Journal of Machine Learning Research 22 (2021) 1-45 Submitted 6/19; Revised 5/21; Published 6/21

Pseudo-Marginal Hamiltonian Monte Carlo

Johan Alenlöv JOHAN.ALENLOV@LIU.SE
Division of Statistics and Machine Learning
Linköping University
Linköping, 581 83, Sweden

Arnaud Doucet DOUCET@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford, OX1 3TG, United Kingdom

Fredrik Lindsten FREDRIK.LINDSTEN@LIU.SE
Division of Statistics and Machine Learning
Linköping University
Linköping, 581 83, Sweden

Editor: Ryan Adams

Abstract

Bayesian inference in the presence of an intractable likelihood function is computationally challenging. When following a Markov chain Monte Carlo (MCMC) approach to approximate the posterior distribution in this context, one typically either uses MCMC schemes which target the joint posterior of the parameters and some auxiliary latent variables, or pseudo-marginal Metropolis—Hastings (MH) schemes. The latter mimic a MH algorithm targeting the marginal posterior of the parameters by approximating unbiasedly the intractable likelihood. However, in scenarios where the parameters and auxiliary variables are strongly correlated under the posterior and/or this posterior is multimodal, Gibbs sampling or Hamiltonian Monte Carlo (HMC) will perform poorly and the pseudo-marginal MH algorithm, as any other MH scheme, will be inefficient for high-dimensional parameters. We propose here an original MCMC algorithm, termed pseudo-marginal HMC, which combines the advantages of both HMC and pseudo-marginal schemes. Specifically, the PM-HMC method is controlled by a precision parameter N, controlling the approximation of the likelihood and, for any N, it samples the *marginal posterior* of the parameters. Additionally, as N tends to infinity, its sample trajectories and acceptance probability converge to those of an ideal, but intractable, HMC algorithm which would have access to the intractable likelihood and its gradient. We demonstrate through experiments that PM-HMC can outperform significantly both standard HMC and pseudo-marginal MH schemes.

Keywords: Hamiltonian Monte Carlo, pseudo-marginal, Markov chain Monte Carlo, latent variable models

1. Introduction

Let $y \in \mathcal{Y}$ denote some observed data and $\theta \in \Theta \subseteq \mathbb{R}^d$ denote parameters of interest. We write $\theta \mapsto p(y \mid \theta)$ for the likelihood of the observations and we assign a prior for θ of density $p(\theta)$ with respect to Lebesgue measure $\mathrm{d}\theta$. Hence the posterior density of interest is given

by

$$\pi(\theta) = p(\theta \,|\, y) \propto p(y \,|\, \theta)p(\theta). \tag{1}$$

For complex Bayesian models, the posterior (1) needs to be approximated numerically. When the likelihood $p(y \,|\, \theta)$ can be evaluated pointwise, this can be achieved using standard MCMC schemes. However, we will consider here the scenario where $p(y \,|\, \theta)$ is intractable, in the sense that it cannot be evaluated pointwise. We detail below two important scenarios where an intractable likelihood occurs.

Example: Latent variable models. Consider the model

$$X_k \overset{\text{i.i.d.}}{\sim} f_\theta(\cdot), \qquad\qquad Y_k \,|\, X_k \sim g_\theta(\cdot \,|\, X_k), \tag{2}$$

where $(X_k)_{k\geq 1}$ are $\mathbb{R}^n$-valued latent variables, $(Y_k)_{k\geq 1}$ are Y-valued. We write $i : j := \{i, i+1, ..., j\}$ for any $i \leq j$. Having observed $Y_{1:T} = y_{1:T} = y$ (thus, $\mathcal{Y} = \mathsf{Y}^T$) the likelihood is given by $p(y_{1:T} \,|\, \theta) = \prod_{k=1}^T p(y_k \,|\, \theta)$, where each term $p(y_k \,|\, \theta)$ satisfies

$$p(y_k \,|\, \theta) = \int_{\mathbb{R}^n} f_\theta(x_k) g_\theta(y_k \,|\, x_k) \mathrm{d}x_k. \tag{3}$$

If the integral (3) cannot be computed in closed form then the likelihood $p(y_{1:T} \,|\, \theta)$ is intractable.

Example: Approximate Bayesian computation (ABC). Consider the scenario where $\theta \mapsto \widetilde{p}(y \,|\, \theta)$ is the "true" likelihood function. We cannot compute it pointwise but we assume we are able to simulate some pseudo-observations $Z \sim \widetilde{p}(\cdot \,|\, \theta)$ using $Z = \gamma(\theta, V)$ where $V \sim \lambda(\cdot)$ for some auxiliary variable distribution λ and mapping $\gamma : \Theta \times \mathcal{V} \to \mathcal{Y}$. Given a kernel $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, the ABC approximation of the posterior is given by (1) where the intractable ABC likelihood is

$$p(y \,|\, \theta) = \int_{\mathcal{Y}} K(y \,|\, z)\widetilde{p}(z \,|\, \theta)\mathrm{d}z = \int K(y \,|\, \gamma(\theta, v))\lambda(v)\mathrm{d}v. \tag{4}$$

Two standard approaches to perform MCMC in these scenarios are:

1. Implement standard MCMC algorithms to sample from the joint distribution of the parameters and auxiliary variables; e.g. in the latent variable context we would target $p(\theta, x_{1:T} \,|\, y_{1:T}) \propto p(\theta)\prod_{k=1}^T f_\theta(x_k)g_\theta(y_k \,|\, x_k)$ and in the ABC context $p(\theta, v \,|\, y) \propto p(\theta)\lambda(v)K(y \,|\, \gamma(\theta, v))$. Gibbs-type approaches alternate between sampling from the parameters conditioned on the data and auxiliary variables, and sampling the auxiliary variables conditioned on the parameters and data. These approaches can converge very slowly if these variables are strongly correlated under the target (Andrieu et al., 2010, Section 2.3). Hamiltonian Monte Carlo (HMC) methods (Duane et al., 1987) offer a possible remedy, but can also struggle in cases where there are strong non-linear dependencies between variables, or when the joint posterior is multimodal (Neal, 2011, Section 5.5.7).

2. Use a pseudo-marginal MH algorithm jointly accepting or rejecting the parameters and auxiliary variables replacing the intractable likelihood term by a non-negative unbiased estimate of the true likelihood; see (Andrieu et al., 2010; Andrieu and Roberts, 2009;

Beaumont, 2003; Flury and Shephard, 2011; Lin et al., 2000). For example, in the ABC context the pseudo-marginal MH algorithm is a MH algorithm targeting $p(\theta, v \mid y)$ using a proposal distribution $q(\theta, \theta')\lambda(v')$. As for any MH algorithm, it can be difficult to select a proposal $q(\theta, \theta')$ which results in an efficient sampler when Θ is high-dimensional.

In many scenarios, the marginal posterior (1) will have a "nicer" structure than the joint posterior of the parameters and auxiliary variables which often exhibit complex patterns of dependence and multimodality. For example, discrete choice models are a widely popular class of models in health economics, e-commerce, marketing and social sciences used to analyze choices made by consumers, individuals or businesses (Train, 2009). When the population is heterogeneous, such models can be represented as (2) where $f_\theta(\cdot)$ is a mixture distribution, the number of components representing the number of latent classes; see e.g. (Burda et al., 2008). In this context, the paucity of data typically available for each individual is such that the joint posterior $p(\theta, x_{1:T} \mid y_{1:T}) = p(\theta \mid y_{1:T}) \prod_{k=1}^{T} p(x_k \mid y_k, \theta)$ will be highly multimodal while the marginal $p(\theta \mid y_{1:T})$ will only have symmetric well-separated modes for T large enough (or one mode if constraints on the mixture parameters are introduced). Such problems also arise in biostatistics (Komárek and Lesaffre, 2008). In these scenarios, current MCMC methods will be inefficient. In this article, we propose a novel HMC scheme, termed pseudo-marginal HMC (PM-HMC), which mimics the HMC algorithm targeting the marginal posterior (1) while integrating out numerically the auxiliary variables. The method is a so called *exact approximation* in the sense that its limiting distribution (marginally in θ) is exactly $\pi(\theta)$, despite the fact that the algorithm is using a finite-precision approximation of the likelihood internally.

1.1 Related Work

Stochastic gradient MCMC (Welling and Teh, 2011; Chen et al., 2014; Ding et al., 2014; Leimkuhler and Shang, 2016)—including HMC-like methods—are a popular class of algorithms for approximate posterior sampling when an unbiased estimate of the *log-likelihood gradient* is available. The typical scenario is when the number of data points T is prohibitively large for evaluating the full gradient, in which case subsampling (mini-batching) can be used to approximate the gradient unbiasedly. This is in contrast with the setting studied in this paper, where we assume that we have access to an unbiased estimate of the likelihood itself, but not of the log-likelihood gradient. Furthermore, these methods are inconsistent for finite step sizes and typically require some type of variance reduction techniques to be efficient (Shang et al., 2015). Recently, Umenberger et al. (2019) have proposed to use debiasing (Jacob et al., 2020) of log-likelihood gradients within stochastic gradient HMC, but this still results in an inconsistent method. Another approach to solving this issue is by Dang et al. (2019), who proposes the use of energy-conserving subsampling which combined with a signed pseudo-marginal algorithm (Quiroz et al., 2021) can be made exact.

Hamiltonian ABC (Meeds et al., 2015) also performs HMC with stochastic gradients but calculates these gradients by using forward simulation. Similarly to stochastic gradient MCMC (and in contrast with PM-HMC), this results in an approximate MCMC which does not preserve the distribution of interest. Constrained HMC (Graham and Storkey, 2017) uses HMC to jointly infer the parameters and auxiliary variables in an auxiliary variable formulation of a Gaussian kernel ABC posterior.

Kernel HMC (Strathmann et al., 2015) is another related approach which approximates the gradients by fitting an exponential family model in a reproducing kernel Hilbert space. If the adaptation of the kernel stops after a finite time, or if the adaptation probability decays to zero, then the method can be shown to attain detailed balance, and thus targets the correct distribution of interest. However, the kernel-based approximation gives rise to a bias in the gradients which is difficult to control and there is no guarantee that the trajectories closely follow the ideal HMC. Kernel HMC requires the selection of a kernel and, furthermore, some appropriate approximation thereof, since the computational cost of a full kernel-based approximation grows cubically with the number of MCMC iterations; see Strathmann et al. (2015) for details.

Pseudo-marginal slice sampling (Murray and Graham, 2016) is closely related to PM-HMC in the sense that both algorithms target the same extended distribution (given by (8) in the consecutive section) as the pseudo-marginal MH and are therefore also *exact approximations* (Andrieu and Roberts, 2009). Pseudo-marginal slice sampling relies instead on alternating the sampling of the parameters given the likelihood estimate and sampling the likelihood estimate given the parameters. In this framework different MCMC methods such as slice sampling (Neal, 2003) and Metropolis—Hastings can be used to sample from the conditional distributions—typically one always uses *elliptical slice sampling* (Murray et al., 2010) for sampling of the likelihood estimates. In comparison our proposed algorithm samples jointly the parameters and the likelihood estimate.

Building on a preprint of the present article Nemeth et al. (2019) have proposed the so called *pseudo-extended HMC* method for sampling multimodal target distributions. This can be viewed as a special case of PM-HMC applied to a model with a single latent variable and no top-level parameter. Osmundsen et al. (2018) have also built on the present work and propose a version of PM-HMC using efficient importance sampling, which they apply to inference in dynamical systems.

1.2 Outline of the Paper

Our paper is organized as follows. In Section 2 we present our algorithm by first introducing in Section 2.1 the standard HMC method. The PM-HMC algorithm is presented in Section 2.2 with an illustration on the latent variable model in Section 2.3. In Section 2.4 we present a customized numerical integrator—an essential component of PM-HMC that is needed for handling the increasing dimension of the latent variable space. This completes the specification of the PM-HMC algorithm. The theoretical justification of our algorithm is presented in Section 3 with the proofs postponed to the appendix. Finally we present numerical results demonstrating the usefulness of our algorithm in Section 4.

2. Pseudo-marginal Hamiltonian Monte Carlo

In this section we present the proposed method. First we give some background on Hamiltonian Monte Carlo (HMC) and, specifically, we consider the case of targeting the marginal posterior $\pi(\theta)$ using HMC. This results in an ideal but intractable algorithm. We then present PM-HMC, an *exact approximation* of the marginal HMC.

2.1 Marginal Hamiltonian Monte Carlo

The Hamiltonian formulation of classical mechanics is at the core of HMC methods. Recall that $\theta \in \Theta \subseteq \mathbb{R}^d$. We identify the potential energy as the negative unnormalized log-target and introduce a momentum variable $\rho \in \mathbb{R}^d$ which defines the kinetic energy $\frac{1}{2}\rho^\top \rho$ of the system.[1] The resulting "exact" Hamiltonian is given by

$$H_{\mathtt{ex}}(\theta, \rho) = -\log\, p(\theta) - \log\, p(y \,|\, \theta) + \frac{1}{2}\rho^\top \rho. \tag{5}$$

We associate a probability density on $\mathbb{R}^d \times \mathbb{R}^d$ to this Hamiltonian through

$$\pi(\theta, \rho) \propto \exp(-H_{\mathtt{ex}}(\theta, \rho)) = \pi(\theta)\mathcal{N}(\rho \,|\, 0_d, I_d), \tag{6}$$

where $\mathcal{N}(z \,|\, \mu, \Sigma)$ denotes the normal density of argument z, mean μ and covariance Σ.

Assuming that the prior density and likelihood function are continuously differentiable, the Hamiltonian dynamics corresponds to the equations of motion

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \nabla_\rho H_{\mathtt{ex}} = \rho, \qquad \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\nabla_\theta H_{\mathtt{ex}} = \nabla_\theta \log\, p(\theta) + \nabla_\theta \log\, p(y \,|\, \theta). \tag{7}$$

A key property of this dynamics is that it preserves the Hamiltonian, i.e. $H_{\mathtt{ex}}(\theta(t), \rho(t)) = H_{\mathtt{ex}}(\theta(0), \rho(0)) = H_0$ for any $t \geq 0$. This enables large moves in the parameter space to be made by simulating the Hamiltonian dynamics. However, to sample from the posterior, it is necessary to explore other level sets of the Hamiltonian; this can be achieved by periodically updating the momentum ρ according to its marginal under π, i.e. $\rho \sim \mathcal{N}(0_d, I_d)$.

The Hamiltonian dynamics only admits a closed-form solution in very simple scenarios, e.g., if $\pi(\theta)$ is normal. Hence, in practice, one usually needs to resort to a numerical integrator. Typically, the Verlet method, also known as the Leapfrog method, is used due to its favourable properties in the context of HMC (Leimkuhler and Matthews, 2015, p. 60; Neal, 2011, Section 5.2.3.3). In particular, this integrator is symplectic which implies that the Jacobian of the transformation $(\theta(0), \rho(0)) \to (\theta(t), \rho(t))$ is unity for any $t > 0$. Because of numerical integration errors, the Hamiltonian is not preserved along the discretized trajectory but this can be accounted for by an MH rejection step. The resulting HMC method is given by the following: at state $\theta := \theta[0]$, *(i)* sample the momentum variable $\rho[0] \sim \mathcal{N}(0_d, I_d)$, *(ii)* simulate approximately the Hamiltonian dynamics over L discrete time steps using a symplectic integrator, yielding $(\theta[L], \rho[L])$, and *(iii)* accept $(\theta[L], \rho[L])$ with probability $1 \wedge \pi(\theta[L], \rho[L])/\pi(\theta[0], \rho[0]) = 1 \wedge \exp(H_{\mathtt{ex}}(\theta[0], \rho[0]) - H_{\mathtt{ex}}(\theta[L], \rho[L]))$. We refer to Neal (2011) for details and a more comprehensive introduction.

2.2 Pseudo-Marginal Hamiltonian dynamics

When the likelihood is intractable, it is not possible to approximate numerically the Hamiltonian dynamics (7), as the integrator requires evaluating $\nabla_\theta \log p(y \,|\, \theta)$ pointwise. We will address this difficult by instead considering a Hamiltonian system defined on an extended phase space when the following assumption holds.

1. For simplicity we assume unit mass. The extension to a general mass matrix is straightforward.

- **Assumption 1**. There exists $(\theta, \mathbf{u}) \mapsto \widehat{p}(y \,|\, \theta, \mathbf{u}) \in \mathbb{R}^+$ where $\mathbf{u} \in \mathcal{U}$ and $m(\cdot)$ a probability density on $\mathcal{U}$ such that $p(y \,|\, \theta) = \int \widehat{p}(y \,|\, \theta, \mathbf{u}) m(\mathbf{u}) \mathrm{d}\mathbf{u}$.

Assumption 1 equivalently states that $\widehat{p}(y \,|\, \theta, \mathbf{U})$ is a non-negative unbiased estimate of $p(y \,|\, \theta)$ when $\mathbf{U} \sim m(\cdot)$. This assumption is at the core of pseudo-marginal methods (Andrieu and Roberts, 2009; Deligiannidis et al., 2018; Lin et al., 2000; Murray and Graham, 2016) which rely on the introduction of an extended target density

$$\bar{\pi}(\theta, \mathbf{u}) = \pi(\theta) \frac{\widehat{p}(y \,|\, \theta, \mathbf{u})}{p(y \,|\, \theta)} m(\mathbf{u}) \propto p(\theta) \widehat{p}(y \,|\, \theta, \mathbf{u}) m(\mathbf{u}). \tag{8}$$

This extended target admits $\pi(\theta)$ as a marginal under Assumption 1. The pseudo-marginal MH algorithm is for example a 'standard' MH algorithm targeting (8) using the proposal $q(\theta, \theta') m(\mathbf{u}')$ when in state $(\theta, \mathbf{u})$, resulting in an acceptance probability

$$1 \wedge \frac{p(\theta') \widehat{p}(y \,|\, \theta', \mathbf{u}')}{p(\theta) \widehat{p}(y \,|\, \theta, \mathbf{u})} \frac{q(\theta', \theta)}{q(\theta, \theta')}. \tag{9}$$

Instead of exploring the extended target distribution $\bar{\pi}(\theta, \mathbf{u})$ using an MH strategy, we will rely here on an HMC mechanism. Our method will use an additional assumption on the distribution of the auxiliary variables and regularity conditions on the simulated likelihood function $\widehat{p}(y \,|\, \theta, \mathbf{u})$.

- **Assumption 2**. $\mathcal{U} = \mathbb{R}^D$, $m(\mathbf{u}) = \mathcal{N}(\mathbf{u} \,|\, 0_D, I_D)$, $(\theta, \mathbf{u}) \longmapsto \widehat{p}(y \,|\, \theta, \mathbf{u})$ is continuously differentiable and $(\theta, \mathbf{u}) \longmapsto \log \nabla \widehat{p}(y \,|\, \theta, \mathbf{u})$ can be evaluated point-wise.

Our algorithm will leverage the fact that $m(\mathbf{u})$ is a normal distribution. Assumptions 1 and 2 will be standing assumptions from now on and allow us to define the following extended Hamiltonian

$$H(\theta, \rho, \mathbf{u}, \mathbf{p}) = -\log p(\theta) - \log \widehat{p}(y \,|\, \theta, \mathbf{u}) + \frac{1}{2}\Big\{\rho^T \rho + \mathbf{u}^T \mathbf{u} + \mathbf{p}^T \mathbf{p}\Big\}, \tag{10}$$

with a corresponding joint probability density on $\mathbb{R}^{2d+2D}$

$$\bar{\pi}(\theta, \rho, \mathbf{u}, \mathbf{p}) = \bar{\pi}(\theta, \mathbf{u}) \mathcal{N}(\rho \,|\, 0_d, I_d) \mathcal{N}(\mathbf{p} \,|\, 0_D, I_D), \tag{11}$$

which also admits $\pi(\theta)$ as a marginal. Here $\mathbf{p} \in \mathbb{R}^D$ are momentum variables associated with $\mathbf{u}$. The corresponding equations of motion associated with this extended Hamiltonian are then given by

$$\frac{d}{dt} \begin{pmatrix} \theta \\ \rho \\ \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \rho \\ \nabla_\theta \log p(\theta) + \nabla_\theta \log \widehat{p}(y \,|\, \theta, \mathbf{u}) \\ \mathbf{p} \\ -\mathbf{u} + \nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta, \mathbf{u}) \end{pmatrix} := \widehat{\Psi}(\theta, \rho, \mathbf{u}, \mathbf{p}). \tag{12}$$

Compared to (7), the intractable log-likelihood gradient $\nabla_\theta \log p(y \,|\, \theta)$ appearing in (7) has now been replaced by the gradient $\nabla_\theta \log \widehat{p}(y \,|\, \theta, \mathbf{u})$ of the log-simulated likelihood $\widehat{p}(y \,|\, \theta, \mathbf{u})$ where $\mathbf{u}$ evolves according to the third and fourth rows of (12).

Remark 1 *The normality assumption is in itself not restrictive as we can think of* $\mathbf{u}$ *as the internal random variables used to compute the likelihood estimator (we can always generate, e.g., a uniform random variate from a normal one using the cumulative distribution function). This assumption has also been used by Deligiannidis et al. (2018) for the correlated pseudo-marginal MH method and Murray and Graham (2016) for pseudo-marginal slice sampling. However, the assumed regularity of the simulated likelihood function* $\widehat{p}(y \mid \theta, \mathbf{u})$ *limits its range of applications. For example, in a state-space model context the likelihood is usually estimated using a particle filter, as in Andrieu et al. (2010), but this results in a discontinuous function* $(\theta, \mathbf{u}) \longmapsto \widehat{p}(y \mid \theta, \mathbf{u})$.

In latent variable models the use of the auxiliary variables will be to construct an importance sampling estimator. In such cases, the user has the freedom to specify the importance distribution such that Assumption 2 is satisfied. In particular, it can be differ from the prior distribution of the latent variables. We use this freedom in the numerical example in Section 4.3.

2.3 Illustration on Latent Variable Models

Consider the latent variable model described by (2) and (3). In this scenario, the intractable likelihood can be unbiasedly estimated using importance sampling. We introduce an importance density $q_\theta(x_k \mid y_k)$ for the latent variable X_k which we assume can be simulated using $X_k = \gamma_k(\theta, V)$ where $\gamma_k : \Theta \times \mathbb{R}^p \to \mathbb{R}^n$ is a deterministic map and $V \sim \mathcal{N}(0_p, I_p)$. We can then approximate the likelihood, using N samples for each k, through

$$\widehat{p}(y_{1:T} \mid \theta, \mathbf{U}) = \prod_{k=1}^{T} \widehat{p}(y_k \mid \theta, \mathbf{U}_k), \qquad \text{where} \qquad \widehat{p}(y_k \mid \theta, \mathbf{U}_k) = \frac{1}{N} \sum_{i=1}^{N} \omega_\theta(y_k, \mathbf{U}_{k,i}), \tag{13}$$

with $\mathbf{U}_k := (\mathbf{U}_{k,1}, \ldots, \mathbf{U}_{k,N})$ and $\omega_\theta(y_k, \mathbf{U}_{k,i}) = \frac{g_\theta(y_k \mid X_{k,i}) f_\theta(X_{k,i})}{q_\theta(X_{k,i} \mid y_k)}$, where $X_{k,i} = \gamma_k(\theta, \mathbf{U}_{k,i})$ and $\mathbf{U}_{k,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_p, I_p)$. We thus have $D = TNp$ in this scenario and

$$\nabla_\theta \log \widehat{p}(y_{1:T} \mid \theta, \mathbf{U}) = \sum_{k=1}^{T} \sum_{i=1}^{N} \frac{\omega_\theta(y_k, \mathbf{U}_{k,i})}{\sum_{j=1}^{N} \omega_\theta(y_k, \mathbf{U}_{k,j})} \nabla_\theta \log \omega_\theta(y_k, \mathbf{U}_{k,i}), \tag{14}$$

$$\nabla_{\mathbf{u}_{k,i}} \log \widehat{p}(y_{1:T} \mid \theta, \mathbf{U}) = \frac{\omega_\theta(y_k, \mathbf{U}_{k,i})}{\sum_{j=1}^{N} \omega_\theta(y_k, \mathbf{U}_{k,j})} \nabla_{\mathbf{u}_{k,i}} \log \omega_\theta(y_k, \mathbf{U}_{k,i}). \tag{15}$$

The pseudo-marginal MH algorithm can mix very poorly if the relative variance of the likelihood estimator is large; e.g. if $N = 1$ in (13). On the contrary, in the context of pseudo-marginal Hamiltonian dynamics, the case $N = 1$ corresponds to Hamiltonian dynamics for a re-parameterization of the original joint model $p(\theta, x_{1:T} \mid y_{1:T})$, which can work well for simple targets; see e.g. Betancourt and Girolami (2015). In some sense, PM-HMC interpolates between joint and marginal Hamiltonian dynamics as we increase N from 1 to ∞. Thus, we expect PM-HMC to be much less sensitive to the choice of N than pseudo-marginal MH; see Section 5 for empirical results.

Remark 2 *Increasing the dimension of the target distribution with* N *can seem counter-intuitive as exploring a higher dimensional space is often perceived to be a harder task.*

However, when extending with multiple Gaussian variables as we do here this can actually help us to explore complex distributions as the extended distribution can be better connected than the marginal. Even a disconnected marginal which is hard to explore for any MCMC method may, when extended in this way, be connected in the extended space and easier to explore. This is exploited in the pseudo-extended HMC algorithm by Nemeth et al. (2019) which builds upon this work and provides some empirical illustrations of this effect. Furthermore, in Section 2.4 we propose a numerical integrator for (12) *which is robust to the increasing dimension of the target distribution when increasing* N*.*

The construction above is based on a standard importance sampling estimator of the likelihood, but more sophisticated estimators could be used. For instance, by building upon our work, Osmundsen et al. (2018) have recently investigated the use of *efficient importance sampling* (Richard and Zhang, 2007) to approximate the likelihood for state-space models within a PM-HMC algorithm.

2.4 Numerical Integration via Operator Splitting

The pseudo-marginal Hamiltonian dynamics (12) can not in general be solved analytically and, as usual, we therefore need to make use of a numerical integrator. The standard choice in HMC is to use the Verlet scheme (Leimkuhler and Matthews, 2015, Section 2.2) which is a symplectic integrator of order $O(h^2)$, where h is the integration step-size. However, the error of the Verlet integrator will also depend on the dimension of the system. For the pseudo-marginal target density (11), we therefore need to take the effect of the D-dimensional auxiliary variable $\mathbf{u}$ into account. For instance, in the context of importance sampling-based PM-HMC for latent variable models discussed above we have $D = TNp$, i.e., the dimension of the extended target increases linearly with the number of importance samples N. This is an apparent problem—by increasing N we expect to obtain solution trajectories closer to those of the true marginal Hamiltonian system. However we also need to integrate numerically an ordinary differential equation of dimension increasing with N so one might fear that the overall numerical integration error increases.

However, it is possible to circumvent this problem by making use of a splitting technique which exploits the structure of the extended target, see (Beskos et al., 2011; Leimkuhler and Matthews, 2015, Section 2.4.1; Neal, 2011, Section 5.5.1; Shahbaba et al., 2014). The idea is to split the Hamiltonian H defined in (11) into two components $H = A + B$, where

$$A(\rho, \mathbf{u}, \mathbf{p}) := \frac{1}{2}\Big\{\rho^\top \rho + \mathbf{u}^\top \mathbf{u} + \mathbf{p}^\top \mathbf{p}\Big\}, \qquad B(\theta, \mathbf{u}) := -\log p(\theta) - \log \widehat{p}(y_{1:T} \,|\, \theta, \mathbf{u}). \tag{16}$$

Separately, the Hamiltonian systems for A and B can both be integrated analytically. Indeed, if we define the mapping $\Phi_h^A : \mathbb{R}^{2d+2D} \mapsto \mathbb{R}^{2d+2D}$ as the solution to the dynamical system with Hamiltonian A simulated for h units of time from a given initial condition, we have the explicit solution

$$\Phi_h^A : \begin{cases} \theta(h) = \theta(0) + h\rho(0), \\ \rho(h) = \rho(0), \\ \mathbf{u}(h) = \mathbf{p}(0)\sin(h) + \mathbf{u}(0)\cos(h), \\ \mathbf{p}(h) = \mathbf{p}(0)\cos(h) - \mathbf{u}(0)\sin(h). \end{cases} \tag{17}$$

Similarly for system B we define the mapping $\Phi_h^B : \mathbb{R}^{2d+2D} \mapsto \mathbb{R}^{2d+2D}$ and get the solution

$$\Phi_h^B : \begin{cases} \theta(h) = \theta(0), \\ \rho(h) = \rho(0) + h\nabla_\theta \{\log p(\theta) + \log \widehat{p}(y_{1:T} \mid \theta, \mathbf{u}(0))\}_{|\theta=\theta(0)}, \\ \mathbf{u}(h) = \mathbf{u}(0), \\ \mathbf{p}(h) = \mathbf{p}(0) + h\nabla_{\mathbf{u}} \log \widehat{p}(y_{1:T} \mid \theta(0), \mathbf{u})_{|\mathbf{u}=\mathbf{u}(0)}. \end{cases} \tag{18}$$

Let the integration time be given as hL, where h is the step-size and L the number of integration steps. To approximate the solution to the original system associated to the vector field $\widehat{\Psi}$ in (12), we then use a symmetric Strang splitting (see, e.g., Leimkuhler and Matthews 2015, p. 108) defined as $\widehat{\Phi}_{hL} = \{\Phi_{h/2}^A \circ \Phi_h^B \circ \Phi_{h/2}^A\}^{\circ L}$. In practice we combine consecutive half-steps of the integration of system A for numerical efficiency, similarly to what is often done for the standard Verlet integrator, i.e., we use

$$\widehat{\Phi}_{hL} = \Phi_{h/2}^A \circ \{\Phi_h^B \circ \Phi_h^A\}^{\circ L-1} \circ \Phi_h^B \circ \Phi_{h/2}^A. \tag{19}$$

In Appendix A the explicit update equations corresponding to a full step of $\widehat{\Phi}_h$ are given.

Using this integration method we can prove (see next section) both that the (θ, ρ)-trajectory of the PM-HMC algorithm converges to the ideal HMC trajectory, and that the acceptance probability of the PM-HMC algorithm converges to that of the ideal HMC algorithm as $N \to \infty$, when using the likelihood estimator defined in (13). This shows that, when using the aforementioned integration technique, the increase in dimension which happens with increased N does not pose any problem with respect to the convergence of the algorithm to its idealized counterpart.

An alternative method which also exploits the structure of the extended Hamiltonian and that could be used in our setting is the exponential integration technique of Chao et al. (2015). In simulations we found the two integrators to perform similarly and we focus on the splitting technique for simplicity.

The proposed PM-HMC method is summarized in Algorithm 1. As this algorithm simulates by design a Markov chain of invariant distribution $\bar{\pi}(\theta, \rho, \mathbf{u}, \mathbf{p})$ defined in (11), it samples asymptotically (in the number of iterations) from its marginal $\pi(\theta)$ under ergodicity conditions. We emphasize that the proposed PM-HMC algorithm is a valid MCMC method for any nonnegative and unbiased likelihood estimator used to define the extended target distribution (i.e., for any $N \geq 1$ when the likelihood estimator is based on importance sampling). Hence, under weak assumptions the generated Markov chain will converge to the correct target distribution $\pi(\theta)$. We formalize this in the following proposition.

Proposition 3 *Algorithm 1 defines a Markov kernel on $(\theta, \mathbf{u})$ with $\bar{\pi}$ as its stationary distribution. The θ-marginal of the stationary distribution is precisely $\pi(\theta)$. In particular, when the likelihood estimator is defined as in* (13) *this property holds for any $N \geq 1$.*

Proof The numerical integrator (19) is a symmetric Strang splitting for the Hamiltonian associated with the extended target distribution $\bar{\pi}(\theta, \rho, \mathbf{u}, \mathbf{p})$. It is therefore reversible and symplectic (Leimkuhler and Matthews, 2015, Section 2.4). With the Metropolis—Hastings

correction on line 3 this implies that the algorithm implements a Markov kernel with stationary distribution $\bar{\pi}(\theta, \mathbf{u})$. Under Assumption 1, $\int \bar{\pi}(\theta, \mathbf{u}) d\mathbf{u} = \pi(\theta)$. When using importance sampling to define the extended target, Assumption 1 holds for any $N \geq 1$. ■

Algorithm 1 Pseudo-marginal HMC (one iteration)

Let $(\theta, \mathbf{u})$ be the current state of the Markov chain. Do:

1. Sample auxiliary variables $\rho \sim \mathcal{N}(0_d, I_d)$ and $\mathbf{p} \sim \mathcal{N}(0_D, I_D)$.
2. Compute $(\theta', \rho', \mathbf{u}', \mathbf{p}') = \widehat{\Phi}_{hL}(\theta, \rho, \mathbf{u}, \mathbf{p})$ using the numerical integrator (19).
3. Accept $(\theta', \mathbf{u}')$ with probability $1 \wedge \exp(H(\theta, \rho, \mathbf{u}, \mathbf{p}) - H(\theta', \rho', \mathbf{u}', \mathbf{p}'))$.

3. Convergence of PM-HMC Towards the Ideal Marginal HMC

In this section we establish the convergence (under suitable assumptions) of the PM-HMC, in the sense that the performance of the algorithm will converge as N increases towards that of the ideal marginal HMC algorithm. We note once again that PM-HMC is a valid MCMC for the target distribution $\pi(\theta)$ for any $N \geq 1$, so we do not rely on a large N for the algorithm to be correct. However, as we show in this section, when N is large it will closely mimic the ideal marginal HMC. We begin by studying the numerical integrator $\widehat{\Phi}_{hL}$ and analyze the error in the first two components when comparing with the results from the ideal HMC algorithm.

For the rest of this section we will adopt the following notation: we use $(\hat{\theta}[\ell], \hat{\rho}[\ell], \hat{\mathbf{u}}[\ell], \hat{\mathbf{p}}[\ell])^\top$ to denote the results of running the PM-HMC, and $(\theta[\ell], \rho[\ell])^\top$ the results of running the ideal HMC algorithm (which is intractable), for ℓ iterations of the corresponding numerical integrator. For the $\mathbf{u}$ and $\mathbf{p}$ variables we use the notation $\hat{\mathbf{u}}[\ell]_{k,i}$ to denote position i in the vector associated with the observation y_k, which is consistent with the notation used in Section 2.3. Further we use the notation $\|\cdot\|$ for the Euclidean norm, $\xrightarrow{d}$ for convergence in distribution, and $\xrightarrow{p}$ for convergence in probability. We focus here on the setting where we use the latent variable model (2)—(3) and the importance sampling estimator (13)—(15). The proofs of these results are postponed to the appendix.

We also present the following assumption on the weight function that will be needed for the proofs.

- **Assumption 3**. The importance weight $\omega_\theta(y, \mathbf{u})$ defined in Section 2.3 satisfies:
  - $(\theta, \mathbf{u}) \to \nabla_{\mathbf{u}} \log \omega_\theta(y, \mathbf{u})$ is Lipschitz with constant M uniformly in y,
  - $(\theta, \mathbf{u}) \to \omega_\theta(y, \mathbf{u})$ is Lipschitz with constant D uniformly in y,
  - there exists constants $0 < \underline{\omega} < \overline{\omega} < \infty$ such that $\underline{\omega} < \omega_\theta(y, \mathbf{u}) < \overline{\omega}$,
  - $\|\nabla_{\mathbf{u}} \log \omega_\theta(y, \mathbf{u})\|$ is bounded from above by $C < \infty$.

This assumption is quite restrictive and does not hold for most practical problems. Nonetheless we have chosen to use it to keep the following theoretical analysis simple and

to the point. In the simulations below we look at models that violate this assumption and show that the algorithm still performs as expected. We believe that it is possible to relax these conditions but that is beyond the scope of this paper.

Proposition 4 *Let $(\theta[L], \rho[L])$ be the value associated with the ideal HMC dynamics and $(\hat{\theta}[L], \hat{\rho}[L])$ be the values associated to the (θ, ρ)-marginal of the PM-HMC dynamics after L steps of the numerical integrator using the Strang splitting with step-size h. Furthermore, let both of the processes start in the same point, i.e. $(\theta[0], \rho[0]) = (\hat{\theta}[0], \hat{\rho}[0])$.*

Assume that Assumption 3 holds and that $\nabla_\theta \log p(\theta \,|\, y)$ is Lipschitz with constant $L_0 < \infty$. Then there exists a constant $\mathcal{L} < \infty$, which does not depend on N or L, such that for any $L \geq 1$ and any choice of initial values $(\theta[0], \rho[0]) = (\hat{\theta}[0], \hat{\rho}[0])$ and $(\hat{\mathbf{u}}[0], \hat{\mathbf{p}}[0])$,

$$\left\| \begin{pmatrix} \hat{\theta}[L] \\ \hat{\rho}[L] \end{pmatrix} - \begin{pmatrix} \theta[L] \\ \rho[L] \end{pmatrix} \right\| \leq h^2 \sqrt{\frac{h^2}{4} + 1} \times \sum_{\ell=0}^{L-1} (1 + h\mathcal{L})^{L-(\ell+1)} \left\| \nabla_\theta \log \left(\frac{\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell] \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell] \cos(\frac{h}{2}))}{p(y \,|\, \theta)} \right)_{| \, \theta = \hat{\theta}[\ell] + \frac{h}{2}\hat{\rho}[\ell]} \right\|.$$

By taking the expected value of both sides, over a distribution of the initial auxiliary variables $\hat{\mathbf{u}}[0]$ and associated momentum $\hat{\mathbf{p}}[0]$, conditioned on the initial values of $(\theta[0], \rho[0]) = (\hat{\theta}[0], \hat{\rho}[0])$ and using Jensen's inequality we get as an immediate corollary that

$$\mathbb{E}\left[\left\| \begin{pmatrix} \hat{\theta}[L] \\ \hat{\rho}[L] \end{pmatrix} - \begin{pmatrix} \theta[L] \\ \rho[L] \end{pmatrix} \right\| \right] \leq h^2 \sqrt{\frac{h^2}{4} + 1} \times \sum_{\ell=0}^{L-1} (1 + h\mathcal{L})^{L-(\ell+1)} \mathbb{E}\left[\left\| \nabla_\theta \log \left(\frac{\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell] \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell] \cos(\frac{h}{2}))}{p(y \,|\, \theta)} \right)_{| \, \theta = \hat{\theta}[\ell] + \frac{h}{2}\hat{\rho}[\ell]} \right\|^2 \right]^{1/2}.$$

That is, the upper bound is directly related to the second moment of the error in the log-likelihood gradient $\nabla_\theta \log(\widehat{p}(y \,|\, \theta, \mathbf{u})/p(y \,|\, \theta))$. The next result establishes a CLT for this error as N grows, in two scenarios. First we study the behavior at stationarity, that is when $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \theta)$. Second we show that a similar CLT holds when $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$, that is at initialization of the algorithm.

In the following we will assume that θ is scalar for notational simplicity. In the multivariate case the results should be interpreted to hold component-wise. We will by $\mathbb{E}_\mathcal{N}$ denote the expected value under the standard normal distribution of appropriate dimension.

Proposition 5 *Suppose that Assumption 3 holds. Let $\varpi_\theta(y, \mathbf{u}) := \omega_\theta(y, \mathbf{u})/p(y \,|\, \theta)$ and assume that $\mathbf{u} \to \nabla_\theta \varpi_\theta(y, \mathbf{u})$ is continuous and that $\|\nabla_\theta \varpi_\theta(y, \mathbf{u})\|$ is bounded from above by a constant $E < \infty$. Further assume that $\mathbb{E}_\mathcal{N}[\varpi_\theta^2(y_k, \mathbf{u})] < \infty$ and $\mathbb{E}_\mathcal{N}[\{\nabla_\theta \varpi(y_k, \mathbf{u})\}^2] < \infty$, there exists a function $g_k(\mathbf{u})$ which may depend on θ and y_k such that $|\nabla_\theta \varpi(y_k, \mathbf{u})| < g_k(\mathbf{u})$ and $\mathbb{E}_\mathcal{N}[g_k(\mathbf{u})] < \infty$ and $\mathbb{E}_\mathcal{N}[\varpi_\theta(y_k, \mathbf{u}) |\nabla_\theta \log \varpi_\theta(y_k, \mathbf{u})|] < \infty$ for all $k = 1, \ldots, T$.*

Then, for any $\ell \geq 1$, the following CLT holds when $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \theta)$ and $\hat{\mathbf{p}}[0] \sim \mathcal{N}(0_D, I_D)$:

$$\sqrt{N} \nabla_\theta \log \frac{\widehat{p}(y_{1:T} \,|\, \theta, \hat{\mathbf{p}}[\ell] \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell] \cos(\frac{h}{2}))}{p(y \,|\, \theta)} \xrightarrow{d} \mathcal{N}(0, \sigma^2(y_{1:T}, \theta)), \quad \text{as } N \to \infty,$$

where the variance is given by

$$\sigma^2(y_{1:T}, \theta) = \sum_{k=1}^{T} \mathbb{E}_{\mathcal{N}}[\{\nabla_\theta \varpi_\theta(y_k, \mathbf{v}_{k,1})\}^2].$$

The same CLT also holds when $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$.

We now look at the acceptance probability of PM-HMC. As we have shown above, the trajectory of the (θ, ρ)-marginal of the extended space used in PM-HMC converges towards the trajectory of the ideal HMC algorithm as N increases. Thus we also expect that the acceptance probability of the PM-HMC algorithm converges towards the acceptance probability of the HMC algorithm at equilibrium. This is established in the following proposition.

Proposition 6 *Let Assumption 3 hold. For any* $\hat{\theta}[0]$ *assume that* $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \hat{\theta}[0])$, $\hat{\rho}[0] \sim \mathcal{N}(0_d, I_d)$, *and* $\hat{\mathbf{p}}[0] \sim \mathcal{N}(0_D, I_D)$, *we then have that*

$$\begin{aligned} H(\hat{\theta}[0], \hat{\rho}[0], \hat{\mathbf{u}}[0], \hat{\mathbf{p}}[0]) - H(\hat{\theta}[L], \hat{\rho}[L], \hat{\mathbf{u}}[L], \hat{\mathbf{p}}[L]) \\ \xrightarrow{p} H_{\mathtt{ex}}(\hat{\theta}[0], \hat{\rho}[0]) - H_{\mathtt{ex}}(\hat{\theta}[L], \hat{\rho}[L]), \quad \text{as } N \to \infty. \end{aligned}$$

Here $H_{\mathtt{ex}}$ *is the Hamiltonian associated with the ideal marginal algorithm given in* (5).

As mentioned in Section 2.4, we again note that as N increases the dimension of the ordinary differential equations one needs to approximate numerically increases. At a glance this might seem problematic, since the integration error typically increases with the dimension, but in this section we have established that the increase of dimension with N does not pose a problem to our algorithm. In fact, for any step-size h and number of integration steps L the (θ, ρ)-marginal of the PM-HMC algorithm will converge towards the values of the ideal HMC algorithm and the acceptance probability (directly related to the difference in the Hamiltonian values) converges to the acceptance probability of the ideal HMC algorithm.

4. Numerical Illustrations

We illustrate the proposed PM-HMC method on three synthetic examples and one real-world data set. Additional details and results are given in the appendix.

For simplicity, we focus on the case when standard importance sampling-based estimators of the likelihood are used to construct the extended Hamiltonian system. However, more efficient estimators can be used and they will intuitively improve the performance of PM-HMC. One illustration of this is given by Osmundsen et al. (2018), who consider the use of *efficient importance sampling* in the context of PM-HMC for inference in dynamical systems. We refer to this article for additional numerical illustrations of PM-HMC.

4.1 Gaussian Model

We consider first the following Gaussian model where $X_k \,|\, \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma_X^2)$ and $Y_k \,|\, (X_k = x_k), \theta \sim \mathcal{N}(x_k, \sigma_Y^2)$ and assign a Gaussian prior on the parameter, $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$. We choose

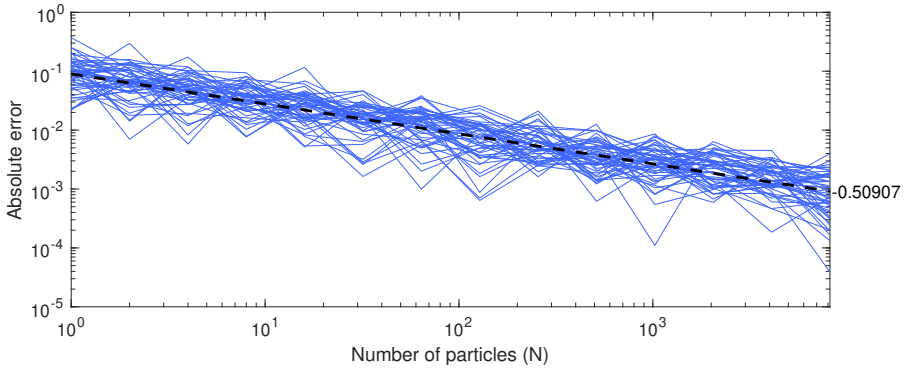

Figure 1: Absolute error $\max_{\ell \in [0,10]} |\theta[\ell] - \hat{\theta}[\ell]|$ for different initial conditions. The slope of the linear fit, shown as a dashed line, is -0.509.

this model as a first illustration since the true posterior will be a Gaussian. This allows us to run the ideal marginal HMC algorithm targeting $p(\theta \,|\, y_{1:T})$ and we can then compare our PM-HMC algorithm with the ideal HMC algorithm. We simulate $T = 30$ i.i.d. observations using $\mu_\theta = 0, \sigma_\theta^2 = 10, \sigma_x^2 = 0.1$, and $\sigma_y^2 = 1$. We apply the PM-HMC algorithm with $N = 2^i$ for i from 0 to 13. First we check the convergence of the trajectories from the numerical integrator $\hat{\Phi}_{hL}$ towards the trajectories for the ideal HMC algorithm. We do this by using $h = 0.1$ and $L = 10$ and look at the maximal position error over the integration period. We run the algorithm 50 times using different starting values and look at how the maximal error depends on the choice of N. The results, displayed in Figure 1, imply a $\sqrt{N}$ convergence rate of the maximal error.

Next we look at the convergence of the acceptance probability as a function of N. First, we run the ideal marginal HMC algorithm for 150 iterations (initialized from a draw from the true posterior) using $h = 0.35$ and $L = 20$. We record the (θ, ρ)-states as well as the average acceptance probability. Then, for each (θ, ρ)-state we perform 500 independent runs of the PM-HMC algorithm using the same values of $\hat{\theta}[0]$ and $\hat{\rho}[0]$, while $\hat{\mathbf{u}}[0]$ and $\hat{\mathbf{p}}[0]$ are randomized for each run. This is done so that we can compare with the acceptance probability given by the ideal HMC algorithm for the same initial values. The results can be seen in Figure 2 where we present the median as well as the upper and lower quartiles across the 500 independent runs. We can clearly see that the acceptance probability converges towards the value of the acceptance probability of the ideal marginal HMC as N increases. For comparison, we perform the same experiment but replace the Strang splitting integrator proposed in Section 2.4 with the standard Verlet integrator. This integrator performs similarly for small values of N, but for $N > 10$ we can see that it deviates from the acceptance probability obtained using Strang splitting, and for large values of N its acceptance probability drops to zero.

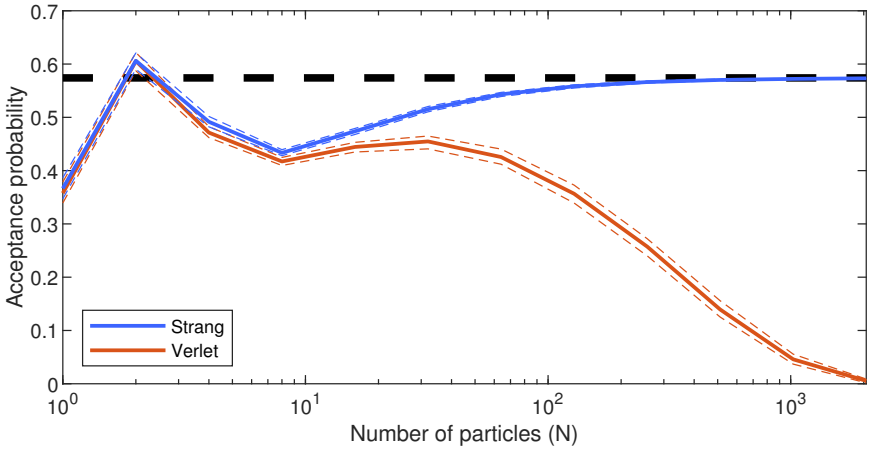

Figure 2: Acceptance probability of the PM-HMC algorithm as a function of N in the Gaussian example, averaged over 150 MCMC iterations. The solid line is the median and the dashed lines the lower and upper quartiles over 500 independent runs. The blue lines correspond to using the Strang splitting integrator from Section 2.4 and the brown lines to the standard Verlet integrator. The thick black dashed line is the average acceptance probability of the ideal marginal HMC algorithm.

4.2 Diffraction Model

Consider a hierarchical model of the form $X_k \,|\, \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $Y_k \,|\, (X_k = x_k), \theta \sim g(\cdot \,|\, x_k, \lambda)$, with $\theta = (\mu, \log(\sigma), \log(\lambda))$, where the observation density is modelled as a diffraction intensity: $g(y \,|\, x, \lambda) = (\lambda\pi)^{-1} \operatorname{sinc}^2(\lambda^{-1}(y - x))$. We simulate $T = 100$ i.i.d. observations using $\mu = 1$, $\sigma = 1$, and $\lambda = 0.1$. We apply the PM-HMC with N ranging from 16 to 256, using a non-centered parameterization and the prior as importance density, as well as a standard HMC working on the joint space using the same non-centered parameterization.[2] For further comparison we also ran, *(i)* a standard pseudo-marginal MH algorithm using $N = 512$ and $N = 1024$ (smaller N resulted in very sticky behavior), *(ii)* the pseudo-marginal slice sampler by Murray and Graham (2016) with N ranging from 16 to 256, and *(iii)* a Gibbs sampler akin to the Particle Gibbs algorithm by Andrieu et al. (2010) (using independent conditional importance sampling kernels to update the latent variables).

Figure 3 shows scatter plots from 40 000 iterations (after a burn-in of 10 000) for the parameters (σ, λ) for standard HMC, PM-HMC ($N = 16$), and pseudo-marginal slice sampling ($N = 128$, smaller values gave very poor mixing). Results for the remaining methods/settings, including the pseudo-marginal MH method and the Particle Gibbs sampler which both performed very poorly, are given in the appendix. For the pseudo-marginal slice sampling we ran a SS+MH algorithm, where we use a random walk for the θ variables and elliptical slice sampling for the $\mathbf{u}$-variables. It is clear from the results that standard HMC fails and generated chains which do not mix. This is not surprising, since the diffraction intensity

2. The standard HMC, here, is equivalent to PM-HMC with $N = 1$. Specifically, it operates on the joint, non-centered $(\theta, \mathbf{u})$ space, and it uses the splitting integrator described in Section 2.4.

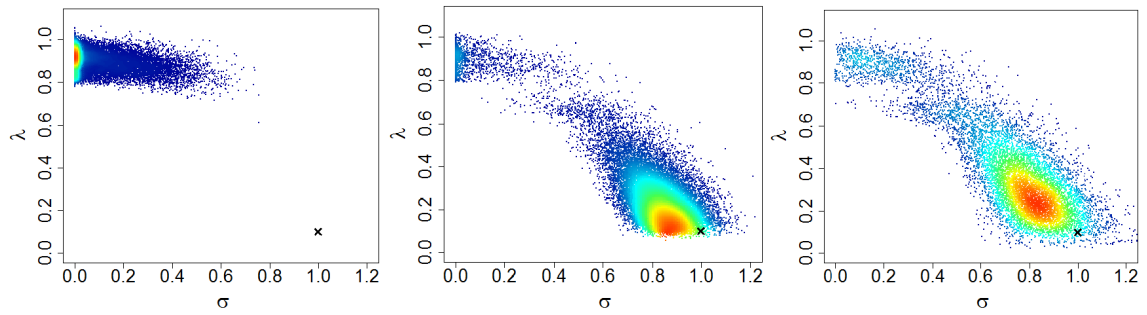

Figure 3: Scatter plots for (σ, λ) for standard HMC using a non-centered parameterisation (left), PM-HMC with $N = 16$ (mid), and pseudo-marginal slice sampling with $N = 128$ (right). The black $\times$ corresponds to the variables used to generate the data.

is exactly zero whenever $\lambda^{-1}(y - x) \bmod \pi = 0$. Consequently, there are infinite potential energy barriers between the lobes of the sinc function. An ideal HMC algorithm operating on the joint space, which exactly simulates the Hamiltonian dynamics, would not be able to pass these barriers. Even though the discretization can result in occasional jumps between modes, it is clear from our results that the discretized HMC algorithm also fails to explore the full posterior. The extended target of the PM-HMC algorithm, on the other hand, circumvents this problem by the introduction of latent variables that smooth out the hard multimodalities of the target distribution. Indeed, when using an importance sampling estimator of the likelihood, the extended target (8) is such that, marginally, the distribution of each $U_{k,i}$ is a mixture between the original posterior and the Gaussian prior (see Remark 9 in the appendix for details). This can also be seen in the expression for the likelihood estimator (13). Due to the summation over the importance weights for $N > 1$, the estimator will be non-zero unless *all* of the auxiliary variables $U_{k,1}, \ldots, U_{k,N}$ result in values $X_{k,i}$ for which the sinc function is zero. While there are still points in the joint space where the extended target distribution is exactly zero, it avoids the hard partitioning of the target space into disjoint regions as is the case for standard HMC. While this type of "hard" multimodality is perhaps not very common in practice, "softer" variants where the observation density terms are strictly positive but close to zero for some inputs are very plausible, and would lead to similar problems for standard HMC.

Pseudo-marginal slice sampling also does well in this example and we found the two methods to be comparable in performance when normalized by computational cost.[3]

4.3 Generalized Linear Mixed Model

Next, we consider inference in a generalized linear mixed model (GLMM), see e.g. Zhao et al. (2006), with a logistic link function: $Y_{ij} \sim \text{Bernoulli}(p_{ij})$, where $\text{logit}(p_{ij}) = X_i + Z_{ij}^{\mathsf{T}}\beta$

3. We do not report effective sample size estimates, as we found these to be very noisy and misleading for this multimodal distribution. Indeed, the "best" effective sample size when normalized by computational cost was obtained for the Particle Gibbs sampler which, by visual inspection, completely failed to converge to the correct posterior.

for $i = 1{:}T$, $j = 1{:}n_i$. Here, Y_{ij} represent the jth observation for the ith "subject", Z_{ij} is a covariate of dimension p, β is a vector of fixed effects and X_i is a random effect for subject i. It has been recognized (Burda et al., 2008; Komárek and Lesaffre, 2008) that it is often beneficial to allow for non-Gaussianity in the random effects. For instance, multimodality can arise as an effect of under-modelling, when effects not accounted for by the covariates result in a clustering of the subjects. To accommodate this we assume X_i to be distributed according to a Gaussian mixture: $X_i \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{K} w_j \mathcal{N}(\mu_j, \lambda_j^{-1})$. For simplicity we fix $K = 2$ for this illustration. The parameters of the model are thus β, $\{\mu_j, \lambda_j\}_{j=1}^2$, and w_1 (as $w_2 = 1 - w_1$), with $X_{1:T}$ being latent variables. We use the parameterisation

$$\theta = (\beta^\mathsf{T}, \mu_1, \mu_2, \log(\lambda_1), \log(\lambda_2), \operatorname{logit}(w_1))^\mathsf{T} \in \mathbb{R}^{p+5}.$$

We used a simulated data set with $T = 500$ and $n_i = 6$, thus a total of 3 000 data points, with $\mu_1 = 0$, $\mu_2 = 3$, $\lambda_1 = 10$, $\lambda_2 = 3$. We set $p = 8$ and generate β as well the covariates Z_{ij} from standard normal distributions (see the appendix for further details on the simulation setup).

We ran PM-HMC, pseudo-marginal slice sampling (Murray and Graham, 2016), and Particle Gibbs (Andrieu et al., 2010) for 7 000 iterations, all using $N = 128$. We note that Gibbs sampling type algorithms (of which our Particle Gibbs sampler is an example) are the *de facto* standard methods for Bayesian GLMMs. For the pseudo-marginal slice sampling we used a SS+MH algorithm where for θ we have one MH step for each component and for $\mathbf{u}$ we used elliptical slice sampling. Figure 4 shows traces for the parameters μ_1 and μ_2, with additional results reported in the appendix. The PM-HMC method clearly outperforms the competing methods in terms of mixing (the computational cost per iteration is about 3.5 times higher for PM-HMC than for pseudo-marginal slice sampling in our implementation). In particular, compared to the results of Section 4.2, we note that PM-HMC handles this more challenging model with a 13-dimensional θ much better than the pseudo-marginal slice sampling which appears here to struggle for high-dimensional θ.

However, we also note that the PM-HMC algorithm gets stuck for many iterations around iteration 1 000 (see Figure 4). Experiments suggest that this stickiness is an issue inherited from the marginal HMC (which PM-HMC mimics) and is not specific to the PM-HMC algorithm. Specifically, we have experienced that the PM-HMC sampler tends to get stuck at large values of λ_1 or λ_2, i.e., in the right tail of the posterior for one of these parameters. A potential solution to address this issue is to use a state-dependent mass matrix as in the Riemannian manifold HMC (Girolami and Calderhead, 2011).

4.4 Respiratory Infection Data

As a final example we consider a version of the generalized linear mixed model from Section 4.3 applied to a real-world data set. This data set is a subset of a cohort study of 275 Indonesian preschool children with 1200 observations. It has previously been studied by Zeger and Karim (1991) using Bayesian mixed models and Schmon et al. (2021) with the model used here. The probability of a respiratory infection is modeled based on 7 covariates: age, sex, height, an indicator for vitamin deficiency, an indicator for subnormal height and two seasonal components. A linear regression model based on the covariates $Z_{i,j}$ is $\eta_{i,j} = Z_{i,j}^\mathsf{T}\beta + X_i$, where $X_i \sim \mathcal{N}(0, \tau)$ denotes the intercept for child $i = 1, \ldots, T$

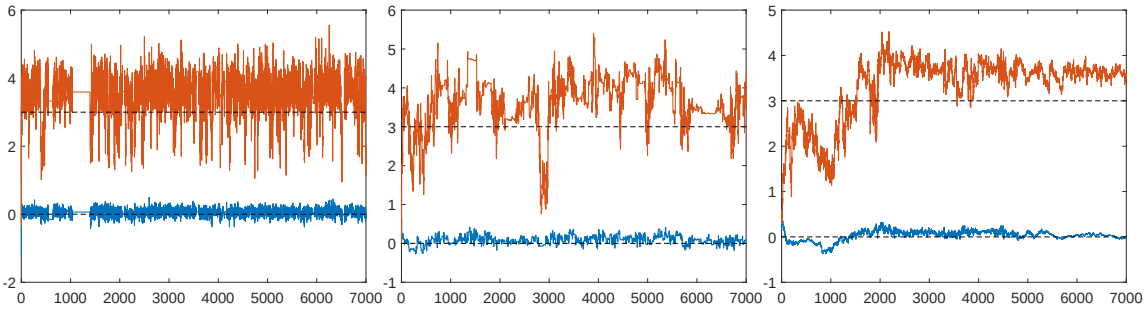

Figure 4: Traces for parameters μ_1 (blue) and μ_2 (orange) for PM-HMC (left), pseudo-marginal slice sampling (mid), and Particle Gibbs sampling (right), all with $N = 128$.

and β the regression parameters. We have J observations for each child given by the vector $Y_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,J}) \in \{0,1\}^J$. The parameters to infer are $\theta = (\beta, \tau) \in \mathbb{R}^9$. Given the random effects, we assume that the observations are conditionally independent given the random effects and satisfy for $i = 1, \ldots, T$,

$$g_\theta(Y_i \mid X_i) = \prod_{j=1}^{J} \frac{\exp(Y_{i,j}\eta_{i,j})}{1+\exp(\eta_{i,j})}, \quad f_\theta(X_i) = \mathcal{N}(X_i; 0, \tau).$$

For the β parameters we set a Gaussian prior with mean 0 and variance 10 000, for τ we use an inverse gamma prior with shape 1 and scale 1.5. The resulting posterior distribution is of dimension $9 + 275 = 284$.

The aim of inference in mixed effects models is to find the population effects θ while integrating out the random effects. As the corresponding marginal likelihood is intractable, we use a pseudo-marginal approach for this problem with $q_\theta = \mathcal{N}(0, 3^2)$ as importance sampling distribution. We ran PM-HMC and three different versions of the pseudo-marginal slice sampling (Murray and Graham, 2016) for 50 000 iterations with a burn-in of 5 000 using different values of $N = \{1, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$.

For the PM-HMC algorithm we use $L = 50$ leapfrog steps with a stepsize $h = 0.01$. For the pseudo-marginal slice sampling algorithm we used elliptical slice sampling (SS+SS), Hamiltonian Monte Carlo (SS+HMC), and Metropolis—Hastings (SS+MH) for sampling the parameter slice. For the SS+HMC algorithm we use 50 leapfrog steps with stepsize 0.01. For the SS+MH algorithm a Gaussian random walk with variance 0.001 is chosen.

In Table 1 we compare the values of the average integrated autocorrelation time (IAT), effective sample size per second (ESS/s), and average acceptance probability ($\mathbb{P}_\mathrm{a}$). As can be seen from these results, the PM-HMC algorithm has much lower IAT than the other algorithms while maintaining a better ESS/s; i.e. we are able to generate more useful samples from the posterior distribution using the PM-HMC algorithm compared the other pseudo-marginal approaches using the same computational budget.

	PM-HMC			SS+HMC			SS+SS			SS+MH		
N	IAT	ESS/s	$\mathbb{P}_a$	IAT	ESS/s	$\mathbb{P}_a$	IAT	ESS/s	$\mathbb{P}_a$	IAT	ESS/s	$\mathbb{P}_a$
1	13.7	0.83	0.67	53.6	0.25	0.69	151	0.48	1	168	0.74	0.13
3	8.96	1.20	0.72	39.0	0.27	0.72	146	0.40	1	162	0.64	0.15
6	6.37	1.40	0.76	30.3	0.29	0.75	139	0.34	1	159	0.55	0.16
9	5.28	1.43	0.77	24.2	0.28	0.77	138	0.29	1	162	0.46	0.16
12	5.08	1.15	0.77	20.9	0.27	0.77	139	0.24	1	165	0.37	0.15
15	5.02	1.07	0.77	16.4	0.31	0.78	140	0.21	1	160	0.35	0.16
18	5.00	0.96	0.77	15.6	0.30	0.78	137	0.20	1	163	0.31	0.17
21	3.82	1.15	0.78	16.3	0.26	0.77	136	0.19	1	161	0.31	0.16
24	3.71	1.10	0.78	12.5	0.32	0.78	137	0.18	1	159	0.29	0.16
27	4.14	0.93	0.78	16.5	0.22	0.77	138	0.17	1	155	0.30	0.17
30	3.33	1.02	0.79	8.97	0.32	0.78	137	0.15	1	159	0.26	0.17

Table 1: Random effects model: average integrated autocorrelation time (IAT), effective sample size per second (ESS/s), and average acceptance probability ($\mathbb{P}_a$). These results are presented for different values of N for the PM-HMC algorithm and three different version of pseudo-marginal slice sampling.

5. Discussion

HMC methods cannot be implemented in scenarios where the likelihood function is intractable. However, we have shown that if we have access to a non-negative unbiased likelihood estimator parameterized by normal random variables, then it is possible to derive an algorithm which mimics the HMC method having access to the exact likelihood. The resulting PM-HMC algorithm replaces the original intractable gradient of the log-likelihood by the gradient of the log-likelihood estimator. Nevertheless, it preserves the target distribution as invariant distribution. Empirically we have observed that this algorithm can work significantly better than the pseudo-marginal MH algorithm as well as the standard HMC method sampling on the joint space for a fixed computational budget.

However, whereas clear guidelines are available for optimizing the performance of the pseudo-marginal MH algorithm (Doucet et al., 2015; Sherlock et al., 2015; Schmon et al., 2021), it is unclear how to address this problem for PM-HMC. When the likelihood estimator of PM-HMC is constructed using importance sampling with N samples, the computational cost per iteration scales linearly with N. Hence, using $N > 1$ must be motivated by a sufficiently large improvement in the mixing speed of the resulting Markov chain. As we increase the number of samples N, the PM-HMC algorithm can be seen as moving from an HMC sampling on the joint space (using a non-centered parameterization) to an HMC sampling on the marginal space. Thus, even the case $N = 1$ results in a method which, in some cases, works well even for large T—we do not experience the same "break down" of the method as for pseudo-marginal MH when using a too small N relative to T.

What we have observed in practice, however, is that the extended target distribution (with $N \gg 1$) can be much easier for the Hamiltonian trajectories to explore in an efficient way, compared to the target distribution on the joint space (that is, with $N = 1$). Specifically, in situations when there is multimodality in the posterior distribution over the T latent

variables,[4] standard HMC on the joint space of parameters and latent variables can easily get stuck in local modes. In such situations, the extended target distribution used by PM-HMC is a way to bridge between the modes, as illustrated empirically by Nemeth et al. (2019). Indeed, as our theoretical analysis shows, the performance of PM-HMC will converge to that of an ideal marginal HMC, for which the latent variables have been marginalized out, as N becomes large.

This explains the large improvement in performance of PM-HMC compared to standard HMC in the examples of Sections 4.2 and 4.3, which have multimodal posteriors. However, for the example studied in Section 4.4, which *does not* have a multimodal posterior, we also observed an improvement in using $N > 1$. We believe that this is largely due to the fact that, in practice, the computational time does not grow linearly with N, due to a computational overhead that can be amortized over multiple samples when using $N > 1$. For instance, in our implementation, going from $N = 1$ to $N = 30$ only increased the computational time by a factor 3.5 (see Table 1). Furthermore, modern computing architectures allow for a high degree of parallelization over multiple samples, which could further improve the results of PM-HMC with $N > 1$ (for instance, we did not make use of GPU acceleration in our experiments). Nevertheless, in practice the number of samples N is a tuning parameter that needs to be chosen based on the geometry of the target posterior and traded off with the computational cost of the method.

Finally, for brevity of presentation, we have restricted ourselves here to the pseudo-marginal approximation of a standard HMC algorithm using a constant mass matrix and a Verlet scheme. However, we believe that the same ideas can be extended to more sophisticated HMC schemes such as the Riemannian manifold HMC (Girolami and Calderhead, 2011) or schemes discretizing the associated Nosé—Hoover dynamics, as described in the appendix.

Acknowledgments

We thank Anthony Caterini for his invaluable help with implementing the random effects model example. Arnaud Doucet's research is supported by the UK Engineering and Physical Sciences Research Council, grants EP/R034710/1 and EP/R013616/1. Fredrik Lindsten's research is supported by the Swedish Research Council via the projects *Learning of Large-Scale Probabilistic Dynamical Models* (contract number: 2016-04278) and *Handling Uncertainty in Machine Learning Systems* (contract number: 2020-04122), by the Swedish Foundation for Strategic Research via the project *Probabilistic Modeling and Inference for Machine Learning* (contract number: ICA16-0015) by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and by ELLIIT.

4. Even for bimodal marginals we could very well have 2^T modes in the joint posterior.

Appendix A. The One Step Integrator

We define $\hat{\Theta}[\ell]$ as the full parameter vector associated with the PM-HMC algorithm after ℓ steps of the integrator. We let $\hat{\Theta}[0]$ be the initial values of the full parameter vector.

Taking one step of the integrator $\hat{\Phi}_h = \Phi^A_{h/2} \circ \Phi^B_h \circ \Phi^A_{h/2}$ (where Φ^A_h and Φ^B_h is given in Equation 17 and 18) gives us the updating scheme from $\hat{\Theta}[\ell] = (\hat{\theta}[\ell], \hat{\rho}[\ell], \hat{\mathbf{u}}[\ell], \hat{\mathbf{p}}[\ell])$ to $\hat{\Theta}[\ell+1]$ through the following equations,

$$\hat{\theta}[\ell+1] = \hat{\theta}[\ell] + h\hat{\rho}[\ell] + \tfrac{h^2}{2}\nabla_\theta \Big\{ \log p(\theta) + \log \widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell]\sin(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]\cos(\tfrac{h}{2})) \Big\}_{\,|\,\theta=\hat{\theta}[\ell]+\frac{h}{2}\hat{\rho}[\ell]},$$

$$\hat{\rho}[\ell+1] = \hat{\rho}[\ell] + h\nabla_\theta \Big\{ \log p(\theta) + \log \widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell]\sin(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]\cos(\tfrac{h}{2})) \Big\}_{\,|\,\theta=\hat{\theta}[\ell]+\frac{h}{2}\hat{\rho}[\ell]},$$

$$\hat{\mathbf{u}}[\ell+1] = \hat{\mathbf{p}}[\ell]\sin(h) + \hat{\mathbf{u}}[\ell]\cos(h) + \sin(\tfrac{h}{2})h\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \hat{\theta}[\ell] + \tfrac{h}{2}\hat{\rho}[\ell], \mathbf{u})_{\,|\,\mathbf{u}=\hat{\mathbf{p}}[\ell]\sin(\frac{h}{2})+\hat{\mathbf{u}}[\ell]\cos(\frac{h}{2})},$$

$$\hat{\mathbf{p}}[\ell+1] = \hat{\mathbf{p}}[\ell]\cos(h) - \hat{\mathbf{u}}[\ell]\sin(h) + \cos(\tfrac{h}{2})h\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \hat{\theta}[\ell] + \tfrac{h}{2}\hat{\rho}[\ell], \mathbf{u})_{\,|\,\mathbf{u}=\hat{\mathbf{p}}[\ell]\sin(\frac{h}{2})+\hat{\mathbf{u}}[\ell]\cos(\frac{h}{2})}.$$

The steps in between are omitted but can easily be checked. Some use of trigonometric identities has to be used to reach the final expressions.

Appendix B. Convergence of Simulated Trajectories

Lemma 7 *Let $f, \hat{f} : \mathbb{R}^n \to \mathbb{R}^n$ be continuous functions and let $x, \hat{x} : \mathbb{N} \to \mathbb{R}^N$ be the solution to the following difference equation:*

$$x[i+1] - x[i] = h \cdot f(x[i]), \quad \hat{x}[i+1] - \hat{x}[i] = h \cdot \hat{f}(\hat{x}[i]), \tag{20}$$

both initialized using $x[0] = \hat{x}[0]$. If f is Lipschitz with constant $\mathcal{L}$ then, for any $\ell \in \mathbb{N}$

$$\|\hat{x}[\ell] - x[\ell]\| \leq \sum_{i=0}^{\ell-1} h(1+h\mathcal{L})^{\ell-(i+1)} \|\hat{f}(\hat{x}[i]) - f(\hat{x}[i])\| \tag{21}$$

Proof Using (20) the proof is straightforward,

$$\begin{aligned}
\|\hat{x}[i+1] - x[i+1]\| &= \|\hat{x}[i] + h \cdot \hat{f}(\hat{x}[i]) - x[i] - h \cdot f(x[i])\| \\
&\leq \|\hat{x}[i] - x[i]\| + h\|\hat{f}(\hat{x}[i]) - f(x[i])\| \\
&= \|\hat{x}[i] - x[i]\| + h\|\hat{f}(\hat{x}[i]) - f(\hat{x}[i]) + f(\hat{x}[i]) - f(x[i])\| \\
&\leq \|\hat{x}[i] - x[i]\| + h\|f(\hat{x}[i]) - f(x[i])\| + h\|\hat{f}(\hat{x}[i]) - f(\hat{x}[i])\| \\
&\leq \|\hat{x}[i] - x[i]\| + h\mathcal{L}\|\hat{x}[i] - x[i]\| + h\|\hat{f}(\hat{x}[i]) - f(\hat{x}[i])\| \\
&= (1+h\mathcal{L})\|\hat{x}[i] - x[i]\| + h\|\hat{f}(\hat{x}[i]) - f(\hat{x}[i])\|.
\end{aligned}$$

Equation (21) then follows by repeating the recursion. ■

For the proof of Proposition 4 we have to compare the result of the ideal HMC algorithm with our proposed PM-HMC algorithm. Since they live on different spaces we augment

the ideal HMC algorithm to incorporate the $\mathbf{u}$ and $\mathbf{p}$ part in such a way that the (θ, ρ)-marginal exactly follows the ideal HMC algorithm. We do this by introducing the following time-stepper that will replace Φ_t^B:

$$\tilde{\Phi}_t^B : \begin{cases} \theta(t) = \theta(0), \\ \rho(t) = \rho(0) + t\nabla_\theta \{\log\, p(\theta) + \log\, p(y \,|\, \theta)\}_{|\theta=\theta(0)}, \\ \mathbf{u}(t) = \mathbf{u}(0), \\ \mathbf{p}(t) = \mathbf{p}(0) + t\nabla_{\mathbf{u}} \log\, \widehat{p}(y \,|\, \theta(0), \mathbf{u})_{|\mathbf{u}=\mathbf{u}(0)}. \end{cases}$$

This differs from Φ^B by using the gradient from the exact posterior when updating $\rho(t)$. Further we introduce $\tilde{\Phi}_h = \Phi_{h/2}^A \circ \tilde{\Phi}_h^B \circ \Phi_{h/2}^A$. Using this splitting operator as numerical integrator we have that the marginal (θ, ρ) will coincide exactly with the ideal HMC algorithm.

Lemma 8 *Let Assumption 3 hold. Also assume that $\nabla_\theta \log p(\theta \,|\, y)$ is Lipschitz with constant L_0, then it follows that $\tilde{\Phi}_h$ is Lipschitz with constant $\mathcal{L} < \infty$ which does not depend on N.*

The proof of Lemma 8 is postponed to Appendix D. Now we have all of the results needed to prove the bound on the difference between the output of the PM-HMC and the HMC algorithm.

Proof [Proof of Proposition 4] Under the assumptions, Lemma 8 holds and thus $\tilde{\Phi}_h$ is Lipschitz with constant $\mathcal{L} < \infty$.

As the space of the ideal HMC algorithm and the pseudo marginal HMC algorithm differ we cannot directly compare them. Thus we augment the space of the ideal HMC algorithm to reach the timestepper $\tilde{\Phi}_h = \Phi_h^A \circ \tilde{\Phi}_h^B \circ \Phi_h^A$ by adding the $\mathbf{u}$ and $\mathbf{p}$ part of the PM-HMC algorithm to the ideal HMC algorithm. Notice that this is no longer a discretization of a Hamiltonian field but it leaves (θ, ρ) unchanged from the ideal HMC algorithm. Let $\hat{\Theta}[\ell] = (\hat{\theta}[\ell], \hat{\rho}[\ell], \hat{\mathbf{u}}[\ell], \hat{\mathbf{p}}[\ell]) = \hat{\Phi}_{h\ell}(\hat{\Theta}[0])$ as above. Let $\Theta[\ell] = (\theta[\ell], \rho[\ell], \mathbf{u}[\ell], \mathbf{p}[\ell]) = \tilde{\Phi}_{h\ell}(\hat{\Theta}[0])$. We have that

$$\begin{aligned} &\|\hat{\Phi}_h(\hat{\Theta}[\ell]) - \tilde{\Phi}_h(\hat{\Theta}[\ell])\| \\ &\quad = \sqrt{\frac{h^4}{4} + h^2}\|\nabla_\theta \log\left(\frac{\widehat{p}(y_{1:T} \,|\, \theta, \hat{\mathbf{p}}[\ell]\sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]\cos(\frac{h}{2}))}{p(y_{1:T} \,|\, \theta)}\right)\|_{\,|\,\theta=\hat{\theta}[\ell]+\frac{h}{2}\hat{\rho}[\ell]}, \end{aligned}$$

by the assumption that $\tilde{\Phi}_h$ is Lipschitz with constant $\mathcal{L}$ we have using Lemma 7 that

$$\begin{aligned} \|\begin{pmatrix} \hat{\theta}[\ell] \\ \hat{\rho}[\ell] \end{pmatrix} - \begin{pmatrix} \theta[\ell] \\ \rho[\ell] \end{pmatrix}\| &\leq \|\hat{\Theta}[\ell] - \Theta[\ell]\| \\ &\leq \sum_{i=0}^{\ell-1} h(1 + h\mathcal{L})^{\ell-(i+1)} \sqrt{\frac{h^4}{4} + h^2} \\ &\qquad \times \|\nabla_\theta \log\left(\frac{\widehat{p}(y_{1:T} \,|\, \theta, \hat{\mathbf{p}}[i]\sin(\frac{h}{2}) + \hat{\mathbf{u}}[i]\cos(\frac{h}{2}))}{p(y_{1:T} \,|\, \theta)}\right)\|_{\,|\,\theta=\hat{\theta}[\ell]+\frac{h}{2}\hat{\rho}[\ell]}. \end{aligned}$$

■

Appendix C. Proof of CLT

Remark 9 *For the proof of the CLT and the coming proof of the convergence of the acceptance probability the following "trick" will be used. Assume that* $\mathbf{u}$ *is drawn from the distribution* $\bar{\pi}(\cdot \,|\, \theta)$*. We use the fact that we can write this distribution as*

$$\bar{\pi}(\mathbf{u} \,|\, \theta) = \frac{1}{N} \sum_{j=1}^{N} \psi(\mathbf{u}_j \,|\, \theta) \prod_{i \neq j} \mathcal{N}(\mathbf{u}_i; 0_p, I_p),$$

where $\psi(\cdot \,|\, \theta) := \mathcal{N}(\cdot\,; 0_p, I_p) \times \varpi_\theta(y, \cdot)$*. That is* $\psi(\cdot \,|\, \theta)$ *is a mixture distribution where we choose one component* j *uniformly at random and sample that variable* $\mathbf{u}_j$ *from* $\psi(\cdot \,|\, \theta)$ *and the rest of the variables* $\mathbf{u}_i : i \neq j$ *from standard Gaussian distributions.*

What can now be done is to introduce new variables $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_N]$ *and* $\hat{\mathbf{w}}_1'$ *such that* $\hat{\mathbf{w}} \sim \mathcal{N}(0_D, I_D)$ *and* $\hat{\mathbf{w}}_1' \sim \psi(\cdot \,|\, \theta)$*. This will be used in the proofs to compute sums over the random variables in the following way, for some function* f *we have that*

$$\sum_j f(\mathbf{u}_j) \overset{d}{=} \sum_j f(\hat{\mathbf{w}}_j) + f(\hat{\mathbf{w}}_1') - f(\hat{\mathbf{w}}_1).$$

The convergence of the right hand side is then split to deal with the sum where all random variables are Gaussian and the difference between a $\psi(\cdot \,|\, \theta)$ *and Gaussian distributed random variable, which is usually much easier then working with the left hand side.*

Proof [Proof of Proposition 5] We start by proving the result for $i = 0$, first we assume that $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$ and secondly we relax this assumption by assuming that $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \theta)$. For ease of notation in the proof we will assume that we only have one observation ($T = 1$), the extension to many observations is immediate.

We assume that $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$ and let $\hat{\mathbf{v}} := \hat{\mathbf{p}}[0] \sin(\frac{h}{2}) + \hat{\mathbf{u}}[0] \cos(\frac{h}{2})$ then $\hat{\mathbf{v}}$ also follows a $\mathcal{N}(0_D, I_D)$ distribution, since by definition we have that $\hat{\mathbf{p}}[0] \sim \mathcal{N}(0_D, I_D)$. Now we write

$$\log \widehat{p}(y \,|\, \theta, \hat{\mathbf{v}}) - \log p(y \,|\, \theta) = \log\left\{1 + \frac{\widehat{p}(y \,|\, \theta, \hat{\mathbf{v}}) - p(y \,|\, \theta)}{p(y \,|\, \theta)}\right\} = \log\left\{1 + \frac{\varepsilon_N(y, \hat{\mathbf{v}}; \theta)}{\sqrt{N}}\right\},$$

where

$$\varepsilon_N(y, \hat{\mathbf{v}}; \theta) := \sqrt{N} \frac{\widehat{p}(y \,|\, \theta, \hat{\mathbf{v}}) - p(y \,|\, \theta)}{p(y \,|\, \theta)} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{\varpi_\theta(y, \hat{\mathbf{v}}_i) - 1\}.$$

Taking the gradient with respect to θ we get that

$$\sqrt{N}\{\nabla_\theta \log \widehat{p}(y \,|\, \theta, \hat{\mathbf{v}}) - \nabla_\theta p(y \,|\, \theta)\} = \frac{\nabla_\theta \varepsilon_N(y, \hat{\mathbf{v}}; \theta)}{1 + \varepsilon_N(y, \hat{\mathbf{v}}; \theta)/\sqrt{N}}.$$

By the definitions we have that

$$\begin{aligned}
&\mathbb{E}[\varepsilon_N(y, \hat{\mathbf{v}}; \theta)^2] = \mathbb{E}[(\tfrac{1}{\sqrt{N}} \sum_{i=1}^{N} \{\varpi_\theta(y, \hat{\mathbf{v}}_i) - 1\})^2] \\
=&\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \{\mathbb{E}[\varpi_\theta(y, \hat{\mathbf{v}}_i) \varpi_\theta(y, \hat{\mathbf{v}}_j)] - \mathbb{E}[\varpi_\theta(y, \hat{\mathbf{v}}_i)] - \mathbb{E}[\varpi_\theta(y, \hat{\mathbf{v}}_j)] + 1\} \\
=&\mathbb{E}[\varpi_\theta(y, \hat{\mathbf{v}}_1)^2] - 1,
\end{aligned}$$

where $\mathbb{E}[\varpi_\theta(y, \hat{\mathbf{v}}_1)^2] < \infty$ by assumption. Under the assumption that $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$ we have that $\varepsilon_N(y, \hat{\mathbf{v}}; \theta)/\sqrt{N} \xrightarrow{p} 0$ by Chebyshev's inequality. Also noting that

$$\mathbb{E}[\nabla_\theta \varepsilon_N(y, \hat{\mathbf{v}}; \theta)] = \sqrt{N}\mathbb{E}[\nabla_\theta \varpi_\theta(y, \hat{\mathbf{v}}_1)] \overset{(1)}{=} \sqrt{N}\nabla_\theta \mathbb{E}[\varpi_\theta(y, \hat{\mathbf{v}}_1)] = 0,$$

where (1) holds under the assumptions of the existence of the function $g(\cdot)$ such that $|\nabla_\theta \varpi_\theta(y, \mathbf{u})| < g(\mathbf{u})$ and $\mathbb{E}_{\mathcal{N}}[g(\mathbf{u})] < \infty$ which then allows us to interchange the differential and integral. By the continuous mapping theorem, Slutsky's lemma and the standard CLT applied to $\nabla_\theta \varepsilon_N(y, \hat{\mathbf{v}}; \theta)$ we get the desired results.

So far we have only proven the result when the algorithm is initialized, that is $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$. At stationarity we will have $(\theta, \hat{\mathbf{u}}) \sim \bar{\pi}(\theta, \hat{\mathbf{u}})$. Now we wish to prove the results under the assumption of having reached this distribution, that is $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \theta)$. We make use of Remark 9 and introduce the variables $\hat{\mathbf{w}} \sim N(0, I_D)$ and $\hat{\mathbf{w}}_1' \sim \psi(\cdot \,|\, \theta)$.

By adding and subtracting the term associated with $\hat{\mathbf{p}}[0]_1 \sin(\frac{h}{2}) + \hat{\mathbf{w}}_1 \cos(\frac{h}{2})$, for ease of notation we let $\hat{\mathbf{p}}[0]^{\hat{\mathbf{w}},h} := \hat{\mathbf{p}}[0]\sin(\frac{h}{2}) + \hat{\mathbf{w}}\cos(\frac{h}{2})$ and write

$$\begin{aligned}
&\sqrt{N}\{\nabla_\theta \log \widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[0]\sin(\tfrac{h}{2}) + \hat{\mathbf{u}}[0]\cos(\tfrac{h}{2})) - \nabla_\theta \log p(y \,|\, \theta)\} \\
&\quad = \frac{\nabla_\theta \varepsilon_N(y, \hat{\mathbf{p}}[0]\sin(\frac{h}{2}) + \hat{\mathbf{u}}[0]\cos(\frac{h}{2}))}{1 + \varepsilon_N(y, \hat{\mathbf{p}}[0]\sin(\frac{h}{2}) + \hat{\mathbf{u}}[0]\cos(\frac{h}{2}))/\sqrt{N}} \\
&\quad \overset{d}{=} \frac{\nabla_\theta \varepsilon_N(y, \hat{\mathbf{p}}[0]^{\hat{\mathbf{w}},h}; \theta) + \frac{1}{\sqrt{N}}(\nabla_\theta \varpi_\theta(y, \hat{\mathbf{p}}[0]_1 \sin(\frac{h}{2}) + \hat{\mathbf{w}}_1' \cos(\frac{h}{2})) - \nabla_\theta \varpi_\theta(\hat{\mathbf{p}}[0]_1^{\hat{\mathbf{w}},h}))}{1 + \frac{1}{\sqrt{N}}\varepsilon_N(y, \hat{\mathbf{p}}[0]^{\hat{\mathbf{w}},h}; \theta) + \frac{1}{N}(\varpi_\theta(y, \hat{\mathbf{p}}[0]_1 \sin(\frac{h}{2}) + \hat{\mathbf{w}}_1' \cos(\frac{h}{2})) - \varpi_\theta(y, \hat{\mathbf{p}}[0]_1^{\hat{\mathbf{w}},h}))},
\end{aligned}$$

as previously we have that

$$\begin{aligned}
\frac{1}{\sqrt{N}}\nabla_\theta \varpi_\theta(y, \hat{\mathbf{p}}[0]_1 \sin(\tfrac{h}{2}) + \hat{\mathbf{w}}_1' \cos(\tfrac{h}{2})) &\xrightarrow{p} 0 \quad \text{as } N \to \infty, \\
\frac{1}{N}\varpi_\theta(y, \hat{\mathbf{p}}[0]_1 \sin(\tfrac{h}{2}) + \hat{\mathbf{w}}_1' \cos(\tfrac{h}{2})) &\xrightarrow{p} 0 \quad \text{as } N \to \infty,
\end{aligned}$$

with the same results holding when we use $\hat{\mathbf{w}}_1$ instead of $\hat{\mathbf{w}}_1'$. By the continuous mapping theorem and Slutsky's lemma we have the result for the initial step.

Assume now that the following results hold at iteration ℓ for any $h > 0$,

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \nabla_\theta \varpi_\theta(y, \hat{\mathbf{p}}[\ell]_i \sin(h) + \hat{\mathbf{u}}[\ell]_i \cos(h)) &\xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta, y)), \qquad \text{as } N \to \infty, \\
\frac{1}{N}\sum_{i=1}^{N} \varpi_\theta(y, \hat{\mathbf{p}}[\ell]_i \sin(h) + \hat{\mathbf{u}}[\ell]_i \cos(h)) &\xrightarrow{p} 1, \qquad \text{as } N \to \infty.
\end{aligned}$$

This gives us using Slutsky's lemma that

$$\sqrt{N}\nabla_\theta \log\left(\frac{\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell]\sin(h) + \hat{\mathbf{u}}[\ell]\cos(h))}{p(y \,|\, \theta)}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta, y)), \quad \text{as } N \to \infty.$$

Taking one step to iteration $\ell + 1$ we get that

$$\begin{aligned}
&\hat{\mathbf{p}}[\ell+1]\sin(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell+1]\cos(\tfrac{h}{2}) \\
&\qquad = \hat{\mathbf{p}}[\ell]\sin(\tfrac{3h}{2}) + \hat{\mathbf{u}}[\ell]\cos(\tfrac{3h}{2}) + \sin(h)h\nabla_{\hat{\mathbf{u}}} \log \widehat{p}(y \,|\, \theta, \hat{\mathbf{u}}) \,|_{\hat{\mathbf{u}} = \hat{\mathbf{p}}[\ell]\sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]\cos(\frac{h}{2})},
\end{aligned}$$

from the assumptions we have that $\varpi_\theta(y, \mathbf{v}) > 0$ and so we can use Lemma 10 and Lemma 11 which gives us that, using Slutsky's lemma

$$\sqrt{N}\nabla_\theta \log\left(\frac{\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell+1]\sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell+1]\cos(\frac{h}{2}))}{p(y \,|\, \theta)}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta, y)), \quad \text{as } N \to \infty,$$

where $\sigma^2(y, \theta) = \mathbb{E}[\{\nabla_\theta \varpi_\theta(y, \mathbf{u})\}^2]$. This finishes the proof. ■

Appendix D. Proof of Lipschitz Continuity

Proof [Proof of Lemma 8] By explicitly writing out $\tilde{\Phi}_h$ we get that

$$\begin{aligned}
&\tilde{\Phi}_h(\Theta) = \tilde{\Phi}_h(\theta, \rho, \mathbf{u}, \mathbf{p}) \\
&= \begin{pmatrix}
\theta + h\rho + \frac{h^2}{2}\nabla_\theta\{\log p(\theta \,|\, y)\}_{\,|\,\theta=\theta+\frac{h}{2}\rho} \\
\rho + h\nabla_\theta\{\log p(\theta \,|\, y)\}_{\,|\,\theta=\theta+\frac{h}{2}\rho} \\
\mathbf{p}\sin(h) + \mathbf{u}\cos(h) + \sin(\frac{h}{2})h\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta + \frac{h}{2}\rho, \mathbf{u})_{\,|\,\mathbf{u}=\mathbf{p}\sin(\frac{h}{2})+\mathbf{u}\cos(\frac{h}{2})} \\
\mathbf{p}\cos(h) - \mathbf{u}\sin(h) + \cos(\frac{h}{2})h\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta + \frac{h}{2}\rho, \mathbf{u})_{\,|\,\mathbf{u}=\mathbf{p}\sin(\frac{h}{2})+\mathbf{u}\cos(\frac{h}{2})}
\end{pmatrix},
\end{aligned}$$

we need to find a $\mathcal{L}$ such that

$$\|\tilde{\Phi}_h(\Theta) - \tilde{\Phi}_h(\hat{\Theta})\| \leq \mathcal{L}\|\Theta - \hat{\Theta}\|,$$

or equivalently find $\mathcal{L}^2$ such that

$$\|\tilde{\Phi}_h(\Theta) - \tilde{\Phi}_h(\hat{\Theta})\|^2 \leq \mathcal{L}^2\Big\{\|\theta - \hat{\theta}\|^2 + \|\rho - \hat{\rho}\|^2 + \|\mathbf{u} - \hat{\mathbf{u}}\|^2 + \|\mathbf{p} - \hat{\mathbf{p}}\|^2\Big\}.$$

We have that

$$\begin{aligned}
&\|\tilde{\Phi}_h(\Theta) - \tilde{\Phi}_h(\hat{\Theta})\|^2 \\
\leq&\|\theta - \hat{\theta}\|^2 + (1+h^2)\|\rho - \hat{\rho}\|^2 + 2\cos^2(h)\|\mathbf{u} - \hat{\mathbf{u}}\| + 2\sin^2(h)\|\mathbf{p} - \hat{\mathbf{p}}\| \\
&+(h^2 + \tfrac{h^4}{4})\|\nabla_\theta \log p(\theta \,|\, y)_{\,|\,\theta=\theta+\frac{h}{2}\rho} - \nabla_\theta \log p(\theta \,|\, y)_{\,|\,\theta=\hat{\theta}+\frac{h}{2}\hat{\rho}}\|^2 \\
&+h^2\|\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta + \tfrac{h}{2}\rho, \mathbf{u})_{\,|\,\mathbf{u}=\mathbf{p}\sin(\frac{h}{2})+\mathbf{u}\cos(\frac{h}{2})} - \nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \hat{\theta} + \tfrac{h}{2}\hat{\rho}, \mathbf{u})_{\,|\,\mathbf{u}=\hat{\mathbf{p}}\sin(\frac{h}{2})+\hat{\mathbf{u}}\cos(\frac{h}{2})}\|^2.
\end{aligned}$$

Under the assumption that $\nabla_\theta \log p(\theta \,|\, y)$ is Lipschitz with constant L_0 and assuming that $\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta, \mathbf{u})$ is Lipschitz with constant L_1 we have that

$$\begin{aligned}
&\|\tilde{\Phi}_h(\Theta) - \tilde{\Phi}_h(\hat{\Theta})\|^2 \\
\leq&(1 + L_0^2(h^2 + \tfrac{h^4}{4}) + h^2L_1^2)\|\theta - \hat{\theta}\|^2 + (1 + h^2 + L_0^2\tfrac{h^2}{4}(h^2 + \tfrac{h^4}{4}) + \tfrac{h^2}{4}L_1^2)\|\rho - \hat{\rho}\|^2 \\
&+(2\cos^2(h) + L_1^2h^2\cos^2(\tfrac{h}{2}))\|\mathbf{u} - \hat{\mathbf{u}}\|^2 + (2\sin^2(h) + L_1^2h^2\sin^2(\tfrac{h}{2}))\|\mathbf{p} - \hat{\mathbf{p}}\|^2,
\end{aligned}$$

it therefore holds that $\tilde{\Phi}_h$ is Lipschitz with constant $\mathcal{L} = \sqrt{2 + h^2 + L_0^2(h^2 + \frac{h^4}{4}) + h^2L_1^2}$. It remains to prove that $\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta, \mathbf{u})$ is Lipschitz with constant L_1 and that L_1 does not grow with N.

To establish this, note that we have

$$\begin{aligned}
&\Big\|\nabla_{\mathbf{u}} \log \widehat{p}\Big(y|\hat{\theta}, \hat{\mathbf{u}}\Big) - \nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \mathbf{u})\Big\| \\
&\quad \leq \|\nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \hat{\mathbf{u}}) - \nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \mathbf{u})\| \qquad (22) \\
&\quad + \Big\|\nabla_{\mathbf{u}} \log \widehat{p}\Big(y|\hat{\theta}, \hat{\mathbf{u}}\Big) - \nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \hat{\mathbf{u}})\Big\|. \qquad (23)
\end{aligned}$$

Consider first term (22) (squared),

$$\begin{aligned}
&\|\nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \hat{\mathbf{u}}) - \nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \mathbf{u})\|^2 \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N}\Big\|\nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \hat{\mathbf{u}}_t) - \nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \mathbf{u}_t)\Big\|^2.
\end{aligned}$$

We have

$$\nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \mathbf{u}_t) = \frac{\omega_\theta(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \mathbf{u}_{t,j})}\nabla_{\mathbf{u}_{t,i}} \log \omega_\theta(y_t, \mathbf{u}_{t,i})$$

so

$$\begin{aligned}
&\Big\|\nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \hat{\mathbf{u}}_t) - \nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \mathbf{u}_t)\Big\|^2 \\
&= \left\|\frac{\omega_\theta(y_t, \hat{\mathbf{u}}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \hat{\mathbf{u}}_{t,j})}\nabla_{\mathbf{u}_{t,i}} \log \omega_\theta(y_t, \hat{\mathbf{u}}_{t,i}) - \frac{\omega_\theta(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \mathbf{u}_{t,j})}\nabla_{\mathbf{u}_{t,i}} \log \omega_\theta(y_t, \mathbf{u}_{t,i})\right\|^2 \\
&\leq \left\{\left\|\frac{\omega_\theta(y_t, \hat{\mathbf{u}}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \hat{\mathbf{u}}_{t,j})}\Big(\nabla_{\mathbf{u}_{t,i}} \log \omega_\theta(y_t, \hat{\mathbf{u}}_{t,i}) - \nabla_{\mathbf{u}_{t,i}} \log \omega_\theta(y_t, \mathbf{u}_{t,i})\Big)\right\|\right. \qquad (24) \\
&\quad \left. + \left\|\left(\frac{\omega_\theta(y_t, \hat{\mathbf{u}}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \hat{\mathbf{u}}_{t,j})} - \frac{\omega_\theta(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \mathbf{u}_{t,j})}\right)\nabla_{\mathbf{u}_{t,i}} \log \omega_\theta(y_t, \mathbf{u}_{t,i})\right\|\right\}^2. \qquad (25)
\end{aligned}$$

Under the Lipschitz assumption on the gradient of the log-weight-function the term on line (24) is bounded by $M\|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\|$. Furthermore, under the boundedness and Lipschitz assumptions on the weight function the term on line (25) is bounded by

$$\begin{aligned}
&C\left|\frac{\omega_\theta(y_t, \hat{\mathbf{u}}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \hat{\mathbf{u}}_{t,j})} - \frac{\omega_\theta(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \mathbf{u}_{t,j})}\right| \\
&\leq C\left|\frac{\omega_\theta(y_t, \hat{\mathbf{u}}_{t,i}) - \omega_\theta(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \hat{\mathbf{u}}_{t,j})}\right| + C\frac{\omega_\theta(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_\theta(y_t, \mathbf{u}_{t,j})}\left|\frac{\sum_{j=1}^{N}\{\omega_\theta(y_t, \mathbf{u}_{t,j}) - \omega_\theta(y_t, \hat{\mathbf{u}}_{t,j})\}}{\sum_{j=1}^{N}\omega_{\hat{\theta}}(y_t, \hat{\mathbf{u}}_{t,j})}\right| \\
&\leq \frac{CD}{N\underline{\omega}}\|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\| + \frac{CD}{N\underline{\omega}}\sum_{j=1}^{N}\|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\| \\
&\leq \frac{2CD}{N\underline{\omega}}\sum_{j=1}^{N}\|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\|.
\end{aligned}$$

Put together we get for the term (22) (squared),

$$
\begin{aligned}
&\Big\|\nabla_{\mathbf{u}} \log \widehat{p}\Big(y|\hat{\theta}, \hat{\mathbf{u}}\Big) - \nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \mathbf{u})\Big\|^2 \\
&\leq \sum_{t=1}^{T}\sum_{i=1}^{N}\left\{ M\|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\| + \frac{2CD}{N\underline{\omega}} \sum_{j=1}^{N} \|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\| \right\}^2 \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N}\left\{ M^2\|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\|^2 + \left(\frac{2CD}{N\underline{\omega}} \sum_{j=1}^{N} \|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\| \right)^2 + \frac{4MCD}{N\underline{\omega}} \|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\| \sum_{j=1}^{N} \|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\| \right\} \\
&= \sum_{t=1}^{T}\left\{ M^2 \sum_{i=1}^{N} \|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\|^2 + \left(\left[\frac{2CD}{\underline{\omega}}\right]^2 + \frac{4MCD}{\underline{\omega}} \right) \times \frac{1}{N} \left(\sum_{j=1}^{N} \|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\| \right)^2 \right\} \\
&\leq \sum_{t=1}^{T}\left\{ M^2 \sum_{i=1}^{N} \|\hat{\mathbf{u}}_{t,i} - \mathbf{u}_{t,i}\|^2 + \left(\left[\frac{2CD}{\underline{\omega}}\right]^2 + \frac{4MCD}{\underline{\omega}} \right) \times \sum_{j=1}^{N} \|\hat{\mathbf{u}}_{t,j} - \mathbf{u}_{t,j}\|^2 \right\} \\
&= \left(M + \frac{2CD}{\underline{\omega}} \right)^2 \|\hat{\mathbf{u}} - \mathbf{u}\|^2,
\end{aligned}
$$

where the inequality on the penultimate line follows from Jensen's inequality.

Next we address the term (23). Analogously to above we have

$$
\begin{aligned}
&\|\nabla_{\mathbf{u}} \log \widehat{p}(y|\theta', \mathbf{u}) - \nabla_{\mathbf{u}} \log \widehat{p}(y|\theta, \mathbf{u})\|^2 \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N} \Big\| \nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta', \mathbf{u}_t) - \nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \mathbf{u}_t) \Big\|^2
\end{aligned} \tag{26}
$$

and

$$
\begin{aligned}
&\Big\| \nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta', \mathbf{u}_t) - \nabla_{\mathbf{u}_{t,i}} \log \widehat{p}(y_t|\theta, \mathbf{u}_t) \Big\|^2 \\
&\leq \left\{ \left\| \frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} \Big(\nabla_{\mathbf{u}_{t,i}} \log \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i}) - \nabla_{\mathbf{u}_{t,i}} \log \omega_{\theta}(y_t, \mathbf{u}_{t,i}) \Big) \right\| \right. \\
&\left. + \left\| \left(\frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} - \frac{\omega_{\theta}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\theta}(y_t, \mathbf{u}_{t,j})} \right) \nabla_{\mathbf{u}_{t,i}} \log \omega_{\theta}(y_t, \mathbf{u}_{t,i}) \right\| \right\}^2 \\
&\leq \left\{ \frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} M\|\hat{\theta} - \theta\| + C \frac{|\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i}) - \omega_{\theta}(y_t, \mathbf{u}_{t,i})|}{\sum_{j=1}^{N} \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} \right. \\
&\quad \left. + C \frac{\omega_{\theta}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\theta}(y_t, \mathbf{u}_{t,j})} \frac{\left| \sum_{j=1}^{N} \omega_{\theta}(y_t, \mathbf{u}_{t,j}) - \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j}) \right|}{\sum_{j=1}^{N} \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} \right\}^2 \\
&\leq \left\{ \frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} M + \frac{CD}{\underline{\omega}} \left(\frac{1}{N} + \frac{\omega_{\theta}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N} \omega_{\theta}(y_t, \mathbf{u}_{t,j})} \right) \right\}^2 \|\hat{\theta} - \theta\|^2.
\end{aligned}
$$

Plugging this expression into (26) we get

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{u}} \log \widehat{p}(\, y|\theta', \mathbf{u}) - \nabla_{\mathbf{u}} \log \widehat{p}(\, y|\theta, \mathbf{u})\right\|^2 \\
&\leq \sum_{t=1}^{T}\sum_{i=1}^{N}\Bigg\{ \left(\frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})}\right)^2 M^2 + \frac{C^2D^2}{\underline{\omega}^2}\left(\frac{1}{N} + \frac{\omega_{\theta}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\theta}(y_t, \mathbf{u}_{t,j})}\right)^2 \\
&\quad + \frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})}\frac{2MCD}{\underline{\omega}}\left(\frac{1}{N} + \frac{\omega_{\theta}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\theta}(y_t, \mathbf{u}_{t,j})}\right)\Bigg\}\|\hat{\theta} - \theta\|^2 \\
&\leq \sum_{t=1}^{T}\sum_{i=1}^{N}\Bigg\{ \frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})} M^2 + \frac{C^2D^2}{\underline{\omega}^2}\left(\frac{1}{N^2} + \left(\frac{2}{N} + 1\right)\frac{\omega_{\theta}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\theta}(y_t, \mathbf{u}_{t,j})}\right) \\
&\quad + \frac{\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,i})}{\sum_{j=1}^{N}\omega_{\hat{\theta}}(y_t, \mathbf{u}_{t,j})}\frac{2MCD}{\underline{\omega}}\left(\frac{1}{N} + 1\right)\Bigg\}\|\hat{\theta} - \theta\|^2 \\
&= \sum_{t=1}^{T}\left\{ M^2 + \frac{C^2D^2}{\underline{\omega}^2}\left(\frac{1}{N} + \frac{2}{N} + 1\right) + \frac{2MCD}{\underline{\omega}}\left(\frac{1}{N} + 1\right)\right\}\|\hat{\theta} - \theta\|^2 \\
&\leq T\left(M + \frac{2CD}{\underline{\omega}}\right)^2\|\hat{\theta} - \theta\|^2.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{u}} \log \widehat{p}\Big(\, y|\hat{\theta}, \hat{\mathbf{u}}\Big) - \nabla_{\mathbf{u}} \log \widehat{p}(\, y|\theta, \mathbf{u})\right\| \\
&\quad \leq \left(M + \frac{2CD}{\underline{\omega}}\right)\|\hat{\mathbf{u}} - \mathbf{u}\| + \sqrt{T}\left(M + \frac{2CD}{\underline{\omega}}\right)\|\hat{\theta} - \theta\| \\
&\quad \leq (\sqrt{T} + 1) \times \left(M + \frac{2CD}{\underline{\omega}}\right)\left\|\begin{pmatrix}\hat{\theta}\\ \hat{\mathbf{u}}\end{pmatrix} - \begin{pmatrix}\theta\\ \mathbf{u}\end{pmatrix}\right\|.
\end{aligned}
$$

■

Appendix E. Proof of Convergence of Acceptance Probability

Proof [Proof of Proposition 6] The proof is divided in two parts. In the first part, we will show that the log-likelihood estimator $\log \widehat{p}(y \,|\, \theta, \mathbf{u})$ converges to the log-likelihood $\log p(y \,|\, \theta)$ as $N \to \infty$. The second part of the proof will show that the remaining $\mathbf{u}$ and $\mathbf{p}$ parts of the acceptance probability vanishes as N increases.

The first part follows by the proof of Proposition 5, see Appendix C. For completeness we repeat that part here, we do the proof by induction to show that for all ℓ we have that,

$$\log \widehat{p}(y \,|\, \theta, \hat{\mathbf{u}}[\ell]) \xrightarrow{p} \log p(y \,|\, \theta), \quad \text{as } N \to \infty.$$

It turns out that it is needed to show this result in a more general setting, that is for any $h > 0$ we wish to show that

$$\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell]\sin(h) + \hat{\mathbf{u}}[\ell]\cos(h)) \xrightarrow{p} p(y \,|\, \theta), \quad \text{as } N \to \infty. \tag{27}$$

When $\ell = 0$ we prove the result in two different settings. First we assume that $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0_D, I_D)$ and the result is clear since (13) is just the likelihood of the importance sampling estimator under the assumption of Gaussian proposal distribution. Secondly we look at the case when $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \theta)$, we again introduce the variables $\hat{\mathbf{w}}_1' \sim \psi(\cdot \,|\, \theta)$ and $\hat{\mathbf{w}} \sim \mathcal{N}(0_D, I_D)$ by the use of Remark 9, we get that

$$\begin{aligned}
&\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[0]\sin(h) + \hat{\mathbf{u}}[0]\cos(h)) = \frac{1}{N}\sum_{i=1}^{N} \omega_\theta(\hat{\mathbf{p}}[0]_i \sin(h) + \hat{\mathbf{u}}[0]_i \cos(h)) \\
&\overset{d}{=} \frac{1}{N}\omega_\theta(\hat{\mathbf{p}}[0]_1 \sin(h) + \hat{\mathbf{w}}_1' \cos(h)) + \frac{1}{N}\sum_{i=2}^{N} \omega_\theta(\hat{\mathbf{p}}[0]_i \sin(h) + \hat{\mathbf{w}}_i \cos(h)),
\end{aligned}$$

here the first part converges to zero and the second part converges to the likelihood. By Slutsky's lemma and the continuous mapping theorem we have the result.

Assume now that (27) holds for any $h > 0$. We then have for $\ell + 1$ and any $h' > 0$ that might differ from the integration step that

$$\begin{aligned}
&\widehat{p}(y \,|\, \theta, \hat{\mathbf{p}}[\ell+1]\sin(h') + \hat{\mathbf{u}}[\ell+1]\cos(h')) \\
&= \widehat{p}\Big(y \,|\, \theta, \hat{\mathbf{p}}[\ell]\sin(h'+h) + \hat{\mathbf{u}}[\ell]\cos(h'+h) \\
&\qquad\qquad + h\sin(h' + \tfrac{h}{2})\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta, \mathbf{u})_{\,|\,\mathbf{u}=\hat{\mathbf{p}}[\ell]\sin(\frac{h}{2})+\hat{\mathbf{u}}[\ell]\cos(\frac{h}{2})}\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N} \omega_\theta\Big(y, \hat{\mathbf{p}}[\ell]\sin(h'+h) + \hat{\mathbf{u}}[\ell]\cos(h'+h) \\
&\qquad\qquad + h\sin(h' + \tfrac{h}{2})\nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta, \mathbf{u})_{\,|\,\mathbf{u}=\hat{\mathbf{p}}[\ell]\sin(\frac{h}{2})+\hat{\mathbf{u}}[\ell]\cos(\frac{h}{2})}\Big),
\end{aligned}$$

which converges to $p(y \,|\, \theta)$ by Lemma 11. The result now follows by the continuous mapping theorem.

Let us take a look at the second part. For this part we will show that

$$\hat{\mathbf{u}}[0]^T\hat{\mathbf{u}}[0] + \hat{\mathbf{p}}[0]^T\hat{\mathbf{p}}[0] - \hat{\mathbf{u}}[L]^T\hat{\mathbf{u}}[L] - \hat{\mathbf{p}}[L]^T\hat{\mathbf{p}}[L] \xrightarrow{p} 0, \quad \text{as } N \to \infty.$$

We do this by rewriting this expression using a telescoping sum to

$$\begin{aligned}
&\hat{\mathbf{u}}[0]^T\hat{\mathbf{u}}[0] + \hat{\mathbf{p}}[0]^T\hat{\mathbf{p}}[0] - \hat{\mathbf{u}}[L]^T\hat{\mathbf{u}}[L] - \hat{\mathbf{p}}[L]^T\hat{\mathbf{p}}[L] \\
&= \sum_{i=1}^{N}\Big(\hat{\mathbf{u}}[0]_i^2 + \hat{\mathbf{p}}[0]_i^2 - \hat{\mathbf{u}}[L]_i^2 - \hat{\mathbf{p}}[L]_i^2\Big) \\
&= \sum_{\ell=0}^{L-1}\sum_{i=1}^{N}\Big(\hat{\mathbf{u}}[\ell]_i^2 + \hat{\mathbf{p}}[\ell]_i^2 - \hat{\mathbf{u}}[\ell+1]_i^2 - \hat{\mathbf{p}}[\ell+1]_i^2\Big).
\end{aligned}$$

What remains to prove is that, for all ℓ,

$$\sum_{i=1}^{N}\Big(\hat{\mathbf{u}}[\ell]_i^2 + \hat{\mathbf{p}}[\ell]_i^2 - \hat{\mathbf{u}}[\ell+1]_i^2 - \hat{\mathbf{p}}[\ell+1]_i^2\Big) \xrightarrow{p} 0, \quad \text{as } N \to \infty.$$

By taking one step of the integrator we have that, see Appendix A

$$\hat{\mathbf{u}}[\ell+1]_i = \hat{\mathbf{p}}[\ell]_i \sin(h) + \hat{\mathbf{u}}[\ell]_i \cos(h) + \sin(\tfrac{h}{2}) h \nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i)_{|\mathbf{u}_i = \hat{\mathbf{p}}[\ell]_i \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \cos(\frac{h}{2})}$$

$$\hat{\mathbf{p}}[\ell+1]_i = \hat{\mathbf{p}}[\ell]_i \cos(h) - \hat{\mathbf{u}}[\ell]_i \sin(h) + \cos(\tfrac{h}{2}) h \nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i)_{|\mathbf{u}_i = \hat{\mathbf{p}}[\ell]_i \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \cos(\frac{h}{2})},$$

combining these we get that, using some trigonometric equalities,

$$\begin{aligned}\hat{\mathbf{u}}[\ell+1]_i^2 + \hat{\mathbf{p}}[\ell+1]_i^2 = \hat{\mathbf{p}}[\ell]_i^2 + \hat{\mathbf{u}}[\ell]_i^2 + h^2 (\nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i)_{|\mathbf{u}_i = \hat{\mathbf{p}}[\ell]_i \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \cos(\frac{h}{2})})^2 \\ + 2h(\hat{\mathbf{p}}[\ell]_i \cos(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \sin(\tfrac{h}{2})) \nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i)_{|\mathbf{u}_i = \hat{\mathbf{p}}[\ell]_i \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \cos(\frac{h}{2})}.\end{aligned}$$

We now need to show two results, both holding as $N \to \infty$,

(i) $\sum_{i=1}^{N} (\nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i)_{|\mathbf{u}_i = \hat{\mathbf{p}}[\ell]_i \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \cos(\frac{h}{2})})^2 \xrightarrow{p} 0,$

(ii) $\sum_{i=1}^{N} (\hat{\mathbf{p}}[\ell]_i \cos(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \sin(\tfrac{h}{2})) \nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i)_{|\mathbf{u}_i = \hat{\mathbf{p}}[\ell]_i \sin(\frac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \cos(\frac{h}{2})} \xrightarrow{p} 0.$

Starting with (i) we have that, by the assumptions on the weight functions, that

$$\sum_{i=1}^{N} (\nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i))^2 \leq \sum_{i=1}^{N} \frac{\overline{\omega}^2}{(\sum_{i=1}^{N} \underline{\omega})^2} C^2 = \frac{\overline{\omega}^2 C}{N \underline{\omega}^2} \to 0, \quad \text{as } N \to \infty.$$

For (ii) using the same bounds related to the gradient we get

$$\begin{aligned}&\left| \sum_{i=1}^{N} (\hat{\mathbf{p}}[\ell]_i \cos(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \sin(\tfrac{h}{2})) \nabla_{\mathbf{u}_i} \log \widehat{p}(y \,|\, \theta, \mathbf{u}_i) \right| \\ &\qquad \leq \frac{1}{N} \frac{C\overline{\omega}}{\underline{\omega}} \sum_{i}^{N} (\hat{\mathbf{p}}[\ell]_i \cos(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \sin(\tfrac{h}{2})).\end{aligned}$$

We now need to prove that

$$\frac{1}{N} \sum_{i}^{N} (\hat{\mathbf{p}}[\ell]_i \cos(\tfrac{h}{2}) + \hat{\mathbf{u}}[\ell]_i \sin(\tfrac{h}{2})) \xrightarrow{p} 0, \quad \text{as } N \to \infty.$$

We will prove this for every $\ell \geq 0$ and for any value of h and this will be done using a proof of induction similar to what is done in the proof of the CLT in Appendix C.

We begin when $\ell = 0$, we have that $\hat{\mathbf{p}}[0] \sim \mathcal{N}(0_D, I_D)$ and assume first that $\hat{\mathbf{u}}[0] \sim \mathcal{N}(0, I_D)$. Then the result is trivial. If we instead assume that $\hat{\mathbf{u}}[0] \sim \bar{\pi}(\cdot \,|\, \theta)$. We again use Remark 9 and introduce the set of variables $\hat{\mathbf{w}}_1 \sim \psi(\cdot \,|\, \theta)$ and $\hat{\mathbf{w}}_{2:N} \sim \mathcal{N}(0_p, I_p)$. We then have that, for any $h > 0$,

$$\begin{aligned}&\frac{1}{N} \sum_{i=1}^{N} (\hat{\mathbf{p}}[0]_i \cos(h) + \hat{\mathbf{u}}[0]_i \sin(h)) \overset{d}{=} \frac{1}{N} \sum_{i=1}^{N} (\hat{\mathbf{p}}[0]_i \cos(h) + \hat{\mathbf{w}}_i \sin(h)) \\ &= \frac{1}{N} (\hat{\mathbf{p}}[0]_1 \cos(h) + \hat{\mathbf{w}}_1 \sin(h)) + \frac{1}{N} \sum_{i=2}^{N} (\hat{\mathbf{p}}[0]_i \cos(h) + \hat{\mathbf{w}}_i \sin(h)).\end{aligned}$$

Both of these terms now converge to 0, the first part trivially does this, while the second part is a sum of standard Gaussian variables and by the law of large numbers this sum converges to 0.

Assume now that for any $h > 0$ we have that

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{p}}[\ell]_i \cos(h) + \hat{\mathbf{u}}[\ell]_i \sin(h)) \xrightarrow{p} 0, \quad \text{as } N \to \infty.$$

We now look at $\ell + 1$, by taking one step of the numerical integrator we get that, for any $h' > 0$ which may be different then the integration step length,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{p}}[\ell+1]_i \cos(h') + \hat{\mathbf{u}}[\ell+1]_i \sin(h')) \\
=&\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{p}}[\ell]_i \sin(h+h') + \hat{\mathbf{u}}[\ell]_i \cos(h+h') \\
&\qquad + h\sin(\tfrac{h}{2}+h')\nabla_{\mathbf{u}_i}\log\widehat{p}(y\,|\,\theta,\mathbf{u}_i)_{\,|\,\mathbf{u}_i=\hat{\mathbf{p}}[\ell]_i\sin(\frac{h}{2})+\hat{\mathbf{u}}[\ell]_i\cos(\frac{h}{2})}) \\
=&\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{p}}[\ell]_i \sin(h+h') + \hat{\mathbf{u}}[\ell]_i \cos(h+h')) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N} h\sin(\tfrac{h}{2}+h')\nabla_{\mathbf{u}_i}\log\widehat{p}(y\,|\,\theta,\mathbf{u}_i)_{\,|\,\mathbf{u}_i=\hat{\mathbf{p}}[\ell]_i\sin(\frac{h}{2})+\hat{\mathbf{u}}[\ell]_i\cos(\frac{h}{2})},
\end{aligned}$$

where the first term converges to 0 in probability by the induction hypothesis and the second sum converges to 0 in probability in the same way as (i) above. This completes the proof. ■

Appendix F. Extension to Nosé-Hoover Dynamics

The Hamiltonian dynamics presented in (7), respectively (12), preserves the Hamiltonian (5), respectively (10). As mentioned earlier, it is thus necessary to randomize periodically the momentum to explore the target distribution of interest. On the contrary, Nosé-Hoover type dynamics do not preserve the Hamiltonian but keep the target distribution of interest invariant (Leimkuhler and Matthews, 2015, Chapter 8). However, they are not necessarily ergodic but can perform well and it is similarly possible to randomize them to ensure ergodicity.

Compared to the Hamiltonian dynamics (7), the Nosé-Hoover dynamics is given by

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \rho, \tag{28}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \nabla_\theta \log p(\theta) + \nabla_\theta \log p(y\,|\,\theta) - \xi\rho, \tag{29}$$

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = \mu^{-1}\Big(\rho^T\rho - d\Big), \tag{30}$$

where $\xi \in \mathbb{R}$ is the so-called thermostat. It is easy to check that this flow preserves

$$\pi(\theta, \rho, \xi) = \pi(\theta, \rho)\mathcal{N}\Big(\xi; 0, \mu^{-1}\Big)$$

invariant; i.e. if $(\theta(0), \rho(0), \xi(0)) \sim \pi$ then $(\theta(t), \rho(t), \xi(t)) \sim \pi$ for any $t > 0$. This dynamics can be straightforwardly extended to the intractable likelihood case to obtain

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \rho, \tag{31}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \nabla_\theta \log p(\theta) + \nabla_\theta \log \widehat{p}(y \,|\, \theta, \mathbf{u}) - \xi\rho, \tag{32}$$

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{p}, \tag{33}$$

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\mathbf{u} + \nabla_{\mathbf{u}} \log \widehat{p}(y \,|\, \theta, \mathbf{u}) - \xi\mathbf{p}, \tag{34}$$

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = \mu^{-1}\Big(\rho^T\rho + \mathbf{p}^T\mathbf{p} - (d + D)\Big). \tag{35}$$

This flow preserves

$$\overline{\pi}(\theta, \rho, \mathbf{u}, \mathbf{p}, \xi) = \overline{\pi}(\theta, \rho, \mathbf{u}, \mathbf{p})\mathcal{N}\Big(\xi; 0, \mu^{-1}\Big).$$

It would also be possible to use the thermostat ξ to only regulate ρ as in (28)-(30).

It is possible to combine Nosé-Hoover numerical integrators with an MH accept-reject step to preserve the invariant distribution Leimkuhler and Reich (2009). A weakness of this approach is that the MH accept-reject step is not only dependent of the difference between the ratio of the target at the proposed state and the target at the initial state but involves an additional Jacobian factor. We will not explore further these approaches here, which will be the subject of future work.

Appendix G. Details About the Numerical Illustrations and Additional Results

Here we will provide additional information about the numerical illustration from Section 4.

G.1 Diffraction Model

A normal $\mathcal{N}(0, 10^2)$ prior was used for each component of θ. We used $h = 0.02$ and $L = 50$ in all cases for simplicity, resulting in average acceptance probabilities in the range 0.6—0.8.

To generate the oberservations accept-reject sampling is used to generate random variables from the distribution

$$g(y \,|\, \lambda) = (\lambda\pi)^{-1} \operatorname{sinc}^2(\tfrac{y}{\lambda}),$$

which are then moved using the simulated x. For proposal we use the following distribution on the positive part of the real line

$$q(y \,|\, \lambda) = \begin{cases} (2\lambda)^{-1} & 0 \leq y < \lambda, \\ \frac{\lambda}{2y^2} & y \geq \lambda, \end{cases}$$

a random sign is then given to the accepted samples to generate samples from the desired distribution.

In Figures 5—9 below we show traces for the parameter λ (which, along with σ, was the most difficult parameter to infer) for the various samplers considered with different settings.

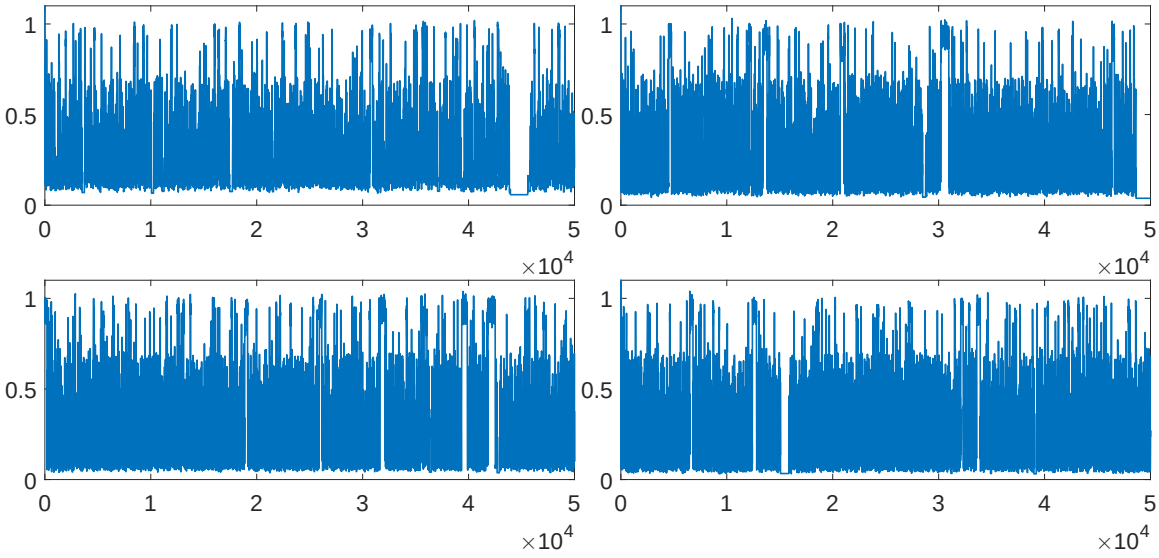

Figure 5: Traces for parameter λ for the **PM-HMC sampler**.From top left to bottom right: $N = 16$, $N = 64$, $N = 128$, $N = 256$.

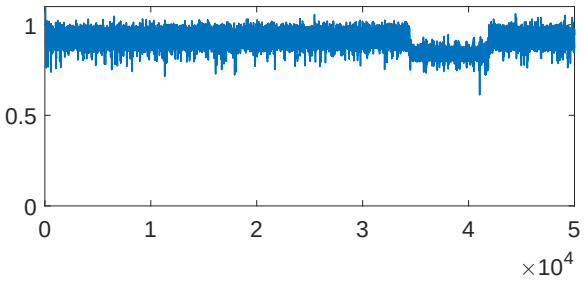

Figure 6: Traces for parameter λ for the **standard HMC sampler**, using a non-centered parameterisation. (Note that the sampler completely misses one mode of the posterior.)

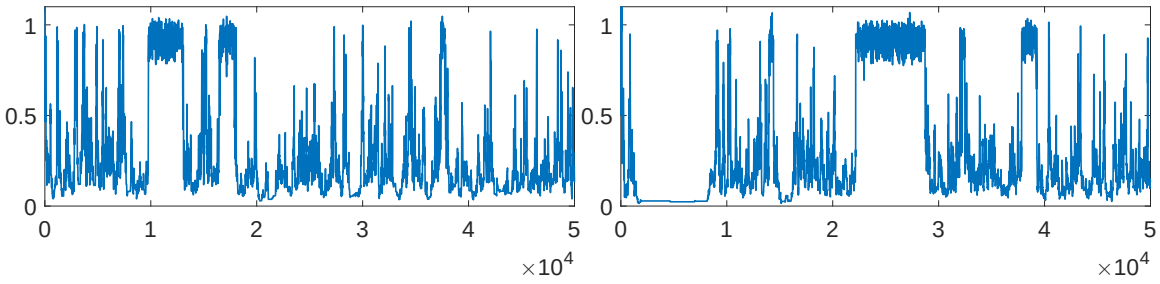

Figure 7: Traces for parameter λ for the **pseudo-marginal MH sampler** with N particles. Left, $N = 512$. Right, $N = 1024$.

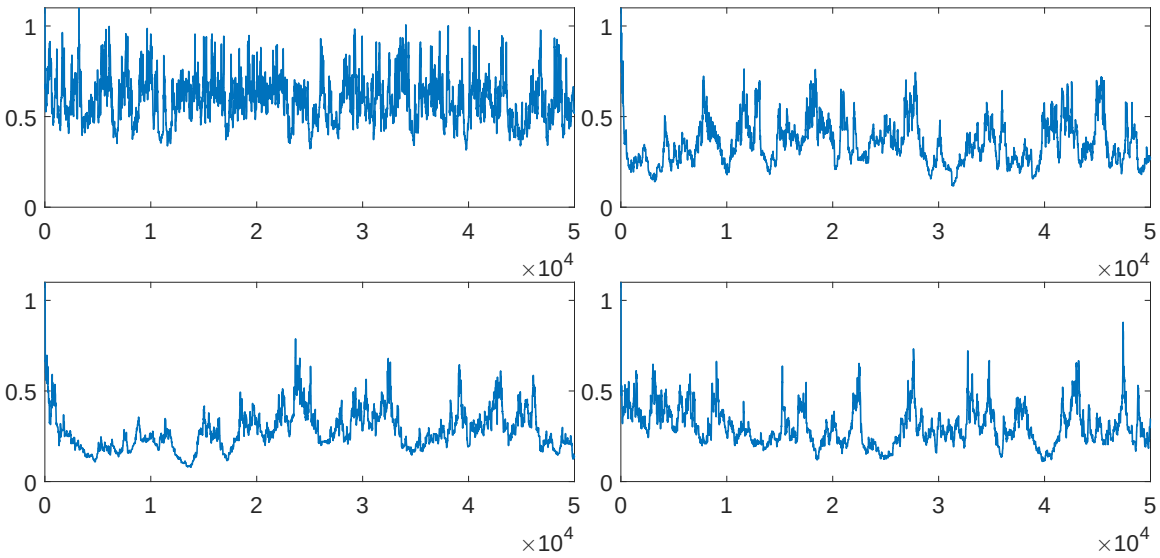

Figure 8: Traces for parameter λ for the **Gibbs sampler** using conditional importance sampling kernels with N particles to update the latent variables. From top left to bottom right: $N = 16$, $N = 64$, $N = 128$, $N = 256$.

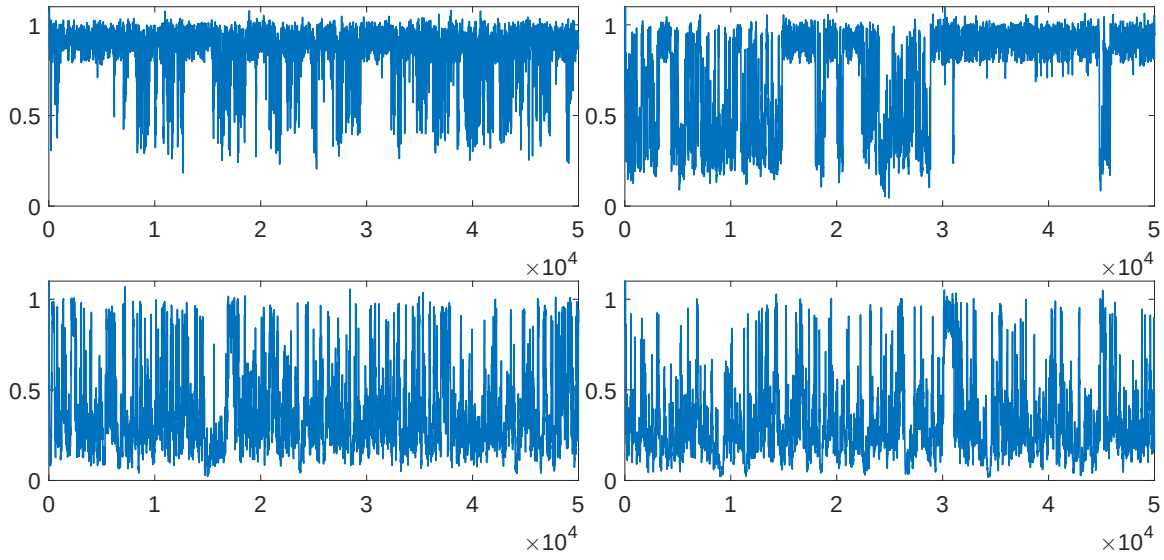

Figure 9: Traces for parameter λ for the **pseudo-marginal slice sampler**. From top left to bottom right: $N = 16$, $N = 64$, $N = 128$, $N = 256$.

G.2 Generalized Linear Mixed Model

In this section we give some additional details and results for the generalized linear mixed model considered in Section 4.3 of the main manuscript.

The parameter values used for the data generation were:

$$\beta = \begin{pmatrix} -1.1671 & 2.4665 & -0.1918 & -1.0080 & 0.6212 & 0.6524 & 1.5410 & 0.2653 \end{pmatrix}^T,$$

$\mu_1 = 0$, $\mu_2 = 3$, $\lambda_1 = 10$, $\lambda_2 = 3$, and $w_1 = 0.8$.

All methods were initialized at the same point in θ-space, as follows: β^{init} was sampled from $\mathcal{N}(0_p, I_p)$, resulting in:

$$\beta^{\text{init}} = \begin{pmatrix} 0.5838 & 0.3805 & -1.5062 & -0.0442 & 0.4717 & -0.1435 & 0.6371 & -0.0522 \end{pmatrix}^T,$$

whereas the remaining parameters were initialized deterministically as $\mu_1^{\text{init}} = 0$, $\mu_2^{\text{init}} = 0$ $\lambda_1^{\text{init}} = 1$, $\lambda_2^{\text{init}} = 0.1$, and $w_1^{\text{init}} = 0.5$.

We used a $\mathcal{N}(0, 100)$ prior for each component of θ. However, for the particle Gibbs sampler we used a different parameterisation and (uninformative) conjugate priors when possible to ease the implementation. Varying the prior did not have any noticeable effect on the (poor) mixing of the Gibbs sampler.

For PM-HMC and the pseudo-marginal slice sampler we used a simple (indeed, naive) choice of importance distribution for the latent variables: $q(x_i) = \mathcal{N}(x_i \,|\, 0, 3^2)$ since this was easily represented in terms of Gaussian auxiliary variables (which is a requirement for both methods). A possibly better choice, which however we have not tried in practice, is to use a Gaussian or t-distributed approximation to the posterior distribution of the latent variables. For the particle Gibbs sampler, which does not require the proposal to be

represented in terms of Gaussian auxiliary variables, we instead used the (slightly better) proposal consisting of sampling from the prior for X_i.

The pseudo-marginal slice sampler made use of elliptical slice sampling for updating the auxiliary variables, as recommended by Murray and Graham (2016). The components of θ were updated one-at-a-time using random walk Metropolis—Hastings kernels. We also updating θ jointly, but this resulted in very poor acceptance rates. The random-walk proposals were tuned to obtain acceptance rates of around 0.2—0.3.

The particle Gibbs sampler used conditional importance sampling kernels for the latent variables $X_{1:T}$, and for the parameters random walk Metropolis—Hastings kernels were used when no conjugate priors were available.

Figures 10—12 show trace plots for the four parameters μ_1, μ_2, $1/\lambda_1$ and $1/\lambda_2$ of the Gaussian mixture model used to model the distribution of the random effects. The three plots correspond to the pseudo-marginal HMC sampler, the pseudo-marginal slice sampler, and the Gibbs sampler with conditional importance sampling kernels, respectively. In Figure 13 we show estimated autocorrelations for the 13 parameters of the model for the three samplers.

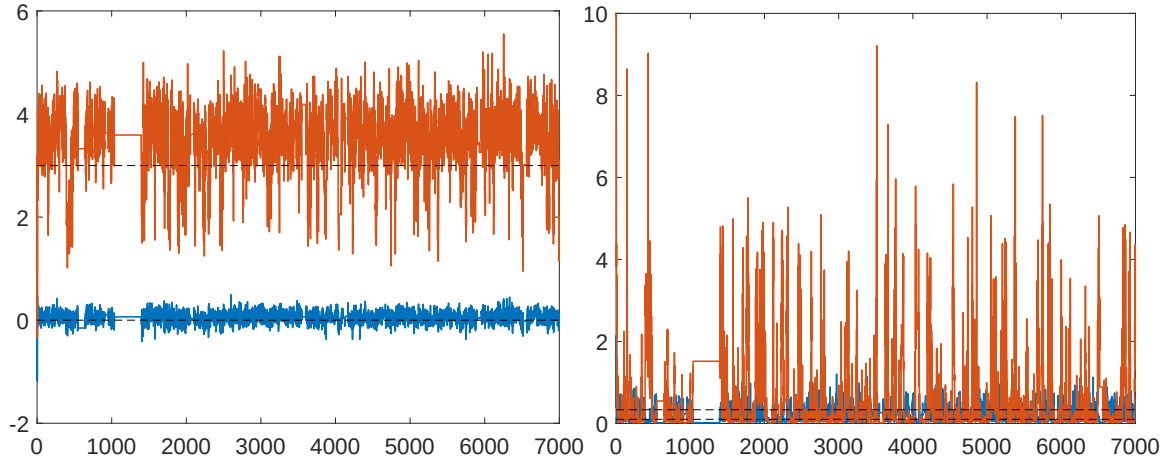

Figure 10: Traces for parameters μ_1 and μ_2 (left) and $1/\lambda_1$ and $1/\lambda_2$ (right) for the **pseudo-marginal HMC sampler** with $N = 128$.

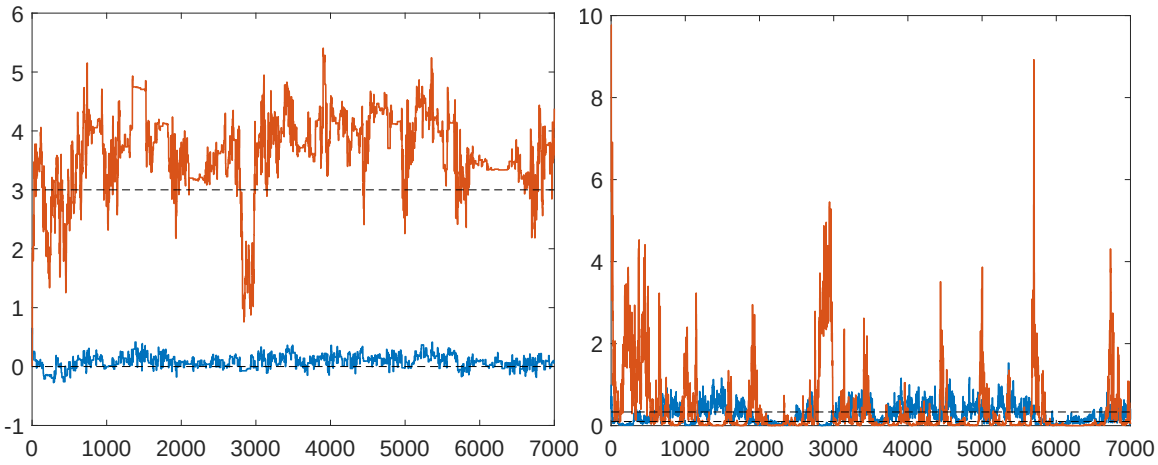

Figure 11: Traces for parameters μ_1 and μ_2 (left) and $1/\lambda_1$ and $1/\lambda_2$ (right) for the **pseudo-marginal slice sampler** with $N = 128$.

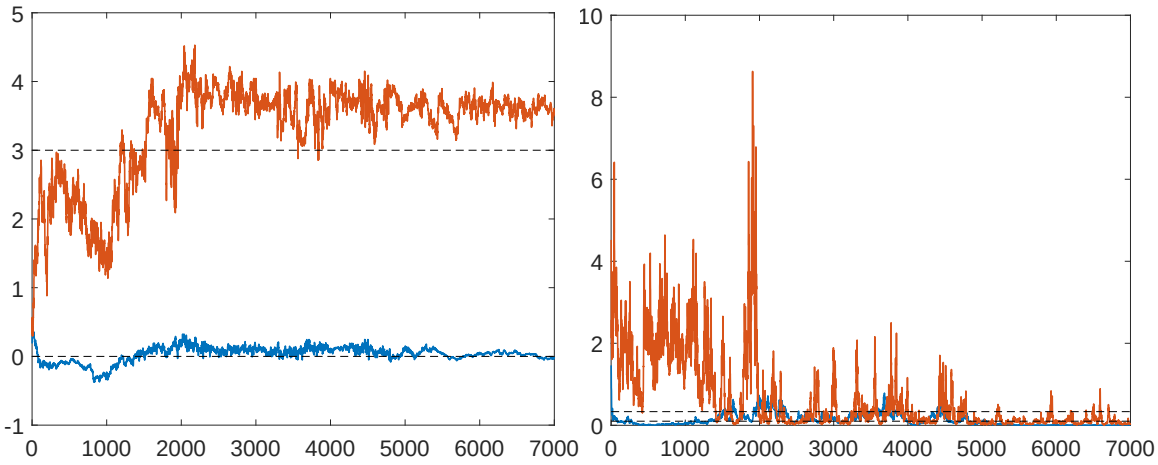

Figure 12: Traces for parameters μ_1 and μ_2 (left) and $1/\lambda_1$ and $1/\lambda_2$ (right) for the **Gibbs sampler** using conditional importance sampling kernels with $N = 128$ particles to update the latent variables.

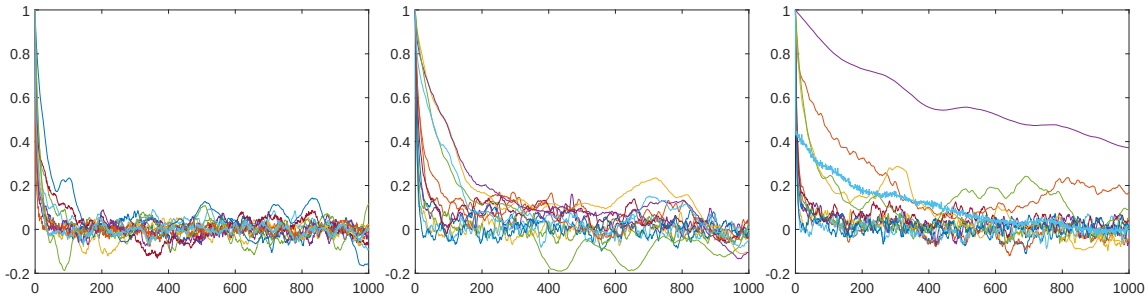

Figure 13: Estimated autocorrelations for the 13 parameters of the model for the **pseudo-marginal HMC sampler** with $N = 128$ (left), **pseudo-marginal slice sampler** with $N = 128$ (mid), and **Gibbs sampler** using conditional importance sampling kernels with $N = 128$ particles to update the latent variables (right).

G.3 Respiratory Infection Data

In this section we give some additional details and results for the random effects model considered in Section 4.4

All methods and runs were initialized at the same point in θ-space, the initial values were

$$\beta^{\text{init}} = \begin{pmatrix} 0.6956 & 0.8695 & 2.2879 & -0.5346 & -0.9756 & -1.8065 & 0.5569 & -0.5209 \end{pmatrix}^{\mathsf{T}},$$

and $\log \tau^{\text{init}} = 1.1049$, while the initial $\mathbf{u}^{\text{init}}$ was randomized for each run.

For the parameters of the different algorithm we set them such that we got something acceptable for the case $N = 21$ and used these values for all runs. It is probably possible to achieve better results for other values of N by tuning the parameters further. In Figure 14—16 show traces for three parameters of the β vector for the different methods and for different values of N. In Figure 17 we see the average time for one iteration of the algorithm in seconds for the different methods for the different values of N used in the simulations, the simulations were done on an i7-6700k running at 4.0GHz with 16GB of memory, no GPU was used to accelerate the computations. Note the different scales on the y-axis, the pseudo-marginal slice-sampling with Metropolis—Hastings is the quickest while the PM-HMC and pseudo-marginal slice-sampling with HMC are the slowest methods.

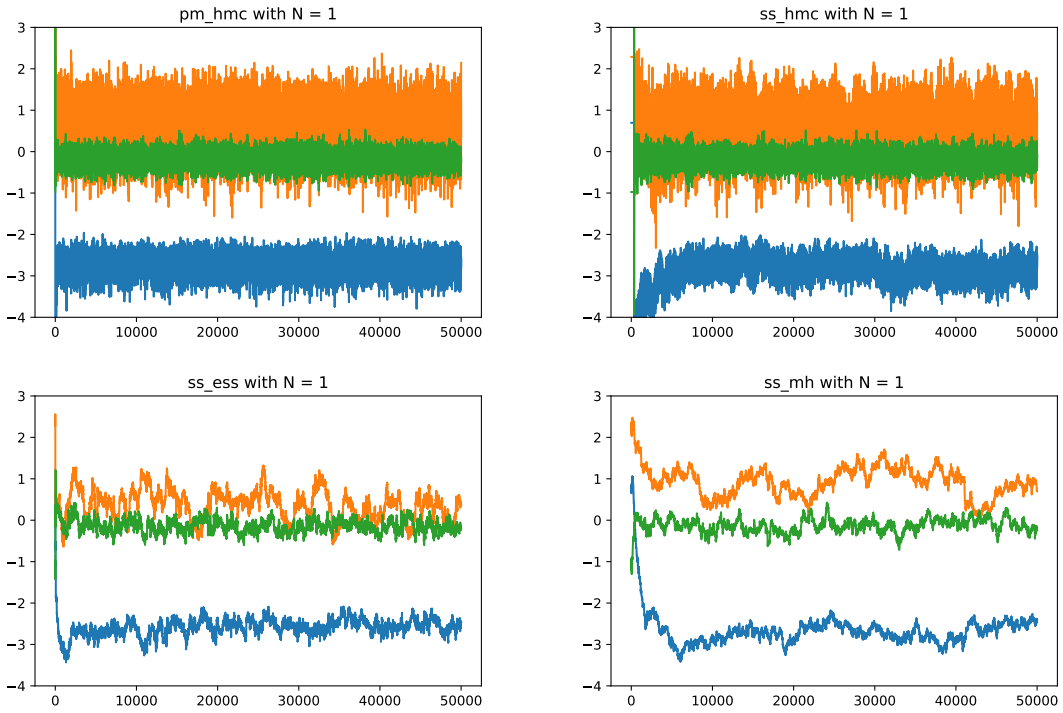

Figure 14: Traces for parameters β_1 (blue), β_3 (orange), and β_5 (green) when $N = 1$ for the **PM-HMC** (top-left) **pseudo-marginal slice sampler** with **HMC** (top-right), **elliptical slice sampling** (bottom-left), and **Metropolis—Hastings** (bottom-right).

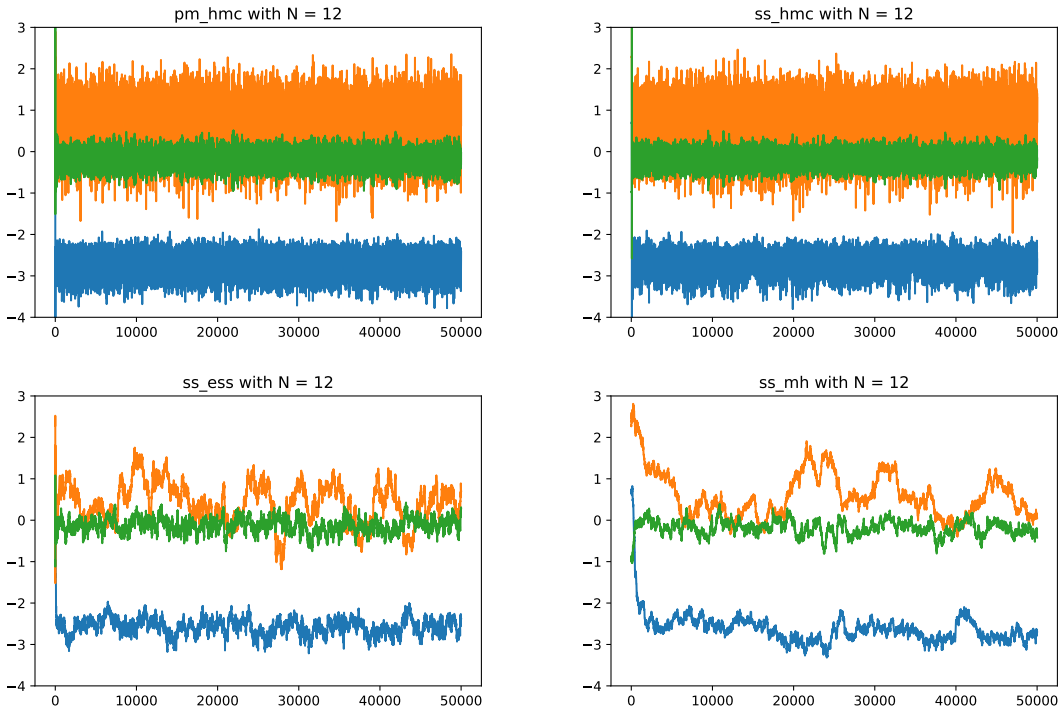

Figure 15: Traces for parameters β_1 (blue), β_3 (orange), and β_5 (green) when $N = 12$ for the **PM-HMC** (top-left) **pseudo-marginal slice sampler** with **HMC** (top-right), **elliptical slice sampling** (bottom-left), and **Metropolis—Hastings** (bottom-right).

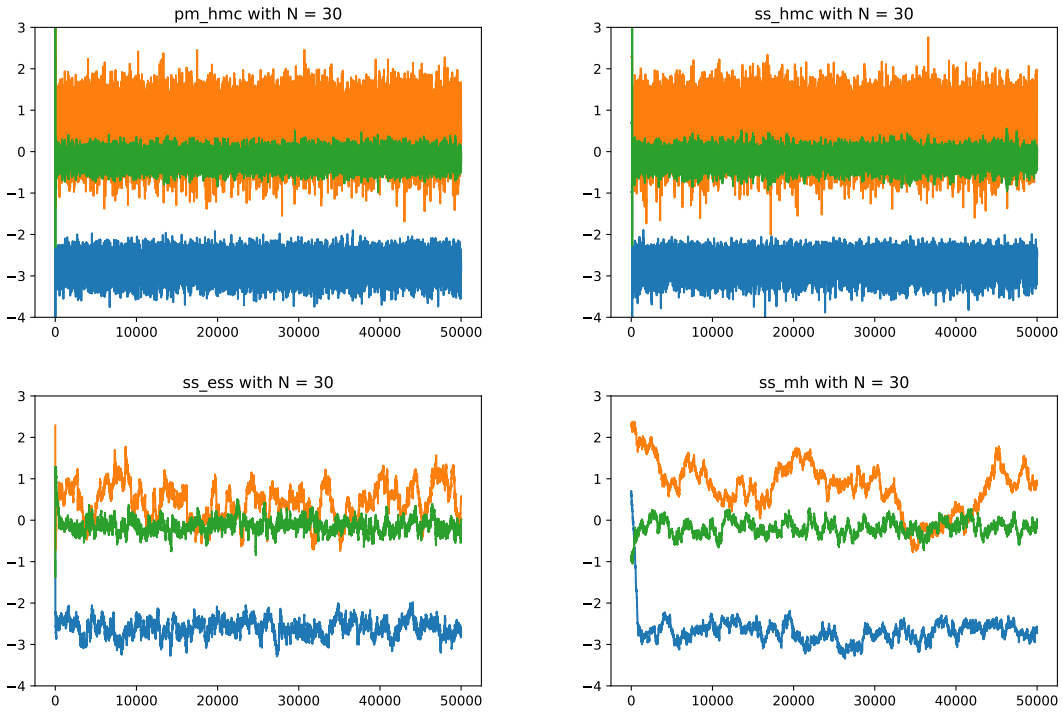

Figure 16: Traces for parameters β_1 (blue), β_3 (orange), and β_5 (green) when $N = 30$ for the **PM-HMC** (top-left) **pseudo-marginal slice sampler** with **HMC** (top-right), **elliptical slice sampling** (bottom-left), and **Metropolis—Hastings** (bottom-right).

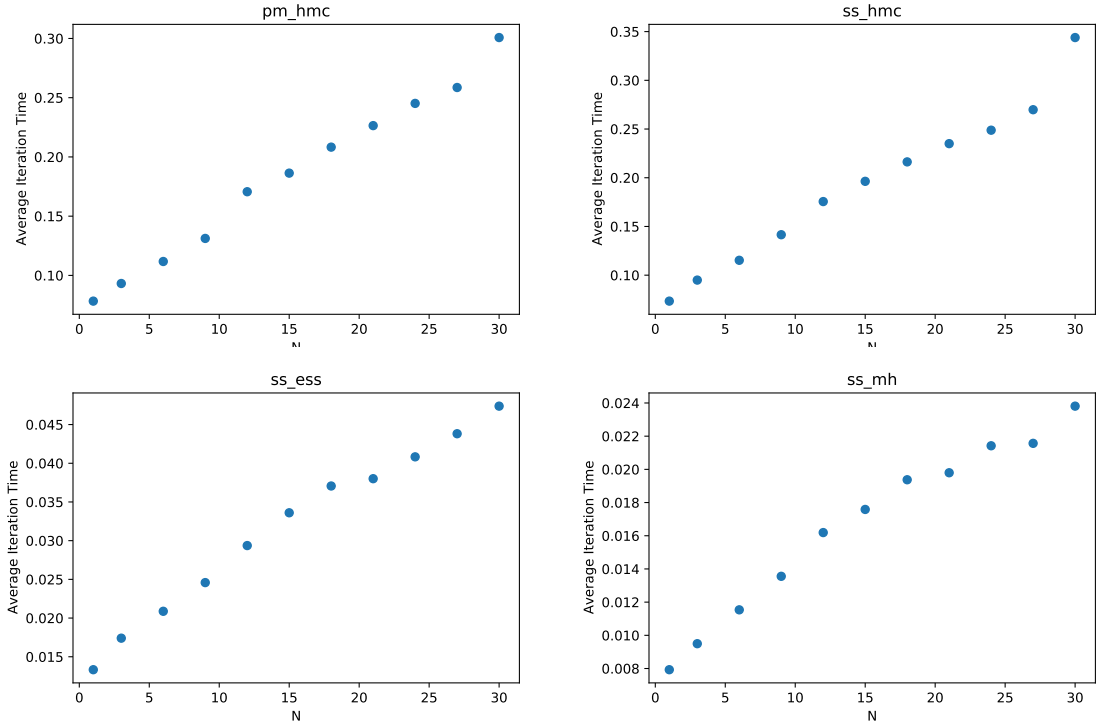

Figure 17: Average iteration time (s) for different values of N for the **PM-HMC** (top-left) **pseudo-marginal slice sampler** with **HMC** (top-right), **elliptical slice sampling** (bottom-left), and **Metropolis—Hastings** (bottom-right).

Appendix H. Auxiliary Results

For the proof of Proposition 5 we need auxiliary results. These results are given here to shorten the proofs above.

Lemma 10 *Let $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ be two sequences of random variables and f a continuous function with continuous and bounded first derivative which satisfies $|f'| < c$. Let h be a continuous bounded function with $|h| < d$ for some constant $d \in \mathbb{R}^+$ and g be a strictly positive continuous bounded function which satisfies $e^{-1} < g < e$ for some $e > 1$. Assume that there exists a random variable Z such that*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} f(X_i) \xrightarrow{d} Z, \quad \text{as } N \to \infty.$$

Then we have that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} f(X_i + \frac{h(Y_i)}{\sum_{j=1}^{N} g(Y_j)}) \xrightarrow{d} Z, \quad \text{as } N \to \infty.$$

Proof For any $\omega \in \Omega$ we have by Taylor's theorem that there exists a

$$\bar{X}_i \in \left[X_i - \frac{h(Y_i)}{\sum_{j=1}^{N} g(Y_j)}, X_i + \frac{h(Y_i)}{\sum_{j=1}^{N} g(Y_j)} \right]$$

such that

$$f(X_i + \frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)}) = f(X_i) + f'(\bar{X}_i)\frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)}.$$

Now we look at

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N f(X_i + \frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N f(X_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^N f'(\bar{X}_i)\frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)},$$

where we use the results above to get the expression on the right hand side. We see that by the assumption the first term converges in distribution to Z while the second term is something that we need to control. From the assumptions we have that

$$-\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{c\cdot d}{\sum_{j=1}^N e^{-1}} < \frac{1}{\sqrt{N}}\sum_{i=1}^N f'(\bar{X}_i)\frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)} < \frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{c\cdot d}{\sum_{j=1}^N e^{-1}}.$$

Since we have that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{c\cdot d}{\sum_{j=1}^N e^{-1}} = \frac{1}{\sqrt{N}} c\cdot d\cdot e \to 0 \quad \text{as } N\to\infty,$$

we have that, by the sandwich property

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N f'(\bar{X}_i)\frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)} \xrightarrow{p} 0, \quad \text{as } N\to\infty.$$

The proof concludes by using Slutsky's lemma. ■

Lemma 11 *Let $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ be two sequences of random variables and f a continuous function with continuous and bounded first derivative which satisfies $|f'| < c$. Let h be a continuous bounded function with $|h| < d$ for some constant $d \in \mathbb{R}^+$ and g be a strictly positive continuous bounded function which satisfies $e^{-1} < g < e$ for some $e > 1$. Assume that there exists a constant μ such that*

$$\frac{1}{N}\sum_{i=1}^N f(X_i) \xrightarrow{p} \mu, \quad \text{as } N\to\infty.$$

Then we have that

$$\frac{1}{N}\sum_{i=1}^N f(X_i + \frac{h(Y_i)}{\sum_{j=1}^N g(Y_j)}) \xrightarrow{p} \mu, \quad \text{as } N\to\infty.$$

The proof of this is analogous to the previous proof and is therefore omitted.

References

C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

M.A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

A. Beskos, F.J. Pinski, J.M. Sanz-Serna, and A.M. Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201–2230, 2011.

M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*, pages 79–101. Chapman & Hall / CRC Press, 2015.

M. Burda, M. Harding, and J. Hausman. A Bayesian mixed logit–probit model for multinomial choice. *Journal of Econometrics*, 147(2):232–246, 2008.

W.L. Chao, J. Solomon, D. Michels, and F. Sha. Exponential integration for Hamiltonian Monte Carlo. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1142–1151, 2015.

T. Chen, E.B. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

K.D. Dang, M. Quiroz, R. Kohn, M.N. Tran, and M. Villani. Hamiltonian Monte Carlo with energy conserving subsampling. *Journal of Machine Learning Research*, 20(100):1–31, 2019.

G. Deligiannidis, A. Doucet, and M.K. Pitt. The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):839–870, 2018.

N. Ding, Y. Fang, R. Babbush, C. Chen, R.D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, pages 3203–3211, 2014.

A. Doucet, M.K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2): 295–313, 2015.

S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

T. Flury and N. Shephard. Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Econometric Theory*, 27(5):933–956, 2011.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (2):123–214, 2011.

M.M. Graham and A.J. Storkey. Asymptotically exact inference in differentiable generative models. *Electronic Journal of Statistics*, 11(2):5105 – 5164, 2017.

P.E. Jacob, F. Lindsten, and T.B. Schön. Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, 115(530):721–729, 2020.

A. Komárek and E. Lesaffre. Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis*, 52 (7):3441–3458, 2008.

B. Leimkuhler and S. Reich. A Metropolis adjusted Nosé-Hoover thermostat. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(4):743–755, 2009.

B. Leimkuhler and X. Shang. Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, 2016.

B.J. Leimkuhler and C. Matthews. *Molecular Dynamics with Deterministic and Stochastic Numerical Methods*. Springer, 2015.

L. Lin, K.F. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Physical Review D*, 61(7): 074505, 2000.

E. Meeds, R. Leenders, and M. Welling. Hamiltonian ABC. In *Proceedings of the 31st conference on Uncertainty in Artificial Intelligence*, pages 582–591, 2015.

I. Murray and M.M. Graham. Pseudo-marginal slice sampling. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 911–919, 2016.

I. Murray, R.P. Adams, and D.J.C. MacKay. Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics*, volume 9, pages 541–548, 2010.

R.M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

R.M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, pages 113–162. Chapman & Hall / CRC Press, 2011.

C. Nemeth, F. Lindsten, M. Filippone, and J. Hensman. Pseudo-extended Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 4314–4324, 2019.

K.K. Osmundsen, T.S. Kleppe, and R. Liesenfeld. Pseudo-marginal Hamiltonian Monte Carlo with efficient importance sampling. *arXiv preprint arXiv:1812.07929*, 2018.

M. Quiroz, M.N. Tran, M. Villani, R. Kohn, and K.D. Dang. The block-Poisson estimator for optimally tuned exact subsampling MCMC. *Journal of Computational and Graphical Statistics*, 2021.

J.F. Richard and W. Zhang. Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385 – 1411, 2007.

S.M. Schmon, G. Deligiannidis, A. Doucet, and M.K. Pitt. Large sample asymptotics of the pseudo-marginal method. *Biometrika*, 108(1):37–51, 2021.

B. Shahbaba, S. Lan, W.O. Johnson, and R.M. Neal. Split hamiltonian monte carlo. *Statistics and Computing*, 24(3):339–349, 2014. doi: 10.1007/s11222-012-9373-1.

X. Shang, Z. Zhu, B. Leimkuhler, and A.J. Storkey. Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling. In *Advances in Neural Information Processing Systems*, pages 37–45, 2015.

C. Sherlock, A.H. Thiery, G.O. Roberts, and J.S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238–275, 2015.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems 28*, pages 955–963, 2015.

K.E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

J. Umenberger, T. Schön, and F. Lindsten. Bayesian identification of state-space models via adaptive thermostats. 2019.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

S.L. Zeger and M.R. Karim. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86, 1991.

Y. Zhao, J. Staudenmayer, B.A. Coull, and M.P. Wand. General design Bayesian generalized linear mixed models. *Statistical Science*, 21(1):35–51, 2006.