Bernoulli **30**(3), 2024, 2301–2325
https://doi.org/10.3150/23-BEJ1674

An asymptotic Peskun ordering and its application to lifted samplers

PHILIPPE GAGNON[a] and FLORIAN MAIRE[b]

Department of Mathematics and Statistics, Université de Montréal, Montréal, Canada,
[a]*philippe.gagnon.3@umontreal.ca*, [b]*florian.maire@umontreal.ca*

A Peskun ordering between two samplers, implying a dominance of one over the other, is known among the Markov chain Monte Carlo community for being a remarkably strong result. It is however also known for being a result that is notably difficult to establish. Indeed, one has to prove that the probability to reach a state $\mathbf{y}$ from a state $\mathbf{x}$, using a sampler, is greater than or equal to the probability using the other sampler, and this must hold for all pairs $(\mathbf{x},\mathbf{y})$ such that $\mathbf{x} \neq \mathbf{y}$. We provide in this paper a weaker version that does not require an inequality between the probabilities for all these states: essentially, the dominance holds asymptotically, as a varying parameter grows without bound, as long as the states for which the probabilities are greater than or equal to belong to a mass-concentrating set. The weak ordering turns out to be useful to compare *lifted* samplers for *partially-ordered* discrete state-spaces with their Metropolis–Hastings counterparts. An analysis in great generality yields a qualitative conclusion: they asymptotically perform better in certain situations (and we are able to identify them), but not necessarily in others (and the reasons why are made clear). A quantitative study in a specific context of graphical-model simulation is also conducted.

Keywords: Bayesian statistics; binary random variables; Ising model; Markov chain Monte Carlo methods; variable selection

1. Introduction

1.1. Peskun ordering: Context, original version and some variants

Let us consider the situation where one is interested in sampling from π, a probability distribution defined on a measurable space $(\mathcal{X},\mathbf{X})$, with $\mathcal{X}$ finite and assumed to correspond to the support of π, and $\mathbf{X}$ a sigma-algebra on $\mathcal{X}$. In a sampling context, π is often referred to as the *target distribution*. Let us consider that, to sample from π, one has access to two Markov chain Monte Carlo (MCMC) algorithms and wonders which one is best. Establishing a Peskun ordering (Peskun, 1973) is possibly the most sought-after route when one wants to prove that a given MCMC algorithm is superior in terms of statistical efficiency to another. The statistical efficiency is measured in terms of asymptotic variances: for any Markov kernel P acting on $(\mathcal{X},\mathbf{X})$ and for any $f : \mathcal{X} \to \mathbb{R}$, we denote by $\mathrm{var}(f,P)$ the asymptotic variance in a central limit theorem for a MCMC estimator of πf, the expectation of $f(\mathbf{X})$ under $\mathbf{X} \sim \pi$. In this paper, all considered Markov kernels are assumed to be irreducible and aperiodic, so that the associated samplers are valid[1]. The original ordering is presented in Theorem 1.

Theorem 1 (Peskun, 1973). *Let P_1 and P_2 be two Markov kernels that are reversible with respect to π. If $P_1(\mathbf{x},\mathbf{y}) \geq P_2(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y}) \in \mathcal{X}^2$ with $\mathbf{x} \neq \mathbf{y}$, then $\mathrm{var}(f,P_1) \leq \mathrm{var}(f,P_2)$ for all $f : \mathcal{X} \to \mathbb{R}$.*

[1]By valid, we mean that a law of large numbers and a central limit theorem hold for time-averages of functionals of Markov chains.

The strength of this result lies in its universality: the order between the asymptotic variances holds for *all* functions f, which explains why we say that a sampler associated with P_1 is superior to a sampler associated with P_2, for the problem at hand. This ordering is however known to be rather challenging to establish. It is indeed only in specific situations that one can establish that the probability to reach $\mathbf{y}$ from $\mathbf{x}$ with P_1 is greater than or equal to that with P_2, and this for all $(\mathbf{x},\mathbf{y}) \in \boldsymbol{\mathcal{X}}^2$ with $\mathbf{x} \neq \mathbf{y}$.

The result of Peskun (1973) was generalized in several ways. First, Tierney (1998) extended it to general state-spaces. Andrieu, Lee and Vihola (2018) then provided a quantitative form requiring that the order on the Markov kernels holds, but up to a multiplicative factor, that is $P_1(\mathbf{x},\mathbf{y}) \geq \omega\, P_2(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y}) \in \boldsymbol{\mathcal{X}}^2$ with $\mathbf{x} \neq \mathbf{y}$, for some $\omega > 0$, while yielding similar conclusions:

$$\operatorname{var}(f, P_1) \leq \frac{\operatorname{var}(f, P_2)}{\omega} + \frac{1-\omega}{\omega}\, \mathbb{V}\mathrm{ar}[f(\mathbf{X})].$$

These results are valid for reversible Markov chains only. Recently, Andrieu and Livingstone (2021) went beyond the reversible scenario. These authors consider a specific type of non-reversibility for which the chains can be seen as being "almost" reversible; they are reversible, up to an involution. This type of non-reversibility nevertheless covers a remarkably large number of known non-reversible MCMC algorithms, including *lifted* algorithms (Chen, Lovász and Pak, 1999, Diaconis, Holmes and Neal, 2000, Gustafson, 1998, Horowitz, 1991).

1.2. Our proposal: A weaker and asymptotic version

With a result as strong as the original ordering, it is somewhat expected to be difficult to establish it. The main result of this paper is that a weaker version of this ordering can lead to similar, but weaker, conclusions. This weaker ordering[2] is particularly well suited for situations where the two Markov chains of interest are well understood, but only on some subsets of the state-space. We believe that this weaker version will allow to compare samplers in situations in which it was not possible before. Indeed, we believe that the difficulty in establishing the original ordering comes from the verification of $P_1(\mathbf{x},\mathbf{y}) \geq P_2(\mathbf{x},\mathbf{y})$ for *all* $(\mathbf{x},\mathbf{y}) \in \boldsymbol{\mathcal{X}}^2$ with $\mathbf{x} \neq \mathbf{y}$.

Recent concepts such as *approximate spectral gaps* introduced in Atchadé (2021) and *large sets* proposed in Yang and Rosenthal (2023) have shown that bounds on the convergence time of Markov chains can be obtained by exploiting the particular behaviour of the process on some subset of the state-space. When the process is particularly efficient on such a subset, resulting bounds can be tighter than traditional ones that account for the whole state-space. We here leverage similar ideas.

Consider that an order on the probabilities $P_1(\mathbf{x},\mathbf{y}) \geq P_2(\mathbf{x},\mathbf{y})$ can be established but only on a subset $\tilde{\boldsymbol{\mathcal{X}}}^2 \subset \boldsymbol{\mathcal{X}}^2$. It is natural to expect that if the mass concentrates on $\tilde{\boldsymbol{\mathcal{X}}}$ and under some mixing guarantees (to guarantee that when the chains exit $\tilde{\boldsymbol{\mathcal{X}}}$ they do not take too long to come back), then $\operatorname{var}(f, P_1) \lessapprox \operatorname{var}(f, P_2)$ for a class of functions f, where the approximation is a consequence of working under a limiting regime to represent a phenomenon of mass concentration. In the following, we prove a result essentially corresponding to that just described. We now provide an overview of a motivating application which is explored in the manuscript.

[2]For brevity, we will use "weaker ordering" or "weak ordering" to refer to the proposed weaker version of Peskun's ordering. As will be seen, using such expressions is however an abuse of terminology because the binary relation defined by our "weak ordering" does not establish an order on the set of reversible Markov kernels in the mathematical sense.

1.3. Lifted samplers: A motivating application

Lifting the state-space is a generic technique which yields what are referred to as *lifted* samplers. The state-space is *lifted* (i.e., extended) to incorporate auxiliary variables. The idea is to think of the random variables we want to sample as *position* variables and to associate to them *direction* variables, to *guide* the Markov chains so as to avoid backtracking, a behaviour often exhibited by reversible schemes that is suspected to increase the autocorrelation of the process. Consider for instance that $\boldsymbol{\mathcal{X}} = \{1, \ldots, K\}$, where K is a positive integer. We associate to the variable $\mathbf{X}$ a direction variable $\nu \in \{-1, +1\}$. A Markov chain is defined on the lifted state-space $\boldsymbol{\mathcal{X}} \times \{-1, +1\}$. The lifted sampler proceeds as a Metropolis–Hastings (MH, (Hastings, 1970, Metropolis et al., 1953)) algorithm in the sense that a proposal is accepted with a given probability, but in this case the proposal is deterministic and given by $\mathbf{y} = \mathbf{x} + \nu$ when $(\mathbf{x}, \nu)$ is the current state. The randomness thus comes from the decision to accept or reject the proposal; in the latter case, the direction is reversed. The lifting technique can be seen as a way to equip the resulting stochastic process with some memory of its past (the direction it comes from), while retaining the Markov property. It can be shown that the resulting Markov chains admit $\pi \otimes \mathcal{U}\{-1, 1\}$ as invariant distribution, where $\mathcal{U}\{-1, 1\}$ denotes the uniform distribution over the set $\{-1, 1\}$ and $\pi \otimes \mathcal{U}\{-1, 1\}$ is the product measure. The sampler is thus valid and expectations under π can be approximated by considering functions $f : \boldsymbol{\mathcal{X}} \times \{-1, +1\} \to \mathbb{R}$ of solely the first argument.

Let P_{lifted} be the Markov kernel associated to this algorithm, and let P_{MH} be the Markov kernel associated to its non-lifted counterpart, which is a MH algorithm proposing $\mathbf{y} = \mathbf{x} + 1$ or $\mathbf{y} = \mathbf{x} - 1$, each with probability $1/2$. Theorem 7 in Andrieu and Livingstone (2021) allows to establish that $\text{var}(f, P_{\text{lifted}}) \leq \text{var}(f, P_{\text{MH}})$, for *any* f of solely the first argument and *any* distribution π. As Peskun's, this result is universal. It is however remarkable that it holds, not only for any f, but also for any π. It is also remarkable to obtain such a result given that the lifted sampler is implemented at no additional computational cost over its non-lifted counterpart, and also with no additional implementation difficulty (lifted samplers often possess these qualities). The result on the order between the asymptotic variances is essentially a consequence of having the same acceptance probabilities in both algorithms. There is thus no lost in terms of acceptance probabilities by using the lifting technique, while there is a potential gain in terms of persistent movement.

The superiority of P_{lifted} over P_{MH} for any π at no additional computational cost motivates an investigation of lifted samplers for other types of discrete state-spaces, especially given the limited number (or rather the absence) of real-world models where the state-space is of the form $\boldsymbol{\mathcal{X}} = \{1, \ldots, K\}$. This latter set is totally ordered; a natural first step in the investigation is thus to consider *partially-ordered* discrete state-spaces. A definition of partially-ordered sets as well as a generic lifted algorithm to sample from distributions defined on such a set are presented in Section 3. Important applications of such an algorithm include simulation of systems formed from binary variables, such as those simulated using the Ising model, and Bayesian variable selection when the posterior model probabilities can be evaluated, up to a normalizing constant.

In the case of partially-ordered discrete state-spaces, Theorem 7 of Andrieu and Livingstone (2021) still allows to prove the superiority of the lifted algorithm over its non-lifted counterpart, which is a reversible sampler; however in this case, the non-lifted counterpart does not correspond to the MH algorithm over which we wish to establish a superiority. This is essentially because, contrarily to the totally-ordered case, the acceptance probabilities in the MH and lifted algorithms are in general different. In certain situations, they can be quite unbalanced in some area of the state-space in the lifted algorithm, while they are not in the MH sampler. In contrast, in some other area of the state-space, the acceptance probabilities are similar. When the mass concentrates on the latter area, we explore the possibility of applying our weak ordering to compare the non-lifted counterpart and the MH algorithm to prove that the lifted sampler is superior to the MH algorithm.

1.4. Organization of the paper

We now describe how the rest of the paper is organized. We introduce our asymptotic Peskun ordering in Section 2. We next use this result to identify situations in which the lifted samplers for partially-ordered discrete state-spaces are expected to outperform (or not) their MH counterparts. Regarding the organization of this part, we first present the lifted samplers in Section 3 and then carry out in Section 4 an analysis in great generality. Given that the analysis is carried out in great generality, we are not in a position to verify the assumptions under which our asymptotic ordering holds. Rather, the analysis aims to establish the results that can be obtained whenever the assumptions are verified. We next conduct in Section 5 a thorough study in a context of a simulation of a simple graphical model. With this simple model, we are in a good position to verify the assumptions of our theoretical result; Section 5 serves as a user guide for applying our asymptotic Peskun ordering. The model corresponds to a Ising model with an external field, but without spatial correlation. The target distribution thus factorizes and the components of $\mathbf{x}$ are independent; the external field defines the marginal distributions. The model can be seen as an approximation to that with weak spatial correlation, referred to in the literature as a *high temperature* model. We will refer to the model studied in Section 5 as *the simple Ising model*. The manuscript finishes in Section 6 with retrospective comments and possible directions for future research. In Section 1 of the Supplementary Material (Gagnon and Maire (2024)), we study more complex problems for which a verification of the assumptions is beyond the scope of the manuscript. The first problem is about the simulation of a Ising model which is more complex than that considered in Section 5 (with spatial correlation). Numerical results are provided and they are consistent with the theoretical ones presented in Section 5. The second problem is that of variable selection in a real-life situation. All proofs of theoretical results are deferred to Section 3 of the Supplementary Material (Gagnon and Maire (2024)). While the paper is concerned with efficient sampling of distributions defined on discrete state-spaces, we stress that numerous results and elements of our analyses translate immediately to general state-space contexts.

2. A weaker and asymptotic version of Peskun's ordering

Before presenting the theoretical result, we provide the intuition behind it (while being more precise than in Section 1.2). This will help justify the assumptions, allow to highlight its relevance, and in fact allow to present a sketch of the proof. Beforehand, we introduce required notation.

In all this section, we consider that the distribution of interest π is parameterized by some $n \in \mathbb{N}$, that is $\pi \equiv \pi_n$. The state-space may also be parameterized by n and is thus denoted by $\boldsymbol{\mathcal{X}}_n$; we assume that, for each $n \in \mathbb{N}$, $\boldsymbol{\mathcal{X}}_n$ is finite. We define two collections of Markov kernels, $\{P_{1,n}\}$ and $\{P_{2,n}\}$, for which $P_{1,n}$ and $P_{2,n}$ are π_n-reversible for all n. We define a collection of subsets $\{\tilde{\boldsymbol{\mathcal{X}}}_n \subset \boldsymbol{\mathcal{X}}_n\}$ which we refer to as *control subsets*. We introduce two collections of restricted kernels $\{\tilde{P}_{1,n}\}$ and $\{\tilde{P}_{2,n}\}$ which, for all n, are defined for any $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}_n^2$ by

$$\tilde{P}_{i,n}(\mathbf{x},\mathbf{y}) := P_{i,n}(\mathbf{x},\mathbf{y}) + P_{i,n}(\mathbf{x},\boldsymbol{\mathcal{X}}_n \backslash \tilde{\boldsymbol{\mathcal{X}}}_n)\mathbb{1}_{\mathbf{y}=\mathbf{x}}, \qquad i \in \{1,2\}.$$

The form of states like $\mathbf{x}$ and $\mathbf{y}$ may depend on n, but we make this dependence implicit to simplify. We let $\tilde{\pi}_n$ be the probability measure defined as $\tilde{\pi}_n := \pi_n(\cdot \cap \tilde{\boldsymbol{\mathcal{X}}}_n)/\pi_n(\tilde{\boldsymbol{\mathcal{X}}}_n)$. It can be readily checked that $\tilde{P}_{1,n}$ and $\tilde{P}_{2,n}$ are both $\tilde{\pi}_n$-reversible, for all n. We define what we call (with some abuse of terminology) the *interior* and the *boundary* of $\tilde{\boldsymbol{\mathcal{X}}}_n$ as $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ := \{\mathbf{x} \in \tilde{\boldsymbol{\mathcal{X}}}_n : P_{i,n}(\mathbf{x},\tilde{\boldsymbol{\mathcal{X}}}_n^\complement) = 0\}$ and $\partial\tilde{\boldsymbol{\mathcal{X}}}_n := \tilde{\boldsymbol{\mathcal{X}}}_n \setminus \tilde{\boldsymbol{\mathcal{X}}}_n^\circ$, respectively, where we assume that the definition of $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$ is the same for $i = 1,2$. The functions for which we want to approximate the expectations may also depend on n and are thus denoted by f_n.

The π_n-weighted scalar product and p-norm are defined as $\langle f_n, g_n \rangle_{\pi_n} := \sum_{\mathbf{x}\in\boldsymbol{\mathcal{X}}} f_n(\mathbf{x})\, g_n(\mathbf{x})\, \pi_n(\mathbf{x})$ and $\|f_n\|_{\pi_n,p} := [\sum_{\mathbf{x}\in\boldsymbol{\mathcal{X}}} |f_n(\mathbf{x})|^p \pi_n(\mathbf{x})]^{1/p}$, respectively, with $\|f_n\|_{\pi_n}$ for the 2-norm. In this section, we consider that the functions are standardized, meaning that $f_n \in \mathcal{L}^2_{0,1}(\pi_n)$, where $\mathcal{L}^2_{0,1}(\pi_n) := \{f_n : \pi_n f_n = 0 \text{ and } \|f_n\|_{\pi_n} = 1\}$. This should not be seen as a restriction given that the magnitude of asymptotic variances, which is proportional to $\|f_n\|^2_{\pi_n}$, is irrelevant when it is of interest to establish an order among them. We note that since for each n, $\boldsymbol{\mathcal{X}}_n$ is finite, $P_{1,n}$ and $P_{2,n}$ admit a non-trivial *right spectral gap* in $\mathcal{L}^2_{0,1}(\pi_n)$, whose variational expression is given by

$$\lambda_i(n) := \inf_{f_n \in \mathcal{L}^2_{0,1}(\pi_n):\|f_n\|_{\pi_n}>0} \frac{\langle f_n, (I_n - P_{i,n}) f_n \rangle_{\pi_n}}{\|f\|^2_{\pi_n}}, \quad i \in \{1,2\}, \tag{1}$$

where I_n is the identity on $\mathcal{L}^2_{0,1}(\pi_n)$. In particular, it can be proved that $\lambda_i(n) \in (0,2)$. We analogously define the right spectral gaps of $\tilde{P}_{i,n}$ and denote them by $\tilde{\lambda}_i(n)$, $i = 1,2$, and we define $\underline{\lambda}(n) := \min\{\lambda_1(n), \lambda_2(n), \tilde{\lambda}_1(n), \tilde{\lambda}_2(n)\}$. In the following, we refer to the *right spectral gap* of a kernel simply as the *spectral gap* to simplify. Finally, we will use o for the little-o notation.

Consider that one wants to establish a Peskun-type ordering between two kernels, but one is only able to establish a (suitable) order on the kernels on a subset of the state-space in the following sense: $P_{1,n}(\mathbf{x},\mathbf{y}) \geq \omega(n) P_{2,n}(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}^2_n$ with $\mathbf{x} \neq \mathbf{y}$ where $\omega(n)$ is a (suitable) positive constant which may depend on n. This ordering implies that $\tilde{P}_{1,n}(\mathbf{x},\mathbf{y}) \geq \omega(n)\tilde{P}_{2,n}(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}^2_n$ with $\mathbf{x} \neq \mathbf{y}$, which in turn implies that

$$\mathrm{var}(f_n, \tilde{P}_{1,n}) \leq \frac{1}{\omega(n)} \mathrm{var}(f_n, \tilde{P}_{2,n}) + \frac{1}{\omega(n)} - 1, \tag{2}$$

by, as mentioned in Section 1.1, Andrieu, Lee and Vihola (2018) (Lemma 33).

Let us consider that π_n concentrates on $\tilde{\boldsymbol{\mathcal{X}}}^\circ_n$. The notion of concentration of π_n naturally implies that we are interested by a certain asymptotic regime, which justifies that we consider a limit $n \to \infty$. Under this regime, $\pi_n(\tilde{\boldsymbol{\mathcal{X}}}^\circ_n) \to 1$, implying that $\pi_n(\tilde{\boldsymbol{\mathcal{X}}}_n) \to 1$. One can imagine that, if the Markov chains associated with $P_{1,n}$ and $P_{2,n}$ do not behave "too badly" outside of $\tilde{\boldsymbol{\mathcal{X}}}_n$, meaning that when they reach the complement $\tilde{\boldsymbol{\mathcal{X}}}^{\mathsf{c}}_n$ they do not stay there for "too long", then $\mathrm{var}(f_n, \tilde{P}_{1,n})$ and $\mathrm{var}(f_n, \tilde{P}_{2,n})$ should be similar to $\mathrm{var}(f_n, P_{1,n})$ and $\mathrm{var}(f_n, P_{2,n})$. This is what we show in order to prove our theoretical result. In fact, if we think of $P_{1,n}, P_{2,n}, \tilde{P}_{1,n}$ and $\tilde{P}_{2,n}$ as samplers, it is seen in the proof that in order to establish a connection between the asymptotic variances, it simplifies to assume that the performance of the worst of these samplers, measured through $\underline{\lambda}(n)$, is not "too poor", which is a stronger assumption than a performance assumption on $P_{1,n}$ and $P_{2,n}$ only. Under these assumptions, we are able to establish that $\mathrm{var}(f_n, P_{i,n})$ is equal to $\mathrm{var}(f_n, \tilde{P}_{i,n})$, up to an error term that depends on n and that vanishes in the large n regime, $i \in \{1,2\}$, which essentially yields our result. The concentration assumption is reasonable given that in practice the mass often concentrates on a subset of the state-space. This is especially true in high dimensions or when the sample size is large in Bayesian statistics contexts (see, e.g., van der Vaart (1998) and Kleijn and Van der Vaart (2012)).

In light of the above, it is understood that three assumptions are required: the order on the kernels on the control subset, the concentration of π_n and a performance guarantee on the samplers. We now state formally the first two assumptions and then present a simplified version of the theoretical result with a strong performance guarantee. We next present a more general version. To simplify the results, yet keeping the focus on most important cases, we consider in the following that $\omega(n) \leq 1$, meaning that we exclude cases where $P_{1,n}$ is overly dominant on $\tilde{\boldsymbol{\mathcal{X}}}_n$.

Assumption 1 (Kernel ordering). For each n, $P_{1,n}(\mathbf{x},\mathbf{y}) \geq \omega(n) P_{2,n}(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}_n^2$ with $\mathbf{x} \neq \mathbf{y}$, where $\omega(n)$ admits a limit, that is $\lim_{n\to\infty} \omega(n) =: \overline{\omega} > 0$.

Assumption 2 (Mass concentration). The mass concentrates on $\tilde{\boldsymbol{\mathcal{X}}}_n^{\circ}$: $\lim_{n\to\infty} \pi_n(\tilde{\boldsymbol{\mathcal{X}}}_n^{\circ}) = 1$.

Given that together Assumptions 1 and 2 correspond to the assumptions of a classic Peskun ordering as in Andrieu, Lee and Vihola (2018) in the limit, one can only hope to establish, under Assumptions 1 and 2, a version of this ordering that holds in some limiting sense.

Theorem 2 (A simple asymptotic Peskun ordering). *Suppose that Assumptions 1 and 2 hold. Assume that the spectral gaps of $P_{i,n}$ and $\tilde{P}_{i,n}$ are bounded away from zero for all n, $i = 1,2$. Assume also that the sequence $\{f_n\}$ is such that $f_n \in \mathcal{L}^2_{0,1}(\pi_n)$ for all n and such that there exist $\delta > 0$ and $\gamma \in (0, \delta/(2+\delta))$ with*

$$\|f_n\|_{\pi_n, 2+\delta} = o\left(\frac{1}{\left(1 - \pi_n(\tilde{\boldsymbol{\mathcal{X}}}_n^{\circ})\right)^{\gamma}} \right). \tag{3}$$

Then, for any $\epsilon \in (0, \overline{\omega})$, there exists $n^* \in \mathbb{N}$, such that for any $n > n^*$

$$\operatorname{var}(f_n, P_{1,n}) \leq \frac{1}{\overline{\omega} - \epsilon} \operatorname{var}(f_n, P_{2,n}) + \frac{1}{2}\left(\frac{1}{\overline{\omega} - \epsilon} + \frac{1}{\overline{\omega}} \right) - 1 + \frac{\epsilon}{2}.$$

We now make a few remarks about Theorem 2. It allows to retrieve (2) in the limit with $\epsilon \to 0$ and $\omega(n) \to \overline{\omega}$. Theorem 2 will be seen to be a special case of the next one in which the spectral gaps are allowed to decrease with n, which is usually the case when n is the dimension of the state-space. As mentioned, considering that the spectral gaps are bounded away from zero simplifies the assumptions, at the price of requiring a strong performance guarantee.

In addition to the three assumptions mentioned earlier, another one is made in (3). This assumption essentially states that the class of functions that satisfies (3) have a $(2+\delta)$-norm that is allowed to grow with n, but not faster (in fact slightly slower) than $1/(1 - \pi_n(\tilde{\boldsymbol{\mathcal{X}}}_n^{\circ}))$. It is thus not all sequences $\{f_n\}$ that are admissible. It could be tempting to consider a collection of large subsets $\tilde{\boldsymbol{\mathcal{X}}}_n$ to encourage a fast concentration of π_n on these sets, thus allowing for a large class of admissible sequences of functions in Theorem 2; however, the larger are the subsets, the more difficult it becomes to obtain a suitable order on the kernels (Assumption 1).

Different values of the limit of $\omega(n)$, that is $\overline{\omega}$, yield different interpretations of the result. The most important case is when $\overline{\omega} = 1$ for which we can state that the sampler associated with $P_{1,n}$ *asymptotically dominates* that associated with $P_{2,n}$ (for the functions that are admissible). When $\overline{\omega} < 1$, Theorem 2 allows to state that $P_{1,n}$ is *asymptotically comparable* to $P_{2,n}$, in the sense that we have a guarantee that the sampler associated with $P_{1,n}$ will asymptotically produce estimators with variances that are at worst roughly $1/\overline{\omega}$ larger than the sampler associated with $P_{2,n}$ (again for the functions that are admissible).

We now present the general asymptotic Peskun ordering.

Theorem 3 (A general asymptotic Peskun ordering). *Suppose that Assumption 1 holds. Consider a sequence $\{f_n\}$ such that $f_n \in \mathcal{L}^2_{0,1}(\pi_n)$ for all n. Assume that there exist $\delta > 0$ and $\gamma \in (0, \delta/(2+\delta))$*

that satisfy

$$\|f_n\|_{\pi_n,2+\delta} = o\left(\frac{1}{\left(1-\pi_n(\tilde{\boldsymbol{X}}_n^\circ)\right)^\gamma}\right), \tag{4}$$

and

$$1-\pi_n(\tilde{\boldsymbol{X}}_n^\circ) = o\left(\underline{\lambda}(n)^{3/(\bar{\delta}-\gamma)}\right), \tag{5}$$

where $\bar{\delta} := \delta/(2+\delta)$. *Then, for any* $\epsilon \in (0,\overline{\omega})$, *there exists* $n^* \in \mathbb{N}$, *such that for any* $n > n^*$

$$\mathrm{var}(f_n, P_{1,n}) \leq \frac{1}{\overline{\omega}-\epsilon}\mathrm{var}(f_n, P_{2,n}) + \frac{1}{2}\left(\frac{1}{\overline{\omega}-\epsilon} + \frac{1}{\overline{\omega}}\right) - 1 + \frac{\epsilon}{2}.$$

We see that the difference between Theorem 3 and Theorem 2 is that Assumption 2 is replaced by (5), an assumption connecting $\pi_n(\tilde{\boldsymbol{X}}_n^\circ)$ to $\underline{\lambda}(n)$, where the latter is now allowed to decrease. After having selected a sequence $\{f_n\}$ and then δ and γ that satisfy (4) (which is equivalent to (3) in Theorem 2), one has to verify that the choice of δ and γ also allows to verify (5). This equation states that the concentration of π_n on $\tilde{\boldsymbol{X}}_n^\circ$ has to be faster than $\underline{\lambda}(n)^{3/(\bar{\delta}-\gamma)}$. Note that when the spectral gaps are bounded away from zero, (5) is equivalent to Assumption 2, showing that Theorem 2 is indeed a special case of Theorem 3.

We acknowledge the fact that estimating certain rates appearing in the conditions of Theorems 2 and 3, especially the rates of spectral quantities, may constitute a problem in itself. We also acknowledge that our sets of assumptions are probably not optimal, but rather a consequence of our proof technique, and may possibly be improved. However, as mentioned, it is understood that the important aspects (the order on the kernels on the control subset, the mass concentration and performance guarantees) together represent necessary conditions. Given the importance of Peskun-type orderings, we believe it is scientifically interesting to understand under which conditions we can establish a result on the asymptotic variances when an order between $P_{1,n}$ and $P_{2,n}$ holds only on a subset of the state-space.

One may be tempted to assume a (non-trivial) relationship between $\tilde{\lambda}_i(n)$ and $\lambda_i(n)$ given that $\tilde{P}_{i,n}$ is a restriction of $P_{i,n}$ on a subset of the state-space $\boldsymbol{\mathcal{X}}_n$. It turns out that counterexamples show that it is not possible to obtain an interesting result in the general case. In regular sampling contexts, we expect the rates at which $\tilde{\lambda}_i(n)$ and $\lambda_i(n)$ decrease to be in the same regime (i.e., both exponential, both polynomial, etc.). For instance, our analysis in a specific context of graphical-model simulation in Section 5 shows that the decay is polynomial for $\tilde{\lambda}_i(n)$ and $\lambda_i(n)$, $i = 1,2$. The analysis also shows that we can select $\{\tilde{\boldsymbol{X}}_n\}$ such that the mass concentrates on $\tilde{\boldsymbol{X}}_n^\circ$ exponentially quickly, implying that Theorem 3 applies, provided $\|f_n\|_{\pi_n,2+\delta}$ does not grow too rapidly.

3. Lifted samplers for partially-ordered discrete state-spaces

In this section, we start by providing a definition of partially-ordered state-spaces in Section 3.1. We next present in Section 3.2 a generic lifted MCMC algorithm for sampling from distributions on partially-ordered discrete sets. In that section, we make another contribution: we make clear that the implementation of lifted samplers for discrete state-spaces is straightforward, as long as a partial order can be established. We put in contrast this contribution with some of other authors by reviewing the literature about sampling on discrete state-spaces in Section 3.3. Note that, in order to match the classical MCMC framework, we consider in this section the target distribution, state-space, and so on, to be fixed, and will thus denote them without a subscript to simplify.

3.1. Partially-ordered state-spaces

In set theory, a partial order on a set $\mathcal{X}$ is a binary relation defined through a set $\mathcal{R} \subset \mathcal{X}^2$ which is reflexive, anti-symmetric, and transitive. A set $\mathcal{X}$ on which a partial order can be defined, is called *partially ordered*. For such a set, pairs $(\mathbf{x},\mathbf{y}) \in \mathcal{X}^2$ with $\mathbf{x} \neq \mathbf{y}$ are *comparable* when either $(\mathbf{x},\mathbf{y}) \in \mathcal{R}$ or $(\mathbf{y},\mathbf{x}) \in \mathcal{R}$ and are said *incomparable* otherwise. This represents the difference with a totally-ordered set such as $\mathbb{N}$ or $\mathbb{R}$ in which every pair of different elements is comparable. We denote $\mathbf{x} \prec \mathbf{y}$ whenever $(\mathbf{x},\mathbf{y}) \in \mathcal{R}$ and $\mathbf{x} \neq \mathbf{y}$, implying that $\mathbf{x}$ and $\mathbf{y}$ are comparable. Of course, this is not the only way to have comparable $\mathbf{x}$ and $\mathbf{y}$ as we can instead have $\mathbf{y} \prec \mathbf{x}$, that is $(\mathbf{y},\mathbf{x}) \in \mathcal{R}$ and $\mathbf{x} \neq \mathbf{y}$.

An important example of such sets is when any $\mathbf{x} \in \mathcal{X}$ can be written as a vector $\mathbf{x} = (x_1, \ldots, x_n)$ for which each component x_i can be of two types, say Type A or Type B, denoted by $x_i \in \{\mathrm{A},\mathrm{B}\}$. In this case, an inclusion-based partial order on $\mathcal{X}$ can be defined through

$$\mathcal{R} = \{(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{X} : \{i : x_i = \mathrm{A}\} \subset \{i : y_i = \mathrm{A}\}\}. \tag{6}$$

It can be readily checked that $\mathcal{R}$ is reflexive, anti-symmetric and transitive. Moreover, defining $n_{\mathrm{A}}(\mathbf{x})$ to be the number of Type A components in $\mathbf{x}$, that is $n_{\mathrm{A}}(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{1}_{x_i=\mathrm{A}}$, we have that a pair $(\mathbf{x},\mathbf{y}) \in \mathcal{X}^2$ such that $\mathbf{x} \neq \mathbf{y}$ and $n_A(\mathbf{x}) = n_A(\mathbf{y})$ is incomparable.

Partially-ordered sets are encountered in many important areas of statistics including the modelling of binary data using networks or graphs and in variable selection. Indeed, for the former, $\mathcal{X}$ can be parameterized such that $\mathcal{X} = \{-1,+1\}^n$, where for example for an Ising model, $x_i \in \{-1,+1\}$ represents the state of a spin. For variable selection, $\mathcal{X} = \{0,1\}^n$ and $x_i \in \{0,1\}$ indicates whether or not the i-th covariate is included in the model employed.

3.2. Generic algorithm

Let us assume that a neighbourhood structure $\{\mathbf{N}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ and a partial order $\mathcal{R}$ have been specified on $\mathcal{X}$. The sampler that we present is a MCMC algorithm that relies on the lifting technique. The state-space is thus extended: we add a direction variable $\nu \in \{-1,+1\}$ to which we assign a uniform distribution $\mathcal{U}\{-1,+1\}$. The target distribution becomes $\pi \otimes \mathcal{U}\{-1,+1\}$. The idea is to generate proposals belonging to a specific subset of the neighbourhood $\mathbf{N}(\mathbf{x})$, where the subset is defined through $\mathcal{R}$ and chosen according to the direction ν, when the current state of the chain is $(\mathbf{x},\nu)$. In particular, the proposal belongs to $\mathbf{N}_{+1}(\mathbf{x}) := \{\mathbf{y} \in \mathbf{N}(\mathbf{x}) : \mathbf{x} \prec \mathbf{y}\} \subset \mathbf{N}(\mathbf{x})$ when the current state of the direction variable is $\nu = +1$ and to $\mathbf{N}_{-1}(\mathbf{x}) := \{\mathbf{y} \in \mathbf{N}(\mathbf{x}) : \mathbf{y} \prec \mathbf{x}\} \subset \mathbf{N}(\mathbf{x})$ when $\nu = -1$. The partial order is thus used to induce directions to follow in the state-space. We assume that $\mathbf{N}(\mathbf{x})$ is formed only of states that are comparable to $\mathbf{x}$ so that $\mathbf{N}_{-1}(\mathbf{x}) \cup \mathbf{N}_{+1}(\mathbf{x}) = \mathbf{N}(\mathbf{x})$. Note that $\mathbf{N}_{-1}(\mathbf{x}) \cap \mathbf{N}_{+1}(\mathbf{x}) = \varnothing$. The underlying assumption $\mathbf{x} \notin \mathbf{N}(\mathbf{x})$ implies that, strictly speaking, $\mathbf{N}$ is not a neighbourhood in a topological sense. We nevertheless carry on with this abuse of terminology.

Recently, successful applications of the lifting technique have been carried out in contexts where the state-space exhibits a one-dimensional discrete parameter which plays a central role in the sampling scheme: the temperature variable in simulated tempering (Sakai and Hukushima, 2016a) and in parallel tempering (Syed et al., 2022), and the model indicator in selection of nested models (Gagnon and Doucet, 2021). When such a one-dimensional feature does not exist, there is no straightforward way of lifting the state-space and inducing directions without facing issues of reducibility or the risk of obtaining inefficient samplers. Leveraging what can be regarded as a directional-neighbourhood structure induced by the partial order on $\mathcal{X}$ allows to break free from the requirement of resorting to an existing one-dimensional parameter to guide the chain.

Algorithm 1 A lifted sampler for partially-ordered discrete state-spaces

1. Generate $\mathbf{y} \sim q_{\mathbf{x},\nu}$ and $u \sim \mathcal{U}[0,1]$.
2. If
$$u \leq \alpha_\nu(\mathbf{x},\mathbf{y}) := 1 \wedge \frac{\pi(\mathbf{y})\, q_{\mathbf{y},-\nu}(\mathbf{x})}{\pi(\mathbf{x})\, q_{\mathbf{x},\nu}(\mathbf{y})}, \tag{7}$$
set the next state of the chain to $(\mathbf{y},\nu)$. Otherwise, set it to $(\mathbf{x},-\nu)$.
3. Go to Step 1.

In what follows, for each $(\mathbf{x},\nu) \in \boldsymbol{\mathcal{X}} \times \{-1,+1\}$, $\mathbf{N}_\nu(\mathbf{x})$ shall be referred to as the *ν-directional neighbourhood* of state $\mathbf{x}$. The proposal distribution, denoted by $q_{\mathbf{x},\nu}$, where $(\mathbf{x},\nu)$ represents the current state of the Markov chain, is assumed to have its support restricted to $\mathbf{N}_\nu(\mathbf{x})$. It will be noticed that the implementation of the generic algorithm is straightforward provided that a partial ordering has been established. Indeed, the required inputs are:

(i) a neighbourhood structure $\{\mathbf{N}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\mathcal{X}}\}$,
(ii) a partial ordering $\mathcal{R}$ on $\boldsymbol{\mathcal{X}}$,
(iii) proposal distributions $q_{\mathbf{x},\nu}$,

and there exist natural candidates for the proposal distributions, as will be explained in Section 3.3 and, in most cases, for the neighbourhood structure as well.

The MCMC algorithm, which bares a strong resemblance with the guided walk (Gustafson, 1998), is presented in Algorithm 1. We use $x \wedge y$ to denote $\min\{x,y\}$. In Section 2 of the Supplementary Material (Gagnon and Maire (2024)), we consider that $\boldsymbol{\mathcal{X}}$ is a model space and propose a trans-dimensional version of Algorithm 1 that can be used for, among others, variable selection when it is not possible to integrate out the model parameters.

Given that $\boldsymbol{\mathcal{X}}$ is finite, there exists a boundary, in the sense that, for some $(\mathbf{x},\nu)$, $\mathbf{N}_\nu(\mathbf{x})$ is the empty set and there is thus no mass beyond state $\mathbf{x}$ when the direction followed is ν. This is for instance the case in the context of variable selection when $\mathbf{x} = (1,\ldots,1)$, meaning that the current model is the full model, and the direction is $\nu = +1$. Algorithm 1 may thus seem incomplete: it does not explicitly specify how the algorithm behaves on the boundary. We can consider that for any $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ on the boundary, the support of $q_{\mathbf{x},\nu}$ is not $\mathbf{N}_\nu(\mathbf{x})$ (because this is the empty set), but instead given by a fictive state outside $\boldsymbol{\mathcal{X}}$. Given that the support of π is $\boldsymbol{\mathcal{X}}$, then any state outside $\boldsymbol{\mathcal{X}}$ has zero mass under π and such a fictive state is automatically rejected at Step 2. As a consequence, when such a state is proposed, the chain remains at $\mathbf{x}$ and the direction is reversed. Note that this is a technical requirement. In practice, one can simply skip Step 1 when $\mathbf{x}$ is on the boundary and directly set the next state to $(\mathbf{x},-\nu)$.

It is possible to establish that the Markov chain defined by Algorithm 1 is $\pi \otimes \mathcal{U}\{-1,+1\}$-invariant by casting it into a more general algorithm framework presented in Andrieu and Livingstone (2021). We present below the associated generalization of Algorithm 1 which has interesting theoretical features. Beforehand, we introduce necessary notation. Let $\rho_\nu : \boldsymbol{\mathcal{X}} \to [0,1]$, for $\nu \in \{-1,+1\}$, be a user-defined function for which we require that for all $(\mathbf{x},\nu) \in \boldsymbol{\mathcal{X}} \times \{-1,+1\}$:

$$0 \leq \rho_\nu(\mathbf{x}) \leq 1 - T_\nu(\mathbf{x},\boldsymbol{\mathcal{X}}), \tag{8}$$

$$\rho_\nu(\mathbf{x}) - \rho_{-\nu}(\mathbf{x}) = T_{-\nu}(\mathbf{x},\boldsymbol{\mathcal{X}}) - T_\nu(\mathbf{x},\boldsymbol{\mathcal{X}}), \tag{9}$$

Algorithm 2 A generalization of Algorithm 1

1. Generate $u \sim \mathcal{U}[0,1]$.
   (i) If $u \leq T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$, generate $\mathbf{y} \sim Q_{\mathbf{x},\nu}$ and set the next state of the chain to $(\mathbf{y}, \nu)$;
   (ii) if $T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}}) < u \leq T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}}) + \rho_\nu(\mathbf{x})$, set the next state of the chain to $(\mathbf{x}, -\nu)$;
   (iii) if $u > T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}}) + \rho_\nu(\mathbf{x})$, set the next state of the chain to $(\mathbf{x}, \nu)$.
2. Go to Step 1.

where, for all $(\mathbf{x}, \nu) \in \boldsymbol{\mathcal{X}} \times \{-1, +1\}$,

$$T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}}) := \sum_{\mathbf{x}' \in \boldsymbol{\mathcal{X}}} q_{\mathbf{x},\nu}(\mathbf{x}')\, \alpha_\nu(\mathbf{x}, \mathbf{x}') = \sum_{\mathbf{x}' \in \mathbf{N}_\nu(\mathbf{x})} q_{\mathbf{x},\nu}(\mathbf{x}')\, \alpha_\nu(\mathbf{x}, \mathbf{x}').$$

These conditions are considered satisfied in the sequel as they guarantee, as established in Proposition 1 below, that the Markov chain $\{(\mathbf{X}, \nu)_k\}$ is $\pi \otimes \mathcal{U}\{-1, +1\}$-invariant and thus that the marginal process $\{\mathbf{X}_k\}$ is π-invariant. Let $Q_{\mathbf{x},\nu}$ be the probability mass function (PMF) defined through $Q_{\mathbf{x},\nu}(\mathbf{x}') \propto q_{\mathbf{x},\nu}(\mathbf{x}')\, \alpha_\nu(\mathbf{x}, \mathbf{x}')$. The generalization of Algorithm 1 is presented in Algorithm 2.

Proposition 1. *The transition kernel of the Markov chain $\{(\mathbf{X}, \nu)_k\}$ simulated by Algorithm 2 admits $\pi \otimes \mathcal{U}\{-1, 1\}$ as invariant distribution.*

One may notice that $T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$ represents the probability to leave the current state $(\mathbf{x}, \nu)$. In Algorithm 2, we thus first decide if we move on from $\mathbf{x}$, in which case, in Step 1.(i), we randomly select the value of $\mathbf{y}$, the state to move to (using the conditional distribution). It can be readily checked that valid choices for ρ_ν include $\rho_\nu(\mathbf{x}) = 1 - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$ and $\rho_\nu(\mathbf{x}) = \max\{0, T_{-\nu}(\mathbf{x}, \boldsymbol{\mathcal{X}}) - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})\}$. If $\rho_\nu(\mathbf{x}) = 1 - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$, the condition for Case (iii) of Step 1 is never satisfied, and the algorithm either accepts the proposal and keeps the same direction, or the proposal is rejected and the direction is reversed. In this case, one can show that Algorithm 2 corresponds to Algorithm 1, which is why Proposition 1 allows ensuring the correctness of Algorithm 1 as well. Setting $\rho_\nu(\mathbf{x})$ otherwise than $\rho_\nu(\mathbf{x}) = 1 - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$ allows in Case (iii) of Step 1 to keep following the same direction, even when the proposal is rejected. Intuitively, this is desirable when the rejection is due to "bad luck", and not because there is low mass in the direction followed. The function $\rho_\nu(\mathbf{x})$ aims to incorporate this possibility in the sampler.

In a typical MCMC framework with continuous state-spaces, the function $\mathbf{x} \mapsto T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$ is intractable. In such a case, it is therefore usually not possible to set $\rho_\nu(\mathbf{x})$ otherwise than $1 - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$. This contrasts with our discrete state-space framework in which it is often possible to directly compute $T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$. Theorem 6 in Andrieu and Livingstone (2021) states that the best choice of function ρ_ν in terms of a mathematical object related to the asymptotic variance is

$$\rho_\nu^*(\mathbf{x}) := \max\{0, T_{-\nu}(\mathbf{x}, \boldsymbol{\mathcal{X}}) - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})\}, \tag{10}$$

and that the worst choice is $\rho_\nu^{\mathrm{w}}(\mathbf{x}) := 1 - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$. Corollary 1 below establishes an order on the asymptotic variances in the context of finite state-spaces of this paper. Denote by P_ρ the transition kernel corresponding to Algorithm 2 for a given function $\rho_\nu : \boldsymbol{\mathcal{X}} \to [0, 1]$.

Corollary 1. *If $\boldsymbol{\mathcal{X}}$ is finite, then for any function ρ_ν satisfying (8)-(9) and for any function $f : \boldsymbol{\mathcal{X}} \times \{-1, +1\} \to \mathbb{R}$ such that $f(\mathbf{x}, -1) = f(\mathbf{x}, +1)$, we have* $\mathrm{var}(f, P_{\rho^*}) \leq \mathrm{var}(f, P_\rho) \leq \mathrm{var}(f, P_{\rho^w})$.

The price to pay for using ρ_ν^* instead of ρ_ν^{w} is that the algorithm is more complicated to implement because it is required to systematically compute $T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$ at each iteration (it is also sometimes required

to compute $T_{-\nu}(\mathbf{x}, \boldsymbol{\mathcal{X}})$). Using ρ_ν^* thus also comes with an additional computational cost. We observed in some numerical experiments that, if we account for this increased computational cost, there is no gain in efficiency of using Algorithm 2 with ρ_ν^* over Algorithm 2 with ρ_ν^{w} (corresponding to Algorithm 1). One may thus opt for simplicity and implement Algorithm 1. Note that the latter and its MH counterpart have essentially the same computational cost.

3.3. Related work about sampling on discrete state-spaces

Sampling on discrete state-spaces is typically performed using uniform proposal distributions in reversible samplers. If we consider for instance that $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_1, \ldots, x_n \in \{\mathrm{A}, \mathrm{B}\}$, Glauber dynamics for graphical models or the tie-no-tie sampler for network models selects uniformly at random one of the coordinate, say x_i, and proposes to change its value from A to B (B to A) when $x_i = \mathrm{A}$ ($x_i = \mathrm{B}$). Such moves are often rejected when the mass concentrates on a subset of the state-space. To address this issue, Zanella (2020) recently proposed a *locally-balanced* generic approach for which the probability to select the i-th coordinate depends on the relative mass of the resulting proposal, that is $\pi(\mathbf{y})/\pi(\mathbf{x})$, aiming to propose less "naive" moves. Zanella (2020) proves that the acceptance probabilities converge to 1 in a high-dimensional regime. This property suggests that locally-balanced samplers are efficient, at least in high dimensions. Indeed, samplers for discrete state-spaces typically use the same neighbourhood structure $\{\mathbf{N}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\mathcal{X}}\}$, implying that the range of the proposal distributions is the same and that higher acceptance probabilities often translate into better mixing properties. Zanella (2020) in fact empirically shows that locally-balanced samplers perform better than alternative solutions to sample from PMFs, and that the difference is highly marked in the high-dimensional regime. Yet, the samplers are reversible, implying that the chains may often go back to recently visited states, or in other words, that the chains exhibit a random-walk behaviour.

In the presented generic algorithms in Section 3.2, there is no restriction on the proposal distributions $q_{\mathbf{x},\nu}$. In Section 4.2, we set them to locally-balanced proposal distributions, thus combining the strengths of the lifting and locally-balanced approaches. An illustration showing the benefit of this combination is provided in Figure 1 in which we measure the performance using the effective sample size (ESS) of a statistic, reported per iteration. ESS per iteration is defined as the inverse of the integrated autocorrelation time. When the chains start in stationarity, integrated autocorrelation time corresponds to the asymptotic variance of a standardized version of the statistic. A small asymptotic variance thus corresponds to a high ESS (and vice versa).

Other (somewhat) generic approaches to non-reversible sampling on discrete state-spaces are (to our knowledge) all contemporary to ours: Bierkens (2016), Sakai and Hukushima (2016b), Power and Goldman (2019), Faizi, Deligiannidis and Rosta (2020) and Herschlag et al. (2020). They rely on the lifting technique as well, except Bierkens (2016). Our work is most closely related to Power and Goldman (2019) in which the approach of Zanella (2020) is also exploited. In fact, when $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_1, \ldots, x_n \in \{\mathrm{A}, \mathrm{B}\}$, Algorithm 1 corresponds to the discrete-time version of a specific sampler independently developed in Power and Goldman (2019). Algorithm 1 can also be seen to be a special case of a sampler presented in Sakai and Hukushima (2016b) in which a general extended transition matrix is defined from lifting the MH one. A similar approach, described in Faizi, Deligiannidis and Rosta (2020), explicitly incorporates the changes in the function f by moving from a state to another in the transition matrix; this latter approach is closely related to ours when $f(\mathbf{x})$ decreases or increases every time we change $\mathbf{x}$ for $\mathbf{y}$ with $\mathbf{x} \prec \mathbf{y}$. We consequently do not claim originality for the samplers presented here. In those papers, however, the notion of partial ordering is not identified nor exploited; the focus is rather on improving state-space exploration through the exploitation of *any* symmetric or algebraic structure of $\boldsymbol{\mathcal{X}}$ identified by users. The focus is the same in Bierkens (2016), but the non-reversibility is

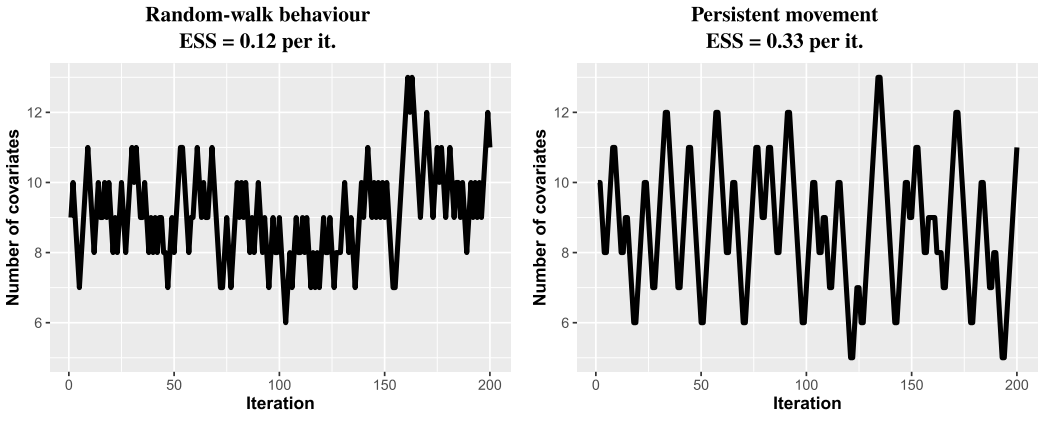

Figure 1. Trace plots for the statistic of number of covariates included in a model for a MH sampler with a locally-balanced proposal distribution on the left panel and its lifted counterpart on the right panel, when applied to solve a real variable-selection problem presented in Section 1 of the Supplementary Material (Gagnon and Maire (2024)).

achieved by directly modifying the acceptance probability in MH, using the notion of vorticity matrix; this approach is valid in general state-space contexts. In Herschlag et al. (2020), the authors generalize non-reversible lifted kernels to *mixed skewed* kernels by means of a series of involutions in a context of undirected graph sampling. In their work, the main application is sampling of districting maps to evaluate the degree of partisan districting. The involutions are created by a series of user-specified vortices that generate non-reversible flows on the state-space. Interestingly, this scheme can be seen as creating directional neighbourhoods.

4. Two specific lifted samplers and their analysis

In this section, we specify two lifted samplers through two different choices of proposal distributions $q_{\mathbf{x},\nu}$ and provide a theoretical analysis using the asymptotic Peskun ordering. We first present and analyse in Section 4.1 a lifted sampler using uniform proposal distributions. As explained in Section 3.3, this sampler is often inefficient, especially in high dimensions, but it is simple enough to allow an easy understanding of the reasons why lifted samplers are *not* expected to always dominate their MH counterparts within our framework. We next turn in Section 4.2 to a more promising choice of proposal distributions, namely the locally-balanced proposal distributions, and study the resulting lifted sampler.

As mentioned, the study here will be conducted in great generality. More precisely, the target distribution will not be specified; we will thus not be in a position to explicitly estimate the rates appearing in the conditions of our theoretical results presented in Section 2. We will make assumptions regarding these rates, but this will not prevent us from defining the control subsets. Making assumptions regarding the rates appearing in the conditions of our theoretical results and defining judiciously the control subsets will allow to gain general insights into the situations in which the lifted samplers are expected to outperform their MH counterparts, and also into those in which there is no guarantee. In Section 5, we conduct a thorough study in a specific context of graphical-model simulation. This will allow to have a concrete example of how the assumptions of our theoretical results can be verified in practice. That study will also allow to improve the understanding of the behaviour of lifted samplers through practical results.

For ease of presentation, we consider in this section the setup where $\mathbf{x} = (x_1, \ldots, x_n)$ and $x_1, \ldots, x_n \in \{\mathrm{A}, \mathrm{B}\}$ with the partial order on $\boldsymbol{\mathcal{X}}$ defined in (6). We consider, additionally, but without loss of generality, a Ising model context where $\mathrm{A} = -1$ and $\mathrm{B} = +1$. Finally, we consider that the neighbourhood structure, used by all samplers, is the typical one, meaning that the neighbourhoods are set to $\mathbf{N}(\mathbf{x}) = \{\mathbf{y} \in \boldsymbol{\mathcal{X}}_n : \sum_i |x_i - y_i| = 2\}$, so that the algorithms propose to flip a single bit at each iteration. Because of the nature of our analysis, we, as in Section 2, highlight a dependency on n of the target distribution, the state-space, and so on, by denoting them by π_n, $\boldsymbol{\mathcal{X}}_n$, etc.

4.1. Uniform proposal distributions

In the reversible MH sampler, it is common, as mentioned in Section 3.3, to set the proposal distribution, denoted by $q_{\mathbf{x}}$ for this algorithm, to the uniform distribution over the neighbourhood of the current state $\mathbf{x} \in \boldsymbol{\mathcal{X}}_n$, that is $q_{\mathbf{x}} = \mathcal{U}\{\mathbf{N}(\mathbf{x})\}$. In the framework of Algorithms 1 and 2, the analogous proposal distribution is naturally defined as $q_{\mathbf{x},\nu} = \mathcal{U}\{\mathbf{N}_\nu(\mathbf{x})\}$. In this case, the acceptance probability (7) of a proposed move becomes

$$\alpha_\nu(\mathbf{x}, \mathbf{y}) = 1 \wedge a_\nu(\mathbf{x}, \mathbf{y}), \qquad a_\nu(\mathbf{x}, \mathbf{y}) := \frac{\pi_n(\mathbf{y})\, |\mathbf{N}_\nu(\mathbf{x})|}{\pi_n(\mathbf{x})\, |\mathbf{N}_{-\nu}(\mathbf{y})|},$$

where we refer to a_ν as the *acceptance ratio*. The function $|\cdot|$ when applied to a set is the cardinality.

In the MH sampler, given that the neighbourhoods are set to $\mathbf{N}(\mathbf{x}) = \{\mathbf{y} \in \boldsymbol{\mathcal{X}}_n : \sum_i |x_i - y_i| = 2\}$, the uniform distribution chooses which bit to flip uniformly at random. Therefore, the size of the neighbourhoods in this sampler is constant for any $\mathbf{x}$ and is given by n. This implies that the acceptance probability in this sampler, denoted by $\alpha(\mathbf{x}, \mathbf{y})$, reduces to

$$\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge a(\mathbf{x}, \mathbf{y}), \qquad a(\mathbf{x}, \mathbf{y}) := \frac{\pi_n(\mathbf{y})\, q_{\mathbf{y}}(\mathbf{x})}{\pi_n(\mathbf{x})\, q_{\mathbf{x}}(\mathbf{y})} = \frac{\pi_n(\mathbf{y})}{\pi_n(\mathbf{x})}.$$

In the lifted case, we have that for any $\nu \in \{-1, +1\}$, $n_\nu(\mathbf{x}) = \sum_{i=1}^n \mathbb{1}_{x_i = \nu}$ and the acceptance probability can thus be rewritten as:

$$\alpha_\nu(\mathbf{x}, \mathbf{y}) = 1 \wedge a_\nu(\mathbf{x}, \mathbf{y}), \qquad a_\nu(\mathbf{x}, \mathbf{y}) = a(\mathbf{x}, \mathbf{y})\, \frac{n_{-\nu}(\mathbf{x})}{n_\nu(\mathbf{y})}. \tag{11}$$

Indeed, $\mathbf{N}_\nu(\mathbf{x}) = \{\mathbf{y} \in \boldsymbol{\mathcal{X}}_n : \text{there exists one } j \text{ such that } y_j = -x_j = \nu\}$ implies that $|\mathbf{N}_\nu(\mathbf{x})| = n_{-\nu}(\mathbf{x})$. The acceptance probability α_ν thus depends on an additional factor $n_{-\nu}(\mathbf{x})/n_\nu(\mathbf{y})$ compared to α in the MH sampler. While the reversible sampler is allowed to backtrack, which makes the size of the neighbourhoods constant, the size of the neighbourhoods diminishes in the lifted sampler as the chain moves further in a given direction (making the neighbourhoods in the reverse direction bigger and bigger). As a consequence, the longer the acceptance streak, the smaller $n_{-\nu}(\mathbf{x})/n_\nu(\mathbf{y})$. On an acceptance streak, this factor eventually becomes less than one and thus shrinks α_ν, relatively to the MH acceptance ratio, until the lifted chain switches its direction. To summarize, the price to pay when considering a Markov chain with persistent dynamic is a shrinking factor $n_{-\nu}(\mathbf{x})/n_\nu(\mathbf{y})$ in the acceptance ratio.

An *ideal* situation, which is incompatible with most statistical models, is one where

$$|\mathbf{N}_{-1}(\mathbf{x})| = |\mathbf{N}_{+1}(\mathbf{x})| = |\mathbf{N}(\mathbf{x})|/2 = n/2, \quad \text{for } \pi_n\text{-almost all } \mathbf{x} \in \boldsymbol{\mathcal{X}}_n. \tag{12}$$

This implies that if the chain is at state $(\mathbf{x}, \nu)$, $a(\mathbf{x}, \mathbf{y}) = a_\nu(\mathbf{x}, \mathbf{y})$ for all $\mathbf{y} \in \mathbf{N}_\nu(\mathbf{x})$. Qualitatively, the persistent dynamic of the lifted chain is no longer counter-balanced by the shrinking factor and is thus

expected to be more efficient than MH. This fact is made rigourous in Corollary 2 below, which follows from Theorem 7 of Andrieu and Livingstone (2021). In the rest of this subsection, the transition kernel corresponding to Algorithm 2 with $q_{\mathbf{x},\nu} = \mathcal{U}\{\mathcal{N}_\nu(\mathbf{x})\}$ is denoted by $P_{\rho,n}$ and that of the MH sampler with $q_{\mathbf{x}} = \mathcal{U}\{\mathcal{N}(\mathbf{x})\}$ by $P_{\mathrm{MH},n}$. Recall that Algorithm 2 with ρ_ν^{w} corresponds to Algorithm 1.

Corollary 2. *Let $n \in \mathbb{N}$. If $\mathcal{X}_n$ is finite and (12) holds, then for any function $f_n : \mathcal{X} \times \{-1,+1\} \to \mathbb{R}$ such that $f_n(\mathbf{x},-1) = f_n(\mathbf{x},+1)$, we have $\mathrm{var}(f_n, P_{\rho,n}) \leq \mathrm{var}(f_n, P_{\mathrm{MH},n})$.*

The proof of Corollary 2 is postponed to Section 3 of the Supplementary Material (Gagnon and Maire (2024)) but its main steps are now presented as they highlight what is important to obtain such an ordering. Central to the proof of Corollary 2 is the idea that once a lifted sampler is defined, it is possible to identify a non-lifted counterpart which differs from Algorithm 2 in that the direction is resampled $\nu \sim \mathcal{U}\{-1,+1\}$ at the beginning of each iteration. At each iteration, a choice between $q_{\mathbf{x},-1}$ and $q_{\mathbf{x},+1}$ is thus first made uniformly at random, and the proposal is next sampled. Non-lifted refers to the fact that, while operating on the extended state-space, the systematic resampling of ν makes the marginal dynamic $\{\mathbf{X}_k\}$ Markov again, and reversible. This scheme, when looking at functions f_n with $f_n(\mathbf{x},-1) = f_n(\mathbf{x},+1)$, makes the extension of state-space to include the direction variable superfluous, explaining how a comparison between $\mathrm{var}(f_n, P_{\rho,n})$ and $\mathrm{var}(f_n, P_{\mathrm{MH},n})$ is possible. Let $P_{\mathrm{rev.},n}$ be the transition kernel of this non-lifted reversible Markov chain. As noted in Andrieu and Livingstone (2021), $P_{\mathrm{rev.},n}$ can indeed be seen as an intermediate kernel through which comparison of the asymptotic variance of $P_{\rho,n}$ and $P_{\mathrm{MH},n}$ is possible if one can establish, perhaps independently, that $\mathrm{var}(f, P_{\rho,n}) \leq \mathrm{var}(f, P_{\mathrm{rev.},n})$ and $\mathrm{var}(f, P_{\mathrm{rev.},n}) \leq \mathrm{var}(f, P_{\mathrm{MH},n})$. While establishing the former essentially follows from Theorem 7 of Andrieu and Livingstone (2021), the latter may prove more difficult. However, under (12), it turns out that $P_{\mathrm{rev.},n} = P_{\mathrm{MH},n}$, trivially establishing that $\mathrm{var}(f, P_{\mathrm{rev.},n}) = \mathrm{var}(f, P_{\mathrm{MH},n})$. Indeed, the sub-stochastic part of $P_{\mathrm{rev.},n}$ associated with accepted proposals is

$$(1/2)q_{\mathbf{x},+1}(\mathbf{y})\alpha_{+1}(\mathbf{x},\mathbf{y}) + (1/2)q_{\mathbf{x},-1}(\mathbf{y})\alpha_{-1}(\mathbf{x},\mathbf{y}), \tag{13}$$

and it can be readily checked that under (12), (13) indeed coincides with the sub-stochastic part of $P_{\mathrm{MH},n}$. These are the same mathematical arguments that allow to prove the dominance mentioned in Section 1.3 of lifted samplers over their MH counterparts when the state-space is totally ordered.

The incompatibility of the condition (12) with most statistical models motivates us to take our analysis one step further, and this is where the asymptotic Peskun ordering presented in Section 2 proves useful. Note that in order to find a model such that (12) is satisfied, one has to be quite creative; an example is provided in Section 4 of the Supplementary Material (Gagnon and Maire (2024)). The next step in our analysis is to establish if the order on the asymptotic variances still holds when (12) is relaxed, and if not, we want to find conditions under which $\mathrm{var}(f_n, P_{\rho,n})$ and $\mathrm{var}(f_n, P_{\mathrm{MH},n})$ can be compared. A modification of our example presented in the supplementary material shows that the order on the asymptotic variances *does not* necessarily hold when (12) is relaxed. This should not come as a surprise in the light of the aforementioned observations about the potentially shrinking factor in α_ν. Comparing the efficiency of $P_{\mathrm{MH},n}$ and $P_{\rho,n}$ beyond the context of Corollary 2 is not an easy task for several reasons:

- $P_{\rho,n}$ is not reversible and most techniques to establish domination results between Markov kernels hold for reversible kernels, Andrieu and Livingstone (2021) being a noteworthy exception;
- the two kernels $P_{\mathrm{MH},n}$ and $P_{\rho,n}$ are not defined on the same state-space.

For these reasons, finding reasonable conditions under which $\mathrm{var}(f_n, P_{\mathrm{rev.},n})$ and $\mathrm{var}(f_n, P_{\mathrm{MH},n})$ can be compared appears to be a suitable route to establish a comparison between $P_{\rho,n}$ et $P_{\mathrm{MH},n}$ (given

that we already know that $\mathrm{var}(f_n, P_{\rho,n}) \leq \mathrm{var}(f_n, P_{\mathrm{rev.},n})$ using similar arguments to those used to prove Corollary 2). We thus employ Theorems 2 and 3.

Note that if one manages to design the distributions $q_{\mathbf{x},\nu}$ such that $q_{\mathbf{x}}(\mathbf{y}) = (1/2)q_{\mathbf{x},-1}(\mathbf{y}) + (1/2)q_{\mathbf{x},+1}(\mathbf{y})$ for all $\mathbf{x},\mathbf{y}$, then one directly has $P_{\mathrm{rev.},n} = P_{\mathrm{MH},n}$ and thus a comparison between $P_{\rho,n}$ and $P_{\mathrm{MH},n}$; this is the approach proposed in Kamatani and Song (2023) for general state-spaces, but it is one that cannot in general be applied in the case of discrete state-spaces. Note also that if one is interested in comparing a lifted sampler using proposal distributions $q_{\mathbf{x},\nu}$ with a MH sampler using proposal distributions defined as $q_{\mathbf{x}}(\mathbf{y}) = (1/2)q_{\mathbf{x},-1}(\mathbf{y}) + (1/2)q_{\mathbf{x},+1}(\mathbf{y})$, then again $P_{\mathrm{rev.},n} = P_{\mathrm{MH},n}$ and a comparison between $P_{\rho,n}$ and $P_{\mathrm{MH},n}$ is direct. In the context of variable selection, the latter MH sampler corresponds to one where it is first chosen to add a covariate or remove one already in the model, and next which covariate to add or delete. In our paper, we focus on the common situation where, in the MH sampler, a proposal is made uniformly at random from $\mathbf{N}(\mathbf{x})$ (or using a locally-balanced weight function as described in Section 4.2), and we want to compare a lifted sampler with the MH one.

The idea that we now explore is to consider situations where the mass concentrates on an area where we have a control over the factor $n_{-\nu}(\mathbf{x})/n_{\nu}(\mathbf{y})$ in α_ν (11), which translates into the existence of a (non-trivial) relationship between the sub-stochastic part of $P_{\mathrm{rev.},n}$ and that of $P_{\mathrm{MH},n}$ on this area. To simplify, we consider situations where the mass concentrates on the centre of the domain, i.e., on states where $n_{-1}(\mathbf{x})$ and $n_{+1}(\mathbf{x})$ are not too far from $n/2$, and set

$$\tilde{\boldsymbol{\mathcal{X}}}_n := \{\mathbf{x} \in \boldsymbol{\mathcal{X}}_n : n/2 - \beta(n) \leq n_{-1}(\mathbf{x}), n_{+1}(\mathbf{x}) \leq n/2 + \beta(n)\}, \tag{14}$$

by choosing a specific function $\beta : \mathbb{N} \to (0, \infty)$. With this definition of $\tilde{\boldsymbol{\mathcal{X}}}_n$ and that of the neighbourhood structure (mentioned at the beginning of Section 4), we are able to state that the interior of $\tilde{\boldsymbol{\mathcal{X}}}_n$ is as follows: $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ = \{\mathbf{x} \in \boldsymbol{\mathcal{X}}_n : n/2 - \beta(n) + 1 \leq n_{-1}(\mathbf{x}), n_{+1}(\mathbf{x}) \leq n/2 + \beta(n) - 1\}$. Note that the analysis can be done by considering instead that the mass concentrates on states where the minimum between $n_{-1}(\mathbf{x})$ and $n_{+1}(\mathbf{x})$ is not too far from n/κ with $\kappa \geq 2$. The difference is that, with control subsets defined as in (14), $\omega(n)$ will be seen to converge to $\overline{\omega} = 1$, whereas in the general framework, $\overline{\omega} \leq 1$ and a function of κ, and the results are more complicated to present. Constructing the control subsets $\{\tilde{\boldsymbol{\mathcal{X}}}_n\}$ as in (14) implies that, remarkably, the analysis is parameterized by the sole function β.

Lemma 1. *Consider the definition of $\tilde{\boldsymbol{\mathcal{X}}}_n$ in (14). Assume that β is such that $\beta(n) = o(n)$. Then for a large enough n, it holds that $\tilde{P}_{\mathrm{rev.},n}(\mathbf{x},\mathbf{y}) \geq \omega(n)\tilde{P}_{\mathrm{MH},n}(\mathbf{x},\mathbf{y})$, for all $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}_n^2$ with $\mathbf{x} \neq \mathbf{y}$, where*

$$\omega(n) = \left(1 - \frac{\beta(n)}{n/2}\right)\left(1 + \frac{\beta(n)}{n/2}\right)^{-2} \to \overline{\omega} = 1. \tag{15}$$

Intuitively, if $\beta(n)$ grows like n or faster, then for a large enough n we have $\boldsymbol{\mathcal{X}}_n = \tilde{\boldsymbol{\mathcal{X}}}_n$ which boils down to the initial Peskun's problem so the assumption $\beta(n) = o(n)$ is sensible. If $\beta(n)$ grows *too* slowly then the control subsets $\tilde{\boldsymbol{\mathcal{X}}}_n$ may eventually fail to track the bulk of $\boldsymbol{\mathcal{X}}_n$, resulting in that the mass of π_n will not concentrate on $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$ and that the restricted kernels will be too different from the original ones to allow the machinery of Section 2 to work. The condition $\beta(n) = o(n)$, together with (15), means that Assumption 1 holds with $\overline{\omega} = 1$. If we assume that the spectral gaps are bounded away from zero, which is realistic, for example, when $\beta(n)$ is constant, and that π_n concentrates on $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$ defined above (implying that Assumption 2 holds), then Theorem 2 can be applied and

$$\mathrm{var}(f_n, P_{\rho,n}) \leq \mathrm{var}(f_n, P_{\mathrm{rev.},n}) \leq \frac{1}{1-\epsilon}\mathrm{var}(f_n, P_{\mathrm{MH},n}) + \frac{\epsilon}{2(1-\epsilon)} + \frac{\epsilon}{2},$$

for any $\epsilon > 0$, provided that n is large enough and that we consider functions $f_n \in \mathcal{L}^2_{0,1}(\pi_n)$ satisfying (3) and such that $f_n(\mathbf{x}, -1) = f_n(\mathbf{x}, +1)$. The assumption on the spectral gaps can be relaxed and Theorem 3 can be instead applied if we are able to establish a connection between the rate at which π_n concentrates on $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$ and that at which $\underline{\lambda}(n)$ decreases, i.e., if (5) can be verified.

To summarize, our analysis suggests that the lifted sampler with uniform proposal distributions dominates its MH counterpart (at least for n large enough and a specific class of functions) when π_n concentrates on states in the centre of the domain. If it concentrates elsewhere, then the lifted sampler is expected to be comparable to its MH counterpart as long as π_n does not concentrate on areas where the neighbourhoods, and thus the additional factors $n_{-\nu}(\mathbf{x})/n_\nu(\mathbf{y})$ in (11), are very unbalanced.

When n is large, uniform proposal distributions, whether they are used in a lifted or MH sampler, are likely to represent a poor strategy. We will thus not focus on samplers with uniform proposal distributions in our study in a context of graphical-model simulation in Section 5. We will rather focus on studying locally-balanced samplers presented in the next subsection which represent efficient alternatives.

4.2. Locally-balanced proposal distributions

In this section, we discuss and analyse samplers using locally-balanced proposal distributions. For simplicity, we will use the same notation as in Section 4.1: $q_{\mathbf{x}}$ and $q_{\mathbf{x},\nu}$ are the proposal distributions in the MH and lifted samplers, respectively, but in this section they are locally-balanced (a definition follows), and $P_{\rho,n}$, $P_{\text{rev.},n}$ and $P_{\text{MH},n}$ are the Markov kernels associated with Algorithm 2, and its non-lifted and MH counterparts, respectively, which are all using locally-balanced proposal distributions. Recall that Algorithm 1 is a special case of Algorithm 2 with $\rho_\nu(\mathbf{x}) = 1 - T_\nu(\mathbf{x}, \boldsymbol{\mathcal{X}})$.

As defined in Zanella (2020) in the MH framework, a proposal distribution is locally-balanced if

$$q_{\mathbf{x}}(\mathbf{y}) = g\left(\frac{\pi_n(\mathbf{y})}{\pi_n(\mathbf{x})}\right) \Big/ c_n(\mathbf{x}), \quad \mathbf{y} \in \mathbf{N}(\mathbf{x}),$$

where $c_n(\mathbf{x})$ is the normalizing constant, that is $c_n(\mathbf{x}) = \sum_{\mathbf{x}' \in \mathbf{N}(\mathbf{x})} g(\pi_n(\mathbf{x}')/\pi_n(\mathbf{x}))$, and g is a positive continuous function such that $g(x)/g(1/x) = x$ for $x > 0$. Such a function g implies that the acceptance probability in the MH algorithm is given by

$$\alpha(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{\pi_n(\mathbf{y})\, q_{\mathbf{y}}(\mathbf{x})}{\pi_n(\mathbf{x})\, q_{\mathbf{x}}(\mathbf{y})} = 1 \wedge \frac{c_n(\mathbf{x})}{c_n(\mathbf{y})}. \tag{16}$$

The name *locally-balanced* comes from the fact that, in the limit, when the state-space becomes larger and larger (but the neighbourhoods have a fixed size and proposed moves are thus local), there is no need for an accept-reject step anymore; the proposal distributions leave the distribution π_n invariant. Indeed, as shown in Zanella (2020), $\sup_{(\mathbf{x},\mathbf{y}) \in \boldsymbol{\mathcal{X}}_n : \mathbf{y} \in \mathbf{N}(\mathbf{x})} c_n(\mathbf{x})/c_n(\mathbf{y}) \to 1$ as $n \to \infty$ under some assumptions. The author more precisely considers that $\mathbf{x} = (x_1, \ldots, x_n)$ and that at any given iteration, only a small fraction of the n components is proposed to change values. The result holds when there exists a uniform bound which does not depend on n on $\pi_n(\mathbf{y})/\pi_n(\mathbf{x})$ for all pairs of neighbouring states $(\mathbf{x}, \mathbf{y})$ and the random variables $X_1, \ldots, X_n$ exhibit a structure of conditional independence, the latter implying that the normalizing constants $c_n(\mathbf{x})$ and $c_n(\mathbf{y})$ share a lot of terms. Note that $c_n(\mathbf{x})$ and $c_n(\mathbf{y})$ are both sums over the same number of terms, which is crucial in showing that $\sup_{(\mathbf{x},\mathbf{y}) \in \boldsymbol{\mathcal{X}}_n : \mathbf{y} \in \mathbf{N}(\mathbf{x})} c_n(\mathbf{x})/c_n(\mathbf{y}) \to 1$.

Two valid choices for g are $g(x) = \sqrt{x}$ and $g(x) = x/(1+x)$, the latter yielding what is referred to as the *Barker proposal distribution* in reference to Barker (1965)'s acceptance probability choice. The advantage of the latter choice is that it is a bounded function of x, which stabilizes the normalizing

constants and thus the acceptance probability, see Zanella (2020), and Livingstone and Zanella (2022) for the continuous-random-variable case.

A locally-balanced proposal distribution in the lifted-sampler framework is naturally defined as

$$q_{\mathbf{x},\nu}(\mathbf{y}) = g\left(\frac{\pi_n(\mathbf{y})}{\pi_n(\mathbf{x})}\right) \Big/ c_{n,\nu}(\mathbf{x}), \quad \mathbf{y} \in \mathbf{N}_\nu(\mathbf{x}),$$

where $c_{n,\nu}(\mathbf{x})$ is the normalizing constant and g is as above. In this case,

$$\alpha_\nu(\mathbf{x},\mathbf{y}) = 1 \wedge \frac{\pi_n(\mathbf{y})\, q_{\mathbf{y},-\nu}(\mathbf{x})}{\pi_n(\mathbf{x})\, q_{\mathbf{x},\nu}(\mathbf{y})} = 1 \wedge \frac{c_{n,\nu}(\mathbf{x})}{c_{n,-\nu}(\mathbf{y})} = 1 \wedge \frac{c_n(\mathbf{x})}{c_n(\mathbf{y})} \frac{c_{n,\nu}(\mathbf{x})/c_n(\mathbf{x})}{c_{n,-\nu}(\mathbf{y})/c_n(\mathbf{y})}. \tag{17}$$

As with the uniform proposal distributions in Section 4.1, we see that the acceptance probability in the lifted sampler (17) differs from that in MH (16). There is thus again a price to pay to use a lifted sampler: there is no guarantee that $c_{n,\nu}(\mathbf{x})/c_{n,-\nu}(\mathbf{y}) \to 1$ for $\mathbf{y} \in \mathbf{N}(\mathbf{x})$, even when $c_n(\mathbf{x})/c_n(\mathbf{y}) \to 1$. A reason is because the sums $c_{n,\nu}(\mathbf{x})$ and $c_{n,-\nu}(\mathbf{y})$ are in this case not over the same number of terms, a consequence of the nature of the lifted sampler.

As previously, the reversible counterpart to the lifted algorithm chooses at each iteration uniformly at random a proposal distribution between $q_{\mathbf{x},-1}$ and $q_{\mathbf{x},+1}$ from which a proposal is sampled. Imagine that $c_n(\mathbf{x})/c_n(\mathbf{y}) = 1$ for all $\mathbf{x},\mathbf{y}$, then one can notice from (17) that the stability of ratios $c_{n,\nu}(\mathbf{x})/c_{n,-\nu}(\mathbf{y})$ is crucial to establish a connection between the sub-stochastic parts of $P_{\text{rev.},n}$ and $P_{\text{MH},n}$ (recall (13)). In fact, in an ideal situation, which is again incompatible with most statistical models, one can establish that $P_{\text{rev.},n} = P_{\text{MH},n}$, guaranteeing a dominance of the lifted sampler.

Corollary 3. *Let $n \in \mathbb{N}$. If $\boldsymbol{\mathcal{X}}_n$ is finite and $c_{n,-1}(\mathbf{x}) = c_{n,+1}(\mathbf{x})$, for all $\mathbf{x} \in \boldsymbol{\mathcal{X}}_n$, then for any function $f_n : \boldsymbol{\mathcal{X}} \times \{-1,+1\} \to \mathbb{R}$ such that $f_n(\mathbf{x},-1) = f_n(\mathbf{x},+1)$, we have* $\text{var}(f_n, P_{\rho,n}) \leq \text{var}(f_n, P_{MH,n})$.

Locally-balanced proposal distributions allow to explore the state-space by often proposing points that belong to the subset on which the mass concentrates. Corollary 3 tells us that, in order to compare $P_{\text{rev.},n}$ to $P_{\text{MH},n}$ (and thus $P_{\rho,n}$ to $P_{\text{MH},n}$), the directional neighbourhoods to which these points belong must have similar mass, implying similar normalizing constants $c_{n,-1}(\mathbf{x})$ and $c_{n,+1}(\mathbf{x})$ over the subset. The analysis can be pushed beyond Corollary 3 by making use of our asymptotic framework. To simplify, we consider, as in Zanella (2020), the situation where $\sup c_n(\mathbf{x})/c_n(\mathbf{y}) \to 1$ where the supremum is over all neighbouring states $\mathbf{x},\mathbf{y}$, and $\overline{\omega} = 1$.

We now turn to the definition of the control subset:

$$\begin{aligned}\tilde{\boldsymbol{\mathcal{X}}}_n &= \{\mathbf{x} \in \boldsymbol{\mathcal{X}}_n : 1 - \beta(n)/(c_n(\mathbf{x})/2) \leq c_{n,\nu}(\mathbf{x})/(c_n(\mathbf{x})/2) \leq 1 + \beta(n)/(c_n(\mathbf{x})/2)\} \\ &= \{\mathbf{x} \in \boldsymbol{\mathcal{X}}_n : |c_{n,-1}(\mathbf{x}) - c_{n,+1}(\mathbf{x})| \leq 2\beta(n)\},\end{aligned} \tag{18}$$

which again is defined through a function $\beta : \mathbb{N} \to (0,\infty)$. The equivalence between the sets follows from the fact that $c_n(\mathbf{x}) = c_{n,-1}(\mathbf{x}) + c_{n,+1}(\mathbf{x})$. Under assumptions on the target such as those in Zanella (2020), the normalizing constants $c_n(\mathbf{x})$ scale linearly with n and below we show that lifted and MH samplers can be compared in terms of asymptotic variances when $\beta(n) = o(n)$, because in this case for states in $\tilde{\boldsymbol{\mathcal{X}}}_n$, $\beta(n)/(c_n(\mathbf{x})/2)$ vanishes and the acceptance probabilities in the lifted sampler are close to 1, as those in MH. Notice that in the case of locally-balanced samplers, we cannot state explicitly what the interior of $\tilde{\boldsymbol{\mathcal{X}}}_n$ is without specifying π_n. With the current level of generality, we cannot go beyond the definition presented in Section 2, which in the framework of this section is $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ := \{\mathbf{x} \in \tilde{\boldsymbol{\mathcal{X}}}_n : q_{\mathbf{x}}(\tilde{\boldsymbol{\mathcal{X}}}_n^{\mathsf{c}}) = 0\}$.

As in the previous section, the analysis can be done by considering instead that the mass concentrates on states where the minimum between $c_{n,-1}(\mathbf{x})$ and $c_{n,+1}(\mathbf{x})$ is not too far from $c_n(\mathbf{x})/\kappa$ with $\kappa \geq 2$. In this case, $\overline{\omega} \leq 1$ and a function of κ, and the definition of the control subset and results are more complex. From the definition of $\tilde{\boldsymbol{\mathcal{X}}}_n$ in (18), we are able to establish a result analogous to Lemma 1.

Lemma 2. *Consider the definition of $\tilde{\boldsymbol{\mathcal{X}}}_n$ in (18) and let $R_n := \{(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}_n^2 : \mathbf{y} \in \mathbf{N}(\mathbf{x})\}$. Assume that*

$$\inf_{(\mathbf{x},\mathbf{y})\in R_n} g\left(\frac{\pi_n(\mathbf{y})}{\pi_n(\mathbf{x})}\right) \geq m, \quad \tau_n := \sup_{(\mathbf{x},\mathbf{y})\in R_n} \frac{c_n(\mathbf{x})}{c_n(\mathbf{y})} \to 1, \quad \beta(n) = o(n),$$

with m independent of n. Then, for a large enough n, it holds that $\tilde{P}_{\text{rev.},n}(\mathbf{x},\mathbf{y}) \geq \omega(n)\tilde{P}_{\text{MH},n}(\mathbf{x},\mathbf{y})$, for all $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{\mathcal{X}}}_n^2$ with $\mathbf{x} \neq \mathbf{y}$, where

$$\omega(n) = \left(1 + \frac{\beta(n)}{nm/2}\right)^{-1} \left(\frac{1 - 2\beta(n)/nm}{1 + 2\tau_n\beta(n)/nm}\right) \to \overline{\omega} = 1\,.$$

Clearly, under the assumptions of Lemma 2 and that π_n concentrates on $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$, Assumptions 1 and 2 are satisfied and we can apply Theorem 2 or Theorem 3 with $\overline{\omega} = 1$, depending on whether the spectral gaps are bounded away from 0 or not. This gives an asymptotic ordering between $P_{\text{MH},n}$ and $P_{\text{rev.},n}$, and thus between $P_{\text{MH},n}$ and $P_{\rho,n}$.

It is expected that lifted samplers only have an advantage when there is room for persistent movement, meaning that they can explore the state-space by using paths of considerable lengths. The analysis conducted in the current section shows that lifted samplers using locally-balanced proposal distributions are expected to have an advantage when, additionally, the mass does not vary much from a directional neighbourhood to another on the subset on which π_n concentrates. These samplers are expected to be comparable to their MH counterparts when, on the subset, the normalizing constants $c_{n,-1}(\mathbf{x})$ and $c_{n,+1}(\mathbf{x})$ are bounded by $c_n(\mathbf{x})/\kappa \pm \beta(n)$ with $\kappa > 2$.

5. Simulation of a simple Ising model: A case study

The sampling method developed in Section 3 and results presented in Section 4 are illustrated through several examples. In this section, we proceed by studying a simple Ising model that allows for an explicit definition of $\tilde{\boldsymbol{\mathcal{X}}}_n$ and $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$ when using locally-balanced samplers, and a verification of the assumptions of Theorem 3. As mentioned in Section 1.4, we study in Section 1 of the Supplementary Material (Gagnon and Maire (2024)) more complex problems (including the simulation of a Ising model which is more complex) for which an explicit definition of $\tilde{\boldsymbol{\mathcal{X}}}_n$ and $\tilde{\boldsymbol{\mathcal{X}}}_n^\circ$ and a verification of the assumptions is beyond the scope of the manuscript.

The model that we study here is the following:

$$\pi_n(\mathbf{x}) = \frac{1}{Z_n} \exp\left(\sum_{i=1}^n \alpha_i x_i\right), \quad \mathbf{x} = (x_1,\ldots,x_n) \in \{-1,+1\}^n, \tag{19}$$

where Z_n is the normalizing constant and $\alpha_i \in \mathbb{R}$. This model can be thought of as an Ising model with a single parameter $\boldsymbol{\alpha}_n := (\alpha_1,\ldots,\alpha_n)$ which is often referred to as the *external field*. This parameter essentially tends to polarize each spin. The difference with classical Ising models like that in Section 1 of the Supplementary Material (Gagnon and Maire (2024)) is that the model defined in (19) does not

possess a spatial-correlation parameter. We can think of this model as being defined on a square lattice (with $x_1, \ldots, x_\eta$ being the values of the components on the first line, $x_{\eta+1}, \ldots, x_{2\eta}$ being the values of the components on the second line, and so on), but by omitting the spatial correlation, the form on which the model is defined is actually not important. As mentioned in Section 1.4, this simplified model can be seen as an approximation to the high temperature model. A common problem in statistical physics is to estimate the average *magnetisation* of an Ising model, the magnetisation being defined as the mapping $\mathbf{x} \mapsto \sum_{i=1}^{n} x_i$.

For the study conducted here, we consider the following simplified situation: n is even, $\alpha_i = \pm c$ with c a positive constant, and $|\{i : \alpha_i = -c\}| = |\{i : \alpha_i = c\}| = n/2$, implying that the number of elements in the external field with the value $-c$ is the same as the number of elements with the value c. In our study, we focus on locally-balanced samplers and consider to simplify that g is a monotonically increasing function, which is the case for the two functions mentioned in Section 4.2, namely $g(x) = \sqrt{x}$ and $g(x) = x/(1+x)$.

In the simplified situation described above, we have that

$$\begin{aligned}\pi_n(\mathbf{x}) \propto \exp\left(\sum_{\{i:\alpha_i x_i = +c\}} c + \sum_{\{i:\alpha_i x_i = -c\}} -c \right) &= \exp\left(c(|\{i : \alpha_i x_i = +c\}| - |\{i : \alpha_i x_i = -c\}|)\right) \\ &= \exp\left(c(n - 2|\{i : \alpha_i x_i = -c\}|)\right) \\ &\propto \exp\left(-2c|\{i : \alpha_i x_i = -c\}|\right). \end{aligned} \tag{20}$$

From the expression in (20), we easily deduce that the mode, denoted by $\mathbf{x}^*$, is such that $|\{i : \alpha_i x_i^* = -c\}| = 0$, and that all the other values of π_n are characterized by $|\{i : \alpha_i x_i = -c\}|$. Let us define $d(\mathbf{x}) := |\{i : \alpha_i x_i = -c\}| \in \{0, \ldots, n\}$, which can be seen as a distance from the mode. We make the dependence on n implicit to simplify. With the expression in (20), we have a better understanding of the model and how to compute probabilities of different events.

To motivate the use of our weak Peskun ordering for a comparison between the lifted and MH samplers, we provide a result about an inequality on the transition probabilities when considering the whole state-space.

Proposition 2. *Within the framework described in this section, we have the following lower bound:*

$$P_{\text{rev.},n}(\mathbf{x}, \mathbf{y}) \geq (1/2) P_{\text{MH},n}(\mathbf{x}, \mathbf{y}),$$

for all $(\mathbf{x}, \mathbf{y}) \in \boldsymbol{\mathcal{X}}_n^2$ *with* $\mathbf{x} \neq \mathbf{y}$*, and for all n. Also, we have the following upper bound:*

$$P_{\text{rev.},n}(\mathbf{x}, \mathbf{y}) \leq (1/2) P_{\text{MH},n}(\mathbf{x}, \mathbf{y}) \left(1 - \frac{g(\exp(2c)) - g(\exp(-2c))}{n g(\exp(-2c))}\right)^{-1},$$

for certain $(\mathbf{x}, \mathbf{y}) \in \boldsymbol{\mathcal{X}}_n^2$ *with* $\mathbf{x} \neq \mathbf{y}$*, when* $n > \exp(2c) - 1$*. It is thus essentially not possible to obtain a better lower bound than that above.*

Proposition 2 implies that the ordering based on Lemma 33 of Andrieu, Lee and Vihola (2018) is the following:

$$\operatorname{var}(f_n, P_{\rho,n}) \leq \operatorname{var}(f_n, P_{\text{rev.},n}) \leq 2 \operatorname{var}(f_n, P_{\text{MH},n}) + 1, \tag{21}$$

for any $f_n \in \mathcal{L}_{0,1}^2(\pi_n)$.

We now turn to an analysis with an objective of applying our weak Peskun ordering. Our analysis allows to show that we can obtain tighter bounds on asymptotic variances when focusing on a subset of the state-space. The first step of such an analysis is to define $\tilde{\mathcal{X}}_n$ and understand which states belong to $\tilde{\mathcal{X}}_n^\circ$. We thus start with a result which will motivate a simple and explicit definition of $\tilde{\mathcal{X}}_n$ that we will connect to that in (18), and from which an explicit characterization of $\tilde{\mathcal{X}}_n^\circ$ will be easily deduced.

Proposition 3. *Within the framework described in this section, we have that for any* $\mathbf{x}$ *and* n*,*

$$1 - \frac{(g(\exp(2c)) + g(\exp(-2c)))\, d(\mathbf{x})/2}{c_n(\mathbf{x})/2} \leq \frac{c_{n,\nu}(\mathbf{x})}{c_n(\mathbf{x})/2} \leq 1 + \frac{(g(\exp(2c)) + g(\exp(-2c)))\, d(\mathbf{x})/2}{c_n(\mathbf{x})/2}.$$

Proposition 3 indicates that setting $\tilde{\mathcal{X}}_n := \{\mathbf{x} : d(\mathbf{x}) \leq \lfloor \phi(n) \rfloor\}$ with ϕ a monotonically increasing function allows to have a control on the ratio of normalizing constants of $q_{\mathbf{x}}$ and $q_{\mathbf{x},\nu}$, and thus on the difference between $P_{\text{rev.},n}$ and $P_{\text{MH},n}$. In particular, it allows to verify the inequality in (18) with $\beta(n) = (g(\exp(2c)) + g(\exp(-2c)))\, \lfloor \phi(n) \rfloor / 2$, even though $\tilde{\mathcal{X}}_n$ is not defined as in (18). This is because

$$\tilde{\mathcal{X}}_n \subset \{\mathbf{x} \in \mathcal{X}_n : 1 - \beta(n)/(c_n(\mathbf{x})/2) \leq c_{n,\nu}(\mathbf{x})/(c_n(\mathbf{x})/2) \leq 1 + \beta(n)/(c_n(\mathbf{x})/2)\}.$$

From our definition of $\tilde{\mathcal{X}}_n$, we can deduce that $\tilde{\mathcal{X}}_n^\circ = \{\mathbf{x} : d(\mathbf{x}) \leq \lfloor \phi(n) \rfloor - 1\}$. With those characterizations of $\tilde{\mathcal{X}}_n$ and $\tilde{\mathcal{X}}_n^\circ$, we easily understand which states belong to those subsets (comparatively to the definition of $\tilde{\mathcal{X}}_n$ in (18) and that of $\tilde{\mathcal{X}}_n^\circ$ that follows from it), and thus how to compute probabilities like $1 - \pi_n(\tilde{\mathcal{X}}_n^\circ)$.

Now that we have define the subsets $\{\tilde{\mathcal{X}}_n\}$, from which $\{\tilde{\mathcal{X}}_n^\circ\}$ are deduced, the next step is to verify whether the mass concentrates on $\tilde{\mathcal{X}}_n^\circ$ (Assumption 2). In our framework, $\tilde{\mathcal{X}}_n^\circ$ depends on the definition of $\phi(n)$. We present a result which indicates how to set $\phi(n)$ to obtain a mass concentration on $\tilde{\mathcal{X}}_n^\circ$.

Proposition 4. *Within the framework described in this section, if* $\lfloor \phi(n) \rfloor \leq n \frac{\exp(-2c)}{1+\exp(-2c)}$*, then* $1 - \pi_n(\tilde{\mathcal{X}}_n^\circ)$ *does not converge to 0. If* $n > \lfloor \phi(n) \rfloor > n \frac{\exp(-2c)}{1+\exp(-2c)}$ *with* $\lfloor \phi(n) \rfloor / n - \frac{\exp(-2c)}{1+\exp(-2c)}$ *converging towards a positive constant, then* $1 - \pi_n(\tilde{\mathcal{X}}_n^\circ)$ *converges to 0 at an exponential rate.*

Proposition 4 indicates that setting $\phi(n) = o(n)$ does not allow for a mass concentration on $\tilde{\mathcal{X}}_n^\circ$. The result is thus somewhat negative as it prevents us to apply the results of Section 4, in particular Lemma 2, and forces us to exploit the structure of the current problem to establish a refined order between $P_{\text{rev.},n}$ and $P_{\text{MH},n}$ (Assumption 1). Proposition 4 indicates that, to obtain a mass concentration (at an exponential rate), we have to enlarge $\tilde{\mathcal{X}}_n$ and include states that are further away from the mode. We now establish a refined order between $P_{\text{rev.},n}$ and $P_{\text{MH},n}$ on $\tilde{\mathcal{X}}_n$, when setting $\phi(n) = n \frac{\exp(-2c)}{1+\exp(-2c)}(1+\varepsilon) + 1$ with $\varepsilon > 0$ arbitrarily small (which is essentially the best choice of $\phi(n)$ that ensures that $\lfloor \phi(n) \rfloor > n \frac{\exp(-2c)}{1+\exp(-2c)}$ with $\lfloor \phi(n) \rfloor / n - \frac{\exp(-2c)}{1+\exp(-2c)}$ converging towards a positive constant).

Proposition 5. *Within the framework described in this section and with* $\phi(n) = n \frac{\exp(-2c)}{1+\exp(-2c)}(1+\varepsilon) + 1$*, we have that*

$$P_{\text{rev.},n}(\mathbf{x}, \mathbf{y}) \geq \omega(n) P_{\text{MH},n}(\mathbf{x}, \mathbf{y}),$$

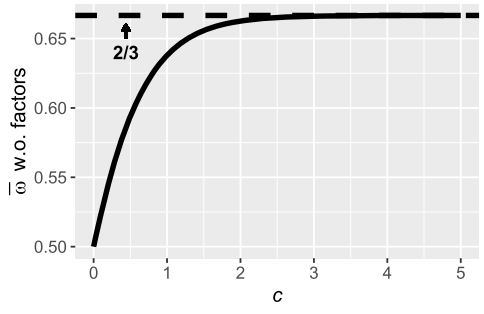

Figure 2. $\overline{\omega}$ as a function of c, without the factors $1+\varepsilon$.

with

$$\omega(n) \to \frac{1}{2} + \frac{1}{2}\frac{1 - 2\frac{\exp(-2c)(1+\varepsilon)}{1+\exp(-2c)}}{1 + 2\frac{1+\varepsilon}{1+\exp(-2c)}} = \overline{\omega},$$

for all $(\mathbf{x},\mathbf{y}) \in \tilde{\boldsymbol{X}}_n^2$ *with* $\mathbf{x} \neq \mathbf{y}$.

Proposition 5 highlights a dependence of $\overline{\omega}$ on the value of c: a smaller value of c yields a larger $\phi(n)$ which results in a larger subset $\tilde{\boldsymbol{X}}_n$ and a possibility of more unbalanced ratios of normalizing constants of $q_{\mathbf{x}}$ and $q_{\mathbf{x},\nu}$, and vice versa. We present in Figure 2 $\overline{\omega}$ as a function of c, without the factors $1+\varepsilon$ that can be made arbitrarily close to 1.

Provided that the spectral gaps of $P_{\text{rev.},n}$, $\tilde{P}_{\text{rev.},n}$, $P_{\text{MH},n}$ and $\tilde{P}_{\text{MH},n}$ do not decrease too quickly as n increases (a result about that follows), Proposition 4 together with Proposition 5 ensure that Theorem 3 can be applied for a class of functions, yielding

$$\text{var}(f_n, P_{\rho,n}) \leq \text{var}(f_n, P_{\text{rev.},n}) \leq \frac{1}{\overline{\omega} - \epsilon}\text{var}(f_n, P_{\text{MH},n}) + \frac{1}{2}\left(\frac{1}{\overline{\omega} - \epsilon} + \frac{1}{\overline{\omega}}\right) - 1 + \frac{\epsilon}{2},$$

for any $\epsilon \in (0, \overline{\omega})$, provided that n is large enough. When c is large enough, we essentially have an upper bound of $(3/2)\text{var}(f_n, P_{\text{MH},n}) + 1/2$, comparatively to what is obtained in (21).

The advantage of this example is that it is standard, easy to understand, and simple enough to prove mass-concentration results, precise orderings between $P_{\text{rev.},n}$ and $P_{\text{MH},n}$, and spectral-gap bounds. This simplicity follows from an independence between the components of $\mathbf{x}$ and the steady decrease in mass by a factor of $\exp(-2c)$ as getting away from the mode, regardless of which components of $\mathbf{x}$ are flipped and become misaligned with the external field. This steady, but relatively slow, decrease in mass forces us to set ϕ to be (essentially) proportional to n. This in turn leads to large subsets $\tilde{\boldsymbol{X}}_n$ and thus an improvement in terms of orderings between $P_{\text{rev.},n}$ and $P_{\text{MH},n}$ which is not optimal, i.e., with $\overline{\omega} < 1$. In Section 4 of the Supplementary Material (Gagnon and Maire (2024)), we *construct* an example (thus an example that is less standard and simple) in which we are able to achieve $\overline{\omega} = 1$ by applying the results of Section 4, in particular Lemma 2.

We now present the last piece of evidence that Theorem 3 can be applied. More specifically, we present a result about lower bounds on the spectral gaps of $P_{\text{rev.},n}$ and $P_{\text{MH},n}$.

Proposition 6. *Within the framework described in this section,* $P_{\text{rev.},n}$ *and* $P_{\text{MH},n}$ *have spectral gaps with lower bounds that decrease to 0 as n increases at a rate of* $n \log n$.

While we do not prove a result about the spectral gaps of $\tilde{P}_{\text{rev.},n}$ and $\tilde{P}_{\text{MH},n}$, there is no reason to believe that these decrease in another regime (for instance, with an exponential rate) given the definition of the sequence of subsets $\{\tilde{\boldsymbol{\mathcal{X}}}_n\}$. For small values of n, we computed the spectral gaps through a spectral decomposition of $\tilde{P}_{\text{rev.},n}$ and $\tilde{P}_{\text{MH},n}$ and the observed rate was polynomial.

To summarize, the analysis in this section shows that Theorem 3 can be applied for any $\delta > 0$ and $\gamma \in (0, \delta/(2+\delta))$, when considering the class of functions f_n with a $(2+\delta)$-norm that grows polynomially with n or slower. An example of functions which satisfies this condition is the standardized version of the magnetisation $\mathbf{x} \mapsto \sum_{i=1}^n x_i$, as indicated by Proposition 7 below.

Proposition 7. *Let f_n be the standardized version of the mapping $\mathbf{x} \mapsto \sum_{i=1}^n x_i$. Within the framework described in this section, $\|f_n\|_{\pi_n,4} \to 3^{1/4}$ as $n \to \infty$.*

6. Discussion

In this paper, we have introduced a weaker version of the celebrated Peskun ordering (Peskun, 1973) and have used it to analyse a class of lifted samplers designed to sample from distributions whose supports are partially-ordered discrete state-spaces. The weaker ordering does not require to establish a relationship between the Markov kernels on the whole state-space; it is only required to establish a relationship on a subset of the state-space, but the order between the asymptotic variances holds asymptotically, as a varying parameter grows without bound, as long as the mass concentrates on the subset (and provided that performance guarantees hold). This weaker requirement turned out to be useful to analyse some aspects of the lifted samplers and in particular how they compare to their MH counterparts. We have also shown that these lifted samplers can be straightforwardly implemented, at no additional computational cost and complexity, whenever a partial ordering on $\boldsymbol{\mathcal{X}}_n$ can be established.

The main contribution of our analysis of the lifted samplers in Section 4 is to provide insights into the situations in which they are expected to outperform their MH counterparts, and also into those in which there is no guarantee. The analysis conducted shows that lifted samplers are expected to have an advantage when the mass does not vary much from a directional neighbourhood to another on the subset on which π_n concentrates and when that subset allows the samplers to experience constant-momentum excursions. It is when they experience constant-momentum excursions of considerable lengths that the lifted samplers shine. While this point was reasonably well understood by the MCMC community, the merit of that part of our research presented in Section 4 has been to provide a rigorous analysis framework, which, de facto, can be used to study similar problems, perhaps some for which one does not have a clear intuition. Our analysis was conducted under a general framework, without focusing on specific statistical models or systems, explaining why we were not in a position to explicitly verify the assumptions of Theorems 2 and 3. We dug deeper and provided a thorough analysis in a context of simulation of a simple Ising model in Section 5, where the normalizing constants $c_{n,\nu}(\mathbf{x})$ and $c_n(\mathbf{x})$ have simple expressions, to take the study of lifted samplers one step further and to provide a concrete example of verification of the assumptions of Theorem 3.

One of the shortcomings of the application of our theoretical results to lifted samplers is that it does not give any quantitative measurement of the improvement offered by a lifted sampler over its MH counterpart when estimating $\pi_n f_n$, meaning that they are not such that $\text{var}(f_n, P_{\rho,n}) \leq (1/\omega_n)\text{var}(f_n, P_{\text{MH},n}) + \text{error}$ for some $\omega_n > 1$. Indeed, our analysis only allows to establish an inequality, but in the case where (essentially) $\omega_n \leq 1$. This a consequence of the route we followed to compare the asymptotic variances of the lifted and MH samplers:

$$\text{var}(f_n, P_{\rho,n}) \leq \text{var}(f_n, P_{\text{rev.},n}) \leq \frac{1}{\omega_n}\,\text{var}(f_n, P_{\text{MH},n}) + \frac{1}{\omega_n} - 1 + \text{error}.$$

In particular, no quantitative reduction factor is provided in the first inequality, which is expected given that this inequality holds in great generality (for any f_n and any π_n). Given that $P_{\text{rev.},n}$ and $P_{\text{MH},n}$ are, at best, similar and in fact, as mentioned in Section 2, ω_n is usually larger than one, a way to have a quantitative variance improvement factor is to obtain a different inequality between $\text{var}(f_n, P_{\rho,n})$ and $\text{var}(f_n, P_{\text{rev.},n})$ by leveraging an advantageous structure of the target distribution when it exists. We believe that this is possible, yet difficult, as the analysis needs to take into account the time duration of constant-momentum excursions conducted by the lifted sampler. This typically involves an analysis of k-step transition kernels with $k > 1$ because it is only after k transitions starting from a state $\mathbf{x}$ that we start to see a significant difference between lifted samplers and their non-lifted and MH counterparts.

Our work can also be extended in another direction: the theoretical result can be generalized to general state-spaces and the lifted samplers can be applied in cases where there exist partial orders on these general state-spaces. However, our proofs implicitly assume that the Markov kernels are uniformly ergodic and it would be interesting to see how this assumption can be relaxed.

A methodological question which has been unaddressed in the paper is that of the choice of the partial order. If a specific state-space admits a partial order, it needs not be unique and its choice may significantly impact the sampler. Indeed, some choices may guarantee more than others those aforementioned constant-momentum excursions. If specifically interested in the estimation of $\pi_n f_n$ for a particular f_n, one could also design the partial order based on f_n, in the spirit of Faizi, Deligiannidis and Rosta (2020).

Finally, in terms of applications of the theoretical work on the weak Peskun ordering, it would be interesting to consider the particular case of Bayesian models where a Bernstein von-Mises theorem holds. Comparing two MCMC methods sampling from the corresponding posterior distribution, our result suggests that one only needs to compare those samplers locally around a realization of a consistent parameter estimator. A question that naturally arises in this context is: is it possible to have a precise estimate of the sample size beyond which the approximate asymptotic-variance ordering holds? From a methodological standpoint this would motivate the design of samplers that are particularly efficient near the parameter estimate, perhaps at the expense of their behaviour in the tails of the distribution.

Acknowledgements

The authors thank two anonymous referees for constructive comments that led to an improved manuscript.

Funding

Philippe Gagnon acknowledges support from NSERC (Natural Sciences and Engineering Research Council of Canada) and FRQNT (Fonds de recherche du Québec – Nature et technologies). Florian Maire acknowledges support from NSERC.

Supplementary Material

Supplement to "An asymptotic Peskun ordering and its application to lifted samplers" (DOI: 10.3150/23-BEJ1674SUPP; .pdf). Numerical experiments to compare lifted samplers with MH ones. Introduction of a lifted trans-dimensional sampler. Proofs of the theoretical results. Example of a model such that (12) is satisfied. An example to illustrate how a careful application of Theorem 3 can allow to conclude that the lifted Markov chain is more efficient than the MH one in certain situations, provided that n is sufficiently large.

References

Andrieu, C., Lee, A. and Vihola, M. (2018). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli* **24** 842–872. MR3706778 https://doi.org/10.3150/15-BEJ785

Andrieu, C. and Livingstone, S. (2021). Peskun–Tierney ordering for Markovian Monte Carlo: Beyond the reversible scenario. *Ann. Statist.* **49** 1958–1981. MR4319237 https://doi.org/10.1214/20-aos2008

Atchadé, Y.F. (2021). Approximate spectral gaps for Markov chain mixing times in high dimensions. *SIAM J. Math. Data Sci.* **3** 854–872. MR4303259 https://doi.org/10.1137/19M1283082

Barker, A.A. (1965). Monte Carlo calculations of the radial distribution functions for a proton–electron plasma. *Aust. J. Phys.* **18** 119–134.

Bierkens, J. (2016). Non-reversible Metropolis–Hastings. *Stat. Comput.* **26** 1213–1228. MR3538633 https://doi.org/10.1007/s11222-015-9598-x

Chen, F., Lovász, L. and Pak, I. (1999). Lifting Markov chains to speed up mixing. In *Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999)* 275–281. New York: ACM. MR1798046 https://doi.org/10.1145/301250.301315

Diaconis, P., Holmes, S. and Neal, R.M. (2000). Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.* **10** 726–752. MR1789978 https://doi.org/10.1214/aoap/1019487508

Faizi, F., Deligiannidis, G. and Rosta, E. (2020). Efficient irreversible Monte Carlo samplers. *J. Chem. Theory Comput.* **16** 2124–2138. https://doi.org/10.1021/acs.jctc.9b01135

Gagnon, P. and Doucet, A. (2021). Nonreversible jump algorithms for Bayesian nested model selection. *J. Comput. Graph. Statist.* **30** 312–323. MR4270506 https://doi.org/10.1080/10618600.2020.1826955

Gagnon, P. and Maire, F. (2024). Supplement to "An asymptotic Peskun ordering and its application to lifted samplers." https://doi.org/10.3150/23-BEJ1674SUPP

Gustafson, P. (1998). A guided walk Metropolis algorithm. *Stat. Comput.* **8** 357–364.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

Herschlag, G., Mattingly, J.C., Sachs, M. and Wyse, E. (2020). Non-reversible Markov chain Monte Carlo for sampling of districting maps. arXiv:2008.07843.

Horowitz, A.M. (1991). A generalized guided Monte Carlo algorithm. *Phys. Lett. B* **268** 247–252.

Kamatani, K. and Song, X. (2023). Non-reversible guided Metropolis kernel. *J. Appl. Probab.* **60** 955–981. MR4624051 https://doi.org/10.1017/jpr.2022.109

Kleijn, B.J.K. and Van der Vaart, A.W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381.

Livingstone, S. and Zanella, G. (2022). The Barker proposal: Combining robustness and efficiency in gradient-based MCMC. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 496–523. MR4412995 https://doi.org/10.1111/rssb.12482

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

Peskun, P.H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60** 607–612. MR0362823 https://doi.org/10.1093/biomet/60.3.607

Power, S. and Goldman, J.V. (2019). Accelerated sampling on discrete spaces with non-reversible Markov Processes. arXiv:1912.04681.

Sakai, Y. and Hukushima, K. (2016a). Irreversible simulated tempering. *J. Phys. Soc. Jpn.* **85** 104002.

Sakai, Y. and Hukushima, K. (2016b). Eigenvalue analysis of an irreversible random walk with skew detailed balance conditions. *Phys. Rev. E* **93** 043318. https://doi.org/10.1103/PhysRevE.93.043318

Syed, S., Bouchard-Côté, A., Deligiannidis, G. and Doucet, A. (2022). Non-reversible parallel tempering: A scalable highly parallel MCMC scheme. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 321–350. MR4412989

Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9. MR1620401 https://doi.org/10.1214/aoap/1027961031

van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

Yang, J. and Rosenthal, J.S. (2023). Complexity results for MCMC derived from quantitative bounds. *Ann. Appl. Probab.* **33** 1259–1300. MR4564431 https://doi.org/10.1214/22-aap1846

Zanella, G. (2020). Informed proposals for local MCMC in discrete spaces. *J. Amer. Statist. Assoc.* **115** 852–865. MR4107684 https://doi.org/10.1080/01621459.2019.1585255

Received February 2023 and revised September 2023