IMH* sampler differ very little in their convergence rates although theoretically we cannot claim one is necessarily better than the other without condition (24). All simulations are based on 10^6 iterations on an Apple M2 chip with 16GB memory, each taking a few minutes.

5.4 A general framework

Inspired by the variants of MTM just discussed, we propose a general framework to combine these variants in Algorithm 7. With $\pi(\cdot)$ as the target distribution on $\mathcal{X}$, we let $p(x, \mathbf{y})$ denote the proposal transition function for multiple correlated proposals, where $x \in \mathcal{X}$ and $\mathbf{y} = (y_1, \dots, y_k) \in \mathcal{X}^k$. We further write the j-th marginal of $p(x, \mathbf{y})$ as $p_j(x, y_j) = \int p(x, \mathbf{y}) d\mathbf{y}_{(-j)}$, and define the jth lab:ssk08 as

$$w_j(y \mid x) = \frac{\pi(y)}{p_j(x, y)} \lambda_j(x, y), \tag{25}$$

for $j = 1, \dots, k$, where λ_j is a symmetric function. Assuming the current state is x, the updating rule is summarized in Algorithm 7.


Algorithm 7 Generalized MTM. Suppose current state is at x.

1: Draw multiple trials $y_1, \dots, y_k$ jointly from $p(x, \mathbf{y})$. Then compute $w_j(y_j \mid x)$ by (25) for $j = 1, \dots, k$.
2: Select index J with probability proportional to $w_j(y_j, x)$ and define $y = y_J$.
3: Draw $x_1^*, x_2^*, \dots, x_{J-1}^*, x_{J+1}^*, \dots, x_k^*$ from the conditional distribution of $p(y, \mathbf{x}^*)$ conditioned on J-th variable as x. And set $x_J^* = x$.
4: Accept y with the ratio $\rho = \min \left\{ 1, \frac{w_J(y \mid x) + \sum_{i \ne J} w_i(y_i \mid x)}{w_J(x \mid y) + \sum_{i \ne J} w_i(x_i^* \mid y)} \right\}$.

If we require $p(x, \mathbf{y})$ to have the same marginals for different y_j's, the algorithm reduces to that of Craiu and Lemieux (2007); if we require $p(x; \mathbf{y})$ to be independent among the y_j's, it reduces to that of Casarin et al. (2013). Note that the balancing proposals are drawn to facilitate the computation of ρ, and this guarantees the detailed balance of the MTM design. The following result is expected and its detailed proof is deferred to the Appendix.

Theorem 6 *The generalized MTM transition rule (Algorithm 7) satisfies the detailed balance condition and hence*
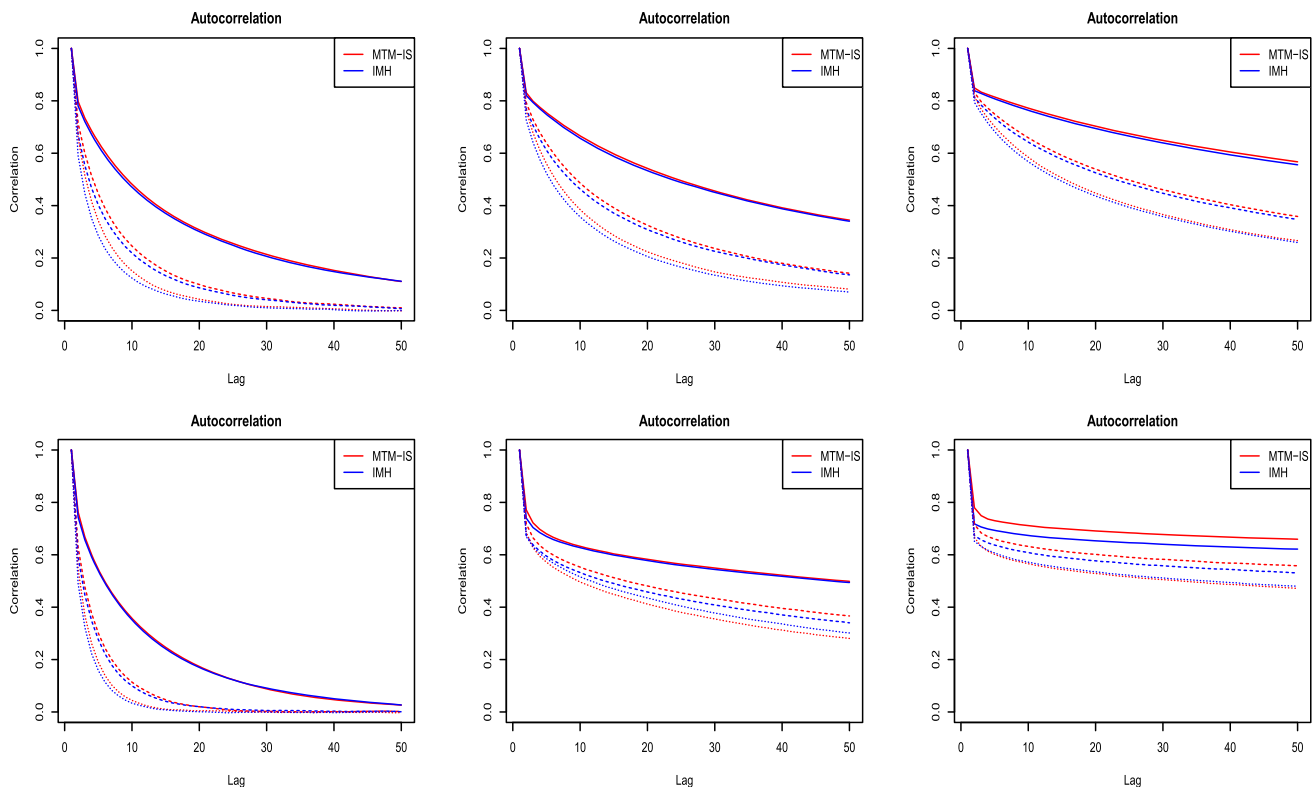
Fig. 7 **Top:** Auto-correlation plots for the Gaussian mixture targets in Example 6 From left to right: dimension $d= 3, 4, 5$, respectively. **Bottom:** Auto-correlation plot for the t-mixture targets in Example 6 From left to right: dimension $d = 1, 2, 3$, respectively. Solid lines: $k = 2$; dashed lines: $k = 6$; dotted lines: $k = 10$

induces a reversible Markov chain with π as its invariant distribution.

Defining $\mathbf{x}^*(j) \triangleq (x_1^*, \ldots, x_{j-1}^*, x, x_{j+1}^*, x_k^*)$, one can determine the transition density of the generalized MTM framework via the same spirit employed in the proof of Theorem 1:

$$A(x, y) = \pi(y) \sum_{j=1}^{k} \Bigg[p_j(x, y) p_j(y, x) \lambda_j(x, y) \times$$

$$\int u_j(\mathbf{x}^*(j), \mathbf{y}) p(x, \mathbf{y}_{(-j)} \mid y_j = y)$$

$$p(y, \mathbf{x}_{(-j)}^* \mid x_j^* = x) \prod_{i \neq j} \mathrm{d} y_i \mathrm{d} x_i^* \Bigg],$$

where we write

$$u_j(\mathbf{x}, \mathbf{y}) \triangleq \min \left\{ \left(\sum_{i=1}^{k} w_i(y_i \mid x_j) \right)^{-1}, \left(\sum_{i=1}^{k} w_i(x_i \mid y_j) \right)^{-1} \right\}$$

for any $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{y} = (y_1, \ldots, y_k)$. A detailed derivation of this formula can be found in the proof of Theorem 6.

As demonstrated in Algorithms 4, 5 and 6, we find that sometimes we do not need to draw balancing trials for MTM to retain the detailed balance. A natural question then arises: can we find a general condition under which which MTM can avoid the drawing of balancing trials? The following theorem provides a sufficient condition that covers all the cases we discussed.

Theorem 7 *If, for any pair (x, y) and $\forall j$, the joint proposal distribution satisfies*

$$p(x, \mathbf{y}_{(-j)} \mid y_j = y) = p(y, \mathbf{y}_{(-j)} \mid y_j = x), \qquad (26)$$

we can maintain the detailed balance by setting $x_j^* \triangleq y_j$ for $j \neq J$ in Algorithm 7.

Remark 3 (Correlated multiple trials) As demonstrated in Sects. 5.1 and 5.2, letting the proposed multiple trials be correlated (especially negatively) can be helpful in improving the chain's convergence. A useful strategy is to use multiple trials as stepping stones to move from one mode of the distribution to another, similar in spirit to Hamiltonian/hybrid Monte Carlo (Qin and Liu 2001; Liu 2008) and the griddy Gibbs MTM (Liu et al. 2000). Indeed, it was shown empirically in Qin and Liu (2001) that applying MTM–HMC

trajectories may further improve the sampling efficiency. However, an in-depth theoretical analysis as carried out here is much more challenging due to the semi-deterministic nature of aforementioned algorithms.

Remark 4 (Employing multiple distributions in MTM) Intuitively, one may hope that using different distributions for each trial could help us explore the state space better. Our results in Sect. 5.3, however, demonstrate that it is still not very useful under the IMH framework if the multiple trials are independent. It may be helpful for the partial MTM framework discussed in Sect. 5.1.

6 Concluding remarks

We have presented a complete eigen-decomposition and convergence rate analysis for the MTM-IS, and compared it with the "thinned" IMH sampler (of the same computational cost). With the exact form of eigenvalues of the MTM-IS, we proved rigorously that the sampler is not as efficient as the simpler "thinned" IMH approach. To the best of our knowledge, this is the first exact rate result known for a MTM type algorithm, although the result's implication is less than encouraging. A good news is that, in a more realistic setting of MTM applications as explained in Sect. 5.1, we can show that MTM improves upon the standard IMH and does not have a suitable competitor.

In a quest for finding advantages MTM may offer, we consider a slightly modified framework that encompasses a few variants of MTM published in the literature. We found that even under the IMH framework, it is possible to construct a MTM algorithm, using either stratified sampling or partial sampling, or sampling without replacement, to gain efficiency. A key to such efficiency gain is to allow multiple trials to be either more dispersed than independent ones (Sect. 5) or applied only to certain "low-cost" parts (Sect. 5.1). Detailed theoretical understanding and guiding principles, however, are still lacking and awaiting further endeavors.

Declarations

Appendix A: Detailed proofs

Proof of Theorem 2 Before proving the theorem, we first define the following additional notations and concepts. Let $A(\cdot, \cdot)$ denote the Markov transition kernel implied by our algorithm. The operator K associated with the resulting Markov chain is defined as follows: for any measurable function f defined on $\mathcal{X}$, operator K maps f to another function defined on $\mathcal{X}$:

$$Kf(x) = \int_{\mathcal{X}} f(y) A(x, \mathrm{d}y).$$

We require that function $f \in L^2(\pi)$. It is easy to see that $Kf \in L^2(\pi)$ as well, meaning that K defines a linear bounded operator on the Hilbert space $L^2(\pi)$ with operator norm 1. For any set $S \subset \mathcal{X}$, we shall also denote $\chi_S : \mathcal{X} \to \{0, 1\}$ as the indicator function which equals 1 if and only if on S. Intuitively, K is just a conditional expectation operator. Note that the constant function 1 is automatically an eigenfunction of eigenvalue 1. We are interested in finding the spectral gap, i.e., the difference between 1 and the second largest eigenvalue. We thus focus on the restricted operator K_0 defined on the orthogonal complement of the constant function:

$$L_0^2(\pi) = \left\{ f \in L^2(\pi) : \int_{\mathcal{X}} f(x)\mathrm{d}x = 0 \right\}.$$

Given Theorem 1, we divide the operator K_0 into two parts: $\forall f \in L_0^2(\pi)$,

$$
\begin{aligned}
K_0 f(x) &= R(x) f(x) \\
&\quad + \int_{\mathcal{X}} \min\left[H[w(x)], H[w(y)]\right] f(y)\pi(y)\mathrm{d}y \\
&=: M_R f(x) + U f(x).
\end{aligned}
$$

Before presenting the formal proof, we remark that this decomposition has the same nature as that in Section 2.1 of Liu (1996), in which the multiplication operator M_R is a low-rank component and the integral-like operator U that resembles the upper triangular matrix in the discrete case. This proof is analogous to that in Atchadé and Perron (2007). The formal proof is divided into the following steps.

Step 1. We first show that operator U is compact. Under the following condition,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \min\{H[w(x)], H[w(y)]\}^2 \pi(x)\pi(y)\mathrm{d}x\mathrm{d}y < \infty,$$

operator U is Hilbert-Schmidt, and therefore compact. Hence, by Weyl's perturbation theorem, we have

$$\sigma_{\mathrm{ess}}(K_0) = \sigma_{\mathrm{ess}}(M_R) \subset \mathrm{ess\text{-}ran}(R).$$

Step 2. Given this, combined with the decomposition

$$\sigma(K_0) = \sigma_{\mathrm{ess}}(K_0) \cup \sigma_d(K_0),$$

we know that it suffices to prove that $\sigma_d(K_0) \subset \mathrm{ess\text{-}ran}(R)$, i.e. all eigenvalues of K_0 are in the essential range of R. To proceed, we assume that there exists $f_0 \in L_0^2(\pi)$ and $\lambda \notin \mathrm{ess\text{-}ran}(R)$, but $K_0 f_0 = \lambda f_0$.

Direct computations yield that for any $f \in L_0^2(\pi)$

$$Uf(x) = \int \min\{H[w(x)], H[w(y)]\} f(y)\pi(y)\mathrm{d}y$$
$$= \int_{\{y:w(y)\geq w(x)\}} H[w(y)]f(y)\pi(y)\mathrm{d}y$$
$$+ \int_{\{y:w(y)<w(x)\}} H[w(x)]f(y)\pi(y)\mathrm{d}y$$
$$= \int_{\{y:w(y)\geq w(x)\}} \{H[w(y)] - H[w(x)]\} f(y)\pi(y)\mathrm{d}y.$$

Since we assume that $\lambda \notin \mathrm{ess\text{-}ran}(R)$, we have $\kappa = \mathrm{ess\,inf}\,(|R(x) - \lambda|) > 0$. We can rearrange equation $K_0 f_0 = \lambda f_0$ to arrive at

$$\int_{\{y:w(y)\geq w(x)\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda} f_0(y)\pi(y)\mathrm{d}y = f_0(x),$$
$$\tag{A1}$$

which can be simplified as $N f_0 = f_0$ with N being an operator well-defined on $L^2(\pi)$ (rather than $L_0^2(\pi)$ in which K_0 is defined). Then, we aim to derive a contradiction about the spectral radius $\mathrm{radii}(G) \triangleq \sup\{|\lambda| : \lambda \in \sigma(G)\}$ for some linear operator G on $L^2(\pi)$ induced by N.

Step 3. Since f_0 is not identically vanishing, we can find $u < w^*$ so that f_0 is not null on $\{x \in \mathcal{X} : u < w(x) \leq w^*\}$. For any partition $I_n = (u = u_n \leq u_{n-1} \leq \ldots \leq u_0 = w^*)$, we denote $D_i = \{x \in \mathcal{X} : u_i < w(x) \leq u_{i-1}\}$ and $L_i^2(\pi) = \{h \in L_0^2(\pi) : h(x) = 0, \forall x \notin D_i\}$ for $i = 1, \ldots, n$. Then $L_i^2(\pi)$ is a closed subspace of $L_0^2(\pi)$, thus a Hilbert space. Moreover, we introduce M_{D_i} as the restriction operator onto D_i on $L^2(\pi)$, by letting $M_{D_i} g(x) = \chi_{D_i}(x)g(x)$ for any $g \in L^2(\pi)$.

We know that

$$M_{D_1} N f_0(x) = \int_{\{y:w(y)\geq w(x)\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda}$$
$$f_0(y)\chi_{D_1}(x)\pi(y)\mathrm{d}y$$
$$= \int_{\{y:w(y)\geq w(x)\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda}$$
$$\chi_{D_1}(y)f_0(y)\chi_{D_1}(x)\pi(y)\mathrm{d}y$$
$$= M_{D_1} N M_{D_1} f_0,$$

where the second inequality follows from the fact that $y \notin D_1$ and $w(y) \geq w(x)$ would together imply that $x \notin D_1$. Obtaining from $N f_0 = f_0$ and $M_{D_1}^2 = M_{D_1}$, we then have $f_{0,D_1} \triangleq M_{D_1} f_0 = M_{D_1} N f_0 = M_{D_1} N M_{D_1} f_0 = M_{D_1} N M_{D_1} f_{0,D_1}$.

In the same manner, we have

$$M_{D_i} N f_0(x)$$
$$= \int_{\{y:w(y)\geq w(x)\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda} f_0(y)\chi_{D_i}(x)\pi(y)\mathrm{d}y$$
$$= \sum_{k=1}^{i-1} \int_{\{y\in D_k\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda} f_0(y)\chi_{D_i}(x)\pi(y)\mathrm{d}y$$
$$+ \int_{\{y:u_{i-1}>w(y)\geq w(x)\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda}$$
$$f_0(y)\chi_{D_i}(x)\pi(y)\mathrm{d}y$$
$$= M_{D_i} h_i(x) + \int_{\{y:u_{i-1}>w(y)\geq w(x)\}}$$
$$\frac{H[w(x)] - H[w(y)]}{R(x) - \lambda} \chi_{D_i}(y) f_0(y)\chi_{D_i}(x)\pi(y)\mathrm{d}y$$
$$= M_{D_i} h_i(x) + M_{D_i} N M_{D_i} f_{0,D_i},$$

where $f_{0,D_i} \triangleq M_{D_i} f_0$ and

$$h_i(x) = \sum_{k=1}^{i-1} \int_{\{y\in D_k\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda} f_0(y)\pi(y)\mathrm{d}y.$$
$$\tag{A2}$$

Rearranging these formulae, we know that

$$M_{D_1} N M_{D_1} f_{0,D_1} = f_{0,D_1}, \tag{A3}$$
$$M_{D_2} N M_{D_2} f_{0,D_2} = f_{0,D_2} - M_{D_2} h_2, \tag{A4}$$
$$\cdots$$
$$M_{D_n} N M_{D_n} f_{0,D_n} = f_{0,D_n} - M_{D_n} h_n. \tag{A5}$$

We claim that (A3) implies that $\mathrm{radii}(M_{D_i} N M_{D_i}) \geq 1$ holds true for at least one index $i \in \{1, \ldots, n\}$. Assuming the converse is true, then $M_{D_1} N M_{D_1} f_{0,D_1} = f_{0,D_1}$ implies that $f_{0,D_1} = 0$ (since 1 cannot be an eigenvalue of $M_{D_1} N M_{D_1}$). Consequently, $h_2 = 0$ follows automatically from its definition (A2), and $M_{D_2} N M_{D_2} f_{0,D_2} = f_{0,D_2}$

implies that $f_{0,D_2} = 0$. This argument can be carried out recursively until n, indicating that f_0 has to vanish on $\{x \in \mathcal{X} : u < w(x) \leq \bar{w}\}$, resulting in a contradiction!

Step 4. Finally, we show that for sufficiently small increments, we can make

$$\text{radii}(M_{D_i} N M_{D_i}) < 1, \ \forall i.$$

First, the mapping

$$H : v \in \mathbb{R}_+ \mapsto k \underbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}_{k-1} \frac{1}{v + \sum_{j=1}^{k-1} w(y_j)} \prod_{j=1}^{k-1} p(x) \mathrm{d}x$$

is continuous, at least on $[u, \bar{w}]$.

Second, $\forall g \in L_i^2(\pi)$ with $\|g\| = 1$, by the Cauchy-Schwarz inequality we have

$$\|M_{D_i} N M_{D_i} g\|^2$$
$$= \int_{\{y \in D_i\}} \left\{ \int_{\{y : w(y) \geq w(x)\}} \frac{H[w(x)] - H[w(y)]}{R(x) - \lambda} g(y) \chi_{D_i}(x) \pi(y) \mathrm{d}y \right\}^2$$
$$\pi(x) \mathrm{d}x$$
$$\leq \left(\frac{\max_{[u_i, u_{i-1}]} H - \min_{[u_i, u_{i-1}]} H}{\kappa} \right)^2 \int_{\{y \in D_i\}} g^2(y) \pi(y) \mathrm{d}y$$
$$\leq \left(\frac{\max_{[u_i, u_{i-1}]} H - \min_{[u_i, u_{i-1}]} H}{\kappa} \right)^2 = \left(\frac{\mathrm{osc}_{[u_i, u_{i-1}]} H}{\kappa} \right)^2,$$

where $\mathrm{osc}_{[u_i, u_{i-1}]} H \triangleq \max_{[u_i, u_{i-1}]} H - \min_{[u_i, u_{i-1}]} H$ denotes the oscillation of H within $[u_i, u_{i-1}]$. Therefore, $\|M_{D_i} N M_{D_i}\| \leq \mathrm{osc}_{[u_i, u_{i-1}]} H / \kappa$. At last, if we choose the partition to be sufficiently small, we would have $\text{radii}(M_{D_i} N M_{D_i}) < 1$ for all i. We then derive a final contradiction to assert that $\sigma_d(K_0) \subset \text{ess-ran}(R)$, ending the proof. $\square$

Proof of Theorem 4 In this proof, every random variable X is taken independently from p. This inequality is proved by induction. First, for $k = 1$, the inequality reduces to equality due to a previous result of Liu (1996) and Atchadé and Perron (2007). For $k = 2$, we see that

$$1 - \mathbb{E}\left[\frac{2}{w^* + w(X)} \right] - \left(1 - \frac{1}{w^*} \right)^2$$
$$= \frac{1}{w^*} \mathbb{E}\left[\frac{2w(X)}{w^* + w(X)} - \frac{1}{w^*} \right] \geq \frac{1}{w^*} \mathbb{E}\left[\frac{2w(X)}{2w^*} - \frac{1}{w^*} \right] = 0.$$

For $k \geq 3$, we will prove the following recursive inequality, which leads to the conclusion of the theorem:

$$1 - \mathbb{E}\left[\frac{k}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] \geq \left(1 - \frac{1}{w^*} \right)$$
$$\left(1 - \mathbb{E}\left[\frac{k-1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] \right). \tag{A6}$$

We prove by simply computing the difference between the two sides:

$$1 - \mathbb{E}\left[\frac{k}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] - \left(1 - \frac{1}{w^*} \right)$$
$$\left(1 - \mathbb{E}\left[\frac{k-1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] \right)$$
$$= \mathbb{E}\left[\frac{k-1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] + \frac{1}{w^*} - \mathbb{E}\left[\frac{k}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right]$$
$$- \mathbb{E}\left[\frac{k-1}{w^*[w^* + \sum_{i=1}^{k-2} w(X_i)]} \right]$$
$$= (k-1) \underbrace{\left(\mathbb{E}\left[\frac{1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] - \mathbb{E}\left[\frac{1}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] \right)}_{(i)}$$
$$+ \underbrace{\frac{1}{w^*} - \mathbb{E}\left[\frac{1}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] - \mathbb{E}\left[\frac{k-1}{w^*[w^* + \sum_{i=1}^{k-2} w(X_i)]} \right]}_{(ii)}.$$

We note that (i) can be modified as

$$(i) = (k-1) \left(\mathbb{E}\left[\frac{1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] - \mathbb{E}\left[\frac{1}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] \right)$$
$$= \sum_{j=1}^{k-1} \left(\mathbb{E}\left[\frac{1}{w^* + \sum_{1 \leq i \leq k-1, i \neq j} w(X_i)} \right] \right.$$
$$\left. - \mathbb{E}\left[\frac{1}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] \right)$$
$$= \sum_{j=1}^{k-1} \mathbb{E}\left\{ \frac{w(X_j)}{[w^* + \sum_{1 \leq i \leq k-1, i \neq j} w(X_i)][w^* + \sum_{i=1}^{k-1} w(X_i)]} \right\}.$$

For (ii), we have

$$(ii) = \mathbb{E}\left\{ \frac{\sum_{j=1}^{k-1} w(X_j)}{w^*[w^* + \sum_{i=1}^{k-1} w(X_i)]} \right\}$$
$$- \mathbb{E}\left[\frac{k-1}{w^*[w^* + \sum_{i=1}^{k-2} w(X_i)]} \right]$$
$$= \sum_{j=1}^{k-1} \mathbb{E}\left\{ \frac{w(X_j)}{w^*[w^* + \sum_{i=1}^{k-1} w(X_i)]} \right\}$$
$$- \sum_{j=1}^{k-1} \mathbb{E}\left\{ \frac{w(X_j)}{w^*[w^* + \sum_{1 \leq i \leq k-1, i \neq j} w(X_i)]} \right\}$$
$$= - \sum_{j=1}^{k-1} \mathbb{E}\left\{ \frac{w(X_j)^2}{w^*[w^* + \sum_{i=1}^{k-1} w(X_i)][w^* + \sum_{1 \leq i \leq k-1, i \neq j} w(X_i)]} \right\}.$$

In conclusion, we have

$$1 - \mathbb{E}\left[\frac{k}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] - \left(1 - \frac{1}{w^*} \right)$$
$$\left(1 - \mathbb{E}\left[\frac{k-1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] \right) = (i) + (ii)$$

$$= \sum_{j=1}^{k-1} \mathbb{E}$$

$$\left\{ \frac{w(X_j)[w^* - w(X_j)]}{w^*[w^* + \sum_{i=1}^{k-1} w(X_i)][w^* + \sum_{1 \le i \le k-1, i \ne j} w(X_i)]} \right\} \ge 0.$$

Consequently, suppose the inequality (15) holds for $k - 1$, i.e.,

$$1 - \mathbb{E}\left[\frac{k-1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] \ge \left(1 - \frac{1}{w^*} \right)^{k-1},$$

from (A6) it immediately follows

$$1 - \mathbb{E}\left[\frac{k}{w^* + \sum_{i=1}^{k-1} w(X_i)} \right] \ge \left(1 - \frac{1}{w^*} \right)$$

$$\left(1 - \mathbb{E}\left[\frac{k-1}{w^* + \sum_{i=1}^{k-2} w(X_i)} \right] \right) \ge \left(1 - \frac{1}{w^*} \right)^k$$

By induction, the final result (15) holds for arbitrary $k \ge 1$.
□

Proof of Theorem 5 Part 1 derives the convergence rate of Algorithm 6. Part 2 derives the convergence rate of the corresponding sequential IMH sampler. Part 3 finishes by deriving the inequality (23) via induction.
Part 1. Via straight forward computation, the transition probability of Algorithm 6 has the following formula ($x \ne y$)

$$A(x, y) = \sum_{j=1}^{k} \underbrace{\int \cdots \int}_{k-1}$$

$$\frac{w_j(y)p_j(y) \prod_{i \ne j} p_i(y_i)\mathrm{d}y_i}{\max\{w_j(y) + \sum_{i \ne j} w_i(y_i), w_j(x) + \sum_{i \ne j} w_i(y_i)\}}$$

$$= \pi(y) \sum_{j=1}^{k} \underbrace{\int \cdots \int}_{k-1} \frac{\prod_{i \ne j} p_i(y_i)\mathrm{d}y_i}{\max\{w_j(y), w_j(x)\} + \sum_{i \ne j} w_i(y_i)}.$$

Plug $\max\{w_j(y), w_j(x)\} \le w_j^*$ into this formula to get

$$A(x, y) \ge \pi(y)\mathbb{E}_p\left[\sum_{j=1}^{k} \frac{1}{w_j^* + \sum_{1 \le i \le k, i \ne j} w_i(X_i)} \right],$$

where X_i is taken independently from $p_i(\cdot)$. Actually this inequality is sufficient to derive a decomposition of $A(x, \cdot)$ as in (6). As shown in the proof of Theorem 3, we upper bound the convergence rate by $1 - \sum_{j=1}^{k} \mathbb{E}\left[\frac{1}{w_j^* + \sum_{i=1, i \ne j}^{k} w_i(X_i)} \right]$ via coupling argument, Lemma 1.

Specifically, when there exists x^* such that $w_j(x^*) = w_j^*$ for all $j = 1, \ldots, k$, we find for any $y \ne x^*$,

$$A(x^*, y) = \pi(y)\mathbb{E}_p\left[\sum_{j=1}^{k} \frac{1}{w_j^* + \sum_{1 \le i \le k, i \ne j} w_i(X_i)} \right].$$

Consequently, the rejection probability at x^* is

$$R(x^*) = 1 - \mathbb{E}_p\left[\sum_{j=1}^{k} \frac{1}{w_j^* + \sum_{1 \le i \le k, i \ne j} w_i(X_i)} \right].$$

Then we lower bound the convergence rate via Lemma 2.
Part 2. Turn to the corresponding sequential IMH sampler. For simplicity, we utilize the concept of L^2 operators introduced in Sect. 2 to derive upper bounds. Within one iteration, the sampler runs an interior loop of length k, with each step as a vanilla IMH step using proposal p_i. The transition probability of a vanilla IMH step is

$$A^{(i)}(x, y) = \frac{1}{\max\{w_i(x), w_i(y)\}}\pi(y)$$
$$+ \left(1 - \int_{\mathcal{X}} \frac{1}{\max\{w_i(x), w_i(y)\}}\pi(y)\mathrm{d}y \right) \delta_x(y).$$

Denote $K^{(i)}$ as the operator defined in $L^2(\pi)$ by $K^{(i)}f(x) = \int f(y)A^{(i)}(x, y)\mathrm{d}y$, and denote $K_0^{(i)}$ as the restriction of $K^{(i)}$ onto $L_0^2(\pi)$, the orthogonal complement of the constant function of $L^2(\pi)$. Theorem 2 implies $\|K_0^{(i)}\| \le 1 - 1/w_i^*$. Denote the whole transition probability of one iteration as $\bar{A}$ and associated operators as $\bar{K}$ and $\bar{K}_0$. Consequently,

$$\|\bar{K}_0\|_2 = \|\bar{K}_0^{(k)} \cdots \bar{K}_0^{(1)}\|_2 \le \prod_{i=1}^{k}(1 - 1/w_i^*).$$

Let $p_n(x) = \bar{A}_n(p_0, x)$ denote the distribution of the n-th state of the Markov chain after n steps from initialization p_0. Liu et al. (1995) establishes

$$\|p_n - \pi\|_{TV} \le 2d_\chi(\pi, p_n) \le 2\|\bar{K}_0^n\|_2 d_\chi(\pi, p_0).$$

Furthermore, we obtain an upper bound on the convergence rate defined in (13): $r \le \|\bar{K}_0\|_0 = \prod_{i=1}^{k}(1 - 1/w_i^*)$.
For a matching lower bound, we consider the special point $x^* \in \mathcal{X}$ such that for all i,

$$A^{(i)}(x^*, y) = \frac{1}{w_i^*}\pi(y) + \left(1 - \frac{1}{w_i^*} \right)\delta_{x^*}(y).$$

Going through the full interior loop within one iteration, the whole rejection probability is at least

$$R(x^*) \geq \prod_{i=1}^{k} \left(1 - \frac{1}{w_i^*}\right).$$

By Lemma 2, a matching lower bound thus obtained.
Part 3. We then establish (23). For $k = 2$,

$$1 - \mathbb{E}\left[\frac{1}{w_1^* + w_2(X_2)}\right] - \mathbb{E}\left[\frac{1}{w_1(X_1) + w_2^*}\right]$$
$$- \left(1 - \frac{1}{w_1^*}\right)\left(1 - \frac{1}{w_2^*}\right)$$
$$= \mathbb{E}\left[\frac{w_1(X_1)}{w_2^*(w_1(X_1) + w_2^*)}\right]$$
$$+ \mathbb{E}\left[\frac{w_2(X_2)}{w_1^*(w_1^* + w_2(X_2))}\right] - \frac{1}{w_1^* w_2^*}$$
$$\geq \frac{1}{w_1^*(w_1^* + w_2^*)} + \frac{1}{w_2^*(w_1^* + w_2^*)} - \frac{1}{w_1^* w_2^*} = 0.$$

For larger $k > 2$, we have, for an arbitrary fixed $l \in \{1, \ldots, k\}$,

$$1 - \sum_{j=1}^{k} \mathbb{E}\left[\frac{1}{w_j^* + \sum_{i=1, i \neq j}^{k} w_i(X_i)}\right]$$
$$- \left(1 - \frac{1}{w_l^*}\right)\left\{1 - \sum_{j=1, j \neq l}^{k} \mathbb{E}\left[\frac{1}{w_j^* + \sum_{i=1, i \neq j, i \neq l}^{k} w_i(X_i)}\right]\right\}$$
$$= \sum_{j=1, j \neq l}^{k} \mathbb{E}\left[\frac{w_j(X_j)}{w_l^*[w_l^* + \sum_{i=1, i \neq l}^{k} w_i(X_i)]}\right]$$
$$- \sum_{j=1, j \neq l}^{k} \mathbb{E}\left[\frac{1}{w_l^*[w_j^* + \sum_{i=1, i \neq j, i \neq l}^{k} w_i(X_i)]}\right]$$
$$+ \sum_{j=1, j \neq l}^{k} \mathbb{E}$$
$$\left[\frac{w_l(X_l)}{[w_j^* + \sum_{i=1, i \neq j}^{k} w_i(X_i)][w_j^* + \sum_{i=1, i \neq j, i \neq l}^{k} w_i(X_i)]}\right]$$
$$\geq \sum_{j=1, j \neq l}^{k} \mathbb{E}$$
$$\left[\frac{1}{w_l^*(w_l^* + w_j^* + B_{jl})} + \frac{1}{(w_l^* + w_j^* + B_{jl})(w_j^* + B_{jl})}\right.$$
$$\left. - \frac{1}{w_l^*(w_j^* + B_{jl})}\right] = 0, \qquad (A7)$$

where we denote $B_{jl} = \sum_{i=1, i \neq j, i \neq l}^{k} w_i(X_i)$ for simplicity. The last inequality is mainly due to

$$w_l^* + \sum_{i=1, i \neq l}^{k} w_i(X_i) \leq w_l^* + w_j^* + \sum_{i=1, i \neq l, i \neq j}^{k}$$
$$w_i(X_i) = w_l^* + w_j^* + B_{jl}$$

applied in the denominators of the two positive terms. The last step of induction is the same as the proof of Theorem 4. Suppose the result holds for $k - 1$, i.e.,

$$1 - \sum_{j=1}^{k-1} \mathbb{E}_p\left[\frac{1}{w_j^* + \sum_{1 \leq i \leq k-1, i \neq j} w_i(X_i)}\right] \geq \prod_{i=1}^{k-1}\left(1 - \frac{1}{w_i^*}\right),$$

it immediately follows from (A7) with $l = k$ that

$$1 - \sum_{j=1}^{k} \mathbb{E}_p\left[\frac{1}{w_j^* + \sum_{1 \leq i \leq k-1, i \neq j} w_i(X_i)}\right]$$
$$\geq \left(1 - \sum_{j=1}^{k-1} \mathbb{E}_p\left[\frac{1}{w_j^* + \sum_{1 \leq i \leq k-1, i \neq j} w_i(X_i)}\right]\right)$$
$$\left(1 - \frac{1}{w_k^*}\right)$$
$$\geq \prod_{i=1}^{k}\left(1 - \frac{1}{w_i^*}\right).$$

The proofs of Theorem 4 and Theorem 5 are essentially the same, both utilizing induction to recursively handle a general integer k. $\qquad \square$

Proof of Theorem 6 To make our notations more explicit, we assume that every distribution mentioned here has a density with respect to the Lebesgue measure. Denote $A(x, y)$ as the actual transition density, we compute directly that

$$\pi(x) A(x, y)$$
$$= \pi(x) \sum_{j=1}^{k} \mathbb{P}(y_j = y, J = j, y_J \text{ gets accepted})$$
$$= \pi(x) \sum_{j=1}^{k} \int p(x, \mathbf{y}_j) \frac{w_j(y, x)}{w_j(y, x) + \sum_{i \neq j} w_i(y_i, x)}$$
$$\rho p(y, \mathbf{x}_{-j}^* \mid x) \prod_{i \neq j} dy_i dx_i^*,$$

where we write $\mathbf{x}_{(-j)}^* = (x_1^*, x_2^*, \ldots, x_{j-1}^*, x_{j+1}^*, \ldots, x_k^*) \in \mathcal{X}^{k-1}$ and $\mathbf{y}(j) = (y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_k) \in \mathcal{X}^k$. Plugging in the definition of ρ, we use the notations $\mathbf{x}^*(j) \triangleq (x_1^*, \ldots, x_{j-1}^*, x, x_{j+1}^*, x_k^*)$ and $u_j(\mathbf{x}, \mathbf{y}) \triangleq \min$

$$\left\{\frac{1}{\sum_{i=1}^{k} w_i(y_i, x_j)}, \frac{1}{\sum_{i=1}^{k} w_i(x_i, y_j)}\right\} \text{ to get}$$

$$\pi(x)A(x, y)$$

$$= \pi(x) \sum_{j=1}^{k} \int p(x, \mathbf{y}(j)) \frac{w_j(y, x)}{w_j(y, x) + \sum_{i \neq j} w_i(y_i, x)}$$

$$\min\left\{1, \frac{w_j(y, x) + \sum_{i \neq j} w_i(y_i, x)}{w_j(x, y) + \sum_{i \neq j} w_i(x_i^*, y)}\right\}$$

$$p(y, \mathbf{x}_{(-j)}^* \mid x) \prod_{i \neq j} \mathrm{d}y_i \mathrm{d}x_i^*$$

$$= \sum_{j=1}^{k} \pi(x) w_j(y, x) p_j(x, y) \int u_j(\mathbf{x}^*(j), \mathbf{y})$$

$$p(x, \mathbf{y}_{(-j)} \mid y_j = y) p(y, \mathbf{x}_{(-j)}^* \mid x_j = x) \prod_{i \neq j} \mathrm{d}y_i \mathrm{d}x_i^*.$$

In the above formula, we use the identity

$$p(x, \mathbf{y}(j)) = p_j(x, y) \times p(x, \mathbf{y}_{(-j)} \mid y_j = y).$$

At last, note that $\pi(x) w_j(y, x) p_j(x, y) = \pi(x)\pi(y) p_j(x, y) p_j(y, x)\lambda_j(x, y)$ is symmetric by our constructions, which implies that $\pi(x)A(x, y)$ is symmetric in x and y, proving the detailed balance condition. $\square$

Proof of Theorem 7 If we simply set $x_j^* := y_j$ for any $j \neq J$ in Algorithm 7, the conditional probability becomes

$$\pi(x)A(x, y)$$

$$= \pi(x) \sum_{j=1}^{k} \left[\int p(x, \mathbf{y}(j)) w_j(y, x) \right.$$

$$\min\left[\frac{1}{w_j(y, x) + \sum_{i \neq j} w_i(y_i, x)}, \right.$$

$$\left. \frac{1}{w_j(x, y) + \sum_{i \neq j} w_i(y_i, y)} \right] \prod_{i \neq j} \mathrm{d}y_i \right]$$

$$= \sum_{j=1}^{k} \left[\int \pi(x) p_j(x, y) w_j(y, x) p(x, \mathbf{y}_{(-j)} \mid y_j = y) \right.$$

$$\min\left[\frac{1}{w_j(y, x) + \sum_{i \neq j} w_i(y_i, x)}, \right.$$

$$\left. \frac{1}{w_j(x, y) + \sum_{i \neq j} w_i(y_i, y)} \right] \prod_{i \neq j} \mathrm{d}y_i \right].$$

Since $\pi(x) w_j(y, x) p_j(x, y)$ is symmetric for x and y, the theorem follows easily from condition (16) in the main text. $\square$

References

Atchadé, Y.F., Perron, F.: On the geometric ergodicity of metropolis-hastings algorithms. Statistics **41**(1), 77–84 (2007)

Bédard, M., Douc, R., Moulines, E.: Scaling analysis of multiple-try mcmc methods. Stoch. Process. Appl. **122**(3), 758–786 (2012)

Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: Handbook of Markov Chain Monte Carlo. CRC Press (2011)

Calderhead, B.: A general construction for parallelizing metropolis-hastings algorithms. Proc. Natl. Acad. Sci. **111**(49), 17408–17413 (2014)

Casarin, R., Craiu, R., Leisen, F.: Interacting multiple try algorithms with different proposal distributions. Stat. Comput. **23**(2), 185–200 (2013)

Chen, X.-H., Dempster, A.P., Liu, J.S.: Weighted finite population sampling to maximize entropy. Biometrika **81**(3), 457–469 (1994)

Craiu, R.V., Lemieux, C.: Acceleration of the multiple-try metropolis algorithm using antithetic and stratified sampling. Stat. Comput. **17**(2), 109–120 (2007)

Dai, C., Liu, J.S.: Monte Carlo approximation of bayes factors via mixing with surrogate distributions. J. Am. Stat. Assoc. **117**, 765 (2020)

Diaconis, P., Khare, K., Saloff-Coste, L.: Gibbs sampling, exponential families and orthogonal polynomials. Stat. Sci. **23**(2), 151–178 (2008)

Diaconis, P., Saloff-Coste, L.: What do we know about the metropolis algorithm? J. Comput. Syst. Sci. **57**(1), 20–36 (1998)

Frenkel, D., Smit, B., Ratner, M.A.: Understanding Molecular Simulation: from Algorithms to Applications. Academic press San Diego (1996)

Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications (1970)

Levin, D.A., Peres, Y.: Markov Chains and Mixing Times, vol. 107. American Mathematical Soc (2017)

Liu, J.S.: Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Stat. Comput. **6**(2), 113–119 (1996)

Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer (2008)

Liu, J.S., Liang, F., Wong, W.H.: The multiple-try method and local optimization in metropolis sampling. J. Am. Stat. Assoc. **95**(449), 121–134 (2000)

Liu, J.S., Wong, W.H., Kong, A.: Covariance structure and convergence rate of the gibbs sampler with various scans. J. R. Stat. Soc. Ser. B **57**(1), 157–169 (1995)

Martino, L.: A review of multiple try mcmc algorithms for signal processing. Digit. Signal Process. **75**, 134–152 (2018)

Martino, L., Leisen, F., Corander, J.: On multiple try schemes and the particle metropolis-hastings algorithm. arXiv preprint arXiv:1409.0051 (2014)

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)

Neal, R.M.: Mcmc using ensembles of states for problems with fast and slow variables such as gaussian process regression. arXiv preprint arXiv:1101.0387 (2011)

Pandolfi, S., Bartolucci, F., Friel, N.: A generalization of the multiple-try metropolis algorithm for bayesian estimation and model selection. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 581–588. JMLR Workshop and Conference Proceedings (2010)

Qin, Z.S., Liu, J.S.: Multipoint metropolis method with application to hybrid monte Carlo. J. Comput. Phys. **172**(2), 827–840 (2001)

Roberts, G.O., Tweedie, R.L.: Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. Biometrika **83**(1), 95–110 (1996)

Tierney, L.: Markov chains for exploring posterior distributions. Ann. Stat. **22**, 1701–1728 (1994)

Wang, G.: Exact convergence analysis of the independent metropolis-hastings algorithms. arXiv preprint arXiv:2008.02455 (2020)

Yang, S., Chen, Y., Bernton, E., Liu, J.S.: On parallelizable Markov chain monte Carlo algorithms with waste-recycling. Stat. Comput. **28**(5), 1073–1081 (2018)