JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Franks, Jordan; Vihola, Matti

Title: Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance

Year: 2020

Version: Accepted version (Final draft)

Copyright: © 2020 Elsevier BV

Rights: CC BY-NC-ND 4.0

Rights url: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the original version:

Franks, J., & Vihola, M. (2020). Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance. Stochastic Processes and Their Applications, 130(10), 6157-6183. https://doi.org/10.1016/j.spa.2020.05.006

arXiv:1706.09873v6 [stat.CO] 24 Mar 2020

IMPORTANCE SAMPLING CORRECTION VERSUS STANDARD AVERAGES OF REVERSIBLE MCMCS IN TERMS OF THE ASYMPTOTIC VARIANCE

JORDAN FRANKS AND MATTI VIHOLA

Abstract. We establish an ordering criterion for the asymptotic variances of two consistent Markov chain Monte Carlo (MCMC) estimators: an importance sampling (IS) estimator, based on an approximate reversible chain and subsequent IS weighting, and a standard MCMC estimator, based on an exact reversible chain. Essentially, we relax the criterion of the Peskun type covariance ordering by considering two different invariant probabilities, and obtain, in place of a strict ordering of asymptotic variances, a bound of the asymptotic variance of IS by that of the direct MCMC. Simple examples show that IS can have arbitrarily better or worse asymptotic variance than Metropolis-Hastings and delayed-acceptance (DA) MCMC. Our ordering implies that IS is guaranteed to be competitive up to a factor depending on the supremum of the (marginal) IS weight. We elaborate upon the criterion in case of unbiased estimators as part of an auxiliary variable framework. We show how the criterion implies asymptotic variance guarantees for IS in terms of pseudo-marginal (PM) and DA corrections, essentially if the ratio of exact and approximate likelihoods is bounded. We also show that convergence of the IS chain can be less affected by unbounded high-variance unbiased estimators than PM and DA chains.

1. Introduction

Let $\nu(\theta, z)\mathrm{d}\theta\mathrm{d}z$ be a probability measure on a jointly measurable space $\mathbf{T} \times \mathbf{Z}$ with σ-finite dominating measure $\mathrm{d}\theta\mathrm{d}z$, and suppose one desires to calculate expectations with respect to ν or its marginal $\nu^*(\theta) := \int \nu(\theta, z)\mathrm{d}z$. In many scenarios of interest, $\nu^*(\theta)$ is intractable to evaluate and $\mathbf{Z}$ is high-dimensional. Even if the full joint density $\nu(\theta, z)$ is tractable up to a normalising constant, high-dimensional Markov chain Monte Carlo (MCMC) based on targeting ν with Metropolis-Hastings (MH) is often inefficient or even useless due to difficulty with the design of proposal distribution [35].

To deal with these issues, often one can transform the high-dimensional MCMC into a pseudo-marginal (PM) [6] or approximate marginal MCMC [51]. Then the proposal distribution can live on the low-dimensional space $\mathbf{T}$, and the resulting chains are often much more efficient. The PM approach is asymptotically exact, while the approximate marginal approach requires an importance sampling (IS) correction to make it so. We next describe these approaches, and give our main result comparing the relative efficiency of these approaches in terms of the asymptotic variance.

Key words and phrases. Asymptotic variance, delayed acceptance, importance sampling, Markov chain Monte Carlo, pseudo-marginal algorithm, unbiased estimator.

Algorithm 1 Pseudo-marginal algorithm, for iteration $k \geq 1$.

(PM 1) Propose a transition $\Theta' \sim q_{\Theta_{k-1}}(\cdot)$.

(PM 2) Given Θ', generate $(\zeta'^{(1:m)}, Z'^{(1:m)})$, and with probability

$$\min\left\{1, \frac{q_{\Theta'}(\Theta_{k-1}) \sum_{i=1}^m \zeta'^{(i)}}{q_{\Theta_{k-1}}(\Theta') \sum_{i=1}^m \zeta_{k-1}^{(i)}}\right\}$$

set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)}) \leftarrow (\Theta', \zeta'^{(i)}, Z'^{(i)})$. Otherwise, set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)})) \leftarrow (\Theta_{k-1}, \zeta_{k-1}^{(i)}, Z_{k-1}^{(i)})$.

1.1. **Pseudo-marginal Markov chain Monte Carlo.** The PM approach is based on replacing $\nu^*(\theta)$ with a non-negative unbiased estimator $\hat{\nu}^*(\theta)$ of $\nu^*(\theta)$ (up to constant) within the standard (but assumed unavailable) MH algorithm targeting ν^* [6, 32]. That is, we assume there is a constant $c_\nu > 0$ such that

$$\mathbb{E}[\hat{\nu}^*(\theta)] = c_\nu \, \nu^*(\theta) \tag{1}$$

for all θ. A standard example of such an estimator is

$$\hat{\nu}^*(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{\nu(\theta, Z_i)}{Q_\theta(Z_i)},$$

where Z_i are sampled i.i.d. from some instrumental distribution $Q_\theta(\cdot)$ satisfying $Q_\theta(z) = 0 \implies \nu(\theta, z) = 0$.

An unbiased estimator satisfying (1) will allow for calculation of marginal expectations with respect to ν^* using the approaches we consider, but often one can do much better, allowing also for joint expectations with respect to ν [2, 51]. That is, suppose one has access to non-negative unbiased estimators for a subclass $\mathcal{L}^1(\nu)$ of functions which we now describe.

With $\theta \in \mathbf{T}$, $\zeta^{(i)} \in [0, \infty)$ and $z^{(i)} \in \mathbf{Z}$ for $i = 1, \ldots, m$, set

$$\zeta(g) := \frac{1}{m} \sum_{i=1}^m \zeta^{(i)} g(\theta, z^{(i)})$$

for $g \in L^1(\nu)$. Let Q be a probability kernel from $\mathbf{T}$ to $\mathbf{V} := [0, \infty)^m \times \mathbf{Z}^m$. Let $\mathcal{L}^1(\nu)$ consist of those functions $f \in L^1(\nu)$ such that for $g \in \{f, |f|\}$,

$$\int Q_\theta(\mathrm{d}v)\zeta(g) = c_\nu \, \nu^*(\theta) \int g(\theta, z)\nu(\mathrm{d}z|\theta) \tag{2}$$

for all $\theta \in \mathbf{T}$ and $v := (\zeta^{(1:m)}, z^{(1:m)}) \in \mathbf{V}$, where $c_\nu > 0$ is some fixed constant not depending on (θ, v), and $\nu(\mathrm{d}z|\theta)$ denotes a regular conditional probability of ν given θ.

Note that if $1 \in \mathcal{L}^1(\nu)$, then (1) is satisfied with $\hat{\nu}^*(\theta) := \frac{1}{m} \sum_{i=1}^m \zeta^{(i)}$, and $L^1(\nu^*)$ is naturally included into $\mathcal{L}^1(\nu)$ via $f(\theta, z) := f(\theta)$ for $f \in L^1(\nu^*)$. In the following, we will always assume that $1 \in \mathcal{L}^1(\nu)$, since the schemes we consider require this for consistency.

Let q be a transition density on $\mathbf{T}$, where $q_\theta(\theta')$ denotes the probability to move from θ to θ'. With $(\Theta_0, \zeta_0^{(1:m)}, Z_0^{(1:m)}) \in \mathbf{T} \times [0, \infty)^m \times \mathbf{Z}^m$ initial values with $\sum_{i=1}^m \zeta_0^{(i)} > 0$, for $k = 1, \ldots, n$, the PM iteration is given in Algorithm 1 [see 6].

Algorithm 2 Delayed-acceptance (DA0), for iteration $k \geq 1$.

(DA0 1) Propose a transition $\Theta' \sim q_{\Theta_{k-1}}(\cdot)$.

(DA0 2) Proceed to Step (DA0 3) with probability

$$\min\left\{1, \frac{\mu_u^*(\Theta')q_{\Theta'}(\Theta_{k-1})}{\mu_u^*(\Theta_{k-1})q_{\Theta_{k-1}}(\Theta')}\right\}.$$

Otherwise, set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)}) \leftarrow (\Theta_{k-1}, \zeta_{k-1}^{(i)}, Z_{k-1}^{(i)}))$ and exit.

(DA0 3) Given Θ', generate $(\zeta'^{(1:m)}, Z'^{(1:m)})$. With probability

$$\min\left\{1, \frac{(\sum_{i=1}^m \zeta'^{(i)})/\mu_u^*(\Theta')}{(\sum_{i=1}^m \zeta_{k-1}^{(i)})/\mu_u^*(\Theta_{k-1})}\right\}$$

set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)}) \leftarrow (\Theta', \zeta'^{(i)}, Z'^{(i)})$. Otherwise, set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)}) \leftarrow (\Theta_{k-1}, \zeta_{k-1}^{(i)}, Z_{k-1}^{(i)})$.

Assuming $1 \in \mathcal{L}^1(\nu)$ and the PM chain is Harris ergodic (see Section 2), for $f \in \mathcal{L}^1(\nu)$ the estimator

$$\frac{1}{n}\sum_{k=1}^n \hat{\zeta}_k^{(i)} f(\Theta_k, Z_k^{(i)}) \xrightarrow{n\to\infty} \nu(f) := \mathbb{E}_\nu[f] \tag{3}$$

with $\hat{\zeta}_k^{(i)} := \zeta_k^{(i)}/\sum_{j=1}^m \zeta_k^{(j)}$, is a consistent estimator for $\nu(f)$ [6].

1.2. **Accelerations based on an approximation.** Suppose one has an approximation μ^* of ν^*, by which we mean that μ^* is some probability measure on $\mathbf{T}$ such that

$$\mu^*(\theta) = 0 \implies \nu^*(\theta) = 0, \tag{4}$$

and there is some constant $c_\mu > 0$ (perhaps unknown) such that we can evaluate unnormalised $\mu_u^*(\theta) = c_\mu \mu^*(\theta)$ for all $\theta \in \mathbf{T}$. The approximation μ^* could arise, for example, when subsampling data [10, 42] or using a more tractable diffusion model instead of a Markov jump process model [26], giving rise to the approximate posterior μ^*. We next describe delayed-acceptance (DA) in two variants, and IS, all of which make use of the approximation μ^*.

1.2.1. *Delayed-acceptance MCMC in two variants.* If one has an approximation μ^* of ν^* as above, one can use a PM acceleration technique known as DA [17, 32, 34], which has garnered considerable interest. With $(\Theta_0, \zeta_0^{(1:m)}, Z_0^{(1:m)}) \in \mathbf{T}\times[0,\infty)^m \times \mathbf{Z}^m$ initial values with $\mu^*(\Theta_0) > 0$ and $\sum_{i=1}^m \zeta_0^{(i)} > 0$, for $k = 1, \ldots, n$, iterate as given in Algorithm 2.

The DA estimator for $\nu(f)$ is given in (3), which is the same as in the PM case. Note that Step (DA0 3), which involves possibly expensive unbiased estimator generation, is only run if the proposal Θ' is assigned sufficient approximate probability in Step (DA0 2) which means it is likely to be accepted in Step (DA0 3).

Consider now Algorithm 3, which is a variant, DA1, of DA0 Algorithm 2. DA1 is considered in [34, 'surrogate transition method,' Section 9.4.3]. In the deterministic case $\frac{1}{m}\sum_{i=1}^m \zeta^{(i)} = c_\nu \nu^*(\theta)$ almost surely for all θ with $(\zeta^{(1:m)}, Z^{(1:m)}) \sim Q_\theta(\cdot)$,

Algorithm 3 Delayed-acceptance (DA1), for iteration $k \geq 1$.

(DA1 1) Propose a transition $\Theta' \sim q_{\Theta_{k-1}}(\cdot)$.

(DA1 2) With probability

$$\min\left\{1, \frac{\mu_u^*(\Theta')q_{\Theta'}(\Theta_{k-1})}{\mu_u^*(\Theta_{k-1})q_{\Theta_{k-1}}(\Theta')}\right\}$$

set $\Theta'' \leftarrow \Theta'$. Otherwise, set $\Theta'' \leftarrow \Theta_{k-1}$.

(DA1 3) Given Θ'', generate $(\zeta''^{(1:m)}, Z''^{(1:m)})$, With probability

$$\min\left\{1, \frac{(\sum_{i=1}^m \zeta''^{(i)})/\mu_u^*(\Theta'')}{(\sum_{i=1}^m \zeta_{k-1}^{(i)})/\mu_u^*(\Theta_{k-1})}\right\}$$

set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)}) \leftarrow (\Theta'', \zeta''^{(i)}, Z''^{(i)})$.
Otherwise, set $(\Theta_k, \zeta_k^{(i)}, Z_k^{(i)}) \leftarrow (\Theta_{k-1}, \zeta_{k-1}^{(i)}, Z_{k-1}^{(i)})$.

Algorithm 4 Importance sampling correction of approximate MCMC

(IS Phase 1) Let $(\Theta_0, \zeta_0^{(1:m)}, Z_0^{(1:m)}) \in \mathbf{X} \times (0,\infty)^m \times \mathbf{Z}^m$ be some initial values with $\mu^*(\Theta_0) > 0$ and $\sum_{i=1}^m \zeta_0^{(i)} > 0$. For $k = 1, \ldots, n$, do:

i. Propose a transition $\Theta' \sim q_{\Theta_{k-1}}(\cdot)$.

ii. With probability

$$\min\left\{1, \frac{\mu_u^*(\Theta')q_{\Theta'}(\Theta_{k-1})}{\mu_u^*(\Theta_{k-1})q_{\Theta_{k-1}}(\Theta')}\right\}$$

set $\Theta_k \leftarrow \Theta'$. Otherwise, set $\Theta_k \leftarrow \Theta_{k-1}$.

(IS Phase 2) For each $k \in \{1, \ldots, n\}$, given Θ_k, generate $(\zeta_k^{(1:m)}, Z_k^{(1:m)})$. With $\xi_k^{(i)} := \zeta_k^{(i)}/\mu_u^*(\theta)$, form the IS estimator

$$E_n^{\mathrm{IS}}(f) := \frac{\sum_{k=1}^n \sum_{i=1}^m \xi_k^{(i)} f(\Theta_k, Z_k^{(i)})}{\sum_{k=1}^n \sum_{i=1}^m \xi_k^{(i)}} \xrightarrow{n\to\infty} \nu(f), \tag{5}$$

consistent if the Phase 1 chain is Harris ergodic and $1, f \in \mathcal{L}^1(\nu)$.

then DA0 and DA1 have the same transition kernels (Proposition 14). But in general DA1 has lower asymptotic variance than DA0 (Proposition 14), although DA0 is probably more computationally efficient. This is evident from the fact that Step (DA1 3) is performed at every iteration of DA1 (Algorithm 3).

1.2.2. *Importance sampling correction of approximate MCMC.* MCMC-IS (Algorithm 4) consists of targeting an approximation μ^* of ν^* with MCMC, and then using importance sampling (IS) correction over the latent states [20, 24, 25, 27, 40, 51]. Let μ^* be an approximation of ν^* as in (4). Note that IS Phase 2, which involves the generation of unbiased estimators, may be done independently for each k which allows for efficient parallelisation.

1.3. **Defining the asymptotic variance.** As PM/DA and MCMC-IS are viable approaches for consistent inference, the central question is which one should be used. The standard measure of statistical efficiency for MCMCs is the *asymptotic variance*.

Definition 1 (Asymptotic variance)**.** Let (X_k) be a ν-Harris ergodic Markov chain with transition L. For $f \in L^2(\nu)$ the *asymptotic variance* of f with respect to L is defined, whenever the limit exists in $[0, \infty]$, as

$$\mathrm{var}(L, f) := \lim_{n\to\infty} \mathbb{E}\Big[\Big(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} [f(X_k^{(s)}) - \nu(f)]\Big)^2\Big], \tag{6}$$

where $(X_k^{(s)})$ denotes a stationary version of the chain (X_k), i.e. $X_0^{(s)} \sim \nu$.

For reversible L, which is the focus of this paper, $\mathrm{var}(L, f)$ always exists in $[0, \infty]$ [see 49]. Moreover, a CLT holds under general conditions.

Proposition 1. *Let $(X_k)_{k\geq 1}$ be an aperiodic ν-reversible Harris ergodic Markov chain with transition L. If $f \in L^2(\nu)$ and $\mathrm{var}(L, f) < \infty$, then, for all initial distributions,*

$$\frac{1}{\sqrt{n}}\Big(\sum_{k=1}^{n} [f(X_k) - \nu(f)]\Big) \xrightarrow{n\to\infty} \mathcal{N}\big(0, \mathrm{var}(L, f)\big), \qquad \textit{in distribution}, \tag{7}$$

where $\mathcal{N}(a, b^2)$ is a normal distribution with mean a and variance b^2.

Proposition 1 follows from [29, Cor. 1.5], where it holds under all initial conditions because of the Harris ergodicity assumption [see 21, Cor. 21.1.6]. Proposition 1 above explains the importance of the asymptotic variance, since it is the CLT limiting variance. The asymptotic variance characterises the *statistical efficiency* of the method in the asymptotic regime, but also characterises the finite sample efficiency in the finite regime [see 46].

1.4. **Comparing the asymptotic variances.** We first define some objects. Given $X_k = (\Theta_k, \zeta_k^{(1:m)}, Z_k^{(1:m)})$ and $f : \mathbf{X} \to \mathbb{R}$, define

$$\zeta_k(f) = \frac{1}{m} \sum_{i=1}^{m} \zeta_k^{(i)} f(\Theta_k, Z_k^{(i)}), \qquad \hat{\zeta}_k^{(i)} := \frac{\zeta_k(f)}{\zeta_k(1)} \qquad \xi_k(f) := \frac{\zeta_k(f)}{\mu_u^*(\Theta_k)}.$$

We also define the MCMC-IS kernel to be

$$\bar{K}_{\theta v}(\mathrm{d}\theta', \mathrm{d}v') := K_\theta(\mathrm{d}\theta') Q_{\theta'}(\mathrm{d}v') \tag{8}$$

where K is the approximate marginal MH kernel in IS (Algorithm 4) Phase (1) with invariant measure μ^* [51]. $\bar{K}$ is $\bar{\mu}$-reversible, where $\bar{\mu}(\mathrm{d}\theta, \mathrm{d}v) := \mu^*(\mathrm{d}\theta) Q_\theta(\mathrm{d}v)$. Define the extended IS weights $w_u(X_k) := \xi_k(1)$ and $w(X_k) := w_u(X_k) * (c_\mu / c_\nu)$.

Assume $1 \in \mathcal{L}^1(\nu)$ and (4) holds. By our discussion of the asymptotic variance, $\mathrm{var}\big(L, \hat{\zeta}(f)\big)$ is assigned to the PM/DA estimator $n^{-1} \sum_{k=1}^{n} \hat{\zeta}_k(f)$ given in (3). Now let $\bar{K}$ be the MCMC-IS kernel defined in (8), and note that the IS estimator (5) can be written as

$$E_n^{\mathrm{IS}}(f) = \frac{\frac{1}{n}\sum_{k=1}^{n} \xi_k(f)}{\frac{1}{n}\sum_{k=1}^{n} \xi_k(1)} = \frac{\frac{1}{n}\sum_{k=1}^{n} w_u(X_k) \hat{\zeta}_k(f)}{\frac{1}{n}\sum_{k=1}^{n} w_u(X_k)}. \tag{9}$$

Since $\mathrm{var}(\bar{K}, w_u \hat{\zeta}(f))$ is assigned to the numerator from the definition of the asymptotic variance, and the denominator converges almost surely to c_ν / c_μ under a Harris ergodicity assumption, the asymptotic variance $\mathrm{var}(\bar{K}, w\hat{\zeta}(f))$ is assigned to the IS estimator by (7) and Slutsky's lemma.

For a function $g : \mathbf{X} \to \mathbb{R}$ and probability ν on $\mathbf{X}$, define the norm

$$\|g\|_{L^\infty(\nu)} := \nu\text{-}\operatorname{ess\,sup}_{x\in\mathbf{X}} |g(x)|. \tag{10}$$

Let us define the marginal weight $w^*(\theta) := \nu^*(\theta)/\mu^*(\theta)$ and note that $\|w^*\|_{L^\infty(\mu^*)} \leq \|w\|_{L^\infty(\bar\mu)}$. Under Harris ergodicity, we remark that a consistent upper bound estimator for $\|w\|_{L^\infty(\bar\mu)}$ is given by

$$\bar c_n := \left(\frac{1}{n-n_b}\sum_{k=n_b+1}^{n} \xi_k(1)\right)^{-1} \max_{n_b<k\leq n} \xi_k(1), \tag{11}$$

which is moreover a consistent estimator for $\|w\|_{L^\infty(\bar\mu)}$ as $n_b, n \to \infty$, where $n_b \geq 1$ denotes the burn-in of the chain.

Let L be the transition corresponding to the PM, DA0, or DA1 chain. Then L has invariant probability

$$\pi(\mathrm{d}\theta, \mathrm{d}\zeta^{(1:m)}, \mathrm{d}z^{(1:m)}) := c_\nu^{-1}\mathrm{d}\theta Q_\theta(\mathrm{d}\zeta^{(1:m)}, \mathrm{d}z^{(1:m)})\zeta(1)$$

Define $\mathcal{L}^2(\nu) := \{f \in \mathcal{L}^1(\nu) : f^2 \in \mathcal{L}^1(\nu)\}$, where we recall $\mathcal{L}^1(\nu)$ was defined through (2).

Our main result (Theorem 12) in the present context says the following.

Corollary 2. *Suppose $1 \in \mathcal{L}^1(\nu)$. Let L be the transition kernel of the PM, DA0 or DA1 chains defined in Algorithm 1-3 respectively. Suppose (4) holds, and let $\bar K$ be the MCMC-IS kernel (8) corresponding to Algorithm 4, and let $f \in \mathcal{L}^2(\nu)$. Suppose K and L are Harris ergodic and $\operatorname{var}(\bar K, w\hat\zeta(f)) < \infty$. Set $\bar f := f - \nu(f)$. The following hold:*

(i) If $\|w^*\|_{L^\infty(\mu^*)} < \infty$ then,

$$\operatorname{var}\big(\bar K, w\hat\zeta(f)\big) \leq \|w^*\|_{L^\infty(\mu^*)}\Big(\operatorname{var}\big(L, \hat\zeta(f)\big) + \operatorname{var}_\pi\big(\hat\zeta(f)\big)\Big) + 3\operatorname{var}_{\bar\mu}\big(w\hat\zeta(\bar f)\big).$$

(ii) If $\|w\|_{L^\infty(\bar\mu)} < \infty$, then

$$\operatorname{var}\big(\bar K, w\hat\zeta(f)\big) + \operatorname{var}_{\bar\mu}\big(w\hat\zeta(\bar f)\big) \leq \|w\|_{L^\infty(\bar\mu)}\Big(\operatorname{var}\big(L, \hat\zeta(f)\big) + \operatorname{var}_\pi\big(\hat\zeta(f)\big)\Big).$$

(iii) If $w\hat\zeta(f) \in L^2(\bar\mu)$, then

$$\operatorname{var}\big(\bar K, w\hat\zeta(f)\big) + \operatorname{var}_{\bar\mu}\big(w\hat\zeta(\bar f)\big) \geq (\bar\mu\text{-}\operatorname{ess\,inf} w)\Big(\operatorname{var}\big(L, \hat\zeta(f)\big) + \operatorname{var}_\pi\big(\hat\zeta(f)\big)\Big).$$

Although these bounds do not provide an ordering of asymptotic variances as in the Peskun-Tierney ordering for direct MCMCs, we could never hope these bounds to do so in our context: we give simple examples showing that PM/DA (resp. IS) can do arbitrarily better than IS (resp. PM/DA) in terms of the asymptotic variance in Appendix D. IS seems to perform better than PM/DA when the approximation is good in the sense that the weight w has a low supremum, as Corollary 2 would suggest. In the usual unbounded space case, this means that $\mu(\theta)$ would need to have fatter tails than $\dot\nu(\theta)$, or at least that $\nu(\theta)/\mu(\theta)$ is bounded, which can often be done by inflating $\mu_u(\theta)$ uniformly by a positive constant [see 51].

The rest of this paper is concerned with proving Corollary 2 and other versions, for example, for general reversible chains, IS jump chains, and when $\mu_u^*(\theta)$ requires unbiased estimators.

1.5. **Previous work.** In various settings and different ways, we are not the first to compare direct MCMC with IS MCMC. A study of self-normalised IS versus the independence MH has been made in [33]. Asymptotic variances are explicitly computed and compared in some discrete examples in [12] who find that IS and MH can be competitive, but that MH can do much better (see also [11, Sect. 4.2]). On the other hand, [50] study independent IS with unbiased estimators, and find that this performs better than PM in their experiments (see also [16]). The IS versus DA question is noted in [18, Sect. 3.3.3], who mention the likely improvement of IS over DA in massive parallelisation. A methodological comparison of the alternatives in the general MCMC and joint inference context is made in [51], who investigate empirically the relative efficiencies, finding that IS and DA can be competitive, with IS doing slightly better than DA in their experiments, with little or no parallelisation. The gap widens with increased parallelisation, a known strength of the IS correction [see 18, 30, 51].

We consider here general reversible Markov chains, in particular PM/DA, and seek a Peskun type ordering of the asymptotic variances.

1.6. **Outline.** After preliminaries in Section 2, we state in Section 3 the Peskun type ordering result for normalised IS (Theorems 3) and augmented IS kernels (Theorem 5). We define jump chains and self-normalised importance sampling (SNIS) in Section 4, before proceeding to Section 5, where we consider a general auxiliary variable framework which accommodates IS and PM type schemes that use unbiased estimators. Specific PM type algorithms and kernels which we consider are given in Section 6, and we compare them with IS (Theorem 16). We discuss some stability considerations in Section 7. Proofs of the Peskun type orderings are given in Appendix A. Dirichlet form bounds and proof of the main comparison application (Theorem 16) are found in Appendix B. Appendix C mentions some properties of augmented chains. Appendix D contains the examples mentioned earlier.

2. Notation and definitions

2.1. **Notation.** The spaces we consider $\mathbf{X}$ are assumed equipped with a σ-algebra, denoted $\mathcal{B}(\mathbf{X})$, and with a σ-finite dominating measure, denoted '$\mathrm{d}x$.' Product spaces will be assumed equipped with their product σ-algebras and corresponding product measures. If μ is a probability density on $\mathbf{X}$, we denote the corresponding probability measure with the same symbol, so that $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$.

For $p \in [1, \infty)$, we denote by $L^p(\mu)$ the Banach space of equivalence classes of measurable $f : \mathbf{X} \to \mathbb{R}$ satisfying $\|f\|_p < \infty$ under the norm $\|f\|_{L^p(\mu)} := \{\int |f(x)|^p \mu(\mathrm{d}x)\}^{1/p}$. We similarly define $L^\infty(\mu)$ under the norm $\|f\|_{L^\infty(\mu)}$ as in (10). We denote by $L_0^p(\mu)$ the subset of $L^p(\mu)$ with $\mu(f) = 0$, where $\mu(f) := \int f(x)\mu(\mathrm{d}x)$. For $f \in L^1(\mu)$ and $K_x(\mathrm{d}x')$ a Markov kernel on $\mathbf{X}$, we define

$\mu K(A) := \int \mu(\mathrm{d}x) K_x(A)$ for $A \in \mathcal{B}(\mathbf{X})$, $Kf(x) := \int K_x(\mathrm{d}x') f(x')$, and inductively $K^n f(x) := K^{n-1}(Kf)(x)$ for $n \geq 2$. For $f, g \in L^2(\mu)$, we define $\langle f, g \rangle_\mu := \int f(x) g(x) \mu(\mathrm{d}x)$, $\|f\|_\mu := (\langle f, f \rangle_\mu)^{1/2}$, and $\mathrm{var}_\mu(f) := \mu(f^2) - \mu(f)^2$.

For $m \in \mathbb{N}$ and $x^{(i)} \in \mathbf{X}$ for $i = 1, \dots m$, we write $x^{(1:m)} := (x^{(1)}, \dots, x^{(m)})$. Throughout, ν will denote the target probability of interest, and for $\varphi \in L^1(\nu)$ we set $\bar{\varphi} := \varphi - \nu(\varphi)$, element of $L^1_0(\nu)$.

2.2. **Definitions.** Let μ and ν be σ-finite measures on $\mathbf{X}$. If $\mu(A) = 0$ implies $\nu(A) = 0$ for all $A \in \mathcal{B}(\mathbf{X})$, we say that ν is *absolutely continuous* with respect to μ, and write $\nu \ll \mu$. Suppose $\nu \ll \mu$. Recall that a *Radon-Nikodým derivative* of ν with respect to μ is a non-negative measurable function $\frac{d\nu}{d\mu}(x)$ on $\mathbf{X}$ such that $\mu(\frac{d\nu}{d\mu} g) = \nu(g)$ for all $g \in L^1(\nu)$. If also μ and ν are probability densities, then it is easy to see that $\frac{d\nu}{d\mu}(x)$ is in $L^1(\mu)$, and is equivalent with $\frac{\nu(x)}{\mu(x)}$.

Let μ be a probability on $\mathbf{X}$. A Markov chain K on $\mathbf{X}$ is *μ-invariant* if $\mu K = \mu$. If also $\langle f, Kf \rangle_\mu \geq 0$ for all $f \in L^2(\mu)$, then K is *positive*. If $\mu(\mathrm{d}x) K_x(\mathrm{d}x') = \mu(\mathrm{d}x') K_{x'}(\mathrm{d}x)$, then K is said to satisfy *detailed balance* with respect to μ, or briefly, K is *μ-reversible*. This implies that K is μ-invariant, and that the *Dirichlet form* $\mathcal{E}_K(f)$ for $f \in L^2(\mu)$ satisfies

$$\mathcal{E}_K(f) := \langle f, (1-K) f \rangle_\mu = \frac{1}{2} \int \mu(\mathrm{d}x) K_x(\mathrm{d}x') \big(f(x) - f(x')\big)^2. \tag{12}$$

We say Markov chain K is *μ-Harris ergodic* if K is μ-invariant, ψ-irreducible, and Harris recurrent. See [36] for the definition of ψ-irreducibility and Harris recurrence, and further details. Most MCMC schemes are Harris ergodic, although a careless implementation can lead to a non-Harris chain [see 43].

3. Peskun type ordering for normalised importance sampling

3.1. **General case.** Let μ and ν be probability measures on a measurable space $\mathbf{X}$, and let $w : \mathbf{X} \to [0, \infty)$ be a non-negative measurable function.

Assumption 1 (Importance sampling)**.** A triplet (μ, ν, w) is such that $\nu \ll \mu$ and $w(x) = \frac{d\nu}{d\mu}(x)$ is the Radon-Nikodým derivative.

Assumption 2. A heptuple $(\mu, \nu, w, K, L, \underline{c}, \overline{c})$ is such that (μ, ν, w) satisfies Assumption 1, K and L are Harris ergodic Markov chains reversible with respect to μ and ν, respectively, and the constants $\underline{c}$, $\overline{c} \geq 0$ satisfy

(a) $\underline{c}\, \mathcal{E}_K(g) \leq \mathcal{E}_L(g) \leq \overline{c}\, \mathcal{E}_K(g)$, for all $g \in L^2(\mu)$, and
(b) $\underline{c} \leq w \leq \overline{c}$, μ-a.e.

Theorem 3. *If Assumption 2 holds, then for all $\varphi \in L^2(\nu)$,*

$$\mathrm{var}(K, w\varphi) + \mathrm{var}_\mu(w\bar{\varphi}) \leq \overline{c} \left[\mathrm{var}(L, \varphi) + \mathrm{var}_\nu(\varphi) \right], \tag{13}$$

$$\mathrm{var}(K, w\varphi) + \mathrm{var}_\mu(w\bar{\varphi}) \geq \underline{c} \left[\mathrm{var}(L, \varphi) + \mathrm{var}_\nu(\varphi) \right]. \tag{14}$$

Remark 4. Here, we recall the notation $\bar{\varphi} := \varphi - \nu(\varphi)$. Regarding Theorem 3, whose proof is given in Appendix A:

(i) If $w = 1$ constant, in which case $\mu = \nu$, it reduces to [4, Lemma 32]. If also $(\underline{c}, \overline{c}) = (0, 1)$, it is the covariance ordering [37, Thm. 4.2], which is a

Peskun [41, 49] type criterion based on the Dirichlet form [see also 49, Proof of Lem. 3].

(ii) The assumptions are the same as those of [31, Lem. 13.22] about comparison of mixing times in the countable state space context.

(iii) (14) holds even if we 'forget' $\overline{c}$, i.e. set $\overline{c} = \infty$ but also require $w\varphi \in L^2(\mu)$. Unless $\mathbf{X}$ is compact, (14) is usually redundant since we can only assume $\underline{c} = 0$.

(iv) At least in the examples we will consider, one can take $\overline{c} := \|w\|_{L^\infty(\mu)}$ and $\underline{c} := \mu - \operatorname{ess\,inf} w$ to satisfy Assumption 2(a-b) (see Remark 19(ii)).

3.2. **Marginalisations and augmented importance sampling kernels.** Let $\mathbf{X} = \mathbf{T} \times \mathbf{Y}$ be a joint space. For a probability μ on $\mathbf{X}$, denote by $\mu^*(\mathrm{d}\theta) = \mu(\mathrm{d}\theta, \mathbf{Y})$ its marginal probability. If (μ, ν, w) on $\mathbf{X}$ satisfies Assumption 1, then $\nu^* \ll \mu^*$, and with $w^*(\theta) := \frac{d\nu^*}{d\mu^*}(\theta)$, the triplet (μ^*, ν^*, w^*) satisfies Assumption 1 on $\mathbf{T}$.

Assumption 3. Assumption 2, with Assumption 2(a–b) replaced with

(a) $\underline{c}\,\mathcal{E}_K(g) \le \mathcal{E}_L(g) \le \overline{c}\,\mathcal{E}_K(g)$, for all $g \in L^2(\mu^*)$, and

(b) $\underline{c} \le w^* \le \overline{c}$, μ^*-a.e.

We introduce the notion of an augmented Markov kernel, as in [9, 51].

Definition 2. Let $\dot{\mu}$ be some probability on $\mathbf{T}$, let $\dot{K}$ be a $\dot{\mu}$-invariant Markov kernel on $\mathbf{T}$, and let $Q_\theta(\mathrm{d}y)$ be a probability kernel from $\mathbf{T}$ to $\mathbf{Y}$. The *Q-augmentation of $\dot{K}$*, or the *Q-augmented kernel K*, is a Markov kernel on $\mathbf{X}$, with transition K and invariant measure μ, given by

$$K_{\theta\, y}(\mathrm{d}\theta', \mathrm{d}y') = \dot{K}_\theta(\mathrm{d}\theta') Q_{\theta'}(\mathrm{d}y'), \quad \text{and} \quad \mu(\mathrm{d}\theta, \mathrm{d}y) = \dot{\mu}(\mathrm{d}\theta) Q_\theta(\mathrm{d}y). \tag{15}$$

Theorem 5. *Suppose Assumption 3 holds, and that K is an augmented kernel as in Definition 2. Let $\varphi \in L^2(\nu)$ with $w\varphi \in L^2(\mu)$. With $\mathscr{N}_K := 0$ if K is positive, and $\mathscr{N}_K := 1$ if not, the following bound holds:*

$$\operatorname{var}(K, w\varphi) \le \overline{c}\big[\operatorname{var}(L, \varphi) + \operatorname{var}_\nu(\varphi)\big] + (1 + 2\mathscr{N}_K) \operatorname{var}_\mu(w\bar{\varphi}) \tag{16}$$

Moreover, if $w\varphi$ only depends on $\theta \in \mathbf{T}$, then (13) *holds.*

Remark 6. Regarding Theorem 5, whose proof is given in Appendix A:

(i) The function φ (and $w\varphi$) is allowed to depend on the auxiliary variable $y \in \mathbf{Y}$, unlike comparison results in the PM setting (see [8, Thm. 7] and [48, Thm. 1]) that are based on the convex order [9, Thm. 10].

(ii) K is positive iff $\dot{K}$ is positive (Lemma 24 of Appendix C). This is the case e.g. if $\dot{K}$ is a random walk MH kernel with normal proposals [13, Lem. 3.1]. See [23, Prop. 3] for more examples.

(iii) See also Remarks 19(ii–iii) in Appendix A about Assumption 3, which also hold for Assumption 2 by trivialising the space $\mathbf{Y}$ (Lemma 20(i)).

4. Jump chains and self-normalised importance sampling

4.1. **Jump chains.** We recall the notion of a jump chain [see 22], which is a Markov chain consisting of the accepted states of the original chain.

Definition 3. Let $(\Theta_k)_{k\geq 1}$ be a Markov chain with transition $K_\theta(\mathrm{d}\theta')$. The *jump chain* $(\tilde{\Theta}_k, \tilde{N}_k)_{k\geq 1}$ with transition $\tilde{K}_{\theta n}(\mathrm{d}\theta', \mathrm{d}n')$ and holding times

$$\tilde{N}_j := \min\left\{ i \geq 1 | \Theta_{\tilde{N}^*_{j-1}+i+1} \neq \Theta_{\tilde{N}^*_{j-1}+1} \right\}, \qquad j \geq 1,$$

is given by $\tilde{\Theta}_1 := \Theta_1$ and $\tilde{\Theta}_{k+1} := \Theta_{\tilde{N}^*_k+1}$, where $\tilde{N}^*_k := \sum_{j=1}^k \tilde{N}_j$, $\tilde{N}^*_0 := 0$.

For a Harris ergodic chain K, $(\tilde{N}_k)_{k\geq 1}$ are independent random variables given $(\tilde{\Theta}_k)_{k\geq 1}$, where $\tilde{N}_k$ is geometrically distributed with parameter $\alpha(\tilde{\Theta}_k)$. Here, $\alpha(\theta) := K(\theta, \mathbf{T}\backslash\{\theta\})$ is the acceptance probability function of K at $\theta \in \mathbf{T}$. See [51, Prop. 24] for this as well as for proof of the following result.

Lemma 7. *Let K be a μ-invariant Markov chain with $\alpha > 0$. The marginal chain $\tilde{K}$ of the jump chain of K has transition $\tilde{K}(\theta, A) = K(\theta, A\backslash\{\theta\})/\alpha(\theta)$, for all $A \in \mathcal{B}(\mathbf{T})$, and is $\tilde{\mu}$-invariant, where $\tilde{\mu}(\mathrm{d}\theta) = \alpha(\theta)\mu(\mathrm{d}\theta)/\mu(\alpha)$. Moreover, K is μ-reversible iff $\tilde{K}$ is $\tilde{\mu}$-reversible, and K is μ-Harris ergodic iff $\tilde{K}$ is $\tilde{\mu}$-Harris ergodic.*

We note that $(\tilde{\Theta}_k, \tilde{N}_k)_{k\geq 1}$ has as its transition the $Q^{(N)}$-augmentation of $\tilde{K}$ (Definition 2), where $\tilde{K}$ is as in Lemma 7 and $Q^{(N)}_\theta(\cdot) \sim \mathrm{Geo}(\alpha(\theta))$ [23].

Different estimators can sometimes be used in place of $(\tilde{N}_k)$, which can lead to lower asymptotic variance of the related MCMC than when not using the jump chain, or when using the jump chain with standard $(\tilde{N}_k)$ [22].

4.2. **Self-normalised importance sampling.** Jump chains can be naturally used with IS estimators, and can lead to improved computational and statistical efficiency [see 51]. To avoid redundancy, we shall adhere to the following convention: when we write $(\Theta_k, \mathsf{N}_k, \mathsf{a}, \mu)$, it shall stand simultaneously for $(\tilde{\Theta}_k, \tilde{N}_k, \alpha, \tilde{\mu})$, corresponding to an IS jump chain (denoted 'ISJ'), and for $(\Theta_k, 1, 1, \mu)$, corresponding to a non-jump IS chain (denoted 'IS0').

Suppose (μ, ν, w) satisfies Assumption 1 and that $(\Theta_k)_{k\geq 1}$ is μ-Harris ergodic. Often one can not evaluate $w(\theta)$. However, one can often evaluate an unnormalised version $w_u(\theta) = c_\xi \cdot w(\theta)$, with $c_\xi > 0$ a (unknown) constant. In this case, for $\varphi \in L^1(\nu)$, one can use the following SNIS estimator,

$$E_n^{SNIS}(\varphi) := \frac{\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)\varphi(\Theta_k)}{\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)} = \frac{\frac{1}{n}\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)\varphi(\Theta_k)}{\frac{1}{n}\sum_{k=1}^n \mathsf{N}_k w_u(\Theta_k)}. \tag{17}$$

By Harris ergodicity, the SNIS estimator is a consistent estimator for $\nu(\varphi)$,

$$E_n^{SNIS}(\varphi) \xrightarrow[\text{a.s.}]{n\to\infty} \frac{\mu(\mathbb{E}[\mathsf{N}_k|\Theta_k]w_u\varphi)}{\mu(\mathbb{E}[\mathsf{N}_k|\Theta_k]w_u)} = \frac{\mu(w_u\varphi/\mathsf{a})}{\mu(w_u/\mathsf{a})} = \nu(\varphi).$$

Next we consider a framework on an extended space, from which a Peskun type ordering for SNIS will trivially follow (Remark 13(ii) of Theorem 12).

5. Unbiased estimators and exact approximation schemes

In many settings, one relies on unbiased estimators in the MCMC [see 6]. We now describe a framework of unbiased estimators which we use and which is suitably general [see 51, 2]. We then describe direct and IS MCMC schemes to calculate $\nu(f)$, and give a general comparison result of their asymptotic variances.

5.1. **Framework.** Recall from Section 1 that our goal is calculation of expectations $\nu(f) = \int f(\theta, z)\nu(\mathrm{d}\theta, \mathrm{d}z)$, with respect to the joint probability ν on $\mathbf{T} \times \mathbf{Z}$, as well as with respect to its marginal probability $\dot{\nu}(\mathrm{d}\theta) = \nu(\mathrm{d}\theta, \mathbf{Z})$. For our unbiased estimators, define the spaces

$$\mathbf{U} := \big\{(\ell, \eta^{(1:\ell)}) : \ell \in \mathbb{N}, \eta^{(i)} \in [0, \infty), \text{ for } i = 1, \dots, \ell\big\}$$
$$\mathbf{V} := \big\{(m, z^{(1:m)}, \zeta^{(1:m)}) : m \in \mathbb{N}, \text{ and } z^{(i)} \in \mathbf{Z}, \zeta^{(i)} \in [0, \infty) \text{ for } i = 1, \dots m\big\}.$$

Let $Q_\theta^{(U)}(\mathrm{d}u)$ be a probability on $\mathbf{U}$ for each $\theta \in \mathbf{T}$, and $Q_{\theta u}^{(V)}(\mathrm{d}v)$ a probability on $\mathbf{V}$ for each $(\theta, u) \in \mathbf{T} \times \mathbf{U}$. Given $\Theta_k \in \mathbf{T}$, $(\ell_k, \eta_k^{(\ell_k)}) \in \mathbf{U}$, $(M_k, Z_k^{(1:m)}, \zeta_k^{(1:m)}) \in \mathbf{V}$, and a function f on $\mathbf{T} \times \mathbf{V}$, we define formally

$$\zeta_k(f) := \frac{1}{M_k} \sum_{i=1}^{M_k} \zeta_k^{(i)} f(\Theta_k, Z_k^{(i)}), \qquad \hat{\zeta}_k(f) = \frac{\zeta_k(f)}{\zeta_k(1)}, \qquad \xi_k := \frac{\zeta_k(f)}{\eta_k(1)}, \tag{18}$$

where $\eta_k(1) := \frac{1}{\ell_k} \sum_{i=1}^{\ell_k} \eta_k^{(i)}$. Let $\mathcal{L}^1(\nu)$ denote the set of functions f on $\mathbf{T} \times \mathbf{Z}$ such that there exists a constant $c_\zeta > 0$ such that for all $\theta \in \mathbf{T}$ and $f \in \mathcal{L}^1(\nu)$,

$$\int Q_\theta^{(U)}(\mathrm{d}u) Q_{\theta u}^{(V)}(\mathrm{d}v) \zeta(g) = c_\zeta \dot{\nu}(\theta) \int g(\theta, z) \nu(\mathrm{d}z|\theta)$$

for $g \in \{f, |f|\}$, where $\nu(\mathrm{d}z|\theta)$ denotes a regular conditional probability of ν given θ. Also, define $\mathcal{L}^2(\nu) = \{f \in \mathcal{L}^1(\nu) : f^2 \in \mathcal{L}^1(\nu)\}$.

Assumption 4. The following hold:

(i) The constant function $1 \in \mathcal{L}^1(\nu)$.
(ii) For all $\Theta_k \in \mathbf{T}$, the variables $U_k = (\ell_k, \eta_k^{(1:\ell_k)}) \sim Q_{\Theta_k}^{(U)}(\cdot)$ and $V_k = (M_k, Z_k^{(1:M_k)}, \zeta_k^{(1:M_k)}) \sim Q_{\Theta_k U_k}^{(V)}(\cdot)$ satisfy
$$\eta_k(1) = 0 \implies \zeta_k(1) = 0.$$

Remark 8. Regarding Assumption 4 and the above definitions:

(i) If $f \in L^1(\nu)$ satisfies $f(\theta, \cdot) = f(\theta)$, then $f \in \mathcal{L}^1(\nu)$. In many settings, $\mathcal{L}^1(\nu)$ may be much larger, or all of $L^1(\nu)$ [see 51, Cor. 28].
(ii) Support condition (ii) holds quite generally, e.g. if $\eta(1) > 0$. In a setting where, given θ, $\eta(1) = \mathrm{pr}(\theta)\eta'(1)$ and $\eta'(1)$ is an unbiased estimator for an approximate likelihood $L^{(U)}(\theta)$, this can be achieved by inflating the likelihood $L^{(U)}(\theta)$ and $\eta'(1)$ uniformly by a constant $\epsilon > 0$: $L^{(U)}(\theta) \mapsto L^{(U)}(\theta) + \epsilon$ and $\eta'(1) \mapsto \eta'(1) + \epsilon$ for all θ [see 51, Prop. 19 and Rem. 20].

5.2. **Pseudo-marginal type schemes and importance sampling schemes.** Define the probability $\pi(\mathrm{d}\theta, \mathrm{d}u, \mathrm{d}v) := c_\zeta^{-1} \mathrm{d}\theta Q_\theta^{(U)}(\mathrm{d}u) Q_{\theta u}^{(V)}(\mathrm{d}v)\zeta(1)$. The following concerns a PM/DA type scheme [6, 17].

Proposition 9. *Suppose a Markov chain $(\Theta_k, U_k, V_k)_{k \geq 1}$ is π-reversible Harris ergodic, where Assumption 4 holds. Then, for all $f \in \mathcal{L}^1(\nu)$,*

$$E_n^{PM}(f) := \frac{1}{n} \sum_{k=1}^{n} \hat{\zeta}_k(f) \xrightarrow[a.s.]{n \to \infty} \nu(f). \tag{19}$$

Proof. Follows by Harris ergodicity, as $\pi(\hat{\zeta}(f)) = \nu(f)$, $f \in \mathcal{L}^1(\nu)$. ■

Algorithm 5 Importance sampling scheme. Suppose Assumption 4 holds.

(Phase 1) Let $(\Theta_k, U_k)_{k\geq 1}$ be a μ-reversible Harris ergodic Markov chain.
(Phase 2) For each $k \geq 1$, let V_k be drawn as follows, for the IS0 and ISJ cases:

(IS0) $V_k \sim Q^{(V)}_{\Theta_k U_k}(\cdot)$. For $f \in \mathcal{L}^1(\nu)$, we define

$$\mathbf{m}_f(\theta, u) := \mathbb{E}[\xi_k(f)|\Theta_k = \theta, U_k = u]. \tag{20}$$

(ISJ) Form a jump chain $(\tilde{\Theta}_k, \tilde{U}_k, \tilde{N}_k)_{k\geq 1}$, and draw V_k from some kernel $V_k \sim Q^{(V|N)}_{\tilde{\Theta}_k \tilde{U}_k \tilde{N}_k}(\cdot)$ from $\mathbf{T} \times \mathbf{U} \times \mathbb{N}$ to $\mathbf{V}$ such that

$$\mathbb{E}[\xi_k(f)|\tilde{\Theta}_k = \theta, \tilde{U}_k = u, \tilde{N}_k = n] = \mathbf{m}_f(\theta, u)$$

for all $n \in \mathbb{N}$ and $f \in \mathcal{L}^1(\nu)$.

Define the probability $\mu(\mathrm{d}\theta, \mathrm{d}u) := c_\eta^{-1} \mathrm{d}\theta Q^{(U)}_\theta(\mathrm{d}u)\eta(1)$ where $c_\eta > 0$ is a normalising constant. Set $\dot{\mu}(\mathrm{d}\theta) = \mu(\mathrm{d}\theta, \mathbf{U})$. Note that under Assumption 4, we have $\dot{\mu}(\theta) = 0$ implies $\dot{\nu}(\theta) = 0$.

Consider now an IS scheme (Algorithm 5) as in [51]. Compared to [51], we additionally assume μ-reversibility of the base chain and nonnegativity of the estimators $\zeta^{(i)} \geq 0$. This is done to facilitate comparison with the previous PM type scheme corresponding to PM and DA algorithms, which are π-reversible and require $\zeta^{(i)} \geq 0$, as $\zeta(1)$ is present in their acceptance ratio (see Section 6). If Assumption 4 (PM kernels) holds, then for all $f \in \mathcal{L}^1(\nu)$,

$$\mu(\mathbf{m}_f) = \frac{1}{c_\eta} \int \mathrm{d}\theta Q^{(U)}_\theta(\mathrm{d}u)\eta(1) Q^{(V)}_{\theta u}(\mathrm{d}v)\xi(f) = c_\xi \nu(f)$$

where $c_\xi := c_\zeta / c_\eta$, and $\mathbf{m}_f$ is defined in (20). This motivates the following consistency result, an instance of [51, Thm. 3] for example for the $\mathsf{N}_k = 1$ case (IS0) and [51, Thm. 13] for the $\mathsf{N}_k = \tilde{N}_k$ case (ISJ).

Proposition 10. *Under Algorithm 5, for all $f \in \mathcal{L}^1(\nu)$,*

$$E_n^{\mathrm{IS}}(f) := \frac{\sum_{k=1}^n \mathsf{N}_k \xi_k(f)}{\sum_{k=1}^n \mathsf{N}_k \xi_k(1)} \xrightarrow[a.s.]{n\to\infty} \nu(f). \tag{21}$$

Remark 11. In the ISJ case, permitting dependence on $\tilde{N}_k$ when drawing V_k in Algorithm 5 allows for variance reduction of $\xi_k(f)$ and hence of the resultant estimator (21) (see Proposition 21), by using larger M_k when $\tilde{N}_k$ is large. For example, M_k could correspond to the number of independent samples drawn from an instrumental or to the number of particles used in a particle filter [2].

5.3. **A Peskun type ordering for importance sampling schemes.** Under Assumption 5 below, the IS estimator $E_n^{\mathrm{IS}}(f)$ (21) satisfies a CLT

$$\sqrt{n}[E_n^{\mathrm{IS}}(f) - \nu(f)] \xrightarrow{n\to\infty} \mathcal{N}\big(0, \mathbb{V}_f^{\mathrm{IS}}\big), \qquad \text{in distribution.} \tag{22}$$

See [51] or Proposition 21 of Appendix A, with a formula for $\mathbb{V}_f^{\mathrm{IS}}$. In analogy with Definition 1 and (7), we refer to $\mathbb{V}_f^{\mathrm{IS}}$ as the *IS asymptotic variance*. The following assumption is sufficient for $\mathbb{V}_f^{\mathrm{IS}} < \infty$.

Assumption 5 (Importance sampling CLT)**.** Suppose Algorithm 5 (IS scheme) and that $(\Theta_k, \mathsf{U}_k, \mathsf{N}_k)_{k\geq 1}$ is aperiodic. Let $f \in \mathcal{L}^2(\nu)$ be a function such that $\mathrm{var}(K, \mathbf{m}_f) < \infty$, where $\mathbf{m}_f$ is defined in (20), and $\mathsf{v}_{\bar{f}}$ by

(IS0) $v_{\bar{f}}(\theta, u) := \mathrm{var}\big(\xi_k(\bar{f})|\Theta_k = \theta, U_k = u\big)$,

(ISJ) $\tilde{v}_{\bar{f}}(\theta, u) := \mathbb{E}[\tilde{N}_k^2 \mathrm{var}\big(\xi_k(\bar{f})|\tilde{\Theta}_k = \theta, \tilde{U}_k = u, \tilde{N}_k\big)|\tilde{\Theta}_k = \theta, \tilde{U}_k = u]$,

satisfies $\mu\big(\mathsf{a}\mathsf{v}_{\bar{f}}\big) < \infty$.

Let us denote the kernel and measure of the IS0 corrected chain of Algorithm 5 by $(\bar{K}, \bar{\mu})$ on the space $\mathbf{X} = (\mathbf{T} \times \mathbf{U}) \times \mathbf{V}$, where,

$$\bar{K}_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v') := K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v')$$
$$\bar{\mu}(\mathrm{d}\theta, \mathrm{d}u, \mathrm{d}v) := \mu(\mathrm{d}\theta, \mathrm{d}u) Q^{(V)}_{\theta u}(\mathrm{d}v). \tag{23}$$

Note that $\bar{K} = K^{(V)}$ is an augmented kernel (Definition 2). Note too that by Slutsky's lemma like in (9), we have $\mathbb{V}_f^{\mathrm{IS}} = \mathrm{var}\big(\bar{K}, w\hat{\zeta}(f)\big)$, where $w = \mathrm{d}\pi/\mathrm{d}\bar{\mu}$.

With definitions as in Assumption 5, we define a 'difference' constant $\mathsf{D}_{\bar{f}}$, for the IS0 and ISJ cases, respectively, by $D_{\bar{f}} := 0$ and

$$\tilde{D}_{\bar{f}} := \mu(a) c_\xi^{-2} \mu(a\tilde{v}_{\bar{f}} - v_{\bar{f}}).$$

Theorem 12. *Suppose the assumptions of Algorithm 5 (IS scheme) hold, and* $\mathbb{V}_f^{IS} < \infty$.

(i) If $(\bar{\mu}, \pi, w, \bar{K}, L, \underline{c}, \overline{c})$ *satisfies Assumption 2 on* $\mathbf{X}$*, then*

$$\mathbb{V}_f^{IS} + \mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \leq \overline{c}\,\mu(\mathsf{a})\big\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\big\} + \mathsf{D}_{\bar{f}}$$
$$\mathbb{V}_f^{IS} + \mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \geq \underline{c}\,\mu(\mathsf{a})\big\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\big\} + \mathsf{D}_{\bar{f}}.$$

(ii) If $(\bar{\mu}, \pi, w, \bar{K}, L, \underline{c}, \overline{c})$ *satisfies Assumption 3 on* $\mathbf{X}$*, then*

$$\mathbb{V}_f^{IS} \leq \overline{c}\,\mu(\mathsf{a})\big\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\big\} + (1 + 2\mathscr{N}_K)\mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) + \mathsf{D}_{\bar{f}}$$

where $\mathscr{N}_K := 0$ *if* K *is positive, and* $\mathscr{N}_K := 1$ *if not.*

Remark 13. Regarding Theorem 12, whose proof is in Appendix A:

(i) Note that $0 \leq \mu(\mathsf{a}) \leq 1$, with a as in Section 4.2, and that $w = c_\xi^{-1}\xi(1)$ and $w^* = c_\xi^{-1}\mathbf{m}_1$, with $\mathbf{m}_f(\theta, u)$ defined in (20).

(ii) As a trivialisation, when $\eta(\Theta_k, U_k) := \eta(1) = \dot{\mu}(\Theta_k)$ a.s., $\mathbf{Z} = \{0\}$, and $\xi_k(f) = w_u(\Theta_k)f(\Theta_k)$ a.s., we obtain a Peskun type ordering for SNIS (17). Here, the simplifications are $\bar{K} \leftrightarrow K$, $\hat{\zeta}(\bar{f}) \leftrightarrow \bar{f}$ and $\xi(\bar{f}) \leftrightarrow c_\xi w\bar{f}$.

6. Pseudo-Marginal and delayed-acceptance MCMC

We define PM and DA type algorithms in the setting of the auxiliary variable framework of Section 5, where PM could be the 'particle marginal MH' [2]; a DA type variant of this algorithm has been implemented e.g. in [26, 42, 51]. After defining the corresponding kernels, we then compare the asymptotic variances of PM/DA with IS (Theorem 16).

Algorithm 6 Pseudo-Marginal parent. Suppose Assumption 4 (PM kernels) holds. Initialise $X_0 \in \mathbf{X}$ with $\zeta_0(1) > 0$. For $k = 1, \dots n$, do:

(1) Draw $\Theta'_k \sim q_{\Theta_{k-1}}(\cdot)$ and $U'_k \sim Q^{(U)}_{\Theta'_k}(\cdot)$ and $V'_k \sim Q^{(V)}_{\Theta'_k U'_k}(\cdot)$. With probability $\min\left\{1, r^{(V)}(X_{k-1}, X'_k)\right\}$ accept X'_k; otherwise, reject.

Algorithm 7 Delayed-acceptance (DA0). Suppose Assumption 4 (PM kernels) holds, and K is a μ-proposal-rejection kernel of the form (26). Initialise $X_0 \in \mathbf{X}$ with $\zeta_0(1) > 0$. For $k = 1, \dots, n$, do:

(1) Draw $\Theta'_k \sim q_{\Theta_{k-1}}(\cdot)$. Construct $U'_k \sim Q^{(U)}_{\Theta'_k}(\cdot)$. With probability $\alpha(\Theta_{k-1}, U_{k-1}; \Theta'_k, U'_k)$, proceed to step (2). Otherwise, reject.
(2) Construct $V'_k \sim Q^{(V)}_{\Theta'_k, U'_k}(\cdot)$. With probability $\min\left\{1, \xi'_k(1)/\xi_k(1)\right\}$, accept (Θ'_k, U'_k, V'_k); otherwise, reject.

Algorithm 8 Delayed-acceptance (DA1). Suppose Assumption 4 (PM kernels) holds. Initialise $X_0 \in \mathbf{X}$ with $\zeta_0(1) > 0$. For $k = 1, \dots, n$, do:

(1) Draw $(\Theta'_k, U'_k) \sim K_{\Theta_{k-1}, U_{k-1}}(\cdot)$.
(2) Construct $V'_k \sim Q^{(V)}_{\Theta'_k, U'_k}(\cdot)$. With probability $\min\left\{1, \xi'_k(1)/\xi_k(1)\right\}$, accept (Θ'_k, U'_k, V'_k); otherwise, reject.

6.1. **Algorithms.** Let $q_\theta(\mathrm{d}\theta') = q_\theta(\theta')\mathrm{d}\theta'$ be a proposal kernel on $\mathbf{T}$. Assume the setup of Assumption 4 (recall that $\eta(1) \geq 0$ and $\zeta(1) \geq 0$). Whenever the denominators are not zero we define the following 'acceptance ratios' for $x, x' \in \mathbf{X} := \mathbf{T} \times \mathbf{U} \times \mathbf{V}$, where $x = (\theta, u, v)$,

$$r^{(U)}(x, x') := \frac{\eta'(1) q_{\theta'}(\theta)}{\eta(1) q_\theta(\theta')}, \quad \text{and} \quad r^{(V)}(x, x') := \frac{\zeta'(1) q_{\theta'}(\theta)}{\zeta(1) q_\theta(\theta')}. \tag{24}$$

Consider Algorithm 6 ('PM parent,' following the terminology of [47]), Algorithm 7 ('DA0'), and Algorithm 8 ('DA1'), with transition kernels given later and which are π-invariant [see 2, 6, 10]. Under Assumption 4 (PM kernels) and the assumption that the resultant chains are π-Harris ergodic, by construction Algorithms (6-8) produce output as in Proposition 9 (PM type scheme). In PM parent (Algorithm 6) and DA1 (Algorithm 8), the computationally expensive V_k-variable is drawn whenever U_k is drawn. This is the essential difference with DA0 (Algorithm 7). The separation of sampling steps can substantially reduce computational cost in DA0 [see 17], even though the asymptotic variance of DA0 is more than PM parent in the case K is the approximate PM kernel (28) [see 10], and more than DA1 in the case K is a 'μ-proposal-rejection chain' (e.g. PM); see Propositions 14 and 15 below).

6.2. **Kernels.** Let K be the transition kernel of a μ-reversible Harris ergodic IS0 base chain $(\Theta_k, U_k)_{k \geq 1}$, with definitions as in Assumption 4 (PM kernels). The *DA1 correction* of K is the π-reversible kernel K^{DA1} corresponding to Algorithm

8, given by,

$$K^{\mathrm{DA1}}_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v') = K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \xi'(1)/\xi(1)\right\} + [1 - \alpha_{\mathrm{DA1}}(\theta, u, v)]\delta_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v'), \tag{25}$$

where $\alpha_{\mathrm{DA1}}(\theta, u, v) := \int K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \xi'(1)/\xi(1)\right\}$.

Let K be 'μ-proposal-rejection kernel,' that is, a μ-reversible kernel of the form

$$K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') = q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta, u; \theta', u') + r_K(\theta, u)\delta_{\theta u}(\mathrm{d}\theta' \mathrm{d}u') \tag{26}$$

for some function $\alpha : (\mathbf{T} \times \mathbf{U})^2 \to [0, 1]$ and $r_K = 1 - \int q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta, u; \theta', u')$. The *DA0 correction* of K is defined to be

$$K^{\mathrm{DA0}}_x(\mathrm{d}x') = q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta, u; \theta', u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \xi'(1)/\xi(1)\right\} + [1 - \alpha_{\mathrm{DA0}}(x)]\delta_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v'), \tag{27}$$

where $\alpha_{\mathrm{DA0}}(x) = \int q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \alpha(\theta, u; \theta', u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, \xi'(1)/\xi(1)\right\}$, and $\mathbf{X} := \mathbf{T} \times \mathbf{U} \times \mathbf{V}$, $x \in \mathbf{X}$, $x := (\theta, u, v)$.

Decreasing the variability of $\xi'(1) = \zeta'(1)/\eta'(1)$ by coupling the u' and v' variables can lead to improved mixing of (27), and is similar in idea to recently proposed 'correlated PM' [19] and 'MHAAR' [3] chains. The mere requirement of reversibility allows the kernel K to be taken to be approximate versions of the two chains listed above, or an approximate DA or 'multi-stage DA' [10]. Regardless, the most straightforward choice for K is the (approximate) PM kernel targeting μ with proposal q, given by,

$$K_{\theta u}(\mathrm{d}\theta', \mathrm{d}u') = q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \min\left\{1, r^{(U)}(x, x')\right\} + [1 - \alpha(\theta, u)]\delta_{\theta u}(\mathrm{d}\theta', \mathrm{d}u'), \tag{28}$$

where $\alpha(\theta, u) := \int q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') \min\left\{1, r^{(U)}(x, x')\right\}$.

The asymptotic variance of DA1 is never more than that of DA0.

Proposition 14. *If K is the μ-proposal-rejection kernel* (26)*, then:*

(i) $\mathrm{var}(K^{DA1}, g) \leq \mathrm{var}(K^{DA0}, g)$ *for all* $g \in L^2(\pi)$.

(ii) If $V \sim Q^{(V)}_{\theta u}(\cdot)$ *with* $V = (M, Z^{(1:M)}, \zeta^{(1:M)})$ *has the property that*

$$\zeta(1) = \varphi(\theta, u) \tag{29}$$

is a deterministic function φ of θ and u, then $K^{DA0} = K^{DA1}$.

However, for the reason discussed in Section 6.1, DA0 is likely more computationally efficient than DA1 in practice.

We define the PM parent kernel P of K^{DA1} to be given by

$$P_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v') = q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, r^{(V)}(x, x')\right\} + [1 - \alpha_{\mathrm{PMP}}(\theta, v)]\delta_{\theta uv}(\mathrm{d}\theta', \mathrm{d}u', \mathrm{d}v'), \tag{30}$$

where $\alpha_{\mathrm{PMP}}(\theta, v) := \int q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\left\{1, r^{(V)}(x, x')\right\}$.

We define a probability kernel from $\mathbf{T}$ to $\mathbf{V}$ by

$$\hat{Q}^{(V)}_\theta(\mathrm{d}v) := \int_{\mathbf{U}} Q^{(U)}_\theta(\mathrm{d}u) Q^{(V)}_{\theta u}(\mathrm{d}v) \tag{31}$$

We then define the following PM kernel with proposal q,

$$M_{\theta v}(\mathrm{d}\theta', \mathrm{d}v') = q_\theta(\mathrm{d}\theta')\hat{Q}^{(V)}_{\theta'}(\mathrm{d}v') \min\left\{1, r^{(V)}(x, x')\right\} + [1 - \alpha_{\mathrm{PM}}(\theta, v)]\delta_{\theta v}(\mathrm{d}\theta', \mathrm{d}v'), \tag{32}$$

targeting $\hat{\pi}(\mathrm{d}\theta, \mathrm{d}v) := \int_{\mathbf{U}} \pi(\mathrm{d}\theta, \mathrm{d}u, \mathrm{d}v)$, where $\alpha_{\mathrm{PM}}(\theta, v) := \int q_\theta(\mathrm{d}\theta')\hat{Q}^{(V)}_{\theta'}(\mathrm{d}v') \min\left\{1, r^{(V)}(x, x')\right\}$.

When U_k and V_k are independent given θ, i.e.

$$Q^{(V)}_{\theta u}(\mathrm{d}v) = Q^{(V)}_{\theta}(\mathrm{d}v), \tag{33}$$

then M (32) is the standard PM with proposal q, since,

$$\hat{Q}^{(V)}_{\theta}(\mathrm{d}v) = Q^{(V)}_{\theta}(\mathrm{d}v).$$

Proposition 15. *If K is the approximate PM* (28)*, and $L \in \{M, P\}$, then:*

(i) $\mathrm{var}(L, \hat{\zeta}(f)) \le \mathrm{var}(K^{DA0}, \hat{\zeta}(f))$ *for all* $f \in L^2_\pi(\nu)$.

(ii) If (29) *holds, then for all* $f \in L^2_\pi(\nu)$,

$$\mathrm{var}(L, \hat{\zeta}(f)) \le \mathrm{var}(K^{DA1}, \hat{\zeta}(f)).$$

6.3. **Comparison with importance sampling correction.** Note that the following result only involves the weight, not the Dirichlet forms.

Theorem 16. *Suppose Assumption 4 (PM kernels) holds, and that one of the following conditions for pairs of kernels holds:*

(I) $L = K^{DA0}$ *is DA0 correction* (27)*, and K is μ-proposal-rejection* (26)*,*

(II) $L = K^{DA1}$ *is DA1 correction* (25)*, and K is μ-reversible,*

(III) $L = P$ *is the PM parent* (30)*, and K is the approx. PM* (28)*, or*

(IV) $L = M$ *is the PM kernel* (32)*, and K is the approx. PM* (28)*.*

Assume K and L are Harris ergodic, and a function $f \in \mathcal{L}^2(\nu)$ is such that $\mathbb{V}^{IS}_f < \infty$. The following statements hold:

(i) The IS asymptotic variance (22) *satisfies, with $\underline{c} := \bar{\mu}\text{-ess}\inf w$,*

$$\mathbb{V}^{IS}_f + \mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \le \mu(\mathsf{a})\,\|w\|_{L^\infty(\bar{\mu})}\left\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\right\} + \mathsf{D}_{\bar{f}}$$

$$\mathbb{V}^{IS}_f + \mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) \ge \mu(\mathsf{a}) \cdot \underline{c} \cdot \left\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\right\} + \mathsf{D}_{\bar{f}}.$$

(ii) With $\mathscr{N}_K := 0$ if K is positive and $\mathscr{N}_K := 1$ if not, the following holds:

$$\mathbb{V}^{IS}_f \le \mu(\mathsf{a})\,\|w^*\|_{L^\infty(\mu)}\left\{\mathrm{var}\big(L, \hat{\zeta}(f)\big) + \mathrm{var}_\pi\big(\hat{\zeta}(f)\big)\right\} + (1 + 2\mathscr{N}_K)\mu(\mathsf{a})\mathrm{var}_{\bar{\mu}}\big(w\hat{\zeta}(\bar{f})\big) + \mathsf{D}_{\bar{f}}.$$

See Remark 13(i) for w and w^*. See Appendix B for the proof of Theorem 16, which follows from Theorem 12, after bounding the Dirichlet forms.

7. Discussion and further stability considerations

A necessary condition for a successful implementation of an IS or PM scheme is a simple support condition, Assumption 4(ii), that can often be easily ensured by Remark 8(ii). On the other hand, Theorem 16 depends on a uniform bound on the marginal weight $w^* \propto \mathbf{m}_1$, with $\mathbf{m}_f(\theta, u)$ as in (20). This bound is much weaker than a bound on w, and can often be ensured. For example, assuming that $\eta(1)\mathbf{m}_1$ is bounded, one can often inflate $\eta(1)$ as in Remark 8(ii) to obtain

an uniform bound on w^*. Other techniques may be applicable if a bounded w^* is particularly desired, such as a combination of cutoff functions, approximations, or tempering [see 39, 51].

When considering a PM/DA implementation, the issue of boundedness of the full weight $w \propto \zeta(1)/\eta(1)$ takes particular importance, more so than in the case with IS. This is because PM and DA are more liable to be poorly mixing, while IS is less affected by noisy estimators. Namely, if $\zeta(1)$ is not bounded, then PM parent and K^{DA0}, with K as in (28), are not geometrically ergodic (Proposition 25).

On the other hand, the IS chain may converge fast, even in the case of unbounded $\zeta(1)$. For example, if K is a random walk MH chain, then K is geometrically ergodic essentially if μ has exponential or lighter tails and a certain contour regularity condition holds [28, 45], where we have said nothing about the exact level estimator $\zeta(1)$. We then apply Lemma 24(v), which says that whenever K is geometrically ergodic then so is $\bar{K}$, to conclude that the IS chain is geometrically ergodic, even in the case of unbounded $\zeta(1)$. This may be beneficial if adaptation is used [5, 7, 44].

Of course, high variability affects also the IS estimator, but we believe this noise to be a smaller issue in IS, as the noise is in the IS output estimator rather than in the acceptance ratio as in PM/DA. This can make a significant difference in the evolution and ergodicity of the chains, as described above.

Acknowledgments

Support has been provided for JF and MV from the Academy of Finland (grants 274740, 284513 and 312605), and for JF from The Alan Turing Institute. JF thanks the organisers of the 2017 SMC course and workshop in Uppsala.

Appendix A. Proofs for the Peskun type orderings

A.1. **Subprobability kernels.** Let K be a μ-reversible Markov kernel on $\mathbf{X}$. For all $\lambda \in (0,1]$, λK is a *subprobability kernel*: $\lambda K(x, \mathbf{X}) \leq 1$ for all $x \in \mathbf{X}$. The *Dirichlet form* $\mathcal{E}_{\lambda K}(f)$ of the subprobability kernel λK is

$$\mathcal{E}_{\lambda K}(f) := \langle f, (1-\lambda K) f \rangle_\mu = \lambda \mathcal{E}_K(f) + (1-\lambda) \|f\|_\mu^2 , \tag{34}$$

defined for $f \in L^2(\mu)$. For $f \in L_0^2(\mu)$, if $(1-K)^{-1} f$ exists in $L^2(\mu)$, then by (6), $\mathrm{var}(K, f) = 2 \langle f, (1-K)^{-1} f \rangle_\mu - \mu(f^2)$ [see 9]. Following [9, 49], we then (formally) extend Definition 1 of the asymptotic variance to subprobability kernels: for $\lambda \in (0,1)$, the operator $(1-\lambda K)$ is always invertible, and we define

$$\mathrm{var}(\lambda K, f) := 2 \left\langle f, (1-\lambda K)^{-1} f \right\rangle_\mu - \mu(f^2). \tag{35}$$

Moreover, (12) and (34) imply for $\lambda \in (0,1]$ that $1-\lambda K$ is a positive operator, i.e. $\mathcal{E}_{\lambda K}(f) \geq 0$ for all $f \in L^2(\mu)$. By a result attributed to Bellman [14, Eq. 14], for positive self-adjoint operators, and used e.g. in [1, 9, 15, 38, 37], we have another asymptotic variance representation: for all $\lambda \in (0,1)$ and $f \in L_0^2(\mu)$,

$$\mathrm{var}(\lambda K, f) = 2 \sup_{g \in L^2(\mu)} \left\{ 2 \langle f, g \rangle_\mu - \mathcal{E}_{\lambda K}(g) \right\} - \mu(f^2). \tag{36}$$

Here, the supremum is attained with $g := (1-\lambda K)^{-1} f$, in which case (36) simplifies to (35). For $\lambda \in (0,1)$, equalities (35–36) hold and are finite for any $f \in L_0^2(\mu)$. The function $\lambda \mapsto \mathrm{var}(\lambda K, f)$ has a limit as $\lambda \uparrow 1$ on the extended real numbers $[0,\infty]$, and $\mathrm{var}(K,f)$ equals this limit [49].

A.2. **Normalised importance sampling ordering.** We set

$$\mathscr{N}_K := -\inf_{\mu(g)=0,\mu(g^2)=1} \langle g, Kg\rangle_\mu \tag{37}$$

for a μ-reversible kernel K, so that the *left spectral gap* of K is $1-\mathscr{N}_K$ [see 9]. We have $\mathscr{N}_K \in [-1,1]$ in general, but $\mathscr{N}_K \in [-1,0]$ if K is positive.

The conditions of the next two lemmas will seem more natural once Lemma 20 is stated.

Lemma 17. *Suppose $(\mu,\nu,w,K,L,\underline{c},\overline{c})$ satisfies Assumption 3 on $\mathbf{X} := \mathbf{T}\times\mathbf{Y}$. Let $\varphi \in L_0^2(\nu)$ be such that $w\varphi \in L^2(\mu)$. Define $u_\lambda := (1-\lambda K)^{-1}(w\varphi)$ and $\check{u}_\lambda := u_\lambda - w\varphi$, in $L^2(\mu)$ for all $\lambda \in (0,1)$. The following hold:*

(i) If $u_\lambda(\theta,y) = u_\lambda(\theta)$, $\lambda\in(0,1)$, then (13) holds.
(ii) If $\check{u}_\lambda(\theta,y) = \check{u}_\lambda(\theta)$, $\lambda\in(0,1)$, then (16) holds, with $\mathscr{N}_K$ as in (37).

Proof. Note that $L^2(\mu^*) \subset L^2(\nu^*)$ by Assumption 3(b). For $g \in L^2(\mu^*)$,

$$\mathcal{E}_{\lambda L}(g) = \lambda\mathcal{E}_L(g) + (1-\lambda)\nu^*(g^2) \le \overline{c}\lambda\mathcal{E}_K(g) + (1-\lambda)\nu^*(g^2),$$

by Assumption 3(a). From the above first equality, now for λK and μ^*,

$$\begin{aligned}\mathcal{E}_{\lambda L}(g) &\le \overline{c}\big[\mathcal{E}_{\lambda K}(g) - (1-\lambda)\mu^*(g^2)\big] + (1-\lambda)\nu^*(g^2)\\ &= \overline{c}\mathcal{E}_{\lambda K}(g) - (1-\lambda)\mu^*\big(g^2[\overline{c}-w^*]\big) \le \overline{c}\mathcal{E}_{\lambda K}(g),\end{aligned} \tag{38}$$

by Assumption 3(b). Since $1-\lambda K$ is self-adjoint on $L^2(\mu)$, we also note that

$$\mathcal{E}_{\lambda K}(\check{u}_\lambda) = \mathcal{E}_{\lambda K}(u_\lambda - w\varphi) = \mathcal{E}_{\lambda K}(u_\lambda) + \mathcal{E}_{\lambda K}(w\varphi) - 2\,\|w\varphi\|_\mu^2,$$

as $\langle v_\lambda, (1-\lambda K)w\varphi\rangle_\mu = \|w\varphi\|_\mu^2$. Regardless of $\lambda\in(0,1)$, $1-\lambda K$ has support of its spectral measure contained in $[0, 1+\mathscr{N}_K]$. Hence, $\mathcal{E}_{\lambda K}(w\varphi) \le (1+\mathscr{N}_K)\,\|w\varphi\|_\mu^2$, so

$$\mathcal{E}_{\lambda K}(\check{u}_\lambda) \le \mathcal{E}_{\lambda K}(u_\lambda) + (\mathscr{N}_K - 1)\,\|w\varphi\|_\mu^2. \tag{39}$$

We now compare the asymptotic variances. By (35),

$$LS := \mathrm{var}(\lambda K, w\varphi) + \|w\varphi\|_\mu^2 = 2\,\langle w\varphi, u_\lambda\rangle_\mu = 2\big[2\,\langle w\varphi, u_\lambda\rangle_\mu - \mathcal{E}_{\lambda K}(u_\lambda)\big].$$

With $\psi := u_\lambda$ for (i), and with $\psi := \check{u}_\lambda$ for (ii) using (39),

$$LS \le 2\big[2\,\langle w\varphi, \psi\rangle_\mu - \mathcal{E}_{\lambda K}(\psi)\big] + E_\psi,$$

where $E_\psi := 0$ if $\psi = u_\lambda$ and $E_\psi := 2(1+\mathscr{N}_K)\,\|w\varphi\|_\mu^2$ if $\psi = \check{u}_\lambda$. Hence,

$$LS \le 2\big[2\,\langle \varphi, \psi\rangle_\nu - \mathcal{E}_{\lambda K}(\psi)\big] + E_\psi \le 2\big[2\,\langle \varphi, \psi\rangle_\nu - (\overline{c})^{-1}\mathcal{E}_{\lambda L}(\psi)\big] + E_\psi,$$

where we have used (38). Since $\psi \in L^2(\mu^*) \subset L^2(\nu)$,

$$\begin{aligned}LS &\le \frac{1}{\overline{c}}\Big(2\sup_{g\in L^2(\nu)}\big\{2\,\langle \overline{c}\varphi, g\rangle_\nu - \mathcal{E}_{\lambda L}(g)\big\} - \|\overline{c}\varphi\|_\nu^2\Big) + \overline{c}\,\|\varphi\|_\nu^2 + E_\psi\\ &= \overline{c}\big(\mathrm{var}(\lambda L, \varphi) + \|\varphi\|_\nu^2\big) + E_\psi,\end{aligned}$$

by (36). We then take the limit $\lambda \uparrow 1$ [49]. Noting that $\|w\varphi\|_\mu^2 = \mathrm{var}_\mu(w\varphi)$ since $\mu(w\varphi) = \nu(\varphi) = 0$, we conclude. ■

Lemma 18. *Suppose the assumptions of Lemma 17 hold, where $\overline{c}$ may be also ∞. If $v_\lambda := (1-\lambda L)^{-1}(\varphi)$ satisfies $v_\lambda(\theta, y) = v_\lambda(\theta)$, then* (14) *holds.*

Proof. The lower bound (14) is trivial if $\underline{c} = 0$. Assume $\underline{c} > 0$. Then $\mu \ll \nu$, $w^{-1} \le \underline{c}^{-1}$ (implying $L^2(\nu) \subseteq L^2(\mu)$), and $\mathcal{E}_K(g) \le \underline{c}^{-1}\mathcal{E}_L(g)$ for all $g \in L^2(\nu)$. The result follows by applying Lemma 17(i). ■

Remark 19. Regarding Lemma 17 and Lemma 18:

(i) The solution v_λ to the Poisson equation [see 36], $(1-\lambda L)g = \varphi$ in $L^2(\mu)$, is also used in [9, Thm. 17] as a lemma for the proof of the convex order criterion Peskun type ordering for PM chains [9, Thm. 10].

(ii) It is reasonable to use a single constant $\overline{c}$ in Assumptions 3(a–b). If one replaces Assumption 3(b) with $w^* \le \overline{c}'\ \mu^* - a.e.$, then, if $\overline{c}' < \overline{c}$, one obtains the same result after bounding a nonpositive quantity by zero in (38). If $\overline{c}' > \overline{c}$, then one would need to impose the unappealing condition that $\sup_{\lambda\in(0,1)} \|u_\lambda\|_{\mu^*}^2 < \infty$ and add a positive constant involving this bound to the final results. Anyways, for the the application in this paper, we have $\overline{c} = \overline{c}'$ (Lemma 23).

(iii) Assumption 3(a) can be replaced with the weaker assumption that $\mathcal{E}_L(g) \le \overline{c}\,\mathcal{E}_K(g)$ for all $g \in \mathcal{G} \subset L^2(\mu^*)$, where $\mathcal{G} := \{u_\lambda : \lambda \in (0,1)\}$ for (i) and $\mathcal{G} := \{\check{u}_\lambda : \lambda \in (0,1)\}$ for (ii).

Lemma 20. *Let K be a μ-reversible chain on $\mathbf{X} = \mathbf{T} \times \mathbf{Y}$. For $h \in L^2(\mu)$ and $\lambda \in (0,1)$, set $h_\lambda := (1-\lambda K)^{-1}h$ and $\check{h}_\lambda := h_\lambda - h$, which are in $L^2(\mu)$.*

(i) If $\mathbf{Y} = \{y_0\}$ is the trivial space, then $h_\lambda(\theta, y) = h_\lambda(\theta)$.

(ii) If K is an augmented kernel, then $\check{h}_\lambda(\theta, y) = \check{h}_\lambda(\theta)$. Moreover, if also $h(\theta, y) = h(\theta)$, then $h_\lambda(\theta, y) = h_\lambda(\theta)$.

Proof. (i) is clear. For (ii), we write the series representation for the inverse of an invertible operator and use Lemma 24(iii), to get that,

$$h_\lambda(\theta, y) = \sum_{n=0}^{\infty} \lambda^n K^n h(\theta, y) = h(\theta, y) + \sum_{n=1}^{\infty} \lambda^n \dot{K}^n (Qh)(\theta).$$

The result then follows. ■

Proof of Theorem 3. The upper bound (13) follows from Lemma 17(i) and Lemma 20(i), while (14) follows from Lemma 18 and Lemma 20(i). ■

Proof of Theorem 5. Follows by Lemma 17 and Lemma 20(ii). ■

A.3. **Importance sampling schemes.** The following CLT, based on Proposition 1, and asymptotic variance formula, are [51, Theorem 7 and 15].

Proposition 21. *Under Assumption 5, the IS estimator* (21) *satisfies the CLT* (22)*, with limiting variance $\mathbb{V}_f^{IS} = \mu(\mathsf{a})\big[\mathrm{var}(K, \mathbf{m}_f) + \mu(\mathsf{av}_{\bar{f}})\big]/c_\xi^2$.*

Proof of Theorem 12. We first note that

$$\xi(f) := \frac{\zeta(f)}{\eta(1)} = \frac{c_\zeta}{c_\eta} \cdot \frac{c_\eta}{c_\zeta}\frac{\zeta(1)}{\eta(1)} \cdot \frac{\zeta(f)}{\zeta(1)} = c_\xi w \hat{\zeta}(f).$$

By Slutsky's lemma applied to (21) in the IS0 case,

$$\mathbb{V}_f^{IS0} = \mathrm{var}\big(\bar{K}, \xi(f)\big)/c_\xi^2 = \mathrm{var}\big(\bar{K}, w\hat{\zeta}(f)\big).$$

Then (i) follows by Theorem 3, and (ii) by Theorem 5, for the IS0 case. To prove the result for the ISJ case, we first note the relationship

$$\mathbb{V}_f^{ISJ} = \mu(\alpha)c_\xi^{-2}\Big[\mathrm{var}(K, \mathbf{m}_f) + \mu(v_{\bar{f}}) + \mu(\alpha\tilde{v}_{\bar{f}} - v_{\bar{f}})\Big] = \mu(\alpha)\mathbb{V}_f^{IS0} + \tilde{D}_{\bar{f}},$$

from Proposition 21. The result then follows from the IS0 case. ■

Appendix B. Proofs for main comparison application

Lemma 22. *Let (K, L) be the pair of kernels as in* (I), (II), *or* (III) *of Theorem 16, where we assume that $(\bar{\mu}, \nu, w)$ satisfies Assumption 1, with $(\bar{K}, \bar{\mu})$ the $Q^{(V)}$-augmentation of K* (23)*. Then, the following hold:*

(i) If $\|w\|_{L^\infty(\bar{\mu})} < \infty$, then $\mathcal{E}_L(g) \le \|w\|_{L^\infty(\bar{\mu})}\mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\bar{\mu})$.
If $\underline{c} := \bar{\mu}\text{-ess}\inf w$, then $\mathcal{E}_L(g) \ge \underline{c}\,\mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\bar{\mu})$.

(ii) If $\|w^*\|_{L^\infty(\mu)} < \infty$, then $\mathcal{E}_L(g) \le \|w^*\|_{L^\infty(\mu)}\mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\mu)$.

Proof. This is done separately below for the cases $L \in \{P, K^{\mathrm{DA0}}, K^{\mathrm{DA1}}\}$. Set $G := [g(x) - g(x')]^2$, $g \in L^2(\bar{\mu})$, with $x, x' \in \mathbf{X} := \mathbf{T} \times \mathbf{U} \times \mathbf{V}$. Then,

$$\begin{aligned}
\mathcal{E}_P(g) &= \frac{1}{2}\int \pi(\mathrm{d}x)q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\big\{1, r^{(V)}(x,x')\big\}G \\
&= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x)q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\big\{w(x), w(x)r^{(V)}(x,x')\big\}G \\
&= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x)q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\big\{w(x), w(x')r^{(U)}(x,x')\big\}G,
\end{aligned}$$

because $w(x)r^{(V)}(x,x') = w(x')r^{(U)}(x,x')$, well-defined on the set of interest. We then use the bounds $\underline{c} \le w \le \|w\|_{L^\infty(\bar{\mu})}$ $\bar{\mu}$-a.e. to conclude (i) for $L = P$.

Now assume $g \in L^2(\mu)$, so $G = [g(\theta, u) - g(\theta', u')]^2$. By Jensen's inequality and concavity of $(x, x') \mapsto \min\{x, x'\}$ when one of x, $x' \ge 0$ is held fixed,

$$\begin{aligned}
\mathcal{E}_P(g) &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x)q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')G\int Q_{\theta'u'}^{(V)}(\mathrm{d}v')\min\big\{w(x), w(x')r^{(U)}(x,x')\big\} \\
&\le \frac{1}{2}\int \bar{\mu}(\mathrm{d}x)q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')G\min\big\{w(x), w^*(\theta',u')r^{(U)}(x,x')\big\}.
\end{aligned}$$

Here, we have used that $r^{(U)}(x,x')$ does not depend on $v' \in \mathbf{V}$, and that

$$\int w(x)Q_{\theta u}^{(V)}(\mathrm{d}v) = \frac{c_\eta}{c_\zeta}\frac{1}{\eta(1)}\int \zeta(1)Q_{\theta u}^{(V)}(\mathrm{d}v) = \frac{\pi^*(\mathrm{d}\theta, \mathrm{d}u)}{\mu(\mathrm{d}\theta, \mathrm{d}u)} = w^*(\theta, u).$$

We then apply Jensen again, this time integrating out $v \in \mathbf{V}$, to get,

$$\begin{aligned}
&\mathcal{E}_P(g) \\
&\le \frac{1}{2}\int \mathrm{d}\theta Q_\theta^{(U)}(\mathrm{d}u)\frac{\eta(1)}{c_\eta}q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')G\int Q_{\theta u}^{(V)}(\mathrm{d}v)\min\big\{w(x), w^*(x')r^{(U)}(x,x')\big\} \\
&\le \frac{1}{2}\int \mathrm{d}\theta Q_\theta^{(U)}(\mathrm{d}u)\frac{\eta(1)}{c_\eta}q_\theta(\mathrm{d}\theta')Q_{\theta'}^{(U)}(\mathrm{d}u')\min\big\{w^*(\theta,u), w^*(\theta',u')r^{(U)}(x,x')\big\}G.
\end{aligned}$$

We then apply the bound $w^* \leq \|w^*\|_{L^\infty(\mu)}$ μ-a.e. and use the fact that $\mathcal{E}_K(g) = \mathcal{E}_{\bar{K}}(g)$ for all $g \in L^2(\mu)$ to conclude (ii) for $L = P$.

Now consider the case $L = K^{\mathrm{DA0}}$. With $G := [g(x) - g(x')]^2$ on $\mathbf{X}^2$,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA0}}}(g) &= \frac{1}{2}\int \pi(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u')\alpha(\theta,u;\theta',u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\Big\{1, \frac{w(x')}{w(x)}\Big\} G \\ &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u')\alpha(\theta,u;\theta',u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\big\{w(x), w(x')\big\} G,\end{aligned}$$

for all $g \in L^2(\bar{\mu})$. As before, this allows us to conclude (i) for $L = K^{\mathrm{DA0}}$.

Now assume $g \in L^2(\mu)$, with $G := [g(\theta,u) - g(\theta',u')]^2$. By Jensen,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA0}}}(g) &\leq \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u')\alpha(\theta,u;\theta',u') G \min\big\{w(x), w^*(\theta',u')\big\} \\ &\leq \frac{1}{2}\int \mu(\mathrm{d}\theta,\mathrm{d}u) q_\theta(\mathrm{d}\theta') Q^{(U)}_{\theta'}(\mathrm{d}u')\alpha(\theta,u;\theta',u') G \min\big\{w^*(\theta,u), w^*(\theta',u')\big\},\end{aligned}$$

which allows us to conclude (ii) as before.

Now consider the case $L = K^{\mathrm{DA1}}$. With $G := [g(x) - g(x')]^2$ on $\mathbf{X}^2$,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA1}}}(g) &= \frac{1}{2}\int \pi(\mathrm{d}x) K_{\theta u}(\mathrm{d}\theta',\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\Big\{1, \frac{w(x')}{w(x)}\Big\} G \\ &= \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) K_{\theta u}(\mathrm{d}\theta',\mathrm{d}u') Q^{(V)}_{\theta' u'}(\mathrm{d}v') \min\big\{w(x), w(x')\big\} G,\end{aligned}$$

for all $g \in L^2(\bar{\mu})$. As before, this allows us to conclude (i) for $L = K^{\mathrm{DA1}}$.

Now assume $g \in L^2(\mu)$, with $G := [g(\theta,u) - g(\theta',u')]^2$. By Jensen,

$$\begin{aligned}\mathcal{E}_{K^{\mathrm{DA1}}}(g) &\leq \frac{1}{2}\int \bar{\mu}(\mathrm{d}x) K_{\theta u}(\mathrm{d}\theta',\mathrm{d}u') G \min\big\{w(x), w^*(\theta',u')\big\} \\ &\leq \frac{1}{2}\int \mu(\mathrm{d}\theta,\mathrm{d}u) K_{\theta u}(\mathrm{d}\theta',\mathrm{d}u') G \min\big\{w^*(\theta,u), w^*(\theta',u')\big\},\end{aligned}$$

which allows us to conclude (ii) as before. ■

Lemma 23. *With assumptions as in Lemma 22, and additionally assuming that K and L determine Harris ergodic chains, the following hold:*

(i) If $\|w\|_{L^\infty(\bar{\mu})} < \infty$, then $(\bar{\mu}, \pi, w, \bar{K}, L, \underline{c}, \|w\|_{L^\infty(\bar{\mu})})$ satisfies Assumption 2.

(ii) If $\|w^*\|_{L^\infty(\mu)} < \infty$, then $(\bar{\mu}, \pi, w, \bar{K}, L, 0, \|w^*\|_{L^\infty(\mu)})$ satisfies Assumption 3.

Proof. Lemma 22(i) and (ii) imply respectively (i) and (ii). ■

Proof of Theorem 16. The support condition Assumption 4(ii) implies that $(\bar{\mu}, \pi, w)$ satisfies Assumption 1. Under conditions (I), (II), or (III), the result follows by Lemma 23 and Theorem 12.

Assume condition (IV). Because $g := \hat{\zeta}(f)$ is a function on $\mathbf{X} = \mathbf{T} \times \mathbf{U} \times \mathbf{V}$ which does not depend on the second coordinate, $P^k g(\theta,u,v) = M^k g(\theta,v)$ for all $(\theta,u,v) \in \mathbf{X}$ and $k \geq 1$. Therefore, $\mathrm{var}(M,g) = \mathrm{var}(P,g)$. ■

Proof of Proposition 14. For any $g \in L^2(\pi)$, set $G := [g(\theta,u,v) - g(\theta',u',v')]^2$. We have

$$\mathcal{E}_{K^{\mathrm{DA1}}}(g) = \mathcal{E}_{K^{\mathrm{DA0}}}(g) + \frac{1}{2}\int \pi(\mathrm{d}x) r_K(\theta,u) Q^{(V)}_{\theta u}(\mathrm{d}v') \min\Big\{1, \frac{\xi'(1)}{\xi(1)}\Big\} G,$$

so (i) follows from the covariance ordering. For (ii), we have

$$K_x^{\mathrm{DA1}}(\mathrm{d}x') = K_x^{\mathrm{DA0}}(\mathrm{d}x') - (1-\alpha_{\mathrm{DA0}}(x))\delta_x(\mathrm{d}x') + r_K(\theta,u)\delta_x(\mathrm{d}x') + (1-\alpha_{\mathrm{DA1}}(x))\delta_x(\mathrm{d}x'),$$

from which we conclude, since $\alpha_{\mathrm{DA1}}(x) = \alpha_{\mathrm{DA0}}(x) + r_K(\theta,u)$. ■

Proof of Proposition 15. (i) is essentially well-known [see 10], and (ii) is straightforward to prove. ■

Appendix C. Properties of augmented kernels and ergodicity

For measurable functions $V : \mathbf{X} \to [1,\infty)$ and $f : \mathbf{X} \to \mathbb{R}$, we set

$$\|\nu\|_V := \sup_{f:|f|\le V} \nu(f), \qquad \text{and} \qquad \|f\|_V := \sup_{x\in\mathbf{X}} \frac{|f(x)|}{V(x)}$$

for any finite signed measure ν on $\mathbf{X}$.

Definition 4. A μ-invariant Markov chain K on $\mathbf{X}$ is said to be

(i) *V-geometrically ergodic* if there is a function $V : \mathbf{X} \to [1,\infty)$ such that

$$\|K^n(x,\cdot) - \mu(\cdot)\|_V \le RV(x)\rho^n$$

for all $n \ge 1$, where $R < \infty$ and $\rho \in (0,1)$ are constants.

(ii) *uniformly ergodic* if K is 1-geometrically ergodic.

Lemma 24. *Let $K_{\theta y}(\mathrm{d}\theta',\mathrm{d}y') = \dot K_\theta(\mathrm{d}\theta')Q_{\theta'}(\mathrm{d}y')$ be an augmented kernel on $\mathbf{T}\times\mathbf{Y}$.*

(i) The invariant measures of K and $\dot K$ satisfy ($\mu K = \mu \implies \mu^*\dot K = \mu^*$), and ($\dot\mu\dot K = \dot\mu \implies \mu K = \mu$), where $\mu(\mathrm{d}\theta,\mathrm{d}y) := \dot\mu(\mathrm{d}\theta)Q_\theta(\mathrm{d}y)$. These implications hold with invariance replaced with reversibility.

(ii) K is μ-Harris ergodic $\iff$ $\dot K$ is $\dot\mu$-Harris ergodic.

(iii) For all $f \in L^1(\mu)$ and $n \ge 1$, $K^n f(\theta,y) = \dot K^n(Qf)(\theta)$.

(iv) K is aperiodic $\iff$ $\dot K$ is aperiodic. K is positive $\iff$ $\dot K$ is positive.

(v) K is geometrically ergodic $\iff$ $\dot K$ is geometrically ergodic.

(vi) K is uniformly ergodic $\iff$ $\dot K$ is uniformly ergodic.

Proof. (i–iii) are [51, Lem. 21]. Proof of (iv) is straightforward.

For (v), consider first the case that $\dot K$ is $\dot V$-geometrically ergodic:

$$\sup_{|f|\le\dot V} |\dot K^n(f)(\theta) - \dot\mu(f)| \le R\dot V(\theta)\rho^n, \qquad n \ge 1,$$

with $\dot V : \mathbf{T} \to [1,\infty)$ and constants R and ρ. Define $V(\theta,y) := \dot V(\theta)$. By (iii),

$$\sup_{|f|\le V} |K^n f(\theta,y) - \mu(f)| = \sup_{|f|\le V} |\dot K^n(Qf)(\theta) - \dot\mu(Qf)|. \tag{40}$$

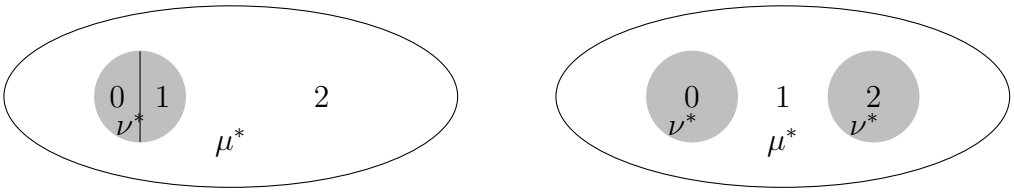
Since $Qf(\theta,y) \le QV(\theta,y) = \dot V(\theta)$, we get that K is V-geometrically ergodic.

Assume now that K is V-geometrically ergodic. Using (40), we have,

$$\sup_{|f|\le V} |K^n f(\theta,y) - \mu(f)| = \sup_{g=Qf:|f|\le V} |\dot K^n g(\theta) - \dot\mu(g)|, \tag{41}$$

for $n \ge 1$. Define $\dot V(\theta) := \inf_y V(\theta,y)$. For all g such that $|g(\theta)| \le \dot V(\theta)$, set $f(\theta,y) := g(\theta)$. Then $|f| \le V$ and $Qf = g$. By (41), $\dot K$ is $\dot V$-geometrically ergodic. This proves (v), and (vi) follows from the form of $\dot V$ and V. ■

FIGURE 1. Two versions and two behaviours



(A) 'PM/DA better' case (B) 'IS better' case

Proposition 25. *Consider the PM parent kernel* (30) *and the DA0 kernel* K^{DA0} (27)*, with* K *as in* (28)*. If* $\zeta(1)$ *is not bounded, then PM parent and* K^{DA0} *are not* V*-geometrically ergodic.*

Proof. This is [6, Thm. 8] for PM chains. To prove that result for PM chains, or in particular for the PM parent chain (30), [6] show that for all $\epsilon > 0$,

$$\nu\big(\mathbf{1}\{\alpha_{\mathrm{PMP}} \leq \epsilon\}\big) > 0. \tag{42}$$

By [45, Thm. 5.1], one concludes that the PM parent is not V-geometrically ergodic [6]. Moreover, from

$$\min\{1, r^{(U)}(x,x')\}\min\{1, w(x')/w(x)\} \leq \min\{1, r^{(V)}(x,x')\}, \tag{43}$$

it follows that $\alpha_{\mathrm{DA0}}(x) \leq \alpha_{\mathrm{PMP}}(x)$. By (42), one concludes that K^{DA0} also is not V-geometrically ergodic. ■

APPENDIX D. TOY EXAMPLES OF TWO EXTREMES

Let $\mathbf{X} := \{0,1,2\}$ and consider the two mass allocations for probabilities μ and ν on $\mathbf{X}$ and function $f \in L^2_0(\nu)$ given pictorially in Figure 1 and precisely in Figure 3. Denote by $q^{(r)}$ the (reflected) random walk proposal on $\mathbf{X}$, given by

FIGURE 3. Mass allocations for μ, ν, and f on $\mathbf{X} = \{0,1,2\}$, $a \in [\frac{1}{2}, 1)$.

$$\begin{array}{rcl} \mu & = & (\ \frac{1-a}{2} \quad \frac{1-a}{2} \quad a\ ) \\ \nu & = & (\ 1/2 \quad 1/2 \quad 0\ ) \\ f & = & (\ 1 \quad -1 \quad 0\ ) \end{array} \qquad \begin{array}{rcl} \mu & = & (\ 1/3 \quad 1/3 \quad 1/3\ ) \\ \nu & = & (\ \frac{a}{2} \quad \frac{1-a}{2} \quad 1/2\ ) \\ f & = & \frac{\sqrt{2}}{\sqrt{a+a^2}}\ (\ 1 \quad 0 \quad -a\ ) \end{array}$$

(A) 'MH/DA better' case (B) 'IS better' case

$q_0^{(r)}(x) = \delta_1(x)$, $q_1^{(r)}(x) = \frac{1}{2}[\delta_0(x) + \delta_2(x)]$, and $q_2^{(r)}(x) = \delta_1(x)$, and by $q_x^{(u)}(x')$ the uniform proposal on $\mathbf{X}$. We set $K :=\mathrm{MH}(q \to \mu)$ and let L be the MH or DA0 kernels, using proposals $q^{(r)}$ or $q^{(u)}$, and targeting ν. We use a parameter $a \in [\frac{1}{2}, 1)$ to allow for continuous intensity shifts in the mass allocations in our examples. Because μ is constant on the support of ν, one can check that the MH and DA0 kernels coincide for $a \in [\frac{1}{2}, 1)$.

The resulting IS and MH/DA asymptotic variances, $\mathrm{var}(K, wf)$ and $\mathrm{var}(L, f)$, can be computed by linear algebra using [see 29, Cor. 1.5]. They are listed in Table 1, and plotted in Figure 5. Here,

$$\mathrm{UB}_a(f) := \max(w)\mathrm{var}(L, f) + \nu(f^2[\max(w) - w]). \tag{44}$$

TABLE 1. Asymptotic variance as a function of $a \in [1/2, 1)$

Proposal	$\mathrm{var}(L,f) \leq$	$\mathrm{var}(K,wf)$	$\mathrm{var}(L,f) \geq$	$\mathrm{var}(K,wf)$
RW $q^{(r)}$	1	$\frac{1}{1-a}$	$\frac{-1+8a+a^2}{a^2-1}$	$\frac{9a}{1+a}$
uniform $q^{(u)}$	2	$\frac{1}{1-a}$	$\frac{-1+10a-a^2}{(1+a)^2}$	$\frac{15a}{4(1+a)}$

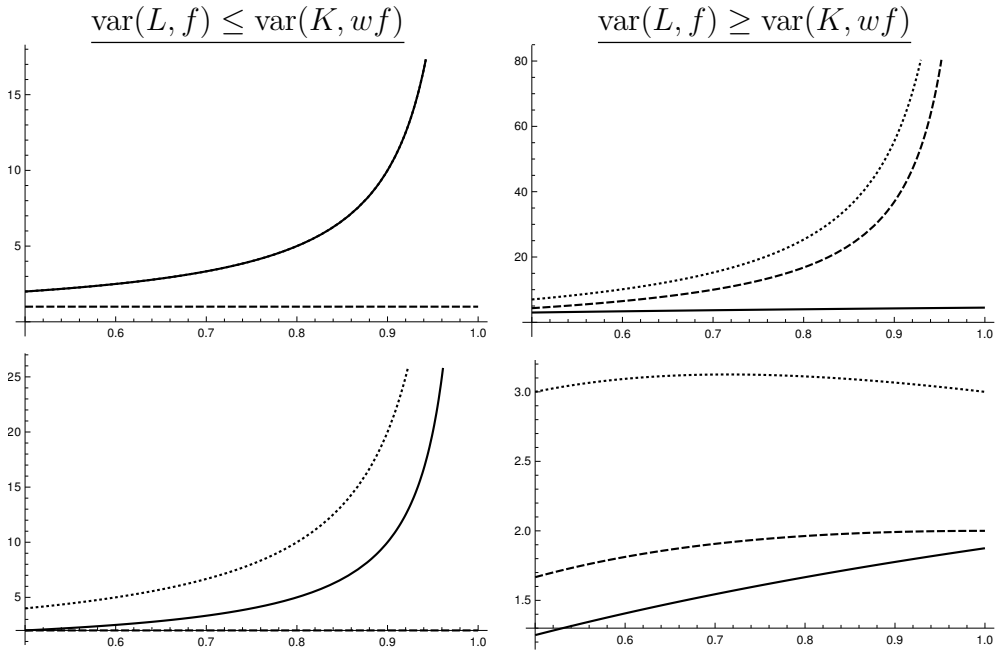


FIGURE 5. Plots from Table 1: $\mathrm{var}(K, wf)$ '—', $\mathrm{var}(L, f)$ '−−', and $\mathrm{UB}_a(f)$ '$\cdots$', vs. $a \in [\frac{1}{2}, 1)$. Here, in the top left, $\mathrm{UB}_a(f)$ exactly coincides with $\mathrm{var}(K, wf)$.

is the upper bound on $\mathrm{var}(K, wf)$ from Corollary 2.

REFERENCES

[1] C. Andrieu. On random- and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016.

[2] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010. (with discussion).

[3] C. Andrieu, A. Doucet, S. Yıldırım, and N. Chopin. On the utility of Metropolis-Hastings with asymmetric acceptance ratio. Preprint arXiv:1803.09527, 2018.

[4] C. Andrieu, A. Lee, and M. Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2), 2018.

[5] C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *J. Appl. Probab.*, 16(3):1462–1505, 2006.

[6] C. Andrieu and G. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.
[7] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373, 2008.
[8] C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 04 2015.
[9] C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, 2016. arXiv:1404.6909.
[10] M. Banterle, C. Grazian, A. Lee, and C. Robert. Accelerating Metropolis-Hastings algorithms by delayed acceptance. Preprint arXiv:1503.00996, 2015.
[11] J. Bardsley, A. Solonen, H. Haario, and M. Laine. Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. *SIAM J. Sci. Comput.*, 36(4):A1895–A1910, 2014.
[12] F. Bassetti and P. Diaconis. Examples comparing importance sampling and the Metropolis algorithm. *Illinois J. Math.*, 50(1-4):67–91, 2006.
[13] P. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.*, 15(1B):700–738, 2005.
[14] R. Bellman. Some inequalities for the square root of a positive definite matrix. *Linear Algebra Appl.*, 1(3):321–324, 1968.
[15] S. Caracciolo, A. Pelissetto, and A. Sokal. Nonlocal Monte Carlo algorithm for self-avoiding random walks with fixed endpoints. *J. Stat. Phys.*, 60:1–53, 1990.
[16] N. Chopin, P. Jacob, and O. Papaspiliopoulos. SMC^2: A sequential Monte Carlo algorithm with particle Markov chain Monte Carlo updates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 75(3):397–426, 2013.
[17] J. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4), 2005.
[18] T. Cui, Y. Marzouk, and K. Willcox. Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction. *J. Comput. Phys.*, 315:363–387, 2016.
[19] G. Deligiannidis, A. Doucet, M. K. Pitt, and R. Kohn. The correlated pseudo-marginal method. Preprint arXiv:1511.04992, 2015.
[20] H. Doss. Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1728–1734, 1994.
[21] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer, 2018.
[22] R. Douc and C. Robert. A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *Ann. Statist.*, 39(1):261–277, 2011.
[23] A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.
[24] W. Gilks and G. Roberts. Strategies for improving MCMC. In *Markov chain Monte Carlo in practice*, volume 6, pages 89–114. 1996.
[25] P. Glynn and D. Iglehart. Importance sampling for stochastic simulations. *Management Sci.*, 35(11):1367–1392, 1989.

[26] A. Golightly, D. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statist. Comput.*, 25, 2015.

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

[28] S. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.

[29] C. Kipnis and S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104(1):1–19, 1986.

[30] A. Lee, C. Yau, M. Giles, A. Doucet, and C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comput. Graph. Statist.*, 19(4):769–789, 2010.

[31] D. Levin, Y. Peres, and E. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2009.

[32] L. Lin, K. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Phys. Rev. D*, 61, 2000.

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

[34] J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2003.

[35] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 03 1994.

[36] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, second edition, 2009.

[37] A. Mira and C. Geyer. Ordering Monte Carlo Markov Chains. Technical report, School of Statistics, University of Minnesota, 1999.

[38] A. Mira and F. Leisen. Covariance ordering for discrete and continuous time Markov chains. *Statist. Sinica*, pages 651–666, 2009.

[39] A. Owen and Y. Zhou. Safe and effective importance sampling. *J. Amer. Statist. Assoc.*, 95(449):135–143, 2000.

[40] P. Parpas, B. Ustun, M. Webster, and Q. K. Tran. Importance sampling in stochastic programming: A Markov chain Monte Carlo approach. *INFORMS J. Comput.*, 27(2):358–377, 2015.

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

[42] M. Quiroz, M.-N. Tran, M. Villani, and R. Kohn. Speeding up MCMC by delayed acceptance and data subsampling. *J. Comput. Graph. Statist.*, 2017. To appear.

[43] G. Roberts and J. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.*, 16(4):2123–2139, 2006.

[44] G. Roberts and J. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007.

[45] G. Roberts and R. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*,

83(1):95–110, 1996.

[46] D. Rudolf. Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Math.*, 485:93 pages, 08 2012.

[47] C. Sherlock and A. Lee. Variance bounding of delayed-acceptance kernels. Preprint arXiv:1706.02142, 2017.

[48] C. Sherlock, A. Thiery, and A. Lee. Pseudo-marginal Metropolis-Hastings using averages of unbiased estimators. Preprint arXiv:1610.09788, 2016.

[49] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.

[50] M.-N. Tran, M. Scharth, M. Pitt, and R. Kohn. Importance sampling squared for Bayesian inference in latent variable models. *arXiv:1309.3339v3*, 2014.

[51] M. Vihola, J. Helske, and J. Franks. Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1609.02541v6, 2016.

School of Mathematics, Statistics and Physics, Newcastle University, NE1 7RU Newcastle, United Kingdom

E-mail address: franks@iki.fi

Department of Mathematics and Statistics, P.O.Box 35, FI-40014 University of Jyväskylä, Finland

E-mail address: matti.vihola@iki.fi