

Gibbs flow for approximate transport with applications to Bayesian computation

Jeremy Heng^{*}, Arnaud Doucet⁺ and Yvo Pokern[†]

^{*}ESSEC Business School, Singapore

⁺Department of Statistics, Oxford University, UK

[†]Department of Statistical Science, University College London, UK

Abstract

Let π_0 and π_1 be two distributions on the Borel space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Any measurable function $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $Y = T(X) \sim \pi_1$ if $X \sim \pi_0$ is called a transport map from π_0 to π_1 . For any π_0 and π_1 , if one could obtain an analytical expression for a transport map from π_0 to π_1 , then this could be straightforwardly applied to sample from any distribution. One would map draws from an easy-to-sample distribution π_0 to the target distribution π_1 using this transport map. Although it is usually impossible to obtain an explicit transport map for complex target distributions, we show here how to build a tractable approximation of a novel transport map. This is achieved by moving samples from π_0 using an ordinary differential equation with a velocity field that depends on the full conditional distributions of the target. Even when this ordinary differential equation is time-discretized and the full conditional distributions are numerically approximated, the resulting distribution of mapped samples can be efficiently evaluated and used as a proposal within sequential Monte Carlo samplers. We demonstrate significant gains over state-of-the-art sequential Monte Carlo samplers at a fixed computational complexity on a variety of applications.

Keywords: Mass transport; Markov chain Monte Carlo; Normalizing constants; Path Sampling; Sequential Monte Carlo.

1 Introduction

The use of the Bayesian formalism of inference is ubiquitous in many areas of science. For statistical models of practical interest, implementation usually relies on Monte Carlo methods to sample from the posterior distribution which might be high dimensional and exhibit complex dependencies. Most available Monte Carlo algorithms rely on proposal distributions and the efficiency of these techniques is crucially dependent on whether these proposals are able to capture important features of the target. In this paper, we leverage ideas from the mass transport literature to develop a new methodology to build efficient proposal distributions which can be used within sequential Monte Carlo (SMC) samplers [41, 12, 20].

Given initial and target distributions π_0 and π_1 defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, which in a Bayesian context may be interpreted as the prior and posterior, a transport map is a measurable function $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $Y = T(X) \sim \pi_1$ if $X \sim \pi_0$. The transport map terminology arises from the fact that one can view T as transporting the probability mass represented by π_0 to the probability mass represented by π_1 . We will use the notation $\pi_1 = (T)_\# \pi_0$ since π_1 is the push-forward measure of π_0 by T . Characterizing the existence of transport maps has generated a large literature in mathematics; see

[58] for a recent review. In particular, much work has been dedicated to the L^2 Monge-Kantorovich problem, where one seeks the optimal transport map T minimizing the expected cost $\mathbb{E}|T(X) - X|^2$.

For the purposes of Monte Carlo simulation, any analytically tractable transport map would allow us to map samples from π_0 to π_1 . However, even without imposing any optimality condition, such transport maps have only been identified in simple scenarios; e.g. when both π_0 and π_1 are Gaussian [46, Remark 2.30]. To obtain an approximate transport map, [37, 22, 44] proposed to minimize some measure of discrepancy between $(T_\beta)_\# \pi_0$ and π_1 , over a set of maps parametrized by a finite-dimensional parameter β , e.g. a linear combination of some basis functions. However, it can be difficult to identify an appropriate subspace of candidate maps, and the resulting optimization problem is generally non-convex unless stringent assumptions are made [31, 45] and high dimensional in the absence of conditional independence structure in the target π_1 [52]. In this article, we circumvent these difficulties by considering a different approach to build approximate transport maps.

The transport maps we will consider are derived from a fluid dynamics interpretation of mass transport. Consider a curve of distributions $\{\pi_t\}_{t \in (0,1)}$ connecting π_0 to π_1 ; e.g. the geometric path $\pi_t \propto \pi_0^{1-\lambda(t)} \pi_1^{\lambda(t)}$ where $\lambda : [0,1] \rightarrow [0,1]$ is an increasing smooth function satisfying $\lambda(0) = 0$ and $\lambda(1) = 1$. The use of bridging distributions between distant π_0 and π_1 is at the core of many state-of-the-art Monte Carlo methods such as path sampling [25, 43] and annealed importance sampling [16, 30, 41, 12]. If we view probability mass as an infinite ensemble of fluid particles, the main idea is to move these particles deterministically, using an ordinary differential equation (ODE) with a carefully designed velocity field, so as to mimic the time evolution of π_t over the time interval $t \in [0,1]$. Loosely speaking, we may think of the movement of particles under such a velocity field as implicitly defining flow transport maps $\{T_t\}_{t \in [0,1]}$ satisfying $\pi_t = (T_t)_\# \pi_0$ for each $t \in [0,1]$.

The idea of constructing transport maps using flows originates from [40]; see also [27, 17, 5] for other early contributions. This approach has since been adopted in application domains ranging from engineering to physics [4, 15, 19, 53, 56]. Noting that, for a given curve of distributions, there could be multiple velocity fields achieving the flow transport, various optimality criteria have been introduced to identify a unique solution [40, 47, 53]; e.g. [47] proposed seeking the velocity field minimizing kinetic energy. In these contributions, the optimal velocity field is given by the solution of an elliptic partial differential equation (PDE). However, when using a full grid, PDE solvers suffer from the curse of dimensionality [18, 42] which could render them impractical. Sparse grid methods may be capable of dealing with sufficiently high dimensions but they come with their own set of approximations, e.g. tensor approximations [11, 18]. Using techniques from differential geometry, [7] constructed a flow transport using contact Hamiltonian flows that also determines λ adaptively, but the velocity field depends on intractable integrals on \mathbb{R}^d which would have to be numerically approximated.

An alternative approach involves building analytically tractable approximations of intractable flow transport maps. For example, in a Bayesian filtering context where π_0 is a Gaussian prior distribution on unknown states X , and the likelihood is also Gaussian distributed with mean vector $\phi(X)$ and a known covariance matrix, [9] proposed linearizing ϕ locally to exploit analytical tractability of Gaussian flows [6, 48, 51]. This article also proposes approximate flow transport maps that are analytically tractable, but the details of our construction are markedly different. Our approach does not require any distributional assumptions on π_0 and π_1 , instead it is based on approximating a novel flow that takes reference to the conditional distributions $\pi_t(x_1|x_2, \dots, x_d)$, $\pi_t(x_2|x_3, \dots, x_d)$, ..., $\pi_t(x_{d-1}|x_d)$ and the marginal distribution $\pi_t(x_d)$ where $x_i \in \mathbb{R}$ for $i = 1, \dots, d$. As these distributions are typically intractable, we propose a tractable approximation which moves particles using a velocity field designed to track the full conditional distributions $\{\pi_t(x_i|x_{-i})\}_{i=1,\dots,p}$, where $x_i \in \mathbb{R}^{d_i}$ denotes the $i = 1, \dots, p$ component of dimension $d_i \in \mathbb{N}$ with $\sum_{i=1}^p d_i$ and $x_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p)$. We shall refer to the latter as the Gibbs flow in reference to the Gibbs sampler. Contrary to existing transport-based methods, Gibbs flow does not require selecting a parametric class of maps, solving a non-convex optimization problem, approximating the solution of a PDE or approximating d -dimensional integrals. Analogous to Gibbs samplers, its implementation allows one to leverage any conditional independence structure in the target π_1 and analytical tractability of any full conditional distribution to move

the corresponding component. For components with intractable Gibbs flow, we will show that by further blocking these components into one-dimensional components, the resulting Gibbs velocity field only involves one-dimensional integrals w.r.t. the corresponding full conditional distributions that can be efficiently approximated using most quadrature routines. We will also introduce a novel time discretization scheme reminiscent of the systematic scan Gibbs sampler to numerically integrate the Gibbs flow. Although other numerical integrators can also be considered, our scheme allows efficient computation of the distribution of resulting mapped samples in high dimensions, which is crucial when employing such distributions as proposals within SMC samplers. Our approach only requires a computational cost of $O(\sum_{i=1}^p d_i^3)$ at each time step without requiring additional approximations to reduce the computational complexity [28]. We establish various theoretical properties of the Gibbs flow and demonstrate significant gains over state-of-the-art methods at a fixed computational complexity on a variety of applications.

The rest of the paper is organized as follows. In Section 2, we introduce the construction of transport maps using flows in a Bayesian context. We present a novel flow transport, the Gibbs flow approximation and its properties in Section 3. We then discuss how the Gibbs flow can be numerically implemented and employed as proposal distributions within SMC samplers in Section 4. Lastly, in Section 5, we illustrate the proposed methodology on a mixture model, a variance component model, and a log-Gaussian Cox point process model. The proofs of all results are given in the Supplementary Material. An R package is available at github.com/jeremyhengjm/GibbsFlow to reproduce all numerical results.

2 Transport with flows

2.1 A curve from prior to posterior

Let $\pi_0(dx)$ be a prior distribution on the Borel space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $L : \mathbb{R}^d \rightarrow \mathbb{R}_+$ denote a likelihood function. To simplify presentation, we shall assume that $\pi_0(dx)$ is absolutely continuous w.r.t. the Lebesgue measure on \mathbb{R}^d , with an everywhere positive density $x \mapsto \pi_0(x)$, and that $x \mapsto L(x)$ is also positive everywhere and satisfies $\lim_{|x| \rightarrow \infty} L(x) = 0$. We will defer a discussion of improper priors to Section 5.2 and suppress all notational dependencies on observations. From Bayes' rule, the resulting posterior distribution $\pi(dx)$ admits the density

$$\pi(x) = \frac{\pi_0(x)L(x)}{Z}, \quad (1)$$

where $Z = \int_{\mathbb{R}^d} \pi_0(u)L(u) du$ denotes the marginal likelihood. Henceforth we shall additionally assume that $\pi_0, L \in C^1(\mathbb{R}^d, \mathbb{R}_+)$, where $C^k(A, B)$ denotes the set of functions from A to B which are k -times continuously differentiable.

We introduce a curve of distributions $\{\pi_t\}_{t \in [0,1]}$ smoothly bridging the prior π_0 to the posterior $\pi_1 = \pi$ by gradually introducing the likelihood using a strictly increasing C^2 -function $\lambda : [0, 1] \rightarrow [0, 1]$ such that $\lambda(0) = 0$ and $\lambda(1) = 1$:

$$\pi_t(x) = \frac{\gamma_t(x)}{Z(t)}, \quad \gamma_t(x) = \pi_0(x)L(x)^{\lambda(t)}, \quad (2)$$

where $Z(t) = \int_{\mathbb{R}^d} \gamma_t(u) du$. The function λ is commonly known as inverse temperature in the context of simulated annealing for optimization problems [32]. By differentiating (2) w.r.t. the time variable t , we obtain its time evolution along the curve

$$\partial_t \pi_t(x) = \lambda'(t) (\log L(x) - I_t) \pi_t(x), \quad (3)$$

where $\lambda' : [0, 1] \rightarrow \mathbb{R}_+$ denotes the time derivative of λ and

$$I_t = \frac{1}{\lambda'(t)} \frac{d}{dt} \log Z(t) = \frac{\frac{d}{dt} \int_{\mathbb{R}^d} \pi_0(u)L(u)^{\lambda(t)} du}{\lambda'(t)Z(t)} = \int_{\mathbb{R}^d} \log L(u) \pi_t(u) du \quad (4)$$

is assumed to be finite for all $t \in [0, 1]$. Under our assumptions, the family of models $\{\pi_t\}_{t \in [0, 1]}$ is regular so interchanging the order of differentiation w.r.t. the time variable and integration w.r.t. the spatial variable in the last equality of (4) is valid. By integrating (4) on the time interval $[0, 1]$, we recover the well-known path sampling identity [25, 43]:

$$\log Z = \int_0^1 \lambda'(t) I_t dt. \quad (5)$$

Equation (3) reveals that the expected log-likelihood I_t plays the role of a reference value which controls the evolution of the density $\pi_t(x)$, i.e. in logarithmic scale, the local behaviour around a point $x \in \mathbb{R}^d$ is such that there is an increase or decrease in density if $\log L(x) > I_t$ or $\log L(x) < I_t$, respectively. In the following, we will see that this difference, when integrated w.r.t. $\pi_t(x)$, provides us with the right direction to move particles at time t . The factors $\lambda'(t)$ and $\pi_t(x)$ in (3) are also intuitive as the change in density must be proportional to how quickly we introduce the likelihood and how much probability mass there is locally. It will be apparent later that these factors dictate the speed of particles. We note that the contact Hamiltonian flow proposed in [7] also depends on the term $\log L(x) - I_t$ which the author therein approximates using Monte Carlo methods.

2.2 Particle dynamics, Liouville's equation and flow transport problem

Consider a particle trajectory $\{X_t\}_{t \in [0, 1]}$ in \mathbb{R}^d , initialized at time $t = 0$ with a random draw $X_0 \sim \pi_0$, and evolved deterministically according to the following ODE

$$\frac{d}{dt}x(t) = f(t, x(t)), \quad t \in [0, 1], \quad (6)$$

with velocity field $f = (f_1, \dots, f_d) : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$. Under appropriate regularity conditions on f which will be detailed later, this ODE admits a unique solution $x(t; X_0)$ for all $t \in [0, 1]$. Therefore we can define the flow map $T_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ as

$$X_t = T_t(X_0) = x(t; X_0) \quad (7)$$

which associates the initial position of the particle to its position at time $t \in [0, 1]$. It can be shown that flow maps are C^1 -diffeomorphisms, i.e. for each $t \in [0, 1]$, T_t is invertible and both T_t and its inverse $T_t^{-1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ are continuously differentiable. These properties render flow maps ideal candidates as transport maps.

Additionally, if we denote the marginal distribution of X_t by $\tilde{\pi}_t = (T_t)_\# \pi_0$, the curve of distributions $\{\tilde{\pi}_t\}_{t \in [0, 1]}$ satisfies, under regularity conditions, the Liouville PDE [23, eq. (3.5.13), p. 54] also known as the continuity equation [2, eq. (8.1.1), p. 169]:

$$\partial_t \tilde{\pi}_t(x) = - \sum_{i=1}^d \partial_{x_i} (\tilde{\pi}_t(x) f_i(t, x)) = - \nabla \cdot (\tilde{\pi}_t(x) f(t, x)) \quad (8)$$

for $(t, x) \in (0, 1) \times \mathbb{R}^d$. Notationally, $\partial_t \varphi(t, x)$ and $\partial_{x_i} \varphi(t, x)$ denote the partial derivatives of $\varphi \in C^1([0, 1] \times \mathbb{R}^d, \mathbb{R})$ w.r.t. t and x_i , respectively, and the divergence operator is defined as $\nabla \cdot \varphi(x) = \sum_{i=1}^d \partial_{x_i} \varphi_i(x)$ for any $\varphi = (\varphi_1, \dots, \varphi_d) \in C^1(\mathbb{R}^d, \mathbb{R}^d)$. The Liouville PDE can be seen as the Fokker-Planck equation in the case of zero diffusivity; an informal but intuitive derivation of this PDE is given in Supplementary Material A.

We can now describe the flow transport problem as identifying a velocity field f such that the curve of target distributions $\{\pi_t\}_{t \in [0, 1]}$ in (2) is the solution of the Liouville PDE (8), i.e. we seek a f that satisfies

$$\partial_t \pi_t(x) = - \nabla \cdot (\pi_t(x) f(t, x)) \quad (9)$$

for $(t, x) \in (0, 1) \times \mathbb{R}^d$. If such a velocity field f is regular enough that the resulting ODE (6) admits a unique solution for all $t \in [0, 1]$ and initial positions $X_0 \sim \pi_0$, then this allows us to construct the flow

maps (7) that satisfy $\pi_t = (T_t)_\# \pi_0$ for all $t \in [0, 1]$. As a consequence, we can obtain samples from $\pi_1 = \pi$ by taking $X_1 = T_1(X_0)$. The following result presents sufficient conditions on velocity fields f that satisfy (9) to ensure the validity of this approach.

Theorem 1. *Suppose $f : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a velocity field that satisfies Liouville equation (9) and the following conditions:*

A1. *(continuously differentiable) $f \in C^1([0, 1] \times \mathbb{R}^d, \mathbb{R}^d)$;*

A2. *(space-time integrability) $\int_0^1 \int_{\mathbb{R}^d} |f(t, x)| \pi_t(x) dx dt < \infty$.*

Then for π_0 -almost every $X_0 \in \mathbb{R}^d$, there exists a unique solution $x(t, X_0)$ to the ODE (6) for all $t \in [0, 1]$. Therefore the flow maps $\{T_t\}_{t \in [0, 1]}$ defined by (7) are flow transports, i.e. $\pi_t = (T_t)_\# \pi_0$ for all $t \in [0, 1]$.

Theorem 1 is a summary of results in [2] written for our purposes; see Supplementary Material B for more details. With Theorem 1 in place, we can now formally define the flow transport problem as identifying a velocity field that satisfies Liouville's equation (9) and Assumptions A1-A2. Although these assumptions are only sufficient conditions, we stress that pathologies can occur when these regularity conditions do not hold. This is illustrated in Supplementary Material G.3, where we exhibit a velocity field that solves (9) and prove that it yields divergent particle trajectories.

3 A novel flow transport and Gibbs flow approximation

As alluded to in the introduction, the flow transport problem is typically underdetermined. Although various optimality criteria could be employed to attain uniqueness, they lead to velocity fields that are implicitly defined by solutions of elliptic PDEs. In this section, we begin by presenting an explicit solution to the flow transport problem before introducing the Gibbs flow approximation.

3.1 A flow transport solution on \mathbb{R}

We first discuss the one-dimensional case before considering the multivariate case. In this case, there is a well-known solution to the flow transport problem; see e.g. [4]. We will also establish that this coincides with the minimal kinetic energy solution considered in [47, 48].

Proposition 1. *Define the velocity field $f : [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$ as*

$$f(t, x) = \frac{-\int_{-\infty}^x \partial_t \pi_t(u) du}{\pi_t(x)} \quad (10)$$

and assume that there exists an $\epsilon > 0$ such that $x \mapsto |f(t, x)| \pi_t(x) = O(|x|^{-1-\epsilon})$ as $|x| \rightarrow \infty$ with a constant that is independent of $t \in [0, 1]$. Then the velocity field (10) solves the flow transport problem on \mathbb{R} and is additionally the minimal kinetic energy solution, i.e. for each $t \in [0, 1]$

$$f(t, \cdot) = \arg \min_{\varphi \in \mathcal{L}(\pi_t)} \frac{1}{2} \int_{\mathbb{R}^d} \varphi^2(x) \pi_t(x) dx, \quad (11)$$

where $\mathcal{L}(\pi_t) = \{\varphi : \mathbb{R} \rightarrow \mathbb{R} : \int_{\mathbb{R}} \varphi^2(x) \pi_t(x) dx < \infty, \varphi(x) \text{ satisfies (9) for all } x \in \mathbb{R} \text{ at } t \in [0, 1]\}$.

To build intuition, we can rewrite (10) using (3) as

$$f(t, x) = \frac{\lambda'(t) I_t (F_t(x) - I_t^x / I_t)}{\pi_t(x)}, \quad (12)$$

where $I_t^x = \int_{-\infty}^x \log L(u) \pi_t(u) du$ and $F_t(x) = \int_{-\infty}^x \pi_t(u) du$ is the cumulative distribution function (CDF) of π_t . The velocity field (12) may be likened to driving a vehicle. The denominator corresponds to the *accelerator*, since, e.g., particles in the tails of π_t need to *speed up* to meet the changing schedule of intermediate distributions. Also, it is intuitive that particle speeds are proportional to the rate $\lambda'(t)$ at which we introduce the likelihood. The numerator amounts to the *steering wheel*: a particle's direction of travel is given by the relative difference between its *current location* x , described by the term $F_t(x)$, and *where the particle needs to go*, prescribed by the term $I_t^x/I_t \in [0, 1]$ which contains information from the likelihood.

3.2 A novel flow transport on \mathbb{R}^d , $d \geq 1$

It is tempting to extend (10) to the multivariate case by simply introducing the velocity field $\bar{f} = (\bar{f}_1, \dots, \bar{f}_d) : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ given for $i = 1, \dots, d$ by

$$\bar{f}_i(t, x) = \frac{-\alpha_i \int_{-\infty}^{x_i} \partial_t \pi_t(u_i, x_{-i}) du_i}{\pi_t(x)}, \quad (13)$$

where $\alpha_i \in \mathbb{R}$ and the integrand of (13) is to be understood as $\partial_t \pi_t(x_1, \dots, x_{i-1}, u_i, x_{i+1}, \dots, x_d)$. This velocity field was previously mentioned in [4] and it can be shown to satisfy Liouville's equation (9) whenever $\sum_{i=1}^d \alpha_i = 1$. However, we show in Supplementary Material G.3 that (13) does not solve the flow transport problem as an ODE with velocity field \bar{f} would yield divergent particle trajectories even on a simple Gaussian example. The main reason for this pathology is the tail behaviour of \bar{f} .

We now give our solution to the flow transport problem in the multivariate case which recovers Proposition 1 when $d = 1$. We will write $x_{i:j} = (x_i, \dots, x_j) \in \mathbb{R}^{j-i+1}$ and denote the marginal distribution of π_t in the $i = 1, \dots, d$ component by $\pi_t(x_i)$ and its CDF by $F_t(x_i) = \int_{-\infty}^{x_i} \pi_t(u_i) du_i$.

Proposition 2. Define the velocity field $f : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ as

$$f_i(t, x) = - \left(\prod_{j=1}^{i-1} \pi_t(x_j) \int_{-\infty}^{x_i} \int_{\mathbb{R}^{i-1}} \partial_t \pi_t(u_{1:i-1}, u_i, x_{i+1:d}) du_{1:i-1} du_i - \prod_{j=1}^{i-1} \pi_t(x_j) F_t(x_i) \int_{\mathbb{R}^i} \partial_t \pi_t(u_{1:i}, x_{i+1:d}) du_{1:i} \right) / \pi_t(x) \quad (14)$$

for $i = 1, \dots, d-1$ (use the convention $\prod_1^0 = 1$) and

$$f_d(t, x) = - \left(\prod_{j=1}^{d-1} \pi_t(x_j) \int_{-\infty}^{x_d} \int_{\mathbb{R}^{d-1}} \partial_t \pi_t(u_{1:d-1}, u_d) du_{1:d-1} du_d \right) / \pi_t(x). \quad (15)$$

If there exists an $\epsilon > 0$ such that $\sup_{\{x \in \mathbb{R}^d : |x| = r\}} |f(t, x)| \pi_t(x) = O(r^{-d-\epsilon})$ as $r \rightarrow \infty$ with a constant that is independent of $t \in [0, 1]$, then the velocity field (14)-(15) solves the flow transport problem on \mathbb{R}^d .

Our construction is a generalization of a method proposed by [8] to build a compactly supported three-dimensional velocity field solving a flow transport problem in the context of molecular quantum chemistry. When the target distributions factorize into independent one-dimensional components, i.e. $\pi_t(x) = \prod_{i=1}^d \pi_t(x_i)$, we establish in Supplementary Material C that the velocity field in (14)-(15) would simply reduce to

$$f_i(t, x_i) = \frac{-\int_{-\infty}^{x_i} \partial_t \pi_t(u_i) du_i}{\pi_t(x_i)}, \quad i = 1, \dots, d, \quad (16)$$

which is the solution of the one-dimensional flow transport problem for each marginal distribution given by Proposition 1. As the integrals in (14)-(15) can be seen as expectations w.r.t. the conditional

distributions $\pi_t(x_1|x_2, \dots, x_d)$, $\pi_t(x_2|x_3, \dots, x_d)$, ..., $\pi_t(x_{d-1}|x_d)$ and the marginal distribution $\pi_t(x_d)$, we see that the flow transport is achieved by taking reference to these conditionals. We refer the reader to Supplementary Material [G](#) for an illustration of flow transport solutions when the curve of distributions [\(2\)](#) lies in the Gaussian family.

3.3 Gibbs flow approximation

Despite the explicit form of the flow transport solution in Proposition [2](#), evaluating the velocity field [\(14\)](#)-[\(15\)](#) would require computing integrals of dimension up to d . For computational tractability, we propose an approximate flow transport that takes reference to the full conditional distributions $\{\pi_t(x_i|x_{-i})\}_{i=1, \dots, p}$, where $x_i \in \mathbb{R}^{d_i}$ and $\sum_{i=1}^p d_i = d$. For component $i = 1, \dots, p$, the time evolution of its full conditional distribution is given by

$$\partial_t \pi_t(x_i|x_{-i}) = \lambda'(t)(\log L(x) - I_t(x_{-i}))\pi_t(x_i|x_{-i}) \quad (17)$$

where $I_t(x_{-i}) = \int_{\mathbb{R}^{d_i}} \log L(u_i, x_{-i}) \pi_t(u_i|x_{-i}) du_i$. We will design Gibbs velocity fields $\tilde{f} = (\tilde{f}_1, \dots, \tilde{f}_p) : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ that track changes in the full conditionals [\(17\)](#) by seeking solutions to the following coupled system of Liouville equations

$$\partial_t \pi_t(x_i|x_{-i}) = -\nabla_{x_i} \cdot (\pi_t(x_i|x_{-i}) \tilde{f}_i(t, x)), \quad i = 1, \dots, p, \quad (18)$$

for $t \in (0, 1)$ and $x = (x_1, \dots, x_p) \in \mathbb{R}^d$. Note that the Liouville equation for each full conditional distribution in [\(18\)](#) is defined on $(0, 1) \times \mathbb{R}^{d_i}$, so the divergence operator $\nabla_{x_i} \cdot \varphi(x_i)$, defined for any $\varphi \in C^1(\mathbb{R}^{d_i}, \mathbb{R}^{d_i})$, only acts on the variables $x_i \in \mathbb{R}^{d_i}$.

For one-dimensional components, i.e. the case $d_i = 1$, the velocity field

$$\tilde{f}_i(t, x) = \frac{-\int_{-\infty}^{x_i} \partial_t \pi_t(u_i|x_{-i}) du_i}{\pi_t(x_i|x_{-i})}, \quad (19)$$

which only involves one-dimensional integrals, can be shown to satisfy [\(18\)](#) for the i^{th} -component, using similar arguments as in Proposition [1](#). For components with dimension $d_i > 1$, one could exploit analytical tractability of full conditional distributions when they lie in the exponential family to determine Gibbs velocity fields, or further block these components into one-dimensional components and employ [\(19\)](#). We will illustrate how to systematically determine Gibbs velocity fields on specific applications in Section [5](#), and will assume for now that we have such a velocity field \tilde{f} satisfying [\(18\)](#). The following result presents sufficient conditions on a Gibbs velocity field \tilde{f} to ensure that, with initial position $X_0 \sim \pi_0$, the ODE

$$\frac{d}{dt} x(t) = \tilde{f}(t, x(t)) \quad (20)$$

admits a unique solution for all $t \in [0, 1]$.

Proposition 3. *Suppose $\tilde{f} : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a velocity field that satisfies the system of Liouville equations [\(18\)](#), Assumption [A3](#) and [A4\(i\)](#) or [A4\(ii\)](#):*

A3. *(continuously differentiable) $\tilde{f} \in C^1([0, 1] \times \mathbb{R}^d, \mathbb{R}^d)$;*

A4(i). *(tail behaviour) there exists $V \in C^1(\mathbb{R}^d, \mathbb{R})$ satisfying $\lim_{|x| \rightarrow \infty} V(x) = \infty$ and $R > 0$ such that $\langle \nabla V(x), \tilde{f}(t, x) \rangle \leq 0$ for all $|x| > R$ and $t \in [0, 1]$, where $\langle \cdot, \cdot \rangle$ denotes the dot product;*

A4(ii). *(global Lipschitz bound) there exists a globally Lipschitz function $g : \mathbb{R}^d \rightarrow \mathbb{R}_+$ such that $\sup_{t \in [0, 1]} |\tilde{f}(t, x)| \leq g(x)$ for all $x \in \mathbb{R}^d$.*

Then for π_0 -almost every $X_0 \in \mathbb{R}^d$, there exists a unique solution $x(t, X_0)$ to the ODE [\(20\)](#) for all $t \in [0, 1]$.

Assumption A3 imposes some regularity on the Gibbs velocity field. Assumption A4(i) requires existence of a Lyapunov function $V \in C^1(\mathbb{R}^d, \mathbb{R})$ so that a particle has non-increasing values of V if it lies in the tails. In some cases, one can choose $V(x) = |x|^2$ as a Lyapunov function; we establish this in the $d = 1$ case in Supplementary Material D. Assumption A4(ii) offers an alternative to finding Lyapunov functions by assuming a globally Lipschitz upper bound on the norm of the Gibbs velocity field; this is verified on a multivariate Gaussian model in Supplementary Material G.6. Under the conclusions of Proposition 3, we can define the Gibbs flow map $\tilde{T}_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ as $X_t = \tilde{T}_t(X_0) = x(t; X_0)$ for each $t \in [0, 1]$. Since the system (18) is only a (tractable) approximation of the desired Liouville equation (9), the marginal distribution of X_t under the Gibbs flow, $\tilde{\pi}_t = (\tilde{T}_t)_\# \pi_0$, will in general not be equal to the target distribution π_t . We now provide a characterization of this error in terms of the following time-dependent local error:

$$\varepsilon_t(x) = \left| \partial_t \pi_t(x) + \nabla \cdot (\pi_t(x) \tilde{f}(t, x)) \right| = \left| \partial_t \pi_t(x) - \sum_{i=1}^p \partial_t \pi_t(x_i | x_{-i}) \pi_t(x_{-i}) \right| \quad (21)$$

which measures how well the Gibbs velocity field mimics the desired change in density (3). The sum over all components in (21) reveals the nature of the Gibbs flow approximation: information about how much probability mass is changing in a particular component is not communicated to other components. In other words, computational tractability is gained at the expense of breaking down a global problem in d dimensions to p many lower dimensional problems. For any function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$, we will write $\|\varphi\|_{L^2}^2 = \int_{\mathbb{R}^d} \varphi^2(x) dx$ if φ is L^2 -integrable, and $\|\varphi\|_\infty = \sup_{x \in \mathbb{R}^d} |\varphi(x)|$ if φ is bounded.

Proposition 4. *Suppose $\tilde{f} : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a velocity field that satisfies the system of Liouville equations (18), Assumptions A3, A4(i) or A4(ii) and*

A5. *(tail decay) there exists $\epsilon > 0$ such that*

$$\sup_{\{x \in \mathbb{R}^d : |x|=r\}} |\tilde{f}(t, x)| \pi_t(x) = O(r^{-d-\epsilon}) \quad \text{and} \quad \sup_{\{x \in \mathbb{R}^d : |x|=r\}} |\tilde{f}(t, x)| \tilde{\pi}_t(x) = O(r^{-d-\epsilon})$$

as $r \rightarrow \infty$ with constants that are independent of $t \in [0, 1]$.

Then the Gibbs flow approximation error is characterized by the following inequality

$$\|\tilde{\pi}_t - \pi_t\|_{L^2}^2 \leq C(t) \int_0^t \|\varepsilon_s\|_{L^2}^2 ds \quad (22)$$

for $t \in [0, 1]$, where $C(t) = t \exp\left(1 + \int_0^t \|\nabla \cdot \tilde{f}(s, \cdot)\|_\infty ds\right)$.

The upper bound (22) is tight in the sense that it is equal to zero when the target distributions have independent components, i.e. $\pi_t(x) = \prod_{i=1}^p \pi_t(x_i)$. When the latter is not the case, we observe that the bound deteriorates with time, which is to be expected as errors can accumulate. To mitigate accumulation of errors, we will combine Gibbs flow with Markov chain Monte Carlo moves in Section 4.3. Rewriting (21) using (3) and (17) reveals that the inverse temperature $\lambda(t)$ should be chosen such that its derivative $\lambda'(t)$ is small at those time instances when the integrated local error $\|\varepsilon_t\|_{L^2}^2$ is large, as this would reduce the magnitude of the resulting L^2 -error in (22). We refer the reader to [25, 43, 60] for other works on how to select $\lambda(t)$. For simplicity, all simulations in Section 5 and the Supplementary Material will employ a quadratic inverse temperature function, i.e. $\lambda(t) = t^2$. Lastly, like with any Gibbs sampler, we expect the use of any appropriate model specific reparameterization to also reduce the L^2 -error in (22).

4 Gibbs flow samplers

4.1 Numerical implementation

Given a target distribution of interest, we advocate exploiting any analytical tractability of full conditional distributions to determine a Gibbs velocity field (e.g. Section 5.2). For components without

such tractability, a generic strategy would be to further block these components into one-dimensional components and rely on (19). We first note that (19) can be computed solely using one-dimensional integrals as the intractable normalizing constant $Z(t)$ cancels in the expression:

$$\tilde{f}_i(t, x) = \frac{\lambda'(t) \left\{ F_t(x_i|x_{-i}) \int_{-\infty}^{\infty} \log L(u_i, x_{-i}) \gamma_t(u_i, x_{-i}) du_i - \int_{-\infty}^{x_i} \log L(u_i, x_{-i}) \gamma_t(u_i, x_{-i}) du_i \right\}}{\gamma_t(x)} \quad (23)$$

and the CDF of $\pi_t(x_i|x_{-i})$ can be rewritten as

$$F_t(x_i|x_{-i}) = \frac{\int_{-\infty}^{x_i} \gamma_t(u_i, x_{-i}) du_i}{\int_{-\infty}^{\infty} \gamma_t(v_i, x_{-i}) dv_i}. \quad (24)$$

The one-dimensional integrals in (23)-(24) are integrals of the form $\int_D \phi(u_i, x_{-i}) du_i$ for some integrand ϕ and domain $D \subseteq \mathbb{R}$. We will treat unbounded domains by imposing a suitable truncation; quadrature routines that determine the truncation adaptively could also be considered. Here we consider the class of composite Newton-Cotes quadrature rules

$$\int_D \phi(u_i, x_{-i}) du_i \approx \sum_{r=1}^R \omega_r \phi(v_r, x_{-i}), \quad (25)$$

where $\{\omega_r\}_{r=1,\dots,R}$ are quadrature weights which depend on the degree of the approximation and $\{v_r\}_{r=1,\dots,R}$ are $R \in \mathbb{N}$ many equispaced quadrature points in D [29, p. 34]. We take (25) to be of the closed type, i.e. v_1 and v_R take the endpoints of D , as this choice will be convenient for domains of the type $D = (-\infty, x_i]$ for $x_i < \infty$. The composite quadrature rule (25) is derived by integrating Lagrange interpolation polynomials on subintervals; the degree of which dictates the accuracy of the approximation on each subinterval. We shall denote the resulting approximation of \tilde{f}_i by \hat{f}_i ; for components $i = 1, \dots, p$ with analytically tractable Gibbs velocity field \tilde{f}_i , we set $\hat{f}_i = \tilde{f}_i$. In Supplementary Material G.2, we examine numerical stability of the Gibbs velocity field computation based on quadrature approximations of (23)-(24).

We now consider how to approximate a particle trajectory driven by the ODE

$$\frac{d}{dt} x(t) = \hat{f}(t, x(t)), \quad t \in [0, 1], \quad (26)$$

with initial condition $X_0 \sim \pi_0$. We will introduce a novel numerical integration scheme that is reminiscent of the systematic Gibbs scan. In the following, we will show that our proposed scheme, in contrast to standard numerical integrators, allows efficient computation of marginal distributions in high dimensions. For simplicity, we discretize the time interval $[0, 1]$ into a regular grid $t_m = mh, m = 0, \dots, M$ with a constant step size $h = 1/M$; non-constant step sizes can also be employed. To evolve a particle with position $X_{m-1} = (X_{m-1,1}, \dots, X_{m-1,p}) \in \mathbb{R}^d$ at time t_{m-1} on the subinterval $[t_{m-1}, t_m]$, we consider

$$\frac{d}{dt} x_1(t) = \hat{f}_1(t, x_1(t), x_{-1}), \quad t \in [t_{m-1}, t_m],$$

for the first component, with the other components fixed as $x_{-1} = (X_{m-1,2}, \dots, X_{m-1,p}) = X_{m-1,2:p}$. If the solution $x_1(t), t \in [t_{m-1}, t_m]$ is analytically tractable, we set $X_{m,1} = x_1(t_m)$; otherwise we will rely on the Euler discretization

$$X_{m,1} = X_{m-1,1} + h \hat{f}_1(t_{m-1}, X_{m-1,1}, X_{m-1,2:p}). \quad (27)$$

Similarly, we update the second component by considering

$$\frac{d}{dt} x_2(t) = \hat{f}_2(t, x_2(t), x_{-2}), \quad t \in [t_{m-1}, t_m], \quad (28)$$

with $x_{-2} = (X_{m,1}, X_{m-1,3:p})$, and setting $X_{m,2} = x_2(t_m)$ if the solution is available or

$$X_{m,2} = X_{m-1,2} + h \hat{f}_2(t_{m-1}, X_{m,1}, X_{m-1,2:p})$$

otherwise. We then iteratively update all other components in a systematic manner to obtain $X_m = (X_{m,1}, \dots, X_{m,p}) \in \mathbb{R}^d$.

In summary, the above procedure defines the maps

$$(X_{m,1:i}, X_{m-1,(i+1):p}) = \Psi_{m,i}(X_{m,1:i-1}, X_{m-1,i:p}), \quad i = 1, \dots, p,$$

(with $X_{m,1:0} = \emptyset$ and $X_{m-1,(p+1):p} = \emptyset$) which update one component at a time. By iterating over all components, the composition

$$X_m = \Phi_m(X_{m-1}) = \Psi_{m,p} \circ \dots \circ \Psi_{m,1}(X_{m-1}) \quad (29)$$

defines our numerical integration scheme. The flow maps $T_{t_m} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ induced by this scheme

$$X_m = T_{t_m}(X_0) = \Phi_m \circ \dots \circ \Phi_1(X_0), \quad m = 0, \dots, M, \quad (30)$$

(with $T_{t_0}(X_0) = X_0$) can be shown to be a first order approximation of the flow maps $\{\hat{T}_t\}_{t \in [0,1]}$ defined by (26) (see Supplementary Material E), i.e. $|T_{t_m}(X_0) - \hat{T}_{t_m}(X_0)| = O(h)$ for all $m = 0, \dots, M$ if the step size h is sufficiently small¹.

4.2 Distribution of approximate Gibbs flow samples

We now detail how to compute the marginal distributions of $X_m, m = 0, \dots, M$, under the numerically approximated Gibbs flow (30). This allows us to utilize these distributions as proposal distributions within a sequential importance sampler.

Under the assumptions of Proposition 3, the Gibbs flow maps $\{\tilde{T}_t\}_{t \in [0,1]}$ are C^1 -diffeomorphisms by construction. Hence their approximation (30) will be injective if the step size h is sufficiently small and quadrature approximations (if employed) are accurate enough - see [9, 36] for similar arguments. Under these conditions, it follows from a change of variables that the density of $q_{t_m} = (T_{t_m})_{\#} \pi_0$, is

$$q_{t_m}(X_m) = \pi_0(X_0) |\det(\nabla T_{t_m}(X_0))|^{-1} \quad (31)$$

where $X_0 = T_{t_m}^{-1}(X_m)$ is given by the inverse map, $|\det(\nabla T_{t_m}(X_0))|$ denotes the absolute value of the determinant of the Jacobian matrix of T_{t_m} . In numerical implementations, monotonicity may be monitored by checking for any sign changes in the Jacobian determinant. From (30), the latter can be computed as

$$\det(\nabla T_{t_m}(X_0)) = \prod_{k=1}^m \det(\nabla \Phi_k(X_{k-1})).$$

Using the structure of our numerical integration scheme (29), the computational cost of computing

$$\det(\nabla \Phi_k(X_{k-1})) = \prod_{i=1}^p \det(\nabla \Psi_{k,i}(X_{k,1:i-1}, X_{k-1,i:p}))$$

is at most $O(\sum_{i=1}^p d_i^3)$. This cost may be even lower in statistical models with conditional independence structure as Gibbs velocity fields will inherit such structures yielding sparse Jacobian matrices (e.g. Section 5.2).

¹This error result holds even if Euler discretizations are employed for some or all components.

In the case of (19) for one-dimensional components and the Euler discretization (27), computing

$$\det(\nabla \Psi_{k,i}(X_{k,1:i-1}, X_{k-1,i:p})) = \det(1 + h \partial_{x_i} \hat{f}_i(t_{k-1}, X_{k,1:i-1}, X_{k-1,i:p}))$$

requires the partial derivative of the approximate Gibbs velocity field $\partial_{x_i} \hat{f}_i(t, x)$. It turns out that we can compute $\partial_{x_i} \hat{f}_i(t, x)$ by simply replacing integrals in the partial derivative of the Gibbs velocity field

$$\partial_{x_i} \tilde{f}_i(t, x) = \lambda'(t) \left\{ \frac{\int_{-\infty}^{\infty} \log L(u_i, x_{-i}) \gamma_t(u_i, x_{-i}) du_i}{\int_{-\infty}^{\infty} \gamma_t(u_i, x_{-i}) du_i} - \log L(x) \right\} - \tilde{f}_i(t, x) \partial_{x_i} \log \gamma_t(x)$$

with approximations based on the same quadrature rule. This follows from the following argument which allows one to compute the partial derivative w.r.t. x_i of approximations of integrals of the form $\int_{-\infty}^{x_i} \phi(u_i, x_{-i}) du_i$. Denote by $\hat{\phi}$ the underlying Lagrange interpolant giving rise to the quadrature rule (25). By the first fundamental theorem of calculus and the closed property of (25)

$$\partial_{x_i} \sum_{r=1}^R \omega_r \phi(v_r, x_{-i}) = \partial_{x_i} \int_{-\infty}^{x_i} \hat{\phi}(u_i, x_{-i}) du_i = \hat{\phi}(x_i, x_{-i}) = \phi(x_i, x_{-i}). \quad (32)$$

To illustrate the computational savings our proposed numerical integrator (29) offers over standard integrators like the forward Euler method

$$X_m = \Phi_m(X_{m-1}) = X_{m-1} + h \hat{f}(t_{m-1}, X_{m-1}), \quad (33)$$

we consider the case of solely one-dimensional components, i.e. $d_i = 1$ for all $i = 1, \dots, p$. In the absence of any sparsity, computing the Jacobian determinant of the mapping in (33) would cost at most $O(d^3)$; in contrast the cost associated to (29) is only $O(d)$.

Given $N \in \mathbb{N}$ independent samples $X_m^n, n = 1, \dots, N$ from (30), the above discussion allows us to employ the marginal distribution q_{t_m} in (31) as a proposal distribution within an importance sampling approximation of π_{t_m} . The importance weights $w_m(X_m^n) = \gamma_{t_m}(X_m^n)/q_{t_m}(X_m^n)$ can be computed recursively using

$$w_m(X_m^n) = w_{m-1}(X_{m-1}^n) \frac{\gamma_{t_m}(X_m^n)}{\gamma_{t_{m-1}}(X_{m-1}^n) |\det(\nabla \Phi_m(X_{m-1}^n))|^{-1}}, \quad m = 1, \dots, M,$$

with $w_0(X_0^n) = 1$. An algorithmic description of the resulting sequential importance sampler is detailed in Algorithm 1. Using the output, we can approximate expectations of the form $\int_{\mathbb{R}^d} \phi(x) \pi(x) dx$ with the weighted sum $\sum_{n=1}^N \phi(X_M^n) W_M^n$, and estimate the marginal likelihood $Z = \int_{\mathbb{R}^d} \pi_0(x) L(x) dx$ unbiasedly with \hat{Z}_M . The adequacy of the importance sampling approximation based on the Gibbs flow can be monitored using the effective sample size (ESS) introduced in [34]. This quantity takes values between 1 and N , and will be equal to N if samples are distributed according to the target distribution.

Algorithm 1 Gibbs flow sequential importance sampler (GF-SIS)

Input: prior π_0 , likelihood L , inverse temperature λ , step size h , and Gibbs velocity field \tilde{f} .

For time step $m = 0$

For $n = 1, \dots, N$

- (a) sample $X_0^n = (X_{0,1}^n, \dots, X_{0,p}^n) \sim \pi_0$;
- (b) set $w_0^n = 1$ and $W_0^n = N^{-1}$;
- (c) set $\text{ESS}_0 = N$ and $\hat{Z}_0 = 1$.

For time step $m = 1, \dots, M$

For $n = 1, \dots, N$

For $i = 1, \dots, p$

- (d) set $(X_{m,1:i}^n, X_{m-1,(i+1):p}^n) = \Psi_{m,i}(X_{m,1:i-1}^n, X_{m-1,i:p}^n)$ using Section 4.1;
- (e) compute $J_{m,i}^n = \det(\nabla \Psi_{m,i}(X_{m,1:i-1}^n, X_{m-1,i:p}^n))$ using Section 4.2;
- (f) set $X_m^n = (X_{m,1}^n, \dots, X_{m,p}^n)$ and $J_m^n = \prod_{i=1}^p J_{m,i}^n$;
- (g) compute unnormalized weights

$$w_m^n = w_{m-1}^n \frac{\gamma_{t_m}(X_m^n)}{\gamma_{t_{m-1}}(X_{m-1}^n) |J_m^n|^{-1}};$$

- (h) compute normalized weights $W_m^n = w_m^n / \sum_{\ell=1}^N w_m^\ell$;
- (i) compute effective sample size $\text{ESS}_m = \left\{ \sum_{n=1}^N (W_m^n)^2 \right\}^{-1}$;
- (j) compute normalizing constant estimator $\hat{Z}_m = N^{-1} \sum_{n=1}^N w_m^n$.

Output: samples $\{X_M^n\}_{n=1, \dots, N}$, normalized weights $\{W_M^n\}_{n=1, \dots, N}$ and normalizing constant estimator \hat{Z}_M .

4.3 Combining Gibbs flow with Markov chain Monte Carlo

State-of-the-art methods based on annealed importance sampling (AIS) simulate $N \in \mathbb{N}$ inhomogeneous Markov chains $X_0^n \sim \pi_0$ and $X_m^n \sim K_m(X_{m-1}^n, \cdot)$, for $m = 1, \dots, M$ and $n = 1, \dots, N$, where K_m is a π_{t_m} -invariant Markov chain Monte Carlo (MCMC) kernel. For each $m = 1, \dots, M$, although the marginal distribution of $\{X_m^n\}_{n=1, \dots, N}$ is typically intractable, one can still use these samples within an importance sampling approximation of π_{t_m} , by associating sample $n = 1, \dots, N$ with the importance weight

$$w_m(X_{0:m-1}^n) = w_{m-1}(X_{0:m-2}^n) \frac{\gamma_{t_m}(X_{m-1}^n)}{\gamma_{t_{m-1}}(X_{m-1}^n)}, \quad m = 1, \dots, M, \quad (34)$$

with $w_0(X_{0:-1}^n) = 1$ and $X_{0:m-1}^n = (X_0^n, \dots, X_{m-1}^n)$. The choice of bridging distributions $\{\pi_{t_m}\}_{m=0, \dots, M}$ and MCMC kernels $\{K_m\}_{m=1, \dots, M}$ can have a large impact on algorithmic performance; if these kernels mix slowly and/or the intermediate distributions are too distant, the variance of the importance weights (34) can be very high.

To improve the performance of AIS, references [56, 57] suggested adding deterministic maps Φ_m which attempt to “push” samples from $\pi_{t_{m-1}}$ to π_{t_m} , but the authors did not propose a generic methodology to construct such transport maps. In our context, we will rely on numerical approximation of the Gibbs flow, as described in Section 4.1, to build these maps. Practically, for $n = 1, \dots, N$, we initialize by sampling $X_0^n \sim \pi_0$ and setting $\tilde{X}_0^n = X_0^n$. For $m = 1, \dots, M$, we then iterate by setting $X_m^n = \Phi_m(\tilde{X}_{m-1}^n)$, as defined in (29), and sampling $\tilde{X}_m^n \sim K_m(X_m^n, \cdot)$ from a π_{t_m} -invariant MCMC kernel. Like in AIS, we can also use the samples $\{\tilde{X}_m^n\}_{n=1, \dots, N}$ within an importance sampling approximation of π_{t_m} . The importance weights are given by

$$w_m(X_{0:m}^n, \tilde{X}_{0:m}^n) = w_{m-1}(X_{0:m-1}^n, \tilde{X}_{0:m-1}^n) \frac{\gamma_{t_m}(X_m^n)}{\gamma_{t_{m-1}}(\tilde{X}_{m-1}^n) |\det(\nabla \Phi_m(\tilde{X}_{m-1}^n))|^{-1}},$$

for $m = 1, \dots, M$ with $w_0(X_0^n, \tilde{X}_0^n) = 1$. We provide an algorithmic description of the resulting annealed importance sampler in Algorithm 2. From the output, expectations $\int_{\mathbb{R}^d} \phi(x) \pi(x) dx$ can be approximated by the weighted sum $\sum_{n=1}^N W_M^n \phi(\tilde{X}_M^n)$ and the marginal likelihood by the unbiased estimator \hat{Z}_M . Although resampling is not considered in Algorithms 1-2 to simplify our exposition, any resampling scheme can also be employed with minor modifications; this is detailed in Supplementary Material F for completeness.

Algorithm 2 Gibbs flow annealed importance sampler (GF-AIS)

Input: prior π_0 , likelihood L , inverse temperature λ , step size h , Gibbs velocity field \tilde{f} , MCMC kernels $\{K_m\}_{m=1, \dots, M}$.

For time step $m = 0$

For $n = 1, \dots, N$

- (a) sample $X_0^n = (X_{0,1}^n, \dots, X_{0,p}^n) \sim \pi_0$ and set $\tilde{X}_0^n = X_0^n$;
- (b) set $w_0^n = 1$ and $W_0^n = N^{-1}$;
- (c) set $\text{ESS}_0 = N$ and $\hat{Z}_0 = 1$.

For time step $m = 1, \dots, M$

For $n = 1, \dots, N$

For $i = 1, \dots, p$

- (d) set $(X_{m,1:i}^n, \tilde{X}_{m-1,(i+1):p}^n) = \Psi_{m,i}(X_{m,1:i-1}^n, \tilde{X}_{m-1,i:p}^n)$ using Section 4.1;
- (e) compute $J_{m,i}^n = \det(\nabla \Psi_{m,i}(X_{m,1:i-1}^n, \tilde{X}_{m-1,i:p}^n))$ using Section 4.2;
- (f) set $X_m^n = (X_{m,1}^n, \dots, X_{m,p}^n)$ and $J_m^n = \prod_{i=1}^p J_{m,i}^n$;
- (g) compute unnormalized weights

$$w_m^n = w_{m-1}^n \frac{\gamma_{t_m}(X_m^n)}{\gamma_{t_{m-1}}(\tilde{X}_{m-1}^n) |J_m^n|^{-1}};$$

- (h) compute normalized weights $W_m^n = w_m^n / \sum_{\ell=1}^N w_m^\ell$;
- (i) sample $\tilde{X}_m^n \sim K_m(X_m^n, \cdot)$ from π_{t_m} -invariant MCMC kernel;
- (j) compute effective sample size $\text{ESS}_m = \left\{ \sum_{n=1}^N (W_m^n)^2 \right\}^{-1}$;
- (k) compute normalizing constant estimator $\hat{Z}_m = N^{-1} \sum_{n=1}^N w_m^n$.

Output: samples $\{\tilde{X}_M^n\}_{n=1, \dots, N}$, normalized weights $\{W_M^n\}_{n=1, \dots, N}$ and normalizing constant estimator \hat{Z}_M .

4.4 Illustration on Gaussian model

We end this section by illustrating our methodology on a toy Gaussian model. We consider a standard Gaussian prior distribution π_0 on \mathbb{R}^d and a likelihood function $L(x; y) = \exp(-\frac{1}{2} \langle x - y, \Omega^{-1}(x - y) \rangle)$ with symmetric positive definite $\Omega \in \mathbb{R}^{d \times d}$ and observation $y \in \mathbb{R}^d$. By conjugacy, the curve of distributions $\{\pi_t\}_{t \in [0,1]}$ defined in (2) lies in the Gaussian family (Supplementary Material G.1). In the following, we consider dimension $d = 4, 8$ and set $y = (14.25, \dots, 14.25)$, $\Omega_{ii} = 1$ for $i = 1, \dots, d$ and $\Omega_{ij} = 0.5$ for $i, j = 1, \dots, d$ satisfying $i \neq j$.

As the full conditional distributions of $\{\pi_t\}_{t \in [0,1]}$ are Gaussian, one can derive the Gibbs flow analytically. Instead of exploiting this analytical tractability, we employ the generic choice (19) for one-dimensional components. To approximate the one-dimensional integrals in the Gibbs velocity field (23)-(24), we truncate \mathbb{R} to the interval $[-10, 10]$ and apply a composite trapezoidal rule with $R = 200$ quadrature points. The Gibbs flow is then approximated using the numerical integration scheme proposed in (29) with the Euler discretization (27) for each component. Using preliminary runs, we choose the number of time steps $M = 100$, which defines the step size $h = 1/M$, to be large enough to ensure that trajectories are stable. As the observation y is rather extreme (i.e. more than 14 prior

standard deviations away from the prior mean), one might expect some stiffness when numerically integrating the Gibbs flow. Figure 1 examines to what extent this is the case by simulating trajectories using an adaptive fourth-order Runge-Kutta numerical integrator from the `pracma` R package. In Supplementary Material G.4, we explore the case where the observation y is more extreme than what is considered here.

In Figure 2, we display the performance of the resulting GF-SIS (Algorithm 1) with $N = 512$ samples. Using the same number of samples N and time steps M , we combine approximate Gibbs flow with Hamiltonian Monte Carlo (HMC) kernels² within GF-AIS (Algorithm 2). Although GF-AIS requires marginally more compute time than GF-SIS in this example, the improvement in algorithmic performance clearly outweighs the computational overhead. As competing algorithm, we consider AIS with the same N , M and HMC kernels as GF-AIS, but we increase the number of HMC iterations at each time step to match the computational time of GF-AIS to ensure a fair comparison.

We also compare our proposed methodology to some existing approaches based on transport maps. Firstly, we consider the methodology developed in [22, 44, 45, 52], which approximates the Knothe-Rosenblatt transport map [49, 33] by minimizing the Kullback-Leibler divergence $\text{KL}((T_\beta)_\# \pi_0 | \pi)$ of the posterior π from the push-forward measure $(T_\beta)_\# \pi_0$, over maps T_β that are parametrized by β using a polynomial basis expansion. We implemented this methodology using the `TransportMaps` Python package³ and will refer to it as TM. Using the same number of importance samples $N = 512$, we choose the order of the polynomial basis as one (which provides the correct parameterization in this Gaussian example) and set the optimization tolerance as 10^{-4} to match the compute time of GF-AIS in dimension $d = 4$. We do not consider TM in dimension $d = 8$ as we observed the substantial computation time required to run the Kullback-Leibler optimization routine to be uncompetitive compared to our proposed methodology. For other problems with conditional independence structures in the posterior distribution, one could exploit such sparsity to obtain a more scalable approach using recent work in [52].

Another methodology we shall consider is the Stein variational gradient descent (SVGD) algorithm introduced by [36], which seeks a velocity field f in a reproducing kernel Hilbert space that minimizes the Kullback-Leibler divergence $\text{KL}((\text{Id} + hf)_\# q | \pi)$ of the posterior π from the push-forward measure $(\text{Id} + hf)_\# q$, where $\text{Id}(x) = x$ denotes the identity map, $h > 0$ is some step size to be specified and q is a given distribution (at particular iteration). We implemented the modification of SVGD described in [28] that allows one to perform importance sampling using the gradient dynamics. In our simulations, we set the number of “leader particles”, used to compute the intractable integral in the SVGD velocity field, as equal to the number of “follower particles”, which represents the sample size of the resulting importance sampler. We manually tuned the bandwidth parameter that defines the radial basis function kernel within SVGD and the step size taken by each SVGD move using preliminary runs. The number of iterations used to refine the adaptive importance sampler is then selected to match the computational cost of GF-AIS and TM.

The results based on 100 independent repetitions of all algorithms are summarized in Figure 2. In dimension $d = 4$, TM clearly outperformed all other competing methods at a fixed computational cost. In dimension $d = 8$, GF-AIS performed the best: its ESS% is the highest and its sample variance of log-marginal likelihood estimates was observed to be 14 and 111 times smaller than AIS and SVGD, respectively. These results also suggest that SVGD is not particularly competitive relative to AIS: its sample variance of log-marginal likelihood estimates was observed to be 2 and 8 times larger than AIS in dimensions $d = 4$ and $d = 8$, respectively. This is consistent with analytical results recently derived in [3] concerning how SVGD suffers from the curse of dimensionality. Given the difficulty of comparing different types of methods in a completely fair manner, we end this discussion by noting that our objective is merely to offer a sense of how our proposed methodology compares with existing transport-based algorithms, and to illustrate the strengths and weaknesses of each method.

²For both dimensions $d = 4$ and $d = 8$, we apply five iterations of a Hamiltonian Monte Carlo kernel at each time step. To achieve suitable acceptance probabilities, we use a step size of 0.25 for the leapfrog integrator and an integration time of 2.5.

³<http://transportmaps.mit.edu/docs/>

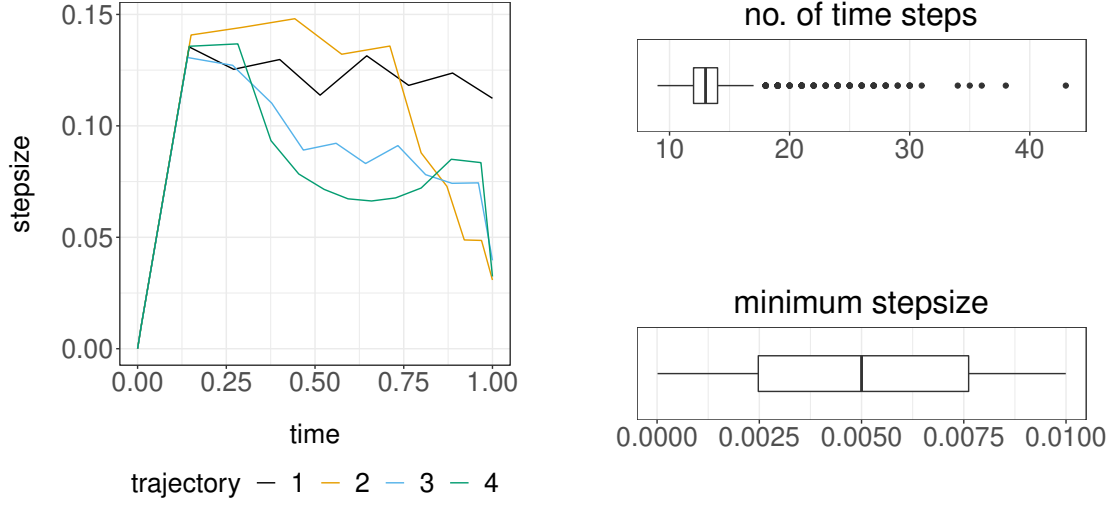


Figure 1: Simulating trajectories under the Gibbs flow using an adaptive fourth-order Runge-Kutta numerical integrator on the toy Gaussian model in Section 4.4. Time evolution of step size taken by adaptive integrator for four trajectories (*left*). Boxplots of the number of time steps taken (*upper-right*) and the minimum step size used over time (*lower-right*) based on 1024 independent trajectories. Note that the boxplot in the lower-right panel considers only 15% of the trajectories which required step sizes that are smaller than the default value.

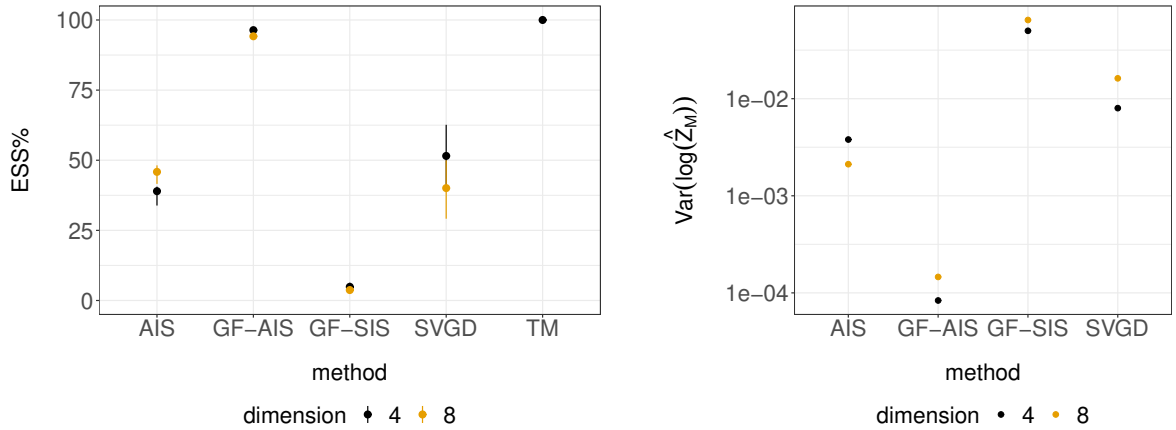


Figure 2: Boxplots of terminal effective sample size percentage (*left*) and variance of log-marginal likelihood estimates (*right*) on the toy Gaussian model of Section 4.4 in dimension $d = 4, 8$, obtained with 100 independent repetitions of AIS, GF-SIS (Algorithm 1), GF-AIS (Algorithm 2), SVGD [36, 28] and TM [22, 44, 45, 52]. The variance of TM log-marginal likelihood estimates in dimension $d = 4$, not shown in the right panel, is 6.58×10^{-18} .

5 Applications

5.1 Bayesian mixture modelling

We now investigate the performance of Gibbs flow samplers on a Bayesian mixture model, where the posterior distribution of mixture means is inferred. This is a canonical example of distributions with multiple well-separated modes.

Consider $J \in \mathbb{N}$ independent observations from a univariate Gaussian mixture model with d components, i.e. for $j = 1, \dots, J$ each observation is distributed according to $Y_j \sim \frac{1}{d} \sum_{i=1}^d \mathcal{N}(x_i, \sigma_i^2)$, where $\mathcal{N}(\mu, \varsigma^2)$ (and $y \mapsto \mathcal{N}(y; \mu, \varsigma^2)$) denotes the Gaussian distribution (and density) with mean μ and variance ς^2 . Following [35], we set $d = 4$, $\sigma_i = \sigma = 0.55$ for $i = 1, \dots, d$ and perform inference only on the mean parameters $x = (x_1, \dots, x_4) \in \mathbb{R}^d$. We generate the data $\{y_j\}_{j=1, \dots, J}$ using $J = 100$ simulations from the model with parameter value $x^* = (-3, 0, 3, 6)$ and stratification between components. We adopt a uniform prior distribution on the d -dimensional hypercube $[-10, 10]^d$. The curve of distributions in (2) is

$$\pi_t(x) = \frac{\mathbb{I}_{[-10, 10]^d}(x) L(x)^{\lambda(t)}}{20^d Z(t)}, \quad t \in [0, 1], \quad (35)$$

where $\mathbb{I}_{[-10, 10]^d}(x) = 1$ if $x \in [-10, 10]^d$ and 0 otherwise, and the likelihood is

$$L(x) = \frac{1}{d^J} \prod_{j=1}^J \sum_{i=1}^d \mathcal{N}(y_j; x_i, \sigma^2). \quad (36)$$

It follows from exchangeability of the prior and non-identifiability of mixture components that the posterior distribution (1) is invariant under “label permutation”. Therefore $\pi_1 = \pi$ admits $d! = 24$ well-separated modes centered approximately around all permutations of x^* . As it is known that simple MCMC and importance sampling methods typically perform poorly for such problems [10], we will determine the quality of the Gibbs flow approximation (18) by examining how well it can explore all 24 modes equally.

As the full conditional distributions of the posterior are not in the exponential family, we employ the Gibbs velocity field (23)-(24) for one-dimensional components. We use a composite trapezoidal rule with $R = 100$ quadrature points to approximate one-dimensional integrals; no additional truncation is necessary as the support of the prior is a hypercube. Due to multimodality, we can expect some stiffness when numerically integrating the Gibbs flow. Based on $N = 1024$ independent trajectories that were simulated using an adaptive integrator: the median number of time steps taken was 37 with an interquartile range (IQR) of 16, and amongst 97% of the trajectories that required step sizes that were smaller than the default value, the median of the minimum step size used over time was 4.4×10^{-3} with an IQR of 3.2×10^{-3} . To identify the time periods when the ODE is stiff, we plot the time evolution of the norm of the Gibbs velocity field along these trajectories in the left panel of Figure 3, and along trajectories generated by combining the Gibbs flow with HMC kernels⁴ in the right panel. These plots show that the ODE is more stiff initially when samples are attracted by multiple local modes, and gradually becomes less stiff as samples concentrate around each mode.

We compare the time evolution of prior samples under the Gibbs flow with the output of a standard SMC sampler with many particles as the reference truth in Figure 4. The performance of the Gibbs flow for this challenging problem is striking as the samples reach all modes. This is also seen in Figure 5 that shows all pairwise marginal posterior distributions on \mathbb{R}^2 (note that each of these marginals admits 12 well-separated modes). To corroborate these observations, we simulate another $N = 16,384$ independent Gibbs flow samples and display the proportion of samples in each of the 24 modes in Figure 6. The uniformity of these proportions is then tested using a Pearson’s Chi-squared goodness-

⁴Here we apply a Hamiltonian Monte Carlo kernel between time intervals of $h = 0.0025$. We use a step size of 0.1 for the leapfrog integrator and an integration time of 1.0.

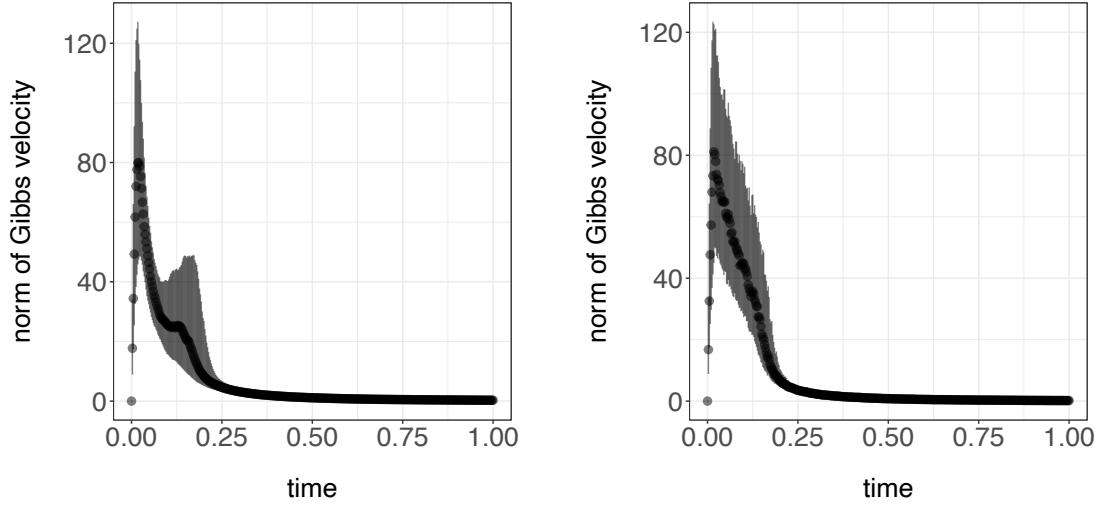


Figure 3: Boxplots illustrating the time evolution of the norm of the Gibbs velocity field along $N = 1024$ trajectories generated by the Gibbs flow (*left*) and by combining the Gibbs flow with HMC kernels (*right*) for the Bayesian mixture model in Section 5.1.

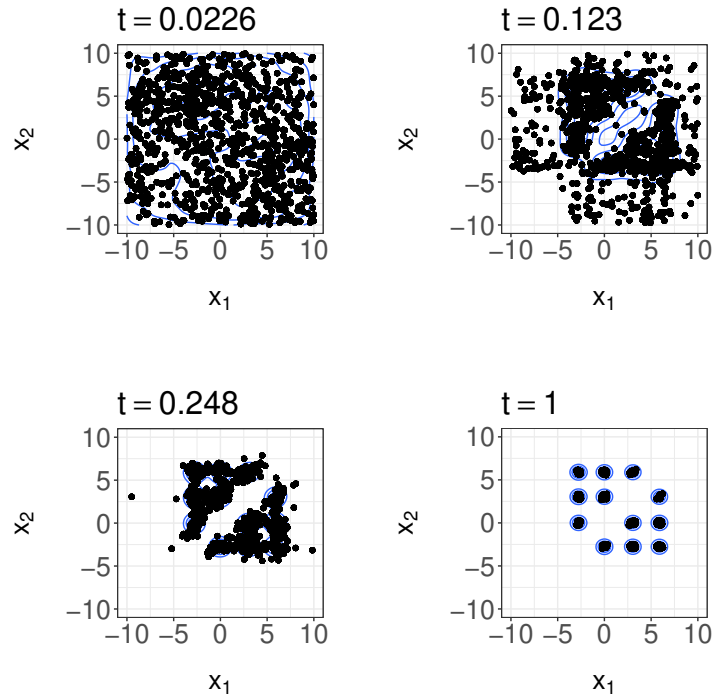


Figure 4: Time evolution of $N = 1024$ prior samples in the first two dimensions under the Gibbs flow (*black dots*) for the Bayesian mixture model in Section 5.1. For each time instance, the superimposed (*blue*) contours represent the marginal of the target distribution obtained as a kernel density estimate from the output of a SMC sampler.

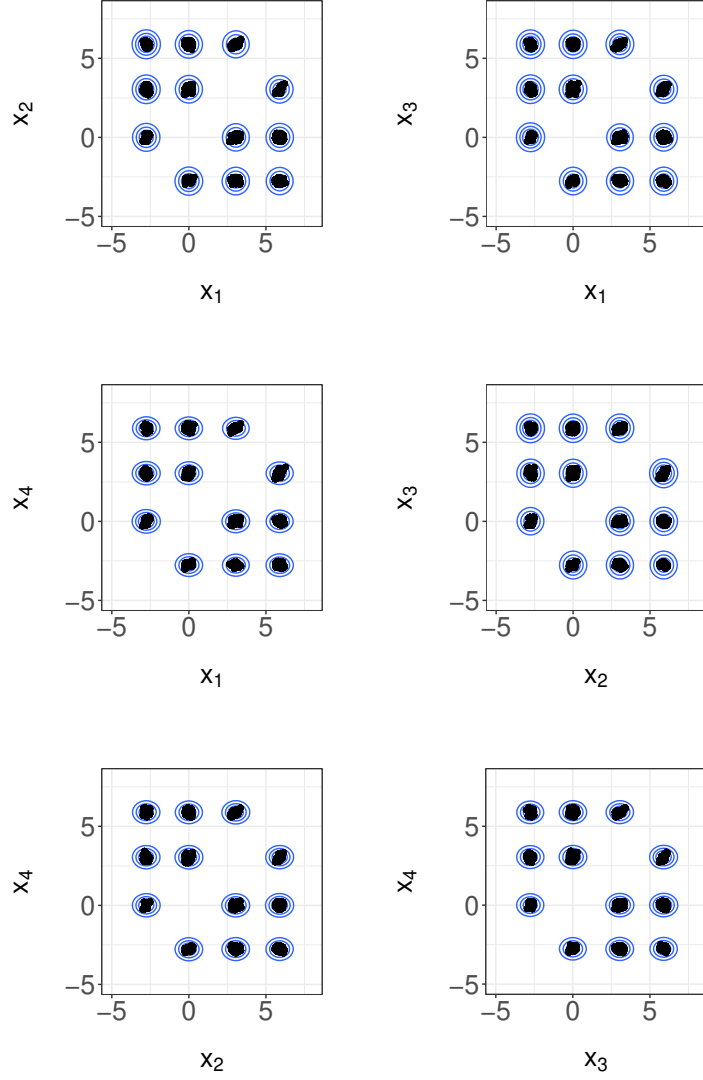


Figure 5: All pairs of marginal posterior distributions on \mathbb{R}^2 for the Bayesian mixture model in Section [5.1](#).

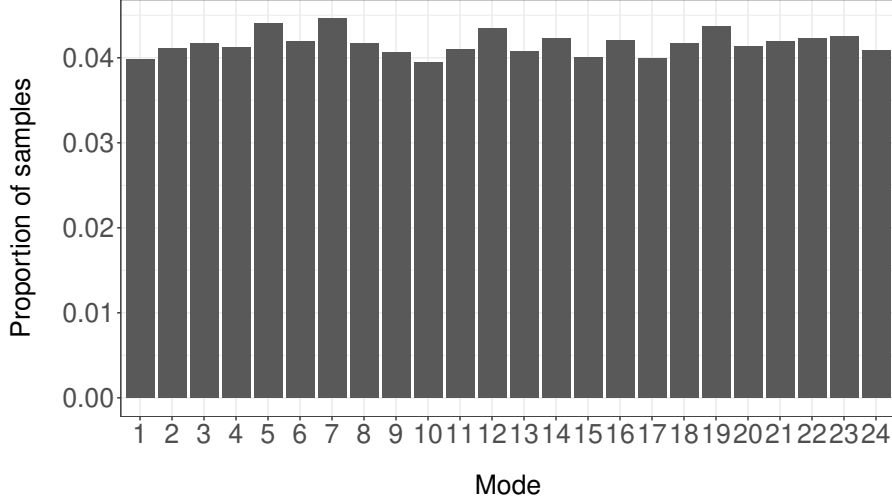


Figure 6: Proportion of Gibbs flow samples in each of the 24 modes for the Bayesian mixture model in Section 5.1.

of-fit test, which gives a p-value of 0.8522. Next, we examine how well the distribution of Gibbs flow samples matches the posterior distribution in the left panel of Figure 7. Although there is good agreement between these distributions, there is still some discrepancy which is analyzed in Proposition 4. In the right panel of Figure 7, we show that this difference can be reduced by combining the Gibbs flow with HMC kernels, as discussed in Section 4.3.

5.2 Variance component models

We now apply our proposed methodology to a variance component model, which is typical of problems in Bayesian statistics where one would employ a Gibbs sampler [24, 50]. There are two hyperparameters with prior distributions $\sigma_\theta^2 \sim \mathcal{IG}(\alpha_0, \beta_0)$ and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, where $\mathcal{IG}(a, b)$ (and $s \mapsto \mathcal{IG}(s; a, b)$) denotes the inverse Gamma distribution (and density) with shape parameter a and scale parameter b . Following [50], we adopt an improper prior for σ_θ^2 and a flat or vague prior for μ . Given these hyperparameters, there are $K \in \mathbb{N}$ location parameters $\theta = (\theta_1, \dots, \theta_K) \in \mathbb{R}^K$ that are conditionally independent and distributed as $\theta_i \sim \mathcal{N}(\mu, \sigma_\theta^2)$ for $i = 1, \dots, K$. With these parameters, $J \in \mathbb{N}$ observations at each location $i = 1, \dots, K$ are modeled as conditionally independent and distributed as $Y_{ij} \sim \mathcal{N}(\theta_i, \sigma_e^2)$ for $j = 1, \dots, J$, where σ_e^2 is estimated empirically. We will write $y = (y_{ij}) \in \mathbb{R}^{K \times J}$ as the observed dataset.

In this application, the improper prior (with a possibly negative value of α_0) is

$$p_0(\sigma_\theta^2, \mu, \theta) = \mathcal{IG}(\sigma_\theta^2; \alpha_0, \beta_0) \mathcal{N}(\mu; \mu_0, \sigma_0^2) \prod_{i=1}^K \mathcal{N}(\theta_i; \mu, \sigma_\theta^2) \quad (37)$$

and the likelihood function is $p(y|\sigma_\theta^2, \mu, \theta) = \prod_{i=1}^K \prod_{j=1}^J \mathcal{N}(y_{ij}; \theta_i, \sigma_e^2)$ for $(\sigma_\theta^2, \mu, \theta) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^K$. To employ the methodology described in Section 2.1, we set $(x_1, x_2, x_3) = (\sigma_\theta^2, \mu, \theta)$ as the parameters to be inferred and consider the following “artificial” prior distribution

$$\pi_0(\sigma_\theta^2, \mu, \theta) = \mathcal{IG}(\sigma_\theta^2; \alpha_1, \beta_1) \mathcal{N}(\mu; \mu_1, \sigma_1^2) \prod_{i=1}^K \mathcal{N}(\theta_i; \mu_2, \sigma_2^2) \quad (38)$$

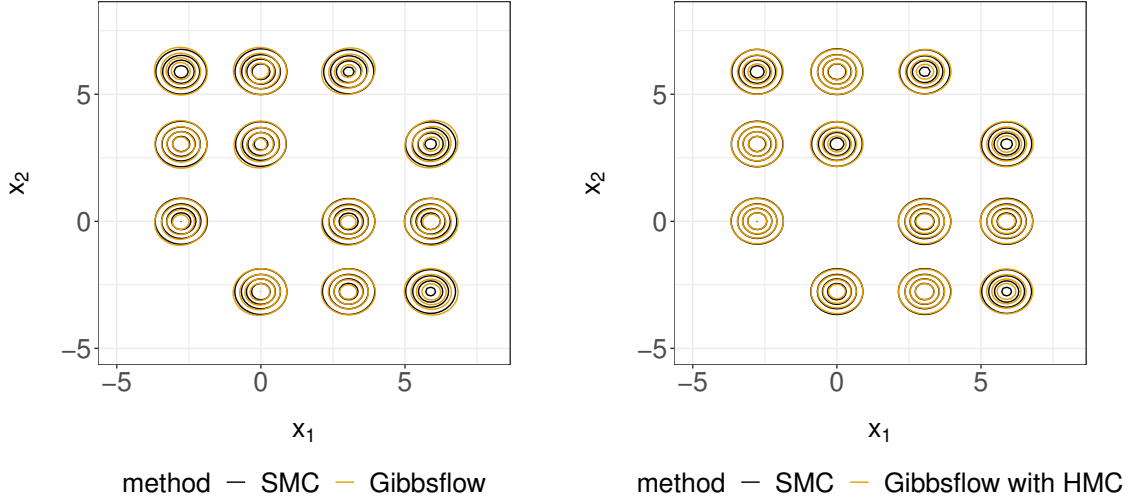


Figure 7: Marginal posterior distribution (*black*), marginal distribution of Gibbs flow samples (*left-orange*), and marginal distribution of samples under the Gibbs flow and Hamiltonian Monte Carlo kernels (*right-orange*) for the Bayesian mixture model in Section 5.1, obtained using kernel density estimates from the output of a SMC sampler and $N = 16,384$ independent samples, respectively.

to initialize our method, for some fixed $\alpha_1 > 0, \beta_1 > 0, \mu_1 \in \mathbb{R}, \sigma_1^2 > 0, \mu_2 \in \mathbb{R}, \sigma_2^2 > 0$. The corresponding “artificial” likelihood function that would yield the desired posterior $p(\sigma_\theta^2, \mu, \theta|y) \propto p_0(\sigma_\theta^2, \mu, \theta)p(y|\sigma_\theta^2, \mu, \theta)$ is

$$L(\sigma_\theta^2, \mu, \theta) = \frac{p_0(\sigma_\theta^2, \mu, \theta)p(y|\sigma_\theta^2, \mu, \theta)}{\pi_0(\sigma_\theta^2, \mu, \theta)}.$$

Given these choices, which are necessary to deal with the improper prior (37), we can then define the curve of distributions $\{\pi_t\}_{t \in [0,1]}$ in (2).

It can be shown that the full conditional distributions of $\pi_t, t \in [0, 1]$ are

$$\begin{aligned} \pi_t(\sigma_\theta^2|\mu, \theta) &= \mathcal{IG}(\sigma_\theta^2; \alpha(t), \beta(t|\mu, \theta)), \quad \pi_t(\mu|\sigma_\theta^2, \theta) = \mathcal{N}(\mu; \nu(t|\sigma_\theta^2, \theta), \varsigma^2(t|\sigma_\theta^2)), \\ \pi_t(\theta|\sigma_\theta^2, \mu) &= \prod_{i=1}^K \mathcal{N}(\theta_i; \xi_i(t|\sigma_\theta^2, \mu, y), \tau^2(t|\sigma_\theta^2)), \end{aligned} \quad (39)$$

where the summary statistics $\alpha, \beta, \nu, \varsigma^2, \xi_1, \dots, \xi_K, \tau^2$ are given in Supplementary Material H. Since these full conditionals lie in the exponential family, we can exploit such analytical tractability to determine a Gibbs velocity field. For the parameter σ_θ^2 , we use (19) which reduces to

$$\tilde{f}_1(t, \sigma_\theta^2, \mu, \theta) = \frac{-\int_0^{\sigma_\theta^2} \{\kappa(t|\mu, \theta) - \alpha'(t) \log(u_1) - \beta'(t|\mu, \theta) u_1^{-1}\} \mathcal{IG}(u_1; \alpha(t), \beta(t|\mu, \theta)) du_1}{\mathcal{IG}(\sigma_\theta^2; \alpha(t), \beta(t|\mu, \theta))} \quad (40)$$

where $\alpha'(t)$ and $\beta'(t|\mu, \theta)$ denote the time derivatives of $\alpha(t)$ and $\beta(t|\mu, \theta)$, respectively,

$$\kappa(t|\mu, \theta) = \alpha'(t)\psi(\alpha(t)) - \alpha'(t) \log(\beta(t|\mu, \theta)) - \alpha(t)\beta^{-1}(t|\mu, \theta)\beta'(t|\mu, \theta)$$

and ψ is the digamma function⁵. For parameters μ and θ which have Gaussian full conditional distri-

⁵We evaluate this function using the `digamma` function in the R `base` package.

butions, the corresponding components of the Gibbs velocity field are more explicit

$$\tilde{f}_2(t, \sigma_\theta^2, \mu, \theta) = \frac{\varsigma'(t|\sigma_\theta^2)}{\varsigma(t|\sigma_\theta^2)}(\mu - \nu(t|\sigma_\theta^2, \theta)) + \nu'(t|\sigma_\theta^2, \theta), \quad (41)$$

$$\tilde{f}_3(t, \sigma_\theta^2, \mu, \theta) = \frac{\tau'(t|\sigma_\theta^2)}{\tau(t|\sigma_\theta^2)}(\theta - \xi(t|\sigma_\theta^2, \mu, y)) + \xi'(t|\sigma_\theta^2, \mu, y), \quad (42)$$

where ς', ν', τ' and $\xi' = (\xi'_1, \dots, \xi'_K)$ denote the time derivatives of ς, ν, τ and $\xi = (\xi_1, \dots, \xi_K)$, respectively. To approximate the Gibbs flow, we update the parameter σ_θ^2 at time step $m = 1, \dots, M$ using the Euler discretization (27), which defines the map

$$\Psi_{m,1}(\sigma_\theta^2, \mu, \theta) = (\sigma_\theta^2 + h\hat{f}_1(t_{m-1}, \sigma_\theta^2, \mu, \theta), \mu, \theta),$$

where \hat{f}_1 denotes an approximation of (40) by truncating the domain of integration $[0, \sigma_\theta^2]$ to $[\varepsilon, \sigma_\theta^2]$, where ε denotes machine epsilon, and applying a composite trapezoidal rule with $R = 50$ quadrature points. To update the parameter μ or θ conditionally on other parameters, since the solution of (28) under the linear velocity (41) or (42) is tractable, we have

$$\begin{aligned} \Psi_{m,2}(\sigma_\theta^2, \mu, \theta) &= \left(\sigma_\theta^2, \frac{\varsigma(t_m|\sigma_\theta^2)}{\varsigma(t_{m-1}|\sigma_\theta^2)}(\mu - \nu(t_{m-1}|\sigma_\theta^2, \theta)) + \nu(t_m|\sigma_\theta^2, \theta), \theta \right), \\ \Psi_{m,3}(\sigma_\theta^2, \mu, \theta) &= \left(\sigma_\theta^2, \mu, \frac{\tau(t_m|\sigma_\theta^2)}{\tau(t_{m-1}|\sigma_\theta^2)}(\theta - \xi(t_{m-1}|\sigma_\theta^2, \mu, y)) + \xi(t_m|\sigma_\theta^2, \mu, y) \right), \end{aligned}$$

for $m = 1, \dots, M$. In contrast to the generic expressions in (23)-(24), approximating the Gibbs flow for this model only requires computing summary statistics and evaluating inverse Gamma densities.

We consider a dataset of $K = 18$ baseball players' batting averages ($J = 1$) taken from [54, Table 1]. In this case, the number of parameters to be inferred is $d = K + 2 = 20$. Following [50], we adopt the empirical estimate $\sigma_e^2 = 4.34 \times 10^{-3}$ and a prior specification corresponding to $\alpha_0 = -1, \beta_0 = 2, \mu_0 = 0, \sigma_0 = 10$. We initialize the Gibbs flow using an “artificial” prior distribution (38) with $\alpha_1 = \beta_1 = 4, \mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 0.1$. By simulating 1024 independent trajectories using an adaptive integrator, we find that the ODE is not particularly stiff in this application. The median number of time steps taken was 11 with an IQR of 1, and amongst 11% of the trajectories that required step sizes that were smaller than the default value, the median of the minimum step size used over time was 5.0×10^{-3} with an IQR of 4.1×10^{-3} . Based on preliminary runs, $M = 50$ time steps was sufficient to ensure that trajectories are stable.

In Figure 8, we display the performance of GF-SIS (Algorithm 1) with $N = 128$ samples and $M = 50$ time steps. To improve performance with fixed N and M , we combine approximate Gibbs flow with HMC kernels⁶ within GF-AIS (Algorithm 2): this increases the ESS% from 63% to 97% on average, and reduces the variance of the log-marginal likelihood estimator by a factor of 22, at the expense of 4 times the compute time of GF-SIS. Plots of the norm of the Gibbs velocity field along trajectories generated by GF-SIS and GF-AIS can be found in Supplementary Material H. As competing algorithm, we consider AIS with the same N, M and HMC kernels as GF-AIS, but we increase the number of HMC iterations at each time step to match the computational time of GF-AIS, so as to ensure a fair comparison. Based on 100 independent repetitions of all three algorithms, the sample variance of log-marginal likelihood estimates relative to AIS was observed to be 1255 and 27,928 times smaller for GF-SIS and GF-AIS, respectively.

We then investigate how the performance of these algorithms behaves with dimension on simulated data with the same model specification. Figure 9 shows that numerical integration of the Gibbs flow becomes more stiff as we increase dimension, which is to be expected. From the left panel, as the dimension d increases, we observe only a logarithmic increase in the median number of time steps

⁶We apply a Hamiltonian Monte Carlo kernel at each time step. To achieve suitable acceptance probabilities, we use a step size of 0.05 for the leapfrog integrator and an integration time of 0.5.

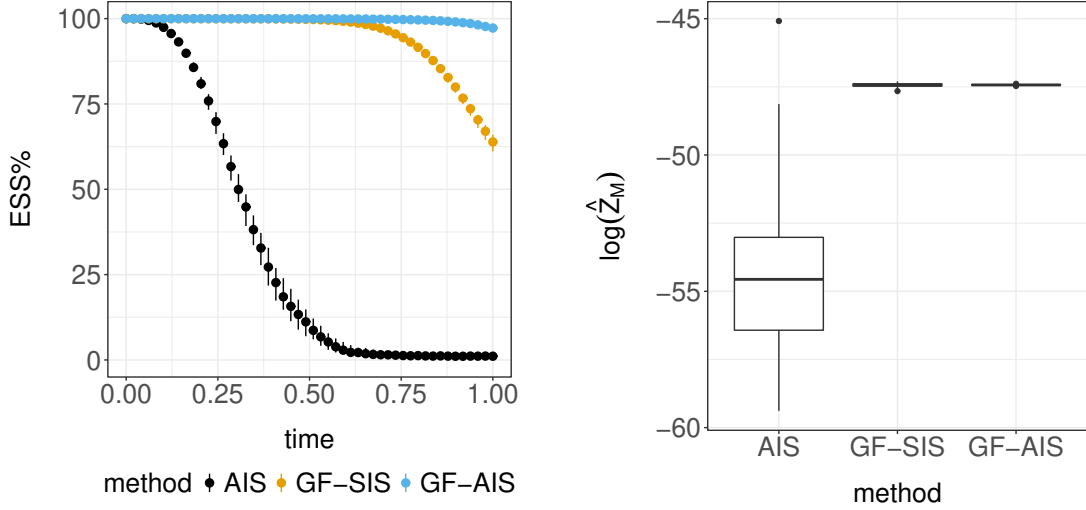


Figure 8: Boxplots of effective sample size percentage (*left*) and log-marginal likelihood estimates (*right*) when fitting the variance component model of Section 5.2 on the baseball dataset, obtained with 100 independent repetitions of AIS, GF-SIS (Algorithm 1) and GF-AIS (Algorithm 2).

taken by an adaptive integrator. Moreover, the proportion of trajectories that require smaller step sizes than the default value appear to be stable with dimension (around 10%), and for these trajectories, the right panel shows that the median of the minimum step size is stable as d increases. Plots of the norm of the Gibbs velocity field along trajectories generated by GF-SIS and GF-AIS in Supplementary Material H indicate that the ODE becomes more stiff in the beginning as dimension increases.

The algorithmic settings remain the same as we scale $d \in \{27, 52, 102, 202, 402\}$, with the exception of increasing time steps $M \in \{125, 250, 500, 750, 1000\}$ logarithmically with d (following the scaling in the left panel of Figure 9) and decreasing the step size of the leapfrog integrator in HMC to achieve stable acceptance probabilities⁷. Like before, we select the number of HMC iterations in AIS to match the compute time of GF-AIS; both algorithms require approximately $\{5, 7, 14, 15, 16\}$ times more compute time than GF-SIS as d varies. Figure 10 summarizes how the performance of these algorithms scale with dimension. Relative to standard AIS, GF-SIS performed better in all of the observed dimensions despite costing less compute time, and GF-AIS offers much better ESS% and variance reduction of several orders at a fixed computational cost.

5.3 Log-Gaussian Cox point processes

Lastly, we present an application of our methodology on a model from spatial statistics. In particular, we consider Bayesian inference for log-Gaussian Cox point processes on a dataset⁸ concerning the locations of 126 Scots pine saplings in a natural forest in Finland [39, 14, 26]. The actual square plot of 10×10 square metres is standardized to the unit square and discretized into a $J \times J$ regular grid. Given a latent intensity process $\Lambda_j, j \in \{1, \dots, J\}^2$, the number of points in each cell Y_j for $j \in \{1, \dots, J\}^2$, is modeled as conditionally independent and Poisson distributed with mean $a\Lambda_j$, where $a = J^{-2}$ is the area of each cell. The prior distribution of the intensity is specified by the relation $\Lambda_j = \exp(X_j)$ for $j \in \{1, \dots, J\}^2$, where $X = (X_j) \in \mathbb{R}^{J \times J}$ is a Gaussian process with

⁷For $d \in \{27, 52, 102, 202, 402\}$, we use 10 steps of the leapfrog integrator with step size $\{0.0125, 0.0100, 0.0075, 0.0050, 0.0025\}$, respectively.

⁸The dataset can be found in the R package `spatstat` as `finpines`.

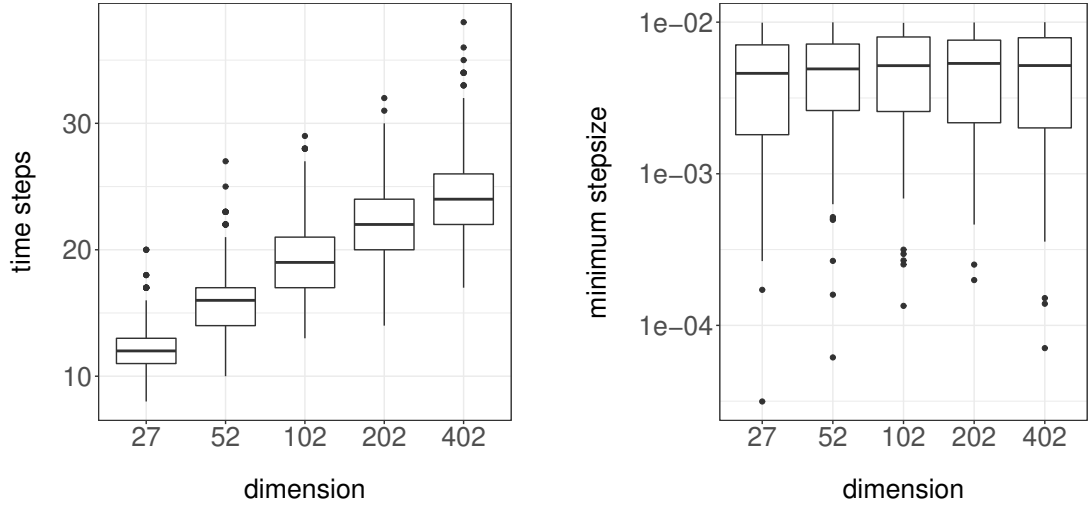


Figure 9: Simulating trajectories under the Gibbs flow using an adaptive fourth-order Runge-Kutta numerical integrator when fitting the variance component model of Section 5.2 on simulated data in various dimensions. Boxplots of the number of time steps taken by the adaptive integrator (*left*) and the minimum step size used over time (*right*) based on 1024 independent trajectories. Note that the boxplots in the right panel consider only trajectories which required step sizes that are smaller than the default value.

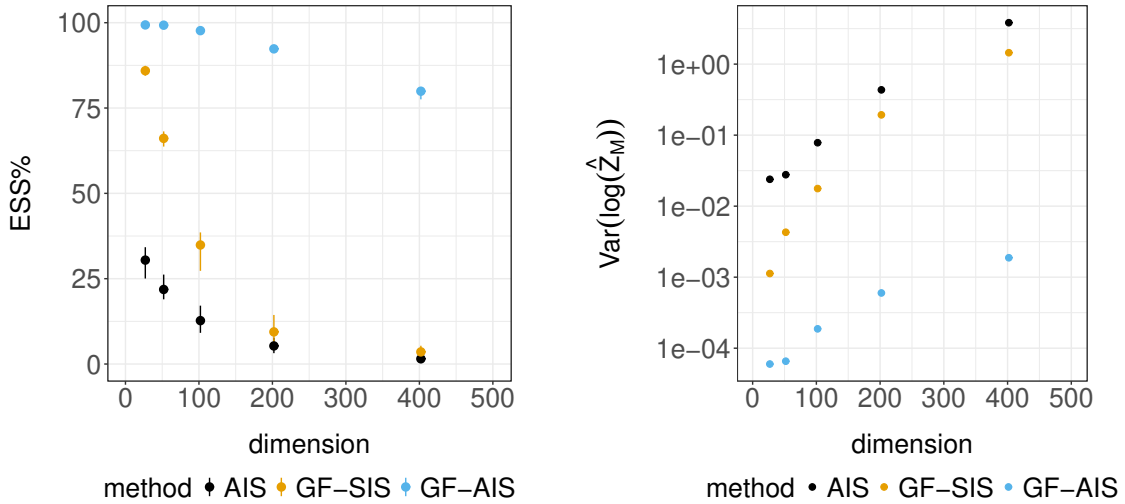


Figure 10: Boxplots of terminal effective sample size percentage (*left*) and variance of log-marginal likelihood estimates (*right*) when fitting the variance component model of Section 5.2 on simulated data in various dimensions, obtained with 100 independent repetitions of AIS, GF-SIS (Algorithm 1) and GF-AIS (Algorithm 2).

constant mean $\mu_0 \in \mathbb{R}$ and exponential covariance function $\Sigma_0(i, j) = \sigma^2 \exp(-|i - j|/(J\beta))$ for $i, j \in \{1, \dots, J\}^2$ and $\sigma^2, \beta > 0$. We will adopt the parameter values $\sigma^2 = 1.91$, $\beta = 1/33$ and $\mu_0 = \log(126) - \sigma^2/2$ estimated by [39]. This application corresponds to working in dimension $d = J^2$ with a prior distribution of $p_0(x) = \mathcal{N}(x; \mu_0 1_d, \Sigma_0)$ where $1_d = (1, \dots, 1) \in \mathbb{R}^d$ and a likelihood function of $p(y|x) = \prod_{j \in \{1, \dots, J\}^2} p(y_j|x_j) = \prod_{j \in \{1, \dots, J\}^2} \exp(x_j y_j - a \exp(x_j))$, where $y = (y_j) \in \mathbb{N}_0^{J \times J}$ denotes the dataset.

We will apply the methodology described in Section 2.1 with initialization from the prior distribution $\pi_0 = p_0$ or a Gaussian approximation of the posterior distribution; given by either a mean field variational Bayes (VB) approximation [55] $\pi_0(x) = \prod_{j \in \{1, \dots, J\}^2} \mathcal{N}(x_j; \mu_j, \sigma_j^2)$, or of the form $\pi_0(x) \propto p_0(x) \prod_{j \in \{1, \dots, J\}^2} \mathcal{N}(x_j; \mu_j, \sigma_j^2)$, where $(\mu_j, \sigma_j^2), j \in \{1, \dots, J\}^2$ are fitted using expectation-propagation (EP) [38], as advocated in [13]. To accommodate these choices, we take an “artificial” likelihood function $L(x) = p_0(x)p(y|x)/\pi_0(x)$ to define the curve of distributions $\{\pi_t\}_{t \in [0,1]}$ in (2). Although the full conditional distributions of $\pi_t, t \in [0, 1]$ are not in the exponential family, computation of the Gibbs velocity field (23)-(24) for one-dimensional components can be greatly simplified by rewriting

$$\tilde{f}_i(t, x) = \frac{\lambda'(t) \left\{ F_t(x_i|x_{-i}) \int_{-\infty}^{\infty} \log L_i(u_i, x_{-i}) \pi_t(u_i|x_{-i}) du_i - \int_{-\infty}^{x_i} \log L_i(u_i, x_{-i}) \pi_t(u_i|x_{-i}) du_i \right\}}{\pi_t(x_i|x_{-i})} \quad (43)$$

where $L_i(x) = p_0(x_i|x_{-i})p(y_i|x_i)/\pi_0(x_i|x_{-i})$ and

$$\frac{\pi_t(u_i|x_{-i})}{\pi_t(x_i|x_{-i})} = \frac{\gamma_t(u_i, x_{-i})}{\gamma_t(x_i, x_{-i})} = \frac{\pi_0(u_i|x_{-i})^{1-\lambda(t)} p_0(u_i|x_{-i})^{\lambda(t)} p(y_i|u_i)^{\lambda(t)}}{\pi_0(x_i|x_{-i})^{1-\lambda(t)} p_0(x_i|x_{-i})^{\lambda(t)} p(y_i|x_i)^{\lambda(t)}}$$

for $i = 1, \dots, d$, and noting that the full conditional distributions $\{p_0(x_i|x_{-i})\}_{i=1, \dots, d}$ and $\{\pi_0(x_i|x_{-i})\}_{i=1, \dots, d}$ are univariate Gaussians that can be precomputed. We approximate (43) by truncating \mathbb{R} to the interval $[\mu_0 - 6\sigma, \mu_0 + 6\sigma]$ and applying a composite trapezoidal rule with $R = 40$ quadrature points. The resulting ODE is then approximated using the Euler discretization (27), which defines the map $\Psi_{m,i}$ for time step $m = 1, \dots, M$ and component $i = 1, \dots, d$.

We first consider initialization from the prior $\pi_0 = p_0$ and vary the spatial resolution by taking $d \in \{10^2, 15^2, 20^2\}$. Figure 11 shows that numerical integration of the Gibbs flow does not become more stiff as we increase the spatial resolution, which is not surprising as the limiting infinite dimensional model is well-defined. Figure 13 displays the performance of GF-SIS (Algorithm 1) using $N = 512$ samples and as we increase the time steps $M \in \{40, 60, 80\}$ with dimension correspondingly. These time steps were chosen using preliminary runs to ensure that trajectories are stable; the norm of the Gibbs velocity field along trajectories generated by GF-SIS, shown in the left panel of Figure 12, also suggest that it may be worth increasing M with d . To obtain better performance for the same number of samples N and time steps M , we combine approximate Gibbs flow with Riemann manifold Hamiltonian Monte Carlo (RM-HMC) kernels⁹ that employ the metric tensor $\Sigma_0^{-1} + a \exp(\mu_0 + \sigma^2/2) I_d$ [26], where $I_d \in \mathbb{R}^{d \times d}$ denotes the identity matrix. Although the resulting GF-AIS (Algorithm 2) requires approximately $\{25\%, 50\%, 100\%\}$ more compute time than GF-SIS as dimension increases, it is apparent from Figure 13 that it improves algorithmic performance by several orders of magnitude. Like before, we compare GF-AIS to an AIS with the same N, M and RM-HMC kernels, but to ensure a fair comparison, the number of RM-HMC iterations at each time step is increased to match computational time. The results summarized in Figure 13 indicate that GF-AIS can offer very significant numerical gains over standard AIS in all three dimensions considered.

Lastly, we investigate the impact of the initial distribution for a spatial resolution of $d = 20^2$. All algorithmic settings are the same as before except slight changes in the step size of the leapfrog integrator in HMC to achieve suitable acceptance probabilities¹⁰. Figures 12 (right panel), 14 and 15

⁹We apply a Riemann manifold Hamiltonian Monte Carlo kernel at each time step, with a leapfrog integrator step size of 0.25 and an integration time of 2.5.

¹⁰For initial distributions given by the prior, a VB and an EP approximation of the posterior distribution, we used 10 steps of the leapfrog integrator with step size 0.25, 0.25 and 0.20, respectively.

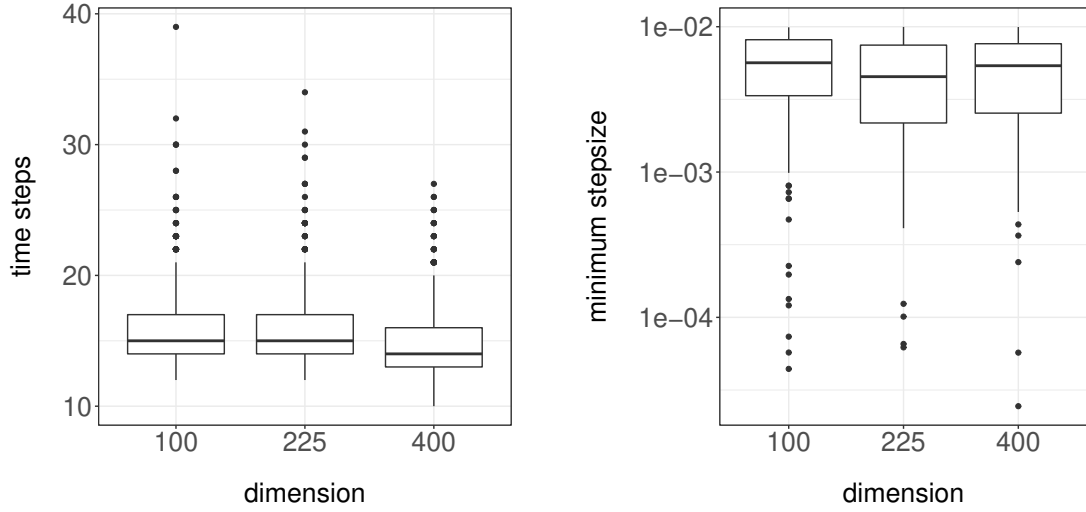


Figure 11: Simulating trajectories under the Gibbs flow using an adaptive fourth-order Runge-Kutta numerical integrator when fitting the log-Gaussian Cox point process model of Section 5.3 on the Scots pine saplings dataset with various spatial resolutions. Boxplots of the number of time steps taken by the adaptive integrator (*left*) and the minimum step size used over time (*right*) based on 1024 independent trajectories. Note that the boxplots in the right panel consider only around 13% of the trajectories which required step sizes that are smaller than the default value.

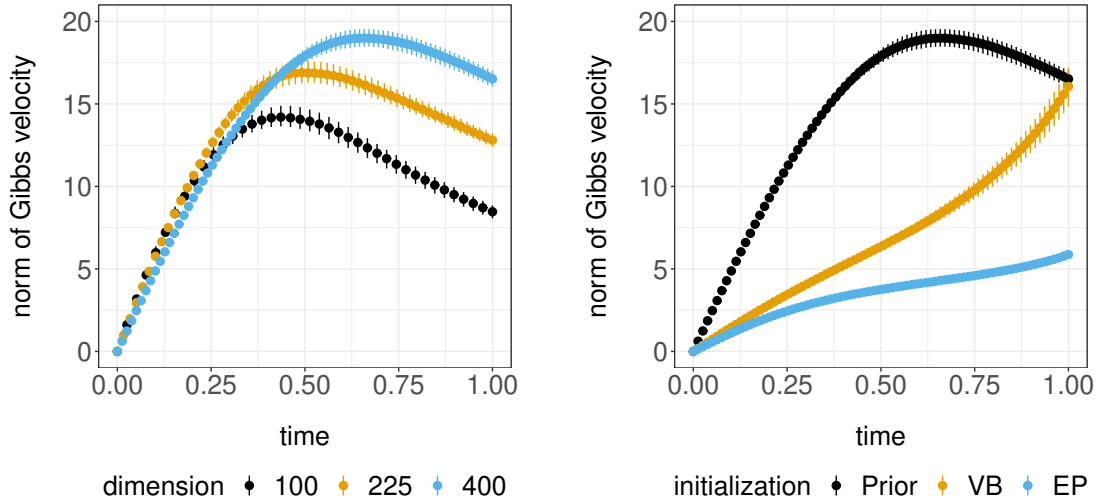


Figure 12: Boxplots illustrating the time evolution of the norm of the Gibbs velocity field along $N = 512$ trajectories generated by GF-SIS when fitting the log-Gaussian Cox point process model of Section 5.3 on the Scots pine saplings dataset with various spatial resolutions (*left*) and initial distributions (*right*).

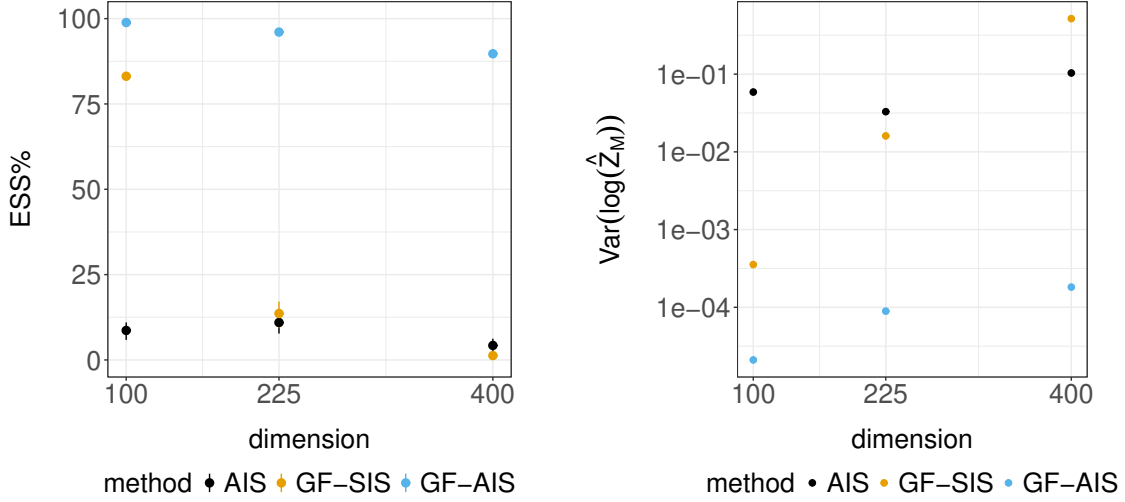


Figure 13: Boxplots of terminal effective sample size percentage (*left*) and variance of log-marginal likelihood estimates (*right*) when fitting the log-Gaussian Cox point process model of Section 5.3 on the Scots pine saplings dataset with various spatial resolutions, obtained with 100 independent repetitions of AIS, GF-SIS (Algorithm 1) and GF-AIS (Algorithm 2).

show that initial distributions that are “closer” to the posterior than the prior typically lead to less stiff ODE systems and better algorithmic performance. Interestingly, due to the nature of the Gibbs flow approximation, we find that an EP approximation of the posterior provides a better initialization than a VB approximation.

References

- [1] L. Ambrosio. Transport equation and Cauchy problem for BV vector fields. *Inventiones Mathematicae*, 158(2):227–260, 2004.
- [2] L. Ambrosio, N. Gigli and G. Savaré. *Gradient Flows in Metric Spaces And in the Space of Probability Measures*. Lectures in Mathematics ETH Zurich. Birkhauser, 2005.
- [3] J. Ba, M. A. Erdogdu, M. Ghassemi, T. Suzuki, S. Sun, D. Wu, and T. Zhang. Towards characterizing the high-dimensional bias of kernel-based particle inference algorithms. In *2nd Symposium on Advances in Approximate Bayesian Inference*, 2019.
- [4] A. R. Barron and X. Luo. Adaptive annealing. In *Proceedings of the Allerton Conference on Communications, Computation and Control*, 665–673, 2007.
- [5] J. D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.
- [6] K. Bergemann and S. Reich. An ensemble Kalman-Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 21(3):213–219, 2012.
- [7] M. J. Betancourt. Adiabatic Monte Carlo. *arXiv preprint arXiv:1405.3489*, 2014.
- [8] O. Bokanowski and B. Grébert. Deformations of density functions in molecular quantum chemistry. *Journal of Mathematical Physics*, 37(4):1553–1573, 1996.

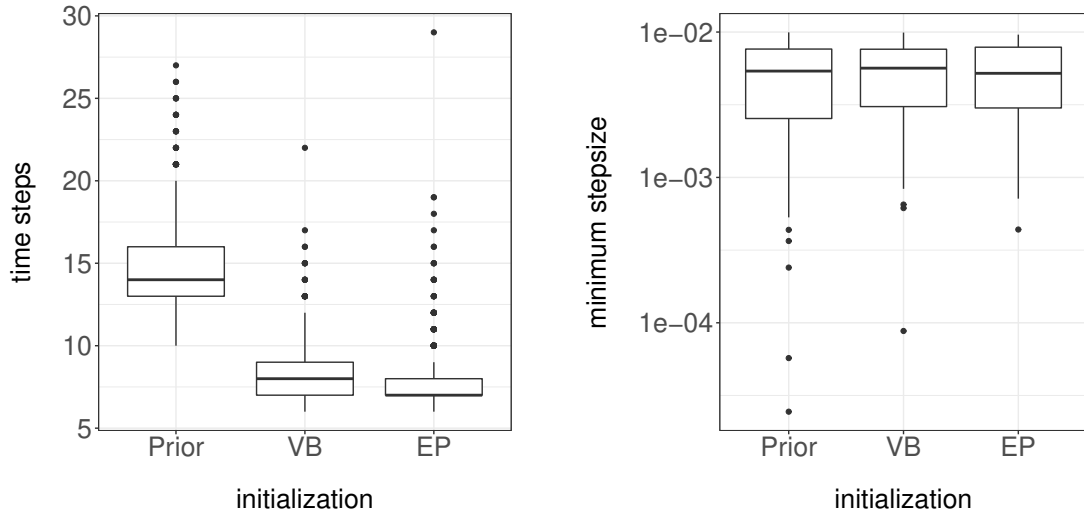


Figure 14: Simulating trajectories under the Gibbs flow using an adaptive fourth-order Runge-Kutta numerical integrator when fitting the log-Gaussian Cox point process model of Section 5.3 on the Scots pine saplings dataset with various initial distributions. Boxplots of the number of time steps taken by the adaptive integrator (*left*) and the minimum step size used over time (*right*) based on 1024 independent trajectories. Note that the boxplots in the right panel corresponding to initial distributions given by the prior, a variational Bayes (VB) and an expectation-propagation (EP) approximation of the posterior distribution, consider only 13%, 5% and 3%, respectively, of the trajectories which required step sizes that are smaller than the default value.

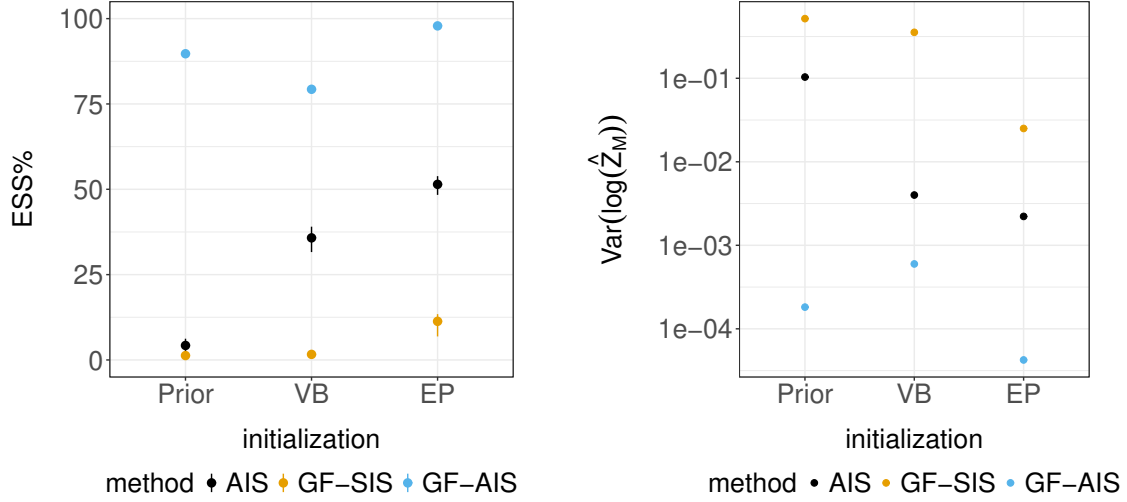


Figure 15: Boxplots of terminal effective sample size percentage (*left*) and variance of log-marginal likelihood estimates (*right*) when fitting the log-Gaussian Cox point process model of Section 5.3 on the Scots pine saplings dataset with various initial distributions, obtained with 100 independent repetitions of AIS, GF-SIS (Algorithm 1) and GF-AIS (Algorithm 2). The initial distributions considered here are the prior, a variational Bayes (VB) and an expectation-propagation (EP) approximation of the posterior distribution.

- [9] P. Bunch and S. J. Godsill. Approximations of the optimal importance density using Gaussian particle flow importance sampling. *Journal of the American Statistical Association*, 111(514): 748–762, 2016.
- [10] G. Celeux, M. Hurn and C. P. Robert. Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.
- [11] A. Chkifa, A. Cohen and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *Journal de Mathématiques Pures et Appliquées*, 103(2):400–428, 2015.
- [12] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.
- [13] N. Chopin and J. Ridgway. Leave Pima Indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science*, 32(1):64–87, 2017.
- [14] O. F. Christensen, G. O. Roberts and J. S. Rosenthal. Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268, 2005.
- [15] D. Crisan and J. Xiong. Approximate McKean–Vlasov representations for a class of SPDEs. *Stochastics*, 82(1):53–68, 2010.
- [16] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*, 90(5–6):1481–1487, 1998.
- [17] B. Dacorogna and J. Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l’Institut Henri Poincaré C (Analyse non linéaire)*, 7:1–26, 1990.

- [18] W. Dahmen, R. Devore, L. Grasedyck and E. Süli. Tensor-sparsity of solutions to high-dimensional elliptic partial differential equations. *Foundations of Computational Mathematics*, 1–62, 2014.
- [19] F. Daum and J. Huang. Particle flow and Monge-Kantorovich transport. In *Proceedings Conference on Information Fusion*, 135–142, 2012.
- [20] P. Del Moral, A. Doucet and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.
- [21] R. J. DiPerna and P. L. Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones Mathematicae*, 98(3):511–547, 1989.
- [22] T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.
- [23] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. 2nd edition, Springer, 2002.
- [24] A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.
- [25] A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.
- [26] 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.
- [27] R. E. Greene and K. Shiohama. Diffeomorphisms and volume-preserving embeddings of noncompact manifolds. *Transactions of the American Mathematical Society*, 255:403–403, 1979.
- [28] J. Han and Q. Liu. Stein Variational Adaptive Importance Sampling. In *Uncertainty in Artificial Intelligence*, 2017.
- [29] A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 2009.
- [30] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14), 2690, 1997.
- [31] S. Kim, R. Ma, D. Mesa and T. P. Coleman. Efficient Bayesian inference methods via convex optimization and optimal transport. In *IEEE International Symposium on Information Theory Proceedings*, pages 2259–2263. IEEE, 2013.
- [32] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [33] H. Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.
- [34] A. Kong, J. S. Liu and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.
- [35] A. Lee, C. Yau, M. B. Giles, A. Doucet and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.
- [36] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, pages 2370–2378, 2016.

- [37] X. L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002.
- [38] T. Minka. Expectation propagation for approximate Bayesian inference. *Proceedings of Uncertainty in Artificial Intelligence*, 17:362–369, 2001.
- [39] J. Møller, A. R. Syversveen and R. P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.
- [40] J. Moser. On the volume elements on a manifold. *Transactions of the American Mathematical Society*, 120(2):286–294, 1965.
- [41] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.
- [42] E. Novak and H. Wozniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *Journal of Complexity*, 25(4):398–404, 2009.
- [43] C. J. Oates, T. Papamarkou and M. Girolami. The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111(514):634–645, 2016.
- [44] M. Parno, T. Moselhy and Y. M. Marzouk. A multiscale strategy for Bayesian inference using transport maps. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1160–1190, 2016.
- [45] M. Parno and Y. M. Marzouk. Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.
- [46] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.
- [47] S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Analysis*, 51(1):235–249, 2011.
- [48] S. Reich. A Gaussian-mixture ensemble transform filter. *Quarterly Journal of the Royal Meteorological Society*, 138(662):222–233, 2012.
- [49] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472, 1952.
- [50] J. S. Rosenthal. Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing*, 6(3):269–275, 1996.
- [51] S. Reich and C. J. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 2015.
- [52] A. Spantini, D. Bigoni and Y. M. Marzouk. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19(66):1–71, 2018.
- [53] Y. Tao, P. G. Mehta and S. P. Meyn. Feedback particle filter. *IEEE Transactions on Automatic Control*, 58(10):2465–2480, 2013.
- [54] C. N. Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.
- [55] M. Teng, F. Nathoo, and T. D. Johnson. Bayesian computation for Log-Gaussian Cox processes: a comparative analysis of methods. *Journal of Statistical Computation and Simulation*, 87(11), 2227–2252, 2017.

- [56] S. Vaikuntanathan and C. Jarzynski. Escorted free energy simulations: Improving convergence by reducing dissipation. *Physical Review Letters*, 100(19), 190601, 2008.
- [57] S. Vaikuntanathan and C. Jarzynski. Escorted free energy simulations. *Journal of Chemical Physics*, 134(5), 054107, 2011.
- [58] C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.
- [59] W. Walter. *Ordinary Differential Equations*. Springer, 1998.
- [60] Y. Zhou, A. M. Johansen and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.