The Annals of Applied Probability
2016, Vol. 26, No. 5, 2661–2696
DOI: 10.1214/15-AAP1158

ESTABLISHING SOME ORDER AMONGST EXACT APPROXIMATIONS OF MCMCS

BY CHRISTOPHE ANDRIEU[1] AND MATTI VIHOLA[2]

University of Bristol and University of Jyväskylä

Exact approximations of Markov chain Monte Carlo (MCMC) algorithms are a general emerging class of sampling algorithms. One of the main ideas behind exact approximations consists of replacing intractable quantities required to run standard MCMC algorithms, such as the target probability density in a Metropolis–Hastings algorithm, with estimators. Perhaps surprisingly, such approximations lead to powerful algorithms which are exact in the sense that they are guaranteed to have correct limiting distributions. In this paper, we discover a general framework which allows one to compare, or order, performance measures of two implementations of such algorithms. In particular, we establish an order with respect to the mean acceptance probability, the first autocorrelation coefficient, the asymptotic variance and the right spectral gap. The key notion to guarantee the ordering is that of the convex order between estimators used to implement the algorithms. We believe that our convex order condition is close to optimal, and this is supported by a counterexample which shows that a weaker variance order is not sufficient. The convex order plays a central role by allowing us to construct a martingale coupling which enables the comparison of performance measures of Markov chain with differing invariant distributions, contrary to existing results. We detail applications of our result by identifying extremal distributions within given classes of approximations, by showing that averaging replicas improves performance in a monotonic fashion and that stratification is guaranteed to improve performance for the standard implementation of the Approximate Bayesian Computation (ABC) MCMC method.

1. Introduction. Consider a probability distribution π defined on some measurable space $(\mathsf{X}, \mathcal{X})$ and assume that sampling realisations from this probability distribution is of interest. A generic and popular way of achieving this consists of using Markov chain Monte Carlo algorithms (MCMC), of which the Metropolis–Hastings (MH) algorithm is the main workhorse.

Let $\{q(x, \cdot)\}_{x\in\mathsf{X}}$ be a family of probability distributions on $(\mathsf{X}, \mathcal{X})$ such that $x \mapsto q(x, A)$ is measurable for all $A \in \mathcal{X}$; in other words $(x, A) \mapsto q(x, A)$ defines

Received August 2014; revised October 2015.

[1]Supported in part by the EPSRC programme grant "Intractable Likelihood: New Challenges from Modern Applications (ILike)" and the EPSRC grant "Bayesian Inference for Big Data with Stochastic Gradient Markov Chain Monte Carlo".

[2]Supported by the Academy of Finland fellowships 250575 and 274740.

MSC2010 subject classifications. Primary 60J22; secondary 60J05, 60E15, 65C05.

Key words and phrases. Asymptotic variance, convex order, Markov chain Monte Carlo, martingale coupling, pseudo-marginal algorithm.

a kernel. (We assume hereafter implicitly that such a measurability condition is satisfied by given families of probability measures, and that given functions are measurable.) For any $x, y \in \mathsf{X}$ we define the "acceptance ratio" $r(x, y)$ as follows: for $(x, y) \in \mathsf{R} \subset \mathsf{X}^2$ (with R the symmetric set as defined in [46], Proposition 1) we let $r(x, y)$ be the Radon–Nikodym derivative

$$0 < r(x, y) := \frac{\pi(\mathrm{d}y)q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)q(x, \mathrm{d}y)} < \infty,$$

and $r(x, y) = 0$ otherwise. The MH algorithm defines a Markov chain with the following transition kernel:

$$P(x, \mathrm{d}y) = q(x, \mathrm{d}y) \min\{1, r(x, y)\} + \delta_x(\mathrm{d}y)\rho(x), \tag{1}$$

where

$$\rho(x) := 1 - \int \min\{1, r(x, y)\} q(x, \mathrm{d}y),$$

and δ_x is the Dirac measure at $x \in \mathsf{X}$. It is standard to show that P is reversible with respect to π and hence leaves π invariant.

In some situations, evaluation of the ratio $r(x, y)$ is either impossible or overly expensive, therefore, rendering the algorithm nonviable. A canonical example is when π is the marginal of a probability density, say $\pi(x) = \int \bar{\pi}(x, z)\,\mathrm{d}z$ for some latent variables z, where the integral is intractable and the "marginal MCMC" targeting π is therefore not implementable. A classical way of addressing this problem consists of running an MCMC targeting the joint distribution $\bar{\pi}$, which may however become prohibitively inefficient in situations where the size of the latent variable is high—this is for example the case in general state-space models [1]. A powerful alternative which has recently attracted substantial interest consists, in simple terms, of replacing the value of $\pi(x)$ with a noisy, but computationally cheap, measurement whenever it is required in the implementation of the MH algorithm above. Although this idea may at first appear naive, it turns out to lead to exact algorithms, that is sampling from π is guaranteed at equilibrium under mild assumptions on the nature of the noise involved.

We now describe in more detail a modification of the "exact", or marginal, MH algorithm above, where one replaces the density values $\pi(x)$ with estimators $\hat{\pi}(x)$ for all $x \in \mathsf{X}$. The estimators $\hat{\pi}(x)$ are assumed to be nonnegative and unbiased (up to a constant), in a sense that there exists $C > 0$ such that for all $x \in \mathsf{X}$, $\mathbb{E}[\hat{\pi}(x)] = C \times \pi(x)$. To fix ideas we first present the method in an algorithmic form, and later define precisely the related mathematical construction. We call this method a pseudo-marginal approximation of P or simply the pseudo-marginal algorithm, following [3, 4].

ALGORITHM 1 (Pseudo-marginal algorithm). Assume $(X_0, \hat{\pi}(X_0)) \in \mathsf{X} \times (0, \infty)$ are some initial values. Then, for $n = 1, 2, \ldots$:

(a) Propose a transition $Y_n \sim q(X_n, \cdot)$.

(b) Given Y_n, generate a random variable $\hat{\pi}(Y_n) \geq 0$ (such that $\mathbb{E}[\hat{\pi}(Y_n)] = C\pi(Y_n)$).

(c) With probability

$$\min\left\{1, \frac{\hat{\pi}(Y_n)}{\hat{\pi}(X_{n-1})} \frac{q(Y_n, X_{n-1})}{q(X_{n-1}, Y_n)}\right\},$$

accept and set $(X_n, \hat{\pi}(X_n)) = (Y_n, \hat{\pi}(Y_n))$; otherwise reject and set $(X_n, \hat{\pi}(X_n)) = (X_{n-1}, \hat{\pi}(X_{n-1}))$.

For example, in the situation where $\pi(x) = \int \bar{\pi}(x, z)\, \mathrm{d}z$ one could use an importance sampling estimator of the integral and define

$$\hat{\pi}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{\bar{\pi}(x, Z_i)}{g_x(Z_i)},$$

where $Z_1, \ldots, Z_N$ are samples from some instrumental probability density g_x. The algorithm resulting from this choice is known as the grouped independence Metropolis–Hastings [5] and has proved useful in the context of inference in phylogenetic trees. Other important instances of pseudo-marginal algorithms include approximate Bayesian computation (ABC) MCMC [33] (see also Section 6.2), the particle marginal MH [1] and algorithms for inference in diffusions [8]. Note that Algorithm 1 is intact if we multiplied all our estimators by a constant $C' > 0$. Therefore, without loss of generality, we will set $C = 1$ in the remainder of the paper for notational simplicity.

We now turn to an abstract representation of the pseudo-marginal algorithm, used in the remainder of this paper, which has the advantage that it simplifies notation and highlights the key structure and quantities underpinning their behaviour [3, 4, 42]. The main idea is to introduce the normalised estimator $\varpi(x) := \hat{\pi}(x)/\pi(x)$ and to view this as a multiplicative perturbation, or noise, of the true density $\pi(x)$, since then $\hat{\pi}(x) = \pi(x)\varpi(x)$. An immediate consequence is that with $W_{n-1} := \varpi(X_{n-1})$ and $U_n := \varpi(Y_n)$ the acceptance ratio in Algorithm 1 takes the form $r(X_{n-1}, Y_n)U_n/W_{n-1}$ which highlights the structure of this approximation, a multiplicative perturbation of the acceptance ratio of the exact algorithm, and the fact that $\{X_n, W_n\}$ is a Markov chain. This leads us to the following abstract model. Let $\{Q_x\}_{x \in \mathsf{X}}$ be a family of probability measures on the nonnegative reals $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ indexed by $x \in \mathsf{X}$ and such that $\int w Q_x(\mathrm{d}w) = 1$ for any $x \in \mathsf{X}$. Consider the following probability distribution on $(\mathsf{X} \times \mathbb{R}_+, \mathcal{X} \times \mathcal{B}(\mathbb{R}_+))$ defined through

$$\tilde{\pi}(\mathrm{d}x \times \mathrm{d}w) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) w.$$

This formalises the multiplicative perturbation idea and includes the distribution of the estimators in the target distribution. Now one can check that the Markov

transition probability $\tilde{P}$ of the pseudo-marginal approximation of the marginal kernel P is a MH algorithm targeting $\tilde{\pi}$ and can be written as follows for any $(x, w) \in \mathsf{X} \times (0, \infty)$ [3, 4]

$$\begin{aligned}\tilde{P}(x, w; \mathrm{d}y \times \mathrm{d}u) := q(x, \mathrm{d}y) Q_y(\mathrm{d}u) \min\left\{1, r(x, y)\frac{u}{w}\right\} \\ + \delta_{x,w}(\mathrm{d}y \times \mathrm{d}u)\tilde{\rho}(x, w),\end{aligned} \tag{2}$$

with rejection probability

$$\tilde{\rho}(x, w) := 1 - \int \min\left\{1, r(x, y)\frac{u}{w}\right\} q(x, \mathrm{d}y) Q_y(\mathrm{d}u).$$

Note that while the distribution of the normalised estimators $Q_y(\mathrm{d}u)$ appears in the kernel and the rejection probability, it does not appear explicitly in the algorithm. The fundamental marginal identity $\tilde{\pi}(\mathrm{d}x \times \mathbb{R}_+) = \pi(\mathrm{d}x)$ is important since it tells us that despite being approximations of P, pseudo-marginal algorithms are exact in the sense that they sample marginally from the desired distribution π at equilibrium.

Often in practice there are various possible choices for the estimators in Algorithm 1, or equivalently the family $\{Q_x\}_{x \in \mathsf{X}}$, depending on the specific application. For example, in case of the simple importance sampling estimator presented above one may choose between various families of importance sampling distributions $\{g_x\}_{x \in \mathsf{X}}$, or vary the number of samples N. The ultimate aim of the present work is to develop simple and general tools for the characterisation of the performance of pseudo-marginal algorithms as a function of the family $\{Q_x\}_{x \in \mathsf{X}}$, therefore, allowing comparisons. We refer the reader to [4] for other quantitative and qualitative properties of such algorithms.

We now recall standard performance measures relevant to the MCMC context. For a generic Markov transition kernel Π with invariant distribution μ, both defined on some measurable space $(\mathsf{E}, \mathcal{E})$, and any function $f : \mathsf{E} \to \mathbb{R}$, we define the asymptotic variance of f for Π as follows. Denote $\mathbb{E}_\mu(\cdot)$ and $\mathrm{var}_\mu(\cdot)$ the expectation and variance operators corresponding to the Markov chain $\{\Phi_i\}_{i \geq 0}$ with transition kernel Π and such that $\Phi_0 \sim \mu$, and denote $\mathrm{var}_\mu(f) := \mathrm{var}_\mu(f(\Phi_0))$. Then the asymptotic variance is defined as

$$\mathrm{var}(f, \Pi) := \lim_{M \to \infty} \mathrm{var}_\mu\left(M^{-1/2} \sum_{i=1}^{M} f(\Phi_i)\right),$$

where the limit is guaranteed to exist for reversible Markov chains, but may be infinite (cf. [46]). When finite, the asymptotic variance naturally appears in the central limit theorem, for example, [26], but it also characterises the finite sample efficiency of MCMC algorithms; see for instance the finite sample bounds given in [29, 40]. When Π is reversible with respect to μ, the so-called right spectral gap $\mathrm{Gap}_R(\Pi)$, defined precisely in Section 3, turns out in some scenarios to be an

indicative criterion of performance. For instance, when Π is a positive operator, the right spectral gap is also the absolute spectral gap, which characterises the geometric convergence to equilibrium [39].

Before focusing on the comparison of pseudo-marginal algorithms, we briefly review here what is known about the standard MH algorithm, since the result will be referred to in several places in this paper and helps to motivate our work. The result is in its first form due to Peskun [37] and was later extended to more general setups in [10, 30, 46].

THEOREM 2 (Peskun). *Let μ be a probability distribution on some measurable space $(\mathsf{E}, \mathcal{E})$ and let Π_1, Π_2 be two μ-reversible Markov transition kernels. Assume that for any $x \in \mathsf{E}$ and any $A \in \mathcal{E}$ such that $x \notin A$, the transitions satisfy $\Pi_1(x, A) \geq \Pi_2(x, A)$.*

(a) *If $f : \mathsf{X} \to \mathbb{R}$ satisfies $\mathrm{var}_\mu(f) < \infty$, then $\mathrm{var}(f, \Pi_1) \leq \mathrm{var}(f, \Pi_2)$.*
(b) *The right spectral gaps satisfy $\mathrm{Gap}_R(\Pi_1) \geq \mathrm{Gap}_R(\Pi_2)$.*

In fact, as pointed out by [10], the off-diagonal order $\Pi_1(x, A) \geq \Pi_2(x, A)$ is stronger than needed for Theorem 2 to hold, and a weaker condition is that

$$\int f(x) f(y) \mu(\mathrm{d}x) \Pi_1(x, \mathrm{d}y) \leq \int f(x) f(y) \mu(\mathrm{d}x) \Pi_2(x, \mathrm{d}y),$$

for any f such that $\mathrm{var}_\mu(f) < \infty$. This condition turns out to be a necessary and sufficient condition. However, the popularity of Theorem 2 stems from the simplicity of its statement providing a simple and intuitive criterion for the comparison of performance of algorithms, which can be checked in practice. As we shall see, Peskun's result is however not directly relevant to pseudo-marginal algorithms when the aim is to compare different approximation strategies. This stems from the fact that changing $\{Q_x\}_{x \in \mathsf{X}}$ changes the invariant distribution of the Markov transition kernel involved.

More specifically, consider two families of distributions, $\{Q_x^{(1)}\}_{x \in \mathsf{X}}$ and $\{Q_x^{(2)}\}_{x \in \mathsf{X}}$ leading to two competing pseudo-marginal approximations of P, say $\tilde{P}_1$ and $\tilde{P}_2$, with distinct invariant distributions $\tilde{\pi}_1$ and $\tilde{\pi}_2$, respectively. Note that both algorithms target $\pi(\cdot)$ marginally and share the same family of proposal distributions $\{q(x, \cdot)\}_{x \in \mathsf{X}}$. In light of Peskun's result, a natural question is to find useful conditions on the families $\{Q_x^{(1)}\}_{x \in \mathsf{X}}$ and $\{Q_x^{(2)}\}_{x \in \mathsf{X}}$ which would ensure that $\mathrm{var}(f, \tilde{P}_1) \leq \mathrm{var}(f, \tilde{P}_2)$ for certain classes of functions $f : \mathsf{X} \times \mathbb{R}_+ \to \mathbb{R}$, or that $\mathrm{Gap}_R(\tilde{P}_1) \leq \mathrm{Gap}_R(\tilde{P}_2)$.

As we shall see, despite the difficulty pointed out earlier, it is possible to prove such results, assuming a classical convex stochastic order between the unit mean random variables $W_x^{(1)} \sim Q_x^{(1)}$ and $W_x^{(2)} \sim Q_x^{(2)}$ for $x \in \mathsf{X}$ (see Theorem 10). Because the convex order is stronger than the simpler variance order between $W_x^{(1)}$ and $W_x^{(2)}$, one may wonder whether the variance order could be sufficient to imply

a similar result. We show, by means of a counterexample, that the variance order is not sufficient to guarantee such results (see Example 13).

Our framework benefits from a considerable body of work on the convex order (see [35, 41] for recent reviews) which allows one to import known results and establish ordering results for pseudo-marginal algorithms with minimal effort. Some applications are discussed in some detail in Sections 6.1, 6.2 and 6.3. We also point to a very recent application of our result, behind an interesting reasoning aimed at deriving a type of quantitative bounds for pseudo-marginal algorithms in some contexts [9], Theorem 3.

We do not know of earlier works with similar Peskun-type orders for Markov chains with different invariant distributions. The recent work [32] establishes Peskun-type orders for inhomogeneous Markov chains, which is applicable to some algorithms which are discussed in the present work as well. However, we emphasise that our results are much more general when applied in the context of pseudo-marginal algorithms.

The remainder of the paper is organised as follows. In Section 2, before embarking on the analysis of the pseudo-marginal algorithms described above, we first focus on related and practically relevant algorithms for which convex ordering is pertinent but the comparison is mathematically much simpler. This allows a gentle and progressive introduction of the material, and helps to explain also why pseudo-marginal algorithms are much more difficult to analyse. In Section 3, we formulate our main findings whose proofs are postponed to Section 5. We prepare the proofs by recording variational bounds for the difference of asymptotic variances in Section 4 which we could not find formulated elsewhere in the literature. In Section 5, we develop our main arguments, which rely on an embedding of the probability distributions involved into a single probability distribution: this is possible thanks to a martingale coupling which is itself a by-product of the convex order. In Section 6, we illustrate the usefulness of the framework in practice. We conclude by discussing other applications and extensions of the present work in Section 7.

2. Convex order and a simple application. In this section, we briefly review well-known equivalent characterisations of the convex order which we will use in the remainder of the paper. We also provide an example of algorithms, related to pseudo-marginal algorithms but different in an essential manner, for which the notion of convex order characterises performance without the need of more sophisticated mathematical developments.

The convex order is a natural way of comparing the "variability" or "dispersion" of two random variables or distributions [35, 41].

DEFINITION 3. The random variables $W_1 \sim F_1$ and $W_2 \sim F_2$ are *convex ordered*, denoted $W_1 \leq_{\text{cx}} W_2$ or $F_1 \leq_{\text{cx}} F_2$ hereafter, if for any convex function

$\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

REMARK 4. For integrable W_1 and W_2, the convex order $W_1 \leq_{\text{cx}} W_2$ clearly implies $\mathbb{E}[W_1] = \mathbb{E}[W_2]$ from the convexity of $x \mapsto x$ and $x \mapsto -x$, and if W_1 and W_2 are square integrable, then $\text{var}(W_1) \leq \text{var}(W_2)$, since $x \mapsto x^2$ is convex. The converse, however, does not generally hold true. This turns out to be an important point when discussing the characterisation of performance of pseudo-marginal approximations.

LEMMA 5. *Suppose that* $\mathbb{E}[W_1] = \mathbb{E}[W_2] \in \mathbb{R}$. *Then* $W_1 \leq_{\text{cx}} W_2$ *is equivalent to*:

(a) $\mathbb{E}[(W_1 - t)_+] \leq \mathbb{E}[(W_2 - t)_+]$ *for all* $t \in \mathbb{R}$, *where* $(\cdot)_+ := \max\{\cdot, 0\}$.
(b) $\mathbb{E}[\min\{W_1, t\}] \geq \mathbb{E}[\min\{W_2, t\}]$ *for all* $t \in \mathbb{R}$.

PROOF. Condition (a) is a characterisation of the *increasing convex order* (Definition 3 restricted to nondecreasing convex ϕ) [41], Theorem 4.A.2, and the increasing convex order with identical expectations implies the convex order [41], Theorem 4.A.35. Similarly, (b) implies $\mathbb{E}[(W_1 - t)_-] \geq \mathbb{E}[(W_2 - t)_-]$ which is equivalent to the *increasing concave order* (Definition 3 with nondecreasing concave functions ϕ), implying the desired convex order [41], Theorem 4.A.35. □

REMARK 6. If $-\infty \leq \underline{w} \leq \bar{w} \leq \infty$ are constants such that $W_1, W_2 \in [\underline{w}, \bar{w}]$ almost surely, then it is sufficient to consider the conditions in Lemma 5(a) or (b) with $t \in [\underline{w}, \bar{w}]$ only.

It should be clear by application of Jensen's inequality that if W_1 and W_2 are defined on the same probability space and $\mathbb{E}[W_2|W_1] = W_1$, then $W_1 \leq_{\text{cx}} W_2$. The following characterisation of the convex order, often referred to as Strassen's theorem [44], Theorem 8, establishes the converse, that is, that the convex order implies the existence of this type of martingale representation. This characterisation turns out to be central to our analysis as it allows us to eventually "embed" $\tilde{\pi}_1$ and $\tilde{\pi}_2$ into a unique probability distribution and open up the possibility to use Hilbert space techniques on a common space.

THEOREM 7. *Suppose that* $\mathbb{E}[W_1]$ *and* $\mathbb{E}[W_2]$ *are well-defined. Then* $W_1 \leq_{\text{cx}} W_2$ *if and only if there exists a probability space with random variables* $\check{W}_1$ *and* $\check{W}_2$ *coinciding with* W_1 *and* W_2 *in distribution, respectively, such that* $(\check{W}_1, \check{W}_2)$ *is a martingale pair, that is,* $\mathbb{E}[\check{W}_2|\check{W}_1] = \check{W}_1$ *(a.s.).*

Before turning to the study of pseudo-marginal algorithms, we consider first another class of related algorithms, which are much simpler to analyse, yet practically relevant [11, 27, 36]. These algorithms are closely related to pseudo-marginal algorithms in that they use noisy measurements of the acceptance ratio $r(x, y)$ of P, but the way the approximation is obtained differs in a fundamental way. For pseudo-marginal algorithms, the approximation of $r(x, y)$ stems at each iteration from a previously sampled, and "recycled", approximation of π at $x \in \mathsf{X}$ and a fresh approximation of π at $y \in \mathsf{X}$ (see Algorithm 1). In contrast for the algorithms, we are now concerned with $r(x, y)$ is approximated afresh whenever it is needed. More precisely, with $(\mathsf{X}, \mathcal{X})$ and π as earlier and using a formalism similar to that used for pseudo-marginal algorithms, we consider the following Markov transition probability on $(\mathsf{X}, \mathcal{X})$:

$$\mathring{P}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int \min\{1, r(x, y)\varpi\} Q_{xy}(\mathrm{d}\varpi) + \delta_x(\mathrm{d}y)\mathring{\rho}(x),$$

where $\{Q_{xy}\}_{(x,y)\in\mathsf{X}^2}$ is a family of probability measures on positive reals, $r(x, y)$ is as in Section 1 and $\mathring{\rho}(x) \in [0, 1]$ ensures that this is a Markov transition probability. We stress again on the fact that while $r(x, y)$ is intractable, the product $r(x, y)\varpi$ is assumed to be computable. It can be shown that the condition

$$\int Q_{xy}(\mathrm{d}\varpi)\varpi \mathbb{I}\{\varpi \in A\} = \int Q_{yx}(\mathrm{d}\varpi)\mathbb{I}\left\{\frac{1}{\varpi} \in A\right\}$$

for all $x, y \in \mathsf{X}$ and all $A \in \mathcal{B}(\mathbb{R}_+)$ ensures that $\mathring{P}$ is reversible with respect to π; Lemma 44 in Appendix B details a slightly more general statement.

EXAMPLE 8. Suppose $a_{xy} = a_{yx} > 0$ for all $x, y \in \mathsf{X}$, then the distributions

$$Q_{xy}(\mathrm{d}\varpi) := \frac{\delta_{a_{xy}}(\mathrm{d}\varpi) + a_{xy}\delta_{a_{xy}^{-1}}(\mathrm{d}\varpi)}{1 + a_{xy}}$$

satisfy the condition above. Another more practically relevant case is the log-normal distribution with suitable parameters, which corresponds to the so-called penalty method [11].

A fundamental consequence of the fact that the acceptance ratio is approximated afresh for $x, y \in \mathsf{X}$ whenever it is needed is that $\mathring{P}$ is a Markov chain on $(\mathsf{X}, \mathcal{X})$ and has π as invariant distribution, independently of $\{Q_{xy}\}_{(x,y)\in\mathsf{X}^2}$. This is in contrast with pseudo-marginal algorithms, for which $\{X_n\}_{n\geq 0}$ is not a Markov chain, but $\{X_n, W_n\}_{n\geq 0}$ is, and the invariant distribution of the latter depends on $\{Q_x\}_{x\in\mathsf{X}}$. As a result, the algorithm corresponding to $\mathring{P}$ is particularly simple to analyse in the context of the convex ordering. Indeed, if for some $x, y \in \mathsf{X}$ we have $Q_{xy}^{(1)} \leq_{\mathrm{cx}}$

$Q_{xy}^{(2)}$ then from Lemma 5(b) we have the following inequality for the acceptance probabilities:

$$\int \min\{1, r(x,y)\varpi_2\} Q_{xy}^{(2)}(\mathrm{d}\varpi_2) \leq \int \min\{1, r(x,y)\varpi_1\} Q_{xy}^{(1)}(\mathrm{d}\varpi_1).$$

In the situation where this inequality holds for any $x, y \in \mathsf{X}$, we can apply Peskun's result stated in Theorem 2 directly. More specifically, if $\mathring{P}_1$ and $\mathring{P}_2$ are the algorithms corresponding to $\{Q_{xy}^{(1)}\}_{(x,y)\in\mathsf{X}^2}$ and $\{Q_{xy}^{(2)}\}_{(x,y)\in\mathsf{X}^2}$, the above inequality implies that the probability of leaving any state $x \in \mathsf{X}$ is larger for $\mathring{P}_1$ than for $\mathring{P}_2$ and, therefore, we conclude that $\mathrm{var}(f, \mathring{P}_1) \leq \mathrm{var}(f, \mathring{P}_2)$ and $\mathrm{Gap}_R(\mathring{P}_1) \geq \mathrm{Gap}_R(\mathring{P}_2)$.

One interest of identifying the convex order as an appropriate concept for the comparison of the asymptotic properties of such algorithms is that it allows one to use the wealth of existing results concerning the convex order. For example, it is direct to establish that the diatomic distribution in Example 8 is the worst possible distribution among all the probability distributions with support included in $[a_{xy}^{-1}, a_{xy}]$ and the choice of $a_{xy} = 1$, the "noiseless" algorithm, leads to the best algorithm; see Section 6.3.

Turning back to the pseudo-marginal algorithms, it will be useful in what follows to consider the expected (or at equilibrium) acceptance probability and particularly the conditional expected acceptance probability defined as

$$\alpha(\tilde{P}) := \int \alpha_{xy}(\tilde{P})\pi(\mathrm{d}x)q(x,\mathrm{d}y) \qquad \text{with}$$

$$\alpha_{xy}(\tilde{P}) := \int \min\left\{1, r(x,y)\frac{u}{w}\right\} Q_x(\mathrm{d}w)\, w\, Q_y(\mathrm{d}u),$$

respectively. It is possible to show directly that if for some $x, y \in \mathsf{X}$ the orders $Q_x^{(1)} \leq_{\mathrm{cx}} Q_x^{(2)}$ and $Q_y^{(1)} \leq_{\mathrm{cx}} Q_y^{(2)}$ hold, then $\alpha_{xy}(\tilde{P}_1) \geq \alpha_{xy}(\tilde{P}_2)$, where $\tilde{P}_1$ and $\tilde{P}_2$ denote pseudo-marginal algorithms with $\{Q_x^{(1)}\}_{x\in\mathsf{X}}$ and $\{Q_x^{(2)}\}_{x\in\mathsf{X}}$ [see also Theorem 10(a) and Theorem 22 for a proof]. If $Q_x^{(1)} \leq_{\mathrm{cx}} Q_x^{(2)}$ for all $x \in \mathsf{X}$, then clearly $\alpha(\tilde{P}_1) \geq \alpha(\tilde{P}_2)$.

It is well known that acceptance probabilities are not always a useful performance criterion for MH algorithms. They turn out, however, to be a relevant in the present context. Indeed for a fixed family $\{q(x,\cdot)\}_{x\in\mathsf{X}}$ a larger value $\alpha_{xy}(\tilde{P})$ indicates that at stationarity the transition from x to y is more likely, which is somewhat reminiscent of the off-diagonal order in Theorem 2. Nevertheless, as pointed out earlier, Peskun's result does not apply here since, among other things, $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution. In the next section, we in fact establish that showing $Q_x^{(1)} \leq_{\mathrm{cx}} Q_x^{(2)}$ for all $x \in \mathsf{X}$ is sufficient to imply the desired orders.

3. Main results: Ordering pseudo-marginal MCMC. Our main results are all based on the following conditional convex order assumption on the weight distributions.

DEFINITION 9. Two families of weight distributions $\{Q_x^{(1)}\}_{x\in\mathsf{X}}$ and $\{Q_x^{(2)}\}_{x\in\mathsf{X}}$ satisfy $\{Q_x^{(1)}\}_{x\in\mathsf{X}} \preceq_{\mathrm{cx}} \{Q_x^{(2)}\}_{x\in\mathsf{X}}$ if

$$Q_x^{(1)} \preceq_{\mathrm{cx}} Q_x^{(2)} \qquad \text{for all } x \in \mathsf{X}.$$

The proofs of our results are based on classical Hilbert space techniques for the analysis of reversible Markov chains. We recall here related definitions which will be useful throughout. Let μ be a probability measure and Π a μ-reversible Markov transition kernel on a measurable space $(\mathsf{E}, \mathcal{F})$. For any probability measure ν on $(\mathsf{E}, \mathcal{F})$ and any function $f : \mathsf{E} \to \mathbb{R}$ let, whenever the integrals are well-defined,

$$\nu(f) := \int f(x)\nu(\mathrm{d}x) \quad \text{and} \quad \Pi f(x) := \int f(y)\Pi(x, \mathrm{d}y),$$

and for $k \geq 2$, by induction,

$$\Pi^k f(x) := \int \Pi(x, \mathrm{d}y)\Pi^{k-1} f(y).$$

We further denote $(\nu\Pi^k)f := \nu(\Pi^k f)$, which can be interpreted as a probability measure. Consider next the spaces of square integrable (and centred) functions defined respectively as

$$L^2(\mathsf{E}, \mu) := \{f : \mathsf{E} \to \mathbb{R} : \mu(f^2) < \infty\},$$
$$L_0^2(\mathsf{E}, \mu) := \{f \in L^2(\mathsf{E}, \mu) : \mu(f) = 0\},$$

endowed with the inner product defined for any $f, g \in L^2(\mathsf{E}, \mu)$ as $\langle f, g\rangle_\mu := \int f(x)g(x)\mu(\mathrm{d}x)$, and the associated norm $\|f\|_\mu := \sqrt{[b]\langle f, f\rangle_\mu}$. The Markov kernel Π defines a self-adjoint operator on $L^2(\mathsf{E}, \mu)$. For $f \in L^2(\mathsf{E}, \mu)$, we consider the Dirichlet forms of f associated with Π

$$\mathcal{E}_\Pi(f) := \langle f, (I - \Pi)f\rangle_\mu = \frac{1}{2}\int \mu(\mathrm{d}x)\Pi(x, \mathrm{d}y)\big(f(x) - f(y)\big)^2,$$

where $I(x, A) := \mathbb{I}\{x \in A\}$ stands for the identity operator. The (right) spectral gap of Π is the distance between 1 and the upper end of the spectrum of Π as an operator on $L_0^2(\mathsf{E}, \mu)$, and has the following variational representation by means of the Dirichlet form:

$$\mathrm{Gap}_R(\Pi) := \inf\{\mathcal{E}_\Pi(f) : f \in L_0^2(\mathsf{E}, \mu), \|f\|_\mu = 1\}.$$

Hereafter, for functions $f : \mathsf{X} \to \mathbb{R}$ we will also denote, whenever necessary, by f the functions from $\mathsf{X} \times \mathbb{R}_+ \to \mathbb{R}$ defined by $f(x, w) := f(x)$. We now state our main result, whose proof is postponed to Section 5.

THEOREM 10. *Let π be a probability distribution on some measurable space $(\mathsf{X}, \mathcal{X})$ and let $\tilde{P}_1$ and $\tilde{P}_2$ be two pseudo-marginal approximations of P aiming to sample from π, sharing a common family of marginal proposal probability distributions $\{q(x, \cdot)\}_{x\in\mathsf{X}}$ but with distinct weight distributions satisfying $\{Q_x^{(1)}\}_{x\in\mathsf{X}} \preceq_{\mathrm{cx}} \{Q_x^{(2)}\}_{x\in\mathsf{X}}$. Then*:

(a) *for any* $x, y \in \mathsf{X}$, *the conditional acceptance rates satisfy* $\alpha_{xy}(\tilde{P}_1) \geq \alpha_{xy}(\tilde{P}_2)$,

(b) *for any* $f : \mathsf{X} \to \mathbb{R}$, *the Dirichlet forms satisfy* $\mathcal{E}_{\tilde{P}_1}(f) \geq \mathcal{E}_{\tilde{P}_2}(f)$,

(c) *for any* $f : \mathsf{X} \to \mathbb{R}$ *with* $\mathrm{var}_\pi(f) < \infty$, *the asymptotic variances satisfy* $\mathrm{var}(f, \tilde{P}_1) \leq \mathrm{var}(f, \tilde{P}_2)$,

(d) *the spectral gaps satisfy* $\mathrm{Gap}_R(\tilde{P}_1) \geq \min\{\mathrm{Gap}_R(\tilde{P}_2), 1 - \tilde{\rho}_2^*\}$, *where* $\tilde{\rho}_2^* := \tilde{\pi}_2 - \mathrm{ess\,sup}_{(x,w)} \tilde{\rho}_2(x, w)$, *the essential supremum of the rejection probability corresponding to* $\tilde{P}_2$,

(e) *if* π *is not concentrated on points, that is,* $\pi(\{x\}) = 0$ *for all* $x \in \mathsf{X}$, *then* $\mathrm{Gap}_R(\tilde{P}_1) \geq \mathrm{Gap}_R(\tilde{P}_2)$.

REMARK 11. Theorem 10(c) does not assume the finiteness of the asymptotic variances, and accommodates the scenarios where either $\mathrm{var}(f, \tilde{P}_2)$ or both $\mathrm{var}(f, \tilde{P}_1)$ and $\mathrm{var}(f, \tilde{P}_2)$ are infinite. Establishing finiteness is a separate, but practically important, problem and we now discuss briefly what are in our view the two most applicable approaches to do so. In the case where $\tilde{P}_2$ admits a spectral gap, then finiteness of $\mathrm{var}(f, \tilde{P}_2)$ is guaranteed, because $\pi(f^2) < \infty$ implies that $\tilde{\pi}_2(f^2) < \infty$. In earlier work [4], Proposition 10, we have shown that if the weight distributions $\{Q_x^{(2)}\}_{x \in \mathsf{X}}$ are uniformly bounded, and the marginal algorithm admits a spectral gap, then so does $\tilde{P}_2$. There are also several results in the literature which guarantee geometric convergence, and thus the existence of spectral gaps for certain classes of Metropolis–Hastings algorithms (e.g., [24, 38]).

In the case where $\tilde{P}_2$ is sub-geometrically ergodic and, therefore, does not admit a spectral gap, then the most successful general technique for guaranteeing the finiteness of $\mathrm{var}(f, \tilde{P}_2)$ consists of establishing sub-geometric drift and minorisation conditions [14, 25, 47]. In earlier work, we have shown that such drifts hold for the pseudo-marginal kernel $\tilde{P}_2$ under general moment conditions on $\{Q_x^{(2)}\}_{x \in \mathsf{X}}$ when the marginal algorithm is strongly uniformly ergodic ([4], Proposition 30 and Corollary 31), and in some more specific scenarios such as the independence sampler ([4], Corollary 27) and the random-walk Metropolis ([4], Theorems 38 and 45).

REMARK 12. In the context of Algorithm 1, if there are two possible estimators of $\pi(Y_n)$ that could be used, say $U_n^{(1)}$ and $U_n^{(2)}$, then $U_n^{(1)} \leq_{\mathrm{cx}} U_n^{(2)}$ is equivalent to our assumption $\{Q_x^{(1)}\}_{x \in \mathsf{X}} \leq_{\mathrm{cx}} \{Q_x^{(2)}\}_{x \in \mathsf{X}}$. This is because the convex order is preserved under scaling.

As pointed out in Remark 4, the convex order $W^{(1)} \leq_{\mathrm{cx}} W^{(2)}$ of square-integrable random variables automatically implies $\mathrm{var}(W^{(1)}) \leq \mathrm{var}(W^{(2)})$, but the reverse is not true in general. A natural question is then to ask if $\mathrm{var}(W_x^{(1)}) \leq \mathrm{var}(W_x^{(2)})$ for all $x \in \mathsf{X}$ could be sufficient to imply, for example, $\mathrm{var}(f, \tilde{P}_2) \geq \mathrm{var}(f, \tilde{P}_1)$ for $f \in L^2(\mathsf{X}, \pi)$? The following counterexample shows that this is not the case.

EXAMPLE 13. Consider the situation where $\mathsf{X} = \{-1, 1\}$, $\pi = (1/2, 1/2)$, and the marginal algorithm is a "perfect" independent Metropolis–Hastings (IMH) algorithm, that is, $q(x, \mathrm{d}y) = \pi(\mathrm{d}y)$, for any $x \in \mathsf{X}$. Suppose that the weight distributions are independent of x and given by

$$Q(\mathrm{d}w) := Q_x(\mathrm{d}w) = \frac{b-1}{b-a}\delta_a(\mathrm{d}w) + \frac{1-a}{b-a}\delta_b(\mathrm{d}w),$$

for some $0 \le a \le 1 \le b < \infty$, and that the function of interest is $f(x) = x$. In this case, $\mathrm{var}_Q(f) = (b-1)(1-a)$ and because of the simple structure of the problem (independence with respect to x of Q_x and the choice of an IMH) one can find an explicit expression for the asymptotic variance

$$\mathrm{var}(f, \tilde{P}) = \frac{a(b-1) + (2b-1)b(1-a)}{b-a}.$$

Now one can easily find numerous counterexamples such as the pairs $(a^{(1)}, b^{(1)}) = (0.9208, 3.0046)$ and $(a^{(2)}, b^{(2)}) = (0.6698, 1.4620)$ for which $\mathrm{var}(W^{(1)}) = 0.1587 \ge \mathrm{var}(W^{(2)}) = 0.1526$ but $\mathrm{var}(f, \tilde{P}_1) = 1.4577 \le \mathrm{var}(f, \tilde{P}_2) = 1.5632$.

REMARK 14. Theorem 10(a)–(c) generalise the findings in [4] which state similar bounds in the special case where $\tilde{P}_1$ corresponds to the marginal algorithm, or equivalently, to the degenerate case $Q_x^{(1)} \equiv \delta_1$. Note also that δ_1 is the unique minimal distribution in the convex order; see Section 6.3.

REMARK 15. In practice, one may be interested in a sequence of estimators $\{W_x^{(i)}\}$, where $i \in \mathbb{N}$ is an "accuracy parameter" such as a number of estimators combined by averaging. Suppose that the estimators are increasingly accurate in the convex order, that is, $W_x^{(i+1)} \le_{\mathrm{cx}} W_x^{(i)}$, then Theorem 10 implies that the following mappings from $\mathbb{N}$ to $\mathbb{R}_+$ have the monotonicity properties:

(a) $i \mapsto \alpha_{xy}(\tilde{P}_i)$ is nondecreasing,
(b) $i \mapsto \mathcal{E}_{\tilde{P}_i}(f)$ is nondecreasing,
(c) $i \mapsto \mathrm{var}(f, \tilde{P}_i)$ is nonincreasing.

We suspect that in addition, in scenarios such as those of Section 6.1, the mappings $i \mapsto \alpha_{xy}(\tilde{P}_i)$ and $i \mapsto \mathcal{E}_{\tilde{P}_i}(f)$ are concave and $i \mapsto \mathrm{var}(f, \tilde{P}_i)$ convex, but we have not yet been able to prove this conjecture. See, however, Proposition 26 for a partial result in that direction.

4. Variational bounds for the asymptotic variance. The first result on our journey to prove Theorem 10 is a variational bound on the difference of asymptotic variances. The result, which is of independent interest, shows that the Dirichlet forms associated with Peskun's variance ordering result [37, 46] need not be ordered for all functions, but only certain subclasses of functions. We note that the

result, Theorem 17, offers also a more direct proof of, for example, the results in [4, 46]. We start by stating a powerful variational representation of the quadratic form of the inverse of a positive self-adjoint operator, attributed to Bellman [7], and used for example by Caracciolo, Pelissetto and Sokal [10].

LEMMA 16. *Let A be a self-adjoint operator on a Hilbert space $\mathcal{H}$, satisfying $\langle f, Af\rangle \geq 0$ for all $f \in \mathcal{H}$ and such that the inverse A^{-1} exists. Then*

$$\langle f, A^{-1} f\rangle = \sup_{g\in\mathcal{H}} \big[2\langle f, g\rangle - \langle g, Ag\rangle\big], \tag{3}$$

where the supremum is attained with $g = A^{-1} f$.

Proof of Lemma 16 is given for the reader's convenience in Appendix A.

Lemma 16 provides us with a quick route to Peskun type ordering. More importantly, it leads to important quantitative bounds on the difference between the asymptotic variances of two μ-reversible Markov transition probabilities in terms of Dirichlet forms. Suppose that Π is a Markov kernel on a measurable space $(\mathsf{E}, \mathcal{F})$, reversible with respect to a probability measure μ, and let $\lambda \in [0, 1)$ be any constant. We may introduce the self-adjoint operator (or sub-probability kernel) $(\lambda\Pi)(x, A) := \lambda\Pi(x, A)$ and we extend the definition of the asymptotic variance to this type of (nonprobabilistic) operator as follows. For any $f \in L^2(\mathsf{E}, \mu)$, we let $\bar{f} := f - \mu(f) \in L_0^2(\mathsf{E}, \mu)$ and define

$$\operatorname{var}(f, \lambda\Pi) := \big\langle \bar{f}, (I - \lambda\Pi)^{-1}(I + \lambda\Pi)\bar{f}\big\rangle_\mu = 2\big\langle \bar{f}, (I - \lambda\Pi)^{-1}\bar{f}\big\rangle_\mu - \|\bar{f}\|_\mu^2,$$

where the inverse $(I - \lambda\Pi)^{-1} := \sum_{k=0}^{\infty} \lambda^k \Pi^k$ is a well-defined bounded operator for any $\lambda \in [0, 1)$. From [46], we know that

$$\lim_{\lambda\uparrow 1} \operatorname{var}(f, \lambda\Pi) = \operatorname{var}(f, \Pi) = \lim_{M\to\infty} \operatorname{var}_\mu\left(M^{-1/2} \sum_{i=1}^{M} f(\Phi_i)\right),$$

even in the case where the expression on the right hand side is infinite. Similarly, we extend the definition of Dirichlet forms to $\mathcal{E}_{\lambda\Pi}(f) := \langle f, (I - \lambda\Pi) f\rangle_\mu$.

THEOREM 17. *Let Π_1 and Π_2 be two Markov transition probabilities defined on some measurable space $(\mathsf{E}, \mathcal{F})$ both reversible with respect to the probability distribution μ, and let $f \in L_0^2(\mathsf{E}, \mu)$. Then*:

(a) *For any $\lambda \in [0, 1)$,*

$$\begin{aligned}\mathcal{E}_{\lambda\Pi_1}\big(\hat{f}_1^\lambda\big) - \mathcal{E}_{\lambda\Pi_2}\big(\hat{f}_1^\lambda\big) &\leq \tfrac{1}{2}\big[\operatorname{var}(f, \lambda\Pi_2) - \operatorname{var}(f, \lambda\Pi_1)\big] \\ &\leq \mathcal{E}_{\lambda\Pi_1}\big(\hat{f}_2^\lambda\big) - \mathcal{E}_{\lambda\Pi_2}\big(\hat{f}_2^\lambda\big),\end{aligned}$$

where $\hat{f}_i^\lambda := (I - \lambda\Pi_i)^{-1} f \in L_0^2(\mathsf{E}, \mu)$.

(b) *The function* $\beta \mapsto \operatorname{var}(f, \beta\Pi_1 + (1-\beta)\Pi_2)$ *defined for* $\beta \in [0,1]$ *is convex, that is, for any* $\beta \in [0,1]$,

$$\operatorname{var}\big(f, \beta\Pi_1 + (1-\beta)\Pi_2\big) \le \beta \operatorname{var}(f, \Pi_1) + (1-\beta)\operatorname{var}(f, \Pi_2).$$

PROOF. In order to prove the results, we use the variational representation of the asymptotic variance, as suggested in [10] and obtained by application of Lemma 16 for $i \in \{1,2\}$,

$$\begin{aligned}\operatorname{var}(f, \lambda\Pi_i) &= 2\big\langle f, (I - \lambda\Pi_i)^{-1} f\big\rangle_\mu - \|f\|_\mu^2 \\ &= 2 \sup_{g \in L_0^2(\mathsf{E},\mu)} \big[2\langle f, g\rangle_\mu - \mathcal{E}_{\lambda\Pi_i}(g)\big] - \|f\|_\mu^2.\end{aligned} \tag{4}$$

Hereafter, we denote $\bar{\imath} = 2$ if $i = 1$ and vice versa. From (4) and Lemma 16 which states that the supremum above is attained for $\hat{f}_i^\lambda$, we have for $i \in \{1,2\}$

$$2\big[2\big\langle f, \hat{f}_{\bar{\imath}}^\lambda\big\rangle_\mu - \mathcal{E}_{\lambda\Pi_i}\big(\hat{f}_{\bar{\imath}}^\lambda\big)\big] \le \operatorname{var}(f, \lambda\Pi_i) + \|f\|_\mu^2 = 2\big[2\big\langle f, \hat{f}_i^\lambda\big\rangle_\mu - \mathcal{E}_{\lambda\Pi_i}\big(\hat{f}_i^\lambda\big)\big].$$

We can now conclude (a) by summing the above inequality with $i = 1$ and with $i = 2$ multiplied by -1, and then dividing by 2. For the second item (b), let $\beta \in (0,1)$ and write for any $g \in L^2(\mathsf{E}, \mu)$,

$$\begin{aligned}&2\langle f, g\rangle_\mu - \mathcal{E}_{\beta\Pi_1 + (1-\beta)\Pi_2}(g) \\ &\quad = \beta\big[2\langle f, g\rangle_\mu - \mathcal{E}_{\Pi_1}(g)\big] + (1-\beta)\big[2\langle f, g\rangle_\mu - \mathcal{E}_{\Pi_2}(g)\big].\end{aligned}$$

The claim follows by taking the supremum over $g \in L^2(\mathsf{E}, \mu)$, separately for the two terms on the right-hand side. □

5. Proofs by a martingale coupling of pseudo-marginal kernels. We preface the proof of Theorem 10 with a key result from [31], which ensures that the conditional convex order implies a conditional martingale coupling of the distributions involved.

THEOREM 18. *Assume* $\{Q_x^{(1)}\}_{x\in\mathsf{X}} \le_{\mathrm{cx}} \{Q_x^{(2)}\}_{x\in\mathsf{X}}$, *then there exists a probability kernel* $(x, A) \mapsto R_x(A)$ *from* X *to* $\mathbb{R}_+^2$ *such that for any* $x \in \mathsf{X}$:

(a) R_x *has marginals* $Q_x^{(1)}$ *and* $Q_x^{(2)}$, *that is,* $R_x(A \times \mathbb{R}_+) = Q_x^{(1)}(A)$ *and* $R_x(\mathbb{R}_+ \times A) = Q_x^{(2)}(A)$ *for all* $A \in \mathcal{B}(\mathbb{R}_+)$,

(b) R_x *is the distribution of a martingale, that is, for all* $A \in \mathcal{B}(\mathbb{R}_+)$, $\int R_x(\mathrm{d}w \times \mathrm{d}v) v \mathbb{I}\{w \in A\} = \int R_x(\mathrm{d}w \times \mathrm{d}v) w \mathbb{I}\{w \in A\}$.

REMARK 19. The property of R_x in Lemma 18(b) holds if and only if $(W, V) \sim R_x(\cdot)$ satisfies $\mathbb{E}[V|W] = W$ almost surely. This means that $\Delta = V - W$ is a martingale difference satisfying $\mathbb{E}[\Delta|W] = 0$, and that $V = W + \Delta$ is "noisier" than W.

The proof of Theorem 18 given in [31], Theorem 1.3, relies on the fundamental martingale characterisation due to Strassen [44] restated in Theorem 7, but involves a nontrivial measurability argument for the case where X is uncountable.

For the rest of this section, we assume that the conditions in Theorem 10 hold, and we denote by $\tilde{\pi}_1$, $\tilde{\pi}_2$ the invariant distributions of $\tilde{P}_1$, $\tilde{P}_2$, respectively. Theorem 18 turns out to be the key instrument in the proof of Theorem 10. It will allow us to circumvent the difficulty of having two distinct invariant distributions $\tilde{\pi}_1$ and $\tilde{\pi}_2$ for $\tilde{P}_1$ and $\tilde{P}_2$, which is incompatible with the Hilbert space setting. Instead, we will be working with two Markov kernels $\breve{P}_1$ and $\breve{P}_2$ equivalent to $\tilde{P}_1$ and $\tilde{P}_2$ in a sense to be made more precise in Lemma 20. The kernels $\breve{P}_1$ and $\breve{P}_2$ introduced below can be thought of as corresponding to two distinct pseudo-marginal implementations, where $\breve{P}_2$ uses the "noisiest" approximation.

LEMMA 20. *Let R_x be the probability kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathbb{R}_+^2, \mathcal{B}(\mathbb{R}_+)^2)$ from Theorem 18. Then the following defines a probability distribution on $(\mathsf{X} \times (0,\infty)^2, \mathcal{X} \times \mathcal{B}((0,\infty)^2))$,*

$$\breve{\pi}(\mathrm{d}x \times \mathrm{d}w \times \mathrm{d}v) := \pi(\mathrm{d}x) R_x(\mathrm{d}w \times \mathrm{d}v) v,$$

and the following define Markov transition probabilities on $(\mathsf{X} \times (0,\infty)^2, \mathcal{X} \times \mathcal{B}((0,\infty)^2))$

$$\begin{aligned}
\breve{P}_1(x,w,v;\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t) &:= q(x,\mathrm{d}y) R_y(\mathrm{d}u \times \mathrm{d}t) \frac{t}{u} \min\left\{1, r(x,y)\frac{u}{w}\right\} \\
&\quad + \delta_{x,w,v}(\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t)\tilde{\rho}_1(x,w), \\
\breve{P}_2(x,w,v;\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t) &:= q(x,\mathrm{d}y) R_y(\mathrm{d}u \times \mathrm{d}t) \min\left\{1, r(x,y)\frac{t}{v}\right\} \\
&\quad + \delta_{x,w,v}(\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t)\tilde{\rho}_2(x,v),
\end{aligned}$$

with the convention $t/u = 0$ for $t = u = 0$, and where $\tilde{\rho}_1$ and $\tilde{\rho}_2$ stand for the rejection probabilities of $\tilde{P}_1$ and $\tilde{P}_2$, respectively. The following marginal equivalences hold between $\breve{\pi}$, $\breve{P}_1$, $\breve{P}_2$ and $\tilde{\pi}_1$, $\tilde{\pi}_2$, $\tilde{P}_1$, $\tilde{P}_2$,

$$\begin{aligned}
\breve{\pi}\big(\mathrm{d}x \times \mathrm{d}w \times (0,\infty)\big) &= \tilde{\pi}_1(\mathrm{d}x \times \mathrm{d}w), \\
\breve{\pi}\big(\mathrm{d}x \times (0,\infty) \times \mathrm{d}v\big) &= \tilde{\pi}_2(\mathrm{d}x \times \mathrm{d}v), \\
\breve{P}_1\big(x,w,v;\mathrm{d}y \times \mathrm{d}u \times (0,\infty)\big) &= \tilde{P}_1(x,w;\mathrm{d}y \times \mathrm{d}u), \\
\breve{P}_2\big(x,w,v;\mathrm{d}y \times (0,\infty) \times \mathrm{d}t\big) &= \tilde{P}_2(x,v;\mathrm{d}y \times \mathrm{d}t),
\end{aligned}$$

where the latter two equalities hold for all $(x,w,v) \in \mathsf{X} \times (0,\infty)^2$. Furthermore, both $\breve{P}_1$ and $\breve{P}_2$ are reversible with respect to $\breve{\pi}$.

PROOF. The measure $\breve{\pi}$ is positive, and by the properties of R_x,

$$\begin{aligned}\breve{\pi}\big(\mathsf{X}\times(0,\infty)^2\big) &= \int_{\mathsf{X}} \pi(\mathrm{d}x) \int_{(0,\infty)\times\mathbb{R}_+} R_x(\mathrm{d}w\times\mathrm{d}v)v \\ &= \int_{\mathsf{X}} \pi(\mathrm{d}x) \int_{(0,\infty)} Q_x^{(1)}(\mathrm{d}w)w = 1,\end{aligned} \tag{5}$$

and the marginal properties follow similarly. The marginal correspondence between $\breve{P}_1$, $\breve{P}_2$ and $\tilde{P}_1$, $\tilde{P}_2$ is also immediate. Clearly, $\breve{P}_2$ is a Metropolis–Hastings algorithm with proposal $q(x,\mathrm{d}y)R_y(\mathrm{d}u\times\mathrm{d}t)$ targeting $\breve{\pi}$, which implies also the reversibility.

We then turn to the reversibility of $\breve{P}_1$. We may focus on the off-diagonal part (cf. [46]) and write for any $A, B \in \mathcal{X}\times\mathcal{B}(\mathbb{R}_+)^2$ with $A, B \subset \{w>0, u>0\}$,

$$\begin{aligned}&\int \pi(\mathrm{d}x)R_x(\mathrm{d}w\times\mathrm{d}v)vq(x,\mathrm{d}y)R_y(\mathrm{d}u\times\mathrm{d}t) \\ &\quad\times\frac{t}{u}\min\Big\{1, r(x,y)\frac{u}{w}\Big\}\mathbb{I}\big\{(x,w,v)\in A, (y,u,t)\in B\big\} \\ &\quad= \int \pi(\mathrm{d}y)q(y,\mathrm{d}x)R_y(\mathrm{d}u\times\mathrm{d}t)tR_x(\mathrm{d}w\times\mathrm{d}v) \\ &\qquad\times\frac{v}{w}\min\Big\{\frac{w}{u}r(y,x), 1\Big\}\mathbb{I}\big\{(y,u,t)\in B, (x,w,v)\in A\big\},\end{aligned}$$

which is enough to conclude. □

We next introduce the spaces of square integrable functions which are constant with respect to the last, the second last and the two last coordinates, respectively,

$$\begin{aligned}L^2_{c2}\big(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}\big) &:= \big\{f : \exists f_1 \in L^2(\mathsf{X}\times\mathbb{R}_+,\tilde{\pi}_1),\, f(x,w,v) = f_1(x,w)\big\}, \\ L^2_{c1}\big(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}\big) &:= \big\{f : \exists f_2 \in L^2(\mathsf{X}\times\mathbb{R}_+,\tilde{\pi}_2),\, f(x,w,v) = f_2(x,v)\big\}, \\ L^2_{c}\big(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}\big) &:= \big\{f : \exists \bar{f} \in L^2(\mathsf{X},\pi),\, f(x,w,v) = \bar{f}(x)\big\},\end{aligned}$$

where $f \in L^2(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi})$ in the definitions. We denote the corresponding classes of zero-mean functions as $L^2_{0,c*}(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}) := \{f \in L^2_{c*}(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}) : \breve{\pi}(f) = 0\}$, where "$*$" is a placeholder. The next corollary of Lemma 20 records properties of $\breve{P}_i$ on the above mentioned classes of functions.

COROLLARY 21. *Let $f_1 \in L^2(\mathsf{X}\times\mathbb{R}_+,\tilde{\pi}_1)$ and $f_2 \in L^2(\mathsf{X}\times\mathbb{R}_+,\tilde{\pi}_2)$, and denote $g_1(x,w,v) := f_1(x,w) \in L^2_{c2}(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi})$ and $g_2(x,w,v) := f_2(x,v) \in L^2_{c1}(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi})$. Then we have the correspondence for $k \geq 1$:*

$$\breve{\pi}(g_1) = \tilde{\pi}_1(f_1), \qquad \breve{P}_1^k g_1(x,w,v) = \tilde{P}_1^k f_1(x,w) \in L^2_{c2}\big(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}\big), \tag{6}$$

$$\breve{\pi}(g_2) = \tilde{\pi}_2(f_2), \qquad \breve{P}_2^k g_2(x,w,v) = \tilde{P}_2^k f_2(x,v) \in L^2_{c1}\big(\mathsf{X}\times\mathbb{R}^2_+,\breve{\pi}\big) \tag{7}$$

and as a result for $k \geq 1$ the following identities hold:

$$\langle g_1, \breve{P}_1^k g_1 \rangle_{\breve{\pi}} = \langle f_1, \tilde{P}_1^k f_1 \rangle_{\tilde{\pi}_1}, \qquad \langle g_2, \breve{P}_2^k g_2 \rangle_{\breve{\pi}} = \langle f_2, \tilde{P}_2^k f_2 \rangle_{\tilde{\pi}_2},$$
$$\operatorname{var}(g_1, \breve{P}_1) = \operatorname{var}(f_1, \tilde{P}_1), \qquad \operatorname{var}(g_2, \breve{P}_2) = \operatorname{var}(f_2, \tilde{P}_2).$$

PROOF. All the claims are direct, except the last two. Recall that (e.g., [46]) for a generic Markov kernel Π reversible with respect to a probability distribution μ and $\psi \in L_0^2(\mathsf{E}, \mu)$ we have

$$\operatorname{var}(\psi, \Pi) = \lim_{M \to \infty} \frac{1}{M} \left(\sum_{i=1}^{M} \mathbb{E}_\mu[\psi^2(\Phi_i)] + 2 \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathbb{E}_\mu[\psi(\Phi_i)\psi(\Phi_j)] \right)$$
$$= \mu(\psi^2) + \lim_{n \to \infty} \frac{2}{M} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \langle \psi, \Pi^{j-i} \psi \rangle_\mu.$$

The last two claims now follow from the correlation equivalences. □

Note that by Corollary 21, $L_c^2(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi}) \subset L_{c1}^2(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi}) \cap L_{c2}^2(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi}) \subset L^2(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi})$ and the same inclusions hold with centred versions $L_{0,c*}(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi})$.

We now state the key result which relates various quantities related to the pseudo-marginal algorithms $\tilde{P}_1$ and $\tilde{P}_2$ and their counterparts $\breve{P}_1$ and $\breve{P}_2$.

THEOREM 22. *Let $\breve{\pi}$, $\breve{P}_1$ and $\breve{P}_2$ be as defined in Lemma 20. Then*:

(a) $\alpha_{xy}(\tilde{P}_1) \geq \alpha_{xy}(\tilde{P}_2)$ *for any* $x, y \in \mathsf{X}$,
(b) $\mathcal{E}_{\breve{P}_1}(g) \geq \mathcal{E}_{\breve{P}_2}(g)$ *for any* $g \in L_{c2}^2(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi})$,
(c) $\operatorname{var}(f, \breve{P}_1) \leq \operatorname{var}(f, \breve{P}_2)$ *for any* $f \in L_{c2}^2(\mathsf{X} \times \mathbb{R}_+^2, \breve{\pi})$,
(d) $\operatorname{Gap}_R(\tilde{P}_1) \geq \operatorname{Gap}_R(\breve{P}_2)$.

PROOF. We first consider (a) and (b). Fix $x, y \in \mathsf{X}$. Then for any bounded function $h : (\mathsf{X} \times \mathbb{R}_+)^2 \to \mathbb{R}_+$ by the properties of R_x and R_y and by Jensen's inequality,

$$\int R_x(\mathrm{d}w \times \mathrm{d}v) v R_y(\mathrm{d}u \times \mathrm{d}t) \frac{t}{u} \min\left\{1, r(x, y)\frac{u}{w}\right\} h(x, w, y, u)$$
$$= \int R_x(\mathrm{d}w \times \mathrm{d}v) R_y(\mathrm{d}u \times \mathrm{d}t) \min\{w, r(x, y)u\} h(x, w, y, u)$$
$$\geq \int R_x(\mathrm{d}w \times \mathrm{d}v) R_y(\mathrm{d}u \times \mathrm{d}t) \min\{v, r(x, y)t\} h(x, w, y, u)$$
$$= \int R_x(\mathrm{d}w \times \mathrm{d}v) v R_y(\mathrm{d}u \times \mathrm{d}t) \min\left\{1, r(x, y)\frac{t}{v}\right\} h(x, w, y, u).$$

We deduce (a) with $h \equiv 1$ and by using the correspondence established in Lemma 20. We also have, with functions such that $h(x, w, x, w) = 0$ for all $(x, w) \in \mathsf{X} \times (0, \infty)$,

$$\int \breve{\pi}(\mathrm{d}x \times \mathrm{d}w \times \mathrm{d}v)\breve{P}_1(x, w, v; \mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t)h(x, w, y, u)$$
$$\geq \int \breve{\pi}(\mathrm{d}x \times \mathrm{d}w \times \mathrm{d}v)\breve{P}_2(x, w, v; \mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t)h(x, w, y, u).$$

Claim (b) is now obtained by letting $h(x, w, y, u) = \min\{m, (g(x, w) - g(y, u))^2\}$ and by monotone convergence as $m \to \infty$.

In (c), we may assume without loss of generality that $f \in L^2_{0,c2}(\mathsf{X} \times \mathbb{R}^2_+, \breve{\pi})$. For any $\lambda \in [0, 1)$, note that by Corollary 21 $\hat{f}^\lambda_1 := (I - \lambda\breve{P}_1)^{-1} f = \sum_{k=0}^\infty \lambda^k (\breve{P}_1)^k f$ satisfies $\hat{f}^\lambda_1(x, w, v) = \hat{f}^\lambda_1(x, w) \in L^2_{0,c2}(\mathsf{X} \times \mathbb{R}^2_+, \breve{\pi})$. We may now apply (b) and Theorem 17 to deduce that

$$0 \leq 2\lambda[\mathcal{E}_{\breve{P}_1}(\hat{f}^\lambda_1) - \mathcal{E}_{\breve{P}_2}(\hat{f}^\lambda_1)] = 2[\mathcal{E}_{\lambda\breve{P}_1}(\hat{f}^\lambda_1) - \mathcal{E}_{\lambda\breve{P}_2}(\hat{f}^\lambda_1)]$$
$$\leq \mathrm{var}(f, \lambda\breve{P}_2) - \mathrm{var}(f, \lambda\breve{P}_1).$$

If $\mathrm{var}(f, \breve{P}_2)$ is infinite, then the claim holds trivially. Suppose now that $\mathrm{var}(f, \breve{P}_2)$ is finite, then taking the limit $\lambda \uparrow 1$ ensures that $\mathrm{var}(f, \breve{P}_2) \geq \mathrm{var}(f, \breve{P}_1)$.

For (d), by the variational definition of the right spectral gap there exists a sequence of functions $\psi_i \in L^2_0(\mathsf{X} \times \mathbb{R}_+, \tilde{\pi}_1)$ with $\mathrm{var}_{\tilde{\pi}_1}(\psi_i) = 1$ such that

$$\lim_{i\to\infty} \mathcal{E}_{\tilde{P}_1}(\psi_i) = \mathrm{Gap}_R(\tilde{P}_1).$$

Notice that denoting $\psi_i(x, w, v) := \psi_i(x, w)$ Corollary 21 implies $\mathrm{var}_{\breve{\pi}}(\psi_i) = \mathrm{var}_{\tilde{\pi}_1}(\psi_i) = 1$ and $\breve{P}_1\psi_i(x, w, v) = \tilde{P}_1\psi_i(x, w)$ and, therefore, $\mathcal{E}_{\breve{P}_1}(\psi_i) = \mathcal{E}_{\tilde{P}_1}(\psi_i)$. Now, (b) allows us to conclude that

$$\mathrm{Gap}_R(\breve{P}_2) \leq \liminf_{i\to\infty} \mathcal{E}_{\breve{P}_2}(\psi_i) \leq \liminf_{i\to\infty} \mathcal{E}_{\breve{P}_1}(\psi_i) = \lim_{i\to\infty} \mathcal{E}_{\tilde{P}_1}(\psi_i). \qquad \square$$

We are now ready to complete the proof of our main result.

PROOF OF THEOREM 10. The acceptance rate order (a) is proved in Theorem 22(a). For what follows, let $K_1(x, u; \cdot)$ and $K_2(x, w; \cdot)$ be (regular) conditional distributions such that

$$R_x(\mathrm{d}w \times \mathrm{d}u) = Q^{(1)}_x(\mathrm{d}w)K_2(x, w; \mathrm{d}u),$$
$$R_x(\mathrm{d}w \times \mathrm{d}u) = Q^{(2)}_x(\mathrm{d}u)K_1(x, u; \mathrm{d}w),$$

and define the following sub-probability kernels corresponding to the acceptance parts of $\tilde{P}_1$ and $\tilde{P}_2$,

$$\tilde{p}_1(x, w; \mathrm{d}y \times \mathrm{d}u) := q(x, \mathrm{d}y)Q^{(1)}_y(\mathrm{d}u) \min\left\{1, r(x, y)\frac{u}{w}\right\},$$

$$\tilde{p}_2(x, v; \mathrm{d}y \times \mathrm{d}t) := q(x, \mathrm{d}y) Q_y^{(2)}(\mathrm{d}t) \min\left\{1, r(x, y)\frac{t}{v}\right\}.$$

With these, we may write

$$\begin{aligned}
\breve{P}_1(x, w, v; \mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t) &= \tilde{p}_1(x, w; \mathrm{d}y \times \mathrm{d}u) K_2(y, u; \mathrm{d}t)\frac{t}{u} \\
&\quad + \delta_{x,w,v}(\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t)\tilde{\rho}_1(x, w), \\
\breve{P}_2(x, w, v; \mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t) &= \tilde{p}_2(x, v; \mathrm{d}y \times \mathrm{d}t) K_1(y, t; \mathrm{d}u) \\
&\quad + \delta_{x,w,v}(\mathrm{d}y \times \mathrm{d}u \times \mathrm{d}t)\tilde{\rho}_2(x, v).
\end{aligned}$$

In both situations, we are now in the setting of Lemma 45 in Appendix C with $\mathsf{E} = \mathsf{X} \times (0, \infty)$ and $\mathsf{S} = (0, \infty)$.

Define $h(x, w, v) := f(x)$ and $g(x, w) := f(x)$, then we have from Theorem 22(b) that $\mathcal{E}_{\tilde{P}_1}(g) = \mathcal{E}_{\breve{P}_1}(h) \geq \mathcal{E}_{\breve{P}_2}(h) = \mathcal{E}_{\tilde{P}_2}(g)$, which completes the proof of (b). Claim (c) follows from Theorem 22(c) and Corollary 21. Now recall that Theorem 22(d) states that $\mathrm{Gap}_R(\tilde{P}_1) \geq \mathrm{Gap}_R(\breve{P}_2)$, and Lemma 45 reads $\mathrm{Gap}_R(\breve{P}_2) \geq \min\{\mathrm{Gap}_R(\tilde{P}_2), 1 - \tilde{\rho}_2^*\}$, which completes the proof of (d). Finally, (e) is a consequence of Remark 46 in Appendix C. □

We conclude this section by a partial result concerning the convexity and concavity of the expected acceptance rates, the Dirichlet forms and the asymptotic variances as discussed in Remark 15. We start by stating simple extensions of Theorem 18 and Lemma 20 to the case of an arbitrary number of distributions, which may be useful also in other contexts. Note that here the indices are reversed in comparison to Remark 15; $Q_x^{(1)}$ corresponds to the least noisy estimate.

LEMMA 23. *Suppose that* $\{Q_x^{(1)}\}_{x\in\mathsf{X}} \leq_{\mathrm{cx}} \cdots \leq_{\mathrm{cx}} \{Q_x^{(n)}\}_{x\in\mathsf{X}}$, *then there exists a kernel* $(x, A) \mapsto R_x(A)$ *from* X *to* $\mathbb{R}_+^n$ *such that, with notation* $w_{1:i} := (w_1, \ldots, w_i)$:

(a) *For all* $i = 1, \ldots, n$, $R_x(\mathbb{R}_+^{i-1} \times A \times \mathbb{R}_+^{n-i-1}) = Q_x^{(i)}(A)$.

(b) *For all* $A \in \mathcal{B}(\mathbb{R}_+^{i-1})$ *and all* $i = 2, \ldots, n$,

$$\int R_x(\mathrm{d}w_{1:n}) w_i \mathbb{I}\{w_{1:i-1} \in A\} = \int R_x(\mathrm{d}w_{1:n}) w_{i-1} \mathbb{I}\{w_{1:i-1} \in A\}.$$

(c) *For all* $i = 3, \ldots, n$, *all* $A_1, \ldots, A_{i-1} \in \mathcal{B}(\mathbb{R}_+)$ *and all bounded measurable* $f : \mathbb{R}_+ \to \mathbb{R}$,

$$\begin{aligned}
&\int R_x(\mathrm{d}w_{1:n}) f(w_i) \mathbb{I}\{w_1 \in A_1, \ldots, w_{i-1} \in A_{i-1}\} \\
&\quad = \int R_x(w_{1:n}) f(w_i) \mathbb{I}\{w_{i-1} \in A_{i-1}\}.
\end{aligned}$$

PROOF. For the existence of R_x, consider Theorem 18 applied to each pair $\{Q_x^{(i-1)}\}_{x\in\mathsf{X}} \leq_{\mathrm{cx}} \{Q_x^{(i)}\}_{x\in\mathsf{X}}$, resulting in $R_x^{(i)}(\mathrm{d}w_{i-1}\times\mathrm{d}w_i)$, and let $K^{(i)}(x,w_{i-1};\mathrm{d}w_i)$ be (regular) conditional probabilities such that $R_x^{(i)}(\mathrm{d}w_{i-1}\times\mathrm{d}w_i) = Q_x^{(i-1)}(\mathrm{d}w_{i-1})K(x,w_{i-1};\mathrm{d}w_i)$. The claim holds for $n=2$ because then $R_x = R_x^{(2)}$. For $n\geq 3$, assuming that the claim holds with $n-1$, it is straightforward to check that the extension

$$R_x(\mathrm{d}w_{1:n}) = R_x(\mathrm{d}w_{1:n-1})K^{(n)}(x,w_{1:n-1};\mathrm{d}w_n)$$

satisfies the required properties. □

REMARK 24. If $\hat{W}_{1:n}\sim R_x$, then $\hat{W}_{1:n}$ is a Markovian martingale, that is, $\mathbb{E}[f(\hat{W}_i)|\hat{W}_{1:i-1}] = \mathbb{E}[\hat{W}_i|\hat{W}_{i-1}]$ and $\mathbb{E}[\hat{W}_i|\hat{W}_{1:i-1}] = \hat{W}_{i-1}$ for $i=2,\dots,n$ and all bounded measurable $f:\mathbb{R}_+\to\mathbb{R}$.

LEMMA 25. *Let $\{Q_x^{(1)}\}_{x\in\mathsf{X}} \leq_{\mathrm{cx}} \cdots \leq_{\mathrm{cx}} \{Q_x^{(n)}\}_{x\in\mathsf{X}}$ and let R_x be as in Lemma 23. Define the probability distribution*

$$\breve{\pi}(\mathrm{d}x\times\mathrm{d}w_{1:n}) := \pi(\mathrm{d}x)R_x(\mathrm{d}w_{1:n})w_n,$$

and the following Markov transition probabilities on $(\mathsf{X}\times(0,\infty)^n, \mathcal{X}\times\mathcal{B}((0,\infty)^n))$:

$$\breve{P}_i(x,w_{1:n};\mathrm{d}y\times\mathrm{d}u_{1:n}) := q(x,\mathrm{d}y)R_y(\mathrm{d}u_{1:n})\frac{u_n}{u_i}\min\left\{1, r(x,y)\frac{u_i}{w_i}\right\} + \delta_{x,w_{1:n}}(\mathrm{d}y\times\mathrm{d}u_{1:n})\tilde{\rho}_i(x,w_i).$$

Then the $\breve{P}_i$s are reversible with respect to $\breve{\pi}$, and satisfy the marginal correspondence

$$\breve{\pi}\big(\mathrm{d}x\times\mathbb{R}_+^{i-1}\times\mathrm{d}w_i\times\mathbb{R}_+^{n-i-1}\big) = \tilde{\pi}_i(\mathrm{d}x\times\mathrm{d}w_i),$$
$$\breve{P}_i\big(x,w_{1:n};\mathrm{d}y\times\mathbb{R}_+^{i-1}\times\mathrm{d}u_i\times\mathbb{R}_+^{n-i-1}\big) = \tilde{P}_i(x,w_i;\mathrm{d}y\times\mathrm{d}u_i).$$

The proof is similar to Lemma 20.

We now give our partial result relying on an abstract condition on the Dirichlet forms of the augmented kernels $\breve{P}_i$ defined in Lemma 25.

PROPOSITION 26. *Let $\{Q_x^{(1)}\}_{x\in\mathsf{X}} \leq_{\mathrm{cx}} \cdots \leq_{\mathrm{cx}} \{Q_x^{(n)}\}_{x\in\mathsf{X}}$ and let $\breve{P}_i$ be as defined in Lemma 25. If for all $i=2,\dots,n$ and any $g_i\in L_0^2(\mathsf{X}\times\mathbb{R}_+^2,\breve{\pi})$ such that $g_i(x,w_{1:n}) = h(x,w_i)\in L^2(\tilde{\pi}_i,\mathsf{X}\times\mathbb{R}_+)$ it holds that*

$$\mathcal{E}_{\breve{P}_{i-1}}(g_i) - \mathcal{E}_{\breve{P}_i}(g_i) \leq \mathcal{E}_{\breve{P}_i}(g_i) - \mathcal{E}_{\breve{P}_{i+1}}(g_i),$$

then for any function $f\in L^2(\mathsf{X},\pi)$,

$$\mathrm{var}(f,\tilde{P}_i) - \mathrm{var}(f,\tilde{P}_{i-1}) \leq \mathrm{var}(f,\tilde{P}_{i+1}) - \mathrm{var}(f,\tilde{P}_i),$$

whenever the quantities above are finite.

PROOF. Without loss of generality, we may assume $f \in L_0^2(\mathsf{X}, \pi)$, so that $\hat{g}_i^\lambda := \sum_{k\geq 0} \lambda \breve{P}_i f \in L_0^2(\mathsf{X} \times \mathbb{R}_+^n, \breve{\pi})$, and $\hat{g}_i^\lambda(x, w_{1:n})$ depends only on x and w_i. By Theorem 17,

$$\begin{aligned}\big(\operatorname{var}(g, \lambda \breve{P}_i) - \operatorname{var}(g, \lambda \breve{P}_{i-1})\big) &\leq 2\big[\mathcal{E}_{\lambda \breve{P}_{i-1}}(\hat{g}_i^\lambda) - \mathcal{E}_{\lambda \breve{P}_i}(\hat{g}_i^\lambda)\big] \\ &= 2\lambda\big[\mathcal{E}_{\breve{P}_{i-1}}(\hat{g}_i^\lambda) - \mathcal{E}_{\breve{P}_i}(\hat{g}_i^\lambda)\big],\end{aligned}$$

and similarly

$$2\lambda\big[\mathcal{E}_{\breve{P}_i}(\hat{g}_i^\lambda) - \mathcal{E}_{\breve{P}_{i+1}}(\hat{g}_i^\lambda)\big] \leq \operatorname{var}(g, \lambda \breve{P}_{i+1}) - \operatorname{var}(g, \lambda \breve{P}_i).$$

Because $\mathcal{E}_{\breve{P}_{i-1}}(\hat{g}_i^\lambda) - \mathcal{E}_{\breve{P}_i}(\hat{g}_i^\lambda) \leq \mathcal{E}_{\breve{P}_i}(\hat{g}_i^\lambda) - \mathcal{E}_{\breve{P}_{i+1}}(\hat{g}_i^\lambda)$, we obtain the desired variance bound for $\breve{P}_{i-1}$, $\breve{P}_i$ and $\breve{P}_{i+1}$. Because the variances are equal to those of $\tilde{P}_{i-1}$, $\tilde{P}_i$ and $\tilde{P}_{i+1}$ as observed in the proof of Theorem 10, we complete the proof. □

6. Applications. The convex order is a well-researched topic with a rich and extensive literature where numerous properties have been established for various purposes; see, for example, [35, 41] for recent book length overviews. For example, the convex order is closed under linear combinations and numerous parametric families of distributions can be convex ordered in terms of their parameters. Conditioning improves on convex order, that is, $\mathbb{E}[W|Z] \leq_{\text{cx}} W$ for some random variable Z (see Theorem 3.A.20 for a more general scenario, in [41]), therefore, establishing that, as expected, "Rao–Blackwellisation" is beneficial in the present context. We detail here applications of such properties directly relevant to the pseudo-marginal context.

We first show in Section 6.1 that the theory of majorisation provides us with a tool to compare algorithms when averaging a number of independent realisations of an approximation. As a by-product, we establish that increasing the number of copies always improves performance. While this result is not entirely surprising, Example 13 is a reminder that the behaviour of these algorithms can be counter-intuitive and surprising. In addition, establishing the result directly seems to be far from obvious while it follows here directly from the convex order. Our second application is more interesting in terms of methodology. It is concerned with stratification, which is often easy to implement without additional computational cost. In particular, we establish in Section 6.2 that the standard application of this variance reduction approach to approximate Bayesian computation (ABC) MCMC always improves performance in this context. We conclude by considering extremal distributions in Section 6.3 and discuss what information they provide on the efficiency under certain constraints. We point again to a recent application of our work in [9] in order to establish quantitative bounds.

6.1. *Averaging and performance monotonicity in the pseudo-marginal algorithm.* A simple and practical way to reduce variability of an estimator is to average multiple independent realisations of this estimator—this is a particularly interesting and relevant strategy given the advent of cheap and widely available parallel computing architectures; see [16] for a recent application of this idea to pseudo-marginal algorithms. It is standard to show that for N independent and identical realisations of an estimator the choice of uniform weights $1/N$ is optimum in terms of variance when linear combinations are considered. It is then a consequence that such equal weight averaging reduces the variance in a monotonic fashion as the number of copies increases. One may wonder whether averaging always improves performance of a pseudo-marginal algorithm, especially in the light of Example 13 where we have showed that the variance is not a reliable criterion in this context. As we shall see, however, the answer to this question is positive, and a direct consequence of the convex order. In fact, we are able to prove this result in a slightly more general scenario where the copies are only assumed to be exchangeable.

We preface our result with some background. Assume $Z(1), Z(2), \ldots, Z(N)$ are exchangeable and nonnegative random variables of unit expectation and denote $Z := (Z(1), Z(2), \ldots, Z(N))$. We introduce the simplex $\mathcal{S}_N := \{\lambda := (\lambda(1), \lambda(2), \ldots, \lambda(N)) \in [0,1]^N : \sum_{i=1}^N \lambda(i) = 1\}$. We consider below convex combinations of the elements of Z in terms of weights in $\mathcal{S}_N$ and to that purpose will use for $a, b \in \mathbb{R}^N$ the notation $(a,b) := \sum_{i=1}^N a(i)b(i)$. We will also denote the components of any $a \in \mathbb{R}^{\mathbb{N}}$ in decreasing order as $\max_i a(i) = a[1] \geq a[2] \geq \cdots \geq a[N] = \min_i a(i)$. We introduce next the notions of Schur concavity and majorisation [34].

DEFINITION 27 (Majorisation and Schur-concavity). Suppose that $\lambda, \mu \in \mathbb{R}^N$:

(a) We say that μ majorises λ, denoted $\lambda \prec \mu$, if $\sum_{i=1}^k \lambda[i] \leq \sum_{i=1}^k \mu[i]$ for all $k = 1, \ldots, N$.

(b) A function $\phi : \mathbb{R}^N \to \mathbb{R}$ is said to be Schur concave if $\lambda \prec \mu$ implies $\phi(\lambda) \geq \phi(\mu)$, and ϕ is Schur convex if $\lambda \prec \mu$ implies $\phi(\lambda) \leq \phi(\mu)$.

We state next a well-known result which establishes that convex combinations of exchangeable random variables with majorised weights imply a convex order on convex linear combinations.

THEOREM 28. *For any $\lambda, \mu \in \mathcal{S}_N$ such that $\lambda \prec \mu$, we have*

$$(\lambda, Z) \leq_{\mathrm{cx}} (\mu, Z).$$

For a proof see, for example, [41], Theorem 3.A.35.

Let then $Z_x = (Z_x(1), Z_x(2), \ldots, Z_x(N))$ for any $x \in \mathsf{X}$ stand for an exchangeable vector as above, and let $\lambda, \mu \in \mathcal{S}_N$. Consider the weights $W_x^{(\lambda)} := (\lambda, Z_x)$

and $W_x^{(\mu)} := (\mu, Z_x)$, which are nonnegative and have unit expectation, and let $\tilde{P}_\lambda$ and $\tilde{P}_\mu$ denote the pseudo-marginal algorithms corresponding to $\{W_x^{(\lambda)}\}_{x\in\mathsf{X}}$ and $\{W_x^{(\mu)}\}_{x\in\mathsf{X}}$, respectively.

THEOREM 29. *Assume that $\lambda, \mu \in \mathcal{S}_N$ satisfy $\lambda \prec \mu$. Then, for any $x, y \in \mathsf{X}$ and any $f \in L^2(\mathsf{X}, \pi)$,*

$$\alpha_{xy}(\tilde{P}_\lambda) \geq \alpha_{xy}(\tilde{P}_\mu) \quad \textit{and} \quad \operatorname{var}(f, \tilde{P}_\lambda) \leq \operatorname{var}(f, \tilde{P}_\mu),$$

that is, the expected acceptance probability is Schur concave, while the asymptotic variance is Schur convex.

Furthermore, if π is not concentrated on points, $\operatorname{Gap}_R(\tilde{P}_\lambda) \geq \operatorname{Gap}_R(\tilde{P}_\mu)$, that is, the right spectral gap is Schur concave.

PROOF. The proof follows directly from Theorems 10 and 28. □

REMARK 30. It is clear that Theorem 29 can be generalised to incorporate state dependent weights, $\lambda = \{\lambda_x\}_{x\in\mathsf{X}}$ and $\mu = \{\mu_x\}_{x\in\mathsf{X}}$ where $\lambda_x, \mu_x \in \mathcal{S}_N$ and use $W_x^{(\lambda)} := (\lambda_x, Z_x)$ and $W_x^{(\mu)} := (\mu_x, Z_x)$. The result also generalises to infinite exchangeable sequences $Z_x = (Z_x(1), Z_x(2), \ldots)$ and $\lambda_x, \mu_x \in \mathcal{S}_\infty := \{\lambda \in [0,1]^\infty : \sum_{i=1}^\infty \lambda(i) = 1\}$ and letting $W_x^{(\lambda)} := (\lambda_x, Z_x) := \sum_{i=1}^\infty \lambda_x(i) Z_x(i)$.

For any $k \in \{1, \ldots, N\}$ we define $u_k \in \mathcal{S}_N$ as the uniform weights $u_k := (1/k, \ldots, 1/k, 0, \ldots, 0)$, that is, the first k components are nonzero and are all equal. The next result shows that the optimal weighting of N estimators is the uniform weighting, and that every extra sample improves performance.

COROLLARY 31. *For any $\lambda \in \mathcal{S}_N$ all $x, y \in \mathsf{X}$ and $f \in L^2(\mathsf{X}, \pi)$,*

$$\alpha_{xy}(\tilde{P}_{u_N}) \geq \alpha_{xy}(\tilde{P}_\lambda) \quad \textit{and} \quad \operatorname{var}(f, \tilde{P}_{u_N}) \leq \operatorname{var}(f, \tilde{P}_\lambda),$$

and the following functions from $\{1, \ldots, N\}$ to $\mathbb{R}_+$ satisfy: $k \mapsto \alpha_{xy}(\tilde{P}_{u_k})$ is nondecreasing and $k \mapsto \operatorname{var}(g, \tilde{P}_{u_k})$ is nonincreasing.

Furthermore, if π is not concentrated on points, $\operatorname{Gap}_R(\tilde{P}_{u_N}) \geq \operatorname{Gap}(\tilde{P}_\lambda)$ and $k \mapsto \operatorname{Gap}_R(\tilde{P}_{u_k})$ is nondecreasing.

PROOF. In proof follows from Theorem 29 by observing that $u_N \prec \lambda$ and that $u_k \prec u_{k-1}$. □

REMARK 32. Note that the convex order $(u_k, Z) \leq_{\mathrm{cx}} (u_{k-1}, Z)$ can be obtained directly by applying [35], Corollary 1.5.24, which is related to the existence of reverse martingales for U-statistics, but our result is slightly stronger.

The monotonicity result in Corollary 31 provides us with some justification for averaging. Because we do not quantify the benefit of increased averaging, we cannot provide guidelines what the optimal number of samples. Instead, we point an interested reader to the recent related work of Sherlock, Thiery, Roberts and Rosenthal [42] and Doucet, Pitt, Deligiannidis and Kohn [15] who give conditions for optimal acceptance rates in some contexts. We note that when parallel architectures are available, averaging may be cheap or free, but our result also form the basis for the justification of validity of adaptive MCMC algorithms which seek for an optimal number of samples in the spirit of those proposed in a different context [2, 18]. For example, it is possible to consider algorithms which increase or decrease the number of samples according to some rule, aiming to reach a predefined average acceptance rate.

6.2. *Stratification.* In the context of Monte Carlo methods, stratification is a technique which aims to reduce variance of estimators of expectations. It turns out that stratification can also imply improved performance in terms of convex order. We refer the reader to very recent and important progress in this area [19, 20], but we start here with a more specific and immediately applicable result.

Approximate Bayesian computation (ABC) [6, 45] are now popular methods which are applicable in Bayesian inference involving intractable (or expensive to evaluate) likelihood function, but where simulation from the model is easy. Consider some data $y^* \in \mathsf{Y}$ and assume that it arises from a family of probability distributions with densities $\{\ell(\cdot|x), x \in \mathsf{X}\}$, with respect to some appropriate reference measure λ. Instead of the exact likelihood $\ell(y^*|x)$, an approximate likelihood function is constructed. Assume $s : \mathsf{Y}^2 \to \mathbb{R}_+$ is a function whose role is to measure dissimilarity between datasets, and consider for some $\varepsilon > 0$ the modified ABC likelihood

$$\ell_{\mathrm{ABC}}(y^*|x) := \int \ell(y|x)\mathbb{I}\{s(y, y^*) \le \varepsilon\}\lambda(\mathrm{d}y).$$

This alternative likelihood function is in general intractable, but naturally lends itself to pseudo-marginal computations [3].

Indeed, for any $x \in \mathsf{X}$ assume $Y_1, Y_2, \ldots, Y_N \sim \ell(y|x)\lambda(\mathrm{d}y)$ are independent samples. Then one can construct a nonnegative and unbiased estimator T_x of $\ell_{\mathrm{ABC}}(y^*|x)$ as follows:

$$T_x := \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{s(Y_i, y^*) \le \varepsilon\}. \tag{8}$$

This leads to the unit expectation estimator $W_x = T_x/\ell_{\mathrm{ABC}}(y^*|x)$. In practice, simulation of the random variables Y on a computer often involves using d (pseudo-) random numbers uniformly distributed on the unit interval $[0, 1]$, which are then mapped to form one Y_i. That is, there is a mapping from the unit cube $[0, 1]^d$ to Y, and with an inconsequential abuse of notation, if $U_i \sim \mathcal{U}([0, 1]^d)$ then $Y(U_i) \sim \ell(y|x)\lambda(\mathrm{d}y)$.

EXAMPLE 33. An extremely simple illustration of this is the situation where $d = 1$ and an inverse c.d.f. method is used, that is $Y(U) = F^{-1}(U)$ where F is the cumulative distribution function (c.d.f.) of Y. For example, in the case of the g-and-k distribution the inverse c.d.f. is given by [17]

$$F^{-1}(u; A, B, c, g, k) = A + B\left(1 + c\frac{1 - \exp(-gz(u))}{1 + \exp(-gz(u))}\right)\left(1 + z(u)^2\right)^k z(u),$$

where $z(u)$ is the standard normal quantile, and A, B, c, g, k are parameters.

In this context, an easily implementable method to improve performance of the corresponding pseudo marginal algorithm is as follows. Let $\mathcal{A} := \{A_1, \ldots, A_N\}$ be a partition of the unit cube $[0, 1]^d$ such that $\mathbb{P}(U_1 \in A_i) = 1/N$, and such that it is possible to sample uniformly from each A_i. Perhaps the simplest example of this is when $\mathcal{A}$ corresponds to the dyadic sub-cubes of $[0, 1]^d$. Let $V_i \sim \mathcal{U}(A_i)$ for $i = 1, \ldots, N$ be independent. We may now replace the estimator in (8) with

$$T_x^{\text{strat}} := \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{s(Y(V_i), y^*) \leq \varepsilon\}. \tag{9}$$

It is straightforward to check that this is a nonnegative unbiased estimator of $\ell_{\text{ABC}}(y^*|x)$ which means that $W_x^{\text{strat}} := T_x^{\text{strat}}/\ell_{\text{ABC}}(y^*|x)$ has unit expectation as required. With $\tilde{P}$ the pseudo-marginal approximation corresponding to using $\{W_x\}_{x\in\mathsf{X}}$ and $\tilde{P}^{\text{strat}}$ the approximation corresponding to $\{W_x^{\text{strat}}\}_{x\in\mathsf{X}}$, we have the following result.

THEOREM 34. *For any $x \in \mathsf{X}$ we have $W_x^{\text{strat}} \leq_{\text{cx}} W_x$ and, therefore, for any $x, y \in \mathsf{X}$ and $f \in L^2(\mathsf{X}, \pi)$,*

$$\alpha_{xy}(\tilde{P}^{\text{strat}}) \geq \alpha_{xy}(\tilde{P}) \quad \textit{and} \quad \operatorname{var}(f, \tilde{P}^{\text{strat}}) \leq \operatorname{var}(f, \tilde{P}).$$

Furthermore, if π is not concentrated on points, $\operatorname{Gap}_R(\tilde{P}^{\text{strat}}) \geq \operatorname{Gap}_R(\tilde{P})$.

PROOF. Notice that $\mathbb{I}\{s(Y(U_i), y^*) \leq \varepsilon\}$ is a Bernoulli random variable of parameter $\bar{p} := \mathbb{P}(s(Y(U_i), y^*) \leq \varepsilon) = \mathbb{P}(U_i \in H)$ where $H := \{u \in [0, 1]^d : s(Y(u), y^*) \leq \varepsilon\}$. Similarly, let $q_i := \mathbb{P}(s(Y(V_i), y^*) \leq \varepsilon) = \mathbb{P}(V_i \in H) = \mathbb{P}(U_i \in H \cap A_i)/\mathbb{P}(U_i \in A_i)$. Note that $\bar{p} = \sum_{i=1}^N q_i/N$. Denoting $q = (q_1, \ldots, q_N)$ and $p = (\bar{p}, \ldots, \bar{p})$, we have the majorisation $p \prec q$ (see Definition 27 in Section 6.1). A well-known result ([21] and [28]) tells us that the sum of the corresponding Bernoulli random variables are convex ordered, that is for independent $W_1, \ldots, W_N \sim \mathcal{U}(0, 1)$,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{W_i \leq q_i\} \leq_{\text{cx}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{W_i \leq p_i\}.$$

The random variable on the left coincides in distribution with T_x^{strat} and the random variable on the right coincides with T_x. Consequently, $W_x^{\mathrm{strat}} = T_x^{\mathrm{strat}}/\ell_{\mathrm{ABC}}(y^*|x) \leq_{\mathrm{cx}} T_x/\ell_{\mathrm{ABC}}(y^*|x) = W_x$. $\square$

Naturally some stratification schemes are going to be better than others, and the majorisation characterisation provides us, in principle, with a criterion for comparisons.

REMARK 35. We note that in some contexts, stratification may also open the possibility for additional computational savings. First, using "early rejection" as suggested in [43] the values of the summands in (9) can be computed progressively until it is possible to decide whether the sample is accepted or rejected. Second, in certain scenarios it may be possible to deduce the values of some indicators in (9), before computing them, from realisations of others. For example, assume $d = 1$, $A_i = [(i-1)/N, i/N]$, $s(y, y^*) = |y - y^*|$, and suppose we use the inverse c.d.f. method. Then due to the monotonicity of the inverse c.d.f., if $\mathbb{I}\{s(Y(V_k), y^*) \leq \varepsilon\} = 1$ and $\mathbb{I}\{s(Y(V_{k+1}), y^*) \leq \varepsilon\} = 0$, then we know that necessarily $\mathbb{I}\{s(Y(V_i), y^*) \leq \varepsilon\} = 0$ for $i = k+2, \ldots, N$.

REMARK 36. It has also been suggested in the literature to replace the indicator function in the ABC likelihood with a more general "kernel" $K : \mathbb{R}_+ \to [0, 1]$ effectively leading, with $\psi = K \circ s$, to

$$\ell_{\mathrm{ABC}}^{\psi}(y^*|x) := \int \ell(y|x)\psi(y, y^*)\lambda(\mathrm{d}y).$$

In such a situation, it is still possible to use stratification, but now inter-related conditions on the stratification scheme and the mapping $u \mapsto \psi(Y(u), y^*)$ are needed. For example, in the scenario $d = 1$ and with a monotone partition, the mapping should be monotone, otherwise the sought convex order may not hold [19].

6.3. *Extremal properties*. In this section, we investigate upper and lower bounds on the performance of pseudo-marginal algorithms by establishing a, perhaps surprising, link to the actuarial science literature in terms of extremal moments and stop-loss functions [12, 22, 23]. More specifically, we consider unit expectation distributions Q_* and Q^* which are minimal and maximal in the convex orders, subject to some constraints. We focus particularly on two types of constraints: a support constraint or a variance constraint. Other constraints such as a kurtosis constraint or a modality constraint are possible, and an interested reader can consult [22].

The link to the actuarial science literature comes from the fact that convex order of distributions of random variables W and V with $\mathbb{E}W = \mathbb{E}V$ is determined by the order of the related stop-loss functions (or integrated survival functions) $\mathbb{E}[(W -$

$t)_+]$ and $\mathbb{E}[(V-t)_+]$; see Lemma 5. The stop-loss links directly with the expected acceptance probability of the algorithm in Example 8 through the identity

$$\begin{aligned}\min\{1, r(x,y)\varpi\} &= r(x,y)\varpi - \max\{0, r(x,y)\varpi - 1\}\\ &= r(x,y)\varpi - r(x,y)\big(\varpi - r^{-1}(x,y)\big)_+.\end{aligned}$$

In the case of the pseudo-marginal algorithm, the above identity with $\varpi = u/w$ provides a connection; see the proof of Theorem 22.

Let $\mathscr{P}$ be some subset of probability distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Well-researched questions about stop-losses involve determining extremal elements $Q \in \mathscr{P}$ maximising or minimising $\mathbb{E}_Q[(W-t)_+]$ for some or all $t \in \mathbb{R}$. We review some of these results particularly relevant to the present setup and apply them to our problem.

THEOREM 37. *Let $\mu \in \mathbb{R}$ and let $\mathscr{P}(\mu)$ stand for the probability distributions Q on $\mathbb{R}$ such that the random variable $W \sim Q$ has expectation $\mathbb{E}_Q[W] = \mu$. Then, for any $t \in \mathbb{R}$,*

$$\delta_\mu(\mathrm{d}w) = \operatorname*{arg\,min}_{Q \in \mathscr{P}(\mu)} \mathbb{E}_Q\big[(W-t)_+\big],$$

with minimum value $(\mu - t)_+$.

THEOREM 38. *Let $a, b, \mu \in \mathbb{R}$ with $a \le \mu \le b$ and let $\mathscr{P}(\mu, [a,b]) \subset \mathscr{P}(\mu)$ be the set of probability distributions Q on $[a,b]$, that is, satisfying $\mathbb{E}_Q[W] = \mu$ and $Q([a,b]) = 1$. Then for any $t \in \mathbb{R}$,*

$$\frac{b-\mu}{b-a}\delta_a(\mathrm{d}w) + \frac{\mu-a}{b-a}\delta_b(\mathrm{d}w) = \operatorname*{arg\,max}_{Q \in \mathscr{P}(\mu,[a,b])} \mathbb{E}_Q\big[(W-t)_+\big]$$

with maximum value $\frac{b-\mu}{b-a}(a-t)_+ + \frac{\mu-a}{b-a}(b-t)_+$.

The proofs of Theorems 37 and 38 can be found in [12, 22, 23].

We state two direct consequences of these results.

THEOREM 39. *Let $a_x, b_x \in \mathbb{R}_+$ be such that $a_x \le 1 \le b_x$ for all $x \in \mathsf{X}$. Consider the class of pseudo-marginal algorithms $\tilde{P}$ such that for any $x \in \mathsf{X}$ the weight distribution Q_x is concentrated on $[a_x, b_x]$, that is, $Q_x \in \mathscr{P}(1, [a_x, b_x])$. Then, for any $f \in L^2(\mathsf{X}, \pi)$,*

$$\operatorname{var}(f, P) \le \operatorname{var}(f, \tilde{P}) \le \operatorname{var}(f, \tilde{P}_{\max}),$$

where $\tilde{P}_{\max}$ is the pseudo-marginal algorithm with noise distributions

$$Q_x^{\max}(\mathrm{d}w) = \frac{1-a_x}{b_x-a_x}\delta_{a_x}(\mathrm{d}w) + \frac{b_x-1}{b_x-a_x}\delta_{b_x}(\mathrm{d}w).$$

Furthermore,

$$\operatorname{var}(f, \tilde{P}_{\max}) \leq \sup_{x \in \mathsf{X}} b_x \operatorname{var}(f, P) + \Big(\sup_{x \in \mathsf{X}} b_x - 1\Big) \operatorname{var}_\pi(f).$$

PROOF. The first claim is direct from Theorems 37 and 38, Lemma 5 and Lemma 10. The last claim follows from [4], Corollary 11. □

We next state for completeness a similar result for algorithms $\mathring{P}$ as discussed in Section 2. In particular, it is direct to check that the diatomic distributions in Example 8 of the form, with $a_{xy} = a_{yx} \geq 1$,

$$Q_{xy}(\mathrm{d}\varpi) = \frac{a_{xy}}{1 + a_{xy}} \delta_{a_{xy}^{-1}}(\mathrm{d}\varpi) + \frac{1}{1 + a_{xy}} \delta_{a_{xy}}(\mathrm{d}\varpi),$$

are maximal among those with support on $[a_{xy}^{-1}, a_{xy}]$. We quote the result without a proof, as it is a direct consequence of the convex order property and Peskun's result.

THEOREM 40. *Let $a_{xy} \in [1, \infty)$ be constants such that $a_{xy} = a_{yx}$ for all $x, y \in \mathsf{X}$. Consider any algorithm $\mathring{P}$ as in Section 2 such that $Q_{xy} \in \mathscr{P}(\mu = 1, [a_{xy}^{-1}, a_{xy}])$ for all $x, y \in \mathsf{X}$. Then, for any $x, y \in \mathsf{X}^2$,*

$$\begin{aligned} &\frac{a_{xy}}{1 + a_{xy}} \min\{1, r(x, y) a_{xy}^{-1}\} + \frac{1}{1 + a_{xy}} \min\{1, r(x, y) a_{xy}\} \\ &\quad \leq \int Q_{xy}(\mathrm{d}\varpi) \min\{1, r(x, y)\varpi\} \\ &\quad \leq \min\{1, r(x, y)\}, \end{aligned}$$

and for any $f \in L^2(\mathsf{X}, \pi)$,

$$\begin{aligned} \operatorname{var}(f, P) &\leq \operatorname{var}(f, \mathring{P}) \\ &\leq \operatorname{var}(f, \mathring{P}_{\max}) \\ &\leq \sup_{x, y \in \mathsf{X}^2} a_{xy} \operatorname{var}(f, P) + \Big(\sup_{x, y \in \mathsf{X}^2} a_{xy} - 1\Big) \operatorname{var}_\pi(f). \end{aligned}$$

We now turn back to pseudo-marginal algorithms. Not surprisingly, it is impossible to find a maximal distribution on $\mathscr{P}(\mu, [a, b])$ whenever either $a = -\infty$ or $b = \infty$. However, as we shall see, with an additional constraint on the variance $\sigma^2 < \infty$ of the distributions, it is possible to find a supremal distribution even when $b = \infty$. More specifically, the stop-loss function can be maximised, but the corresponding class of distributions is not closed and maximising distribution will not have a finite variance. We first state the following results which can be found in [12, 22, 23].

TABLE 1
Characterisation of the maxima of (10)

$\mathbb{E}_{Q^*}[(W-t)_+]$	**Atoms of Q^***	**Range of t**
$\frac{1}{2}(\mu - t + \sigma_\mu(t))$	$t - \sigma_\mu(t), t + \sigma_\mu(t)$	$t \le c, \sigma_\mu(t) \le t - a$
$(\mu - a)\frac{(\mu-t)(\mu-a)+\sigma^2}{(\mu-a)^2+\sigma^2}$	$a, \mu + \frac{\sigma^2}{\mu-a}$	$t \le c, \sigma_\mu(t) \ge t - a$
$\frac{1}{2}(\mu - t + \sigma_\mu(t))$	$t - \sigma_\mu(t), t + \sigma_\mu(t)$	$t \ge c, \sigma_\mu(t) \le b - t$
$\frac{(b-t)\sigma^2}{(\mu-b)^2+\sigma^2}$	$\mu - \frac{\sigma^2}{b-\mu}, b$	$t \ge c, \sigma_\mu(t) \ge b - t$

THEOREM 41. *Let $a, b, \mu \in \mathbb{R}$ such that $a \le \mu \le b$ and let $\sigma^2 \in [0, (\mu - a)(b - \mu)]$. Define $\mathscr{P}(\mu, \sigma^2, [a, b]) \subset \mathscr{P}(\mu, [a, b])$ be the set of probability distributions Q such that $\mathbb{E}_Q[W] = \mu$, $\operatorname{var}_Q(W) = \sigma^2$ and $Q([a, b]) = 1$. Denote $\sigma_\mu^2(t) := \sigma^2 + (\mu - t)^2$ and $c := (a + b)/2$. Then the maximisation problem*

$$Q^* := \operatorname*{arg\,max}_{Q \in \mathscr{P}(\mu, \sigma^2, [a,b])} \mathbb{E}_Q\big[(W - t)_+\big], \tag{10}$$

has diatomic solutions Q^* for different values of t, which are given in Table 1.

The following result is a restatement of [35], Theorem 1.5.10(b), and it gives us a way to extend Theorem 41 to unbounded supports.

THEOREM 42. *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is nonincreasing and convex, and satisfies $\lim_{t\to\infty} \phi(t) = 0$ and $\lim_{t\to-\infty} \phi(t) = \mu \in \mathbb{R}$. Then there exists a random variable X with $\mathbb{E}[X] = \mu$ such that $\phi(t) = \mathbb{E}[(X - t)_+]$, and the c.d.f. of X can be written as $F_X(t) = 1 + \phi'(t)$, where ϕ' stands for the right derivative of ϕ.*

EXAMPLE 43. In our case, we are interested in distributions on the positive real line and with unit mean, for which we notice that

$$\begin{aligned}\phi(t) &:= \lim_{b\to\infty} \sup_{\mathscr{P}(1,\sigma^2,[0,b])} \mathbb{E}\big((W - t)_+\big) \\ &= \begin{cases} \dfrac{(1-t)+\sigma^2}{1+\sigma^2}, & 0 \le t \le \dfrac{\sigma^2+1}{2}, \\ \dfrac{1}{2}\big(\sqrt{\sigma^2 + (1-t)^2} + 1 - t\big), & t \ge \dfrac{\sigma^2+1}{2}. \end{cases}\end{aligned}$$

It is straightforward to check that ϕ satisfies the conditions in Theorem 42, so the "supremal" distribution $Q^*_{\sigma^2}$ on $\mathscr{P}(1, \sigma^2, [0, \infty))$ has the following cumulative

distribution function:

$$Q^*_{\sigma^2}([0,t]) = \begin{cases} \dfrac{\sigma^2}{1+\sigma^2}, & 0 \le t < \dfrac{\sigma^2+1}{2}, \\ \dfrac{1}{2} + \dfrac{1}{2}\dfrac{t-1}{\sqrt{\sigma^2+(1-t)^2}}, & t \ge \dfrac{\sigma^2+1}{2}. \end{cases}$$

Note that $Q^*_{\sigma^2}$ does not belong to $\mathscr{P}(1, \sigma^2, [0,\infty))$ because it has infinite variance, but the expectation is one, which can be verified by noting that $\phi(0) = 1$.

7. Discussion and perspectives. In this paper, we have shown that the convex (partial) order of distributions is a natural and useful tool in order to compare various performance measures of competing implementations of exact approximations of the Metropolis–Hastings algorithms. As examples of applications of our theory, we have shown that it is possible to identify extremal behaviours of such algorithms under various distributional constraints. More importantly, from a practical point of view, we have shown that averaging of independent estimators improves performance monotonically. Even though averaging may not always be useful in the context of sequential implementations, it turns out to be particularly relevant when parallel architectures are available (e.g., [16]). Prompted by our theory and other recently established results, we have also proposed to use stratification in ABC MCMC and beyond, which has the advantage to provide better performance at no additional computational cost.

There are many other results from the stochastic ordering literature relevant to the present context we have not yet investigated. For example, introducing negative dependence when averaging two weights could further improve performance. This is a direct application of the result on positive and negative quadrant dependence of [13], Lemma 2. Other dependence orders could be exploited, such as the supermodular order [35] which can be used to characterise the (positive) dependence order of the components of random vectors of arbitrary length. In this scenario, the supermodular order

$$\bigl(W_1^{(1)}, W_2^{(1)}, \ldots, W_N^{(1)}\bigr) \le_{\mathrm{sm}} \bigl(W_1^{(2)}, W_2^{(2)}, \ldots, W_N^{(2)}\bigr),$$

implies the convex order $\sum_{i=1}^N W_i^{(1)} \le_{\mathrm{cx}} \sum_{i=1}^N W_i^{(2)}$; see, for example, the results in [41], Section 9.A.

We would like to point out here another promising and useful avenue of research related to the discussion of [1] by Lee and Holmes to which our current theory does not seem to apply directly. First, we notice that pseudo-marginal algorithms can be extended to the situation where we can define a joint distribution $Q_{xy}(\mathrm{d}w \times \mathrm{d}u)$ with marginals $Q_x(\mathrm{d}w)$ and $Q_y(\mathrm{d}u)$ and which satisfies the following symmetry condition for any $x, y \in \mathsf{X}$ and $A, B \in \mathcal{B}(\mathbb{R}_+)$, $Q_{xy}(A \times B) = Q_{yx}(B \times A)$. The proposal distribution used in this algorithm is the corresponding conditional distribution $Q_{xy}(\mathrm{d}u|w)$, which now depends on x, y and w, and the acceptance ratio

remains as in (2). Standard pseudo-marginal algorithms correspond to the choice $Q_{xy}(\mathrm{d}w \times \mathrm{d}u) = Q_x(\mathrm{d}w)Q_y(\mathrm{d}u)$. This formalism allows one to disentangle the dependence structure from the variability of the marginal distributions. One can easily establish that this results in a MH kernel with $\tilde{\pi}$ as invariant distribution.

Intuitively, inducing positive dependence should reduce the variability of the acceptance probability and, therefore, lead to better performance—this is the motivation behind the work of Lee and Holmes. We note, however, that this is likely to reduce "mixing" on the noise component w. An order which seems suitable to rank such algorithms is the concordance order, also known as the correlation order; see [13], which coincides with the upper orthant, concordance and supermodular order in the bivariate scenario [35]. Using, for simplicity, our earlier notation for the present scenario, one can show that if for some $x, y \in \mathsf{X}$ $(W_{xy}^{(2)}, U_{xy}^{(2)}) \leq_{\mathrm{c}} (W_{xy}^{(1)}, U_{xy}^{(1)})$ then $\alpha_{xy}(\tilde{P}^{(2)}) \leq \alpha_{xy}(\tilde{P}^{(1)})$ and, therefore, $\mathcal{E}_{\tilde{P}^{(1)}}(g) \geq \mathcal{E}_{\tilde{P}^{(2)}}(g)$ for any $g \in L^2(\mathsf{X}, \pi)$. However, we do not know whether or not this implies $\mathrm{var}(f, \tilde{P}^{(1)}) \leq \mathrm{var}(f, \tilde{P}^{(2)})$ for $f \in L^2(\mathsf{X}, \pi)$.

APPENDIX A: PROOF OF LEMMA 16

It is easy to see that also A^{-1} is self-adjoint, and for any $g \in \mathcal{H}$,

$$0 \leq \langle A^{-1}f - g, A(A^{-1}f - g)\rangle = \langle f, A^{-1}f\rangle - 2\langle f, g\rangle + \langle g, Ag\rangle.$$

This implies (3) with inequality "$\geq$". We conclude by taking $g = A^{-1}f$.

APPENDIX B: PERTURBED METROPOLIS–HASTINGS ALGORITHMS

Assume π, q and r are as defined in Section 1, and assume $\mathring{\pi}_{x_1}(\mathrm{d}x_2)$ are probability distributions on $(\mathsf{X}_2, \mathcal{X}_2)$ for all $x_1 \in \mathsf{X}$, and denote with a slight abuse of notation the distribution $\mathring{\pi}(\mathrm{d}x_1 \times \mathrm{d}x_2) = \pi(\mathrm{d}x_1)\mathring{\pi}_{x_1}(\mathrm{d}x_2)$ on $(\mathring{\mathsf{X}}, \mathring{\mathcal{X}}) = (\mathsf{X} \times \mathsf{X}_2, \mathcal{X} \times \mathcal{X}_2)$. Assume that we are interested in approximating a Metropolis–Hastings algorithm targeting $\mathring{\pi}(\mathrm{d}x_1 \times \mathrm{d}x_2)$ using a family of proposal distributions of the following form, with $x := (x_1, x_2) \in \mathring{\mathsf{X}}$:

$$\mathring{q}(x, \mathrm{d}y_1 \times \mathrm{d}y_2) := q(x_1, \mathrm{d}y_1)\mathring{\pi}_{y_1}(\mathrm{d}y_2).$$

This covers, for example, the algorithm used in [27]. Note that this algorithm is also of the type discussed in Appendix C.

LEMMA 44. *Let* $\{Q_{xy_1}(\mathrm{d}\varpi \times \mathrm{d}y_2)\}_{(x,y_1)\in\mathring{\mathsf{X}}\times\mathsf{X}}$ *be a family of probability distributions on* $((0,\infty) \times \mathsf{X}_2, \mathcal{B}((0,\infty)) \times \mathcal{X}_2)$. *Consider the Markov transition kernel* $\mathring{P}$ *on* $(\mathring{\mathsf{X}}, \mathring{\mathcal{X}})$ *defined through*

$$\mathring{P}(x; \mathrm{d}y_1 \times \mathrm{d}y_2) := q(x_1, \mathrm{d}y_1) \int Q_{xy_1}(\mathrm{d}\varpi \times \mathrm{d}y_2) \min\{1, r(x_1, y_1)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}(x),$$

where the probability of rejection $\mathring{\rho}(x) \in [0,1]$ *is such that* $\mathring{P}(x,\cdot)$ *defines a probability distribution on* $(\mathring{\mathsf{X}}, \mathring{\mathcal{X}})$ *for all* $x \in \mathring{\mathsf{X}}$. *Assume further that for any* $x_1, y_1 \in \mathsf{X}$ *and any* $A, B \in \mathcal{X}_2$ *and* $C \in \mathcal{B}((0,\infty))$,

$$\int \mathring{\pi}_{x_1}(\mathrm{d}x_2) Q_{xy_1}(\mathrm{d}\varpi \times \mathrm{d}y_2) \varpi \mathbb{I}\left\{x_2 \in A, y_2 \in B, \frac{1}{\varpi} \in C\right\}$$
$$= \int \mathring{\pi}_{y_1}(\mathrm{d}y_2) Q_{yx_1}(\mathrm{d}\varpi \times \mathrm{d}x_2) \mathbb{I}\{x_2 \in A, y_2 \in B, \varpi \in C\}.$$

Then $\mathring{P}$ *is reversible with respect to* $\mathring{\pi}$.

PROOF. Let $A, B \in \mathring{\mathcal{X}}$, then we may write, restricting the integrals below to the set $\{r(x_1,y_1) > 0\}$ (see [46]),

$$\int \mathring{\pi}(\mathrm{d}x) q(x_1, \mathrm{d}y_1) \int Q_{xy_1}(\mathrm{d}\varpi \times \mathrm{d}y_2) \min\{1, r(x_1,y_1)\varpi\} \mathbb{I}\{x \in A, y \in B\}$$
$$= \int \pi(\mathrm{d}x_1) q(x_1, \mathrm{d}y_1) r(x_1,y_1) \int \mathring{\pi}_{x_1}(\mathrm{d}x_2) Q_{xy_1}(\mathrm{d}\varpi \times \mathrm{d}y_2) \varpi$$
$$\times \min\left\{\frac{r(y_1,x_1)}{\varpi}, 1\right\} \mathbb{I}\{x_2 \in A_{x_1}, y_2 \in B_{y_1}\}$$
$$= \int \pi(\mathrm{d}y_1) q(y_1, \mathrm{d}x_1) \int \mathring{\pi}_{y_1}(\mathrm{d}y_2) Q_{yx_1}(\mathrm{d}\varpi \times \mathrm{d}x_2) \min\{1, r(y_1,x_1)\varpi\}$$
$$\times \mathbb{I}\{x_2 \in A_{x_1}, y_2 \in B_{y_1}\},$$

where $A_{x_1} := \{x_2 \in \mathsf{X}_2 : (x_1,x_2) \in A\}$ and $B_{y_1} := \{y_2 \in \mathsf{X}_2 : (y_1,y_2) \in B\}$. We can now conclude. □

APPENDIX C: SPECTRAL GAPS OF AUGMENTED KERNELS

Recall that the left spectral gap of a μ-reversible Markov kernel Π can be defined as

$$\mathrm{Gap}_L(\Pi) := \inf_{\|f\|_\mu = 1} \langle f, (I + \Pi_\nu) f \rangle_\mu,$$

and the absolute spectral gap is defined as $\mathrm{Gap}(\Pi) := \min\{\mathrm{Gap}_L(\Pi), \mathrm{Gap}_R(\Pi)\} \in [0,1]$.

LEMMA 45. *Assume* Π *is a Markov kernel on a measurable space* $(\mathsf{E}, \mathcal{F})$ *and reversible with respect to a probability measure* μ. *Suppose* Π *has the form*

$$\Pi(x, \mathrm{d}y) = p(x, \mathrm{d}y) + \delta_x(\mathrm{d}y) r(x),$$

where $p(x, \mathrm{d}y) \geq 0$ *is a sub-probability kernel and* $r(x) \in [0, 1]$ *for all* $x \in \mathsf{E}$. *Assume* $\nu(x, A)$ *is a probability kernel from* $(\mathsf{E}, \mathcal{F})$ *to another measurable space* $(\mathsf{S}, \mathcal{S})$, *and define the Markov kernel*

$$\Pi_\nu(x, w; \mathrm{d}y \times \mathrm{d}u) := p(x, \mathrm{d}y)\nu(y, \mathrm{d}u) + \delta_{x,w}(\mathrm{d}y \times \mathrm{d}u)r(x).$$

Then, denoting $r^* := \mu - \operatorname{ess\,sup}_x r(x)$ *and* $r_* := \mu - \operatorname{ess\,inf}_x r(x)$:

(a) Π_ν *is reversible with respect to* $\mu_\nu(\mathrm{d}x \times \mathrm{d}w) = \mu(\mathrm{d}x)\nu(x, \mathrm{d}w)$,
(b) $\mathcal{E}_\Pi(f) = \mathcal{E}_{\Pi_\nu}(f)$ *for all* $f \in L^2(\mathsf{E}, \mu)$, *with* $f(x, w) = f(x)$,
(c) $\min\{\mathrm{Gap}_R(\Pi), 1 - r^*\} \leq \mathrm{Gap}_R(\Pi_\nu) \leq \mathrm{Gap}_R(\Pi)$,
(d) $\min\{\mathrm{Gap}_L(\Pi), 1 + r_*\} \leq \mathrm{Gap}_L(\Pi_\nu) \leq \mathrm{Gap}_L(\Pi)$.

PROOF. The reversibility (a) follows from (cf. [46])

$$\mu_\nu(\mathrm{d}x \times \mathrm{d}w)p(x, \mathrm{d}y)\nu(y, \mathrm{d}u) = \mu_\nu(\mathrm{d}y \times \mathrm{d}u)p(y, \mathrm{d}x)\nu(x, \mathrm{d}w).$$

Consider then a function $f \in L^2(\mathsf{E} \times \mathsf{S}, \mu_\nu)$. We write any such f as $f = \bar{f} + f_0$ where $f_0(x, w) = f_0(x) := \int \nu(x, \mathrm{d}w) f(x, w)$ and $\bar{f} = f - f_0$. It is straightforward to check that $\Pi_\nu f_0(x, w) = \Pi f_0(x)$ and that $\Pi_\nu f = \Pi f_0 + r\bar{f}$, implying $\Pi_\nu \bar{f}(x, w) = r\bar{f}$. These allow us to write

$$\begin{aligned} \langle f, \Pi_\nu f\rangle_{\mu_\nu} &= \langle f_0, \Pi_\nu f_0\rangle_{\mu_\nu} + \langle \bar{f}, \Pi_\nu \bar{f}\rangle_{\mu_\nu} + 2\langle f_0, \Pi_\nu \bar{f}\rangle_{\mu_\nu} \\ &= \langle f_0, \Pi f_0\rangle_\mu + \langle \bar{f}, r\bar{f}\rangle_{\mu_\nu}. \end{aligned} \tag{11}$$

For f constant in the second variable, we have $f_0 = f$ and $\bar{f} = 0$ and $\|f_0\|_\mu = \|f\|_{\mu_\nu}$, implying (b).

For the spectral gap bounds, assume $\|f\|_{\mu_\nu} = 1$ and note that $1 = \|f\|^2_{\mu_\nu} = \|f_0\|^2_{\mu_\nu} + \|\bar{f}\|^2_{\mu_\nu}$. This, with (11), allows us to deduce

$$\begin{aligned} \mathrm{Gap}_R(\Pi_\nu) &= \inf_{\|f\|_{\mu_\nu}=1} \big\langle f, (I - \Pi_\nu) f\big\rangle_{\mu_\nu} \\ &= \inf_{\|f\|_{\mu_\nu}=1} 1 - \langle f, \Pi_\nu f\rangle_{\mu_\nu} \\ &= \inf_{\|f\|_{\mu_\nu}=1} \big\langle f_0, (I - \Pi) f_0\big\rangle_\mu + \big\langle \bar{f}, (1 - r)\bar{f}\big\rangle_{\mu_\nu}, \end{aligned}$$

where the inf and sup are taken over functions $f \in L^2_0(\mathsf{E} \times \mathsf{S}, \mu_\nu)$. The right-hand side inequality in (c) follows by restricting to functions constant in the second variable. For the first inequality, note that $\langle \bar{f}, (1 - r)\bar{f}\rangle_{\mu_\nu} \geq (1 - r^*)\|\bar{f}\|^2_{\mu_\nu}$ and $\langle f_0, (I - \Pi) f_0\rangle_\mu \geq \mathrm{Gap}(\Pi)\|f_0\|^2_\mu$. The claim follows because $\|f_0\|^2_{\mu_\nu} + \|\bar{f}\|^2_{\mu_\nu} = 1$.

Similarly, for the left gap (d),

$$\begin{aligned} \mathrm{Gap}_L(\Pi_\nu) &= \inf_{\|f\|_{\mu_\nu}=1} \big\langle f, (I + \Pi_\nu) f\big\rangle_{\mu_\nu} \\ &= \inf_{\|f\|_{\mu_\nu}=1} \big\langle f_0, (I + \Pi) f_0\big\rangle_\mu + \big\langle \bar{f}, (1 + r)\bar{f}\big\rangle_{\mu_\nu}, \end{aligned}$$

and $\langle \bar{f}, (1 + r)\bar{f}\rangle_{\mu_\nu} \geq (1 + r_*)\|\bar{f}\|^2_{\mu_\nu}$. □

REMARK 46. In the following, special scenarios some of the conclusions of Lemma 45 take a simpler form:

(a) If μ is not concentrated on points, that is, $\mu(\{x\}) = 0$ for all $x \in \mathsf{E}$, then $\mathrm{Gap}_R(\Pi) \leq 1 - r^*$ and, therefore, $\mathrm{Gap}_R(\Pi_\nu) = \mathrm{Gap}_R(\Pi)$.

(b) If Π is positive, that is, $\langle g, \Pi g \rangle_\mu \geq 0$ for all $g \in L^2(\mathsf{E}, \mu)$, then Π_ν is positive, and consequently $\mathrm{Gap}(\Pi) = \mathrm{Gap}_R(\Pi)$ and $\mathrm{Gap}(\Pi_\nu) = \mathrm{Gap}_R(\Pi_\nu)$.

Item (a) is a restatement of [4], Theorem 54, and (b) follows because positivity of Π is equivalent to $\mathrm{Gap}_L(\Pi) \geq 1$, implying $\mathrm{Gap}_L(\Pi_\nu) \geq 1$.

Acknowledgement. The authors would like to thank Lasse Leskelä for illuminating conversations about convex orders and peacock processes.

REFERENCES

[1] ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 269–342. MR2758115

[2] ANDRIEU, C. and ROBERT, C. P. (2001). Controlled MCMC for optimal sampling. Technical Report No. Ceremade 0125, Univ. Paris Dauphine.

[3] ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. MR2502648

[4] ANDRIEU, C. and VIHOLA, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.* **25** 1030–1077. MR3313762

[5] BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.

[6] BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

[7] BELLMAN, R. (1968). Some inequalities for the square root of a positive definite matrix. *Linear Algebra Appl.* **1** 321–324. MR0234977

[8] BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G. O. and FEARNHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 333–382. MR2278331

[9] BORNN, L., PILLAI, N., SMITH, A. and WOODARD, D. (2015). The use of a single pseudo-sample in approximate Bayesian computation. Preprint. Available at arXiv:1404.6298v4.

[10] CARACCIOLO, S., PELISSETTO, A. and SOKAL, A. D. (1990). Nonlocal Monte Carlo algorithm for self-avoiding walks with fixed endpoints. *J. Stat. Phys.* **60** 1–53. MR1063212

[11] CEPERLEY, D. M. and DEWING, M. (1999). The penalty method for random walks with uncertain energies. *J. Chem. Phys.* **110** 9812.

[12] DE VYLDER, F. and GOOVAERTS, M. J. (1982). Analytical best upper bounds on stop-loss premiums. *Insurance Math. Econom.* **1** 197–211. MR0665626

[13] DHAENE, J. and GOOVAERTS, M. J. (1996). Dependency of risks and stop-loss order. *Astin Bull.* **26** 201–212.

[14] DOUC, R., FORT, G., MOULINES, E. and SOULIER, P. (2004). Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab.* **14** 1353–1377. MR2071426

[15] DOUCET, A., PITT, M. K., DELIGIANNIDIS, G. and KOHN, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102** 295–313. MR3371005

[16] DROVANDI, C. C. (2014). Pseudo-marginal algorithms with multiple CPUs. Working Paper No. 61505. Queensland Univ. Technology, Brisbane.

[17] FEARNHEAD, P. and PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 419–474. MR2925370

[18] GILKS, W. R., ROBERTS, G. O. and SAHU, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93** 1045–1054. MR1649199

[19] GOLDSTEIN, L., RINOTT, Y. and SCARSINI, M. (2011). Stochastic comparisons of stratified sampling techniques for some Monte Carlo estimators. *Bernoulli* **17** 592–608. MR2787606

[20] GOLDSTEIN, L., RINOTT, Y. and SCARSINI, M. (2012). Stochastic comparisons of symmetric sampling designs. *Methodol. Comput. Appl. Probab.* **14** 407–420. MR2966302

[21] HOEFFDING, W. (1956). On the distribution of the number of successes in independent trials. *Ann. Math. Statist.* **27** 713–721. MR0080391

[22] HÜRLIMANN, W. (2008). Extremal moment methods and stochastic orders: Application in actuarial science. Chapters IV, V and VI. *Bol. Asoc. Mat. Venez.* **15** 153–301. MR2566879

[23] JANSSENS, G. K. and RAMAEKERS, K. M. (2008). On the use of bounds on the stop-loss premium for an inventory management decision problem. *J. Interdiscip. Math.* **11** 115–126. MR2406170

[24] JARNER, S. F. and HANSEN, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.* **85** 341–361. MR1731030

[25] JARNER, S. F. and ROBERTS, G. O. (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Probab.* **12** 224–247. MR1890063

[26] JONES, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surv.* **1** 299–320. MR2068475

[27] KARAGIANNIS, G. and ANDRIEU, C. (2013). Annealed importance sampling reversible jump MCMC algorithms. *J. Comput. Graph. Statist.* **22** 623–648. MR3173734

[28] KARLIN, S. and NOVIKOFF, A. (1963). Generalized convex inequalities. *Pacific J. Math.* **13** 1251–1279. MR0156927

[29] ŁATUSZYŃSKI, K., MIASOJEDOW, B. and NIEMIRO, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli* **19** 2033–2066. MR3129043

[30] LEISEN, F. and MIRA, A. (2008). An extension of Peskun and Tierney orderings to continuous time Markov chains. *Statist. Sinica* **18** 1641–1651. MR2469328

[31] LESKELÄ, L. and VIHOLA, M. (2014). Conditional convex orders and measurable martingale couplings. Preprint. Available at arXiv:1404.0999v3.

[32] MAIRE, F., DOUC, R. and OLSSON, J. (2014). Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Ann. Statist.* **42** 1483–1510. MR3262458

[33] MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.

[34] MARSHALL, A. W., OLKIN, I. and ARNOLD, B. (2010). *Inequalities*: *Theory of Majorization and Its Applications*. Springer, Berlin.

[35] MÜLLER, A. and STOYAN, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester. MR1889865

[36] NICHOLLS, G. K., FOX, C. and WATT, A. M. (2012). Coupled MCMC with a randomized acceptance probability. Preprint. Available at arXiv:1205.6857v1.

[37] PESKUN, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60** 607–612. MR0362823

[38] ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25 (electronic). MR1448322

[39] ROBERTS, G. O. and TWEEDIE, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *J. Appl. Probab.* **38A** 37–41.

[40] RUDOLF, D. (2012). Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Math.* (*Rozprawy Mat.*) **485** 1–93. MR2977521

[41] SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. Springer, New York. MR2265633

[42] SHERLOCK, C., THIERY, A. H., ROBERTS, G. O. and ROSENTHAL, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43** 238–275. MR3285606

[43] SOLONEN, A., OLLINAHO, P., LAINE, M., HAARIO, H., TAMMINEN, J. and JÄRVINEN, H. (2012). Efficient MCMC for climate model parameter estimation: Parallel adaptive chains and early rejection. *Bayesian Anal.* **7** 715–736. MR2981633

[44] STRASSEN, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** 423–439. MR0177430

[45] TAVARE, S., BALDING, D. J., GRIFFITHS, R. C. and DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.

[46] TIERNEY, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9. MR1620401

[47] TUOMINEN, P. and TWEEDIE, R. L. (1994). Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. in Appl. Probab.* **26** 775–798. MR1285459

SCHOOL OF MATHEMATICS
UNIVERSITY WALK
BRISTOL, BS8 1TW
UNITED KINGDOM
E-MAIL: C.Andrieu@bristol.ac.uk

DEPARTMENT OF MATHEMATICS AND STATISTICS
P.O. BOX 35
FI-40014 UNIVERSITY OF JYVÄSKYLÄ
FINLAND
E-MAIL: matti.vihola@iki.fi