Submitted to the Annals of Statistics

COMPARISON OF ASYMPTOTIC VARIANCES OF INHOMOGENEOUS MARKOV CHAINS WITH APPLICATION TO MARKOV CHAIN MONTE CARLO METHODS

By Florian Maire [‡], Randal Douc[‡] and Jimmy Olsson[§]

Institut Télécom/Télécom SudParis, CNRS UMR 5157 SAMOVAR [‡] and KTH Royal Institute of Technology[§]

In this paper we study the asymptotic variance of sample path averages for inhomogeneous Markov chains that evolve alternatingly according to two different π-reversible Markov transition kernels P and Q. More specifically, our main result allows us to compare directly the asymptotic variances of two inhomogeneous Markov chains associated with different kernels P_i and Q_i, $i \in \{0, 1\}$, as soon as the kernels of each pair (P_0, P_1) and (Q_0, Q_1) can be ordered in the sense of lag-one autocovariance. As an important application we use this result for comparing different data-augmentation-type Metropolis-Hastings algorithms. In particular, we compare some pseudo-marginal algorithms and propose a novel exact algorithm, referred to as the *random refreshment* algorithm, which is more efficient, in terms of asymptotic variance, than the Grouped Independence Metropolis Hastings algorithm and has a computational complexity that does not exceed that of the Monte Carlo Within Metropolis algorithm.

1. Introduction. *Markov chain Monte Carlo* (MCMC) *methods* allow samples from virtually any target distribution π, known up to a normalizing constant, to be generated. In particular, the celebrated *Metropolis-Hastings algorithm* (introduced in [8] and [6]) simulates a Markov chain evolving according to a π-reversible Markov transition kernel by first generating, using some instrumental kernel, a candidate and then accepting or rejecting the same with a probability adjusted to satisfy the detailed balance condition [16]. When choosing between several Metropolis-Hastings algorithms, it is desirable to be able to compare the efficiencies, in terms of the asymptotic variance of sample path averages, of different π-reversible Markov chains. Despite the practical importance of this question, only a few results in this direction exist the literature. P. H. Peskun [12] defined a partial ordering for finite state space Markov chains, where one transition kernel has a higher

[§]This work is supported by the Swedish Research Council.

[‡]This work has been supported partially by the ONERA, the French Aerospace Lab and the DGA, the French Procurement Agency

order than another if the former dominates the latter on the off-diagonal (see Definition 1). This ordering was extended later by L. Tierney [16] to general state space Markov chains and another even more general ordering, the covariance ordering, was proposed in [9]. In general it holds that if a homogeneous π-reversible Markov transition kernel is greater than another according to one of these orderings, then the asymptotic variance of sample path averages for a Markov chain evolving according to the former is smaller for all square integrable (with respect to π) target functions.

We provide an extension of this result to inhomogeneous Markov chains that evolve alternatingly according to two different π-reversible Markov transition kernels. To the best of our knowledge, this is the first work dealing with systematic comparison of asymptotic variances of inhomogeneous Markov chains. The approach is linked with the operator theory for Markov chains but does not make use of any spectral representation. After some preliminaries (Section 2), our main result, Theorem 4, is stated in Section 3. In Section 4 we apply Theorem 4 in the context of MCMC algorithms by comparing the efficiency, in terms of asymptotic variance, of some existing data-augmentation-type algorithms. Moreover, we propose a novel pseudo-marginal algorithm (in the sense of [1]), referred to as the *random refreshment* algorithm, which—on the contrary to the pseudo-marginal version of the *Monte Carlo Within Metropolis* (MCWM) algorithm—turns out to be exact and more efficient than the pseudo-marginal version of the *Grouped Independence Metropolis Hastings* (GIMH) algorithm. Here the analysis is again driven by Theorem 4. The proof of Theorem 4 is given in Section 5 and some technical lemmas are postponed to Appendix A. Finally, Appendix B relates some existing MCMC algorithms to the framework considered in this paper.

2. Preliminaries. We denote by $\mathbb{N} := \{0, 1, 2, \ldots\}$ and $\mathbb{N}^* := \{1, 2, \ldots\}$ the sets of non-negative and positive integers, respectively. In the following all random variables are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $(\mathsf{X}, \mathcal{X})$ be a measurable space; then we denote by $\mathcal{M}(\mathcal{X})$ and $\mathcal{F}(\mathcal{X})$ the spaces of positive measures and measurable functions on $(\mathsf{X}, \mathcal{X})$, respectively. The Lebesgue integral of $f \in \mathcal{F}(\mathcal{X})$ over X with respect to the measure $\mu \in \mathcal{M}(\mathcal{X})$ is, when well-defined, denoted by $\mu f := \int f(x)\, \mu(\mathrm{d}x)$. Recall that a *Markov transition kernel* P on $(\mathsf{X}, \mathcal{X})$ is a mapping $P : \mathsf{X} \times \mathcal{X} \to [0, 1]$ such that

- for all $\mathsf{A} \in \mathcal{X}$, $\mathsf{X} \ni x \mapsto P(x, \mathsf{A})$ is a measurable function,
- for all $x \in \mathsf{X}$, $\mathcal{X} \ni \mathsf{A} \mapsto P(x, \mathsf{A})$ is a probability measure.

A kernel P induces two integral operators, one acting on $\mathcal{M}(\mathcal{X})$ and the other on $\mathcal{F}(\mathcal{X})$; more specifically, for $\mu \in \mathcal{M}(\mathcal{X})$ and $f \in \mathcal{F}(\mathsf{X})$, we define the measure

$$\mu P : \mathcal{X} \ni \mathsf{A} \mapsto \int P(x, \mathsf{A})\mu(\mathrm{d}x)$$

and the measurable function

$$Pf : \mathsf{X} \ni x \mapsto \int f(x')P(x, \mathrm{d}x').$$

Moreover, the *composition* (or *product*) of two kernels P and Q on $(\mathsf{X}, \mathcal{X})$ is the kernel defined by

$$PQ : \mathsf{X} \times \mathcal{X} \ni (x, \mathsf{A}) \mapsto \int Q(x', \mathsf{A})P(x, \mathrm{d}x').$$

We will from now on fix a distinguished probability measure π on $(\mathsf{X}, \mathcal{X})$. Given π, we denote by $\mathsf{L}^2(\pi) := \{f \in \mathcal{F}(\mathcal{X}) : \pi f^2 < \infty\}$ the space of square integrable functions with respect to π and furnish the same with the scalar product

$$\langle f, g\rangle := \int f(x)g(x)\pi(\mathrm{d}x) \quad (f \in \mathsf{L}^2(\pi), g \in \mathsf{L}^2(\pi))$$

and the associated norm

$$\|f\|_{\mathsf{L}^2} := \left(\pi f^2\right)^{1/2} \quad (f \in \mathsf{L}^2(\pi)).$$

Here we have expunged the measure π from the notation for brevity. If P is a Markov kernel on $(\mathsf{X}, \mathcal{X})$ admitting π as an invariant distribution, then the mapping $f \mapsto Pf$ defines an operator on $\mathsf{L}^2(\pi)$, and by Jensen's inequality it holds that

$$\|P\| := \sup_{f \in \mathsf{L}^2(\pi) : \|f\|_{\mathsf{L}^2} \leq 1} \|Pf\|_{\mathsf{L}^2} \leq 1. \tag{1}$$

Recall that a kernel P is π-*reversible* if and only if the detailed balance relation

$$\pi(\mathrm{d}x)P(x, \mathrm{d}x') = \pi(\mathrm{d}x')P(x', \mathrm{d}x)$$

holds. If the Markov kernel P is π-reversible, then $f \mapsto Pf$ defines a self-adjoint operator on $\mathsf{L}^2(\pi)$, i.e. for all f and g belonging to $\mathsf{L}^2(\pi)$,

$$\langle f, Pg\rangle = \langle Pf, g\rangle . \tag{2}$$

The following off-diagonal ordering of Markov transition kernels on a common state space was, in the case of Markov chains in a finite state space, proposed in [12]. The ordering was extended later in [16] to the case of general state space Markov chains.

Definition 1. *Let P_0 and P_1 be Markov transition kernels on $(\mathsf{X}, \mathcal{X})$ with invariant distribution π. We say that P_1* dominates *P_0 on the off-diagonal, denoted $P_1 \succeq P_0$, if for all $\mathsf{A} \in \mathcal{X}$ and π-a.s. all $x \in \mathsf{X}$,*

$$P_1(x, \mathsf{A} \setminus \{x\}) \geq P_0(x, \mathsf{A} \setminus \{x\}).$$

The previous ordering allows for comparing the asymptotic efficiency of different reversible kernels. More specifically, the following seminal result was established in [12, Theorem 2.1.1] for Markov chains in discrete state space and extended later in [16, Theorem 4] to Markov chains in general state space.

Theorem 2. *Let P_0 and P_1 be two π-reversible kernels on $(\mathsf{X}, \mathcal{X})$. If $P_1 \succeq P_0$, then for all $f \in \mathsf{L}^2(\pi)$,*

$$v(f, P_1) \leq v(f, P_0),$$

where we have defined, for a Markov chain $\{X_k; k \in \mathbb{N}\}$ with π-reversible transition kernel P and initial distribution π,

$$v(f, P) := \lim_{n \to \infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n-1} f(X_k) \right). \tag{3}$$

Note that according to [16], if $\{X_k; k \in \mathbb{N}\}$ is a π-reversible Markov chain and $f \in \mathsf{L}^2(\pi)$, then $\lim_{n\to\infty} n^{-1}\mathrm{Var}(\sum_{k=0}^{n-1} f(X_k))$ is guaranteed to exist (but may be infinite). Nevertheless, the ordering in question does not allow Markov kernels lacking probability mass on the diagonal, i.e. kernels P satisfying $P(x, \{x\}) = 0$ for all $x \in \mathsf{X}$, to be compared. This is in particular the case for Gibbs samplers in general state space. To overcome this limitation, one may consider instead the following covariance ordering based on lag-one autocovariances.

Definition 3. *Let P_0 and P_1 be Markov transition kernels on $(\mathsf{X}, \mathcal{X})$ with invariant distribution π. We say that P_1* dominates *P_0 in the covariance ordering, denoted $P_1 \succcurlyeq P_0$, if for all $f \in \mathsf{L}^2(\pi)$,*

$$\langle f, P_1 f \rangle \leq \langle f, P_0 f \rangle .$$

The covariance ordering, which was introduced implicitly in [16, p. 5] and formalized in [9], is an extension of the off-diagonal ordering since according to [16, Lemma 3], $P_1 \succeq P_0$ implies $P_1 \succcurlyeq P_0$. Moreover, it turns out that for

reversible kernels, $P_1 \succcurlyeq P_0$ implies $v(f, P_0) \geq v(f, P_1)$ (see the proof of [16, Theorem 4]).

All these results concern homogeneous Markov chains, whereas many MCMC algorithms such as the Gibbs or the Metropolis-within-Gibbs samplers use several kernels, e.g. P and Q in the case of two kernels [13]. A natural idea would then be to apply Theorem 2 to the homogeneous Markov chain having the block kernel PQ as transition kernel; however, even when the kernels P and Q are both π-reversible, the product PQ of the same is usually not π-reversible, except in the particular case when P and Q commute, i.e. $PQ = QP$. Thus, Theorem 2 cannot in general be applied directly in this case.

3. Main assumptions and results. In the following, let P_i and Q_i, $i \in \{0, 1\}$, be Markov transition kernels on $(\mathsf{X}, \mathcal{X})$. Define $\{X_k^{(0)}; k \in \mathbb{N}\}$ and $\{X_k^{(1)}; k \in \mathbb{N}\}$ as the Markov chains evolving as follows:

$$X_0^{(i)} \xrightarrow{P_i} X_1^{(i)} \xrightarrow{Q_i} X_2^{(i)} \xrightarrow{P_i} X_3^{(i)} \xrightarrow{Q_i} \cdots \tag{4}$$

This means that for all $k \in \mathbb{N}$, $i \in \{0, 1\}$, and $\mathsf{A} \in \mathcal{X}$,

- $\mathbb{P}\left(X_{2k+1}^{(i)} \in \mathsf{A} \mid \mathcal{F}_{2k}^{(i)}\right) = P_i(X_{2k}^{(i)}, \mathsf{A})$,
- $\mathbb{P}\left(X_{2k+2}^{(i)} \in \mathsf{A} \mid \mathcal{F}_{2k+1}^{(i)}\right) = Q_i(X_{2k+1}^{(i)}, \mathsf{A})$,

where $\mathcal{F}_n^{(i)} := \sigma(X_0^{(i)}, \ldots, X_n^{(i)})$, $n \in \mathbb{N}$. We impose the following assumption.

(A1) (*i*) P_i and Q_i, $i \in \{0, 1\}$, are π-reversible,
(*ii*) $P_1 \succcurlyeq P_0$ and $Q_1 \succcurlyeq Q_0$.

As mentioned above, $P_1 \succeq P_0$ implies $P_1 \succcurlyeq P_0$; thus, in practice, a sufficient condition for (**A1**)(ii) is that $P_1 \succeq P_0$ and $Q_1 \succeq Q_0$.

THEOREM 4. *Assume that P_i and Q_i, $i \in \{0, 1\}$, satisfy (**A1**) and let $\{X_k^{(i)}; k \in \mathbb{N}\}$, $i \in \{0, 1\}$, be Markov chains evolving as in (9) with initial distribution π. Then for all $f \in \mathsf{L}^2(\pi)$ such that for $i \in \{0, 1\}$,*

$$\sum_{k=1}^{\infty} \left(|\mathrm{Cov}(f(X_0^{(i)}), f(X_k^{(i)}))| + |\mathrm{Cov}(f(X_1^{(i)}), f(X_{k+1}^{(i)}))| \right) < \infty, \tag{5}$$

it holds that

$$v_1(f) \leq v_0(f), \tag{6}$$

where

$$v_i(f) := \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n-1} f(X_k^{(i)})\right) \quad (i \in \{0,1\}). \tag{7}$$

Remark 5. *At present, we have not been able to extend the arguments of our current proof of Theorem 4 (see Section 5) to inhomogeneous Markov chains evolving alternatingly according to* more *than two different kernels. On the other hand, we have not been able to find a counterexample rejecting the hypothesis that a similar result would hold true also in that case. We leave this as an open problem.*

Remark 6. *The condition* (5) *is* not *a necessary condition for* (6)*; indeed, letting* $\mathsf{X} = \{-1, 1\}$, $\pi(\mathrm{d}x') = P_0(x, \mathrm{d}x') = (\delta_1(\mathrm{d}x') + \delta_{-1}(\mathrm{d}x'))/2$, $Q_1 = Q_0 = P_1$, *where, as in [4, Example 5],* $P_1(x, \mathrm{d}x') = \delta_{-x}(\mathrm{d}x')$, *provides a straightforward counterexample.*

When verifying if a given f satisfies the condition (5) it may be convenient to consider the homogeneous Markov chains $\{X_{2k}; k \in \mathbb{N}\}$ or $\{X_{2k+1}; k \in \mathbb{N}\}$ or even $\{(X_{2k}, X_{2k+1}); k \in \mathbb{N}\}$. Typically, none of these chains are π-reversible. Nevertheless, π-reversibility is not needed for checking conditions of type (5), which can be established using upper bounds on the V*-norm* between the distribution given by the nth iterate of a homogeneous kernel and its stationary distribution. This will be developed in the following section.

3.1. *Sufficient conditions for the absolute summability assumption* (5). For any measurable real-valued function f on $(\mathsf{X}, \mathcal{X})$, define the V*-norm* of the *function* f by

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

Moreover, let ξ be a finite signed measure on $(\mathsf{X}, \mathcal{X})$. Then by the Jordan decomposition theorem there exists a unique pair of positive, finite, and singular measures ξ_+ and ξ_- on $(\mathsf{X}, \mathcal{X})$ such that $\xi = \xi_+ - \xi_-$. The pair $\xi_\pm$ is referred to as the *Jordan decomposition* of the signed measure ξ. The finite measure $|\xi| := \xi_+ + \xi_-$ is called the *total variation* of ξ. Let V be a nonnegative function taking values in $[1, \infty)$; then the V*-norm* of the *signed measure* ξ is defined by

$$\|\xi\|_V := |\xi|(V) = \sup_{f:|f|_V\le 1} \xi f.$$

Definition 7. *A Markov kernel P on $(\mathsf{X}, \mathcal{X})$ is V-*geometrically ergodic *if it admits a unique invariant distribution π and there exists a measurable function $V : \mathsf{X} \to [1, \infty)$ satisfying $\pi V < \infty$ and such that the following holds.*

(a) There exist constants $(C, \rho) \in \mathbb{R}^+ \times (0, 1)$ such that for all $x \in \mathsf{X}$ and all $n \in \mathbb{N}$,

$$\|P^n(x, \cdot) - \pi\|_V \leq C\rho^n V(x). \tag{8}$$

(b) There exist constants $(b, \lambda) \in \mathbb{R}^+ \times (0, 1)$ such that $PV \leq \lambda V + b$.

Remark 8. *[5, Theorem 1.2] provides sufficient conditions, in terms of drift towards a* small set*, for (a) in Definition 7 to hold; see also [14, Fact 10] for necessary and sufficient conditions under the assumption of aperiodicity and irreducibility. Moreover, the coming developments require only the bound (8) to hold π-a.s.*

We have now all necessary tools for giving sufficient conditions that imply the absolute summability assumption (5). Let the chain $\{X_k; k \in \mathbb{N}\}$ evolve according to

$$X_0 \xrightarrow{P} X_1 \xrightarrow{Q} X_2 \xrightarrow{P} X_3 \xrightarrow{Q} \cdots \tag{9}$$

with $X_0 \sim \pi$, for some Markov kernels P and Q.

Proposition 9. *If the Markov kernel PQ is V-geometrically ergodic, then for all functions f such that $|f|_{V^{1/2}} < \infty$ and $|Pf|_{V^{1/2}} < \infty$,*

$$\sum_{k=1}^{\infty} \left(|\mathrm{Cov}(f(X_0), f(X_k))| + |\mathrm{Cov}(f(X_1), f(X_{k+1}))| \right) < \infty,$$

where $\{X_k; k \in \mathbb{N}\}$ evolves as in (9).

The proof of Proposition 9 is found in Section A.1.

4. Application to data-augmentation-type algorithms. Before considering some applications of Theorem 4 we recall the following proposition, describing how to obtain a π-reversible Markov chain using some instrumental kernel K. Although this result is fundamental in the Metropolis-Hastings literature (see for example [14], [13], [3], and the references therein), it is restated here as it will be used in various situations in the sequel (especially when there is no fixed reference measure dominating all the distributions $\{K(x, \cdot); x \in \mathsf{X}\}$).

Proposition 10. *Let K be a Markov transition kernel on $\mathsf{X} \times \mathcal{X}$ and π a probability measure on $(\mathsf{X}, \mathcal{X})$. Define the probability measures $\mu(\mathrm{d}x \times \mathrm{d}x') := \pi(\mathrm{d}x)K(x, \mathrm{d}x')$ and $\nu(\mathrm{d}x \times \mathrm{d}x') := \pi(\mathrm{d}x')K(x', \mathrm{d}x)$. Assume that the measures ν and μ are equivalent and such that for μ-a.s. all $(x, x') \in \mathsf{X}^2$,*

$$0 < \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x, x') < \infty, \tag{10}$$

where $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ denotes the Radon-Nikodym derivative. Then the Markov kernel $P(x, \mathrm{d}x') := K(x, \mathrm{d}x')\alpha(x, x') + \delta_x(\mathrm{d}x')\beta(x)$, where

$$\alpha(x, x') := 1 \wedge \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x, x') \quad \textit{and} \quad \beta(x) := 1 - \int K(x, \mathrm{d}x')\alpha(x, x'),$$

is π-reversible.

A natural application of Theorem 4 consists in using the result for comparing different data-augmentation-type algorithms. In the following we wish to target a probability distribution π^* defined on $(\mathsf{Y}, \mathcal{Y})$ using a sequence $\{Y_k; k \in \mathbb{N}\}$ of Y-valued random variables. To this aim, M. A. Tanner and W. H. Wong [15] suggest writing π^* as the marginal of some distribution π defined on the product space $(\mathsf{Y} \times \mathsf{U}, \mathcal{Y} \otimes \mathcal{U})$ in the sense that $\pi(\mathrm{d}y \times \mathrm{d}u) = \pi^*(\mathrm{d}y)\, R(y, \mathrm{d}u)$, where R is some Markov transition kernel on $\mathsf{Y} \times \mathcal{U}$. In most cases the marginal π^* is of sole interest, while the component u is introduced for convenience as a means of coping with analytic intractability of the marginal. (It could also be the case that the marginal π^* is too computationally expensive to evaluate.) A first solution consists in letting $\{Y_k; k \in \mathbb{N}\}$ be the first-component process $\{Y_k^{(1)}; k \in \mathbb{N}\}$ of the π-reversible Markov chain $\{(Y_k^{(1)}, U_k^{(1)}); k \in \mathbb{N}\}$ defined as follows. Let S and T be instrumental Markov transition kernels on $\mathsf{Y} \times \mathsf{U} \times \mathcal{Y}$ and $\mathsf{Y} \times \mathsf{U} \times \mathsf{Y} \times \mathcal{U}$, respectively, and define a transition of the chain $\{(Y_k^{(1)}, U_k^{(1)}); k \in \mathbb{N}\}$ by Algorithm 1.

Remark 11. *In the expression (11) of α we assume implicitly that the families $\{S(y, u; \cdot); (y, u) \in \mathsf{Y} \times \mathsf{U}\}$ and $\{T(y, u, \hat{y}; \cdot); (y, u, \hat{y}) \in \mathsf{Y} \times \mathsf{U} \times \mathsf{Y}\}$ of probability measures are dominated by a fixed nonnegative measure and we denote by s and t the corresponding transition kernel densities, respectively. In some cases (see e.g. [10]) it may however happen (typically when some Dirac mass is involved) that these kernels are not dominated by a nonnegative measure; nevertheless, Algorithm 1 as well as Algorithm 2 defined below*

Algorithm 1 The *freeze* algorithm

Given $(Y_k^{(1)}, U_k^{(1)}) = (y, u)$,

(i) draw $\hat{Y} \sim S(y, u; \cdot)$ and call the outcome $\hat{y}$ (abbr. $\rightsquigarrow \hat{y}$),

(ii) draw $\hat{U} \sim T(y, u, \hat{y}; \cdot) \rightsquigarrow \hat{u}$,

(iii) set

$$(Y_{k+1}^{(1)}, U_{k+1}^{(1)}) \leftarrow \begin{cases} (\hat{y}, \hat{u}) & \text{with probability } \alpha(y, u, \hat{y}, \hat{u}) \\ & \qquad := 1 \wedge \dfrac{\pi^*(\hat{y}) r(\hat{y}, \hat{u}) s(\hat{y}, \hat{u}; y) t(\hat{y}, \hat{u}, y; u)}{\pi^*(y) r(y, u) s(y, u; \hat{y}) t(y, u, \hat{y}; \hat{u})}, \\ (y, u) & \text{otherwise.} \end{cases} \tag{11}$$

remain valid provided that the ratio in α is replaced by the corresponding Radon-Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(y, u, \hat{y}, \hat{u})$, where, in this case,

$$\mu(\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}\hat{y} \times \mathrm{d}\hat{u}) := \pi(\mathrm{d}y) R(y, \mathrm{d}u) S(y, u; \mathrm{d}\hat{y}) T(y, u, \hat{y}; \mathrm{d}\hat{u}),$$
$$\nu(\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}\hat{y} \times \mathrm{d}\hat{u}) := \pi(\mathrm{d}\hat{y}) R(\hat{y}, \mathrm{d}\hat{u}) S(\hat{y}, \hat{u}; \mathrm{d}y) T(\hat{y}, \hat{u}, y; \mathrm{d}u).$$

By applying Proposition 10 we deduce that the output $\{(Y_k^{(1)}, U_k^{(1)}); k \in \mathbb{N}\}$ is a π-reversible Markov chain. As a consequence, the sequence $\{Y_k^{(1)}; k \in \mathbb{N}\}$ targets, although it is not itself a Markov chain, the marginal distribution π^*. Note that the method requires the product $\pi^*(y) r(y, u) s(y, u; \hat{y}) t(y, u, \hat{y}; \hat{u})$ to be known at least up to a multiplicative constant to guarantee the computability of the acceptance probability α in (11).

Example 12 (Grouped-independence Metropolis-Hastings). *The Grouped-independence Metropolis-Hastings (GIMH) algorithm (see [1]) is used in situations where π^* is analytically intractable. In this algorithm, the quantity $\pi^*(y)$ is in the acceptance probability replaced by an importance sampling estimate*

$$\pi_N^*(y) := \frac{1}{N} \sum_{\ell=1}^{N} \frac{\bar{\pi}(y, v_\ell)}{q_y(v_\ell)}, \tag{12}$$

where $\bar{\pi}(y, v)$ is the density of some augmented target distribution $\bar{\pi}(\mathrm{d}y \times \mathrm{d}v)$ defined on the product space $(\mathsf{Y} \times \mathsf{V}, \mathcal{Y} \otimes \mathcal{V})$, known up to a normalizing constant and allowing π^* as marginal distribution, and $\{v_1, \ldots, v_N\}$ are i.i.d. draws from the proposal q_y. Denoting by $s(y, \cdot)$ the density used for proposing new candidates $\hat{y}$, one obtains the acceptance probability ratio

$$\frac{\pi_N^*(\hat{y}) s(\hat{y}, y)}{\pi_N^*(y) s(y, \hat{y})} = \frac{\pi^*(\hat{y}) r(\hat{y}, \hat{u}) s(\hat{y}, y) t(y, u)}{\pi^*(y) r(y, u) s(y, \hat{y}) t(\hat{y}, \hat{u})},$$

where $u := (v_1, \ldots, v_N)$ *and*

$$\pi^*(y)r(y,u) = \frac{1}{N}\sum_{\ell=1}^{N}\left(\bar{\pi}(y,v_\ell)\prod_{m\neq\ell} q_y(v_m)\right),$$

$$t(y,u) = \prod_{\ell=1}^{N} q_y(v_\ell).$$

Consequently, the GIMH algorithm can be perfectly cast into the framework of the freeze algorithm, with the auxiliary variable U *playing the role of the* N*-dimensional Monte Carlo sample and* $\mathsf{U} = \mathsf{V}^n$.

In the following we use Theorem 4 for comparing the performance of Algorithm 1 to that of different modifications of the same obtained in the cases where

(I) simulating R-transitions is feasible,
(II) simulating R-transitions is infeasible.

Case I: simulating R*-transitions is feasible.* In this case, an alternative to Algorithm 1 consists in letting $\{Y_k; k \in \mathbb{N}\}$ be the sequence $\{Y_k^{(2)}; k \in \mathbb{N}\}$ generated through Algorithm 2. Note that Algorithm 2 "refreshes", in

Algorithm 2 The *systematic refreshment* algorithm

Given $Y_k^{(2)} = y$,

(i) draw $U \sim R(y,\cdot) \rightsquigarrow u$,
(ii) draw $\hat{Y} \sim S(y,u;\cdot) \rightsquigarrow \hat{y}$,
(iii) draw $\hat{U} \sim T(y,u,\hat{y};\cdot) \rightsquigarrow \hat{u}$,
(iv) set $Y_{k+1}^{(2)} \leftarrow \begin{cases} \hat{y} & \text{with probability } \alpha(y,u,\hat{y},\hat{u}) \quad \text{(defined in (11))}, \\ y & \text{otherwise.} \end{cases}$

Step (i), systematically the second component of the Markov chain, which advocates Algorithm 2 to have better mixing properties than Algorithm 1. The main task of the present section is to establish rigorously this heuristics. The output $\{Y_k^{(2)}; k \in \mathbb{N}\}$ of Algorithm 2 is, on the contrary to $\{Y_k^{(1)}; k \in \mathbb{N}\}$, a Markov chain. It is not a classical Metropolis-Hastings Markov chain due to the auxiliary variables U and $\hat{U}$ that appear explicitly in the acceptance probability. However, as established in the following proposition, whose proof is found in Section A.2, the π-reversibility of $\{(Y_k^{(1)}, U_k^{(1)}); k \in \mathbb{N}\}$ implies π^*-reversibility of $\{Y_k^{(2)}; k \in \mathbb{N}\}$.

PROPOSITION 13. *The sequence* $\{Y_k^{(2)}; k \in \mathbb{N}\}$ *generated in Algorithm 2 is a* π^**-reversible Markov chain.*

EXAMPLE 14 (Randomized MCMC [10]). *In [10], the authors use the terminology* Randomized MCMC *(r-MCMC) for a* π^**-reversible Metropolis-Hastings chain* $\{Y_k; k \in \mathbb{N}\}$ *generated using a set of auxiliary variables* $\{U_k; k \in \mathbb{N}\}$ *with a particular expression of the acceptance probability. Although only one of these auxiliary variables is sampled at each time step, one may actually cast this approach into the framework of Algorithm 2 by creating artificially another auxiliary variable according to the deterministic kernel*

$$T(y, u, \hat{y}; \mathrm{d}\hat{u}) = \delta_{f(u)}(\mathrm{d}\hat{u}),$$

where f *is any continuously differentiable involution on* U*. Even though* T *is not dominated, it is possible to verify* (10) *using that* f *is an involution. We prove in Section B.1 that the r-MCMC algorithm is a special case of Algorithm 2 with this particular choice of* T *and with the general form of the acceptance probability described in Remark 11.*

EXAMPLE 15 (Generalized Multiple-try Metropolis [11]). *The* Generalized Multiple-try Metropolis *(GMTM)* algorithm *[11] is an extension of the* Multiple-try Metropolis Hastings algorithm *proposed in [7]. Given* $Y_k = y$*, one draws* n *i.i.d. possible moves* $V_1, \ldots, V_n$ *according to* $\check{R}(y, \cdot)$*. After this, a random index* J *taking the value* $j \in \{1, \ldots, n\}$ *with probability proportional to* $\omega(y, V_j)$ *is generated, whereupon a candidate is constructed as* $\hat{Y} = V_J$*. The candidate is then accepted with some probability that is computed using* n *additional random variables* $\hat{V}_1, \ldots, \hat{V}_n$*, where* $\hat{V}_1, \ldots, \hat{V}_{n-1}$ *are i.i.d. draws from* $\check{R}(\hat{y}, \cdot)$*, and* $\hat{V}_n$ *is set deterministically to* $\hat{V}_n = y$ *(see Section B.2 for more details concerning the acceptance probability). In Section B.2, Proposition 27, it is shown that the GMTM algorithm is in fact a special case of Algorithm 2 with* $U = (V_1, \ldots, V_{J-1}, V_{J+1}, \ldots, V_n)$ *and* $\hat{U} = (\hat{V}_1, \ldots, \hat{V}_{n-1})$*.*

When the function $k : (y, \hat{y}) \mapsto \int R(y, \mathrm{d}u) s(y, u; \hat{y})$ is known explicitly, one may obtain another π^*-reversible Markov chain by means of the classical Metropolis-Hastings ratio, i.e. we use again Algorithm 2 but replace the acceptance probability $\alpha(y, u, \hat{y}, \hat{u})$ by

$$\hat{\alpha}(y, \hat{y}) := 1 \wedge \frac{\pi^*(\hat{y}) k(\hat{y}, y)}{\pi^*(y) k(y, \hat{y})}. \tag{13}$$

The following proposition, which generalizes a similar result obtained in [10, Section 2.3] for the r-MCMC algorithm, shows, when combined with

[16, Theorem 4], that the asymptotic variance of the classical Metropolis-Hastings estimator is smaller than that of the estimator based on Algorithm 2.

PROPOSITION 16. *The Metropolis-Hastings kernel associated with the acceptance probability* (13) *is larger, in the sense of Definition 1, than the transition kernel associated with Algorithm 2.*

PROOF. Set

$$\mu(\mathrm{d}u \times \mathrm{d}\hat{u}) := \frac{R(y, \mathrm{d}u)s(y, u; \hat{y})T(y, u, \hat{y}; \mathrm{d}\hat{u})}{k(y, \hat{y})}$$

and note that μ is a probability measure. Hence, as the mapping $\mathbb{R} \ni v \mapsto 1 \wedge v$ is concave, Jensen's inequality implies that

$$\begin{aligned}
&\frac{\iint R(y, \mathrm{d}u)s(y, u; \hat{y})T(y, u, \hat{y}; \mathrm{d}\hat{u})\alpha(y, u, \hat{y}, \hat{u})}{k(y, \hat{y})\hat{\alpha}(y, \hat{y})} \\
&\quad = \frac{\iint \mu(\mathrm{d}u \times \mathrm{d}\hat{u})\alpha(y, u, \hat{y}, \hat{u})}{\hat{\alpha}(y, \hat{y})} \\
&\quad \leq \left(1 \wedge \iint \mu(\mathrm{d}u \times \mathrm{d}\hat{u})\frac{\pi(\hat{y}, \hat{u})s(\hat{y}, \hat{u}; y)t(\hat{y}, \hat{u}, y; u)}{\pi(y, u)s(y, u; \hat{y})t(y, u, \hat{y}; \hat{u})}\right) \Big/ \hat{\alpha}(y, \hat{y}) \\
&\quad = 1 \wedge \frac{\pi^*(\hat{y})k(\hat{y}, y)}{\pi^*(y)k(y, \hat{y})} \Big/ \hat{\alpha}(y, \hat{y}) = 1
\end{aligned}$$

(a similar technique was used in the proof of [2, Lemma 1]). The previous computation shows that the off-diagonal transition density function of the Metropolis-Hastings Markov chain associated with the acceptance probability (13) is larger than that of the chain in Algorithm 2. This completes the proof. □

However, in practice a closed-form expression of k is rarely available, which prevents the classical Metropolis-Hastings algorithm from being implemented. Thus, if the transition density r is known explicitly and can be sampled we have to choose between Algorithm 1 and Algorithm 2 for approximating π^*. The classical tools (such as the ordering in Definition 1) for comparing $\{Y_k^{(1)}; k \in \mathbb{N}\}$ and $\{Y_k^{(2)}; k \in \mathbb{N}\}$ cannot be applied here, since $\{Y_k^{(1)}; k \in \mathbb{N}\}$ is not even a Markov chain. Nevertheless, Theorem 4 allows the theoretical comparison between these two algorithms by embedding $\{Y_k^{(1)}; k \in \mathbb{N}\}$ and $\{Y_k^{(2)}; k \in \mathbb{N}\}$ into inhomogeneous π-reversible Markov chains. The construction, which will be carried through in full detail below, leads to the following result.

THEOREM 17. *Let* $\{Y_k^{(1)}; k \in \mathbb{N}\}$ *and* $\{Y_k^{(2)}; k \in \mathbb{N}\}$ *be sequences of random variables generated by Algorithm 1 and Algorithm 2, respectively, where* $(Y_0^{(1)}, U_0^{(1)}) \sim \pi$ *and* $Y_0^{(2)} \sim \pi^*$. *Then for all* $h \in \mathsf{L}^2(\pi^*)$ *satisfying*

$$\sum_{k=1}^{\infty} |\mathrm{Cov}(h(Y_0^{(i)}), h(Y_k^{(i)}))| < \infty \quad (i \in \{1,2\}) \tag{14}$$

it holds that

$$\lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n-1} h(Y_k^{(2)})\right) \le \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n-1} h(Y_k^{(1)})\right).$$

We preface the proof of Theorem 17 by the following lemma, which may serve as a basis for the comparison of *homogeneous* Markov chains evolving according to P_iQ_i (or Q_iP_i), $i \in \{0,1\}$, where P_i and Q_i, $i \in \{0,1\}$, are kernels satisfying(**A1**)on some product space.

LEMMA 18. *Let* P_i *and* Q_i, $i \in \{0,1\}$, *be kernels satisfying (**A1**) on* $(\mathsf{X}, \mathcal{X})$, *with* $\mathsf{X} = \mathsf{Y} \times \mathsf{U}$ *and* $\mathcal{X} = \mathcal{Y} \otimes \mathcal{U}$. *In addition, assume that for all* $(y,u) \in \mathsf{X}$,

$$P_i(y, u; \{y\} \times \mathsf{U}) = 1 \quad (i \in \{0,1\}). \tag{15}$$

Then for all $f \in \mathsf{L}^2(\pi)$ *depending on only the first argument (i.e.* $f(y,u) = h(y)$ *for some* h*) and such that*

$$\sum_{n=1}^{\infty} |\langle f, (P_iQ_i)^n f\rangle| < \infty \quad (i \in \{0,1\}) \tag{16}$$

it holds that

$$v(f, P_1Q_1) = v(f, Q_1P_1) \le v(f, P_0Q_0) = v(f, Q_0P_0).$$

REMARK 19. *The assumption* (15) *is essential in Lemma 18. Indeed, let* $\mathsf{X} = \{-1, 1\}$ *and* $\pi(\{1\}) = \pi(\{-1\}) = 1/2$, *and define the kernels* $P_0(x, \mathrm{d}x) = \delta_x(\mathrm{d}x')$, $Q_0(x, \mathrm{d}x') = \varepsilon\pi(\mathrm{d}x') + (1-\varepsilon)\delta_{-x}(\mathrm{d}x')$ *for some* $\varepsilon \in (0,1)$, $P_1(x, \mathrm{d}x') = \pi(\mathrm{d}x')$, *and* $Q_1 = Q_0$. *Then the kernels* P_i *and* Q_i, $i \in \{0,1\}$, *satisfy(**A1**), and consequently Theorem 4 applies to the inhomogeneous chains evolving alternatingly according to the same. However, the similar result does not hold true for chains evolving according to the product kernels* P_iQ_i *and* Q_iP_i, $i \in \{0,1\}$, *as*

$$v(f, P_0Q_0) = v(f, Q_0P_0) = \frac{\varepsilon}{2-\varepsilon} < 1 = v(f, P_1Q_1) = v(f, Q_1P_1),$$

with f *being the identity mapping on* X.

Proof of Lemma 18. Define Markov chains $\{X_k^{(i)}; k \in \mathbb{N}\}$, $i \in \{0,1\}$, evolving as

$$\cdots \xrightarrow{Q_i} X_{2k}^{(i)} = \begin{pmatrix} Y_k^{(i)} \\ U_k^{(i)} \end{pmatrix} \xrightarrow{P_i} X_{2k+1}^{(i)} = \begin{pmatrix} \check{Y}_k^{(i)} \\ \check{U}_k^{(i)} \end{pmatrix} \xrightarrow{Q_i} X_{2k+2}^{(i)} = \begin{pmatrix} Y_{k+1}^{(i)} \\ U_{k+1}^{(i)} \end{pmatrix} \xrightarrow{P_i} \cdots$$

with $X_0^{(i)} \sim \pi$. By construction,

$$\begin{aligned} (17) \quad & \sum_{k=1}^{\infty} \Big(|\mathrm{Cov}(f(X_0^{(i)}), f(X_k^{(i)}))| + |\mathrm{Cov}(f(X_1^{(i)}), f(X_{k+1}^{(i)}))| \Big) \\ & \qquad = \pi f^2 - \pi^2 f + 4 \sum_{k=1}^{\infty} |\mathrm{Cov}(h(Y_0^{(i)}), h(Y_k^{(i)}))| < \infty \quad (i \in \{0,1\}), \end{aligned}$$

where finiteness follows from the assumption (16). Moreover, for all $n \in \mathbb{N}^*$ and $i \in \{0,1\}$,

$$\mathrm{Var}\left(\sum_{k=0}^{n-1} h(Y_k^{(i)}) \right) = \mathrm{Var}\left(\sum_{k=0}^{n-1} h(\check{Y}_k^{(i)}) \right) = \frac{1}{4} \mathrm{Var}\left(\sum_{k=0}^{2n-1} f(X_k^{(i)}) \right),$$

which implies, by (17),

$$v(f, P_i Q_i) = v(f, Q_i P_i) = \frac{1}{2} \lim_{n \to \infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n} f(X_k^{(i)}) \right) \quad (i \in \{0,1\}).$$

Finally, by (17) we may now apply Theorem 4 to the chains $\{X_k^{(i)}; k \in \mathbb{N}\}$, $i \in \{0,1\}$, which establishes immediately the statement of the lemma. □

Proof of Theorem 17. We introduce the kernels

- $P_1(y, u; \mathrm{d}y' \times \mathrm{d}u') = \delta_{(y,u)}(\mathrm{d}y' \times \mathrm{d}u')$,
- $P_2(y, u; \mathrm{d}y' \times \mathrm{d}u') = \delta_y(\mathrm{d}y') R(y, \mathrm{d}u')$,
- $Q_1 = Q_2$ being defined implicitly as the transition kernel associated with the *freeze algorithm* (Algorithm 1).

It can be checked readily that the two sequences $\{Y_k^{(1)}; k \in \mathbb{N}\}$ and $\{Y_k^{(2)}; k \in \mathbb{N}\}$ generated by Algorithm 1 and Algorithm 2, respectively, have indeed the same distributions as the marginal processes (with respect to the first component) of homogeneous chains evolving according to the products $P_1 Q_1$ and $P_2 Q_2$, respectively. In addition, all kernels P_i and Q_i, $i \in \{1,2\}$, are π-reversible, as

- P_1 is reversible with respect to any probability measure (in particular, it is π-reversible),
- P_2 is π-reversible as a Gibbs-sampler sub-step transition kernel,
- $Q_1 = Q_2$ is π-reversible as a classical Metropolis-Hastings transition kernel.

Since P_1 has no off-diagonal component, it holds that $P_2 \succeq P_1$; moreover, trivially, $Q_2 = Q_1 \succeq Q_1$. Thus, we may complete the proof by applying Lemma 18 to the function $f(y,u) = h(y)$, for which the condition (16) is satisfied (by (14)). □

Case II: simulating R-transitions is infeasible. Pseudo-marginal algorithms (see [1] and [2]) are implemented using a Markov kernel $\check{R}$ on $\mathsf{Y} \times \mathcal{U}$ and a family $\{w_u; u \in \mathsf{U}\}$ of real-valued nonnegative functions on Y such that $\int \check{R}(y, \mathrm{d}u) w_u(y) = 1$ for all $y \in \mathsf{Y}$. We denote by $\check{r}$ the transition density of the kernel $\check{R}$ with respect to some dominating measure. Note that $R(y, \mathrm{d}u) := \check{R}(y, \mathrm{d}u) w_u(y)$ is a Markov transition kernel as well. The problem at hand is to sample the target distribution

$$\pi(\mathrm{d}y \times \mathrm{d}u) := \pi^*(\mathrm{d}y) R(y, \mathrm{d}u) = \pi^*(\mathrm{d}y) \check{R}(y, \mathrm{d}u) w_u(y)$$

under the assumption that

- for all $(y,u) \in \mathsf{Y} \times \mathsf{U}$, $\pi^*(y)\check{r}(y,u)w_u(y)$ is known up to a normalizing constant,
- for all $y \in \mathsf{Y}$, $\check{R}(y, \cdot)$ can be sampled from.

The particular case where $w_u(y) = 1$ for all $(y,u) \in \mathsf{Y} \times \mathsf{U}$ was discussed in the previous section, and we now turn to the case $w_u(y) \neq 1$ (i.e. sampling directly from R is infeasible). The solution provided by pseudo-marginal algorithms consists in replacing, in Algorithm 2, the operation (i) by the sampling $U \sim \check{R}(y, \cdot)$, and the computing the acceptance probability α (as defined in (11)) via the formula

$$\alpha(y, u, \hat{y}, \hat{u}) := 1 \wedge \frac{\pi^*(\hat{y})\check{r}(\hat{y}, \hat{u}) w_{\hat{u}}(\hat{y}) s(\hat{y}, \hat{u}; y) t(\hat{y}, \hat{u}, y; u)}{\pi^*(y)\check{r}(y, u) w_u(y) s(y, u; \hat{y}) t(y, u, \hat{y}; \hat{u})}.$$

The output of this algorithm, which will be referred to as the *noisy algorithm* in the following, is typically not—on the contrary to Algorithm 2—π^*-reversible due to the replacement of R by $\check{R}$. This justifies the denomination. However, when w is close to unity the noisy algorithm is close to Algorithm 2, which is, according to Theorem 17, more efficient than Algorithm 1 in terms of asymptotic variance.

Example 20 (Monte Carlo within Metropolis). *The Monte Carlo within Metropolis algorithm (MCWM; see [1]) resembles closely the GIMH algorithm (see Example 12), however with the important difference that the importance sampling estimates $\pi_N^*(Y_k)$ (given by (12)) are* not *stored and propagated through the algorithm along with the Y_k-values. Instead each estimate of the marginal density is recomputed using a "fresh" MC sample before the calculation of the acceptance probability. Thus, the MCWM algorithm can be cast into the framework of the noisy algorithm with $T = \check{R}$ and with the auxiliary variables U and $\hat{U}$ playing the roles of N-dimensional Monte Carlo samples.*

Considering this, we now propose a novel algorithm which will be referred to as the *random refreshment algorithm* and which is a hybrid between Algorithm 2 and the noisy algorithm. This novel algorithm, which is described in Algorithm 3 below, targets *exactly* π^* and turns out to be more efficient than Algorithm 1.

Algorithm 3 The *random refreshment* algorithm

Given $(Y_k^{(3)}, U_k^{(3)}) = (y, u)$,

(i) (i.1) draw $U' \sim \check{R}(y, \cdot) \rightsquigarrow u'$,

(i.2) set

$$\check{U} \leftarrow \begin{cases} u' & \text{with probability } \varrho(y, u, u') := 1 \wedge \dfrac{w_{u'}(y)}{w_u(y)}, \\ u & \text{otherwise,} \end{cases} \rightsquigarrow \check{u}. \tag{18}$$

(ii) draw $\hat{Y} \sim S(y, \check{u}; \cdot) \rightsquigarrow \hat{y}$,

(iii) draw $\hat{U} \sim T(y, \check{u}, \hat{y}; \cdot) \rightsquigarrow \hat{u}$,

(iv) set $(Y_{k+1}^{(3)}, U_{k+1}^{(3)}) \leftarrow \begin{cases} (\hat{y}, \hat{u}) & \text{with probability } \alpha(y, \check{u}, \hat{y}, \hat{u}), \\ (y, \check{u}) & \text{otherwise.} \end{cases}$

In Step (i) in Algorithm 3, the auxiliary variable $\check{U}$ can be either "refreshed", i.e. replaced by a new candidate U', or kept at the previous state $U_k^{(3)}$ according to an acceptance probability that turns out to be a standard Metropolis-Hastings acceptance probability (which will be seen in the proof of Theorem 21 below). This allows the desired distribution π as the target distribution of $\{(Y_k^{(3)}, U_k^{(3)}); k \in \mathbb{N}\}$. In comparison, the noisy algorithm described above differs only from Algorithm 3 by Step (i), in that the new candidate is always accepted in the noisy algorithm. This "systematic refreshment" makes actually the noisy algorithm imprecise in the sense

that π is no longer the target distribution except when $w_u(y) = 1$ for all $(y, u) \in \mathsf{Y} \times \mathsf{U}$ (in which case $\varrho(y, u, u')$ in (18) becomes identically equal to unity and Algorithm 3 translates into Algorithm 2. Compared to Algorithm 1, Step (i) allows the second component to be refreshed randomly according to the probability $\varrho(y, u, \check{u})$ whereas this component remains unchanged in Algorithm 1. Thus, in conformity with Algorithm 2, it is likely that Algorithm 3 has better mixing properties than Algorithm 1. That this is indeed the case may be established by reapplying the embedding technique developed in the previous part. This is formalized in the following.

THEOREM 21. *Let $\{Y_k^{(1)}; k \in \mathbb{N}\}$ and $\{Y_k^{(3)}; k \in \mathbb{N}\}$ be the sequences of random variables generated by Algorithm 1 and Algorithm 3, respectively, where $(Y_0^{(i)}, U_0^{(i)}) \sim \pi$, $i \in \{1, 3\}$. Then the following holds true.*

(i) The output of Algorithm 3 is π-reversible.
(ii) For all $h \in \mathsf{L}^2(\pi^*)$ satisfying

$$\sum_{k=1}^{\infty} |\mathrm{Cov}(h(Y_0^{(i)}), h(Y_k^{(i)}))| < \infty \quad (i \in \{1, 3\})$$

it holds that

$$\lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n-1} h(Y_k^{(3)})\right) \leq \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{k=0}^{n-1} h(Y_k^{(1)})\right).$$

PROOF. Let the kernels P_1 and Q_1 be defined as in the proof of Theorem 17 and introduce furthermore

- P_3 defined implicitly by the transition $(Y_k^{(3)}, U_k^{(3)}) \to (Y_k^{(3)}, \check{U})$ according to Step (i) in Algorithm 3 (note that the first component is held fixed throughout the transition),
- $Q_3 = Q_1$.

In conformity with the proof of Theorem 17 it can be checked readily that the two sequences $\{Y_k^{(1)}; k \in \mathbb{N}\}$ and $\{Y_k^{(3)}; k \in \mathbb{N}\}$ generated by Algorithm 1 and Algorithm 3, respectively, have indeed the same distributions as the marginal processes (with respect to the first component) of homogeneous chains evolving according to the products P_1Q_1 and P_3Q_3, respectively. The π-reversibility of the kernels P_1 and $Q_1 = Q_3$ was established in the proof of Theorem 17. To verify π-reversibility of P_3 as well, note that P_3 is a Metropolis-Hastings kernel associated with the target distribution π, whose acceptance probability includes a Radon-Nikodym derivative of the type

given in Proposition 10; it is therefore π-reversible. Indeed, note that P_3 updates only the second component according to $\check{R}(y, \mathrm{d}u')$ with the acceptance probability $\varrho(y, u, u')$. Assuming first that $\check{R}$ is dominated and denoting by $\check{r}$ its transition density, we have

$$\varrho(y, u, u') = 1 \wedge \frac{w_{u'}(y)}{w_u(y)} = 1 \wedge \frac{\pi(y, u')\check{r}(y, u)}{\pi(y, u)\check{r}(y, u')},$$

where $\pi(y, u) = \pi^*(y)\check{r}(y, u)w_u(y)$ in the density of the target π. This shows that $\varrho(y, u, u')$ is indeed the acceptance probability of a Metropolis-Hastings Markov chain targeting π, with proposal kernel $\check{R}(y, \mathrm{d}u')\delta_y(\mathrm{d}y')$; the π-reversibility of P_3 follows. The proof can be adapted easily to the case where $\check{R}$ is not dominated. As a consequence, the product P_3Q_3 is also π-reversible, which establishes the statement (i) of the theorem. Finally, since P_1 has zero mass on the off-diagonal, it holds that $P_3 \succeq P_1$ and, clearly, $Q_3 = Q_1 \succeq Q_1$. The proof of (ii) is now concluded by applying Lemma 18 along the lines of the proof of Theorem 17. □

5. Proof of Theorem 4. We preface the proof of Theorem 4 by some preliminary lemmas.

LEMMA 22. *Assume that $P_1, P_2, \ldots, P_n$ are π-reversible Markov transition kernels. Then, for all $(f, g) \in \mathsf{L}^2(\pi) \times \mathsf{L}^2(\pi)$,*

$$\langle f, P_1P_2 \cdots P_n g\rangle = \langle P_n \cdots P_2P_1 f, g\rangle .$$

PROOF. As each P_ℓ is π-reversible, it holds that $\langle P_\ell f, g\rangle = \langle f, P_\ell g\rangle$ for all $(f, g) \in \mathsf{L}^2(\pi) \times \mathsf{L}^2(\pi)$ and $\ell \in \{1, \ldots, n\}$. Applying repeatedly this relation yields

$$\begin{aligned}\langle f, P_1P_2 \cdots P_n g\rangle &= \langle P_1 f, P_2 \cdots P_n g\rangle \\ &= \langle P_2P_1 f, P_3 \cdots P_n g\rangle \cdots = \langle P_n \cdots P_2P_1 f, g\rangle .\end{aligned}$$

□

LEMMA 23. *Let P and Q be Markov transition kernels on $(\mathsf{X}, \mathcal{X})$ such that $\pi P = \pi Q = \pi$ and let $\{X_k; k \in \mathbb{N}\}$ be a Markov chain evolving as*

$$X_0 \xrightarrow{P} X_1 \xrightarrow{Q} X_2 \xrightarrow{P} X_3 \xrightarrow{Q} \cdots$$

with initial distribution $X_0 \sim \pi$. Then, for all $f \in \mathsf{L}^2(\pi)$ such that

$$\sum_{k=1}^{\infty} \left(|\mathrm{Cov}(f(X_0), f(X_k))| + |\mathrm{Cov}(f(X_1), f(X_{k+1}))|\right) < \infty, \tag{19}$$

the limit, as n tends to infinity, of $n^{-1}\mathrm{Var}(\sum_{k=0}^{n-1} f(X_k))$ exists, and

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left(\sum_{k=0}^{n-1} f(X_k)\right) = \pi f^2 - \pi^2 f + \sum_{k=1}^{\infty} \mathrm{Cov}(f(X_0), f(X_k)) + \sum_{k=1}^{\infty} \mathrm{Cov}(f(X_1), f(X_{k+1})). \tag{20}$$

Proof. As covariances are symmetric,

$$\frac{1}{n}\mathrm{Var}\left(\sum_{k=0}^{n-1} f(X_k)\right) = \pi f^2 - \pi^2 f + 2n^{-1} \sum_{0\le i<j\le n-1} \mathrm{Cov}(f(X_i), f(X_j)).$$

We now consider the limit, as n tends to infinity, of the last term on the right hand side. Let $\mathcal{E}$ and $\mathcal{O}$ denote the two complementary subsets of $\mathbb{N}$ consisting of the even and odd numbers, respectively. For all $(i,j) \in \mathbb{N}^2$ such that $i < j$, we have

$$\mathrm{Cov}(f(X_i), f(X_j)) = \begin{cases} \mathrm{Cov}(f(X_0), f(X_{j-i})) & \text{if } i \in \mathcal{E}, \\ \mathrm{Cov}(f(X_1), f(X_{j-i+1})) & \text{if } i \in \mathcal{O}. \end{cases}$$

This implies that

$$n^{-1} \sum_{\substack{0\le i<j\le n-1 \\ i\in\mathcal{E}}} \mathrm{Cov}(f(X_i), f(X_j)) = \sum_{k=1}^{n-1} n^{-1} \left(\left\lfloor \frac{n-1-k}{2} \right\rfloor + 1 \right) \mathrm{Cov}(f(X_0), f(X_k))$$

and

$$n^{-1} \sum_{\substack{0\le i<j\le n-1 \\ i\in\mathcal{O}}} \mathrm{Cov}(f(X_i), f(X_j)) = \sum_{k=1}^{n-2} n^{-1} \left(\left\lfloor \frac{n-2-k}{2} \right\rfloor + 1 \right) \mathrm{Cov}(f(X_1), f(X_{k+1})).$$

Under (19), the dominated convergence theorem applies, which provides that the limit, as n goes to infinity, of $n^{-1}\mathrm{Var}(\sum_{k=0}^{n-1} f(X_k))$ exists and is equal to (20). □

LEMMA 24. *Let P_i and Q_i, $i \in \{0,1\}$, be π-reversible Markov kernels on $(\mathsf{X}, \mathcal{X})$ such that $P_0 \succcurlyeq P_1$ and $Q_0 \succcurlyeq Q_1$. For all $n \in \mathbb{N}$ and $i \in \{0,1\}$, denote by $R_n^{(i)}$ the Markov kernel $R_n^{(i)} := P_i \mathbb{1}_{\mathcal{E}}(n) + Q_i \mathbb{1}_{\mathcal{O}}(n)$. In addition, let $f \in \mathsf{L}^2(\pi)$ be such that for $i \in \{0,1\}$,*

$$\sum_{k=1}^{\infty} \left| \left\langle f, R_0^{(i)} \cdots R_{k-1}^{(i)} f \right\rangle \right| < \infty. \tag{21}$$

Then for all $\lambda \in (0,1)$,

$$\sum_{k=1}^{\infty} \lambda^k \left(\left\langle f, R_0^{(1)} \cdots R_{k-1}^{(1)} f \right\rangle + \left\langle f, R_1^{(1)} \cdots R_k^{(1)} f \right\rangle \right) \\ \leq \sum_{k=1}^{\infty} \lambda^k \left(\left\langle f, R_0^{(0)} \cdots R_{k-1}^{(0)} f \right\rangle + \left\langle f, R_1^{(0)} \cdots R_k^{(0)} f \right\rangle \right).$$

PROOF. For all $n \in \mathbb{N}$ and all $\alpha \in (0,1)$, define $R_n^{(\alpha)} := (1-\alpha) R_n^{(0)} + \alpha R_n^{(1)}$. In addition, set, for $\lambda \in (0,1)$, $K_\lambda(\alpha) := K_\lambda^{(\mathcal{E})}(\alpha) + K_\lambda^{(\mathcal{O})}(\alpha)$, where

$$K_\lambda^{(\mathcal{E})}(\alpha) := \sum_{k=1}^{\infty} \lambda^k \left\langle f, R_0^{(\alpha)} \cdots R_{k-1}^{(\alpha)} f \right\rangle,$$
$$K_\lambda^{(\mathcal{O})}(\alpha) := \sum_{k=1}^{\infty} \lambda^k \left\langle f, R_1^{(\alpha)} \cdots R_k^{(\alpha)} f \right\rangle.$$

Now, fix a distinguished $\lambda \in (0,1)$; we want show that for all $\alpha \in [0,1]$,

$$\frac{\mathrm{d}K_\lambda}{\mathrm{d}\alpha}(\alpha) \leq 0. \tag{22}$$

Thus, we start with differentiating $K_\lambda^{(\mathcal{E})}$:

$$\frac{\mathrm{d}K_\lambda^{(\mathcal{E})}}{\mathrm{d}\alpha}(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{k=1}^{\infty} \lambda^k \left\langle f, R_0^{(\alpha)} \cdots R_{k-1}^{(\alpha)} f \right\rangle. \tag{23}$$

To interchange $\frac{\mathrm{d}}{\mathrm{d}\alpha}$ and $\sum_{k=1}^{\infty}$ in the previous equation, we first note that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \left\langle f, R_0^{(\alpha)} \cdots R_{k-1}^{(\alpha)} f \right\rangle = \sum_{\ell=0}^{k-1} \frac{\partial}{\partial \alpha_\ell} \left\langle f, R_0^{(\alpha_0)} \cdots R_{k-1}^{(\alpha_{k-1})} f \right\rangle \Big|_{(\alpha_0,\ldots,\alpha_{k-1})=(\alpha,\ldots,\alpha)} \\ = \sum_{\ell=0}^{k-1} \left\langle f, R_{0\nearrow \ell-1}^{(\alpha)} (R_\ell^{(1)} - R_\ell^{(0)}) R_{\ell+1\nearrow k-1}^{(\alpha)} f \right\rangle,$$

where $R^{(\alpha)}_{s\nearrow t} := R^{(\alpha)}_s R^{(\alpha)}_{s+1} \cdots R^{(\alpha)}_t$ for $s \le t$ and $R^{(\alpha)}_{s\nearrow t} := \mathrm{id}$ otherwise. By (1), $\|R^{(\alpha)}_n\| \le 1$, which implies that $\sup_{\alpha\in[0,1]} |\frac{\mathrm{d}}{\mathrm{d}\alpha}\langle f, R^{(\alpha)}_0 \cdots R^{(\alpha)}_{k-1} f\rangle| \le 2k\pi(f^2)$. Thus, as $\sum_{k=1}^\infty \lambda^k k < \infty$ we may interchange, in (23), $\frac{\mathrm{d}}{\mathrm{d}\alpha}$ and $\sum_{k=1}^\infty$, yielding

$$\frac{\mathrm{d}K^{(\mathcal{E})}_\lambda}{\mathrm{d}\alpha}(\alpha) = \sum_{k=1}^\infty \lambda^k \sum_{\ell=0}^{k-1} \left\langle f, R^{(\alpha)}_{0\nearrow \ell-1}(R^{(1)}_\ell - R^{(0)}_\ell) R^{(\alpha)}_{\ell+1\nearrow k-1} f \right\rangle .$$

Similarly, it can be established that

$$\frac{\mathrm{d}K^{(\mathcal{O})}_\lambda}{\mathrm{d}\alpha}(\alpha) = \sum_{k=1}^\infty \lambda^k \sum_{\ell=1}^{k} \left\langle f, R^{(\alpha)}_{1\nearrow \ell-1}(R^{(1)}_\ell - R^{(0)}_\ell) R^{(\alpha)}_{\ell+1\nearrow k} f \right\rangle .$$

We now apply Lemma 22 to the two previous sums. For this purpose we will use the following notation $R^{(\alpha)}_{s\searrow t} := R^{(\alpha)}_s R^{(\alpha)}_{s-1} \cdots R^{(\alpha)}_t$ for $s \ge t$ and $R^{(\alpha)}_{s\searrow t} := \mathrm{id}$ otherwise. Then,

$$\begin{aligned}
&\frac{\mathrm{d}K_\lambda}{\mathrm{d}\alpha}(\alpha) \\
&= \sum_{k=1}^\infty \lambda^k \Bigg\{ \sum_{\ell=0}^{k-1} \left\langle R^{(\alpha)}_{\ell-1\searrow 0} f, (R^{(1)}_\ell - R^{(0)}_\ell) R^{(\alpha)}_{\ell+1\nearrow k-1} f \right\rangle \\
&\quad + \sum_{\ell=1}^{k} \left\langle R^{(\alpha)}_{\ell-1\searrow 1} f, (R^{(1)}_\ell - R^{(0)}_\ell) R^{(\alpha)}_{\ell+1\nearrow k} f \right\rangle \Bigg\} \\
&= \sum_{\ell=0}^\infty \sum_{m=0}^\infty \lambda^{\ell+m+1} \left\langle R^{(\alpha)}_{\ell-1\searrow 0} f, (R^{(1)}_\ell - R^{(0)}_\ell) R^{(\alpha)}_{\ell+1\nearrow \ell+m} f \right\rangle \\
&\quad + \sum_{\ell=1}^\infty \sum_{m=1}^\infty \lambda^{\ell+m-1} \left\langle R^{(\alpha)}_{\ell-1\searrow 1} f, (R^{(1)}_\ell - R^{(0)}_\ell) R^{(\alpha)}_{\ell+1\nearrow \ell+m-1} f \right\rangle .
\end{aligned}$$

Now, note that $R^{(\alpha)}_n = R^{(\alpha)}_{n'}$ for all $(n, n') \in \mathcal{O}^2$ and $R^{(\alpha)}_m = R^{(\alpha)}_{m'}$ for all $(m, m')^2 \in \mathcal{E}^2$; hence, separating, in the two previous sums, odd and even

indices ℓ provides

$$\begin{aligned}
&\frac{\mathrm{d}K_\lambda}{\mathrm{d}\alpha}(\alpha)\\
&=\sum_{\ell\in\mathcal{E}}\sum_{m=0}^{\infty}\lambda^{\ell+m+1}\left\langle R^{(\alpha)}_{1\nearrow\ell}f,(R^{(1)}_0-R^{(0)}_0)R^{(\alpha)}_{1\nearrow m}f\right\rangle\\
&\quad+\sum_{\ell\in\mathcal{E}\setminus\{0\}}\sum_{m=1}^{\infty}\lambda^{\ell+m-1}\left\langle R^{(\alpha)}_{1\nearrow\ell-1}f,(R^{(1)}_0-R^{(0)}_0)R^{(\alpha)}_{1\nearrow m-1}f\right\rangle\\
&\quad+\sum_{\ell\in\mathcal{O}}\sum_{m=0}^{\infty}\lambda^{\ell+m+1}\left\langle R^{(\alpha)}_{0\nearrow\ell-1}f,(R^{(1)}_1-R^{(0)}_1)R^{(\alpha)}_{0\nearrow m-1}f\right\rangle\\
&\quad+\sum_{\ell\in\mathcal{O}}\sum_{m=1}^{\infty}\lambda^{\ell+m-1}\left\langle R^{(\alpha)}_{0\nearrow\ell-2}f,(R^{(1)}_1-R^{(0)}_1)R^{(\alpha)}_{0\nearrow m-2}f\right\rangle.
\end{aligned}$$

Finally, by combining the even and the odd sums,

$$\begin{aligned}
&\frac{\mathrm{d}K_\lambda}{\mathrm{d}\alpha}(\alpha)\\
&=\left\langle\sum_{\ell=0}^{\infty}\lambda^\ell R^{(\alpha)}_{1\nearrow\ell}f,(R^{(1)}_0-R^{(0)}_0)\sum_{m=0}^{\infty}\lambda^m R^{(\alpha)}_{1\nearrow m}f\right\rangle\\
&\quad+\left\langle\sum_{\ell=0}^{\infty}\lambda^\ell R^{(\alpha)}_{0\nearrow\ell-1}f,(R^{(1)}_1-R^{(0)}_1)\sum_{m=0}^{\infty}\lambda^m R^{(\alpha)}_{0\nearrow m-1}f\right\rangle.
\end{aligned}$$

Since $R^{(1)}_n \succcurlyeq R^{(0)}_n$, the operator $R^{(0)}_n - R^{(1)}_n$ is nonnegative on $\mathsf{L}^2(\pi)$ (by [16, Lemma 3]), and for all $f \in \mathsf{L}^2(\pi)$ it holds that $\langle f,(R^{(1)}_n - R^{(0)}_n)f\rangle \le 0$. This shows (22), which implies that the function $\alpha \mapsto K_\lambda(\alpha)$ is non-increasing on $(0,1)$. The proof is complete. □

Proof of Theorem 4. According to Lemma 23, for all functions $f \in \mathsf{L}^2(\pi)$ and $i \in \{0,1\}$,

(24)
$$v^{(i)}(f) = \pi f^2 - \pi^2 f + \sum_{k=1}^{\infty}\left(\mathrm{Cov}(f(X^{(i)}_0), f(X^{(i)}_k)) + \mathrm{Cov}(f(X^{(i)}_1), f(X^{(i)}_{k+1}))\right).$$

For the kernels P_i and Q_i, $i \in \{0,1\}$, in the statement of the theorem, let $\{R^{(i)}_k; k \in \mathbb{N}\}$, $i \in \{0,1\}$, be defined as in Lemma 24, which then implies

that for all $\lambda \in (0,1)$,

$$\sum_{k=1}^{\infty}\left(\lambda^k \mathrm{Cov}(f(X_0^{(1)}), f(X_k^{(1)})) + \lambda^k \mathrm{Cov}(f(X_1^{(1)}), f(X_{k+1}^{(1)}))\right) \leq \sum_{k=1}^{\infty}\left(\lambda^k \mathrm{Cov}(f(X_0^{(0)}), f(X_k^{(0)})) + \lambda^k \mathrm{Cov}(f(X_1^{(0)}), f(X_{k+1}^{(0)}))\right). \tag{25}$$

We conclude the proof by letting λ tend to one on each side of the previous inequality. Under (5), we may, by the dominated convergence theorem, interchange limits with summation, which establishes the inequality (25) also in the case $\lambda = 1$. Combining this with (24) completes the proof. □

6. Conclusion. In this paper, we have extended successfully the theoretical framework proposed in [12] and [16] as a means of comparing the asymptotic variance of sample path averages for different Markov chains and, consequently, the efficiency of different MCMC algorithms to the context of inhomogeneous Markov chains evolving alternatingly according to two different Markov transition kernels. It turned out that this configuration covers, although not apparently, several popular MCMC algorithms such as Randomized MCMC [10], Multiple-try Metropolis [7] and its generalization [11], and the pseudo-marginal algorithms [1, 2]. It should be remarked however that our results do not take possible additional computational cost into consideration, which may be of importance in practical applications. While these algorithms are inapproachable for the standard tools provided in [12] and [16], our results allow, without heavy technical developments, rigorous theoretical justifications advocating the use of these algorithms. As illustrated by our novel *random refreshment* algorithm in the context of pseudo-marginal algorithms, the results of the present paper can also be used for designing new algorithms and improving, in terms of asymptotic variance, existing ones.

Acknowledgement. We thank the anonymous referees for insightful comments that improved significantly the presentation of the paper. A special thank goes to the referee who provided the two counterexamples in Remark 6 and Remark 19.

APPENDIX A: PROOFS OF PROPOSITIONS 8 AND 12

A.1. Proof of Proposition 9. First, set $\xi = P^n(x,\cdot) - \pi$; then by Jensen's inequality,

$$\|\xi\|_{V^{1/2}} = |\xi|(\mathsf{X})\frac{|\xi|(V^{1/2})}{|\xi|(\mathsf{X})} \le |\xi|(\mathsf{X})\left(\frac{|\xi|(V)}{|\xi|(\mathsf{X})}\right)^{1/2} = |\xi|^{1/2}(\mathsf{X})\,\|\xi\|_V^{1/2}\,,$$

and since $|\xi|(\mathsf{X}) \le 2$,

$$\|P^n(x,\cdot) - \pi\|_{V^{1/2}} \le (2C\rho^n V(x))^{1/2}\,. \tag{26}$$

Now, without loss of generality we may assume that $\pi f = 0$, $|f|_{V^{1/2}} \le 1$, and $|Pf|_{V^{1/2}} \le 1$. Then applying (26) yields for all $x \in \mathsf{X}$,

$$|(PQ)^n f(x)| \le (2C\rho^n V(x))^{1/2}\,.$$

Hence, for all $n \in \mathbb{N}$,

$$\begin{aligned}|\mathrm{Cov}(f(X_0), f(X_{2n}))| &= |\mathbb{E}(f(X_0)(PQ)^n f(X_0))| \\ &\le (2C\rho^n)^{1/2}\,\mathbb{E}\left(|f(X_0)|V^{1/2}(X_0)\right) \le (2C\rho^n)^{1/2}\,\pi V.\end{aligned}$$

In the same way, for all $n \ge 0$,

$$|\mathrm{Cov}(f(X_0), f(X_{2n+1}))| = |\mathbb{E}(f(X_0)(PQ)^n Pf(X_0))| \le (2C\rho^n)^{1/2}\,\pi V.$$

By applying successively the Cauchy-Schwarz and Jensen inequalities we obtain

$$\mathbb{E}\left(|f(X_1)|QV^{1/2}(X_1)\right) \le \left[\mathbb{E}\left(f^2(X_1)\right)\mathbb{E}\left(QV(X_1)\right)\right]^{1/2} \le \pi V,$$

where the last inequality follows from $f^2 \le V$ and $\pi P = \pi Q = \pi$. This implies that for all $n \in \mathbb{N}^*$,

$$\begin{aligned}|\mathrm{Cov}(f(X_1), f(X_{2n}))| &= |\mathbb{E}(f(X_1)Q(PQ)^{n-1}f(X_1))| \\ &\le \left(2C\rho^{n-1}\right)^{1/2}\mathbb{E}\left(|f(X_1)|QV^{1/2}(X_1)\right) \le \left(2C\rho^{n-1}\right)^{1/2}\pi V.\end{aligned}$$

In the same way, for all $n \in \mathbb{N}^*$ we have, using that $|Pf(x)| \le V^{1/2}(x)$,

$$|\mathrm{Cov}(f(X_1), f(X_{2n+1}))| = |\mathbb{E}(f(X_1)Q(PQ)^{n-1}Pf(X_1))| \le \left(2C\rho^{n-1}\right)^{1/2}\pi V.$$

The statement of the proposition follows.

A.2. Proof of Proposition 13. Let K be the transition kernel of the Markov chain $\{Y_k^{(2)}; k \in \mathbb{N}\}$, i.e. for all $f \in \mathcal{F}(\mathcal{Y})$,

$$\int f(y')K(y, \mathrm{d}y') = f(y)\beta(y) + \int f(y')R(y, \mathrm{d}u)S(y, u; \mathrm{d}y')T(y, u, y'; \mathrm{d}u')\alpha(y, u, y', u'),$$

where $\beta(y) := 1 - \int R(y, \mathrm{d}u)S(y, u; \mathrm{d}y')T(y, u, y'; \mathrm{d}u')\alpha(y, u, y', u')$. Thus, establishing π^*-reversibility of K amounts to verifying, for all f and g in $\mathcal{F}(\mathcal{Y})$,

$$\int f(y)g(y')\pi^*(\mathrm{d}y) \int R(y, \mathrm{d}u)S(y, u; \mathrm{d}y')T(y, u, y'; \mathrm{d}u')\alpha(y, u, y', u') = \int f(y)g(y')\pi^*(\mathrm{d}y') \int R(y', \mathrm{d}u')S(y', u'; \mathrm{d}y)T(y', u', y; \mathrm{d}u)\alpha(y', u', y, u). \tag{27}$$

Indeed, by π-reversibility of $\{(Y_k^{(1)}, U_k^{(1)}); k \in \mathbb{N}\}$ it holds, for all $\bar{f}$ and $\bar{g}$ in $\mathcal{F}(\mathcal{Y} \otimes \mathcal{U})$,

$$\iint \bar{f}(y, u)\bar{g}(y', u')\pi(\mathrm{d}y \times \mathrm{d}u)S(y, u; \mathrm{d}y')T(y, u, y'; \mathrm{d}u')\alpha(y, u, y', u') = \iint \bar{f}(y, u)\bar{g}(y', u')\pi(\mathrm{d}y' \times \mathrm{d}u')S(y', u'; \mathrm{d}y)T(y', u', y; \mathrm{d}u)\alpha(y', u', y, u),$$

which establishes (27) by letting $\bar{f}(y, u) = f(y)$ and $\bar{g}(y, u) = g(y)$. This completes the proof.

APPENDIX B: RELATION BETWEEN ALGORITHM 2 AND THE R-MCMC AND GMTM ALGORITHMS

B.1. r-MCMC as a special case of Algorithm 2. As proposed initially by [10], the r-MCMC algorithm generates a Markov chain $\{Y_k^{(2)}; k \in \mathbb{N}\}$ with transitions given by Algorithm 4 below. Denote by $\left|\frac{\partial f}{\partial u}(u)\right|$ the Jacobian determinant of a vector-valued transformation f. In this algorithm, f is any continuously differentiable involution on $\mathsf{U} = \mathbb{R}^d$. In addition, $\check{R}$ and $\check{S}$ are instrumental kernels on $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{Y}^2, \mathcal{U})$, respectively, having transition densities $\check{r}$ and $\check{s}$ with respect to some dominating measure and Lebesgue measure on $\mathbb{R}^d$, respectively.

PROPOSITION 25. *The r-MCMC algorithm is a special case of Algorithm 2.*

Algorithm 4 r-MCMC [10]

Given $Y_k^{(2)} = y$,

(i) draw $\hat{Y} \sim \check{R}(y, \cdot) \rightsquigarrow \hat{y}$,

(ii) draw $U \sim \check{S}(y, \hat{y}; \cdot) \rightsquigarrow u$,

(iii) set

$$Y_{k+1}^{(2)} \leftarrow \begin{cases} \hat{y} & \text{w. pr. } \alpha^{(\mathrm{r})}(y, u, \hat{y}) := 1 \wedge \dfrac{\pi^*(\hat{y})\check{r}(\hat{y}, y)\check{s}(\hat{y}, y; f(u))}{\pi^*(y)\check{r}(y, \hat{y})\check{s}(y, \hat{y}; u)} \left| \dfrac{\partial f}{\partial u}(u) \right|, \\ y & \text{otherwise.} \end{cases} \tag{28}$$

Proof. Since $\hat{Y}$ and U, obtained in Steps (i) and (ii) of Algorithm 4, are not drawn in the same order as in Algorithm 2, we first derive the expression of the corresponding kernels R and S, i.e.

$$R(y, \mathrm{d}u) = \left(\int \check{R}(y, \mathrm{d}\hat{y})\check{s}(y, \hat{y}; u) \right) \lambda_d(\mathrm{d}u) = r(y, u)\lambda_d(\mathrm{d}u),$$
$$S(y, u; \mathrm{d}\hat{y}) = \frac{\check{R}(y, \mathrm{d}\hat{y})\check{s}(y, \hat{y}; u)}{\int \check{R}(y, \mathrm{d}\hat{y})\check{s}(y, \hat{y}; u)},$$

where λ_d is Lebesgue measure on $\mathbb{R}^d$. Also note that

$$R(y, \mathrm{d}u)S(y, u; \mathrm{d}\hat{y}) = \check{R}(y, \mathrm{d}\hat{y})\check{s}(y, \hat{y}; u)\lambda_d(\mathrm{d}u). \tag{29}$$

Moreover, introduce another auxiliary variable $\hat{U}$ taking values in U and being drawn according to $T(y, u, \hat{y}; \mathrm{d}\hat{u}) = \delta_{f(u)}(\mathrm{d}\hat{u})$. Note that the kernel T is not dominated by a common nonnegative measure regardless the value of u; still, following Remark 11, the r-MCMC algorithm may be covered by Algorithm 2, provided that the ratio in the acceptance probability $\alpha^{(\mathrm{r})}(y, u, \hat{y})$ corresponds to the Radon-Nikodym derivative in Proposition 10 for

$$K^{(\mathrm{r})}(y, u; \mathrm{d}\hat{y} \times \mathrm{d}\hat{u}) = S(y, u; \mathrm{d}\hat{y})T(y, u, \hat{y}; \mathrm{d}\hat{u}) = S(y, u; \mathrm{d}\hat{y})\delta_{f(u)}(\mathrm{d}\hat{u})$$

and

$$\pi^{(\mathrm{r})}(\mathrm{d}y \times \mathrm{d}u) = \pi^*(\mathrm{d}y)R(y, \mathrm{d}u).$$

The proof is completed by applying Lemma 26 below. □

Lemma 26. *The acceptance probability $\alpha^{(\mathrm{r})}$ in (28) is equal to*

$$\alpha^{(\mathrm{r})}(y, u, \hat{y}) = 1 \wedge \frac{\mathrm{d}\nu^{(\mathrm{r})}}{\mathrm{d}\mu^{(\mathrm{r})}}(x, \hat{x}), \tag{30}$$

where $x := (y, u)$, $\hat{x} := (\hat{y}, \hat{u})$, *and* $\frac{\mathrm{d}\nu^{(\mathrm{r})}}{\mathrm{d}\mu^{(\mathrm{r})}}$ *denotes the Radon-Nikodym derivative between the measures* $\nu^{(\mathrm{r})}$ *and* $\mu^{(\mathrm{r})}$ *defined by*

$$\nu^{(\mathrm{r})}(\mathrm{d}x \times \mathrm{d}\hat{x}) := \pi^{(\mathrm{r})}(\mathrm{d}\hat{y} \times \mathrm{d}\hat{u}) K^{(\mathrm{r})}(\hat{y}, \hat{u}; \mathrm{d}y \times \mathrm{d}u),$$
$$\mu^{(\mathrm{r})}(\mathrm{d}x \times \mathrm{d}\hat{x}) := \pi^{(\mathrm{r})}(\mathrm{d}y \times \mathrm{d}u) K^{(\mathrm{r})}(y, u; \mathrm{d}\hat{y} \times \mathrm{d}\hat{u}).$$

Proof. Write $\alpha^{(\mathrm{r})}(y, u, \hat{y}) = 1 \wedge \gamma^{(\mathrm{r})}(y, u, \hat{y})$, where

$$\gamma^{(\mathrm{r})}(y, u, \hat{y}) := \frac{\pi^*(\hat{y})\check{r}(\hat{y}, y)\check{s}(\hat{y}, y; f(u))}{\pi^*(y)\check{r}(y, \hat{y})\check{s}(y, \hat{y}; u)} \left| \frac{\partial f}{\partial u}(u) \right|.$$

To show (30) we will prove that for all bounded measurable functions G on $(\mathsf{Y} \times \mathsf{U})^2$ it holds that

$$\mathbb{E}_{\nu^{(\mathrm{r})}}[G(X, \hat{X})] = \int G(x, \hat{x}) \nu^{(\mathrm{r})}(\mathrm{d}x \times \mathrm{d}\hat{x}) = \int G(x, \hat{x}) \gamma^{(\mathrm{r})}(y, u, \hat{y}) \mu^{(\mathrm{r})}(\mathrm{d}x \times \mathrm{d}\hat{x})$$

(where $x = (y, u)$ and $\hat{x} = (\hat{y}, \hat{u})$). Now, using the change of variables $u = f(\hat{u})$, which is equivalent to $\hat{u} = f(u)$ (since f is an involution) and using the relation (29) we obtain

$$\begin{aligned}
&\mathbb{E}_{\nu^{(\mathrm{r})}}[G^{(\mathrm{r})}(X, \hat{X})] \\
&= \int G^{(\mathrm{r})}(y, f(\hat{u}), \hat{y}, \hat{u}) \pi^*(\mathrm{d}\hat{y}) r(\hat{y}, \hat{u}) S(\hat{y}, \hat{u}; \mathrm{d}y) \lambda_d(\mathrm{d}\hat{u}), \\
&= \int G^{(\mathrm{r})}(y, u, \hat{y}, f(u)) \pi^*(\mathrm{d}\hat{y}) r(\hat{y}, f(u)) S(\hat{y}, f(u); \mathrm{d}y) |(\partial f / \partial u)(u)| \lambda_d(\mathrm{d}u), \\
&= \int G^{(\mathrm{r})}(y, u, \hat{y}, f(u)) \frac{\pi^*(\hat{y})\check{r}(\hat{y}, y)\check{s}(\hat{y}, y; f(u))}{\pi^*(y)\check{r}(y, \hat{y})\check{s}(y, \hat{y}; u)} \left| \frac{\partial f}{\partial u}(u) \right| \\
&\qquad\qquad \times \pi^*(\mathrm{d}y) \check{R}(y, \mathrm{d}\hat{y}) \check{S}(y, \hat{y}; \mathrm{d}u) \\
&= \int G^{(\mathrm{r})}(x, \hat{x}) \gamma^{(\mathrm{r})}(y, u, \hat{y}) \mu^{(\mathrm{r})}(\mathrm{d}x \times \mathrm{d}\hat{x}),
\end{aligned}$$

which completes the proof. □

B.2. GMTM as a special case of Algorithm 2. The GMTM algorithm proposed in [11] generates a Markov chain $\{Y_k^{(2)}; k \in \mathbb{N}\}$ with transitions given by Algorithm 5 below. In Algorithm 5, the auxiliary variables $V_1, \ldots, V_n$ are defined on Y and for all $y \in \mathsf{Y}$ and $(v_1, \ldots, v_n) \in \mathsf{Y}^n$, $\{\omega(y, v_k) / \sum_{\ell=1}^n \omega(y, v_\ell)\}_{k=1}^n$ are sample weights. Moreover, $\check{R}$ is an instrumental kernel defined on $(\mathsf{Y}, \mathcal{Y})$ having the transition density $\check{r}$ with respect to some dominating measure on $(\mathsf{Y}, \mathcal{Y})$.

Algorithm 5 GMTM [11]

Given $Y_k^{(2)} = y$,

(i) draw $(V_1, \dots, V_n) \sim_{\text{i.i.d.}} \check{R}(y, \cdot) \rightsquigarrow (v_1, \dots, v_n)$,

(ii) let J take the value $j \in \{1, 2, \dots, n\}$ w. pr. $\omega(y, v_j) / \sum_{\ell=1}^n \omega(y, v_\ell)$,

(iii) let $\hat{y} \leftarrow v_j$,

(iv) draw $(\hat{V}_1, \dots, \hat{V}_{n-1}) \sim_{\text{i.i.d.}} \check{R}(\hat{y}, \cdot) \rightsquigarrow (\hat{v}_1, \dots, \hat{v}_{n-1})$,

(v) let $\hat{v}_n \leftarrow y$,

(vi) let

$$Y_{k+1}^{(2)} \leftarrow \begin{cases} \hat{y} & \text{with probability } \alpha^{(\mathrm{m})}(y, v, \hat{y}, \hat{v}) \\ & \qquad := 1 \wedge \dfrac{\pi^*(\hat{y})\check{r}(\hat{y}, y)\omega(\hat{y}, y) \sum_{k=1}^n \omega(y, v_k)}{\pi^*(y)\check{r}(y, \hat{y})\omega(y, \hat{y}) \sum_{k=1}^n \omega(\hat{y}, \hat{v}_k)}, \\ y & \text{otherwise.} \end{cases} \tag{31}$$

Proposition 27. *The GMTM algorithm is a special case of Algorithm 2.*

Proof. Denoting by $V_1, \dots, V_n$ the random variables generated in Step (i) in Algorithm 5, the proposed candidate $\hat{Y}$ is obtained as V_J, where J is generated in Step (ii). Let $U = V_{-J}$, where

$$v_{-j} := (v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_n).$$

To obtain the joint distribution of $(\hat{Y}, U)$ conditionally on $Y_k^{(2)}$, write, for any bounded measurable function G on Y^n,

$$\begin{aligned} &\mathbb{E}[G(\hat{Y}, U) \mid Y_k^{(2)} = y] \\ &= \sum_{j=1}^n \mathbb{E}[G(V_j, V_{-j}) \mathbb{1}_{J=j} \mid Y_k^{(2)} = y] \\ &= \int \cdots \int \check{R}(y, \mathrm{d}\hat{y}) \prod_{k=1}^{n-1} \check{R}(y, \mathrm{d}u_k) \frac{n\omega(y, \hat{y})}{\sum_{\ell=1}^{n-1} \omega(y, u_\ell) + \omega(y, \hat{y})} G(\hat{y}, u) \\ &= \int \cdots \int R(y, \mathrm{d}u) S(y, u; \mathrm{d}\hat{y}) G(\hat{y}, u), \end{aligned}$$

where we introduced the kernels

$$R(y, \mathrm{d}u) := n \prod_{k=1}^{n-1} \check{R}(y, \mathrm{d}u_k) \int \frac{\check{R}(y, \mathrm{d}\hat{y})\omega(y, \hat{y})}{\sum_{\ell=1}^{n-1} \omega(y, u_\ell) + \omega(y, \hat{y})}, \tag{32}$$

$$S(y, u; \mathrm{d}\hat{y}) := \frac{\check{R}(y, \mathrm{d}\hat{y})\omega(y, \hat{y})}{\sum_{\ell=1}^{n-1} \omega(y, u_\ell) + \omega(y, \hat{y})} \Bigg/ \int \frac{\check{R}(y, \mathrm{d}\hat{y})\omega(y, \hat{y})}{\sum_{\ell=1}^{n-1} \omega(y, u_\ell) + \omega(y, \hat{y})}. \tag{33}$$

Now, set $\hat{U} = (\hat{V}_1, \ldots, \hat{V}_{n-1})$ where $\hat{V}_i$s are sampled in Step (iv). The distribution of $\hat{U}$ conditionally on $(Y_k^{(2)}, U, \hat{Y}) = (y, u, \hat{y})$ is given by

$$T(y, u, \hat{y}; \mathrm{d}\hat{u}) = \prod_{k=1}^{n-1} \check{R}(\hat{y}, \mathrm{d}\hat{u}_k). \tag{34}$$

If $\check{R}$ is dominated by a nonnegative measure, then (32), (33), and (34) show that the kernels R, S, and T are dominated as well. Denoting by r, s, and t the corresponding transition densities, it can be readily checked that

$$\frac{\pi^*(\hat{y})r(\hat{y}, \hat{u})s(\hat{y}, \hat{u}; y)t(\hat{y}, \hat{u}, y; u)}{\pi^*(y)r(y, u)s(y, u; \hat{y})t(y, u, \hat{y}; \hat{u})} = \frac{\pi^*(\hat{y})\check{r}(\hat{y}, y)\omega(\hat{y}, y)(\sum_{k=1}^{n-1} \omega(y, u_k) + \omega(y, \hat{y}))}{\pi^*(y)\check{r}(y, \hat{y})\omega(y, \hat{y})(\sum_{k=1}^{n-1} \omega(\hat{y}, \hat{u}_k) + \omega(\hat{y}, y))},$$

so that $\alpha^{(\mathrm{m})}$ defined in (31) corresponds to the acceptance probability α defined in (11) with these particular choices of r, s, and t. Consequently, the GMTM algorithm is a special case of Algorithm 2. □

Note that in the previous proof, we have chosen the auxiliary variable U as the vector of rejected candidates after Step (ii). Another natural idea would consist in choosing $U = (V_1, \ldots, V_n)$, where the V_is are obtained in Step (i); however, since $\hat{Y}$ belongs to this set of candidates, the model would then not be dominated, which would make the proof more intricate.

REFERENCES

[1] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[2] C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *arXiv preprint arXiv:1210.1484*, 2012.

[3] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*, volume 2. CRC press, 1996.

[4] O. Häggström and J. S. Rosenthal. On variance conditions for Markov chain CLTs. *Elect. Comm. in Probab.*, 12:454–464, 2007.

[5] M. Hairer and J. C. Mattingly. Yet another look at Harris ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer, 2011.

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

[7] J. S. Liu, F. Liang, and W. H. Wong. The Multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.

[8] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[9] A. Mira. Ordering and improving the performance of Monte Carlo Markov chains. *Statistical Science*, pages 340–350, 2001.

[10] G. K. Nicholls, C. Fox, and A. M. Watt. Coupled MCMC with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*, 2012.

[11] S. Pandolfi, F. Bartolucci, and N. Friel. A generalization of the Multiple-try Metropolis algorithm for Bayesian estimation and model selection. In *International Conference on Artificial Intelligence and Statistics*, pages 581–588, 2010.

[12] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, December 1973.

[13] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

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

[15] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

[16] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, pages 1–9, 1998.

Telecom SudParis
9 rue Charles Fourier
91011 Evry
France
E-mail: florian.maire@telecom-sudparis.eu
randal.douc@telecom-sudparis.eu

SE-100 44 Stockholm
Sweden
E-mail: jimmyol@kth.se