# A fast non-reversible sampler for Bayesian finite mixture models

Filippo Ascolani[*] and Giacomo Zanella[†]

October 6, 2025

## Abstract

Finite mixtures are a cornerstone of Bayesian modelling, and it is well-known that sampling from the resulting posterior distribution can be a hard task. In particular, popular reversible Markov chain Monte Carlo schemes are often slow to converge when the number of observations $n$ is large. In this paper we introduce a novel and simple non-reversible sampling scheme for Bayesian finite mixture models, which is shown to drastically outperform classical samplers in many scenarios of interest, especially during convergence phase and when components in the mixture have non-negligible overlap. At the theoretical level, we show that the performance of the proposed non-reversible scheme cannot be worse than the standard one, in terms of asymptotic variance, by more than a factor of four; and we provide a scaling limit analysis suggesting that the non-reversible sampler can reduce the convergence time from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. We also discuss why the statistical features of mixture models make them an ideal case for the use of non-reversible discrete samplers.

[*]Duke University, Department of Statistical Science, Durham, NC, United States (filippo.ascolani@duke.edu)

[†]Bocconi University, Department of Decision Sciences and BIDSA, Milan, Italy (giacomo.zanella@unibocconi.it)

# 1 Introduction

## 1.1 Bayesian finite mixture models

Let $K \in \mathbb{N}$ and consider a finite mixture model (Marin et al., 2005; Frühwirth-Schnatter, 2006; McLachlan et al., 2019) defined as

$$
\begin{aligned}
Y_i \mid \boldsymbol{\theta}, \boldsymbol{w} &\overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k f_{\theta_k}(\cdot) & i &= 1, \ldots, n \\
\theta_k &\overset{\text{i.i.d.}}{\sim} p_0 & k &= 1, \ldots, K \\
\boldsymbol{w} &\sim \mathrm{Dir}(\boldsymbol{\alpha}),
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, $\boldsymbol{w} = (w_1, \ldots, w_K)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and $\mathrm{Dir}(\boldsymbol{\alpha})$ denotes the Dirichlet distribution on the $(K-1)$-dimensional simplex $\Delta_{K-1} \subset \mathbb{R}^K$ with parameters $\boldsymbol{\alpha}$. Here $f_\theta$ is a probability density on a space $\mathcal{Y}$ depending on a parameter $\theta \in \Theta \subset \mathbb{R}^d$, to which a prior distribution with density $p_0$ is assigned. For example, one could have $\mathcal{Y} = \Theta = \mathbb{R}^d$ and $f_\theta(y) = N(y \mid \theta, \Sigma)$ for some fixed $\Sigma$, where $N(y \mid \theta, \Sigma)$ denotes the density of the multivariate normal with mean vector $\theta$ and covariance matrix $\Sigma$ at a point $y$.

A popular strategy to perform posterior computations with model (1) (also for maximum likelihood estimation, as originally noted in Dempster et al. (1977)) consists in augmenting the model as follows

$$
Y_i \mid c, \boldsymbol{\theta}, \boldsymbol{w} \overset{\text{i.i.d.}}{\sim} f_{\theta_{c_i}}(y), \quad c_i \mid \boldsymbol{\theta}, \boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(\boldsymbol{w}), \quad \boldsymbol{w} \sim \mathrm{Dir}(\boldsymbol{\alpha}), \quad \theta_k \overset{\text{i.i.d.}}{\sim} p_0,
\tag{2}
$$

where $c = (c_1, \ldots, c_n) \in [K]^n$, with $[K] = \{1, \ldots, K\}$, is the set of allocation variables and $\mathrm{Cat}(\boldsymbol{w})$ denotes the Categorical distribution with probability weights $\boldsymbol{w}$. Given a sample $Y = (Y_1, \ldots, Y_n)$, the joint posterior distribution of $(c, \boldsymbol{\theta}, \boldsymbol{w})$ then reads

$$
\begin{aligned}
\pi(c, \boldsymbol{\theta}, \boldsymbol{w}) &\propto \left[ \prod_{i=1}^{n} w_{c_i} f_{\theta_{c_i}}(Y_i) \right] \mathrm{Dir}(\boldsymbol{w} \mid \boldsymbol{\alpha}) \prod_{k=1}^{K} p_0(\theta_k) \\
&\propto \left[ \prod_{k=1}^{K} w_k^{n_k(c) + \alpha_k - 1} \right] \prod_{k=1}^{K} \prod_{i \,:\, c_i = k} f_{\theta_k}(Y_i) p_0(\theta_k),
\end{aligned}
\tag{3}
$$

where $n_k(c) = \sum_{i=1}^{n} \mathbb{1}(c_i = k)$ and $\mathbb{1}$ denotes the indicator function. In particular, it is possible to integrate out $(\boldsymbol{\theta}, \boldsymbol{w})$ from (3) to obtain the marginal posterior distribution of $c$ given by

$$
\pi(c) \propto \left[ \prod_{k=1}^{K} \Gamma\left(\alpha_k + n_k(c)\right) \right] \prod_{k=1}^{K} \int_{\Theta} \prod_{i \,:\, c_i = k} f_{\theta_k}(Y_i) p_0(\theta_k) \, \mathrm{d}\theta_k,
\tag{4}
$$

2

from which we deduce the so-called full-conditional distribution of $c_i$

$$(5) \qquad \pi(c_i = k \mid c_{-i}) \propto [\alpha_k + n_k(c_{-i})]\, p(Y_i \mid Y_{-i}, c_{-i}, c_i = k) \qquad\qquad k \in [K]$$

where $c_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$, $Y_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$ and

$$p(Y_i \mid Y_{-i}, c_{-i}, c_i = k) = \int_\Theta f_\theta(Y_i) \frac{\prod_{j \neq i : c_j = k} f_\theta(Y_j) p_0(\theta)}{\int_\Theta \prod_{j \neq i : c_j = k} f_{\theta'}(Y_j) p_0(\theta')\, \mathrm{d}\theta'}\, \mathrm{d}\theta$$

is the predictive distribution of observation $Y_i$ given $Y_{-i}$ and the allocation variables. If $p_0$ is conjugate with respect to the density $f_\theta$, then $p(Y_i \mid Y_{-i}, c_{-i}, c_i = k)$ and thus $\pi(c_i = k \mid c_{-i})$ are available in closed form. For example, if $f_\theta(y) = N(y \mid \theta, \Sigma)$ and $p_0(\theta) = N(\theta \mid \theta_0, \Sigma_0)$ it holds that $p(Y_i \mid Y_{-i}, c_{-i}, c_i = k) = N(Y_i \mid \bar{\mu}, \bar{\Sigma})$, where

$$\bar{\Sigma} = \Sigma + \left(\Sigma_0^{-1} + n_k(c_{-i})\Sigma^{-1}\right)^{-1}, \quad \bar{\mu} = \left(\Sigma_0^{-1} + n_k(c_{-i})\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\theta_0 + n_k(c_{-i})\Sigma^{-1}\bar{Y}_{k,-i}\right)$$

and $\bar{Y}_{k,-i} = n_k^{-1}(c_{-i}) \sum_{j \neq i : c_j = k} Y_j$. Analogous expressions are available for likelihoods in the exponential family, see e.g. Robert (2007, Sec.3.3) for details.

## 1.2 The Marginal Gibbs (MG) sampler

Most popular algorithms for finite mixture models are based on the augmentation in (2), see e.g. Diebolt and Robert (1994). Here we consider the random-scan[1] Gibbs sampler which iterates updates from $\pi(c_i \mid c_{-i})$ for randomly chosen $i \in [n]$. Its Markov kernel $P_{\mathrm{MG}}$ is defined as

$$P_{\mathrm{MG}}(c, c') = \frac{1}{n} \sum_{i=1}^{n} P_{\mathrm{MG},i}(c, c'), \qquad\qquad c, c' \in [K^n]$$

with $P_{\mathrm{MG},i}(c, c') = \delta_{c_{-i}}(c'_{-i})\pi(c_i \mid c_{-i})$. The associated pseudocode is given in Algorithm 1. We refer to $P_{\mathrm{MG}}$ as *marginal* sampler, since it targets the marginal posterior distribution of $c$ defined in (4). Once a sample from $\pi(c)$ is available, draws from $\pi(c, \boldsymbol{\theta}, \boldsymbol{w})$ can be obtained by sampling from $\pi(\boldsymbol{\theta}, \boldsymbol{w} \mid c)$, so that Algorithm 1 can be used to perform full posterior inference on $\pi(c, \boldsymbol{\theta}, \boldsymbol{w})$.

Being an irreducible and aperiodic Markov kernel on a finite space, $P_{\mathrm{MG}}$ is uniformly ergodic for every fixed $n$, see e.g. Levin and Peres (2017, Theorem 4.9) and Roberts and Rosenthal (2004, Sec.3.3) for discussion about uniform ergodicity. However, as we will see later on, its rate of convergence tends to deteriorate quickly as $n$ increases.

---

[1]Here we consider the random-scan strategy since it simplifies some of the proofs and comparisons below. We expect the behaviour of $P_{\mathrm{MG}}^n$, where $P^k = P \ldots P$ denotes the $k$-th power of a Markov kernel $P$, to be roughly comparable to the one of the deterministic-scan version (which updates $c_i$ for $i = 1, \ldots, n$ sequentially at each iteration) in most cases of interest, and thus stick to the random-scan for simplicity.

**Algorithm 1** (Marginal sampler $P_{\mathrm{MG}}$)

---

Initialize $c^{(0)} \in [K]^n$.

**for** $t \geq 1$ **do**

    Sample $i \sim \mathrm{Unif}(\{1, \ldots, n\})$, where Unif denotes the uniform distribution.

    Sample $c_i^{(t)} \sim \pi(c_i \mid c_{-i}^{(t-1)})$, with $\pi(c_i = k \mid c_{-i})$ as in (5), and set $c_{-i}^{(t)} = c_{-i}^{(t-1)}$.

**end for**

---

A popular alternative to the marginal sampler is the so-called *conditional* sampler introduced in Diebolt and Robert (1994), which directly targets $\pi(c, \boldsymbol{\theta}, \boldsymbol{w})$ defined in (3) alternating updates of $(\boldsymbol{\theta}, \boldsymbol{w}) \mid c$ and $c \mid (\boldsymbol{\theta}, \boldsymbol{w})$. We postpone the discussion of this algorithm to Section 3.2, since the latter is always dominated by $P_{\mathrm{MG}}$ in terms of mixing speed (see e.g. Proposition 3.3).
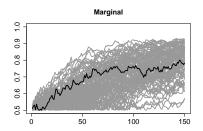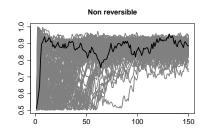
## 1.3  Illustrative example

It is well-known that $P_{\mathrm{MG}}$ can be slow to converge when $n$ is large (Celeux et al., 2000; Lee et al., 2009). As a first illustrative example, we take model (1) with $K = 2$, $f_\theta(y) = N(y \mid \theta, 1)$, $p_0(\theta) = N(\theta \mid 0, 1)$, $\boldsymbol{\alpha} = (0.5, 0.5)$, and we consider the posterior distribution given $(Y_1, \ldots, Y_n)$ generated as

$$(6) \qquad Y_i \overset{\text{i.i.d.}}{\sim} 0.9N(0.9, 1) + 0.1N(-0.9, 1), \qquad\qquad i = 1, \ldots, n$$

with $n = 2000$. This is a relatively simple one-dimensional problem, with data generated from two components with a reasonable degree of separation between them (the two means are almost two standard deviations away from each other).

We use $P_{\mathrm{MG}}$ to sample from the resulting posterior $\pi(c)$, leading to a Markov chain $\{c^{(t)}\}_{t=0,1,2,\ldots}$ on $[K]^n$. The left panel of Figure 1 displays 100 independent traceplots of the size of the largest cluster in $\{c^{(t)}\}_t$, with all chains independently initialized by sampling $c^{(0)}$ uniformly from $[K]^n$. Most runs are still far from 0.9 (value around which we expect the posterior to concentrate) after $150 \times n$ iterations. Indeed trajectories exhibit a typical random-walk behaviour, with slow convergence to stationarity. The center panel instead shows the traceplots generated by the same number of runs and iterations of $P_{\mathrm{NR}}$, the non-reversible scheme we propose in Section 2.2 (see Algorithm 6 therein for pseudocode and full details). The chain appears to reach the high-probability region and forget the starting configuration much faster. This is also clear from the right panel, which displays empirical estimates of the marginal distributions of the Markov chains induced by $P_{\mathrm{MG}}$ and $P_{\mathrm{NR}}$ over time.
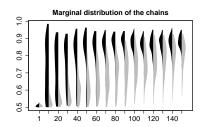
4

Figure 1: Left and center: traceplots of 100 independent runs of the size of the largest cluster for 150 iterations (after a thinning of size $n$, i.e. after $150 \times n$ total updates) of $P_{\mathrm{MG}}$ (left) and $P_{\mathrm{NR}}$ (center) in Algorithm 6. Right: empirical estimates of the marginal distribution at every $10 \times n$ iterations for $P_{\mathrm{MG}}$ (gray) and $P_{\mathrm{NR}}$ (black). The model is the one in (1) with $K = 2$, $f_\theta(y) = N(y \mid \theta, 1)$, $p_0(\theta) = N(\theta \mid 0, 1)$ and $\boldsymbol{\alpha} = (0.5, 0.5)$, while the data are generated as in (6). Initial configurations $c^{(0)}$ are sampled uniformly from $[K]^n$.

## 1.4 Lifted samplers for discrete spaces

Our proposed sampler is inspired by classical non-reversible MCMC constructions (Diaconis et al., 2000; Fearnhead et al., 2018), which loosely speaking force the algorithm to persistently move in one direction as much as possible. To illustrate the idea, consider an arbitrary probability distribution $\pi$ on a countable space $\mathcal{C}$ and an augmented distribution

$$\tilde{\pi}(c, v) = \frac{\pi(c)}{2}, \quad c \in \mathcal{C}, v \in \{-1, +1\},$$

on the space $\mathcal{X} = \mathcal{C} \times \{-1, +1\}$, so that $\pi$ is the marginal distribution of $\tilde{\pi}$ over $\mathcal{C}$. Given two Markov kernels $\{q_{+1}(c, \cdot)\}_{c \in \mathcal{C}}$ and $\{q_{-1}(c, \cdot)\}_{c \in \mathcal{C}}$ on $\mathcal{C}$, let $P_{\mathrm{lift}}$ be the non-reversible $\tilde{\pi}$-invariant Markov kernel defined in Algorithm 2. The kernels are usually

---

**Algorithm 2** Generating a sample $(c', v') \sim P_{\mathrm{lift}}((c, v), \cdot)$

Generate $\tilde{c} \sim q_v(c, \cdot)$.
Set $(c', v') = (\tilde{c}, v)$ with probability

$$\min\left\{ 1, \frac{\pi(\tilde{c}) q_{-v}(\tilde{c}, c)}{\pi(c) q_v(c, \tilde{c})} \right\}.$$

Otherwise set $(c', v') = (c, -v)$.

---

chosen so that $q_v(c, \cdot)$ and $q_{-v}(c, \cdot)$ have disjoint support and the variable $v \in \{-1, +1\}$ encodes a direction (or velocity) along which the chain is exploring the space: such direction is reversed only when a proposal is rejected (see Algorithm 2). As a simple

example, take $\mathcal{C} = \mathbb{N}$ and $q_v(c, \cdot) = \delta_{c+v}(\cdot)$, so that $v = +1$ implies that the chain is moving towards increasing values and viceversa with $v = -1$. Within this perspective $v$ can be seen as a "memory bank" which keeps track of the past history of the chain and introduces momentum. The kernel $P_{\text{lift}}$ is often referred to as a *lifted* version of the (reversible) Metropolis-Hastings kernel with proposal distribution $q = 0.5q_{+1} + 0.5q_{-1}$ and target distribution $\pi$. Lifted kernels can mix significantly faster than their reversible counterparts (Diaconis et al., 2000) and are in general at least as efficient as the original method under mild assumptions (Bierkens, 2016; Andrieu and Livingstone, 2021; Gagnon and Maire, 2024b). However, whether or not lifting techniques achieve a notable speed-up depends on the features of $\pi$ and the choice of $q_v$. For example, if proposed moves are often rejected, then the direction $v$ will be reversed frequently and the chain will exhibit an almost reversible behaviour; while if the sampler manages to make long 'excursions' (i.e. consecutive moves without flipping direction) one expects to observe significant gains obtained by lifting.

**Non-reversible samplers for mixture models** General techniques to construct non-reversible samplers for discrete spaces have been proposed in the literature, see e.g. Gagnon and Maire (2024a, Sec.3) for constructions on partially-ordered discrete spaces and Power and Goldman (2019) for discrete spaces with algebraic structures. We are, however, not aware of successful applications of these methodologies to mixture models. Part of the reason could be that, in order to build a lifted counter-part of $P_{\text{MG}}$ for mixture models, one would need to define some notion of direction or partial ordering on $[K]^n$, or more generally on the space of partitions and it is not obvious how to do so in a way that is computationally efficient and results in long excursions with persistence in direction (thus leading to substantial speed-ups). For example, one could directly rely on the cartesian product structure of $[K]^n$ and attach a velocity component to each coordinate, applying for example the discrete Hamiltonian Monte Carlo algorithm of Nishimura et al. (2020): this however would not pair well with the geometry of posterior distributions $\pi(c)$ arising in mixture models and likely result in short excursions and little speed-up.

To tackle this issue, we take a different perspective on $[K]^n$, moving from the kernel $P_{\text{MG}}$, which is a mixture over data-points (i.e. over $i \in [n]$), to a kernel $P_{\text{R}}$ which is a mixture over pairs of clusters (see Section 2.1 for definition). This allows us to derive an effective non-reversible sampler $P_{\text{NR}}$ targeting $\pi(c)$, built as a mixture of lifted samplers associated to pairs of clusters (see Section 2.2 for definition). Note that, while we designed our sampler to be effective for posterior distributions of mixture models, the proposed scheme can in principle be used with any distribution $\pi$ on $[K]^n$.

## 1.5 Related literature

**Bayesian mixture models** Bayesian finite mixture models are a classical topic which has received a lot of attention in the last decades, see Marin et al. (2005);

Frühwirth-Schnatter (2006) for some reviews. The challenges related to sampling from the resulting posterior distribution have been also discussed extensively, see e.g. early examples in (Diebolt and Robert, 1994; Celeux et al., 2000; Stephens, 2000; Lee et al., 2009; Hobert et al., 2011), and the marginal and conditional samplers we compare with are arguably the most popular schemes that are routinely used to accomplish such tasks (Marin et al., 2005, Section 1.4).

**Lifted MCMC for statistical models with discrete parameters**  Non-reversible MCMC samplers have been previously designed for and applied to Bayesian statistical models with discrete parameters, such as variable selection, permutation-based and graphical models; see e.g. Power and Goldman (2019); Gagnon and Maire (2024a); Schauer and Wienöbst (2024) and references therein. However, posterior distributions arising from such models are usually strongly concentrated and highly non-smooth, limiting the length of excursions and speed-ups obtained with lifted chains. As a result, one often ends up observing large gains (e.g. hundred-fold) when targeting uniform or prior distributions (which are usually quite flat) and more modest gains (e.g. two-fold) when targeting actual posterior distributions used in practice; see e.g. Schauer and Wienöbst (2024, Figures 1 and 3), Power and Goldman (2019, Table 1) or Gagnon and Maire (2024a, Figure 1)[2]. Instead, a key feature of our proposed sampler is that, in many cases of interest, the speed-up relative to its reversible counter-part remains large even in the presence of observed data (i.e. for the actual posterior). We argue that this is due to statistical features of mixture models that make them well-suited to appropriately designed non-reversible samplers (such as $P_{\mathrm{NR}}$); see Section 2.2.1 for more details.

## 1.6   Structure of the paper

In Section 2 we introduce our proposed non-reversible Markov kernel $P_{\mathrm{NR}}$, which targets $\pi(c)$ in (4). In Section 3 we first show that, after accounting for computational cost, $P_{\mathrm{NR}}$ cannot perform worse than $P_{\mathrm{MG}}$, in terms of asymptotic variance, by more than a factor of four. This is done by combining some variations of classical results on asymptotic variances of lifted samplers with a Peskun comparison between $P_{\mathrm{MG}}$ and an auxiliary reversible kernel $P_{\mathrm{R}}$. We then provide analogous results for the conditional sampler by showing that it is dominated by the marginal one (Section 3.2). In Section 4 we continue the comparison between $P_{\mathrm{MG}}$ and $P_{\mathrm{NR}}$, showing that in the prior case the latter improves on the former by an order of magnitude, i.e. reducing the convergence time from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. This is done through a scaling limit analysis, which

---

[2]This is in contrast with applications of lifting techniques to discrete models arising in Statistical Physics (see e.g. Vucelja, 2016), which often feature a higher degree of symmetry and smoothness, thus making non-reversible MCMC methods more effective; see e.g. Power and Goldman (2019, Table 1) for numerical examples and Faulkner and Livingstone (2024) for a recent review.

proves that, after rescaling time by a factor of $n^2$, the evolution of the frequencies $n_k(c)$ evolving according to $P_{\mathrm{MG}}$ converges to a Wright-Fisher process (Ethier, 1976), which is a diffusion on the probability simplex. In contrast, when the chain evolves according to $P_{\mathrm{NR}}$, we obtain convergence to a non-singular piecewise deterministic Markov process (Davis, 1984) after rescaling time by only a factor of $n$. Section 5 discusses a variant of $P_{\mathrm{NR}}$ and, finally, Section 6 provides various numerical simulations, where $P_{\mathrm{NR}}$ is shown to significantly outperform $P_{\mathrm{MG}}$ in sampling from mixture models posterior distributions, both in low and high-dimensional cases. The Supplementary Material contains additional simulation studies, together with the proofs of all the theoretical results. R code to replicate all the numerical experiments can be found at https://github.com/gzanella/NonReversible_FiniteMixtures.

# 2   A non-reversible marginal sampler

## 2.1   A reversible sampler that operates over pairs of clusters

Let $\pi(c)$ be an arbitrary probability distribution on $[K]^n$, such as (4) in the context of finite mixtures, and denote the set of ordered pairs in $[K]$, with cardinality $K(K-1)/2$, as

$$(7) \qquad \mathcal{K} = \left\{ (k, k') \in [K]^2 \,:\, k < k' \right\}.$$

As an intermediate step towards defining $P_{\mathrm{NR}}$, we first consider a $\pi$-reversible Markov kernel on $[K]^n$ defined as

$$(8) \qquad P_{\mathrm{R}}(c, c') = \sum_{(k,k')\in\mathcal{K}} p_c(k, k') P_{k,k'}(c, c') \qquad\qquad c, c' \in [K]^n,$$

where

$$(9) \qquad p_c(k, k') = \frac{n_k(c) + n_{k'}(c)}{(K-1)n}, \qquad\qquad (k, k') \in \mathcal{K}$$

is a probability distribution on $\mathcal{K}$ for each $c \in [K]^n$, i.e. $\sum_{(k,k')\in\mathcal{K}} p_c(k, k') = 1$, and $P_{k,k'}$ is the $\pi$-reversible Metropolis-Hastings (MH) kernel that proposes to move a uniformly drawn point from cluster $k$ to cluster $k'$ or viceversa with probability $1/2$. The resulting kernel $P_{\mathrm{R}}$ is defined in Algorithm 3 where, for ease of notation, for every $c \in [K]^n$, $i \in [n]$ and $k \in [K]$ we write $(c_{-i}, k) \in [K]^n$ for the vector $c$ with the $i$-th entry $c_i$ replaced by $k$.

Despite the fact that $P_{\mathrm{R}}$ is a mixture with weights $p_c$ depending on the current state $c$, invariance with respect to $\pi$ is preserved, as proven in the next lemma. The

**Algorithm 3** Generating a sample $c' \sim P_R(c, \cdot)$

Sample $(k, k') \sim p_c$ as in Algorithm 4.
Set $(k_-, k_+) = (k, k')$ with probability $1/2$ and $(k_-, k_+) = (k', k)$ otherwise
If $n_{k_-}(c) = 0$ set $c' = c$.
If $n_{k_-}(c) > 0$ sample $i \sim \text{Unif}(\{i' \in [n] : c_{i'} = k_-\})$ and set $c' = (c_{-i}, k_+)$ with probability $\min\{1, r(c, i, k_-, k_+)\}$ and $c' = c$ otherwise, where

$$(10) \qquad r(c, i, k_-, k_+) = \left(\frac{n_{k_-}(c)}{n_{k_+}(c) + 1}\right) \frac{\pi(c_i = k_+ \mid c_{-i})}{\pi(c_i = k_- \mid c_{-i})},$$

---

key point is that $p_c(k, k')$ only depends on $n_k(c) + n_{k'}(c)$ and $P_{k,k'}$ leaves the latter quantity unchanged.

**Lemma 2.1.** *The Markov kernel $P_R$ defined in Algorithm 3 is $\pi$-reversible. Moreover, if $\pi(c) > 0$ for every $c \in [K]^n$ it is also irreducible, aperiodic and uniformly ergodic.*

Sampling from $p_c$ can be performed efficiently using Algorithm 4, where one cluster is selected with probability proportional to its size and the other uniformly at random from the remaining ones. Validity is proved in the next lemma.

**Lemma 2.2.** *For each $c \in [K]^n$, Algorithm 4 produces a sample $(k, k')$ from the probability distribution $p_c$ defined in (9).*

---

**Algorithm 4** Sampling $(k, k') \sim p_c$ defined in (9)

Sample $k_1$ from $\{1, \dots, K\}$ with probabilities $(n_1(c)/n, \dots, n_K(c)/n)$
Sample $k_2$ uniformly at random from $\{1, \dots, K\}\backslash\{k_1\}$
Set $k = \min\{k_1, k_2\}$ and $k' = \max\{k_1, k_2\}$

---

**Comparison between $P_{MG}$ and $P_R$** Both $P_{MG}$ and $P_R$ can be interpreted as reversible Metropolis-Hastings schemes that propose single-point moves. Specifically, $P_{MG}$ and $P_R$ propose moving datapoint $i$ to cluster $k$ with, respectively, probabilities

$$a_{MG}(i, k) = \frac{\pi(c_i = k \mid c_{-i})}{n} \quad \text{and} \quad a_R(i, k) = \frac{n_{c_i}(c) + n_k(c)}{n_{c_i}(c)} \frac{\mathbb{1}(c_i \neq k)}{2(K-1)n},$$

for $(i, k) \in [n] \times [K]$. For $P_{MG}$ the Metropolis-Hastings acceptance probability is always one, while for $P_R$ it is not. It is interesting to note that

$$a_R(i, k) \geq \frac{1}{2(K-1)n} \geq \frac{1}{2(K-1)} a_{MG}(i, k),$$

9

meaning that the proposal probabilities of $P_{\mathrm{R}}$ can be at most $2(K-1)$ times smaller than the ones of $P_{\mathrm{MG}}$. This connection will help providing formal comparison results between $P_{\mathrm{R}}$ and $P_{\mathrm{MG}}$ in Section 3 (see Theorem 3.1 and Remark 3.2 for more details). We postpone details on these comparisons to Section 3 and now focus on how to leverage the mixture representation of $P_{\mathrm{R}}$ in (8) to build effective non-reversible algorithms targeting $\pi(c)$.

**Cost per iteration of $P_{\mathrm{MG}}$ and $P_{\mathrm{R}}$**   For both $P_{\mathrm{MG}}$ and $P_{\mathrm{R}}$ the cost per iteration is usually dominated by the computation of the conditional distribution $\pi(c_i = k \mid c_{-i})$ in (5), which will depend on the specific combination of kernel $f_\theta$ and prior $p_0$. Indeed, Algorithm 1 requires in addition only to sample from a uniform distribution on a discrete set (which has a fixed cost). Similar considerations hold for Algorithm 3, since sampling from $p_c$ with Algorithm 4 entails again only sampling from two uniform distributions. Thus, we measure cost per iteration of these samplers in terms of the number of conditional distribution evaluations, which is $K$ for $P_{\mathrm{MG}}$ and 2 for $P_{\mathrm{R}}$: therefore the ratio of costs of $P_{\mathrm{MG}}$ versus $P_{\mathrm{R}}$ is $K/2$. The same will hold for $P_{\mathrm{NR}}$ in Algorithm 6 below, which requires essentially the same computations of Algorithm 3.

## 2.2   The proposed non-reversible sampler

Consider the extended target distribution

$$(11) \quad \tilde{\pi}(c,v) := \pi(c) \left(\frac{1}{2}\right)^{K(K-1)/2} \qquad c \in [K]^n, \ v = (v_{k,k'})_{(k,k')\in\mathcal{K}} \in \{-1,+1\}^{K(K-1)/2}$$

and the $\tilde{\pi}$-invariant Markov kernel $P_{\mathrm{NR}}$ defined as

$$(12) \qquad P_{\mathrm{NR}}((c,v),(c',v')) = \sum_{(k,k')\in\mathcal{K}} p_c(k,k') \tilde{P}_{k,k'}((c,v),(c',v')) \,,$$

with $p_c$ defined as in (9) and $\tilde{P}_{k,k'}$ being the $\tilde{\pi}$-invariant kernel defined in Algorithm 5. The kernel $\tilde{P}_{k,k'}$ operates on the $k$-th and $k'$-th clusters, and it is a lifted counter-part of $P_{k,k'}$, with associated velocity component $v_{k,k'}$. In this construction, the velocity vector $v$ is $K(K-1)/2$ dimensional and only one of its component is actively used to move $c$ at each iteration. The pseudo-code associated to $P_{\mathrm{NR}}$ is given in Algorithm 6.

The algorithm depends on a parameter $\xi \geq 0$, which can be interpreted as the refresh rate at which directions are flipped. While useful to take $\xi > 0$ for technical reasons (i.e. to ensure aperiodicity), we do not expect the value of $\xi$ to have significant impacts in practice provided it is set to a small value, and in the simulations we always set $\xi = 1/2$.

The next lemma shows that $P_{\mathrm{NR}}$ is a valid $\tilde{\pi}$-invariant kernel.

10

---

**Algorithm 5** Generating a sample $(c', v') \sim \tilde{P}_{k,k'}((c,v), \cdot)$

---

With probability $\xi/n$ flip $v_{k,k'}$ to $-v_{k,k'}$

Set $(k_-, k_+) = (k, k')$ if $v_{k,k'} = +1$, and $(k_-, k_+) = (k', k)$ if $v_{k,k'} = -1$

If $n_{k_-}(c) = 0$, set $(c', v') = (c, v^{(flip)})$, with $v^{(flip)} = (v_{-(k,k')}, -v_{k,k'})$

If $n_{k_-}(c) > 0$, sample $i \sim \text{Unif}(\{i' \in [n] \mid c_{i'} = k_-\})$ and set $(c', v') = ((c_{-i}, k_+), v)$ with probability $\min\{1, r(c, i, k_-, k_+)\}$ and otherwise $(c', v') = (c, v^{(flip)})$, with $r(c, i, k_-, k_+)\}$ defined in (10)

With probability $\xi/n$ flip $v'_{k,k'}$ to $-v'_{k,k'}$

---

---

**Algorithm 6** One step of the non-reversible kernel $(c', v') \sim P_{\text{NR}}((c,v), \cdot)$

---

Sample $(k, k') \sim p_c$ as in Algorithm 4.

Sample $(c', v') \sim \tilde{P}_{k,k'}((c,v), \cdot)$ as in Algorithm 5.

---

**Lemma 2.3.** *For any probability distribution $\pi$ on $[K]^n$, the Markov kernel $P_{\text{NR}}$ defined in Algorithm 6 is $\tilde{\pi}$-invariant, with $\tilde{\pi}$ as in (11). Moreover, if $\xi > 0$ and $\pi(c) > 0$ for every $c \in [K]^n$, then $P_{\text{NR}}$ is irreducible, aperiodic and uniformly ergodic.*

### 2.2.1 Specificities of mixture models that make $P_{\text{NR}}$ work well

We now discuss at an informal level some of the specificities of the posterior distribution $\pi(c)$ arising from mixture models that make $P_{\text{NR}}$ work well.

**Lack of identifiability and concentration** An important statistical feature of mixture models is that cluster labels are in general not identifiable as $n \to \infty$, meaning that even when $n$ is large there is non-vanishing uncertainty on the value of $c_i$. In other words, while the posterior distribution of $\boldsymbol{w}$ and $\boldsymbol{\theta}$ concentrates as $n \to \infty$, the one of $c$ does not (not even at the level of marginals, meaning that, for every fixed $i$, $\pi(c_i)$ does not converge to a point mass as $n \to \infty$); see e.g. Nguyen (2013); Guha et al. (2021) and references therein for asymptotic results for mixture model posteriors. Intuitively, lack of concentration occurs because the information about each individual $c_i$ does not grow with $n$ (since each $c_i$ is associated to a single datapoint). This also tends to make posteriors flatter, i.e. moving one observation from one cluster to another usually leads to a small change in the target distribution. By contrast, many models with discrete parameters lead to posteriors that become increasingly more concentrated and rough as $n \to \infty$, which has a major impacts on the convergence properties of MCMC algorithms targeting them, including making standard MCMC converge faster (see e.g. Yang et al., 2016; Zhou et al., 2022; Zhou and Chang, 2023) and lifting techniques less effective (as already discussed in Section 1.5).

**Cancellations in the acceptance ratio**   For $\pi(c)$ as in (4), the MH ratio $r(c, i, k_-, k_+)$ reads

$$(13) \quad r(c, i, k_-, k_+) = \left( \frac{\alpha_{k_+} + n_{k_+}(c)}{n_{k_+}(c) + 1} \right) \left( \frac{n_{k_-}(c)}{\alpha_{k_-} + n_{k_-}(c) - 1} \right) \frac{p(Y_i \mid Y_{-i}, c_{-i}, c_i = k_+)}{p(Y_i \mid Y_{-i}, c_{-i}, c_i = k_-)}.$$

Interestingly, the proposal probability almost matches the term $\alpha_k + n_k(c)$ arising from the prior. In particular, by writing

$$\left( \frac{\alpha_{k_+} + n_{k_+}(c)}{n_{k_+}(c) + 1} \right) \left( \frac{n_{k_-}(c)}{\alpha_{k_-} + n_{k_-}(c) - 1} \right) = \left( 1 + \frac{\alpha_{k_+} - 1}{n_{k_+}(c) + 1} \right) \left( 1 + \frac{\alpha_{k_-} - 1}{n_{k_-}(c)} \right)^{-1},$$

we see that the first part of (13) goes to 1 as $n_{k_+}(c)$ and $n_{k_-}(c)$ increase, for every fixed value of $\boldsymbol{\alpha}$. Notice that with $\alpha_k = 1$ for every $k$ this ratio is always equal to 1. This cancellation contributes to make (13) closer to 1 and thus to make excursions of $P_{\mathrm{NR}}$ longer.

**Flatness in the tails and behavior out of stationarity**   Interestingly, also the ratio of predictive distributions in (13) tends to get close to 1 for partitions that do not correspond to well-identified and separate clusters, meaning that mixture model posteriors $\pi(c)$ become increasingly flatter in the tails. To illustrate this, consider the common situation when labels are initialized uniformly at random, i.e. $c_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([K])$. In this situations, by construction, clusters are similar to each other under $c^{(0)}$, resulting in ratios of predictive distributions that are close to 1 (and converge to 1 as $n \to \infty$). As a consequence, the non-reversible chain will proceed almost deterministically without flipping directions until clusters start to differentiate significantly. This is indeed the behavior observed in the middle panel of Figure 1, as well as in Section 6.2 and B of the Supplementary Material with different values of $K$ and likelihood kernels. More generally, in mixture model contexts, non-reversibility is particularly helpful during the transient phase, where the algorithm has not yet reached the high-probability region under $\pi$ and has to explore large flat regions of the state space[3].

**Overlapping components and the overfitted case**   Another situation that makes ratios of predictive distributions close to 1 is when the actual clusters in the data have a considerable overlap. An extreme case of this situation is when the true number of components $K^*$ (assuming data were actually generated by a well-specified mixture model) is strictly smaller than $K$, which amounts to assuming that a plausible upper bound on $K^*$ is known and $K$ is set to such value (instead of the less plausible scenario where $K^*$ itself is known). This is often called the overfitted case, see e.g. Rousseau

---

[3]This is, again, in contrast with typical Bayesian discrete models that lead to posteriors with large "discrete gradients" in the tails providing strong enough signal for reversible schemes to converge fast in the first part of the transient phase (Yang et al., 2016; Zhou et al., 2022; Zhou and Chang, 2023).

and Mengersen (2011) for a theoretical exploration of its asymptotic properties, and it is a common situation since in many context (e.g. density estimation) it is preferable to overshoot rather than undershoot the value of $K$ and thus practitioners often set $K$ to some conservative, moderately large value. See Section 6.4 for more discussion on the overfitted case and empirical evidence that in this setting the improvement of $P_{NR}$ over the latter is particularly apparent.

# 3 Asymptotic variance comparison results

In this section we compare the various kernels discussed above in terms of asymptotic variances. Among other results we show that, after accounting for computational cost, $P_{NR}$ cannot be worse than $P_{MG}$ by more than a factor of 4. Given a Markov chain $\{X_t\}_t$ with a $\pi$-invariant Markov kernel $P$ started in stationarity, the asymptotic variance of the associated MCMC estimator is given by

$$\mathrm{Var}(g, P) = \lim_{T \to \infty} T \mathrm{Var}\left(\frac{1}{T} \sum_{t=1}^{T} g(X_t)\right) = \mathrm{Var}_\pi(g) + 2 \sum_{t=1}^{\infty} \mathrm{Cov}\left(g(X_0), g(X_t)\right),$$

for every function $g$ such that $\mathrm{Var}_\pi(g)$ is well-defined.

## 3.1 Ordering of reversible and non-reversible schemes

The next theorem provides ordering results for the asymptotic variances of $P_{MG}$, $P_R$ and $P_{NR}$. Technically speaking these kernels are not directly comparable, since $P_{MG}$ and $P_R$ are defined on $[K]^n$ while $P_{NR}$ is defined on $[K]^n \times \{-1, +1\}^{K(K-1)/2}$. Nonetheless, we are only interested in estimating expectations of test functions that depend on $c$ alone, so that we can restrict attention to those, as usually done in non-reversible MCMC literature (Andrieu and Livingstone, 2021; Gagnon and Maire, 2024b). Given a test function $g : [K]^n \to \mathbb{R}$, with a slight abuse of notation, we also use $g$ in $\mathrm{Var}(g, P_{NR})$ to denote the function defined as $g(c, v) = g(c)$ for all $(c, v) \in [K]^n \times \{-1, +1\}^{K(K-1)/2}$.

**Theorem 3.1.** *Let $\pi$ be a probability distribution on $[K]^n$ and $g : [K]^n \to \mathbb{R}$. Then*

(14) $\quad Var(g, P_{NR}) \leq Var(g, P_R) \leq c(K) \, Var(g, P_{MG}) + [c(K) - 1] \, Var_\pi(g),$

*where $c(K) = 2(K - 1)$ and $Var_\pi(g)$ denotes $Var(g(X_0))$ for $X_0 \sim \pi$.*

Since in most realistic applications $\mathrm{Var}(g, P_{MG})$ is much larger than $\mathrm{Var}_\pi(g)$, the inequality in (14) implies that $P_{NR}$ can be worse than $P_{MG}$, in terms of variance of the associated estimators, only by a factor of $2(K - 1)$. Moreover, since the cost per iteration of $P_{MG}$ is $K/2$ times the one of $P_{NR}$ (see Section 2.1) the overall worsening is at most by a factor of 4.

**Remark 3.2.** The proof of the second inequality in (14) relies on showing that $P_R(c, c') \geq c^{-1}(K)P_{MG}(c, c')$ for every $c \neq c'$, which means that the probability of changing state according to $P_R$ is not too low compared to the one of $P_{MG}$. Interestingly, the converse is not true, in the sense that there is no $d > 0$ independent from $n$ such that $P_{MG}(c, c') \geq dP_R(c, c')$. Indeed, let $\pi$ be as in (4) with $K = 3$, $\boldsymbol{\alpha} = (1, 1, 1)$ and $f_\theta = f$. Then if $c = (1, \ldots, 1, 2, 3)$ and $c' = (1, \ldots, 1, 2, 2)$ it is easy to see that

$$P_{MG}(c, c') = \frac{2}{n(3 + n - 1)} \quad \text{and} \quad P_R(c, c') = \frac{1}{6n}.$$

The first inequality in (14) instead relies on extending classical asymptotic variance comparison results for lifted kernels to the case of state-dependent mixtures such as in $P_{NR}$, as shown in Section C.1.1 of the supplement.

$\square$

We stress that the results of Theorem 3.1 hold uniformly, in the sense that no assumptions on $\pi$ are needed. Thus using $P_{NR}$ is guaranteed to provide performances which never get significantly worse than the ones of $P_{MG}$ in terms of asymptotic variances. In the next sections, we will see that on the contrary $P_{NR}$ can lead to significant improvements (e.g. by a factor of $n$) relative to $P_{MG}$.

## 3.2 Comparison with conditional sampler

We now define the *conditional* sampler targeting $\pi(c, \boldsymbol{\theta}, \boldsymbol{w})$ mentioned in Section 1.2. The pseudocode is given in Algorithm 7 and we denote with $P_{CD}$ the associated Markov kernel on $[K]^n \times \Theta^K \times \Delta_{K-1}$. Also here we consider the random-scan case, which allows for an easier comparison with $P_{MG}$ and $P_{NR}$. We expect the main take-away messages to remain valid for the arguably more popular deterministic-scan scheme, even if theoretical results there are less neat (see e.g. Roberts and Rosenthal (2015); He et al. (2016); Gaitonde and Mossel (2024); Ascolani et al. (2024) and references therein).

The next proposition, whose proof is inspired by the one of (Liu, 1994, Thm.1), shows that $P_{MG}$ always yields a smaller asymptotic variance than $P_{CD}$. Again with an abuse of notation we use $g$ to denote both $g : [K]^n \to \mathbb{R}$ and $g : [K]^n \times \Theta^K \times \Delta_{K-1} \to \mathbb{R}$ function of the first argument alone.

**Proposition 3.3.** *Let $\pi$ be as in (3) and $g : [K]^n \to \mathbb{R}$. Then for every $f_\theta$, $n$, $Y$ we have that $Var(g, P_{MG}) \leq Var(g, P_{CD})$.*

Combined with Theorem 3.1, the above result implies that $\text{Var}(g, P_{NR}) \leq c(K)\text{Var}(g, P_{CD}) + [c(K) - 1]\text{Var}_\pi(g)$, so that if $P_{NR}$ outperforms $P_{MG}$ then it should also be preferred to $P_{CD}$. Thus in the following we restrict to the comparison between $P_{MG}$ and $P_{NR}$.

14

**Algorithm 7** (Conditional sampler $P_{\text{CD}}$)

---

Initialize $(c^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{w}^{(0)}) \in [K]^n \times \Theta^K \times \Delta_{K-1}$

**for** $t \geq 1$ **do**

    Sample $i \sim \text{Unif}\left(\{1, \ldots, n+1\}\right)$.

    **if** $i \leq n$ **then**

        Sample $c_i^{(t)} \sim \pi(c_i \mid \boldsymbol{\theta}^{(t-1)}, \boldsymbol{w}^{(t-1)})$ with

$$\pi(c_i = k \mid \boldsymbol{\theta}, \boldsymbol{w}) = \frac{w_k f_{\theta_k}(Y_I)}{\sum_{k'=1}^K w_{k'} f_{\theta_{k'}}(Y_i)}, \quad k = 1, \ldots, K.$$

    **end if**

    **if** $i = n + 1$ **then**

        Sample $\boldsymbol{w}^{(t)} \sim \text{Dir}\left(\alpha_1 + n_1(c^{(t-1)}), \ldots, \alpha_K + n_K(c^{(t-1)})\right)$.

        Sample $\theta_k^{(t)} \sim \pi(\theta_k \mid c^{(t-1)}) \propto \prod_{j : c_j^{(t-1)} = k} f_{\theta_k}(Y_j) p_0(\theta_k)$ for $k = 1, \ldots, K$.

    **end if**

**end for**

---

# 4  Scaling limit analysis

In this section we derive scaling limit results for $P_{\text{MG}}$ and $P_{\text{NR}}$ as $n \to \infty$. In general, given a sequence of discrete-time Markov chains $\{X_t^{(n)}\}_{t \in \mathbb{N}}$, scaling limit results (Gelman et al., 1997; Roberts and Rosenthal, 2001b) consist in showing that a time-changed process $\{Z_t^{(n)}\}_{t \in \mathbb{R}}$ defined as $Z_t^{(n)} = X_{\lceil h(n)t \rceil}^{(n)}$, with $h(n) \to \infty$ and $\lceil \cdot \rceil$ denoting the ceiling function, converges in a suitable sense to a non-degenerate process $\{Z_t\}_{t \in \mathbb{R}_+}$ as $n \to \infty$. Provided the limiting process is non-singular and ergodic, this is usually interpreted as suggesting that $\mathcal{O}(h(n))$ iterations of the discrete-time Markov chain are needed to mix. In other words, the time rescaling required to obtain a non-trivial limit is a measure of how the process speed scales as $n$ grows.

In order to derive such results we restrict to the prior case, where the likelihood is uninformative and the posterior distribution of $c$ coincides with the prior (2). This can be formally described by setting $f_\theta(y) = f(y)$, with $f$ probability density on $\mathcal{Y}$, so that the data do not provide any information on the labels. The joint distribution and full conditionals become

$$(15) \qquad \pi(c) = \frac{\prod_{k=1}^K \Gamma(\alpha_k + n_k(c))}{\Gamma(|\boldsymbol{\alpha}| + n)}, \quad \pi(c_i = k \mid c_{-i}) = \frac{\alpha_k + n_k(c_{-i})}{|\boldsymbol{\alpha}| + n - 1},$$

with $|\boldsymbol{\alpha}| = \sum_{k=1}^K \alpha_k$. This is clearly a simplified setting, which allows an explicit mathematical treatment and it can be considered as an extreme case of un-identifiability and overlapping components (which are indeed all the same). Extending the analysis

to the more realistic case of informative likelihood is an interesting direction for future research, see Section 7 for more details.

## 4.1   Marginal sampler

Consider a Markov chain $\{c^{(t)}\}_{t \in \mathbb{N}}$ with kernel $P_{\mathrm{MG}}$ and invariant distribution (15), where we suppress the dependence on $n$ for simplicity. Let

$$X_k(c) = \frac{n_k(c)}{n} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(c_i = k), \quad c \in [K]^n,$$

be the multiplicity of component $k$ and

$$(16) \qquad \boldsymbol{X}_t = (X_{t,1}, \dots, X_{t,K}) = \left( X_1\left(c^{(t)}\right), \dots, X_K\left(c^{(t)}\right) \right).$$

Crucially, since $\pi(c_i = k \mid c_i)$ defined in (15) only depends on the multiplicities, i.e. $(c_1, \dots, c_n)$ are exchangeable a priori, it follows that $(\boldsymbol{X}_t)_{t=0,1,2,\dots}$ is itself a Markov chain. Moreover, $\{\boldsymbol{X}_t\}_{t \in \mathbb{N}}$ is de-initializing for $\{c^{(t)}\}_{t \in \mathbb{N}}$ in the sense of Roberts and Rosenthal (2001a), so that the convergence properties of the former are equivalent to the one of the latter (by e.g. Corollary 2 therein). With an abuse of notation, we denote the kernel of $\{\boldsymbol{X}_t\}_{t \in \mathbb{N}}$ also as $P_{\mathrm{MG}}$.

In the proof of Theorem 4.1 we show that

$$\mathbb{E}\left[X_{t+1,k} - x_k \mid \boldsymbol{X}_t = \boldsymbol{x}\right] = \frac{2}{n^2}\left[\frac{\alpha_k}{2} - |\boldsymbol{\alpha}|\frac{x_k}{2} + o(1)\right]$$

and

$$\mathbb{E}\left[(X_{t+1,k} - x_k)^2 \mid \boldsymbol{X}_t = \boldsymbol{x}\right] = \frac{2}{n^2}\left[x_k(1 - x_k) + o(1)\right],$$

as $n \to \infty$. The above suggests that a rescaling of order $\mathcal{O}(n^2)$ is needed to have a non-trivial limit, as we will formally show below. In particular, let $\{\boldsymbol{Z}_t\}_{t \in \mathbb{R}_+}$ be the continuous-time process with generator

$$(17) \qquad \mathcal{A}g(\boldsymbol{x}) = \frac{1}{2}\sum_{k=1}^K (\alpha_k - |\boldsymbol{\alpha}|x_k)\frac{\partial}{\partial x_k}g(\boldsymbol{x}) + \frac{1}{2}\sum_{k,k'=1}^K x_k(\delta_{kk'} - x_{k'})\frac{\partial^2}{\partial x_k \partial x_{k'}}g(\boldsymbol{x}),$$

for every $g : \Delta_{K-1} \to \mathbb{R}$ twice differentiable and where $\boldsymbol{x} = (x_1, \dots, x_K)$. Such process exists (Ethier, 1976) and is called Wright-Fisher with mutation rates given by $\boldsymbol{\alpha}$. In particular, $\{\boldsymbol{Z}_t\}_{t \in \mathbb{R}_+}$ is a diffusion taking values in $\Delta_{K-1}$ whose stationary density is exactly $\pi(\boldsymbol{x}) = \mathrm{Dir}(\boldsymbol{x} \mid \boldsymbol{\alpha})$. The next theorem shows that, choosing $h(n) = n^2/2$, the continuous-time rescaling of $\{\boldsymbol{X}_t\}_{t \in \mathbb{N}}$ converges to $\{\boldsymbol{Z}_t\}_{t \in \mathbb{R}_+}$.

**Theorem 4.1.** *Let* $\{\boldsymbol{Z}_t^{(n)}\}_{t\in\mathbb{R}_+}$ *such that* $\boldsymbol{Z}_t^{(n)} = \boldsymbol{X}_{\lceil\frac{n^2}{2}t\rceil}$, *where* $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ *is the Markov chain in* (16) *with kernel* $P_{\mathrm{MG}}$ *and invariant distribution* $\pi$ *as in* (15). *Let* $\{\boldsymbol{Z}_t\}_{t\in\mathbb{R}_+}$ *be a diffusion with generator as in* (17). *Then if* $\boldsymbol{Z}_0^{(n)} \to \boldsymbol{Z}_0$ *weakly as* $n \to \infty$, *we have that* $\{\boldsymbol{Z}_t^{(n)}\}_{t\in\mathbb{R}_+} \to \{\boldsymbol{Z}_t\}_{t\in\mathbb{R}_+}$ *weakly as* $n \to \infty$, *according to the Skorokhod topology.*

**Remark 4.2.** The proof relies on convergence of generators, which is a standard technique when dealing with sequences of stochastic processes: we refer to (Ethier and Kurtz, 1986, Chapter 4) for details. While this approach is common in the MCMC literature (see e.g. Gelman et al. (1997); Roberts and Rosenthal (2001b) and related works), we are not aware of applications of it to mixture model contexts. On the contrary, the Wright-Fisher process often arises as the scaling limit of models for populations subjected to genetic drift and mutation (Ethier and Kurtz, 1986; Etheridge, 2011). Connections between sampling schemes and diffusions in population genetics have been also explored in other context, especially for sequential Monte Carlo techniques (Koskela et al., 2020; Brown et al., 2021). □

**Remark 4.3.** Theorem 4.1 suggests that $\mathcal{O}(n^2)$ iterations are needed for $P_{\mathrm{MG}}$ to converge. This is coherent with Khare and Zhou (2009, Prop.14.10.1) where, albeit motivated by a different problem, the authors show that, when targeting the prior distribution $\pi(c)$ in (15), the second largest eigenvalue of $P_{\mathrm{MG}}$ is

$$1 - \frac{|\boldsymbol{\alpha}|}{n(n+|\boldsymbol{\alpha}|-1)}.$$

This implies that the so-called relaxation time of $P_{\mathrm{MG}}$ scales as $\mathcal{O}(n^2)$ as $n \to \infty$, which means that $\mathcal{O}(n^2)$ iterations are required to mix; see e.g. Levin and Peres (2017, Thm.12.5) for more details on relaxation times. □

In order to see why an $\mathcal{O}(n^2)$ convergence is slower than desired, consider for example the case $K = 2$. Then $\{X_{t,1}\}_{t\in\mathbb{N}}$ is a Markov chain on $\{0, 1/n, \ldots, 1\}$ and thus $P_{\mathrm{MG}}$ requires $n^2$ iterations to sample from a distribution on a state space with cardinality $n$. Moreover, $\{X_{t,1}\}_{t\in\mathbb{N}}$ can be seen as a random walk with transition probabilities

$$\mathbb{P}\left(X_{t+1,1} = x_1 + \frac{1}{n} \mid X_{t,1} = x_1\right) = (1-x_1)\frac{\alpha_1 + nx_1}{\alpha_1 + \alpha_2 + n - 1} \approx x_1(1-x_1)$$

and

$$\mathbb{P}\left(X_{t+1,1} = x_1 - \frac{1}{n} \mid X_{t,1} = x_1\right) = x_1\frac{\alpha_2 + n(1-x_1)}{\alpha_1 + \alpha_2 + n - 1} \approx x_1(1-x_1),$$

when $n$ is large. Thus the probability of going up and down is almost the same, leading to the observed random-walk behaviour. This is reminiscent of classical examples studied in the non-reversible MCMC literature (Diaconis et al., 2000), where a faster algorithm is devised by considering a lifted version of the standard random walk.

## 4.2 Non-reversible sampler $P_{\mathrm{NR}}$

Consider now a Markov chain $\{c^{(t)}, v^{(t)}\}_{t\in\mathbb{N}}$ with kernel $P_{\mathrm{NR}}$ and invariant distribution (15). Define $\boldsymbol{X}_t$ as in (16) and $\boldsymbol{V}_t = \left\{V_{t,k,k'}\right\}_{(k,k')\in[K]^2}$ as

$$
V_{t,k,k'} = \begin{cases} 0 & \text{if } k = k' \\ v_{k,k'}^{(t)} & \text{if } k < k' \\ -v_{k',k}^{(t)} & \text{if } k > k' \,. \end{cases}
$$

This means that $V_{t,k',k} = +1$ implies that we are proposing from cluster $k'$ to $k$, for every pair $(k, k')$. This allows for a simpler statement in the theorem to follow.

By exchangeability arguments as above, it is simple to see that $\{(\boldsymbol{X}_t, \boldsymbol{V}_t)\}_{t\in\mathbb{N}}$ is de-initializing for $\{c^{(t)}, v^{(t)}\}_{t\in\mathbb{N}}$ and thus it has the same convergence properties. In the proof of Theorem 4.4 we show that

$$
\mathbb{E}\left[X_{t+1,k} - x_k \mid \boldsymbol{X}_t = \boldsymbol{x}, \boldsymbol{V}_t = v\right] = \frac{1}{n}\left[\sum_{k' : v_{k',k}=+1} \frac{x_k + x_{k'}}{K-1} - \sum_{k' : v_{k',k}=-1} \frac{x_k + x_{k'}}{K-1} + o(1)\right],
$$

which suggest that rescaling time by $n$ is sufficient for a non-trivial limit. A technical issue is that, when $X_{t,k} = 0$ for some $k$ then one of the velocities jumps deterministically to $V_{t,k',k} = +1$ with $k' \neq k$. To avoid complications related to such boundary effects, we study the scaling of the process in the set

$$
E_M \times V = \left\{\boldsymbol{x} \in \Delta_{K-1} \mid x_k > \frac{1}{M} \text{ for every } k\right\} \times \{-1, 0, +1\}^{[K]^2},
$$

with $M > 0$ arbitrarily large but fixed.

Let $\left\{\boldsymbol{Z}_t^{(M)}\right\}_{t\in\mathbb{R}_+} = \left\{\boldsymbol{Z}_{1,t}^{(M)}, \boldsymbol{Z}_{2,t}^{(M)}\right\}_{t\in\mathbb{R}_+}$ be a piecewise deterministic Markov process (Davis, 1984) on $E_M \times V$ defined as follows. Consider a inhomogeneous Poisson process $\Lambda_t$ with rate

$$
(18) \quad \lambda\left(\boldsymbol{Z}_t^{(M)}\right) = \frac{1}{2(K-1)}\sum_{k\neq k'}\left(Z_{1,t,k}^{(M)} + Z_{1,t,k'}^{(M)}\right)\beta\left(Z_{1,t,k}^{(M)}, Z_{1,t,k'}^{(M)}, Z_{2,t,k,k'}^{(M)}\right) + 2\xi,
$$

where

$$
\beta(x_k, x_{k'}, v_{k',k}) = \max\left\{0, \frac{\alpha_{k_-} - 1}{x_{k_-}} + \frac{1 - \alpha_{k_+}}{x_{k_+}}\right\}
$$

with $k_- = k'$ and $k_+ = k$ if $v_{k',k} = +1$ and viceversa if $v_{k',k} = -1$. In between events,

$\{\boldsymbol{Z}_t^{(M)}\}_{t\in\mathbb{R}_+}$ evolves deterministically as

(19)

$$\frac{\mathrm{d}Z_{1,t,k}^{(M)}}{\mathrm{d}t} = \Phi_k\left(\boldsymbol{Z}_t^{(M)}\right)$$

$$= \frac{1}{K-1}\left[\sum_{k':Z_{2,t,k',k}^{(M)}=+1}\left(Z_{1,t,k}^{(M)} + Z_{1,t,k'}^{(M)}\right) - \sum_{k':Z_{2,t,k',k}^{(M)}=-1}\left(Z_{1,t,k}^{(M)} + Z_{1,t,k'}^{(M)}\right)\right]$$

and

$$\frac{\mathrm{d}Z_{2,t,k',k}^{(M)}}{\mathrm{d}t} = 0,$$

with $(k',k) \in [K]^2$. The system of differential equations in (19) admits a unique solution by linearity in its arguments. Instead, at each event of $\Lambda_t$, say at $\tau > 0$, a pair $(k,k') \in [K]^2$ is selected with probability

$$q(k,k') \propto \frac{Z_{1,t,k}^{(M)} + Z_{1,t,k'}^{(M)}}{2(K-1)}\left[\beta\left(Z_{1,\tau_-,k}^{(M)}, Z_{1,\tau_-,k'}^{(M)}, Z_{2,\tau_-,k',k}^{(M)}\right) + 2\xi\right]\mathbb{1}(k \neq k')$$

and then the process jumps as follows:

(20)
$$Z_{2,\tau,k',k}^{(M)} = -Z_{2,\tau_-,k',k}^{(M)} \quad \text{and} \quad Z_{2,\tau,k,k'}^{(M)} = -Z_{2,\tau_-,k,k'}^{(M)},$$

where $\tau_-$ denotes the the left-limit at $\tau$. It follows that $\left\{\boldsymbol{Z}_t^{(M)}\right\}_{t\in\mathbb{R}_+}$ is a continuous-time process with generator

(21)

$$\mathcal{B}^{(M)}g(\boldsymbol{z}) = \mathbb{1}(z_1 \in E_M)\left\{\sum_{k=1}^K \Phi_k(\boldsymbol{z})\frac{\partial}{\partial z_{1,k}}g(\boldsymbol{z}) + \lambda(\boldsymbol{z})\sum_{k\neq k'}q(k,k')\left[g(\boldsymbol{z}_{(\boldsymbol{k},\boldsymbol{k'})}) - g(\boldsymbol{z})\right]\right\},$$

for every $g : E_M \times V \to \mathbb{R}$ twice continuously differentiable in the first argument, where $\boldsymbol{z}^{(k,k')} \in E_M \times V$ is equal to $\boldsymbol{z}$ except for

$$\boldsymbol{z}_{2,k,k'}^{(k,k')} = -\boldsymbol{z}_{2,k',k}^{(k,k')} = -\boldsymbol{z}_{2,k,k'}.$$

Such a process exists for every $M > 0$ since the rates $\lambda(\boldsymbol{z})$ are bounded (Davis, 1984). We can think of $\left\{\boldsymbol{Z}_t^{(M)}\right\}_{t\in\mathbb{R}_+}$ as a process with an absorbing boundary, which remains constant as soon as $Z_{1,t,k}^{(M)} \leq 1/M$ for some $k$.

Analogously, define $\left\{\boldsymbol{X}_t^{(M)}, \boldsymbol{V}_t^{(M)}\right\}_{t\in\mathbb{N}}$ as a modification of $\{\boldsymbol{X}_t, \boldsymbol{V}_t\}_{t\in\mathbb{N}}$, which remains constant as soon as $X_{t,k}^{(M)} \leq 1/M$ for some $k$. The next theorem shows that, choosing $h(n) = n$, the continuous-time rescaling of $\left\{\boldsymbol{X}_t^{(M)}, \boldsymbol{V}_t^{(M)}\right\}_{t\in\mathbb{N}}$ converges to $\left\{\boldsymbol{Z}_t^{(M)}\right\}_{t\in\mathbb{R}_+}$.

19

**Theorem 4.4.** *Fix $M > 0$ and let $\left\{ \boldsymbol{Z}_t^{(M,n)} \right\}_{t \in \mathbb{R}_+}$ such that $\boldsymbol{Z}_t^{(M,n)} = \left( \boldsymbol{X}_{\lceil nt \rceil}^{(M)}, \boldsymbol{V}_{\lceil nt \rceil}^{(M)} \right)$, where $\{ \boldsymbol{X}_t, \boldsymbol{V}_t \}_{t \in \mathbb{N}}$ is a Markov chain with operator $P_{\mathrm{NR}}$ and invariant distribution as in* (11)*, with $\pi$ in* (15)*. Let $\left\{ \boldsymbol{Z}_t^{(M)} \right\}_{t \in \mathbb{R}_+}$ be a piecewise deterministic Markov process with generator* (21)*. Then if $\boldsymbol{Z}_0^{(M,n)} \to \boldsymbol{Z}_0^{(M)}$ weakly as $n \to \infty$, we have that $\left\{ \boldsymbol{Z}_t^{(M,n)} \right\}_{t \in \mathbb{R}_+} \to \left\{ \boldsymbol{Z}_t^{(M)} \right\}_{t \in \mathbb{R}_+}$ weakly as $n \to \infty$, according to the Skorokhod topology.*

**Remark 4.5.** Looking at the process only in the interior of the simplex is inspired by other works on diffusion approximations, see e.g. Barton et al. (2004) where they use a similar technique to deal with explosive behaviour in the boundary. If $\alpha_k > 1$ for every $K$, we could proceed as in Theorem 4.2 therein to show that the boundary is never reached and thus the limit can be extended to the whole space. □

Theorem 4.4 suggests that the overall computational cost of Algorithm 6 is $\mathcal{O}(n)$ and, combined with Theorem 4.1, this suggest an $\mathcal{O}(n)$ speed-up relative to $P_{\mathrm{MG}}$ in the prior case. In Section 6 we will show empirically that large improvements are also present in more realistic and interesting settings where the likelihood is informative.

## 5   A variant of $P_{\mathrm{NR}}$

The kernels $P_{\mathrm{R}}$ and $P_{\mathrm{NR}}$ sample a new pair $(k, k')$ at every iteration. While this is natural and allows for direct theoretical comparisons with $P_{\mathrm{MG}}$ (see Theorem 3.1), an alternative in the non-reversible case is to keep the same value of $(k, k')$ for multiple iterations. We thus define the following, non-reversible and $\tilde{\pi}$-invariant kernel

$$(22) \qquad Q_{\mathrm{NR}} = \sum_{(k,k') \in \mathcal{K}} \frac{2}{K(K-1)} \sum_{t=1}^{\infty} q_{m_c(k,k')}(t) \tilde{P}_{k,k'}^t,$$

with $m_c(k, k') = (n_k(c) + n_{k'}(c))/s$ for some fixed $s \in (0, 1)$ and $q_m(t)$ being the probability mass function of a geometric random variable with parameter $1/m$. The algorithm picks a couple $(k, k')$ uniformly at random and then takes a random number of steps of the lifted kernel $\tilde{P}_{kk'}$, with average number of steps proportional to the total number of points in the two clusters, i.e. $n_k(c) + n_{k'}(c)$. The associated pseudo-code is presented in Algorithm 8. Reasoning as in Lemma 2.3 it is easy to see that $Q_{\mathrm{NR}}$ is

---

**Algorithm 8** Modified non-reversible sampler $(c', v') \sim Q_{\mathrm{NR}}((c, v), \cdot)$

---

Sample $(k, k') \sim \mathrm{Unif}(\mathcal{K})$
Sample $t \sim \mathrm{Geom}\left( s/(n_k(c) + n_{k'}(c)) \right)$ for some fixed $s \in (0, 1)$
Sample $(c', v') \sim \tilde{P}_{k,k'}^t((c, v), \cdot)$

---

$\tilde{\pi}$-invariant and uniformly ergodic, as stated in the next lemma.

**Lemma 5.1.** *For any probability distribution $\pi$ on $[K]^n$, the Markov kernel $Q_{\mathrm{NR}}$ defined in Algorithm 8 is $\tilde{\pi}$-invariant, with $\tilde{\pi}$ as in (11). Moreover, if $\pi(c) > 0$ for every $c \in [K]^n$, then $Q_{\mathrm{NR}}$ is irreducible, aperiodic and uniformly ergodic.*

The distinction with the main algorithm is that $P_{\mathrm{NR}}$ resamples the pair $(k, k')$ at each iteration with probability proportional to $n_k(c) + n_{k'}(c)$, while $Q_{\mathrm{NR}}$ keeps the same $(k, k')$ for $O(n_k(c) + n_{k'}(c))$ iterations and then resamples the pair $(k, k')$ uniformly from $\mathcal{K}$. Indeed we expect $P_{\mathrm{NR}}$ and $Q_{\mathrm{NR}}$ to perform similarly for fixed values of $K$, but we empirically observe that $Q_{\mathrm{NR}}$ tends to yield slower mixing as $K$ increases: see Section A in the Supplementary Material for a simulative comparison in the prior case. This motivated us to focus on $P_{\mathrm{NR}}$ as the main scheme of interest in this paper.

**Remark 5.2.** In the prior case of Section 4, where the invariant distribution is given by (15), it is possible to find a corresponding scaling limit for $Q_{\mathrm{NR}}$. The proof is analogous to the case of $P_{\mathrm{NR}}$ and we omit it for brevity, just limiting ourselves to identifying the candidate limit and discussing its implications. Consider a Markov chain $\{(\boldsymbol{X}_t, \boldsymbol{V}_t)\}_{t \in \mathbb{N}}$ with kernel $Q_{\mathrm{NR}}$. With similar calculations as in Theorem 4.4, the process $\left\{\boldsymbol{Z}_t^{(M,n)}\right\}_{t \in \mathbb{R}_+}$ defined as $\boldsymbol{Z}_t^{(M,n)} = \left(\boldsymbol{X}_{\lceil nt \rceil}^{(M)}, \boldsymbol{V}_{\lceil nt \rceil}^{(M)}\right)$ can be shown to converge to $\{\boldsymbol{Z}_t\}_{t \in \mathbb{R}_+}$ with generator

$$
\mathcal{C}^{(M)} g(\boldsymbol{z}) = \mathbb{1}(\boldsymbol{z}_1 \in E_M) \left\{ \frac{\partial}{\partial z_{1,k_+}} g(\boldsymbol{z}) - \frac{\partial}{\partial z_{1,k_-}} g(\boldsymbol{z}) \right.
$$
$$
+ \max\left\{0, \frac{\alpha_{k_-} - 1}{z_{1,k_-}} + \frac{1 - \alpha_{k_+}}{z_{1,k_+}}\right\} [g(\boldsymbol{z}_1, -\boldsymbol{z}_2) - g(\boldsymbol{z})]
$$
$$
\left. + \frac{s}{z_{1,k_-} + z_{2,k_+}} \sum_{k \neq k'} \frac{z_{1,k} + z_{1,k'}}{2(K-1)} \left[g\left(\boldsymbol{z}_1, \boldsymbol{z}_2^{(k,k')}\right) - g(\boldsymbol{z})\right] \right\},
$$

with $k_- = k'$ and $k_+ = k$ if $z_{2,k,k'} = +1$ and viceversa if $z_{2,k,k'} = -1$. Moreover $\boldsymbol{z}_2^{(k,k')}$ is the vector with $z_{2,k,k'} = -z_{2,k',k} = +1$ and zero otherwise. Interestingly, $\mathcal{C}^{(M)}$ coincides with the generator of the so-called Coordinate Sampler, introduced in Wu and Robert (2020), with target distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. $\qquad \square$

## 5.1 The random projection sampler being approximated

The main feature of $Q_{\mathrm{NR}}$ is that, after sampling a pair $(k, k') \in \mathcal{K}$, the operator $\tilde{P}_{k,k'}$ is applied for a random number of iterations. If $s \to 0$ the latter diverges almost surely, meaning that after selecting the pair the sampler will behave as $\tilde{P}_{k,k'}^t$ with $t \to \infty$. By definition of $\tilde{P}_{k,k'}^t$ and ergodicity, this converges to the kernel $\Pi_{k,k'}$ that updates the sub-partition of points in clusters $k$ and $k'$ conditional on the rest, i.e.

$$
(23) \quad \lim_{t \to \infty} P_{k,k'}^t(c, c') = \tilde{\Pi}_{k,k'}(c, c') \propto \left( \prod_{i \,:\, c_i \notin \{k, k'\}} \mathbb{1}\left(c_i = c_i'\right) \right) \pi(c') \quad c, c' \in [K]^n.
$$

21

Note that $\Pi_{k,k'}$ is a projection kernel, i.e. $\Pi^2_{k,k'} = \Pi_{k,k'}$. Analogously, again as $s \to 0$, $Q_{\mathrm{NR}}$ converges to the random projection kernel defined as

$$(24) \qquad P_{\mathrm{RP}}(c,c') = \frac{2}{K(K-1)} \sum_{(k,k') \in \mathcal{K}} \Pi_{k,k'}(c,c') \qquad\qquad c,c' \in [K]^n \,,$$

whose structure resembles the one of a random-scan Gibbs Sampler that updates the sub-partition of two randomly chosen pairs of clusters given the configuration of the other clusters. In this perspective, $Q_{\mathrm{NR}}$ can be interpreted as a Metropolis-within-Gibbs sampler approximating $P_{\mathrm{RP}}$.

**Remark 5.3.** In the prior case, as $n \to \infty$, we expect $P_{\mathrm{RP}}$ in turn to approximate a Gibbs sampler on the $(K-1)$-dimensional simplex, which at every iteration updates two coordinates chosen at random. In the special case of $\boldsymbol{\alpha} = (1,\ldots,1)$, the latter has been studied in Smith (2014) and shown to require $\mathcal{O}(K \log(K))$ iterations for mixing. $\square$

# 6 Simulations

## 6.1 Prior case

First of all we consider the prior case, where $f_\theta = f$ and the target distribution is given by (15). We let $K = 3$, $n = 1000$ and we run Algorithms 1 and 6 for 300 independent runs, first with $\boldsymbol{\alpha} = (1,1,1)$ and then with $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$. Initial configurations are independently generated, so that $c_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([K])$. For each run we store the value of the chains after $T = 100 \times n$ iterations and plot the corresponding proportion of labels of the first two components, i.e. $(n_1(c^{(T)})/n, n_2(c^{(T)})/n)$ in Figure 2. If the chains had reached convergence by then, these should be 300 independent samples approximately following a Dirichlet-Multinomial distribution with parameters $\boldsymbol{\alpha}$ (since $n$ is large, this is visually close to drawing samples directly from a $\mathrm{Dir}(\boldsymbol{\alpha})$ distribution).

From the results in Figure 2, it is clear that the non-reversible scheme (second column) leads to faster convergence: this is particularly manifest in the second row (corresponding to $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$), where the mass should be concentrated around the borders of the simplex. Indeed, both chains associated to $P_{\mathrm{MG}}$ remain stuck close to the initial configuration, where the proportion within each group is close to $1/3$. This is also clear from the last column of Figure 2, which shows that the marginal distribution of $P_{\mathrm{NR}}$ (in black) converges to the stationary one after fewer iterations.
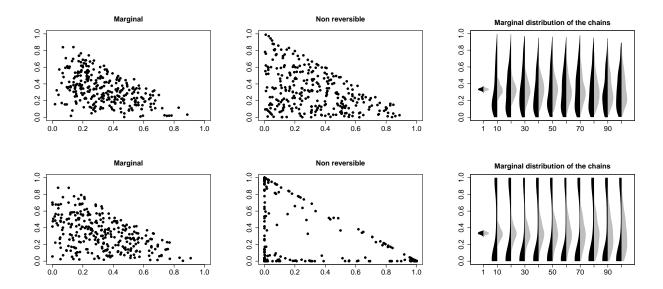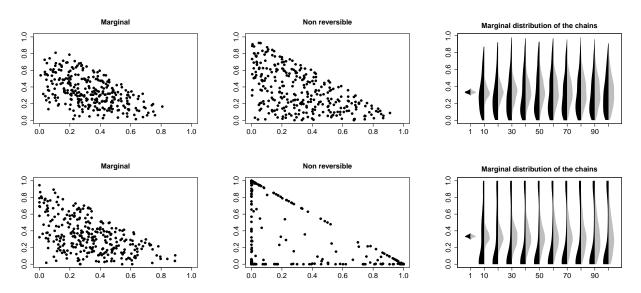
Figure 2: Left and center column: plot of the proportions of the first two components in the last of 100 iterations (after a thinning of size $n$) over 300 independent runs for $P_{\mathrm{MG}}$ (left) and $P_{\mathrm{NR}}$ (center). Right column: plot of the marginal distribution of the proportion of the first component at every 10 iterations (after thinning) for $P_{\mathrm{MG}}$ (gray) and $P_{\mathrm{NR}}$ (black). The first and second rows refer to $\boldsymbol{\alpha} = (1, 1, 1)$ and $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$, respectively. The target distribution is given in (15).

## 6.2  Posterior case

We now consider model (1) with $\mathcal{Y} = \Theta = \mathbb{R}$, $K = 3$,

(25) $$f_\theta(y) = N(y \mid \theta, \sigma^2), \quad p_0(\theta) = N(\theta \mid \mu_0, \sigma_0^2).$$

and hyperparameters set to $\mu_0 = 0$ and $\sigma^2 = \sigma_0^2 = 1$. We then generate 300 independent data sets of size $n = 1000$, each generated from the model as follows:

1. Sample $\boldsymbol{w} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ and $\theta_k \overset{\text{i.i.d.}}{\sim} p_0$ for $k = 1, \ldots, K$.

2. Sample $Y_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^K w_k f_{\theta_k}(y)$ for $i = 1, \ldots, n$.

For each dataset we target the associated posterior using $P_{\mathrm{MG}}$ and $P_{\mathrm{NR}}$. As before we initialize each chain uniformly, i.e. $c_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([K])$, and store its value after $T = 100 \times n$ iterations. Since the data are generated from the (Bayesian) model, the resulting distribution of the proportions within each component should be close to the prior one, i.e. again a Dirichlet-multinomial with parameter $\boldsymbol{\alpha}$. This test for convergence, discussed for example in Geweke (2004), relies on the fact that sampling from the *prior* distribution is equivalent to sampling from the *posterior*, given data generated according to the marginal distribution induced by the model.

23

The resulting samples are displayed in Figure 3, with the same structure as in Figure 2. Again the non-reversible scheme is much closer to the correct distribution, while $P_{\text{MG}}$ remains close to the initial configuration. Indeed, the results are remarkably close to the ones presented in Section 6.1: this suggests that the behaviour observed in the prior case is informative also of the actual behaviour observed in the posterior case, at least in this setting. In Section B of the Supplementary Material similar results are shown for the Poisson kernel.



Figure 3: Left and center column: plot of the proportions of the first two components in the last of 100 iterations (after a thinning of size $n$) over 300 independent runs for $P_{\text{MG}}$ (left) and $P_{\text{NR}}$ (center). Right column: plot of the marginal distribution of the proportion of the first component at every 10 iterations (after thinning) for $P_{\text{MG}}$ (gray) and $P_{\text{NR}}$ (black). The rows refer to $\alpha = 1$ and $\alpha = 0.1$ and the target distribution is given by the posterior of model (1), with $f_\theta(y)$ as in (25), $\mu_0 = 0$ and $\sigma^2 = \sigma_0^2 = 1$.

## 6.3 A high dimensional example

We now consider a higher dimensional version of the previous setting, where

$$(26) \qquad f_\theta(y) = N(y \mid \theta, \sigma_p^2 I_p), \quad p_0(\theta) = N(\theta \mid \mu_0, \sigma_0^2 I_p),$$
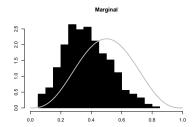
where now $y \in \mathbb{R}^p$ and $\theta \in \mathbb{R}^p$ with $p \geq 1$. We rescale the likelihood variance as $\sigma_p^2 = cp$ which guarantees that
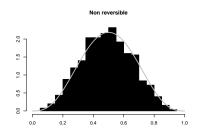
$$\frac{1}{\sigma_p^2} \sum_{j=1}^{p} (\theta_{1j} - \theta_{2j})^2 = \mathcal{O}(1).$$

24

In other words, we ask that the distance across components, rescaled by the variance, does not diverge as $p$ grows: this implies that some overlap between components is retained and that the problem is statistically non-trivial (see e.g. Chandra et al. (2023) for more discussion of Bayesian mixture models with high-dimensional data).

We generate 500 independent samples of size $n = 1000$ from model (26) with $p = 18$, $K = 5$, $\mu_0 = 0$, $\sigma_0^2 = 0.5$, $c = 2$ and $\boldsymbol{\alpha} = (4, 1, \ldots, 1)$. The data are generated as explained in the previous section and we run both $P_{\mathrm{MG}}$ and $P_{\mathrm{NR}}$, retaining only the last iteration for every chain: the initialization is again uniform at random.

In Figure 4 we plot the histograms of the last iteration for the proportion associated to the first component of $P_{\mathrm{MG}}$ and $P_{\mathrm{NR}}$ for 500 independent runs. Comparing the latter with the prior density, given by a Dirichlet-Multinomial with parameters $(4, 4)$ (approximately $\mathrm{Beta}(4, 4)$), it is evident that the non-reversible scheme is able to forget the initialization while the reversible is not. Indeed, as also clear from the right plot of Figure 4, the marginal distribution of $P_{\mathrm{MG}}$ significantly underestimates the size of the first cluster after $T = 100 \times n$ iterations.
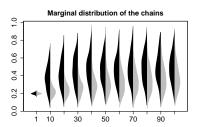


Figure 4:   Left and center column: histogram of the proportion of the first component in the last of 100 iterations (after a thinning of size $n$) over 500 independent runs for $P_{\mathrm{MG}}$ (left) and $P_{\mathrm{NR}}$ (center). The gray line corresponds to the density of a  $\mathrm{Beta}(4, 4)$. Right column: plot of the marginal distribution of the chains at every 10 iterations (after thinning) for $P_{\mathrm{MG}}$ (gray) and $P_{\mathrm{NR}}$ (black). The target distribution is given by the posterior of model (1), with $f_\theta(y)$ as in (26), $p = 18$, $K = 5$, $\mu_0 = 0$, $\sigma_0^2 = 0.5$, $c = 2$ and $\boldsymbol{\alpha} = (4, 1, \ldots, 1)$.

## 6.4    Overfitted setting

Finally, we consider an overfitted case, previously discussed in Section 2.2.1. We take a one-dimensional Gaussian kernel as in (25) and take $\alpha_k = \alpha$ for all $k \in \{1, \ldots, K\}$. In this setting, using the notation of Section 2.2.1, Rousseau and Mengersen (2011, Thm.1) implies that

1. if $\alpha > 1/2$, then more than $K^*$ atoms have non-negligible mass, i.e.  multiple atoms are associated to the same "true" component,

2. if $\alpha \leq 1/2$, then the posterior concentrates on configurations with exactly $K^*$ components, up to $n^{-1/2}$ posterior mass.

We take $K = 2$ and $K^* = 1$, with $Y_i \overset{\text{i.i.d.}}{\sim} N(y \mid 2, 1)$ and $n = 1000$. The first two columns of Figure 5 plot the histogram of the proportion of the first component after $T = 100 \times n$ iterations (and thinning of size $n$) for $\alpha = 1$ (top row) and $\alpha = 0.1$ (bottom row). The two algorithms are initialized according to the "incorrect" scenario, i.e. all the observations in the first component in the first row and uniformly at random in the bottom row. The figure illustrates that only $P_{\text{NR}}$ is able to reach the high probability region: this means that, despite its locality, the persistence of $P_{\text{NR}}$ allows for significantly faster traveling across the space. On the contrary, $P_{\text{MG}}$ remain stuck in the initial configuration (which yields a similar likelihood) for both the scenarios. This is also confirmed by the right column, which depicts the marginal distribution of the chains: after few iterations, the distribution associated to $P_{\text{NR}}$ stabilizes and yields the correct behaviour.
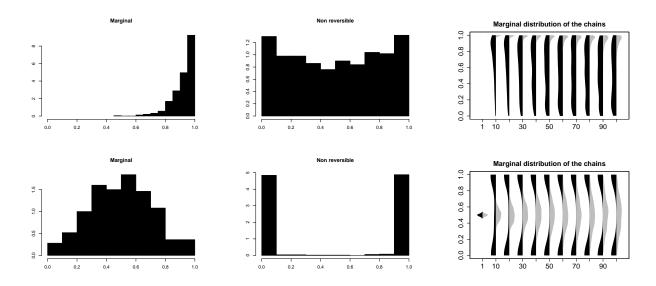


Figure 5: Left and center column: histogram of the proportion of the first component in the last of 100 iterations (after a thinning of size $n$) over 300 independent runs for $P_{\text{MG}}$ (left) and $P_{\text{NR}}$ (center). Right column: plot of the marginal distribution of the proportion of the first component at every 10 iterations (after thinning) for $P_{\text{MG}}$ (gray) and $P_{\text{NR}}$ (black). First row: $\alpha = 3/2$ and initialization uniformly at random. Second row: $\alpha = 0.1$ and $c_i^{(0)} = 1$ for every $i$. The target distribution is given by the posterior of model (1), with $Y_i \overset{\text{i.i.d.}}{\sim} N(y \mid 2, 1)$ and $f_\theta(y)$ as in (25), $\mu_0 = 0$ and $\sigma^2 = \sigma_0^2 = 1$.

# 7  Discussion

In this work we introduced a novel, simple and effective non-reversible MCMC sampler for mixture models, which enjoys three favourable features: (i) it is a simple modification of the original marginal scheme of Algorithm 1, (ii) its performance cannot be worse than the reversible chain by more than a factor of four (Theorem 3.1), (iii) it is shown to drastically speed-up convergence in various scenarios of interest.

Both the theory and methodology presented in this work could be extended in many interesting directions, and we now discuss some of those, starting from algorithmic and methodological ones. First, in the current formulation of Algorithm 6, the pair of clusters to update and the observation to move are selected with probabilities that do not depend on the actual observations within the clusters (except for their sizes). A natural extension would be to consider informed proposal distributions, as in e.g. Zanella (2020); Power and Goldman (2019); Gagnon and Maire (2024b): we expect this to lead to a potentially large decrease of the number of iterations needed for mixing, but with an additional cost per iteration. We leave the discussion and exploration of this tradeoff to future work. Second, one could also consider schemes that adaptively modify the probabilities $p_c(k, k')$ in (9) in order to propose more often clusters with higher overlap (or higher acceptance rates of proposed swaps), thus reducing computational waste associated to frequently proposing swaps across clusters with little overlap.

From the theoretical point of view, it would be highly valuable to extend the scaling limit analysis to the posterior case. While interesting, we expect this to require working with measure-valued processes and, more crucially, to require significant work in combining the MCMC analysis part with currently available results about posterior asymptotic behaviour of mixture models (Nguyen, 2013; Guha et al., 2021).

In this paper we stick to the case of a fixed number of components. A natural generalization regards the case of $K$ random or infinite (e.g. Dirichlet process mixtures, see Ferguson (1973); Lo (1984)). This presents additional technical difficulties that we leave to future work: for example, since no upper bound is available on the number of components, it would be more natural to define a Markov chain over the full space of partitions of $[n]$. Finally, mixture models are an instance of the broader framework of latent class models (Goodman, 1974) and it would be interesting to explore the effectiveness of the methodology developed here in such broader settings.

# References

Andrieu, C. and S. Livingstone (2021). Peskun–Tierney ordering for Markovian Monte Carlo: beyond the reversible scenario. *The Annals of Statistics 49*(4), 1958–1981.

Ascolani, F., H. Lavenant, and G. Zanella (2024). Entropy contraction of the Gibbs sampler under log-concavity. *arXiv preprint arXiv:2410.00858*.

Barton, N. H., A. M. Etheridge, and A. K. Sturm (2004). Coalescence in a random background. *Annals of Applied Probability 14*(3), 754 – 785.

Bierkens, J. (2016). Non-reversible Metropolis-Hastings. *Statistics and Computing 26*(6), 1213–1228.

Brown, S., P. A. Jenkins, A. M. Johansen, and J. Koskela (2021). Simple conditions for convergence of sequential Monte Carlo genealogies with applications. *Electronic Journla of Probability 26*, 1 – 22.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Chandra, N. K., A. Canale, and D. B. Dunson (2023). Escaping the curse of dimensionality in Bayesian model-based clustering. *Journal of Machine Learning Research 24*(144), 1–42.

Chen, T.-L. and C.-R. Hwang (2013). Accelerating reversible Markov chains. *Statistics & Probability Letters 83*(9), 1956–1962.

Davis, M. H. (1984). Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological) 46*(3), 353–376.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological) 39*(1), 1–22.

Diaconis, P., S. Holmes, and R. M. Neal (2000). Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability*, 726–752.

Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological) 56*(2), 363–375.

Etheridge, A. (2011). *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009*. Springer.

Ethier, S. N. (1976). A class of degenerate diffusion processes occurring in population genetics. *Communications on Pure and Applied Mathematics 29*(5), 483–493.

Ethier, S. N. and T. G. Kurtz (1986). *Markov processes: characterization and convergence*. John Wiley & Sons.

Faulkner, M. F. and S. Livingstone (2024). Sampling algorithms in statistical physics: a guide for statistics and machine learning. *Statistical Science 39*(1), 137–164.

Fearnhead, P., J. Bierkens, M. Pollock, and G. O. Roberts (2018). Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Statistical Science 33*(3), 386–412.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Stat. 1*(2), 209–230.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Volume 425. Springer.

Gagnon, P. and F. Maire (2024a). An asymptotic Peskun ordering and its application to lifted samplers. *Bernoulli 30*(3), 2301–2325.

Gagnon, P. and F. Maire (2024b). Theoretical guarantees for lifted samplers. *arXiv preprint arXiv:2405.15952*.

Gaitonde, J. and E. Mossel (2024). Comparison Theorems for the Mixing Times of Systematic and Random Scan Dynamics. *arXiv preprint arXiv:2410.11136*.

Gelman, A., W. R. Gilks, and G. O. Roberts (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of applied probability 7*(1), 110–120.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association 99*(467), 799–804.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*(2), 215–231.

Guha, A., N. Ho, and X. Nguyen (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli 27*(4), 2159–2188.

He, B. D., C. M. De Sa, I. Mitliagkas, and C. Ré (2016). Scan order in Gibbs sampling: Models in which it matters and bounds on how much. *Advances in neural information processing systems 29*.

Hobert, J. P., V. Roy, and C. P. Robert (2011). Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modeling. *Statistical Science 26*(3), 332–351.

Khare, K. and H. Zhou (2009). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. App. Probab. 19*, 737–777.

Koskela, J., P. A. Jenkins, A. M. Johansen, and D. Spanó (2020). Asymptotic genealogies of interacting particle systems with an application to sequential Monte Carlo. *The Annals of Statistics 1*, 560 – 583.

Lee, K., J.-M. Marin, K. Mengersen, and C. Robert (2009). Bayesian inference on finite mixtures of distributions. In *Perspectives in mathematical sciences I: Probability and statistics*, pp. 165–202. World Scientific.

Levin, D. A. and Y. Peres (2017). *Markov chains and mixing times*, Volume 107. American Mathematical Soc.

Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association 89*(427), 958–966.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *Ann. Stat. 12*(1), 351–357.

Marin, J.-M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics 25*, 459–507.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual review of statistics and its application 6*(1), 355–378.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture

models. *Annals of Statistics 41*(1), 370–400.

Nishimura, A., D. B. Dunson, and J. Lu (2020). Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods. *Biometrika 107*(2), 365–380.

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

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Volume 2. Springer.

Roberts, G. O. and J. S. Rosenthal (2001a). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics 28*(3), 489–504.

Roberts, G. O. and J. S. Rosenthal (2001b). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science 16*(4), 351–367.

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

Roberts, G. O. and J. S. Rosenthal (2015). Surprising convergence properties of some simple Gibbs samplers under various scans. *International Journal of Statistics and Probability 5*(1), 51–60.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 73*(5), 689–710.

Schauer, M. and M. Wienöbst (2024). Causal structure learning with momentum: Sampling distributions over Markov Equivalence Classes. *Proceedings of Machine Learning Research 246*, 382–400.

Smith, A. (2014). A Gibbs sampler on the n-simplex. *Annals of Applied Probability 24*(1), 114–130.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied probability*, 1–9.

Vucelja, M. (2016). Lifting—a nonreversible Markov chain Monte Carlo algorithm. *American Journal of Physics 84*(12), 958–968.

Wu, C. and C. P. Robert (2020). Coordinate sampler: a non-reversible Gibbs-like MCMC sampler. *Statistics and Computing 30*(3), 721–730.

Yang, Y., M. J. Wainwright, and M. I. Jordan (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics 44*(6), 2497–2532.

Zanella, G. (2020). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association 115*(530), 852–865.

Zhou, Q. and H. Chang (2023). Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *The Annals of Statistics 51*(3), 1058–1085.

Zhou, Q., J. Yang, D. Vats, G. O. Roberts, and J. S. Rosenthal (2022). Dimension-free mixing for high-dimensional Bayesian variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 84* (5), 1751–1784.

# A  Comparison between $P_{\mathrm{NR}}$ and $Q_{\mathrm{NR}}$

In this section we consider the same setting of Section 6.1, where the target distribution is given in (15). We run both $P_{\mathrm{NR}}$ and $Q_{\mathrm{NR}}$ (with $s = 1$) for 300 independent trials with initialization uniformly at random. We consider $n = 1000$, $K = 3, 10, 20, 50$ and $\boldsymbol{\alpha} = (1, 1/(K-1), \ldots, 1/(K-1))$, so that the marginal distribution on the proportion of the first component is a Dirichlet-Multinomial with parameters $(1, 1)$ and thus close to a uniform distribution on $(0, 1)$.

Figure 6 plots the corresponding empirical marginal distribution obtained by the chains (black corresponds to $P_{\mathrm{NR}}$ and gray to $Q_{\mathrm{NR}}$). Even if both schemes correctly reach stationarity, it seems that $Q_{\mathrm{NR}}$ yields slower mixing as $K$ increases: this is particularly evident in the case $K = 50$, where $Q_{\mathrm{NR}}$ remains close to the initial configuration.

# B  Simulations for the Poisson kernel

Here we consider model (1) with $K = 3$ and

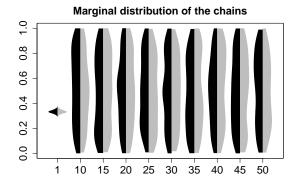$$(27) \qquad f_\theta(y) = \mathrm{Po}(y \mid \theta), \quad p_0(\theta) = \mathrm{Gamma}(\theta \mid \beta_1, \beta_2).$$
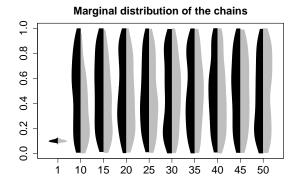
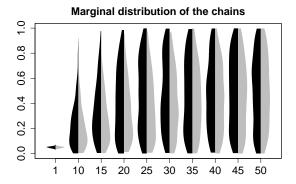It is easy to show that the predictive distribution reads

$$p(Y_{n+1} = y \mid Y) = \frac{\Gamma(\beta_1 + \sum_{i=1}^n Y_i + y)}{\Gamma(\beta_1 + \sum_{i=1}^n Y_i)\Gamma(y+1)} \frac{(n+\beta_2)^{\beta_1 + \sum_{i=1}^n Y_i}}{(n+\beta_2+1)^{\beta_1 + \sum_{i=1}^n Y_i + y}}.$$

We consider $\beta_1 = \beta_2 = 1$ and we draw 300 independent samples from the model above with $n = 1000$, following the same procedure illustrated in Section 6.2. For each dataset we run Algorithms 1 and 6, initialized uniformly at random, and we retain only the last iteration.

The results of the simulations are similar to the ones of Section 6.2, as shown in Figure 7: again the non-reversible scheme is much closer to the prior distribution, while $P_{\mathrm{MG}}$ remains close to the initial configuration.
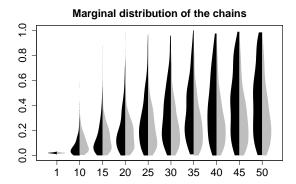
Figure 6:  Plots of the marginal distribution of the proportion of the first component for the chains associated to $P_{\mathrm{NR}}$ (black) and $Q_{\mathrm{NR}}$ (gray) at every. From top left to bottom right, the plots refer to $K = 3, 10, 20, 50$, where the target distribution is as in (15) with $\boldsymbol{\alpha} = (1, 1/(K-1), \ldots, 1/(K-1))$ and $n = 1000$.

# C    Proofs

## C.1    General results about lifting and mixtures

In order to prove results below, especially Theorem 3.1, we first need to generalize some classical results about lifting of Markov chains (see e.g. Chen and Hwang, 2013; Bierkens, 2016; Andrieu and Livingstone, 2021) to our mixture case, which can be seen as a way to construct 'multi-dimensional' lifted chains. We will make use of the following classical lemma, which for example follows by results in Chen and Hwang (2013) as detailed below.

**Lemma C.1.** *Let $\mu$ be a probability distribution on a finite space $\mathcal{X}$, $P$ a $\mu$-invariant and irreducible Markov transition matrix, $P^*$ the $\mu$-adjoint of $P$ and $K = (P + P^*)/2$.*
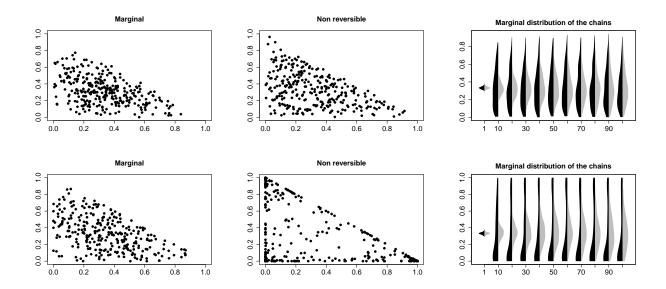
Figure 7: Left and center column: plot of the proportions of the first two components in the last of 100 iterations (after a thinning of size $n$) over 300 independent runs for $P_{\mathrm{MG}}$ (left) and $P_{\mathrm{NR}}$ (center). Right column: plot of the marginal distribution of the proportion of the first component at every 10 iterations (after thinning) for $P_{\mathrm{MG}}$ (gray) and $P_{\mathrm{NR}}$ (black). The rows refer to $\alpha = 1$ and $\alpha = 0.1$ and the target distribution is given by the posterior of model (1), with $f_\theta(y)$ as in (27) and $\beta_1 = \beta_2 = 1$.

*Then $Var(g, P) \leq Var(g, K)$ for all $g : \mathcal{X} \to \mathbb{R}$.*

*Proof.* Consider the decomposition $P = K + Q$, $Q = \frac{1}{2}P - \frac{1}{2}P^*$. By construction, $K$ is a $\mu$-reversible transition matrix. Moreover, by definition of adjoint we have that $Q$ is antisymmetric with respect to $\mu$, which means that $\mu(x)Q(x,y) = -\mu(y)Q(y,x)$ for all $x, y \in \mathcal{X}$. Finally, for every $x \in \mathcal{X}$ we have that

$$\sum_{y \in \mathcal{X}} Q(x,y) = \frac{1}{2} \sum_{y \in \mathcal{X}} P(x,y) - \frac{1}{2} \sum_{y \in \mathcal{X}} P^*(x,y) = 0$$

and thus each row of $Q$ sums up to zero. Therefore by Lemma 2 in Chen and Hwang (2013) we have that $Var(g, P) = Var(g, K + Q) \leq Var(g, K)$ for all $g : \mathcal{X} \to \mathbb{R}$. $\square$

### C.1.1  Result with general notation

Let $\pi$ be a probability distribution on a finite space $\mathcal{C}$. Let $D \in \mathbb{N}$, and $(K_{d,+1})_{d \in \{1,\dots,D\}}$ and $(K_{d,-1})_{d \in \{1,\dots,D\}}$ be Markov transition kernels on $\mathcal{C}$ such that

$$(28) \qquad \pi(c)K_{d,+}(c, c') = \pi(c')K_{d,-}(c', c) \qquad \text{for all } c \neq c' \text{ and all } d = 1, \dots, D.$$

33

Define the Markov transition kernel on $\mathcal{C}$

$$(29) \qquad K_R(c, c') = \sum_{d=1}^{D} p_c(d) K_d\left(c, c'\right),$$

where $K_d = (K_{d,+} + K_{d,-})/2$ and $p_c$ are weights such that $\sum_{d=1}^{D} p_c(d) = 1$ for all $c \in \mathcal{C}$ and

$$(30) \qquad p_c(d) = p_{c'}(d) \qquad\qquad \text{if } K_d(c, c') > 0.$$

Define the Markov transition kernel on $\mathcal{C} \times \{-1, 1\}^D$

$$(31) \qquad K_{NR}((c, v), (c', v')) = \sum_{d=1}^{D} p_c(d)\, (F_d K_{\text{lift},d} F_d)\, ((c, v), (c', v')),$$

where $F_d$ is the flipping operator defined as

$$F_d((c, v), (c', v')) = \mathbb{1}(c = c')\left[(1 - \alpha)\mathbb{1}(v = v') + \alpha\mathbb{1}(v_{-d} = v'_{-d}, v'_d = -v_d)\right]$$

for some fixed $\alpha \in [0, 1]$ and

$$(32) \qquad K_{\text{lift},d}((c, v), (c', v')) = K_{d,v_d}\left(c, c'\right) Q_{d,c,c'}(v, v')$$

with

$$(33) \quad Q_{d,c,c'}(v, v') = \mathbb{1}(v_{-d} = v'_{-d})\left[\mathbb{1}(c \neq c')\mathbb{1}(v_d = v'_d) + \mathbb{1}(c = c')\mathbb{1}(v_d = -v'_d)\right].$$

Here $\alpha$ plays the role of a refresh rate. One could also think at the case $\alpha = 0$ for simplicity, where $F_d$ becomes the identity operator and can thus be ignored.

**Lemma C.2.** *Under* (28)-(33), *we have that*

(a) $K_R$ *is $\pi$-reversible.*

(b) $K_{NR}$ *is $\tilde{\pi}$-invariant, with $\tilde{\pi}(c, v) = \pi(c)2^{-D}$.*

(c) $Var(\tilde{g}, K_{NR}) \leq Var(g, K_R)$ *for all $g : \mathcal{C} \times \{-1, +1\}^D \to \mathbb{R}$ and $\tilde{g} : \mathcal{C} \to \mathbb{R}$ such that $g(c, v) = \tilde{g}(c)$ for all $(c, v) \in \mathcal{C} \times \{-1, +1\}^D$.*

*Proof.* Consider first part (a). By (28), for every $c \neq c'$ we have

$$\begin{aligned} 2\pi(c)K_d(c, c') &= \pi(c)K_{d,+}(c, c') + \pi(c)K_{d,-}(c, c') \\ &= \pi(c')K_{d,-}(c', c) + \pi(c')K_{d,+}(c', c) = 2\pi(c')K_d(c', c). \end{aligned}$$

34

and thus by (29) and (30) we have

$$\pi(c)K_R(c,c') = \sum_{d=1}^{D} p_c(d)\pi(c)K_d(c,c')$$

$$= \sum_{d=1}^{D} p_{c'}(d)\pi(c')K_d(c',c) = \pi(c')K_d(c',c),$$

meaning that $K_R$ is $\pi$-reversible.

Consider now point (b). Let $(c',v') \in \mathcal{C} \times \{-1,+1\}^D$. If $v = v'$, by (32), $Q_{d,c,c'}(v,v') = \mathbb{1}(c \neq c')$ and (28) we have

$$\sum_{c \in \mathcal{C}} \tilde{\pi}(c,v)K_{\text{lift},d}((c,v),(c',v')) = \sum_{c \in \mathcal{C}} \pi(c)K_{d,v_d}(c,c')Q_{d,c,c'}(v,v')2^{-D}$$

$$= \sum_{c \neq c'} \pi(c)K_{d,v'_d}(c,c')2^{-D}$$

$$= \sum_{c \neq c'} \pi(c')K_{d,-v'_d}(c',c)2^{-D} = \tilde{\pi}(c',v')\left[1 - K_{d,-v'_d}(c',c')\right].$$

Similarly, if $v_{-d} = v'_{-d}$ and $v_d = -v'_d$, by $Q_{d,c,c'}(v,v') = \mathbb{1}(c = c')$ we have that

$$\sum_{c \in \mathcal{C}} \tilde{\pi}(c,v)K_{\text{lift},d}((c,v),(c',v')) = \sum_{c \in \mathcal{C}} \pi(c)K_{d,v_d}(c,c')Q_{d,c,c'}(v,v')2^{-D}$$

$$= \tilde{\pi}(c')K_{d,-v'_d}(c',c')2^{-D} = \tilde{\pi}(c',v')K_{d,-v'_d}(c',c').$$

Summing the two expressions above, and using the fact that $K_{\text{lift},d}((c,v),(c',v')) = 0$ if $v_{-d} \neq v'_{-d}$, we have

$$\sum_{c,v} \tilde{\pi}(c,v)K_{\text{lift},d}((c,v),(c',v')) = \tilde{\pi}(c',v'),$$

which implies that $K_{\text{lift},d}$ is $\tilde{\pi}$-invariant. Since $F_d$ is also trivially $\tilde{\pi}$-invariant and composition of invariant kernels remains invariant, then $F_d K_{\text{lift},d} F_d$ is $\tilde{\pi}$-invariant. Finally, using (30), we have

$$\sum_{c,v} \tilde{\pi}(c,v)K_{NR}((c,v),(c',v')) = \sum_{d=1}^{D} \sum_{c,v} p_c(d)\tilde{\pi}(c,v)\left(F_d K_{\text{lift},d} F_d\right)((c,v),(c',v'))$$

$$= \sum_{d=1}^{D} p_{c'}(d) \sum_{c,v} \tilde{\pi}(c,v)\left(F_d K_{\text{lift},d} F_d\right)((c,v),(c',v'))$$

$$= \sum_{d=1}^{D} p_{c'}(d)\tilde{\pi}(c',v') = \tilde{\pi}(c',v'),$$

and therefore $K_{NR}$ is $\tilde{\pi}$-invariant.

Consider now point (c). Let $\bar{K}_R = (K_{NR} + K_{NR}^*)/2$, where $K_{NR}^*$ is the $\tilde{\pi}$-adjoint of $K_{NR}$. Since $F_d^* = F_d$, which is easy check by definition of $F_d$, we have that $(F_d K_{\text{lift},d} F_d)^* = F_d^* K_{\text{lift},d}^* F_d^* = F_d K_{\text{lift},d}^* F_d$, which implies

$$K_{NR}^*((c,v),(c',v')) = \sum_{d=1}^{D} p_c(d) \left( F_d K_{\text{lift},d}^* F_d \right)((c,v),(c',v'))$$

and thus

$$\bar{K}_R((c,v),(c',v')) = \frac{1}{2}\sum_{d=1}^{D} p_c(d) \left(F_d K_{\text{lift},d} F_d\right)((c,v),(c',v')) + \frac{1}{2}\sum_{d=1}^{D} p_c(d) \left(F_d K_{\text{lift},d}^* F_d\right)((c,v),(c',v'))$$

$$= \sum_{d=1}^{D} p_c(d) \left(F_d \bar{K}_{NR,d} F_d\right)((c,v),(c',v'))$$

with $\bar{K}_{NR,d} := \frac{1}{2}K_{\text{lift},d} + \frac{1}{2}K_{\text{lift},d}^*$. By (28) we have that for $c' \neq c$

$$K_{\text{lift},d}^*((c,v),(c',v')) = \frac{\tilde{\pi}(c',v')}{\tilde{\pi}(c,v)} K_{\text{lift},d}((c',v'),(c,v))$$

$$= \frac{\pi(c')}{\pi(c)} K_{d,v_d'}\left(c',c\right) Q_{d,c',c}(v',v)$$

$$= K_{d,-v_d'}\left(c,c'\right) Q_{d,c',c}(v',v) = K_{d,-v_d}\left(c,c'\right) Q_{d,c,c'}(v,v'),$$

where we used the definition of $Q_{d,c',c}(v',v)$. For $c' = c$ we have that

$$K_{\text{lift},d}^*((c,v),(c,v')) = K_{\text{lift},d}((c,v'),(c,v))$$

$$= K_{d,v_d'}\left(c,c\right) Q_{d,c,c}(v',v) = K_{d,-v_d}\left(c,c\right) Q_{d,c,c'}(v,v')$$

where we used that $Q_{d,c,c}(v',v) > 0$ implies $v_d' = -v_d$. Thus

$$\bar{K}_{NR,d}((c,v),(c',v')) = \frac{1}{2}K_{d,v_d}\left(c,c'\right) Q_{d,c,c'}(v,v') + \frac{1}{2}K_{d,-v_d}\left(c,c'\right) Q_{d,c,c'}(v,v')$$

$$= \left[\frac{1}{2}K_{d,+}\left(c,c'\right) + \frac{1}{2}K_{d,-}\left(c,c'\right)\right] Q_{d,c,c'}(v,v')$$

$$= K_d(c,c')Q_{d,c,c'}(v,v').$$

Let now $g : \mathcal{C} \times \{-1,+1\}^D \to \mathbb{R}$ and $\tilde{g} : \mathcal{C} \to \mathbb{R}$ such that $g(c,v) = \tilde{g}(c)$. Then, since $F_d$ leaves the first coordinate invariate and $K_d$ does not depend on $v$, we have that

$$\left(F_d \bar{K}_{NR,d} F_d g\right)(c,v) = \sum_{c',v'} \left(F_d \bar{K}_{NR,d} F_d\right)((c,v),(c',v'))\tilde{g}(c')$$

$$= \sum_{c'} K_d(c,c')\tilde{g}(c') = K_d\tilde{g}(c),$$

36

which implies

$$\bar{K}_{NR}g(c,v) = \sum_{c',v'} \bar{K}_{NR}((c,v),(c',v'))g(c',v') = K_R\tilde{g}(c).$$

By simple induction then $\bar{K}_{NR}^t g(c,v) = K_R^t \tilde{g}(c)$ for every $t$. It thus follows that $\text{Var}(g,\bar{K}_R) = \text{Var}(\tilde{g},K_R)$. Point (c) then follows from $\text{Var}(g,K_{NR}) \leq \text{Var}(g,\bar{K}_R)$, which is a consequence of Lemma C.1. □

## C.2  Proof of Lemma 2.1

*Proof.* Reversibility follows by Lemma C.2 (point (a)), with $K_R = P_R$ and $(k,k') \in \mathcal{K}$ in place of $d \in [D]$. The only delicate condition to verify is given by (30), which follows since $P_{k,k'}(c,c') > 0$ implies that $n_k(c) + n_{k'}(c) = n_k(c') + n_{k'}(c')$ and therefore $p_c(k,k') = p_{c'}(k,k')$.

Since $\pi(c) > 0$ for every $c \in [K]^n$, we have $\pi(c_i = k \mid c_{-i}) > 0$ for every $c_{-i} \in [K]^{n-1}$. Combining this with the fact that $p_c(k,k') > 0$ for every $(k,k') \in \mathcal{K}$ such that $n_k(c) + n_{k'}(c) > 0$, we get that for every pair $c \neq c' \in [K]^n$ there exists a $T \in \mathbb{N}$ and a sequence $c = c^{(0)}, c^{(1)}, \ldots, c^{(T)} = c'$ such that $P_R(c^{(t-1)}, c^{(t)}) > 0$ for every $t = 1, \ldots, T-1$. Thus, $P_{NR}$ is irreducible. It is also easy to see that $P_R$ is aperiodic. Uniform ergodicity then follows from Levin and Peres (2017, Theorem 4.9). □

## C.3  Proof of Lemma 2.2

*Proof.* Fix $(k,k') \in \mathcal{K}$ and let $(k_1,k_2)$ be the pair sampled in the first two lines of Algorithm 4. Then a draw from the latter will have $(k,k')$ as realization if and only if $(k_1,k_2) = (k,k')$ or $(k_1,k_2) = (k',k)$. By construction

$$\mathbb{P}(k_1 = k, k_2 = k') = \frac{n_k(c)}{(K-1)n}, \quad \mathbb{P}(k_1 = k', k_2 = k) = \frac{n_{k'}(c)}{(K-1)n},$$

and thus $p_c(k,k') = \mathbb{P}(k_1 = k, k_2 = k') + \mathbb{P}(k_1 = k', k_2 = k)$, as desired. □

## C.4  Proof of Lemma 2.3

*Proof.* Invariance follows by Lemma C.2 (point (b)), with $K_{NR} = P_{NR}$ and $(k,k') \in \mathcal{K}$ in place of $d \in [D]$. The condition (30) is satisfied as shown in the proof of Lemma 2.1. Consider then irreducibility. For ease of notation, we use the notation $\mathcal{X} = [K]^n \times \mathcal{K}$ and $x = (c,v) \in \mathcal{X}$. If $\pi(c) > 0$, this implies that $\pi(c_i = k \mid c_{-i}) > 0$ for every $c_{-i} \in [K]^{n-1}$. Combining this with the fact that $p_c(k,k') > 0$ for every $(k,k') \in \mathcal{K}$ such that $n_k(c) + n_{k'}(c) > 0$, we get that for every pair $x \neq x' \in \mathcal{X}$ there exists a $T \in \mathbb{N}$ and a sequence $x = x^{(0)}, x^{(1)}, \ldots, x^{(T)} = x'$ such that $P_{NR}(x^{(t-1)}, x^{(t)}) > 0$ for every $t = 1, \ldots, T-1$. Thus, $P_{NR}$ is irreducible. Moreover, if $\xi > 0$ it is immediate to

deduce that $P_{\text{NR}}$ is aperiodic. Uniform ergodicity then follows from Levin and Peres (2017, Theorem 4.9).  $\square$

## C.5    Proof of Theorem 3.1

*Proof of Theorem 3.1.* The first inequality $\text{Var}(g, P_{\text{NR}}) \leq \text{Var}(g, P_{\text{R}})$ follows by point (c) of Lemma C.2, with $K_{NR} = P_{\text{NR}}$ and $K_R = P_{\text{R}}$.

In order to prove the other inequality in (14) it suffices to show that

$$(34) \qquad P_{\text{R}}(c, c') \geq \frac{1}{c(K)} P_{\text{MG}}(c, c'), \quad c \neq c' \in [K]^n,$$

by, e.g., Theorem 2 in Zanella (2020).

In order to prove (34), fix $c$ and $c'$ such that $c = (c_{-i}, k)$ and $c' = (c_{-i}, k')$ with $i \in [n]$ and $(k, k') \in \mathcal{K}$. Indeed for every other pair $(c, c')$ such that $c \neq c'$ we have that $P_{\text{R}}(c, c') = P_{\text{MG}}(c, c') = 0$. By definition of $P_{\text{MG}}$ and $P_{\text{R}}$ we have

$$P_{\text{MG}}(c, c') = \frac{1}{n} \pi(c_i = k' \mid c_{-i}).$$

and

$$P_{\text{R}}(c, c') = \frac{1}{2n_k} \frac{n_k + n_{k'}}{(K-1)n} \min\left\{1, \frac{n_k}{n_{k'} + 1} \frac{\pi(c_i = k' \mid c_{-i})}{\pi(c_i = k \mid c_{-i})}\right\},$$

where $n_j = n_j(c)$ for every $j \in [K]$ and $n_k \geq 1$ by definition of $c$. Thus

$$\begin{aligned}
P_{\text{R}}(c, c') &= \frac{1}{2(K-1)n} \min\left\{\frac{n_k + n_{k'}}{n_k}, \frac{n_k + n_{k'}}{n_{k'} + 1} \frac{\pi(c_i = k' \mid c_{-i})}{\pi(c_i = k \mid c_{-i})}\right\} \\
&\geq \frac{1}{2(K-1)n} \min\left\{1, \frac{\pi(c_i = k' \mid c_{-i})}{\pi(c_i = k \mid c_{-i})}\right\} \\
&= \frac{\pi(c_i = k' \mid c_{-i})}{2(K-1)n} \min\left\{\frac{1}{\pi(c_i = k' \mid c_{-i})}, \frac{1}{\pi(c_i = k \mid c_{-i})}\right\} \\
&\geq \frac{1}{2(K-1)} \frac{\pi(c_i = k' \mid c_{-i})}{n} = \frac{1}{2(K-1)} P_{\text{MG}}(c, c'),
\end{aligned}$$

which is exactly (34).  $\square$

## C.6    Proof of Proposition 3.3

*Proof.* Let $\tilde{P}_{\text{MG}}$ be the $\pi(c, \boldsymbol{w}, \boldsymbol{\theta})$-reversible Markov kernel on $[K]^n \times \Theta^K \times \Delta_{K-1}$ that, given $(c^{(t)}, \boldsymbol{w}^{(t)}, \boldsymbol{\theta}^{(t)})$ generates $(c^{(t+1)}, \boldsymbol{w}^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$ by

$$c^{(t+1)} \sim P_{\text{MG}}\left(c^{(t)}, \cdot\right), \quad (\boldsymbol{w}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}) \sim \pi\left(\boldsymbol{w}, \boldsymbol{\theta} \mid c = c^{(t)}\right).$$

By construction, $\mathrm{Var}(g, P_{\mathrm{MG}}) = \mathrm{Var}(g, \tilde{P}_{\mathrm{MG}})$ for any $g$ that is a function of $c$ alone, because the marginal process on $[K]^n$ induced by $\tilde{P}_{\mathrm{MG}}$ is a Markov chain with kernel $P_{\mathrm{MG}}$.

We now compare $\tilde{P}_{\mathrm{MG}}$ and $P_{\mathrm{CD}}$. Let

$$\langle f, g \rangle_\pi = \int_{\mathcal{X}} f(x) g(x) \pi(\mathrm{d}x), \quad \mathcal{X} = [K]^n \times \Theta^K \times \Delta_{K-1},$$

be the $L^2(\pi)$ inner product. Then for any $g \in L^2(\pi)$ and $(c, \boldsymbol{w}, \boldsymbol{\theta}) \sim \pi$ we have

$$
\begin{aligned}
\langle (I &- P_{\mathrm{CD}})g, g \rangle_\pi \\
&= \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c_{-i}, \boldsymbol{w}, \boldsymbol{\theta})] + \frac{1}{n+1} \mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c)] \\
&\leq \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c_{-i})] + \frac{1}{n+1} \sum_{i=1}^{n} \frac{\mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c_{-i})]}{n} \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c_{-i})] = \langle (I - \tilde{P}_{\mathrm{MG}})g, g \rangle_\pi \,,
\end{aligned}
$$

(35)

where the middle inequality follows from the fact that

$$\mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c)] = \mathbb{E}\left[\mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c) \mid c_{-i}]\right] \leq \mathbb{E}[\mathrm{Var}(g(c, \boldsymbol{w}, \boldsymbol{\theta}) \mid c_{-i})],$$

for every $i = 1, \ldots, n$ by the law of total variance. We thus have $\langle (I - P_{\mathrm{CD}})g, g \rangle_\pi \leq \langle (I - \tilde{P}_{\mathrm{MG}})g, g \rangle_\pi$ for every $g \in L^2(\pi)$, which implies $\mathrm{Var}(g, \tilde{P}_{\mathrm{MG}}) \leq \mathrm{Var}(g, P_{\mathrm{CD}})$ for all $g$ (see e.g. the proof of Tierney, 1998, Theorem 4). We thus have $\mathrm{Var}(g, P_{\mathrm{MG}}) = \mathrm{Var}(g, \tilde{P}_{\mathrm{MG}}) \leq \mathrm{Var}(g, P_{\mathrm{CD}})$ for all $g$ for functions $g$ that depend only on $c$. $\qquad\square$

## C.7 Proof of Theorem 4.1

*Proof.* By (15) for every $\boldsymbol{x} \in \Delta_{K-1}$ we have that, as $n \to \infty$,

$$
\begin{aligned}
\mathbb{E}\left[X_{t+1,k} - x_k \mid \boldsymbol{X}_t = \boldsymbol{x}\right] &= \frac{1 - x_k}{n} \frac{\alpha_k + n x_k}{|\boldsymbol{\alpha}| + n - 1} - \frac{x_k}{n} \frac{|\boldsymbol{\alpha}| - \alpha_k + n(1 - x_k)}{|\boldsymbol{\alpha}| + n - 1} \\
&= \frac{\alpha_k - |\boldsymbol{\alpha}| x_k}{n(|\boldsymbol{\alpha}| + n - 1)} = \frac{2}{n^2} \left[\frac{\alpha_k}{2} - |\boldsymbol{\alpha}| \frac{x_k}{2} + o(1)\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[(X_{t+1,k} - x_k)^2 \mid \boldsymbol{X}_t = \boldsymbol{x}\right] &= \frac{1 - x_k}{n^2} \frac{\alpha_k + n x_k}{|\boldsymbol{\alpha}| + n - 1} + \frac{x_k}{n^2} \frac{|\boldsymbol{\alpha}| - \alpha_k + n(1 - x_k)}{|\boldsymbol{\alpha}| + n - 1} \\
&= \frac{2}{n^2} \left[x_k(1 - x_k) + o(1)\right]
\end{aligned}
$$

39

and

$$\mathbb{E}\left[(X_{t+1,k} - x_k)\left(X_{t+1,k'} - x_{k'}\right) \mid \boldsymbol{X}_t = \boldsymbol{x}\right] = \frac{-x_k}{n^2}\frac{\alpha_{k'} + nx_{k'}}{|\boldsymbol{\alpha}| + n - 1} + \frac{x_{k'}}{n^2}\frac{\alpha_k + nx_k}{|\boldsymbol{\alpha}| + n - 1}$$
$$= \frac{2}{n^2}\left[-x_k x_{k'} + o(1)\right],$$

and $n^2\mathbb{E}\left[(X_{t+1,k} - x_k)^3 \mid \boldsymbol{X}_t = \boldsymbol{x}\right] = o(1)$ for $k \neq k' \in [K]$. By a second-order Taylor expansion, this means that

$$\sup_{\boldsymbol{x} \in \Delta_{K+1}} |\mathbb{E}\left[g(\boldsymbol{X}_{t-1}) \mid \boldsymbol{X}_t = \boldsymbol{x}\right] - g(\boldsymbol{x}) - \mathcal{A}g(\boldsymbol{x})| \to 0,$$

as $n \to \infty$ for every $g$ twice differentiable real-valued function. The result then follows by Corollary 8.9 in (Ethier and Kurtz, 1986, Chapter 4). $\square$

## C.8 Proof of Theorem 4.4

*Proof.* Fix $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{v}) \in E_M \times V$. Notice that

$$\mathbb{E}\left[X^{(M)}_{t+1,k} - x_k \mid \boldsymbol{X}^{(M)}_t = \boldsymbol{x}, \boldsymbol{V}^{(M)}_t = \boldsymbol{v}\right]$$
$$= \left(1 - \frac{\xi}{n}\right)\frac{1}{n}\left[\sum_{k' : v_{k',k}=+1}\frac{x_k + x_{k'}}{K - 1}\alpha(\boldsymbol{x}, k, k') - \sum_{k' : v_{k',k}=-1}\frac{x_k + x_{k'}}{K - 1}\alpha(\boldsymbol{x}, k', k)\right]$$
$$- \frac{\xi}{n}\frac{1}{n}\left[\sum_{k' : v_{k',k}=+1}\frac{x_k + x_{k'}}{K - 1}\alpha(\boldsymbol{x}, k', k) - \sum_{k' : v_{k',k}=-1}\frac{x_k + x_{k'}}{K - 1}\alpha(\boldsymbol{x}, k, k')\right],$$

where

$$\alpha(\boldsymbol{x}, k, k') = \min\left\{1, \left(\frac{\alpha_k + nx_k}{nx_k + 1}\right)\left(\frac{nx_{k'}}{\alpha_{k'} + nx_{k'} - 1}\right)\right\}$$
$$= 1 - \frac{\beta(x_k, x_{k'}, v_{k',k})}{n} + o\left(\frac{1}{n}\right),$$

from which we deduce that

$$(36) \qquad \mathbb{E}\left[X^{(M)}_{t+1,k} - x_k \mid \boldsymbol{X}^{(M)}_t = \boldsymbol{x}, \boldsymbol{V}^{(M)}_t = \boldsymbol{v}\right] = \frac{\Phi_k(\boldsymbol{z})}{n} + o\left(\frac{1}{n}\right).$$

Similarly we get that

$$(37) \qquad \mathbb{E}\left[\left(X^{(M)}_{t+1,k} - x_k\right)\left(X^{(M)}_{t+1,k'} - x_{k'}\right) \mid \boldsymbol{X}^{(M)}_t = \boldsymbol{x}, \boldsymbol{V}^{(M)}_t = \boldsymbol{v}\right] = o\left(\frac{1}{n}\right),$$

for every $(k, k') \in [K]^2$. Moreover
(38)
$$
\mathbb{E}\left[g\left(\boldsymbol{x}, \boldsymbol{V}_{t+1}^{(M)}\right) - g(\boldsymbol{x}, \boldsymbol{v}) \mid \boldsymbol{X}_t^{(M)} = \boldsymbol{x}, \boldsymbol{V}_t^{(M)} = \boldsymbol{v}\right]
$$

$$
= \sum_{k \neq k'} [g(\boldsymbol{z}_{(\boldsymbol{k},\boldsymbol{k'})}) - g(\boldsymbol{z})] \frac{x_k + x_{k'}}{2(K-1)} \left[ \left(1 - \alpha(\boldsymbol{x}, k, k')\right) \left(1 - \frac{\xi}{n}\right)^2 + \alpha(\boldsymbol{x}, k, k') \left(1 - \frac{\xi}{n}\right) \frac{\xi}{n} \right.
$$

$$
\left. + \left(1 - \alpha(\boldsymbol{x}, k', k)\right) \left(\frac{\xi}{n}\right)^2 + \alpha(\boldsymbol{x}, k', k) \frac{\xi}{n} \left(1 - \frac{\xi}{n}\right) \right]
$$

$$
= \sum_{k \neq k'} \frac{g(\boldsymbol{z}_{(\boldsymbol{k},\boldsymbol{k'})}) - g(\boldsymbol{z})}{n} \frac{x_k + x_{k'}}{2(K-1)} \left[\beta(x_k, x_{k'}, v_{k,k'}) + 2\xi\right] + o\left(\frac{1}{n}\right)
$$

$$
= \frac{\lambda(\boldsymbol{z})}{n} \sum_{k \neq k'} q(k, k') \left[g(\boldsymbol{z}_{(\boldsymbol{k},\boldsymbol{k'})}) - g(\boldsymbol{z})\right] + o\left(\frac{1}{n}\right).
$$

By a Taylor expansion we have that

$$
\mathbb{E}\left[g(\boldsymbol{Z}_{t+1}^{(M)}) \mid \boldsymbol{Z}_t^{(M)} = \boldsymbol{z}\right] - g(\boldsymbol{z}) = \mathbb{E}\left[g(\boldsymbol{x}, \boldsymbol{V}_{t+1}^{(M)}) \mid \boldsymbol{Z}_t^{(M)} = \boldsymbol{z}\right] - g(\boldsymbol{z})
$$

$$
+ \sum_{k=1}^K \mathbb{E}\left[\left(X_{t+1,k}^{(M)} - x_k\right) \mid \boldsymbol{Z}_t^{(M)} = \boldsymbol{z}\right] \frac{\partial}{\partial z_{1,k}} g(\boldsymbol{z}) + o\left(\frac{1}{n}\right),
$$

that, combined with (36), (37) and (38), implies

$$
\sup_{\boldsymbol{z} \in E_M \times [K]^2} \left| \mathbb{E}\left[g(\boldsymbol{Z}_{t+1}^{(M)}) \mid \boldsymbol{Z}_t = \boldsymbol{z}\right] - g(\boldsymbol{z}) - \mathcal{B}g(\boldsymbol{z}) \right| \to 0,
$$

as $n \to \infty$ for every $g : \Delta_{K-1} \times [K]^2 \to \mathbb{R}$ twice continuously differentiable in the first argument. The result then follows by Corollary 8.9 in Ethier and Kurtz (1986, Chapter 4). $\qquad \square$