Markov Interacting Importance Samplers

Eduardo F. Mendes Marcel Scharth Robert Kohn

June 25, 2015

Abstract

We introduce a new Markov chain Monte Carlo (MCMC) sampler called the Markov Interacting Importance Sampler (MIIS). The MIIS sampler uses conditional importance sampling (IS) approximations to jointly sample the current state of the Markov Chain and estimate conditional expectations, possibly by incorporating a full range of variance reduction techniques. We compute Rao-Blackwellized estimates based on the conditional expectations to construct control variates for estimating expectations under the target distribution. The control variates are particularly efficient when there are substantial correlations between the variables in the target distribution, a challenging setting for MCMC. An important motivating application of MIIS occurs when the exact Gibbs sampler is not available because it is infeasible to directly simulate from the conditional distributions. In this case the MIIS method can be more efficient than a Metropolis-within-Gibbs approach. We also introduce the MIIS random walk algorithm, designed to accelerate convergence and improve upon the computational efficiency of standard random walk samplers. Simulated and empirical illustrations for Bayesian analysis show that the method significantly reduces the variance of Monte Carlo estimates compared to standard MCMC approaches, at equivalent implementation and computational effort.

Keywords: Bayesian inference; Control variate; Mixed Logit; PMCMC; Markov Modulated Poisson Process; Rao-Blackwellization; Variance reduction.

1 Introduction

This paper introduces *Markov interacting importance samplers* (MIIS), a general Markov Chain Monte Carlo (MCMC) algorithm that iterates by sampling the current state from a conditional importance sampling approximation to a target distribution. An importance sampling (IS) approximation consists of a set of weighted samples from a proposal distribution that approximates the target. Markov interacting importance samplers are conditional in the sense that the importance distribution may depend on the previous state of the Markov chain. The marginal distribution of the states converges to the target distribution for any number of importance samples at each iteration of the Markov chain; the algorithm does not induce an approximation error.

We adopt importance sampling as a basic tool from the perspective that it can be more efficient than a Metropolis-Hastings sampler based on an identical proposal. Importance sampling naturally incorporates the information from all generated samples, while standard Metropolis-Hastings estimates lose information from rejected draws. In addition, importance sampling estimates are based on independent samples and as a consequence the method is immediately amenable to a range of variance reduction techniques (such as antithetic sampling and stratified mixture sampling), as well as convenient to implement and parallelize. It is not standard practice in applied work to incorporate these features into Metropolis-Hastings approaches as they are more challenging to design and use efficiently in an MCMC framework. See for example Craiu and Lemieux (2007), Hammer and Tjelmeland (2008), Jacob et al. (2011), and Dellaportas and Kontoyiannis (2012).

Importance sampling can be efficient when we are able to construct numerically accurate and computationally fast approximations to a full target distribution. Richard and Zhang (2007), Hoogerheide et al. (2012) and Li et al. (2013) are recent contributions in this area that have led to the application of IS to challenging problems: see for example Liesenfeld et al. (2013) and Tran et al. (2014). We motivate MIIS by observing that even if the joint target density is intractable by global approximation, we can frequently obtain efficient importance samplers for the conditional distributions. MCMC methods provide a natural way of handling large dimensional problems by sampling from conditional distributions (Gibbs sampling) or by generating samples from complex target densities through local exploration. The MIIS algorithm leverages the advantages of importance sampling in this setting.

As a leading application, we consider the case in which it is not possible to implement an exact Gibbs sampler due to infeasibility of direct simulation from the conditional distributions. The MIIS method relies on IS approximations of the conditional distributions to sample the current state of the Markov Chain. The advantage of importance

sampling is that we can additionally use the approximation (that is, all the generated samples) to estimate conditional expectations, possibly by incorporating the full range of variance reduction methods available for standard importance sampling. We compute Rao-Blackwellized estimates based on the conditional expectations to construct control variates for estimating expectations under the target distribution. The control variates are particularly effective when there are substantial correlations between the variables in the target distribution. This is a challenging setting for standard MCMC approaches because the conditioning scheme may imply strong serial correlation in the Markov chain.

We introduce the general MIIS algorithm and present four examples that demonstrate its flexibility. The first two examples present the implementation of MIIS based on simple importance sampling targeting the full and conditional distributions. We derive conditions for the ergodicity and uniform ergodicity of the sampler. The third example introduces antithetic variables and is also uniformly ergodic under general conditions. The final example introduces the MIIS random walk algorithm, designed to accelerate convergence and improve upon the computational efficiency of standard random walk samplers. The random walk sampler is uniformly ergodic assuming that the importance weights are bounded. Ergodicity holds under milder constraints.

Our method relates to the Particle Gibbs (PG) algorithm developed for Bayesian inference in general state space models by Andrieu et al. (2010). The PG algorithm iteratively draws the latent state trajectories from its high-dimensional smoothing distribution using a particle filter approximation, and the parameters of the model from their conditionals given the state trajectories. Lindsten and Schön (2012), Lindsten et al. (2014b), Mendes et al. (2014) and Carter et al. (2014) present extensions, while Chopin and Singh (2013), Andrieu et al. (2013) and Lindsten et al. (2014a) study the theoretical aspects of the algorithm. We can show that the particle Gibbs algorithm is a particular type of MIIS. Compared to PG, the MIIS algorithm addresses a wider class of sampling problems and the use of variance reduction methods.

We illustrate Markov interacting importance samplers in a range of examples. We consider the estimation of the posterior mean for a Bayesian Mixed Logit model using the health dataset studied by Fiebig et al. (2010). The presence of unobserved heterogeneous preferences in this discrete choice model motivates the use of MCMC methods that iteratively sample the model parameters and the latent choice attribute weights conditional on each other. The results show that the MIIS algorithm with control variates increases efficiency in mean squared error by a factor of four to twenty compared to the Metropolis-within-Gibbs algorithm, which is a standard tool for problems that are not amenable to exact Gibbs sampling. We also implement the MIIS random walk importance sampler for carrying out posterior inference for Markov modulated Poisson

processes, a problem considered for example by Fearnhead and Sherlock (2006). Our analysis reveals four to hundredfold gains in efficiency over the standard random walk Metropolis algorithm and the multiple-try Metropolis algorithm of Liu et al. (2000). In this context, the improvements are mainly due to parallelization and better convergence of the Markov chain.

2 Markov Interacting Importance Samplers

To focus on the main ideas, we use densities in our mathematical discussion up to Section 6. We assume that the densities are defined with respect to measures that we leave unspecified for now. We provide a more precise treatment in Section 7 and the appendix.

2.1 Notation and basic definitions

This subsection presents some of the notation used in the article. We define the basic random variables on a set A that is a subset of Euclidean space. Suppose that $f(x)$ is a real function with $x \in A$. We take any density $\nu(x)$ on A to be with respect to some measure on A, which we denote as $\mathrm{d}x$. We define the expected value of f with respect to the density ν as

$$E_\nu(f) := \int f(x)\nu(x)dx \tag{1}$$

provided the integral exists.

In our article, $\pi(x)$ is the target density. We often can evaluate $\pi(x)$ only up to a constant of proportionality $m(x)$, with $\pi(x) = m(x)/Z_m$, where $Z_m = \int_A m(x)dx$ is the normalizing constant. Suppose that $x_i \in A, i = 1, \ldots, N$. Then, for $1 \leq i \leq j \leq N$, we define $i\!:\!j := \{i, i+1, \ldots, j\}$, $x_{i:j} := (x_i, \ldots, x_j)$ and $x_{\setminus k} := (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_N)$.

2.2 Conditional Importance Sampler

This section introduces the *conditional importance sampler* (CIS) which is the basic building block of the MCMC algorithms in this article. The CIS is motivated by the question: "*how to implement an importance sampler approximation to π that provides unbiased samples?*"The CIS is our solution to this problem. We go beyond simple importance sampler and construct a general framework that not only covers the simple importance sampling approximation with variance reduction techniques, but also extends the basic importance sampling paradigm, allowing local exploration of the target inside an MCMC setting, for instance, by using a random-walk approach.

At each iterate of an MCMC algorithm, the CIS constructs an empirical approximation to the target density $\pi(\cdot)$. It generates an auxiliary variable ξ and N particles $X_{1:N}$ conditional on the previous iterate y, in such a way that one particle X_k is generated through a Markov transition kernel and the other $N-1$ particles are generated conditional on X_k.

We now present a more precise description of the CIS. Let $\eta(\xi|y)$ be the conditional density of the auxiliary variable ξ, with $\xi, y \in A$, and take $\eta(\xi) = \int \eta(\xi|y)\pi(y)dy$ so that $\pi(y|\xi) = \eta(\xi|y)\pi(y)/\eta(\xi)$. Let $T(y, x; \xi)$ be the density of a Markov transition kernel from y to $x \in A$, conditional on ξ, that is reversible with respect to $\pi(y|\xi)$; i.e., $\pi(y|\xi)T(y, x; \xi) = \pi(x|\xi)T(x, y; \xi)$, or equivalently,

$$\pi(y)\eta(\xi|y)T(y, x; \xi) = \pi(x)\eta(\xi|x)T(x, y; \xi). \tag{2}$$

Given $\xi \in A$, let $\mathbf{q}(x_{1:N}|\xi)$ be a joint importance distribution with marginals $q_i(x_i|\xi)$ $(i = 1, \ldots, N)$. For any $1 \leq k \leq N$, define the conditional density

$$\mathbf{q}_{\backslash k}(x_{\backslash k}|x_k, \xi) := \frac{\mathbf{q}(x_{1:N}|\xi)}{q_k(x_k|\xi)}\,. \tag{3}$$

Definition 1 (Conditional Importance Sampler)**.** *For any given $y \in A$ and $1 \leq k \leq N$, the* Conditional Importance Sampler *generates $X_{1:N}, \xi|(y, k)$ from the probability distribution*

$$\Gamma^N(x_{1:N}, \xi|y, k) := \eta(\xi|y)T(y, x_k; \xi)\, \mathbf{q}_{\backslash k}(x_{\backslash k}|x_k, \xi). \tag{4}$$

The auxiliary variable η introduces dependence in the importance sampling approximation. Moreover, we can often choose the auxiliary density η so that $w_i(x; \xi)$ is bounded. For instance, the random-walk importance sampling algorithm chooses $\eta(\xi|x) = q(x|\xi) = \phi(|\xi - x|)$. The weights are $w_i(x; \xi) = m(x)$, which are bounded if $m(x)$ is bounded. The dependence on ξ can be easily dropped if one takes $\eta(\cdot|y) = \eta(\cdot)$ and each $q_i(\cdot|\xi) = q_i(\cdot)$. The Markov transition kernel $T(y, \cdot; \xi)$ can be taken as the identity kernel, i.e., $T(y, \cdot; \xi) = \delta(\cdot - y)$, which is our choice in Sections 3 and 6. A Metropolis-Hastings kernel targeting $\pi(\cdot|\xi)$ is also a valid choice.

The CIS generates $(X_{1:N}, \xi)$ using the following algorithm.

Algorithm 1 (Conditional Importance Sampler)**.** *Given (y, k),*

1. *sample $\xi \sim \eta(\xi|y)$;*
2. *sample $X_k \sim T(y, x_k; \xi)$; i.e., generate the particle x_k using the Markov kernel.*

3. *sample* $X_{\backslash k} \sim \mathbf{q}_{\backslash k}(x_{\backslash k}|x_k, \xi)$*; i.e., generate all the remaining particles conditional on* ξ *and the propagated particle* x_k*.*

From the output of the Conditional Importance Sampler we define the weights for $i = 1, \ldots, N$

$$W_i(x_{1:N}; \xi) := \frac{w_i(x_i; \xi)}{\sum_{j=1}^N w_j(x_j; \xi)} \quad \text{where} \quad w_i(x; \xi) := \frac{m(x)}{q_i(x|\xi)} \eta(\xi|x) \tag{5}$$

and let $\widehat{\pi}^N_{CIS} := \{(x_1, W_1(x_{1:N}, \xi)), \ldots, (x_N, W_N(x_{1:N}, \xi))\}$ be the empirical approximation to π. The weights depend on the marginals $q_i(\cdot|\xi)$ $(i = 1, \ldots, N)$ of $\mathbf{q}(x_{1:N}|\xi)$, the auxiliary distribution $\eta(\xi|\cdot)$ and the target distribution $\pi(\cdot) \propto m(\cdot)$. Based on $\widehat{\pi}^N_{CIS}$, we define the estimator of $E_\pi(f)$ as

$$\widehat{E}^N_{CIS}(f) := \sum_{i=1}^N W_i(x_{1:N}, \xi) f(x_i) = E_{\widehat{\pi}^N_{CIS}}(f). \tag{6}$$

Define the joint density

$$\widetilde{\pi}^N(k, y, x_{1:N}, \xi) := N^{-1} \pi(y) \Gamma^N(x_{1:N}, \xi|y, k). \tag{7}$$

Lemma 1 gives some fundamental properties of $\widetilde{\pi}^N(k, y, x_{1:N}, \xi)$ and shows that the expectation of $\widehat{E}^N_{CIS}(f)$ is $E_\pi(f)$ if the marginal distribution $\widetilde{\pi}^N(y, k) = N^{-1}\pi(y)$. We use $\widehat{E}^N_{CIS}(f)$, additively, within an MCMC scheme to construct unbiased estimators of $E_\pi(f)$. The unbiasedness property is critical for the variance reduction techniques in Section 5.

Theorem 1. *Suppose that* $E_\pi(|f|)$ *is finite,* (k, y) *is a sample from* $N^{-1}\pi(y)$*, and that* $(x_{1:N}, \xi)$ *is generated from* $\Gamma^N(x_{1:N}, \xi|y, k)$*. Then,*

(i) $\widetilde{\pi}^N(y) = \pi(y)$.

(ii)

$$\widetilde{\pi}^N(k, y|x_{1:N}, \xi) = \sum_{i=1}^N W_i(x_{1:N}, \xi) I(k = i) T(x_i, y; \xi), \tag{8}$$

or equivalently,

$$\widetilde{\pi}^N(K = i|x_{1:N}, \xi) = W_i(x_{1:N}, \xi) \quad \textit{and} \quad \widetilde{\pi}^N(y|x_{1:N}\xi, k) = T(x_k, y; \xi). \tag{9}$$

(iii) $E_{\widetilde{\pi}^N}\left(\widehat{E}^N_{CIS}(f)\right) = E_\pi(f)$.

Remark 1. *We now compare importance sampling to conditional importance sampling. In importance sampling, we draw particles $x_{1:N}$ from an importance or proposal density $\mathbf{q}(x_{1:N})$ with marginal densities $q_i(x_i)$ and calculate their importance weights*

$$W_i(x_{1:N}) := \frac{w_i(x_i)}{\sum_{j=1}^N w_j(x_j)}, \quad \textit{where} \quad w_i(x_i) := \frac{m(x_i)}{q_i(x_i)},$$

to obtain the approximation $\widehat{\pi}_{IS}^N := \{W_{1:N}(x_{1:N}), x_{1:N}\}$ to π. The IS sampling estimate of $E_\pi(f)$ is

$$\widehat{E}_{IS}^N(f) := \sum_{i=1}^N W_i(x_{1:N}) f(x_i) = E_{\widehat{\pi}_{IS}^N}(f) \tag{10}$$

In the simplest case, the particles $x_{1:N}$ are sampled independently from the same proposal distribution q, i.e., $q_1 = \cdots = q_N = q$ and $\mathbf{q}(x_{1:N}) = \prod_{i=1}^N q(x_i)$. Despite similarities, there fundamental differences between using $\widehat{\pi}_{CIS}^N$ and $\widehat{\pi}_{IS}^N$.

1. *The marginal distribution of a sample X from $\widehat{\pi}_{IS}^N$ is not $\pi(X)$, while the distribution of Y from $\widehat{\pi}_{CIS}^N$ is $\pi(Y)$. Similarly,*

$$E_q\left(\widehat{E}_\pi^{IS}(f)\right) \neq E_\pi(f), \tag{11}$$

   whereas $E_{\widetilde{\pi}^N}\left(\widehat{E}_{CIS}^N(f)\right) = E_\pi(f)$.

2. *The weights w_i in the CIS may depend on an auxiliary variable ξ, with density $\eta(\cdot|y)$, that incorporates past information in the proposal opening the possibility for using local proposals. Moreover, it can be used as a mechanism to bound the weights and provide more robust estimators.*

2.3 Markov Interacting Importance Sampling Algorithm

The MIIS algorithm simulates from the target distribution π on A. It iterates by first constructing a discrete approximation to π using the CIS, conditional on the previous state (y, k) of the Markov Chain, and then samples from the approximation. It requires specifying a joint proposal distribution $\mathbf{q}(x_{1:N}; \xi)$, an auxiliary distribution $\eta(\xi|y)$, and a Markov transition kernel $T(y, x; \xi)$.

Algorithm 2 (Markov Interacting Importance Sampler)**.** *Given $y^{(0)} \in A$ and $1 \leq k^{(0)} \leq N$, at step $t = 1, 2, \ldots$*

1. *Generate $\xi^{(t)}|y^{(t-1)} \sim \eta(\xi|y^{(t-1)})$.*

2. *Generate* $X^{(t)}_{k^{(t-1)}}|(y^{(t-1)}, \xi^{(t)}) \sim T\left(y^{(t-1)}, x^{(t)}_{k^{(t-1)}}; \xi^{(t)}\right)$.

3. *Generate*

$$X^{(t)}_{\setminus k^{(t-1)}} \left| \left(x^{(t-1)}_{k^{(t-1)}}, k^{(t-1)}, \xi^{(t)}\right) \sim \mathbf{q}_{\setminus k^{(t-1)}} \left(x^{(t)}_{\setminus k^{(t-1)}} \middle| x^{(t)}_{k^{(t-1)}}, k^{(t-1)}, \xi^{(t)}\right)\right. .$$

4. *For* $k = 1, \ldots, N$, *calculate*

$$w_k\left(x^{(t)}_k; \xi\right) = \frac{m(x^{(t)}_k)}{q_k(x^{(t)}_k|\xi^{(t)})}\eta(\xi^{(t)}|x^{(t)}_k), \quad \textit{and} \quad W_k(x^{(t)}_{1:N}, \xi^{(t)}) = \frac{w_k\left(x^{(t)}_k; \xi\right)}{\sum_{j=1}^N w_j\left(x^{(t)}_j; \xi\right)} .$$

Draw $K^{(t)} = k|(x^{(t)}_{1:N}, \xi^{(t)})$ *with probability* $W_k(x^{(t)}_{1:N}, \xi^{(t)})$.

5. *Generate* $Y^{(t)}|(x^{(t)}_{1:N}, k^{(t)}, \xi^{(t)}) \sim T\left(x^{(t)}_{k^{(t)}}, x^{(t)}; \xi^{(t)}\right)$.

We divide the algorithm into two blocks. The first block consists of steps 1 to 3 and uses the CIS to draw an approximation to π. It corresponds to Algorithm 1 in Section 2.2. The second block consists of steps 4 and 5 and draws an element from this approximation. It corresponds to part (ii) of Theorem 1.

The MIIS algorithm is a Gibbs sampler on an augmented space that contains all variables sampled in the CIS step, i.e., it is a Gibbs sampler targeting (7). It also follows that if $(k^{t-1}, y^{(t-1)}) \sim N^{-1}\pi(\cdot)$, the marginal distribution of $y^{(t)}$ is the original target π; the MIIS algorithm generates samples from π without the approximation error induced by the CIS step.

Theorem 2 (Target Distribution)**.** *The Markov Interacting Importance Sampler is a Gibbs sampler targeting the augmented density* (7) *that has* $\pi(y)$ *as a marginal density.*

3 Examples

This section illustrates the MIIS methodology in three useful examples. For simplicity, the Markov transition density is set to the identity density, i.e., $T(y, x; \xi) = \delta_y(x)$, which denotes a density in$x$ that integrates to 1 and which is zero exact at $x = y$; we will sometimes write it as $\delta(x - y)$. We do not use the auxiliary variable ξ in the first two examples, which is equivalent to assuming that $\eta(\xi|x) = \eta(\xi)$ and $\mathbf{q}(x_{1:N}|\xi) = \mathbf{q}(x_{1:N})$. Section 7.2 gives formal convergence results for all three examples.

3.1 Simple Importance Sampling

This specification corresponds to the iterated Sampling Importance Resampling algorithm (i-SIR) in Andrieu et al. (2013). In importance sampling algorithms we generate particles independently from importance distributions $q_i(x) = q(x)$ $(i = 1, \ldots, N)$, i.e., $X_{1:N} \sim \prod_{i=1}^{N} q(x_i)$. Hence $\mathbf{q}(x_{1:N}|\xi) = \prod_{i=1}^{N} q(x_i)$ and

$$\mathbf{q}_{\backslash k}(x_{\backslash k}|x_k, k, \xi) = \prod_{i \neq k}^{N} q(x_i).$$

The CIS in this case is

$$\Gamma^N(x_{1:N}\xi|y, k) = \eta(\xi)\delta(y - x_k) \prod_{i \neq k}^{N} q(x_i).$$

Algorithm 3 follows from Algorithm 2.

Algorithm 3 (MIIS with Simple Importance Sampling)**.** *Given* $y^{(t-1)}$ *and* $k^{(t-1)} = k$*,*

1. *Generate* $X_i^{(t)} \sim q(x)$*, for* $i = \{1{:}N\} \setminus k$*, and set* $x_k^{(t)} = y^{(t-1)}$*.*
2. *Draw* $K^{(t)} = k|x_{1:N}^{(t)}$ *with probability proportional to* $w_k(x_k^{(t)}) = m(x_k^{(t)})/q(x_k^{(t)})$*.*
3. *Set* $y^{(t)} = x_{k^{(t)}}^{(t)}$*.*

3.2 Importance Sampling with Antithetic Variables

In the importance sampling literature, the method of antithetic variables consists of drawing perfectly negatively correlated particles to reduce the variance of the Monte Carlo estimate. We can use this method within the MIIS framework. The importance sampler with antithetic variables draws the particles in pairs from a proposal distribution. Suppose that N is even. For $k \leq N/2$, let $q_k(x_k)$ be the density of x_k with corresponding cumulative distribution function $Q_k(\cdot)$ and let $x_{N/2+k} = Q_k^{-1}(1 - Q_k(x_k))$, where Q_k^{-1} is the inverse of Q_k. We write the joint density of $x_k, x_{N/2+k}$ as

$$q_{k,N/2+k}(x_k, x_{N/2+k}) = q_k(x_k)\delta_{Q_k^{-1}(1-Q_k(x_k))}(x_{N/2+k}).$$

The marginals are $q_k(x) = q_{N/2+k}(x)$ and the conditional density of X_k given $x_{N/2+k}$ is $q_k(x_k|x_{N/2+k}) = \delta_{Q_k^{-1}(1-Q_k(x_{N/2+k}))}(x_k)$. For notational simplicity assume $k \leq N/2$. We

sample the particle system given (x_k, k) from

$$\begin{aligned}\mathbf{q}_{\backslash k}(x_{\backslash k}|x_k,\xi,k) &= \delta_{Q_k^{-1}(1-Q_k(x_k))}(x_{N/2+k})\prod_{i\neq k}^{N/2} q_{i,N/2+i}(x_i,x_{N/2+i}) \\ &= \frac{\prod_{i=1}^{N/2} q_{i,N/2+i}(x_i,x_{N/2+i})}{q_k(x_k)} \\ &= \frac{\mathbf{q}(x_{1:N})}{q_k(x_k)},\end{aligned}$$

and the CIS is

$$\Gamma^N(x_{1:N}\xi|y,k) = \eta(\xi)\delta_y(x_k)\frac{\prod_{i=1}^{N/2} q_{i,N/2+i}(x_i,x_{N/2+i})}{q_k(x_k)}.$$

Algorithm 4 (MIIS with Antithetic Variables)**.** *Given $y^{(t-1)}$ and $k^{(t-1)} = k$,*

1. *Generate $(X_i^{(t)}, X_{N/2+i}) \sim q_{i,N/2+i}(x_i, x_{N/2+i})$, for $i = \{1{:}N/2\} \setminus k$.*
2. *If $k \leq N/2$, set $x_k^{(t)} = y^{(t-1)}$, and $x_{N/2+k} = Q_k^{-1}(1 - Q_k(x_k^{(t)}))$. If $k > N/2$, set $x_k^{(t)} = y^{(t-1)}$, and $x_{k-N/2} = Q_{k-N/2}^{-1}(1 - Q_{k-N/2}(x_k^{(t)}))$.*
3. *Draw $K^{(t)} = k|x_{1:N}^{(t)}$ with probability proportional to $m(x_k^{(t)})/q_k(x_k^{(t)})$.*
4. *Set $y^{(t)} = x_{k^{(t)}}^{(t)}$.*

3.3 Random Walk Importance Sampler

The random walk importance sampler draws particles from a symmetric proposal dependent on its past. The advantage is that the method bounds the weights by construction. The random walk proposal performs local exploration around the auxiliary variable ξ, which we sample conditionally on the previous state.

Let $q(\cdot|y) = \eta(\cdot|y) = \phi(\cdot - y)$ denote the proposal functions for q_i and η. Then

$$\mathbf{q}_{\backslash k}(x_{\backslash k}|x_k,k,\xi) = \prod_{i\neq k}^{N}\phi(x_i-\xi)$$

The CIS is

$$\Gamma^N(x_{1:N},\xi|y,k) = \delta_y(x_k)\phi(\xi-x_k)\prod_{i\neq k}^{N}\phi(x_i-\xi).$$

The random walk importance sampler bounds the weights if $m(x)$ is bounded. The sampling algorithm follows from Algorithm 2

Algorithm 5 (MIIS with Random Walk proposal)**.** *Given* $x^{(t-1)}$ *and* $k^{(t-1)} = k$*,*

1. *Generate* $\xi^{(t)}|y^{(t-1)} \sim \phi(\xi - x^{(t-1)})$
2. *Generate* $X_i^{(t)} \sim \phi(x - \xi^{(t)})$*, for* $i = \{1{:}N\} \setminus k$*, and set* $x_k^{(t)} = y^{(t-1)}$*.*
3. *Draw* $K^{(t)} = k|x_{1:N}^{(t)}$ *with probability proportional to* $m(x_k^{(t)})$*.*
4. *Set* $y^{(t)} = x_{k^{(t)}}^{(t)}$*.*

4 MIIS Targeting Conditional Distributions

This section shows how to use use the MIIS algorithm within a Gibbs sampling framework. We use the following notation. Suppose we partition $x \in A$ as $\{x(1), \ldots, x(d)\}$. Then, for $1 \leq s \leq t \leq d$, $x(s{:}t) := \{x(s), x(s+1), \ldots, x(t)\}, x_i(s{:}t) := \{x_i(s), \ldots, x_i(t)\}$, etc. We define $A_s := \{x(s){:}\, x \in A\}$ and $A_{\setminus s} := \{x(\setminus s){:}\, x \in A\}$. For a density $\nu(x)$, $x \in A$, we define the conditional density $\nu_s(x(s)|x(\setminus s)) := \nu(x)/\nu(x(\setminus s))$ and the conditional expectation

$$E_{\nu_s(\cdot|x(\setminus s))}(f) := \int_{A_s} f(x)\nu_s(x(s)|x(\setminus s))\mathrm{d}x(s). \tag{12}$$

4.1 Conditional Importance Sampler for conditional distributions

The CIS for conditional distributions is similar to the CIS in Section 2.2, but now targets $\pi_s(x(s)|x(\setminus s))$, $s = 1, \ldots, d$. Given $y \in A$, $s \in \{1{:}d\}$ and $k_s \in \{1{:}N\}$, let $\eta_s(\xi(s)|y(s), y(\setminus s))$ be the density of the auxiliary variable $\xi(s)$, conditional on y. Let $T_s(y(s), x_{k_s}(s); \xi(s), y(\setminus s))$ be a the density of a Markov transition kernel, conditional on $(\xi(s), y(\setminus s))$, that is reversible with respect to $\pi_s(y(s)|\xi(s), y(\setminus s)) \propto \pi_s(y(s)|y(\setminus s))\eta_s(\xi(s)|y(s), y(\setminus s))$.

Given $\xi(s)$ and $y(\setminus s)$, let $\mathbf{q}_s(x_{1:N}(s)|\xi(s), y(\setminus s))$ be a joint importance density with marginals $q_{s,i}(x_i(s)|\xi(s), y(\setminus s))$ $(i = 1, \ldots, N)$, and

$$\mathbf{q}_{s,\setminus k_s}(x_{\setminus k_s}(s)|x_{k_s}, \xi(s), y(\setminus s)) := \frac{\mathbf{q}_s(x_{1:N}(s)|\xi(s), y(\setminus s))}{q_{s,k_s}(x_{k_s}(s)|\xi(s), y(\setminus s))}. \tag{13}$$

Definition 2 (Conditional Importance Sampler for conditional distributions:)**.** *For* $1 \leq s \leq d$*,* $y \in A$*, and* $k_s \in \{1{:}N\}$*, the* Conditional Importance Sampler for conditional

distributions *generates* $X_{1:N}(s), \xi(s)|(y(s), k_s, y(\backslash s))$ *from the probability distribution*

$$\Gamma_s^N(x_{1:N}(s), \xi(s)|y(s), k_s, y(\backslash s)) = \eta_s(\xi(s)|y(s), y(\backslash s))\, T_s(y(s), x_{k_s}(s); \xi(s), y(\backslash s)) \times \mathbf{q}_{s,\backslash k_s}(x_{\backslash k_s}(s)|x_{k_s}(s), k_s, \xi(s), y(\backslash s)). \quad (14)$$

In the CIS for conditional densities, we first generate $\xi(s)$, then we generate $x_{k_s}(s)$ conditional on $\xi(s)$, and finally the remaining particles $x_{\backslash k_s}(s)$ conditional $\xi(s)$ and $x_{k_s}(s)$

Suppose we express the target $\pi_s(x(s)|x(\backslash s)) \propto m_s(x(s)|x(\backslash s))$, where we can evaluate $m_s(x(s)|x(\backslash s))$. From the output of the CIS for conditional distributions, we define the weights

$$W_{s,i}(x_{1:N}(s); \xi(s)|y(\backslash s)) = \frac{w_{s,i}(x_i(s); \xi(s)|y(\backslash s))}{\sum_{j=1}^N w_{s,j}(x_j(s); \xi(s)|y(\backslash s))}, \quad (15)$$

where

$$w_{s,i}(x_i(s); \xi(s)|y(\backslash s)) = \frac{m_s(x_i(s)|y(\backslash s))}{q_{s,i}(x_i(s)|\xi(s), y(\backslash s))}\, \eta_s(\xi(s)|x_i(s), y(\backslash s)) \quad (16)$$

and consider $\widehat{\pi}_{s,CIS}^N(\cdot|y(\backslash s)) := \{(W_{s,1}, x_1(s)), \ldots, (W_{s,N}, x_N(s))\}$ as an empirical approximation of $\pi_s(\cdot|y(\backslash s))$. Based on $\widehat{\pi}_{s,CIS}^N$, we define the estimator of $E_{\pi_s(\cdot|y(\backslash s))}(f)$ as

$$\widehat{E}_{s,CIS}^N(f|y(\backslash s)) := \sum_{i=1}^N W_{s,i}(x_i(s); \xi(s), y(\backslash s)) f(x_i(s), y(\backslash s)) = E_{\widehat{\pi}_{s,CIS}^N(\cdot|y(\backslash s))}(f). \quad (17)$$

Analogously to the CIS, define the joint density of $(K_s, Y(s), X_{1:N}(s), \xi(s)$ conditional on $Y(\backslash s)$ as

$$\widetilde{\pi}_s^N(k_s, y(s), x_{1:N}(s), \xi(s)|y(\backslash s)) := \frac{\pi_s(y(s)|y(\backslash s))}{N} \Gamma_s^N(x_{1:N}(s), \xi(s)|y(s), k_s, y(\backslash s)). \quad (18)$$

Lemma 3 gives some properties of the density (18) and shows that if $(k_s, y(\backslash s))$ is generated from $N^{-1}\pi_s(y(s))|y(\backslash s))$ then the expectation of $\widehat{E}_{s,CIS}^N(f)$ is $E_{\pi_s(\cdot|y(\backslash s))}(f)$.

Theorem 3. *Suppose* $(k_s, y(s))$ *be a sample from* $N^{-1}\pi_s(y(s))|y(\backslash s))$*, and* $(x_{1:N}(s), \xi(s))$ *a sample from* $\Gamma_s^N(x_{1:N}(s), \xi(s)|y(s), k_s, y(\backslash s))$*. Then, conditional on* $y(\backslash s)$*,*

(i) $\widetilde{\pi}_s^N(y(s)) = \pi_s(y(s))$.

(ii) The conditional density of $k_s, y(s)$ *given* $x_{1:N}(s), \xi(s)$ *is*

$$\widetilde{\pi}_s^N(k_s, y(s)|x_{1:N}(s), \xi(s)) = W_{s,k_s} T(x_{s_k}(s), y(s); \xi(s))$$

or equivalently

$$\widetilde{\pi}_s^N(k_s|x_{1:N}(s),\xi(s)) = W_{s,k_s} \quad \textit{and}$$
$$\widetilde{\pi}_s^N(y(s)|x_{1:N}(s),\xi(s),k_s) = T_s(x_{k_s}(s),y(s);\xi(s)).$$

(iii) $E_{\widetilde{\pi}_s^N(\cdot|y(\backslash s))}\left(\widehat{E}_{s,CIS}^N(f)\right) = E_{\pi_s(\cdot|y(\backslash s))}(f)$.

4.2 The Markov Interacting Importance Sampler within Gibbs

The algorithm extends the MIIS sampler targeting the full density. It simulates sequentially from the conditional distributions $\pi_1(y(1)|y(\backslash 1)),\ldots,\pi_d(y(d)|y(\backslash d))$, using the CIS approximation to the conditionals. The method is an alternative to the Metropolis-within-Gibbs algorithm that is is suitable for the application of the variance reduction techniques in Section 5. The MIIS within Gibbs sampler requires the specification of joint proposal distributions $\{\mathbf{q}_s(x_{1:N}(s)|\xi(s),y(s),y(\backslash s)\}$, auxiliary distributions $\{\eta_s(\xi(s)|y(s),y(\backslash s))\}$, and Markov transition kernels $\{T_s(y(s),x_{k_s}(s);\xi(s),y(\backslash s))\}$, for each $s=1,\ldots,d$. The general form of the MIIS Gibbs sampler is given by Algorithm 6

Algorithm 6 (The Markov Interacting Importance Sampler within Gibbbs)**.** *Given $y^{(0)}\in A$ and $1\leq k_s^{(0)}\leq N$, $s=1,\ldots,d$, the algorithm at step $t=1,2,\ldots,$ is described as follows, with all terms conditional on $y^{(t)}(1{:}s{-}1)$ and $y^{(t-1)}(s{+}1{:}d)$.*

1. *For $s=1,\ldots,d$,*

   1.1. *Generate $\xi^{(t)}(s)\sim\eta_s(\xi(s)|y^{(t-1)}(s))$.*

   1.2. *Generate*
   $$X_{k_s^{(t-1)}}^{(t)}(s)\sim T\left(y^{(t-1)}(s),x_{k_s^{(t-1)}}^{(t)}(s);\xi^{(t)}(s)\right).$$

   1.3. *Generate*
   $$X_{\backslash k_s^{(t-1)}}^{(t)}(s)\sim \mathbf{q}_{s,\backslash k_s^{(t-1)s}}\left(x_{\backslash k_s^{(t-1)}}^{(t)}(s)\middle|x_{k_s^{(t-1)}}^{(t)}(s),k_s^{(t-1)},\xi^{(t)}(s)\right),$$
   conditional on $x_{k_s^{(t-1)}}^{(t)}(s),k_s^{(t-1)},\xi^{(t)}(s),y^{(t)}(\backslash s)$.

   1.4. *Draw $K_s^{(t)}=k|(x_{1:N}^{(t)}(s),\xi^{(t)}(s))$ with probability proportional to*
   $$w_{s,k}\left(x_k^{(t)}(s);\xi^{(t)}(s)\right)=\frac{m_s\left(x_k^{(t)}(s)\right)\eta_s\left(\xi^{(t)}(s)|x_k^{(t)}(s)\right)}{q_{s,k}\left(x_k^{(t)}(s)|\xi^{(t)}(s)\right)}.$$

1.5. Generate

$$Y^{(t)}(s) \sim T\left(x^{(t)}_{k^{(t)}}(s), y^{(t)}(s); \xi^{(t)}(s)\right).$$

2. Set $y^{(t)} = (y^{(t)}(1), \ldots, y^{(t)}(d))'$.

For each partition $s = 1, \ldots, d$, the algorithm iterates as in the MIIS algorithm. Steps 1.1 – 1.3 construct an approximation $\widehat{\pi}^N_{s,CIS}$ to $\pi_s(\cdot|y(\backslash s))$. Steps 1.4 and 1.5 then draw an element from this approximation. As before, the MIIS for conditional distributions is a Gibbs sampler on an augmented space that contains all variables sampled in the CIS step. It also follows that the marginal distribution of $y^{(t)}$ is the original target π. Theorem 4 shows the augmented target distribution and that it generates samples from π.

Theorem 4 (Target Distribution)**.** *The Markov Interacting Importance Sampler is a Gibbs sampler targeting the augmented distribution given by*

$$\widetilde{\pi}^N(y, \xi, x_{1:N}(1), \ldots, x_{1:N}(d), k_{1:d}) = \frac{\pi(y)}{N^d} \prod_{s=1}^{d} \Gamma^N_s(x_{1:N}(s), \xi(s)|y(s), k_s, y(\backslash s)), \tag{19}$$

and has $N^{-d}\pi(y)$ *as a marginal distribution of* $(k_{1:d}, y)$.

4.3 Example: MIIS within Gibbs with Simple Importance Sampling

The MIIS sampler takes the conditional distributions in the Gibbs sampler as the target distributions for the conditional importance samplers. Suppose that we use a simple importance sampling algorithm to construct the CIS approximation. Then, for each $s = 1, \ldots, d$,

$$\Gamma^N_s(x_{1:N}(s), \xi(s)|y(s), k_s, y(\backslash s)) = \eta(\xi(s))\delta(y(s) - x_k(s)) \prod_{i \neq k_s}^{N} q_{s,i}(x_i(s)),$$

for proposal distributions $q_{s,i}(x_i(s)) = q_s(x_i(s))$.

The distribution of the marginal sequence $x^{(t)}$ generated by this algorithm converges to the full target π as the number of iterations increases under suitable regularity conditions that are given in Section 7.

Next algorithm follows from Algorithm 6. Corollary 4 in Section 7.3 gives formal convergence result for Algorithm 7.

Algorithm 7 (MIIS for Gibbs Sampler with Simple Importance Sampling)**.** *Given* $y^{(0)}$ *and* $k_{1:d}^{(0)}$*,*

1. *for* $s = 1, \ldots, d$

   (a) *Generate* $X_i(s)|y^{(t)}(\backslash s) \sim q_s(x_i(s))$*, for* $i = \{1\!:\!N\}\backslash k_s$*, and set* $x_{k_s}^{(t)}(s) = y^{(t-1)}(s)$*.*

   (b) *Draw* $K_s^{(t)} = k|(x_{1:N}^{(t)}(s), y^{(t)}(\backslash s))$ *with probability proportional to the weight* $(m_s(x_k^{(t)}(s)|y^{(t)}(\backslash s))/q_s(x_k^{(t)}(s))$*.*

   (c) *Update* $y^{(t)}(s) = x_{k_s^{(t)}}^{(t)}(s)$*.*

2. $t = t{+}1$

5 Estimation of expectations using variance reduction methods

Variance reduction techniques play a central role in Monte Carlo integration. We can directly embed variance reduction methods such as antithetic sampling into the conditional importance sampling approximation. This section takes a step further and considers variance reduction methods based on the output of the MIIS algorithm. Suppose that the algorithm targeting π runs for M iterations. The simplest estimator of $E_\pi(f)$, which uses only the output $\{x^{(t)}\}$ from the Markov Chain, is

$$\widehat{E}_{MC}^{M}(f) := \frac{1}{M}\sum_{t=1}^{M} f\left(x^{(t)}\right) = E_{\widehat{\pi}_{MC}^{M}}(f) \tag{20}$$

where $\widehat{\pi}_{MC}^{M} = \{(1/N, x^{(1)}), \ldots, (1/N, x^{(M)})\}$.

We can improve efficiency by *reusing all the particles*, constructing *Rao-Blackwellized* estimators, and using *control variates*. Section 7.4 shows that all the estimators in this section are consistent under ergodicity. We assume throughout this section that the chain has reached the stationary distribution before running M iterations of the algorithm. In this case the estimators are also unbiased. In the practical situation where the initialization is arbitrary, the estimators are asymptotically unbiased in M for a fixed N.

5.1 Reusing all the particles

The MIIS algorithm constructs an unbiased approximation

$$\widehat{E}^N_{CIS,t}(f) := \sum_{i=1}^{N} W_i \left(x^{(t)}_{1:N}; \xi^{(t)}\right) f\left(x^{(t)}_i\right) \tag{21}$$

to $E_\pi(f)$ at each iteration t of the Markov chain, after the chain has converged. The MIIS estimator that averages over the terms $\widehat{E}^N_{CIS,t}(f)$ is

$$\widehat{E}^{M,N}_{MIIS}(f) := \frac{1}{M}\sum_{t=1}^{M} \widehat{E}^N_{CIS,t}(f) = \widehat{E}^M_{MC}\Big(\widehat{E}^N_{CIS,t}(f)\Big) \tag{22}$$

5.2 Rao-Blackwellization

The motivation for Rao-Blackwellized estimators is that the variance of $f(x(s))$ is larger than the variance of $E_{\pi_s(\cdot|x(\backslash s)}(f)$. However, the latter requires knowledge of the conditional expectation in closed form. The MIIS for the Gibbs sampler overcomes this limitation by using an unbiased approximation of the unknown conditional expectation. It follows from Theorem 3 that, at each iteration t of the Markov chain, the term $\widehat{E}^N_{s,CIS}(f)$ is an unbiased estimator of $E_{\pi_s(\cdot|x(\backslash s)}(f)$. For each $s = 1, \ldots, d$, define

$$\widehat{E}^{M,N}_{s,RB}(f) = \frac{1}{M}\sum_{t=1}^{M} \widehat{E}^N_{s,CIS,t}(f) \tag{23}$$

where

$$\widehat{E}^N_{s,CIS,t}(f) = \sum_{i=1}^{N} W_{s,i} \left(x^{(t)}_{1:N}(s); \xi^{(t)}_s | x^{(t)}(\backslash s)\right) f\left(x^{(t)}_i\right) \tag{24}$$

and $x^{(t)}_i = \{x_i(s), x^{(t)}(\backslash s)\}$ and $x^{(t)}(\backslash s) = \{x^{(t)}(1\!:\!s\!-\!1), x^{(t-1)}(s\!+\!1\!:\!d)\}$.

We define the Rao-Blackwellized MIIS estimator for the Gibbs sampler as the average of the marginal Rao-Blackwellized estimators in (23),

$$\widehat{E}^{M,N}_{MIIS}(f) = \frac{1}{d}\sum_{s=1}^{d} \widehat{E}^{M,N}_{s,RB}(f). \tag{25}$$

Both the marginal Rao-Blackwellized MIIS estimators $\widehat{E}^{M,N}_{s,RB}(f)$ and the Rao-Blackwellized MIIS estimator for the Gibbs sampler $\widehat{E}^{M,N}_{MIIS}(f)$ are unbiased estimators of $E_\pi(f)$ and converge to $E_\pi(f)$ with probability one as $M \to \infty$, for any $N \geq 2$.

5.3 Control Variates

It is optimal to further combine the simple Monte Carlo estimator and the MIIS estimator. For $j = 1, \ldots, p$, suppose that $g_j(x)$ is an integrable function with respect to the density π and $U(g_j)$ a real function such that $E_{\widetilde{\pi}^N}\big(U(g_j)\big) = 0$. Let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$ be a $p \times 1$ vector of parameters and let $F = f - \sum_{j=1}^{p} \kappa_j U(g_j)$. For an optimal choice of $\boldsymbol{\kappa}$, we would like the variance of the estimate of the posterior mean of F to be smaller than that of f. The variables $U(g_i)$ are the *control variates*. The Monte Carlo estimator using F in place of f is studied in many settings; Robert and Casella (2004) and Liu (2001), among others, discuss the standard case. Control variates are not commonly used in an MCMC setting because the Markov sampling scheme makes it more difficult to find suitable candidate control variates with mean zero.

Define $\widehat{E}^N_{CIS,t}(g_j)$ similarly to (21) and

$$U_t(g_j) := g_j(x^{(t)}) - \widehat{E}^N_{CIS,t}(g_j) \tag{26}$$

Assuming ergodicity, the samples from the MIIS Markov chain are eventually distributed as $\widetilde{\pi}^N$ and $\widetilde{\pi}^N[U_t(g_i)] = 0$ as required. The estimator with control variates is

$$\begin{aligned}
\widehat{E}^{M,N}_{CV}(f;\kappa) &= \frac{1}{M}\sum_{t=1}^{M}\left\{ f\left(x^{(t)}\right) - \sum_{j=1}^{p}\kappa_j\left[g_j(x^{(t)} - \widehat{E}^N_{CIS,t}(g_j)\right]\right\} \\
&= \frac{1}{M}\sum_{t=1}^{M}\left\{ f\left(x^{(t)}\right) - \sum_{j=1}^{p}\kappa_j U_t(g_j)\right\} \\
&= \widehat{E}^M_{MC}\left[f - \sum_{j=1}^{p}\kappa_j U(g_j)\right] = \widehat{E}^M_{MC}(F).
\end{aligned} \tag{27}$$

An alternative compact notation shows how we combine the previous estimators,

$$\widehat{E}^{M,N}_{CV}(f;\kappa) = \widehat{E}^M_{MC}(f) - \sum_{j=1}^{p}\kappa_j\left[\widehat{E}^M_{MC}(g_j) - \widehat{E}^{M,N}_{MIIS}(g_j)\right]. \tag{28}$$

In a simple case we may have for example $p = 1$ and $g_1(x) = f(x)$, which allows us to take advantage of the typically high correlations between the simple MC and MIIS estimators of $E_\pi(f)$.

The optimal choice of coefficients $\boldsymbol{\kappa}$ (in the sense of minimizing the variance of the estimator) solves the problem of projecting $\widehat{E}^M_{MC}(f)$ on $\sum_{j=1}^{p}\kappa_j\widehat{E}^M_{MC}(U(g_j))$. The solution is $\boldsymbol{\kappa}^* = \Sigma^{-1}_{UU}\Sigma_{Uf}$, where $\Sigma_{UU} = E(\widehat{E}^M_{MC}(U)\times\widehat{E}^M_{MC}(U)')$ and $\Sigma_{Uf} = E(\widehat{E}^M_{MC}(U)\times\widehat{E}^M_{MC}(f))$, where the expectations are with respect to all the random variables generated by a MIIS

Markov Chain with M iterations. In our applications we estimate the covariances by using the overlapping batch means method as in Flegal and Jones (2011).

We can also use control variates in a Gibbs sampler setting. Our estimator generalizes the control variates approach used by Dellaportas and Kontoyiannis (2012), which *only* applies to exact Gibbs samplers. For a function f and functions $g_{s,j}$ that are integrable with respect to π,

$$\widehat{E}^{M,N}_{s,CV}(f;\kappa) := \widehat{E}^{M}_{MC}(f) - \sum_{s=1}^{d}\sum_{j=1}^{p_s} \kappa_{s,j}\left[\widehat{E}^{M}_{MC}(g_{s,j}) - \widehat{E}^{M,N}_{s,RB}(g_{s,i})\right]. \tag{29}$$

We estimate the optimal parameter $\boldsymbol{\kappa} = \{\kappa_{1,1},\ldots,\kappa_{1,p_1},\kappa_{2,1},\ldots,\kappa_{2,p_2},\ldots,\kappa_{d,1},\ldots,\kappa_{d,p_d}\}$ as above.

6 Illustrations

6.1 Gibbs sampler with importance sampling

6.1.1 Sampling from a bivariate normal distribution

In this example we sample from a simple bivariate normal distribution to compare the performance of the MIIS sampler with control variates to the Metropolis-within-Gibbs (MwG) sampler in a setting in which the exact Gibbs sampler is available as a reference. Dellaportas and Kontoyiannis (2012) adopt this example to illustrate their use of control variates for the Gibbs sampler. The purpose of this example is to show, in a simple setting, that the MIIS sampler with control variates performs well relative to the MwG and Gibbs samplers. We also present results for the Gibbs sampler with control variates as in (Dellaportas and Kontoyiannis, 2012), which we regard as the 'gold standard'for this problem. Beyond this example, we make the important point that the MIIS and MwG samplers *do not* require being able to sample from exact conditional distributions, whereas it is necessary to sample from the exact conditional distributions for the Gibbs sampler. All the methods are very simple to implement for this example. The target distribution is

$$\pi(x) \propto \exp\left(-\frac{1}{2}x'\Sigma^{-1}x\right), \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where $\rho \in \{0.25, 0.5, 0.99\}$ represent low, moderate and high correlation.

We are interested in MCMC estimators of the mean, variance, covariance, a tail probability of the marginal distribution of $x(1)$, i.e., $E_\pi(X(1))$, $E_\pi(X(1)^2) - E_\pi(X(1))^2$, $E_\pi(X(1)X(2)) - E_\pi(X(1))E_\pi(X(2))$, $E_\pi(I[X(1) < -2.32]) = \Pr(X(1) < -2.32)$,

We implement the MIIS algorithm of Section 4.3 (Algorithm 6). We separately consider the standard case and the use of antithetic variables as in Section 3.2 (Algorithm 4). The importance distribution $q_{s,i}(x_{s,i})$ for the MIIS method is a Student t with 5 degrees of freedom, shifted and rescaled to have the same mean and variance as the target conditional distribution $\pi_s(x(s)|x(\backslash s))$. We use the same proposal for the MwG sampler. The number of particles in the IS approximation is $N = 50$. To make the Gibbs and MwG algorithms comparable to MIIS, in these methods we sample 50 iterates of $X(1)$ ($X(2)$) conditional on the current state of $X(2)$ ($X(1)$) in the chain.

We use control variates of MIIS as in Section 5.3. The estimator is given by (28), where we consider at least two control variates for each moment estimate

$$U_1 = \pi^M_{MC}(f(x(1)))-\pi^{M,N}_{MIIS}(f(x(1))) = M^{-1}\sum_{t=1}^{M} f(x^{(t)}(1))-M^{-1}\sum_{t=1}^{M}\sum_{i=1}^{N} W_i(x^{(t)}_{1:N})f(x^{(t)}_i(1))$$

with $U_2 = \pi^M_{MC}(f(x(2)))-\pi^{M,N}_{MIIS}(f(x(2)))$ expressed similarly. The control variates are the differences between the standard MCMC estimates and the corresponding Rao Blackwellized MIIS estimates. We consider additional control variates for estimating the tail probability and $E_\pi(X(1)X(2))$. For the tail probability, we include the same control variates used for mean estimation. For estimating $E_\pi(X(1)X(2))$, we incorporate the control variates used for estimating the mean and variance. We apply the overlapping batch means method in Flegal and Jones (2011) to estimate the covariance matrix of the standard estimator (20) and the control variates based on the output of each chain. That allows us to estimate the optimal coefficients for the control variates as described in Section 5.3.

Table 1 summarizes the results. We report the estimated mean square error (MSE) relative to the MwG sampler based on 500 independent Markov Chains with 10,000 iterations (after a burn-in period of 1,000 iterations) . The results reveal that when the correlation in the target bivariate normal distribution is pronounced ($\rho = 0.99$), the MIIS method with control variates improves the MSEs for estimating the mean, variance, and covariance by 98-99% compared to the MwG sampler. The control variates efficiently explore the information in the chain and the high correlation between the two variables to reduce variance. The results for the covariance estimators show that the MIIS approach can work well when estimating expectations which involve variables in different blocks of the sampler. Introducing antithetic variables in the conditional importance sampler leads to a 99.8% reduction in MSE compared to MwG. Despite the high correlation in the target distribution, the MIIS estimator with antithetic variables takes advantage of the fact that the mean of the proposal is the exact conditional mean. As ρ becomes lower, the MIIS-CV method displays a lower but still large reduction in MSE in comparison

to MwG. The table also shows that as in Dellaportas and Kontoyiannis (2012), the use of control variates in the Gibbs sampler is highly efficient. The disadvantage with the Gibbs-CV method is the requirement that the Gibbs sampler is feasible in the first place, whereas the MIIS-CV applies generally. This simulation exercise illustrates that in many situations, accurate estimation of conditional expectations using MIIS will translate into accurate estimation of expectations under the target distribution with the use of control variates.

Table 1: Bivariate Gaussian simulation – Monte Carlo MSE of target density expectation estimates relative to MwG.

			$\rho = 0.99$		
	Gibbs	Gibbs-CV	MwG	MIIS-CV	MIIS/A-CV
Mean	1.087	0.002	1.000	0.011	0.002
Variance	0.805	0.001	1.000	0.011	0.001
Covariance	0.789	0.001	1.000	0.022	0.002
$P(X(1) < -2.32)$	0.942	0.746	1.000	0.966	0.874
			$\rho = 0.5$		
	Gibbs	Gibbs-CV	MwG	MIIS-CV	MIIS/A-CV
Mean	0.931	0.000	1.000	0.025	0.000
Variance	0.974	0.000	1.000	0.177	0.225
Covariance	0.988	0.000	1.000	0.066	0.022
$P(X(1) < -2.32)$	0.906	0.148	1.000	0.270	0.240
			$\rho = 0.25$		
	Gibbs	Gibbs-CV	MwG	MIIS-CV	MIIS/A-CV
Mean	0.944	0.000	1.000	0.073	0.000
Variance	0.830	0.000	1.000	0.493	0.850
Covariance	0.973	0.000	1.000	0.167	0.025
$P(X(1) < -2.32)$	0.810	0.020	1.000	0.179	0.179

6.1.2 Mixed Logit Model

We consider posterior simulation for the Mixed Logit (MIXL) model as a substantive applied example where it is necessary to apply a method such as importance sampling within Gibbs or Metropolis-within-Gibbs. The binary Mixed Logit model specifies the probability that an individual chooses a certain alternative $j = 1$ (over $j = 0$) at occasion t as

$$p(i \text{ chooses } j = 1 \text{ at } t | Z_{it}, \beta_i) = \frac{\exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}{1 + \exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}, \tag{30}$$

where $\delta_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{Li})'$ is the vector of utility weights for individual i and $Z_{it} = (z_{1it}, \ldots, z_{Lit})'$ is the corresponding vector of attributes for the choice. The individual specific constants are $\beta_{0i} = \beta_0 + \eta_{0i}$ with $\eta_{0i} \sim \mathbf{N}(0, \sigma_0^2)$ and the attribute weights for each individual are latent variables with specification

$$\beta_{li} = \beta_l + \eta_{li}, \qquad l = 1, \ldots, L, \tag{31}$$

with $\eta_{li} \sim \mathbf{N}(0, \sigma_l^2)$.

The parameter vector is $\theta = (\beta_0, \sigma_0^2, \beta_1, \ldots, \beta_L, \sigma_1^2, \ldots, \sigma_L^2)'$, while the vector of latent variables for each individual is $\zeta_i = (\beta_{0i}, \ldots, \beta_{Li})$. The Mixed Logit model captures heterogeneity in preferences by allowing individuals to weight the choice attributes differently. By introducing taste heterogeneity, the MIXL specification avoids the restrictive independence of irrelevant alternatives (IIA) property of the standard multinomial logit model (Fiebig et al., 2010).

We consider an empirical application to the Pap smear data set used for simulated maximum likelihood estimation in Fiebig et al. (2010). In this data set, $I = 79$ women choose whether or not to have a Pap smear test on $T = 32$ choice scenarios. We let the observed choice for individual i at occasion t be $y_{it} = 1$ if the woman chooses to take the test and $y_{it} = 0$ otherwise. Table 2 lists the choice attributes and the associated coefficients. We impose the restriction that $\sigma_5^2 = 0$ in our illustrations since we have found no evidence of heterogeneity for this attribute. To simplify the computational algorithm for this example given this restriction, we fix β_5 at the maximum likelihood estimate.

Table 2: Choice attributes for the pap smear data set

Choice attributes	Values	Associated parameters
Alternative specific constant for test	1	β_0, σ_0
Whether patient knows doctor	0 (no), 1 (yes)	β_1, σ_1
Whether doctor is male	0 (no), 1 (yes)	β_2, σ_2
Whether test is due	0 (no), 1 (yes)	β_3, σ_3
Whether doctor recommends test	0 (no), 1 (yes)	β_4, σ_4
Test cost	{0, 10, 20, 30}/10	β_5

We specify the priors as $\beta_0 \sim \mathbf{N}(0, 100)$, $\sigma_0 \propto (1+\sigma_0^2)^{-1}$, $\beta_l \sim \mathbf{N}(0, 100)$, $\sigma_l \propto (1+\sigma_l^2)^{-1}$, for $l = 1, \ldots, L$. We follow Gelman (2006) and impose half-Cauchy priors on the standard deviation parameters.

In the general notation of the paper, we want to simulate the posterior distribution of $x = \{\boldsymbol{\theta}', \zeta_1', \ldots, \zeta_I'\}'$.

6.1.3 Results

We focus on the estimation of the posterior mean of the model parameters, that is

$$E_\pi(\beta_0),\ E_\pi(\sigma_0),\ E_\pi(\beta_1),\ \ldots,\ E_\pi(\beta_4),\ E_\pi(\sigma_1),\ \ldots,\ E_\pi(\sigma_4).$$

We implement MIIS and Metropolis-within-Gibbs algorithms that iteratively sample the parameters ($x(1) = \theta$) and the choice attributes for all individuals ($x(2) = \{\zeta_1', \ldots, \zeta_I'\}'$) conditional on each other. Equation (31) implies that conditional on β_{li} for all i and $l = 0, 1, \ldots, 4$, the posterior of θ factorises into five components with Gaussian conditional likelihoods from which we can independently sample the corresponding mean and standard deviation parameters. As before, the number of importance samples for the MIIS method is $N = 50$. We generate 50 iterates of $x(s)$ conditional of the previous value of $x(\backslash s)$ in the MwG algorithm to make the two approaches comparable. The proposal for the individual choice attributes combines the efficient importance sampling (EIS) method of Richard and Zhang (2007) with the defensive sampling approach of Hesterberg (1995). The importance density is the two component defensive mixture

$$q(\zeta_i|y_{i1}, \ldots, y_{iT}) = \omega q^{\text{EIS}}(\zeta_i|y_{i1}, \ldots, y_{iT}) + (1-\omega)p(\zeta_i),$$

where $q^{\text{EIS}}(x_i|y_{i1}, \ldots, y_{iT})$ is a multivariate Gaussian importance density obtained using the EIS method. Following Hesterberg (1995), the inclusion of the state prior $p(\zeta_i)$ in the mixture ensures that the importance weights are bounded. We set the mixture weight as $\omega = 0.5$. We also use the EIS method to obtain the importance parameters for the five bivariate parameter proposals (the conditional maximum likelihood estimates are easy to implement alternatives which we use to initialise the EIS method) and incorporate antithetic variables throughout.

We consider the same set of twenty control variates for each MIIS estimate. The first set of control variates are based on the parameters $\boldsymbol{\theta}$,

$$\widehat{E}^M_{MC}(\theta_j) - \widehat{E}^{M,N}_{MIIS}(\theta_j), \qquad \text{for } j = 1, \ldots 10,$$

These control variables are the differences between the standard MCMC posterior mean estimates and the MIIS Rao-Blackwellised estimates. We additionally use two types of control variates based on the individual choice attributes. The first group of control variates based on the attributes is

$$I^{-1} \sum_{i=1}^{I} \widehat{E}^M_{MC}(\beta_{ki}) - \widehat{E}^{M,N}_{MIIS}(\beta_{ki}), \qquad k = 0, \ldots, 4$$

and the second is

$$I^{-1}\sum_{i=1}^{I}\widehat{E}^{M}_{MC}(\beta^2_{ki})-\widehat{E}^{M,N}_{MIIS}(\beta^2_{ki}), \qquad k=0,\ldots,4.$$

The motivation for this second set of control variates is that the parameters of the model are the means and variances of the individual choice attributes, see equation (31). Since there are I individuals, we construct the control variates by averaging the posterior moment estimates of β_{ki}. Because of the correlation between the parameters ($x(1)$) and the choice attributes ($x(2)$) in the Markov chain, we expect these control variates to be highly correlated with the posterior mean estimates of the parameters. Moreover, the use of all twenty control variates simultaneously allows us to leverage the high posterior correlations for variance reduction. We estimate the optimal control variate coefficients as in the last section.

Table 3 reports the estimated MSE for each method relative to MwG. The results are based on 500 independent Markov Chains with 20,000 iterations after 1,000 burn-in draws. The MIIS column in the table corresponds to the Rao-Blackwellized estimate $\widehat{E}^{M,N}_{MIIS}(\theta_j)$ given by (25). We initialize every chain at the maximum likelihood estimate and approximate the "true" posterior means by averaging all the 500 MwG and MIIS estimates (without control variables). The results show that the benefits of using the MIIS Rao-Blackwellized estimates by themselves may be small or negligible because the autocorrelation in the MIIS chain is the main determinant of the total variance of the estimates in this example. When we use the Rao-Blackwellized estimates to construct the control variates, we obtain 75-95% reductions in MSE relative to the MwG algorithm. The two methods have similar computational cost and implementation effort.

Table 3: Mixed Logit Application – Monte Carlo MSE of posterior mean estimates relative to MwG.

Parameter	MwG	MIIS	MIIS-CV
β_0	1.00	0.91	0.07
β_1	1.00	1.23	0.06
β_2	1.00	0.92	0.05
β_3	1.00	0.98	0.06
β_4	1.00	0.66	0.08
σ_0	1.00	0.95	0.07
σ_1	1.00	1.02	0.16
σ_2	1.00	0.94	0.08
σ_3	1.00	1.17	0.08
σ_4	1.00	0.54	0.25

6.2 Random Walk Importance Sampler

6.2.1 Markov Modulated Poisson Process

A Markov Modulated Poisson Process (MMPP) Y_t is a Poisson process whose intensity λ_t takes on a discrete number d of values $\psi = (\psi_1, \ldots, \psi_d)'$, with the intensity at any time point determined by the state of an unobserved continuous-time Markov chain with generator Q. We identify the model by imposing the parameter restriction $\psi_d > \ldots > \psi_1$. Sherlock et al. (2010) recently considered the MMPP as a challenging case study for comparing a range of Random Walk Metropolis (RWM) algorithms proposed in the literature. We replicate their setting to illustrate how the random walk importance sampler of Section 3.3 can lead to more efficient and robust MCMC simulation compared to standard RW samplers.

Suppose that we observe a realisation of the process over a certain time window and record n event times. Fearnhead and Sherlock (2006) derived the likelihood for the model as

$$L(Q, \psi, t) = \nu' \exp^{(Q-\Psi)t_1} \Psi \ldots \exp^{(Q-\Psi)t_n} \Psi \exp^{(Q-\Psi)t_{n+1}} \iota, \tag{32}$$

where

$$Q = \begin{pmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{pmatrix},$$

ν is the initial distribution of the latent state Z_t (which we take to be the stationary distribution of the chain implied by Q), i.e., $\Pr(Z_t = j) = \nu(j)$, $\Psi = \text{diag}(\psi)$, ι is a vector of ones, t_1 is the time from the start of the observation window until the first event, t_i is the time between events $i-1$ and i, and t_{n+1} is time between event n and the end of the observation window.

6.2.2 Simulation Study

We replicate the simulation study in Sherlock et al. (2010). We simulate the MMPP model with $d = 2$ over an observation window of 100 seconds. The generator matrix Q has parameters $q_{12} = q_{21} = 1$. The intensity vector is $\psi = (10, 17)'$. As in the Sherlock et al. (2010) application, we complete the model by specifying exponential priors for all the parameters. The means of the priors are the true parameters.

We consider three different methods: the standard RWM algorithm, the multiple-try RWM (MTM) of Liu et al. (2000), and the MIIS random walk method (Algorithm 7). We consider a random walk on the transformed parameter vector $\widetilde{\theta} = (\log(\psi_1), \log(\psi_2 - \psi_1), \log(q_{12}), \log(q_{21})$, which is more efficient than working on the original scale. Let i

index the current iteration of the Markov Chain. The proposal for all methods is

$$q_{i+1}(\theta^*_{i+1}|\theta^*_i) = N\left(\widetilde{\theta}_i, \frac{2.38^2}{4}\widetilde{\Sigma}\right),$$

where $\widetilde{\Sigma}$ is an estimate of the posterior covariance matrix of $\widetilde{\theta}$ based on a trial run of the RWM algorithm. The scale of the proposal aims to achieve an acceptance rate of 0.234 in the standard RWM algorithm, which is optimal rate under certain assumptions; see the discussion in Sherlock et al. (2010). We consider four control variates associated with each parameter for estimating each posterior mean

$$\widehat{E}^M_{MC}(\theta_j) - \widehat{E}^{M,N}_{MIIS}(\theta_j), \qquad \text{for } j = 1, \dots 4.$$

As before, the control variates are the differences between the standard MCMC estimates and the Rao-Blackwellized estimates that reuse all the particles.

We parallelize the likelihood evaluations over eight cores at every iteration of the MIIS Markov chain and set the number of particles to $N = 8$ and $N = 16$. Our discussion treats the MIIS method with $N = 8$ draws as being comparable to the standard RWM method, which is difficult to parallelize. This implies that the MIIS algorithm performs eight times as many likelihood evaluations as the standard RWM algorithm in total, but in the same amount of time under perfect parallelization. We report the actual computing times in the tables. We configure the MTM method such that it performs the same number of likelihood evaluations per iteration as the MIIS algorithm with $N = 8$ particles. However, the parallelization for the MTM method is less efficient as every iteration of the method requires two separate stages.

The simulation study averages results over ten independent realisations of the DGP. For every realisation, we simulated 500 independent Markov chains for each method and ran each Markov Chain for 10,000 iterations after discarding a burn-in of 1,000 iterations. We consider two cases for initialisation. We initialize the algorithm at the maximum likelihood estimate for half the chains. For the other half, we initialize the chain by drawing from the prior. We use the same draw from the prior to initialize all the methods at each replication. Initializing from the prior allows us to compare the convergence performance of each method. We then compute the posterior mean and variance estimates based on each chain. We combine all chains initialized at the true parameters to obtain precise approximations to the true posterior means and variances.

Table 4 reports the MSE efficiency of the posterior estimates relative to the performance of the RWM method. We average the results over the four parameters for conciseness. We also present the actual computing times, and average acceptance rate

and average integrated autocorrelation time (IACT). We base the last two results on longer independent chains with 100,000 iterations (one per simulated dataset). In the case of MIIS, the "acceptance" rate is the proportion of iterations in which the sampled particle is not the previous iterate. We also report the relative time adjusted MSE for each method, which we define as $(\text{time}_{alg} \times \text{mse}_{alg})/(\text{time}_{MH} \times \text{mse}_{MH})$. This estimate approximates the MSE relative to the MH algorithm for the same amount of computing time. We average all these results over realisations.

The results show that the MIIS method reduces the time adjusted MSEs by 70% compared to MH when we initialize the chain at the true parameters. This gain in performance comes both from reductions in IACT and the use of Rao-Blackwellization to estimate the posterior moments. The MIIS method also outperforms the MTM method, at a lower computational cost. Using control variates further reduces the MSE, leading to a 78-83 time adjusted improvement over MH. To put these gains in the context of the random walk literature, the best performing algorithm in the simulation of Sherlock et al. (2010) for the same DGP, a MH random walk in the log scale with with an adaptively tuned mixture proposal, leads to a 84% reduction in variance (measured by IACT) compared the least efficient algorithm in their analysis, a MH random walk with tuned proposal $N(0, \lambda^2 I)$. The table also shows that when we initialize all algorithms from the prior, the MIIS algorithm with $N = 8$ particles and control variates generates 80-99% reductions in time adjusted MSE compared to the standard RWM algorithm. This result suggests that the MIIS algorithm is more robust to the initial conditions than the standard RWM and MTM algorithms.

6.2.3 Empirical Example

We now apply the RWM, MTM and MIIS methods using data from the empirical example in Fearnhead and Sherlock (2006). The data consists of positions (in bases) of Chi sites (a DNA motif) in the genome of Escherichia coli bacteria. The specification of the MMPP model is the same as in the simulation study above. We follow the procedure described in Fearnhead and Sherlock (2006) to obtain data-based parameters for the exponential priors. We estimate the MMPP for the lagging part of the outer ring of the E. coli genome strand, which has 117 observations in total. We ran each Markov Chain for 50,000 iterations after discarding a burn-in of 1,000 iterations and initialized all chains at the maximum likelihood estimate. We also use the Hessian of the likelihood at the maximum likelihood estimate to obtain the shape of the random walk proposal.

Table 5 displays the Monte Carlo MSEs over 500 replications of each algorithm. The results show that the MIIS-CV algorithm with $N = 16$ has $83-90\%$ lower MSEs than the standard RWM algorithm. Adjusting for the actual computational times, the improve-

Table 4: MMPP DGP with $\psi_1 = 10$ and $\psi_2 = 17$. Monte Carlo MSE of posterior estimates relative to MH (average across parameters).

We define the time adjusted MSE as $(\text{time}_{alg} \times \text{mse}_{alg})/(\text{time}_{MH} \times \text{mse}_{MH})$, which approximates the MSE relative to the MH algorithm for the same amount of computing time.

			MIIS		MIIS-CV	
	MH	MTM	N=8	N=16	N=8	N=16
Acceptance	0.30	0.46	0.32	0.47	0.32	0.47
IACT	15.0	9.3	7.5	5.6	7.5	5.6
Time	8	15	9	14	9	14
	Initializing at the true parameters					
Mean	1.00	0.47	0.24	0.16	0.15	0.12
Variance	1.00	0.67	0.25	0.14	0.20	0.15
	Time adjusted MSE					
Mean	1.00	0.88	0.27	0.29	0.17	0.21
Variance	1.00	1.26	0.29	0.25	0.22	0.27
	Initializing from the prior					
Mean	1.00	0.07	0.06	0.02	0.01	0.01
Variance	1.00	0.24	0.19	0.06	0.05	0.04
	Time adjusted MSE					
Mean	1.00	0.13	0.07	0.03	0.01	0.01
Variance	1.00	0.46	0.21	0.10	0.06	0.06

ments are between 44−66%. The MIIS algorithm also outperforms the MTM method. We note from the table that the practical computational cost of adding particles tends to be low, so that we can consider a higher N to increase robustness.

Table 5: Empirical example for the MMPP model – Monte Carlo MSE of posterior estimates relative to MH.

We define the time adjusted MSE as $(\text{time}_{alg} \times \text{mse}_{alg})/(\text{time}_{MH} \times \text{mse}_{MH})$, which approximates the MSE relative to the MH algorithm for the same amount of computing time.

			MIIS		MIIS-CV	
	MH	MTM	N=8	N=16	N=8	N=16
Acceptance	0.24	0.54	0.42	0.57	0.42	0.57
IACT	55.2	20.2	13.3	9.4	13.3	9.4
Time	20	62	55	59	58	63
			Posterior mean			
ψ_1	1.00	0.45	0.25	0.17	0.18	0.13
ψ_2	1.00	0.50	0.21	0.15	0.15	0.10
q_{12}	1.00	0.39	0.20	0.14	0.16	0.10
q_{21}	1.00	0.52	0.27	0.15	0.15	0.12
			Time adjusted MSE			
ψ_1	1.00	1.39	0.68	0.50	0.51	0.43
ψ_2	1.00	1.55	0.59	0.45	0.45	0.33
q_{12}	1.00	1.21	0.54	0.42	0.46	0.33
q_{21}	1.00	1.60	0.73	0.45	0.44	0.37
			Posterior variance			
ψ_1	1.00	0.53	0.34	0.19	0.22	0.17
ψ_2	1.00	0.65	0.29	0.21	0.18	0.11
q_{12}	1.00	0.45	0.20	0.12	0.17	0.12
q_{21}	1.00	0.50	0.23	0.14	0.15	0.10
			Time adjusted MSE			
ψ_1	1.00	1.66	0.93	0.56	0.64	0.56
ψ_2	1.00	2.02	0.81	0.63	0.51	0.36
q_{12}	1.00	1.39	0.55	0.35	0.50	0.38
q_{21}	1.00	1.55	0.63	0.41	0.45	0.33

7 Theory

This section presents our theoretical results for the MIIS estimators and restates some of the definitions in previous sections in a more general setting.

Let (A, Ω) denote a measurable space and π some given *target* probability distribution on (A, Ω). Assume that a reference measure μ dominates π ($\pi \ll \mu$) and that $\pi(\mathrm{d}x) = \pi(x)\mu(\mathrm{d}x)$. With a small abuse of notation, we write $\pi(x)$ for the density of the probability

measure π with respect to μ. In most situations $A \subseteq \Re^d$ ($d \in \mathbb{N}$, the set of positive integers), $\Omega = \mathcal{B}(A)$ is the Borel σ-algebra of the set A, and the majorizing measure μ is either the counting measure, the Lebesgue measure, or a combination of both. We assume that the distributions $\eta(\mathrm{d}\xi|y)$ and $q_k(\mathrm{d}x|\xi)$ ($k = 1, \ldots, N$) admit densities $\eta(\xi|y)$ and $q_k(x|\xi)$ with respect to the same measure μ. We work interchangeably with other distributions and their corresponding densities. Let $A = \times_{i=1}^{d} A_i$ and $\Omega = \Omega_1 \otimes \cdots \otimes \Omega_d$. In the conditional case, for all $s = 1, \ldots, d$, let (A_s, Ω_s) denote a measurable spaces. We assume that $\pi_s(\mathrm{d}y(s)|y(\backslash s))$, $q_{s,k}(\mathrm{d}y(s)|\xi_s, y(\backslash s)$, and $\eta_s(\mathrm{d}\xi_s|y(s), y(\backslash s))$ are defined on (A_s, Ω_s) and have densities with respect to some majorizing measure μ_s, that may depend on $y(\backslash s) \in A_{\backslash s}$ for $A_{\backslash s} = \times_{i \neq s}^{d} A_i$.

7.1 Convergence of the marginal MIIS chain

If (y, k) is marginally distributed as $N^{-1}\pi$, then the CIS estimator is unbiased by Theorem, 1(iii). Theorem 5 below shows that the MIIS Algorithm (Algorithm 2 in Section 2.3) samples from the target density $N^{-1}\pi$ asymptotically, i.e., as the number of iterations $t \to \infty$. In other words, the marginal distribution of $(y^{(t)}, k^{(t)})$ is $N^{-1}\pi$, asymptotically.

For all $l, k \in \{1{:}N\}$ and $y, z \in A$, define

$$\begin{aligned} S_{l,k}(\xi, y, z) &:= \int_{A^2} T(y, \mathrm{d}x_l; \xi) T(z, \mathrm{d}x_k; \xi) \frac{\mathbf{q}_{l,k}(x_l, x_k|\xi)}{q_l(x_l|\xi) q_k(x_k|\xi)}, \quad l \neq k \\ S_{l,l}(\xi, y, z) &:= \int_{A} \frac{T(y, \mathrm{d}x_l; \xi) T(z, x_l; \xi)}{q_l(x_l|\xi)}, \quad l = k \end{aligned} \tag{33}$$

where $\mathbf{q}_{l,k}(x_l, x_k|\xi)$ is the joint marginal of (x_l, x_k) for $l \neq k$.

The proof of Theorem 5 is based on the following assumption discussed in Section 7.2.

Assumption 1. *(i) There exists a constant C, $0 < C < \infty$, such that the marginal densities $q_k(x|\xi)$ satisfy $\pi(x)\eta(\xi|x) \leq C\, q_k(x|\xi)$, for each k and all $x, \xi \in A$.*

(ii) (a) For each $k, l \in \{1{:}N\}$ and $y, z \in A$, there exist functions $h_{k,l}(y, z)$ such that

$$\int_A S_{l,k}(\xi, y, z)\, \eta(\xi|y)\eta(\xi|z)\mu(\mathrm{d}\xi) \geq h_{l,k}(y, z).$$

(b) For each $l \in \{1{:}N\}$ there exists a set $\mathcal{J}_l \subseteq \{1{:}N\}\backslash\{l\}$ such that: $\mathcal{J}_l \cap \mathcal{J}_k \neq \emptyset$ for $l \neq k$; and

(c) for all $j \in \mathcal{J}_l$ and $y, z \in A$ $h_{l,j}(y, z) > 0$ and $h_{j,l}(y, z) > 0$.

Assumption 1 (i) requires the weights to be uniformly bounded and it is often used in the particle literature. This condition is not restrictive and can be enforced by choosing

suitable q_k and η. Part (ii) is a technical condition that imposes regularity conditions on the pairwise dependence of the particles, on the kernel T, and the auxiliary distribution η.

Theorem 5. *(i) If Assumption 1 holds then the marginal chain $\{(y^{(t)}, k^{(t)})\}$, sampled using MIIS, is Markov and ergodic, i.e.,*

$$\lim_{t\to\infty} \mid P^t(l, y; \cdot) - N^{-1}\pi(\cdot)|_{TV} = 0,$$

where $P(y, l; B\times\{k\})$ is the Markov transition kernel from (y, l) to $B\times\{k\}$, $B \in \Omega$, $k \in \{1{:}N\}$.

(ii) If, in addition, for $k \in \mathcal{J}_l$, $h_{l,k}(y, z) \geq \underline{h}_{l,k}(z) > 0$, i.e., $\underline{h}_{l,}$ does not depend on the initial value $y \in A$, then the marginal chain is uniformly ergodic.

The distribution of the marginal chain $\{(y^{(t)}, k^{(t)})\}$ converges to the target distribution $N^{-1}\pi$ as the number of iterations increases. It means that, after a warm up period, the marginal distribution of samples from the chain is $N^{-1}\pi$ and, hence, $\widehat{E}^M_{MC}(f)$ is an unbiased estimator of $E_\pi(f)$ for any integrable f. If $E_\pi(|f|) < \infty$, then by Theorem 3 of Tierney (1994), ergodicity implies that $\widehat{E}^M_{MC}(f)$ is also a consistent estimator of $E_\pi(f)$. If $E_\pi(f^2) < \infty$ and uniform ergodicity holds then by Theorem 5 of Tierney (1994) we also obtain a central limit theorem for $\widehat{E}^M_{MC}(f)$.

7.2 Convergence results for the examples in Section 3

This section discusses the application of Theorem 5 to the examples in Section 3. In all three examples T is the identity kernel, i.e., $T(y, dz; \xi) = \delta_y(dz)$. This gives $S_{l,k}(\xi, y, z) = \mathbf{q}_{l,k}(y, z|\xi)/q_l(y|\xi)q_k(z|\xi)$ for $k \neq l$ and $S_{l,l}(\xi, y, z) = I(z = y)/q_l(y|\xi)$. Hence, we require $h_{l,k}(y, z) \geq 0$ functions such that

$$\int_A \eta(\xi|y)\eta(\xi|z)\frac{\mathbf{q}_{l,k}(y, z|\xi)}{q_l(y|\xi)q_k(z|\xi)}\mu(d\xi) \geq h_{l,k}(y, z), \quad \text{for} \quad l \neq k$$
$$I(z = y)\int_A \frac{\eta(\xi|y)\eta(\xi|z)}{q_l(y|\xi)}\mu(\mathrm{d}\xi) \geq h_{l,l}(y, z).$$

Part (i) is assumed explicitly and we choose $\mathcal{J}_l$ to satisfy Assumption 1(ii).

Simple importance sampling example

This example is discussed in Section 3.1.

Corollary 1. *Suppose that there is no dependence of T and the q_i on ξ and (i) $T(x, \mathrm{d}y) = \delta_x(\mathrm{d}y)$, (ii) $q(\mathrm{d}x_{1:N}) = \prod_{i=1}^N q_i(\mathrm{d}x_i)$ and (iii) $\pi(\mathrm{d}x_i) \leq C\, q_i(\mathrm{d}x_i)$, where $C > 0$ is a positive constant. Then, the marginal chain $\{(y^{(t)}, k^{(t)})\}$ is uniformly ergodic for $N \geq 3$.*

Let $\eta(\mathrm{d}\xi|y) = \delta_0(\mathrm{d}\xi)$, without loss of generality. It is easy to see that $h_{l,k}(y, z) = I(l \neq k)$ is a valid choice. Assumption 1 is satisfied by taking $\mathcal{J}_l = \{1{:}N\}\backslash\{l\}$, and $N \geq 3$. Uniform ergodicity follows from Theorem 5 part (ii), because $h_{k,l}(y, \cdot)$ does not depend on y for $l \in \mathcal{J}_k$.

Importance sampling with antithetic variables

This example is discussed in Section 3.2.

Corollary 2. *Suppose there is no dependence on ξ and (i) $T(x, \mathrm{d}y) = \delta_x(\mathrm{d}y) = \delta_x(y)\mathrm{d}y$, (ii) $q(\mathrm{d}x_{1:N}) = \prod_{i=1}^{N/2} q_{i,i+N/2}(\mathrm{d}x_i, \mathrm{d}x_{i+N/2})$ such that $q_{i+N/2}(\mathrm{d}x_{i+N/2}|x_i) = \delta_{Q_i^{-1}(1-Q_i(x_i))}(\mathrm{d}x_{i+N/2})$, where Q_i is the cdf of $q_i(x_i)$. (iii) $\pi(\mathrm{d}x_i) \leq C\, q_i(\mathrm{d}x_i)$, where $C > 0$ is a positive constant Then, the marginal chain $\{(y^{(t)}, k^{(t)})\}$ is uniformly ergodic for $N/2 \geq 3$.*

It is straightforward to check that,

$$\frac{\mathbf{q}_{l,k}(y, z)}{q_l(y)q_k(z)} = \begin{cases} \frac{\delta_{Q^{-1}(1-Q(y))}(z)}{q(z)} & k \in \{1{:}N\} \cap \{l{-}N/2, N/2{+}l\},\ k \neq l \\ 1 & k \in \{1{:}N\}\backslash\{l, l{-}N/2, N/2{+}l\} \end{cases}.$$

Choose $\mathcal{J}_l = \{1{:}N\}\backslash\{l, l{-}N/2, N/2{+}l\}$. It is easy to see that $h_{l,k} = I(k \in \mathcal{J}_l)$ is a valid choice. Assumption 1 is satisfied and the MIIS sampler is uniformly ergodic using the same arguments as in the previous example.

Random walk importance sampler

This example is discussed in Section 3.3.

Corollary 3. *Suppose that (i) $T(x, \mathrm{d}y|\xi) = \delta_x(\mathrm{d}y) = \delta_x(y)\mathrm{d}y$, (ii) $\mathbf{q}(\mathrm{d}x_{1:N}|\xi) = \prod_{i=1}^N q_i(\mathrm{d}x_i|\xi)$ and (iii) $q_i(\mathrm{d}x_i|\xi) = \phi(x_i{-}\xi)\mathrm{d}x_i$; (iv) $\eta(\mathrm{d}\xi|y) = \phi(\xi{-}y)\mathrm{d}\xi$; (v) $\phi(x{-}y) > 0$ for any $x, y \in A$. (vi) $\pi(x_i) \leq C$. Then, the marginal chain $\{(y^{(t)}, k^{(t)})\}$ is ergodic for $N \geq 3$. If $\inf_{z,y\in A} \int \phi(\xi{-}z)\phi(\xi{-}y)\mathrm{d}\xi > \varepsilon > 0$ Then $\{(y^{(t)}, k^{(t)})\}$ is uniformly ergodic.*

For $l \neq k$, $S_{l,k}(\xi, y, z) = 1$ because the proposals are independent, and

$$h(y, z) := h_{k,l}(y, z) = \int_A \phi(\xi{-}y)\phi(z{-}\xi)\mathrm{d}\xi > 0.$$

Choose $\mathcal{J}_l = \{1{:}N\}\backslash\{l\}$. Then Assumption 1 holds and ergodicity follows from Theorem 5. By assumption there exists $\epsilon > 0$ such that $h(y, z) \geq \epsilon$ for all $y, z \in A$. By defining $\underline{\mathrm{h}}_{l,k}(z) = \epsilon$ for $k \in \mathcal{J}_l$, uniform ergodicity follows from part (ii) of Theorem 5.

7.3 The MIIS Gibbs Sampler

This section shows that the marginal chain $\{l_{1:d}^{(t)}, Y^{(t)}\}$ generated by the MIIS Gibbs sampler (Algorithm 6 in Section 4.2) is ergodic if (i) the *ideal Gibbs sampler*, i.e., the Gibbs sampler drawing variables from the conditionals $\pi_s(\mathrm{d}y(s)|y(\backslash s))$, is irreducible and aperiodic; (ii) the CIS Gibbs sampler satisfies regularity conditions that are similar to Assumption 1, but hold for each $s = 1, \ldots, d$; (iii) The space A is Euclidean with Lebesgue measure the underlying measure.

Our notation assumes that we condition on $y(\backslash s)$ when dealing with the sth component and do not usually show this conditioning explicitly. The transition kernel for the ideal Gibbs sampler is

$$P_G(y; \mathrm{d}z) := \prod_{i=1}^{m} \pi_s(\mathrm{d}z(s)|z(1\!:\!s\!-\!1), y(s\!+\!1\!:\!d)) \tag{34}$$

For all $l, k \in \{1{:}N\}$ and $\xi(s), y(s), z(s) \in A_s$, define

$$\begin{aligned} S_{s,l,k}(\xi(s), y(s), z(s)) &:= \int_{A^2} T_s(y(s), \mathrm{d}x_l(s); \xi(s)) T_s(z(s), \mathrm{d}x_k(s); \xi(s)) \\ &\quad \times \frac{\mathbf{q}_{s,l,k}(x_l(s), x_k(s)|\xi(s))}{q_{s,l}(x_l(s)|\xi(s)) q_{s,k}(x_k(s)|\xi(s))}, \quad l \neq k \\ S_{s,l,l}(\xi(s), y(s), z(s)) &:= \int_{A} \frac{T_s(y(s), \mathrm{d}x_l(s); \xi(s)) T_s(z(s), x_l(s); \xi(s))}{q_{s,l}(x_l(s)|\xi(s))}, \quad l = k \end{aligned} \tag{35}$$

where $\mathbf{q}_{s,l,k}(x_l(s), x_k(s)|\xi(s))$ is the joint marginal of $(x_l(s), x_k(s))$ for $l \neq k$.

The proof of Theorem 6 is based on the following assumption, which generalizes Assumption 1 to the Gibbs case.

Assumption 2. *The following condition holds for all $s = 1, \ldots, d$. All terms are conditional on $y(\backslash s)$, unless stated otherwise.*

(i) There exists a constant C, $0 < C < \infty$, such that the marginal densities $q_{s,k}(x(s)|\xi(s))$ satisfy $\pi_s(x(s))\eta(\xi(s)|x(s)) \leq C^{1/d} q_{s,k}(x(s)|\xi(s))$, for each k and all $x(s), \xi(s) \in A_s$.

(ii) For each $k, l \in \{1{:}N\}, y(s), z(s) \in A_s$ and $y(\backslash s) \in A_{\backslash s}$,

(a) There exist functions $h_{s,k,l}(y(s), z(s)) \geq 0$ such that

$$\int_{A_s} S_{s,l,k}(\xi(s), y(s), z(s))\, \eta(\xi(s)|y(s))\eta(\xi(s)|z(s))\mu(\mathrm{d}\xi(s)) \geq h_{s,l,k}(y(s), z(s));$$

(b) for each $l \in \{1{:}N\}$, there exists a set $\mathcal{J}_{s,l} \subseteq \{1{:}N\}\backslash\{l\}$ $\mathcal{J}_{s,l} \cap \mathcal{J}_{s,k} \neq \emptyset$ for $l \neq k$;

(c) for each $j \in \mathcal{J}_{s,l}$, $h_{s,l,j}(z(s), y(s)) > 0$ *and* $h_{s,j,l}(z(s), y(s)) > 0$ *on* $y \in \{x \in A : \pi(x) > 0\}$ *and* $z(s) \in A_s$.

Define $\mathbf{l} := l_{1:d}$ and $\mathbf{k} := k_{1:d}$.

Theorem 6. *Suppose Assumption 2 holds. If* P_G *is irreducible and aperiodic, then so is the marginal kernel* $P_M(y, \mathbf{l}; \mathrm{d}z \times \mathbf{k})$*, and for any starting values* $y \in A$ *with* $\pi(y) > 0$ *and* $\mathbf{l} \in \{1{:}N\}^d$*,*

$$\lim_{t \to \infty} \left| P_M^t(y, \mathbf{l}; \cdot) - N^{-d}\pi(\cdot) \right|_{TV} = 0.$$

Gibbs Sampler with simple importance sampling example

We now consider the example of the Gibbs sampler with simple importance sampling discussed in Section 4.3.

Corollary 4. *Suppose that there is no dependence on* ξ *and the following conditions hold for* $s = 1, \ldots, d$*. (i)* $T_s(x(s), \mathrm{d}y(s)|y(\backslash s)) = \delta_{x(s)}(\mathrm{d}y(s))$*, (ii)* $q_s(\mathrm{d}y_{1:N}(s)|y(\backslash s)) = \prod_{i=1}^N q_{s,i}(\mathrm{d}y_i(s)|y(\backslash s))$*, (iii) There is a* $C > 0$ *such that* $q_{s,i}(\mathrm{d}y_i(s)|y(\backslash s)) \geq C^{1/d}\pi_s(\mathrm{d}y_i(s)|y(\backslash s))$*. If we further assume that the ideal Gibbs sampler,* P_G*, is irreducible and aperiodic, then the distribution of the marginal chain* $\{\mathbf{l}^{(t)}, y^{(t)}, t \geq 1\}$ *converges to the full target* $N^{-d}\pi(\cdot)$ *as* $t \to \infty$ *for any fixed* $N \geq 3$*.*

This corollary follows after the same arguments used in the marginal case. The functions $h_{s,l,k} = I(l \neq k)$ and the sets $\mathcal{J}_{s,l} = \{1{:}N\}\backslash\{l\}$ for each $s = 1, \ldots, d$. The result follows from Theorem 3.

7.4 Consistent estimation of expectations

Using all the particles

The next theorem shows that the MIIS estimator $\widehat{E}_{MIIS}^{M,N}(f)$ discussed in Section 5 converge to $E_\pi(f)$.

Corollary 5. *Let* $f : A \mapsto \mathbb{R}$ *be such that* $E_\pi(|f|) < \infty$ *and suppose Assumption 1 holds. Then the MIIS estimator* $\widehat{E}_{MIIS}^{M,N}(f) \to E_\pi(f)$ *with probability one as* $M \to \infty$*, for any* $N \geq 2$*.*

Using Rao Blackwellized estimators

Define the Rao-Blackwellized estimators $\widehat{E}_{s,RB}^{M,N}(f)$ and $\widehat{E}_{RB}^{M,N}(f)$ as in Section xxx. Then,

Corollary 6. *Let* $f : A \mapsto \mathbb{R}$ *be such that* $E_\pi(|f|) < \infty$*. Suppose Assumption 2 holds. Then, the Rao-Blackwellized estimators* $\widehat{E}_{s,RB}^{M,N}(f)$ *and* $\widehat{E}_{RB}^{M,N}(f)$ *converge to* $E_\pi(f)$ *with probability 1 as* $M \to \infty$*.*

Using Control Variates

The following two results shows that the estimators based on control variates discussed in Section 5.3 are consistent under ergodicity.

Corollary 7. *Let $f : A \mapsto \mathbb{R}$ be such that $E_\pi(|f|) < \infty$. Suppose Assumption 1 holds. Then the estimator using control variates $\widehat{E}_{CV}^{M,N}(f,\theta) \to E_\pi(f)$ with probability one as $M \to \infty$, for any $\boldsymbol{\kappa} \in \mathbb{R}^p$.*

Corollary 8. *For any $s = 1, \ldots, d$, let $f : A \mapsto \mathbb{R}$ be such that $E_\pi(|f|) < \infty$. Suppose Assumption 2 holds. Then the estimator using control variates $\widehat{E}_{s,CV}^{M,N}(f,\theta) \to E_\pi(f)$ with probability one as $M \to \infty$ and any $\boldsymbol{\kappa} \in \mathbb{R}^{p_1+\cdots+p_d}$.*

References

Andrieu, C., A. Doucet, and R. Holenstein (2010): "Particle Markov chain Monte Carlo methods," *J. R. Statist. Soc B*, 72, 269–342.

Andrieu, C., A. Lee, and M. Vihola (2013): "Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of Particle Gibbs samplers," ArXiv:1312.6432 [math.PR].

Carter, C. K., E. F. Mendes, and R. Kohn (2014): "An extended space approach for particle Markov chain Monte Carlo methods," ArXiv:1406.5795v2.

Chopin, N. and S. S. Singh (2013): "On the particle Gibbs sampler," ArXiv preprint arXiv:1304.1887.

Craiu, R. V. and C. Lemieux (2007): "Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling," *Statistics and computing*, 17, 109–120.

Dellaportas, P. and I. Kontoyiannis (2012): "Control variates for estimation based on reversible Markov chain Monte Carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 133–161.

Fearnhead, P. and C. Sherlock (2006): "An exact Gibbs sampler for the Markov-modulated Poisson process," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 767–784.

Fiebig, D. G., M. P. Keane, J. Louviere, and N. Wasi (2010): "The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity," *Marketing Science*, 29, 393–421.

Flegal, J. M. and G. L. Jones (2011): "Implementing MCMC: estimating with confidence," *Handbook of Markov chain Monte Carlo, Boca Raton, Florida: Chapman & Hall/CRC*, 175–197.

Gelman, A. (2006): "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515–534.

Hammer, H. and H. Tjelmeland (2008): "Control Variates for the Metropolis–Hastings Algorithm," *Scandinavian Journal of Statistics*, 35, 400–414.

Hesterberg, T. (1995): "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194.

Hoogerheide, L., A. Opschoor, and H. K. van Dijk (2012): "A class of adaptive importance sampling weighted EM algorithms for efficient and robust posterior and predictive simulation," *Journal of Econometrics*, 171, 101–120.

Jacob, P., C. P. Robert, and M. H. Smith (2011): "Using parallel computation to improve independent Metropolis–Hastings based estimation," *Journal of Computational and Graphical Statistics*, 20, 616–635.

Li, W., Z. Tan, and R. Chen (2013): "Two-Stage Importance Sampling with Mixture Proposals," *Journal of the American Statistical Association*, 108, 1350–1365.

Liesenfeld, R., J.-F. Richard, and J. Vogler (2013): "Analysis of discrete dependent variable models with spatial correlation," Working paper.

Lindsten, F., R. Douc, and E. Moulines (2014a): "Uniform ergodicity of the Particle Gibbs sampler," ArXiv:1401.0683 [math.ST].

Lindsten, F., M. I. Jordan, and T. B. Schön (2014b): "Particle Gibbs with Ancestor Sampling," arxiv:1401.0604.

Lindsten, F. and T. B. Schön (2012): "On the use of backward simulation in particle Markov chain Monte Carlo methods," arxiv:1110.2873.

Liu, J. S. (2001): *Monte Carlo strategies in scientific computing*, Springer.

Liu, J. S., F. Liang, and W. H. Wong (2000): "The Multiple-Try Method and Local Optimization in Metropolis Sampling," *Journal of the American Statistical Association*, 95, 121–134.

Mendes, E. F., C. K. Carter, and R. Kohn (2014): "On general sampling schemes for particle Markov chain Monte Carlo methods," Arxiv:1401.1667.

Richard, J.-F. and W. Zhang (2007): "Efficient High-Dimensional Importance Sampling," *Journal of Econometrics*, 141, 1385–1411.

Robert, C. P. and G. Casella (2004): *Monte Carlo statistical methods*, Springer.

Sherlock, C., P. Fearnhead, and G. Roberts (2010): "The random walk Metropolis : linking theory and practice through a case study." *Statistical Science*, 25, 172–190.

Tierney, L. (1994): "Markov chains for exploring posterior distributions," *The Annals of Statistics*, 22, 1701–1728.

Tran, M.-N., M. Scharth, M. K. Pitt, and R. Kohn (2014): "Importance Sampling Squared for Bayesian Inference in Latent Variable Models," .

A Proofs

The notation in this section is the same as in Section 7.

A.1 Markov Interacting Importance Sampler

Proof of Theorem 1. Part (i): From (7), $\widetilde{\pi}^N$ is a proper distribution function that integrates to 1 and has marginal $\widetilde{\pi}^N(dy) = \pi(dy)$. Part (ii): The joint distribution $\widetilde{\pi}^N(\mathrm{d}x_{1:N}, \mathrm{d}\xi, k)$ is

$$\begin{aligned}
\widetilde{\pi}^N(\mathrm{d}x_{1:N}, \mathrm{d}\xi, k) &= \int_A \widetilde{\pi}^N(\mathrm{d}x_{1:N}, \mathrm{d}\xi, \mathrm{d}y, k) \\
&= \int_A N^{-1}\pi(\mathrm{d}y)\eta(\mathrm{d}\xi|x)T(y, \mathrm{d}x_k; \xi)\, \mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k, \xi) \\
&= \int_A N^{-1}\pi(\mathrm{d}x_k)\eta(\mathrm{d}\xi|x_k)T(x_k, \mathrm{d}y; \xi)\, \mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k, \xi) \\
&= N^{-1}\pi(\mathrm{d}x_k)\eta(\mathrm{d}\xi|x_k)\, \mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k, \xi) \\
&= \frac{\pi(x_k)\eta(\mathrm{d}\xi|x_k)}{Nq_k(x_k|\xi)}\, \mathbf{q}(\mathrm{d}x_{1:N}|x_k, \xi) \\
&= \frac{w_k(x_k|\xi)}{N\int_A m(x)\mu(\mathrm{d}x)}\mathbf{q}(\mathrm{d}x_{1:N}|\xi),
\end{aligned}$$

The second line is the joint distribution, the third line follows from reversibility of the Markov kernel, the fourth line integrates out y, and the last line follows from the definition

of the weights. The conditional distribution

$$\widetilde{\pi}^N(K=k|x_{1:N},\xi)=\frac{\widetilde{\pi}^N(x_{1:N},\xi,k)}{\sum_{i=1}^N\widetilde{\pi}^N(x_{1:N},\xi,i)}=\frac{w_k(x_k|\xi)}{\sum_{i=1}^N w_i(x_i|\xi)}=W_k(x_{1:N},\xi),$$

Similarly,

$$\begin{aligned}\widetilde{\pi}^N(\mathrm{d}y|x_{1:N},\xi,k)&=\frac{\widetilde{\pi}^N(\mathrm{d}x_{1:N},\mathrm{d}\xi,\mathrm{d}y,k)}{\widetilde{\pi}^N(\mathrm{d}x_{1:N},\mathrm{d}\xi,k)}\\&=\frac{N^{-1}\pi(\mathrm{d}y)\eta(\mathrm{d}\xi|x)T(x,\mathrm{d}x_k;\xi)\,\mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|k,\xi)}{N^{-1}\pi(\mathrm{d}x_k)\eta(\mathrm{d}\xi|x_k)\,\mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k,\xi)}\\&=T(x_k,\mathrm{d}y;\xi)\frac{\pi(\mathrm{d}x_k)\eta(\mathrm{d}\xi|x_k)\,\mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k,\xi)}{\pi(\mathrm{d}x_k)\eta(\mathrm{d}\xi|x_k)\,\mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k,\xi)}\\&=T(x_k,\mathrm{d}y;\xi).\end{aligned}$$

Part (iii):

$$\begin{aligned}\widetilde{\pi}^N(k,\mathrm{d}x_k)&=\int\widetilde{\pi}^N(k,\mathrm{d}y,\mathrm{d}\xi,\mathrm{d}x_{1:N})=N^{-1}\int\pi(\mathrm{d}y)\eta(\mathrm{d}\xi|y)T(y,\mathrm{d}x_k;\xi)\mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k,\xi)\\&=N^{-1}\pi(\mathrm{d}x_k)\int\eta(\mathrm{d}\xi|x_k)T(x_k,\mathrm{d}y;\xi)\mathbf{q}_{\backslash k}(\mathrm{d}x_{\backslash k}|x_k,\xi)=N^{-1}\pi(\mathrm{d}x_k)\,.\end{aligned}$$

Hence,

$$E_{\widetilde{\pi}^N}(f(X_K))=\sum_{k=1}^N N^{-1}\int\pi(\mathrm{d}x_k)f(x_k)=E_\pi(f).$$

Similarly, by first conditioning on $X_{1:N}$ and ξ, we obtain

$$\begin{aligned}E_{\widetilde{\pi}^N}(f(X_K))&=E_{\widetilde{\pi}^N}\Big(E_{\widetilde{\pi}^N(\cdot|x_{1:N},\xi)}f(X_K)\Big)\\&=E_{\widetilde{\pi}^N}\Big(N^{-1}\sum_{k=1}^N\int f(x_k)W_k(x_{1:N},\xi)T(x_k,\mathrm{d}y;\xi)\Big)\\&=E_{\widetilde{\pi}^N}\Big(\widehat{E}^N_{CIS}(f)\Big)\end{aligned}$$

□

Proof of Theorem 2. The proof follows from Part (ii) of Theorem 1 and because $\widetilde{\pi}^N(\mathrm{d}\xi,dx_{1:N}|y,k)=\Gamma^N(\mathrm{d}\xi,dx_{1:N}|y,k)$ □

A.2 Markov Interacting Importance Sampler for Conditional Distributions

Proof of Theorem 3. The proof is analogous to the proof of Theorem 1, with π replaced by $\pi_s(\cdot|x(\backslash s))$. □

Proof of Theorem 4. We write the MIIS Gibbs sampler as a Gibbs sampler in an augmented space. Each step of the algorithm consists in sampling from the following collapsed Gibbs sampler.

Algorithm 8. *For $s = 1, \ldots, d$,*

(i) Sample $X_{1:N}(s), \xi(s)|(y(s), k_s, y(\backslash s), \xi(\backslash s), (x_{1:N}(\backslash s)), k_{\backslash s})$ from $\Gamma_s^N(\mathrm{d}x_{1:N}(s), \mathrm{d}\xi(s)|y(s), k_s)$; and

(ii) Sample $Y(s), K_s|x_{1:N}(s), \xi(s), (y(\backslash s), \xi(\backslash s))$ from

$$\sum_{i=1}^{N} W_{s,i}(x_{1:N}(s); \xi(s)))I(K_s = i)T(x_i(s), \mathrm{d}y(s); \xi(s)).$$

To prove the theorem it is sufficient to show that the conditional density

$$\widetilde{\pi}^N(\mathrm{d}y(s), \mathrm{d}\xi(s), k_s, \mathrm{d}x_{1:N}(s)|y(\backslash s), \xi(\backslash s), k_{\backslash s}, x_{1:N}(\backslash s))$$

gives the sth step in Algorithm 8 above. The proof uses the same arguments as those in Theorem 2. The joint distribution

$$\widetilde{\pi}^N(\mathrm{d}y, \mathrm{d}\xi(\backslash s), (\mathrm{d}x_{1:N}(i), i \in \backslash s), k_{1:d}) = \frac{\pi(\mathrm{d}y)}{N^d} \prod_{i \neq s}^{d} \Gamma_i^N(\mathrm{d}x_{1:N}(i), \mathrm{d}\xi(i)|x(i), k_i, y(\backslash i)),$$

after integrating out $(x_{1:N}(s), \xi(s))$. Hence, the conditional joint distribution

$$\widetilde{\pi}^N(\mathrm{d}x_{1:N}(s), \mathrm{d}\xi(s)|\mathrm{d}y, \xi(\backslash s), (x_{1:N}(i), i \in \backslash s), k_{1:d}) = \Gamma_s^N(\mathrm{d}x_{1:N}(s), \mathrm{d}\xi(s)|x(s), k_s, x(\backslash s)),$$

which is consistent with part (i) of Algorithm 8. Similarly, $\widetilde{\pi}^N(\mathrm{d}y, \mathrm{d}\xi(s), \mathrm{d}x_{1:N}(s), k_s) =$

$N^{-1}\pi(\mathrm{d}y)\times\Gamma_s^N(\mathrm{d}x_{1:N}(s),\mathrm{d}\xi(s)|y(s),k_s,y(\backslash s))$, so

$$\begin{aligned}\widetilde{\pi}^N(\mathrm{d}y(\backslash s),\mathrm{d}\xi(s),\mathrm{d}x_{1:N}(s),k_s) &= \int_{A_s} N^{-1}\pi(\mathrm{d}y(s),\mathrm{d}y(\backslash s))\times\Gamma_s^N(\mathrm{d}x_{1:N}(s),\mathrm{d}\xi(s)|x(s),k_s,y(\backslash s))\\ &= N^{-1}\pi_{\backslash s}(\mathrm{d}y(\backslash s))\mathbf{q}_s(\mathrm{d}x_{1:N}(s)|\xi(s),\mathrm{d}y(\backslash s))\\ &\times\int_{A_s}\frac{\eta_s(\mathrm{d}\xi(s)|y(s),y(\backslash s))\pi_s(\mathrm{d}y(s)|y(\backslash s))}{q_{s,k_s}(\mathrm{d}x_{k_s}(s)|\xi(s),y(\backslash s))}T(y(s),\mathrm{d}x_{k_s}(s);\xi(s),y(\backslash s))\\ &= N^{-1}\pi_{\backslash s}(\mathrm{d}y(\backslash s))\mathbf{q}_s(\mathrm{d}x_{1:N}(s)|\xi(s),y(\backslash s))\\ &\frac{\eta_s(d\xi(s)|x_{k_s}(s),y(\backslash s))\pi_s(x_{k_s}(s)|\mathbf{x}(\backslash s))}{q_{s,k_s}(x_{k_s}(s)|\xi(s),x(\backslash s))}\\ &\propto\frac{\pi_{\backslash s}(\mathrm{d}y(\backslash s))}{N}\mathbf{q}_s(x_{1:N}(s)|\xi(s),x(\backslash s))\,w_{s,k_s}(x_{k_s}(s),\xi(s),x(\backslash s)).\end{aligned}$$

Hence, $\Pr(K_s=k_s|y(\backslash s),\xi(s),x_{1:N}(s))=W_{s,k_s}(x_{1:N}(s),\xi(s),y(\backslash s))$, which is consistent with Step (ii) of Algorithm 8 Following the same arguments as in the proof of Theorem 2, we can check that $\widetilde{\pi}^N(\mathrm{d}y(s)|y(\backslash s),\xi(s),x_{1:N}(s),k_s)=T(x_{k_s}(s),\mathrm{d}y(s),\xi(s),x(\backslash s))$. Finally, one can verify that the algorithm targets π by first integrating out $(x_{1:N}(i),\xi(i))$, $i=1,\ldots,d$, and then summing over $k_1,\ldots,k_d$. □

A.3 Convergence of MIIS

Before proving Theorem 5, we obtain a preliminary lemma.

Lemma 1. *Suppose Assumption 1 holds. Then,*

(i)

$$P(y,l;\mathrm{d}z\times\{k\})\geq\frac{1}{C}\frac{\pi(\mathrm{d}z)}{N}h_{l,k}(y,z).$$

(ii) Recursively define $H_{l,k}(y,z)=h_{l,k}(y,z)$ *and*

$$H_{l,k}^{t+1}(y,z):=\mathbb{E}_{N^{-1}\pi}[H_{l,J}^t(y,V)h_{J,k}(V,z)]=\sum_{j=1}^N N^{-1}\int_A H_{l,j}^t(y,v)h_{j,k}(v,z)\pi(dv).$$

Then,

$$P^t(y,l;\mathrm{d}z\times\{k\})\geq\left(\frac{1}{C}\right)^t\frac{\pi(\mathrm{d}z)}{N}H_{l,k}^t(y,z).\tag{36}$$

(iii) $H_{l,k}^t(y,z)>0$ *for* $t\geq 2$ *for all* $y,z\in A$.

Proof. We first obtain Part (i). Assumption 1 Part (i) implies that $W_k(x_{1:N};\xi)\geq$

$w_k(x_k;\xi)/CN$. Hence, for $k \neq l$,

$$\begin{aligned}
(CN)P(y,l;\mathrm{d}z\times\{k\}) &\geq \int_{A^{N+1}} \frac{\pi(\mathrm{d}x_k)\eta(\xi|x_k)}{q_k(\mathrm{d}x_k|\xi)} T(x_k,\mathrm{d}z;\xi)\Gamma^N(\mathrm{d}x_{1:N},\mathrm{d}\xi|y,l) \\
&= \pi(\mathrm{d}z)\int_{A^{N+1}} \frac{\eta(\xi|z)}{q_k(\mathrm{d}x_k|\xi)} T(z,\mathrm{d}x_k;\xi)\Gamma^N(\mathrm{d}x_{1:N},\mathrm{d}\xi|y,l) \\
&= \pi(\mathrm{d}z)\int_{A^{N+1}} \frac{\eta(\xi|z)\eta(\mathrm{d}\xi|y)}{q_k(\mathrm{d}x_k|\xi)q_l(\mathrm{d}x_l|\xi)} T(z,\mathrm{d}x_k;\xi)T(y,\mathrm{d}x_l;\xi)\mathbf{q}(\mathrm{d}x_{1:N}|\xi) \\
&= \pi(\mathrm{d}z)\int_{A^3} \frac{\mathbf{q}_{l,k}(\mathrm{d}x_l,\mathrm{d}x_k|\xi)}{q_k(\mathrm{d}x_k|\xi)q_l(\mathrm{d}x_l|\xi)} T(z,\mathrm{d}x_k;\xi)T(y,\mathrm{d}x_l;\xi)\eta(\xi|z)\eta(\mathrm{d}\xi|y) \\
&= \pi(\mathrm{d}z)\int_A \eta(\xi|z)\eta(\mathrm{d}\xi|y)\times\left[\int_{A^2} \frac{\mathbf{q}_{l,k}(\mathrm{d}x_l,\mathrm{d}x_k|\xi)}{q_k(\mathrm{d}x_k|\xi)q_l(\mathrm{d}x_l|\xi)} T(z,\mathrm{d}x_k;\xi)T(y,\mathrm{d}x_l;\xi)\right] \\
&= \pi(\mathrm{d}z)\int_A \eta(\xi|z)\eta(\xi|y)S_{l,k}(\xi,y,z)\mu(\mathrm{d}\xi) \\
&\geq \pi(\mathrm{d}z)h_{l,k}(y,z).
\end{aligned}$$

We can similarly result for $k = l$. We now prove part (ii). By part (i), Eq. (36) holds for $t = 1$. Suppose that (36) holds for some t. Then,

$$\begin{aligned}
P^{t+1}(y,l;\mathrm{d}z\times\{k\}) &= \sum_{j=1}^N \int_A P^t(y,l;\mathrm{d}v\times\{j\})P(v,j;\mathrm{d}z\times\{k\}) \\
&\geq \left(\frac{1}{C}\right)^t \sum_{j=1}^N \int_A \frac{\pi(\mathrm{d}v)}{N} H_{l,j}^t(y,v)\times\frac{\pi(\mathrm{d}z)}{CN}h_{j,k}(v,z) \\
&= \left(\frac{1}{C}\right)^{t+1} \frac{\pi(\mathrm{d}z)}{N}\,\frac{1}{N}\sum_{j=1}^N \int_A \pi(\mathrm{d}v)H_{l,j}^t(y,v)h_{j,k}(v,z) \\
&= \left(\frac{1}{C}\right)^{t+1} \frac{\pi(\mathrm{d}z)}{N}\,\mathbb{E}_{N^{-1}\pi}[H_{l,J}^t(y,V)h_{J,k}(V,z)] \\
&= \left(\frac{1}{C}\right)^{t+1} \frac{\pi(\mathrm{d}z)}{N}\,H_{l,k}^{t+1}(y,z).
\end{aligned}$$

Hence, the bound holds for all t. We now prove Part (iii). We first show that $H_{l,k}^t(y,z) > 0$ for $t = 2$ and then, recursively, for all $t \geq 2$. For any pair $y, z \in A$, and $l, k \in \{1{:}N\}$,

$$H_{l,k}^2(y,z) = \mathbb{E}_{N^{-1}\pi}[h_{l,J}(y,V)h_{J,k}(V,z)] \geq \mathbb{E}_{N^{-1}\pi}[h_{l,J}(y,V)h_{J,k}(V,z)I(J \in \mathcal{J}_l \cap \mathcal{J}_k)] > 0,$$

where the last inequality follows from Assumption 1 Part (ii). If $H_{l,j}^t(y,\cdot) > 0$, then

$$H_{l,k}^{t+1}(y,z) = \mathbb{E}_{N^{-1}\pi}[H_{l,J}^t(y,V)h_{J,k}(V,z)] \geq \mathbb{E}_{N^{-1}\pi}[H_{l,J}^t(y,V)h_{J,k}(V,z)I(J \in \mathcal{J}_k)] > 0.$$

□

Proof of Theorem 5. The sequence $\{(y^{(t)}, k^{(t)})\}$ from the MIIS algorithm is Markov, because the MIIS algorithm is a two component Gibbs sampler, and has transition kernel

$$P(y, l; B\times\{k\}) = \int_{A^{N+1}} W_k(x_{1:N}, \xi) T(x_k, B; \xi) \Gamma^N(\mathrm{d}x_{1:N}, \mathrm{d}\xi | y, l). \tag{37}$$

The proof shows that for all starting values $(y, l) \in A\times\{1{:}N\}$, the t^{th} step Markov transition kernel $P^t(y, l; B\times\{k\})$ is positive for all $t \geq 2$, and any $B \in \Omega$ such that $\pi(B) > 0$ and $k \in \{1{:}N\}$.

Suppose that $y \in A$, $B \in \Omega$ and $k, l \in \{1{:}N\}$. If $\pi(B) = 0$ then $P^t(y, l; B\times\{k\}) = 0$ for $t \geq 1$; if $\pi(B) > 0$ then $P^t(y, l; B\times\{k\}) > 0$ for all $t \geq 2$ by Lemma 1. This means that the marginal chain is $N^{-1}\pi$-irreducible and aperiodic and that $P(y, l; \mathrm{d}z\times\{k\})$ is absolutely continuous with respect to $N^{-1}\pi(\mathrm{d}z)$. It then follows from Theorem 1 and Corollary 1 in Tierney (1994) that for all $(y, l) \in A\times\{1{:}N\}$, $\lim_{t\to\infty} |P^t(y, l; \cdot - N^{-1}\pi(\cdot)|_{TV} = 0$, proving the first part of the theorem. Proof of second part. Define $g_l(z) := \min_{k\in\mathcal{J}_l} \underline{\mathrm{h}}_{l,k}(z) > 0$. Then,

$$H^2_{l,k}(y, z) = \sum_{k'\in\mathcal{J}_l} \int \pi(\mathrm{d}z') h_{l,k'}(y, z') h_{k',k}(z', z) \geq \Big(\int \pi(\mathrm{d}z') g_l(z') \Big) \sum_{k'\in\mathcal{J}_l} \underline{\mathrm{h}}_{k',k}(z).$$

Let $D_1 := \int \pi(\mathrm{d}z') g_l(z')$,

$$D_2 := \sum_{k'\in\mathcal{J}_l} \int \underline{\mathrm{h}}_{k',k}(z)\pi(\mathrm{d}z) \quad \text{and} \quad \nu(B) := D_2^{-1} \sum_{k'\in\mathcal{J}_l} \int_B \underline{\mathrm{h}}_{k',k}(z)\pi(\mathrm{d}z).$$

Then, from (36),

$$P^2(y, l; \mathrm{d}z, \{k\}) \geq C^{-2} D_1 D_2 N^{-1} \nu(\mathrm{d}z).$$

and uniform ergodicity follows from Proposition 2 in Tierney (1994). □

A.4 Convergence of the MIIS Gibbs Sampler

We again consider the marginal chain $\{y_t, \mathbf{l}_t, t \geq 0\}$ of the MIIS sampler, where $\mathbf{l}_t := (l_{1:d})_t$. Let $P_{s,M}(y(s), l_s; \mathrm{d}z(s)\times\{k_s\}|y(\backslash s))$ be the transition kernel for the sth component of the marginal chain. The transition kernel for the marginal chain is

$$P_M(y, \mathbf{l}; \mathrm{d}z\times\{\mathbf{k}\}) = \prod_{s=1}^{d} P_{s,M}(y(s), l_s; \mathrm{d}z(s)\times\{k_s\}|z(<s), y(>s)),$$

where we use the shorthand notation $z(<s) = z(1{:}s{-}1)$ and $y(>s) = y(s{+}1{:}d)$. Define

$$h_{\mathbf{l},\mathbf{k}}(y,z) := \prod_{s=1}^{d} h_{s,l_s,k_s}(y(s),z(s);z(<s),y(>s)). \tag{38}$$

We require the definition of the sub-stochastic kernel $H_{\mathbf{l},\mathbf{k}}(y,\mathrm{d}z) = C^{-1}N^{-d}h_{\mathbf{l},\mathbf{k}}(y,z)P_G(y,\mathrm{d}z)$ and, iteratively,

$$\begin{aligned}
H^{t+1}_{\mathbf{l},\mathbf{k}}(y,\mathrm{d}z) &= \frac{1}{CN^d}\sum_{\mathbf{j}\in\{1:N\}^d}\int_A H^t_{\mathbf{l},\mathbf{j}}(y,\mathrm{d}v)\, h_{\mathbf{j},\mathbf{k}}(v,z)P_G(v,\mathrm{d}z)\\
&= \frac{1}{CN^d}\sum_{\mathbf{j}\in\{1:N\}^d}\int_A h_{\mathbf{l},\mathbf{j}}(y,v)\, H^t_{\mathbf{j},\mathbf{k}}(v,\mathrm{d}z)P_G(y,\mathrm{d}v)\\
&= \mathbb{E}_{P_G(y,\cdot)/N^d}\left[h_{\mathbf{l},\mathbf{J}}(y,V)\, H^t_{\mathbf{J},\mathbf{k}}(V,\mathrm{d}z)\right].
\end{aligned}$$

Lemma 2. *Suppose Assumption 2 holds. Then,*

(i) The marginal chain $\{y^(t), \mathbf{l}^{(t)}, t\geq 0\}$ is Markov.

(ii) For $t = 1, 2, ...$

$$P^t_M(y,\mathbf{l};\mathrm{d}z{\times}\{\mathbf{k}\}) \geq H^t_{\mathbf{l},\mathbf{z}}(y,\mathrm{d}z).$$

(iii) Suppose $t\geq 2$, $B\in\Omega$, and $\pi(y)>0$. If $P^t_G(y,B)>0$ then $H^t_{\mathbf{l},\mathbf{k}}(y,B)>0$.

Proof. Part (i) follows from the construction of the MIIS sampler.

We show part (ii) by induction. By part (i) of Lemma 1, for each $s = 1, ..., d$,

$$P_{s,M}(y(s),l_s;k_s,\mathrm{d}z(s)|y(\backslash s)) \geq C^{-1/d}N^{-1}h_{s,l_s,k_s}(y(s),z(s);y(\backslash s))\pi_s(\mathrm{d}z(s)|y(\backslash s)).$$

Hence, for $t=1$, part (ii) follows form the definition of $P_M(y,\mathbf{l};z{\times}\{\mathbf{k}\})$ and $H_{\mathbf{l},\mathbf{k}}(y,\mathrm{d}z)$. Suppose $P^t_M(y,\mathbf{l};\mathrm{d}v{\times}\{\mathbf{j}\}) \geq H^t_{\mathbf{l},\mathbf{j}}(y,\mathrm{d}v)$, for $\mathrm{d}v\in\Omega$ and $\mathbf{j}\in\{1{:}N\}^d$. Then

$$\begin{aligned}
P^{t+1}_M(y,\mathbf{l};\mathrm{d}z{\times}\{\mathbf{k}\}) &= \sum_{\mathbf{j}\in\{1:N\}^d}\int_A P^t_M(y,\mathbf{l};\mathrm{d}v{\times}\{\mathbf{j}\})P_M(v,\mathbf{j};\mathrm{d}z{\times}\{\mathbf{k}\})\\
&\geq C^{-1}N^{-d}\int_A \sum_{\mathbf{j}\in\{1:N\}} h_{\mathbf{j},\mathbf{k}}(v,z)\, H^t_{\mathbf{l},\mathbf{j}}(y,\mathrm{d}v)P_G(v,\mathrm{d}z)\\
&= H^{t+1}_{\mathbf{l},\mathbf{z}}(y,\mathrm{d}z).
\end{aligned}$$

Then part (ii) also holds for $t{+}1$, proving the result.

Part (iii) follows By induction. We first show that the result holds for $t = 2$ and, then, we show that if the result holds for some $t \geq 2$, it also holds for t+1. Let $\mathcal{J}_{\mathbf{l}} = \times_{s=1}^{d} \mathcal{J}_{s,l_s}$ and verify that, under assumption 2(ii) part b, $\mathcal{J}_{\mathbf{l}} \cap \mathcal{J}_{\mathbf{k}} \neq \emptyset$, for any pair $\mathbf{l}, \mathbf{k} \in \{1{:}N\}$.

Suppose $t = 2$. If $P_G^2(y, B) > 0$, there is a set $F' \in \Omega$ such that $P_G(y, F') > 0$ and $P_G(x, B) > 0$ for $x \in F'$. Let $F' \supseteq F = F_1 \times \cdots \times F_d$. For $v \in F$ (i.e., each $v(s) \in F_s$), $s = 1, \ldots, d$, and j in $\mathcal{J}_{s,l} \cap \mathcal{J}_{s,k}$,

$$h_{s,l,j}(y(s), v(s); v(< s), y(> s)) h_{s,j,k}(v(s), z(s); z(< s), v(> s)) > 0,$$

from Assumption 2(ii) part (c). Therefore,

$$\sum_{\mathbf{j} \in \{1:N\}} h_{\mathbf{l},\mathbf{j}}(y, v) h_{\mathbf{j},\mathbf{k}}(v, z) \geq \prod_{s=1}^{d} \sum_{j \in \mathcal{J}_{s,k_s} \cap \mathcal{J}_{s,l_s}} h_{s,l_s,j}(y(s), v(s); v(< s), y(> s)) \times h_{s,j,k_s}(v(s), z(s); z(< s), v(> s)) > 0$$

for $v \in F$, and any $\mathbf{l}, \mathbf{k} \in \{1{:}N\}^d$. Hence

$$\begin{aligned} H_{\mathbf{l},\mathbf{k}}^2(y, z) &= \frac{1}{C^2} \int_B \int_A \sum_{\mathbf{j} \in \{1:N\}} h_{\mathbf{l},\mathbf{j}}(y, v) h_{\mathbf{j},\mathbf{k}}(v, z) P_G(y, \mathrm{d}v) P_G(v, \mathrm{d}z) \\ &\geq \frac{1}{C^2} \int_B \left\{ \int_F \sum_{\mathbf{j} \in \mathcal{J}_{\mathbf{l}} \cap \mathcal{J}_{\mathbf{k}}} h_{\mathbf{l},\mathbf{j}}(y, v) h_{\mathbf{j},\mathbf{k}}(v, z) P_G(y, \mathrm{d}v) P_G(v, \mathrm{d}z) \right\} > 0, \end{aligned}$$

where the last line follows from calculating each integral between brackets over F_s, $s = 1, \cdots, d$.

Suppose that part (iii) holds for some $t \geq 2$ and that $P_G^{t+1}(y, B) > 0$. Then,

$$\begin{aligned} H_{\mathbf{l},\mathbf{k}}^{t+1}(y, B) &= \frac{1}{CN^d} \sum_{\mathbf{j} \in \{1:N\}} \int_B \int_A H_{\mathbf{l},\mathbf{j}}^{t}(y, \mathrm{d}v) h_{\mathbf{j},\mathbf{k}}(v, z) P_G(v, \mathrm{d}z) \\ &\geq \frac{1}{CN^d} \int_B \int_F \left[\sum_{\mathbf{j} \in \mathcal{J}_{\mathbf{k}}} H_{\mathbf{l},\mathbf{j}}^{t}(y, \mathrm{d}v) h_{\mathbf{j},\mathbf{k}}(v, z) \right] P_G(v, \mathrm{d}z) > 0, \end{aligned}$$

where $F \in \Omega$ is such that $P_G(x, B) > 0$ for $x \in F$ and $P_G^t(y, F) > 0$. The result holds for any $\mathbf{l}$ and $\mathbf{k}$ in $\{1{:}N\}^d$. □

Proof of Theorem 6. The result follows from Lemma 2 and Theorem 1 in Tierney (1994). First define the Markov kernel $[N^{-d} P_G](y, \mathbf{l}; B \times \{\mathbf{k}\})$, that is the kernel of the Gibbs sampler that draws $(z(s), k_s) | (z(< s), y(> s), \mathbf{k}_{<s}, \mathbf{l}_{>s})$ from $N^{-1} \pi_s(z(s) | z(< s), y(> s))$, sequentially. If the Gibbs kernel P_G is irreducible and aperiodic, so it is the kernel $[N^{-d} P_G]$, since all $k_s \in \{1{:}N\}$, $s = 1, \ldots, d$, are accessible at each iteration. The proof

consists in showing that accessible sets from $[N^{-d}P_G]^t$, the ideal Gibbs in $t \geq 2$ steps, are also accessible by the MIIS-Gibbs kernel after t iterations, P_M^t. Lemma 2 (i) shows that P_M is a Markov kernel. Parts (ii) and (iii) together show that $P_G(y, B) > 0$ implies that $P_M(y, \mathbf{k}; B \times \{\mathbf{l}\}) > 0$ for any pair $(\mathbf{l}, \mathbf{k})$. Hence, all sets accessible by $[N^{-d}P_G]$ are also accessible by P_M, which implies that P_M is also irreducible. To show that P_M is aperiodic, we assume by contradiction that P_M is not aperiodic. In this case, $[N^{-d}P_G]$ would have to be periodic as well, which contradicts with the assumption that $[N^{-d}P_G]$ is aperiodic. The result follows from Theorem 1 of Tierney (1994). □

It also follows from Theorem 6 that, $\lim_{t\to\infty} P_M^t(y(s), l_s; \cdot|y(\backslash s)) = N^{-1}\pi_s(\cdot|y(\backslash s)$, which implies that the control variates in Section 5.3 can be safely used.

Proof of Corollary 4. We can check that the conditions of Assumption 2 hold in a similar way to the proof of Corollary 1. The result follows from Theorem 6. □

A.5 Proofs of consistency

Proof of Corollary 5. The distribution of $\{l^{(t)}, y^{(t)}\}$ converges to $N^{-1}\pi(\cdot)$ by Theorem 5. Let $\widehat{E}^N_{CIS,t}(f)$ be defined by Equation (22) in Section 5.1. The result now follows from Lemma 1, which shows that each $\widehat{E}^N_{CIS,t}(f)$ is unbiased and by the strong law of large numbers for ergodic sequences (Tierney, 1994, Theorem 3). □

Proof of Corollary 6. The distribution of $(l^{(t)}_{1:d}, y^{(t)})$ converges to $N^{-d}\pi(\cdot)$ by Theorem 6 The result follows from Lemma 3 and the strong law of large numbers for ergodic sequences (Tierney, 1994, Theorem 3). □

Proof of Corollary 7. For any f with $E_\pi(|f|) < \infty$, it follows from Corollary 5 that $\widehat{E}^M_{MC}(f) \to \pi(f)$, and $\widehat{E}^{M,N}_{MIIS}(f) \to \pi(f)$ with probability one. This means that $\widehat{E}^M_{MC}(f) - \widehat{E}^{M,N}_{MIIS}(f) \to 0$, with probability one. Hence, for any constant $\boldsymbol{\kappa} \in \mathbb{R}^p$, and π-integrable functions $g_1, \ldots, g_p$, the linear combination $\sum_{i=1}^p \kappa_i[\pi^M_{MC}(g_i) - \pi^{M,N}_{MIIS}(g_i)] \to 0$ with probability one. The proof now follows from Corollary 5 □

Proof of Corollary 8. The proof of this corollary follows the same arguments used in the proof of Corollary 7. □