Variational Formulation of the Particle Flow Particle Filter

Yinzhuang Yi YIYI@UCSD.EDU
Contextual Robotics Institute
University of California, San Diego
La Jolla, CA 92093, USA

Jorge Cortés CORTES@UCSD.EDU
Contextual Robotics Institute
University of California, San Diego
La Jolla, CA 92093, USA

Nikolay Atanasov NATANASOV@UCSD.EDU
Contextual Robotics Institute
University of California, San Diego
La Jolla, CA 92093, USA

Abstract

This paper provides a formulation of the particle flow particle filter from the perspective of variational inference. We show that the transient density used to derive the particle flow particle filter follows a time-scaled trajectory of the Fisher-Rao gradient flow in the space of probability densities. The Fisher-Rao gradient flow is obtained as a continuous-time algorithm for variational inference, minimizing the Kullback-Leibler divergence between a variational density and the true posterior density. When considering a parametric family of variational densities, the function space Fisher-Rao gradient flow simplifies to the natural gradient flow of the variational density parameters. By adopting a Gaussian variational density, we derive a Gaussian approximated Fisher-Rao particle flow and show that, under linear Gaussian assumptions, it reduces to the Exact Daum and Huang particle flow. Additionally, we introduce a Gaussian mixture approximated Fisher-Rao particle flow to enhance the expressive power of our model through a multi-modal variational density. Simulations on low- and high-dimensional estimation problems illustrate our results.

Keywords: Variational Inference, Fisher-Rao Gradient Flow, Particle Flow Particle Filter, Bayesian Inference, Nonlinear Filtering

1 Introduction

This work aims to unveil a relationship between variational inference (Jordan et al., 1999) and particle flow particle filtering (Daum and Huang, 2007, 2009) by providing a variational formulation of the particle flow particle filter. These techniques consider Bayesian inference problems in which we start with a prior density function $p(\mathbf{x})$, and given an observation $\mathbf{z}$ and associated likelihood density function $p(\mathbf{z}|\mathbf{x})$, we aim to compute a posterior density function $p(\mathbf{x}|\mathbf{z})$, following Bayes' rule.

Variational inference (Jordan et al., 1999; Wainwright et al., 2008; Blei et al., 2017) is a method for approximating complex posterior densities $p(\mathbf{x}|\mathbf{z})$ by optimizing a simpler variational density $q(\mathbf{x})$ to closely match the true posterior. The particle flow particle filter (Daum and Huang, 2007, 2009; Daum et al., 2010) approximates the posterior density $p(\mathbf{x}|\mathbf{z})$ using a set of discrete samples $\{\mathbf{x}_i\}_i$ called particles. The particles are initially sampled from

the prior density $p(\mathbf{x})$ and propagated according to a particle dynamics function satisfying the Fokker-Planck equation (Jazwinski, 2007).

Uncovering a variational formulation of the particle flow particle filter offers several advantages. It provides an expression for the approximate density after particle propagation, thereby enhancing the expressive capability of the particle flow particle filter. It offers an alternative to linearizaiton for handling nonlinear observation models. Additionally, by using a mixture density as the variational density $q(\mathbf{x})$, the variational formulation can capture multimodal posterior densities, making it well-suited for inference tasks that need to capture the likelihood of several possible outcomes.

Traditional approaches to Bayesian inference include the celebrated Kalman filter (Kalman, 1960), which makes linear Gaussian assumptions, and its extensions to nonlinear observation models — the extended Kalman filter (EKF) (Anderson and Moore, 2005), which linearizes the observation model, and the unscented Kalman filter (UKF) (Julier et al., 1995), which propagates a set of discrete samples from the prior, called sigma points, through the nonlinear observation model. While the EKF and UKF consider nonlinear observation models, they use a single Gaussian to approximate the posterior and, hence, cannot capture multimodal behavior. To handle multimodal posteriors, the Gaussian sum filter (Alspach and Sorenson, 1972) approximates the posterior density using a weighted sum of Gaussian components. Alternatively, particle filters, or sequential Monte Carlo methods, have been proposed, in which the posterior is approximated by a set of particles. The bootstrap particle filter (Gordon et al., 1993) is a classical example, which draws particles from a proposal distribution, commonly chosen to be the prior, and updates their weights using the observation likelihood. Particle filters can suffer from particle degeneracy, particularly when the state dimension is high or the measurements are highly informative (Bickel et al., 2008). To alleviate particle degeneracy, the Hamiltonian Monte Carlo (HMC) method (Neal, 2011; Betancourt and Girolami, 2015) generates the particles by simulating a hypothetical Hamiltonian system, which leverages the first-order gradient information of the posterior. However, as pointed out in Alenlöv et al. (2021), the HMC method can get stuck in local modes if there is multi-modality in the posterior density. Another method that actively migrates the particles while keeping particle weights unchanged is the particle flow particle filter (Daum and Huang, 2007, 2009; Daum et al., 2010). Unlike the HMC method, the particles are migrated according to the particle dynamics function satisfying the Fokker-Planck equation (Jazwinski, 2007). To further improve the performance of the particle flow particle filter, the invertible particle flow (Li and Coates, 2017) uses the particle ensemble after the flow as samples from a proposal distribution and updates weights based on the unnormalized posterior and the proposal distribution. However, the particle flow particle filter's reliance on a Gaussian prior density limits its effectiveness for multimodal densities. To address this, Gaussian mixture models have been incorporated into particle flow particle filter variants (Pal and Coates, 2017, 2018; Li et al., 2019; Zhang and Meyer, 2024).

Variational inference formulates the Bayesian inference problem as an optimization problem, whose solution minimizes the Kullback–Leibler divergence between the variational density and the posterior. The optimization problem can be solved with a closed-form coordinate descent algorithm only when the likelihood function is suitably structured (Wainwright et al., 2008). For more complex models, computing the necessary gradients can become intractable and, to overcome this challenge, Hoffman et al. (2013); Ranganath et al. (2014)

employ stochastic gradients obtained by automatic differentiation. To further improve the efficiency of variational inference, the natural gradient method has been applied (Hoffman et al., 2013; Lin et al., 2019a; Martens, 2020; Huang et al., 2022; Khan and Rue, 2023). Moreover, the optimization problem can be defined over the space of probability densities, i.e., the optimization variables belong to an infinite-dimensional manifold. This includes the Wasserstein gradient flow (Jordan et al., 1998; Ambrogioni et al., 2018; Lambert et al., 2022), which models the gradient dynamics on the space of probability densities with respect to the Wasserstein metric. As an alternative to parametric variational inference methods, particle-based variational inference methods represent the approximating distribution using a set of particles. This includes the Stein variational gradient descent (Liu and Wang, 2016) and diffusion-based variational inference (**??**). Our work here is based upon a particular instance of the variational inference method, namely Gaussian variational inference, where the variational density is restricted to Gaussian or Gaussian mixture densities (Barber and Bishop, 1997; Seeger, 1999; Opper and Archambeau, 2009; Lin et al., 2019a).

The main contribution of this paper is the development of a variational formulation for the particle flow particle filter by identifying the transient density used in its derivation as a time-scaled trajectory of the Fisher-Rao gradient flow minimizing the Kullback-Leibler divergence. This variational approach removes the Gaussian assumptions on the prior and the likelihood, allowing for flexibility to approach estimation problems with any prior-likelihood pair. Utilizing this flexibility, we propose an approximated Gaussian mixture particle flow particle filter to capture the multi-modal behavior of the posterior. Finally, we explore several important identities related to the particle flow particle filter, leading to an inverse- and derivative-free formulation of the Fisher-Rao particle flows.

The paper is organized as follows. In Section 2, we provide the necessary background on the particle flow particle filter and variational inference. Section 3 presents our first main result, identifying the transient density used to derive the particle flow particle filter as a time-scaled trajectory of the Fisher-Rao gradient flow. Building on this, Section 4 introduces a Gaussian Fisher-Rao particle flow, where the variational density is constrained to a single Gaussian distribution, and demonstrates that, under linear Gaussian assumptions, the Gaussian Fisher-Rao particle flow reduces to the Exact Daum and Huang flow (Daum et al., 2010). Section 5, then, proposes the approximated Gaussian mixture Fisher-Rao particle flow by restricting the variational density to a Gaussian mixture density. Section 6 develops an inverse- and derivative-free formulation of the proposed particle flows and shows that they preserve Gauss-Hermite particles and the Mahalanobis distance. Finally, Section 7 provides simulations on low- and high-dimensional examples to illustrate our results.

Notation

We let $\mathbb{R}$ denote the real numbers. We use boldface letters to denote vectors and capital letters to denote matrices. We denote the probability density function (PDF) of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance Σ by $p_{\mathcal{N}}(\,\cdot\,;\boldsymbol{\mu},\Sigma)$. We use the notation $f \in C^{\infty}$ to indicate that $\mathbf{x} \mapsto f(\mathbf{x})$ is a smooth (infinitely differentiable) function. We use $\nabla_{\mathbf{x}} f(\mathbf{x})$ to denote the gradient of $f(\mathbf{x})$ and $\nabla_{\mathbf{x}} \cdot f(\mathbf{x})$ to denote the divergence of $f(\mathbf{x})$. For a random variable $\mathbf{y}$ with PDF $p(\mathbf{y})$, we use $\mathbb{E}_{p(\mathbf{y})}[f(\mathbf{y})]$ to denote the expectation of $f(\mathbf{y})$. We follow the numerator layout when calculating derivatives. For example, for a function

$f : \mathbb{R}^n \to \mathbb{R}^m$, we have $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}$ and $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^\top} \in \mathbb{R}^{n \times m}$. Note that $\nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^\top}$. For a matrix $A \in \mathbb{R}^{n \times n}$, we use $\text{tr}(A)$ to denote its trace and $|A|$ to denote its determinant. We let $\mathbb{I}_k(\omega)$ denote the indicator function, which is 1 when $\omega = k$ and 0 otherwise.

2 Background

We consider a Bayesian estimation problem where $\mathbf{x} \in \mathbb{R}^n$ is a random variable with a prior PDF $p(\mathbf{x})$. Given a measurement $\mathbf{z} \in \mathbb{R}^m$ with a known measurement likelihood $p(\mathbf{z}|\mathbf{x})$, the posterior PDF of $\mathbf{x}$ conditioned on $\mathbf{z}$ is determined by Bayes' theorem (Bayes, 1763):

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})}, \quad (1)$$

where $p(\mathbf{z})$ is the marginal measurement PDF computed as $p(\mathbf{z}) = \int p(\mathbf{z}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$. Calculating the posterior PDF in (1) is often intractable since the marginal measurement PDF $p(\mathbf{z})$ cannot typically be computed in closed form, except when the prior $p(\mathbf{x})$ and the likelihood $p(\mathbf{z}|\mathbf{x})$ are a conjugate pair. To calculate the posterior of non-conjugate prior and likelihood, approximation methods are needed. This problem can be approached using two classes of methods, which we review next: particle-based methods (Ducet et al., 2009), where the posterior is approximated using a set of weighted particles, and variational inference methods (Blei et al., 2017), where the posterior is approximated using a parametric variational PDF.

2.1 Particle-Based Inference

We review two particle-based inference methods: discrete-time and continuous-time particle filters. The major difference between these methods, as their name suggests, is the type of time evolution of their update step. A concrete example of a discrete-time particle filter method is the bootstrap particle filter (Gordon et al., 1993), which employs a single-step update on the particle weights once a measurement is received. On the other hand, the particle flow particle filter (Daum and Huang, 2007) updates the particle location following an ordinary differential equation upon receiving a measurement.

2.1.1 Discrete-Time Formulation: Particle Filter

We review the bootstrap particle filter as an example of a discrete-time particle-based inference method but note that other methods exist (Särkkä and Svensson, 2023). The bootstrap particle filter is a classical particle-based inference method that approximates the Bayes' posterior (1) using a set of weighted particles, where we draw N independent identically distributed particles $\{\mathbf{x}_i\}_{i=1}^N$ from the prior $\mathbf{x}_i \sim p(\mathbf{x})$. The particle weights are determined by the following equation:

$$w_i = \frac{p(\mathbf{z}|\mathbf{x}_i)p(\mathbf{x}_i)}{\sum_{j=1}^N p(\mathbf{z}|\mathbf{x}_j)p(\mathbf{x}_j)}.$$

The empirical mean and covariance of the posterior can be calculated using the weighted particles:

$$\tilde{\boldsymbol{\mu}} = \sum_{i=1}^{N} w_i \mathbf{x}_i, \qquad \tilde{\Sigma} = \sum_{i=1}^{N} w_i (\mathbf{x}_i - \tilde{\boldsymbol{\mu}})(\mathbf{x}_i - \tilde{\boldsymbol{\mu}})^\top.$$

One would immediately notice that since the particles are sampled from the prior and their positions remain unchanged, if the high-probability region of the posterior differs significantly from the prior, then the particle set may either poorly represent the posterior or result in most particles having negligible weights. This challenge is known as *particle degeneracy* (Bickel et al., 2008). The issue of particle degeneracy arises from a significant mismatch between the density from which the particles are drawn and the Bayes' posterior. To alleviate this issue, particles should be sampled from a distribution that more closely aligns with the posterior. This leads to the sequential importance sampling particle filter (Kitagawa, 1996), where particles are sampled from a proposal density. However, constructing an effective proposal density that accurately aligns with the posterior is a challenging task. An alternative is to move particles sampled from the prior to regions with high likelihood, which leads to the particle flow particle filter method, explained next.

2.1.2 Continuous-Time Formulation: Particle Flow Particle Filter

We review the particle flow particle filter (PFPF) proposed in Daum and Huang (2007, 2009); Daum et al. (2010) as an example of a continuous-time particle-based inference method. This method seeks to avoid the particle degeneracy faced by the bootstrap particle filter. The particles are moved based on a deterministic or stochastic differential equation, while the particle weights stay unchanged. We follow the exposition in Crouse and Lewis (2019), where the interested reader can find a detailed derivation. Before introducing the PFPF, it is important to understand the Liouville equation (Wibisono et al., 2017), as it provides the necessary background for its derivation. Consider a random process $\mathbf{x}_t \in \mathbb{R}^n$ that satisfies a drift-diffusion stochastic ordinary differential equation:

$$\mathrm{d}\mathbf{x}_t = \phi(\mathbf{x}_t, t)\,\mathrm{d}t + B(\mathbf{x}_t, t)\,\mathrm{d}\mathbf{w}_t,$$

where $\phi(\mathbf{x}_t, t) \in \mathbb{R}^n$ is a drift function, $B(\mathbf{x}_t, t) \in \mathbb{R}^{n \times m}$ is a diffusion matrix function and $\mathbf{w}_t$ is standard Brownian motion. The PDF $p(\mathbf{x}, t)$ of $\mathbf{x}_t$ evolves according to the Fokker–Planck equation (Jazwinski, 2007):

$$\frac{\partial p(\mathbf{x};t)}{\partial t} = -\nabla_{\mathbf{x}} \cdot \big(p(\mathbf{x};t)\phi(\mathbf{x},t)\big) + \frac{1}{2}\nabla_{\mathbf{x}} \cdot \Big(\nabla_{\mathbf{x}} \cdot \big(B(\mathbf{x},t)B(\mathbf{x},t)^\top p(\mathbf{x};t)\big)\Big).$$

In this work, we only consider the case when the diffusion term B is zero. In such a case, the random process is described by the following ordinary differential equation:

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \phi(\mathbf{x}_t, t), \quad (2)$$

where we term $\phi(\mathbf{x}_t, t)$ the *particle dynamics function*. The Fokker-Planck equation then reduces to the Liouville equation (Wibisono et al., 2017):

$$\frac{\partial p(\mathbf{x};t)}{\partial t} = -\nabla_{\mathbf{x}} \cdot \big(p(\mathbf{x};t)\phi(\mathbf{x},t)\big). \quad (3)$$

In other words, for a particle evolving according to (2) with initialization $\mathbf{x}_0$ sampled from PDF $p_0(\mathbf{x})$, at a later time t, the particle is distributed according to the PDF $p(\mathbf{x};t)$, which satisfies (3) with $p(\mathbf{x};t=0)=p_0(\mathbf{x})$. Conversely, if the time evolution of a PDF $p(\mathbf{x};t)$ is specified, the corresponding particle dynamics can be determined by finding a particle dynamics function $\phi(\mathbf{x},t)$ that satisfies (3). This observation facilitates the development of an alternative approach to particle filtering, where particles are actively moved towards regions of higher probability.

To derive the PFPF, we first take the natural logarithm of both sides of Bayes' theorem (1):

$$\log(p(\mathbf{x}|\mathbf{z})) = \log(p(\mathbf{z}|\mathbf{x})) + \log((p(\mathbf{x})) - \log(p(\mathbf{z})).$$

Our objective is to find a particle dynamics function $\phi(\mathbf{x},t)$ such that, if a particle is sampled from the prior and evolves according to (2), then it will be distributed according to the Bayes' posterior at a certain later time. Building on the discussion above, finding this particle dynamics function requires specifying a transformation from the prior to the posterior. To specify the transformation, we introduce a pseudo-time parameter $0 \leq \lambda \leq 1$ and define a log-homotopy of the form:

$$\log(p(\mathbf{x}|\mathbf{z};\lambda)) = \lambda \log(p(\mathbf{z}|\mathbf{x})) + \log(p(\mathbf{x})) - \log(p(\mathbf{z};\lambda)),$$

where the marginal measurement density $p(\mathbf{z};\lambda)$ has been parameterized by the pseudo-time λ so that $p(\mathbf{x}|\mathbf{z};\lambda)$ is a valid PDF for all values of λ. This defines the following *transient density* describing the PFPF process:

$$p(\mathbf{x}|\mathbf{z};\lambda) = \frac{p(\mathbf{z}|\mathbf{x})^\lambda p(\mathbf{x})}{p(\mathbf{z};\lambda)}, \qquad p(\mathbf{z};\lambda) = \int p(\mathbf{z}|\mathbf{x})^\lambda p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{4}$$

The transient density (4) defines a smooth and continuous transformation from the prior $p(\mathbf{x}|\mathbf{z};0)=p(\mathbf{x})$ to the posterior $p(\mathbf{x}|\mathbf{z};1)=p(\mathbf{x}|\mathbf{z})$. Based on our discussion above, in order to obtain a particle flow describing the evolution of the transient density from the prior to the posterior, we need to find a particle dynamics function $\boldsymbol{\phi}(\mathbf{x},\lambda)$ such that the following equation holds:

$$\frac{\partial p(\mathbf{x}|\mathbf{z};\lambda)}{\partial \lambda} = -\nabla_{\mathbf{x}} \cdot \big(p(\mathbf{x}|\mathbf{z};\lambda)\boldsymbol{\phi}(\mathbf{x},\lambda)\big), \tag{5}$$

where $p(\mathbf{x}|\mathbf{z};\lambda)$ is the transient density given in (4).

Assumption 1 (Linear Gaussian Assumptions) *The prior PDF is Gaussian, given by $p(\mathbf{x}) = p_{\mathcal{N}}(\mathbf{x};\hat{\mathbf{x}},P)$ with mean $\hat{\mathbf{x}} \in \mathbb{R}^n$ and covariance $P \in \mathbb{R}^{n\times n}$, and the likelihood is also Gaussian, given by $p(\mathbf{z}|\mathbf{x}) = p_{\mathcal{N}}(\mathbf{z};H\mathbf{x},R)$ with mean $H\mathbf{x} \in \mathbb{R}^m$ and covariance $R \in \mathbb{R}^{m\times m}$.*

Finding a particle dynamics function $\boldsymbol{\phi}(\mathbf{x},\lambda)$ satisfying (5) for a general transient density is challenging. This is why we consider Assumption 1. Under linear Gaussian assumptions, the transient density is Gaussian with PDF given by $p(\mathbf{x}|\mathbf{z};\lambda) = p_{\mathcal{N}}(\mathbf{x};\boldsymbol{\mu}_\lambda,\Sigma_\lambda)$, where

$$\boldsymbol{\mu}_\lambda = \Sigma_\lambda(P^{-1}\hat{\mathbf{x}} + \lambda H^\top R^{-1}\mathbf{z}), \quad \Sigma_\lambda = P - \lambda PH^\top(R+\lambda HPH^\top)^{-1}HP. \tag{6}$$

Daum et al. (2010) provide a closed-form expression for the particle dynamics function $\boldsymbol{\phi}(\mathbf{x},\lambda)$ satisfying (3):

$$\phi(\mathbf{x},\lambda) = A_\lambda \mathbf{x} + \mathbf{b}_\lambda, \tag{7}$$

with

$$A_\lambda = -\frac{1}{2}PH^\top(R+\lambda HPH^\top)^{-1}H, \qquad \mathbf{b}_\lambda = (I+2\lambda A_\lambda)(A_\lambda\hat{\mathbf{x}} + (I+\lambda A_\lambda)PH^\top R^{-1}\mathbf{z}).$$

Within the particle flow literature, the solution above is termed the Exact Daum and Huang (EDH) flow. A detailed derivation of the EDH flow is missing in Daum et al. (2010). In subsequent works, the separation of variables method has been widely adopted to derive the EDH flow (Crouse and Lewis, 2019; Ward and DeMars, 2022; Zhang and Meyer, 2024), where the pseudo-time rate of change of the marginal density $p(\mathbf{z};\lambda)$ is ignored due to its independence of $\mathbf{x}$, indicating the EDH flow satisfies (3) up to some constant. For completeness, we show in the following result that the EDH flow satisfies (3) exactly.

Lemma 1 (EDH Flow is Exact) *Consider the transient density $p(\mathbf{x}|\mathbf{z};\lambda)$ given by* (4) *with $p(\mathbf{x}) = p_{\mathcal{N}}(\mathbf{x};\hat{\mathbf{x}},P)$ and $p(\mathbf{z}|\mathbf{x}) = p_{\mathcal{N}}(\mathbf{z};H\mathbf{x},R)$. Then, the particle dynamics function $\phi(\mathbf{x},\lambda)$ given by the EDH flow* (7) *satisfies* (5).

Proof Expanding both sides of equation (5) leads to:

$$\frac{\partial p(\mathbf{x}|\mathbf{z};\lambda)}{\partial\lambda} = p(\mathbf{x}|\mathbf{z};\lambda)\left(\log(p(\mathbf{z}|\mathbf{x})) - \frac{\partial\log(p(\mathbf{z};\lambda))}{\partial\lambda}\right),$$

$$-\nabla_{\mathbf{x}}\cdot(p(\mathbf{x}|\mathbf{z};\lambda)(A_\lambda\mathbf{x}+\mathbf{b}_\lambda)) = -p(\mathbf{x}|\mathbf{z};\lambda)\,\mathrm{tr}(A_\lambda) - \frac{\partial p(\mathbf{x}|\mathbf{z};\lambda)}{\partial\mathbf{x}}(A_\lambda\mathbf{x}+\mathbf{b}_\lambda).$$

Since $p(\mathbf{x}|\mathbf{z};\lambda)$ is a Gaussian density, we have:

$$\frac{\partial p(\mathbf{x}|\mathbf{z};\lambda)}{\partial\mathbf{x}} = -p(\mathbf{x}|\mathbf{z};\lambda)\left((\mathbf{x}-\boldsymbol{\mu}_\lambda)^\top\Sigma_\lambda^{-1}\right).$$

As a result, we only need to show that the following equality holds:

$$\log(p(\mathbf{z}|\mathbf{x})) - \frac{\partial\log(p(\mathbf{z};\lambda))}{\partial\lambda} = (\mathbf{x}-\boldsymbol{\mu}_\lambda)^\top\Sigma_\lambda^{-1}(A_\lambda\mathbf{x}+\mathbf{b}_\lambda) - \mathrm{tr}(A_\lambda).$$

Expanding both sides and re-arranging leads to

$$\frac{\partial\log(p(\mathbf{z};\lambda))}{\partial\lambda} + \frac{1}{2}\log(\|2\pi R\|) + \frac{1}{2}\mathbf{z}^\top R^{-1}\mathbf{z} = \mathrm{tr}(A_\lambda) + \lambda\mathbf{b}_\lambda^\top H^\top R^{-1}\mathbf{z} + \mathbf{b}_\lambda^\top P^{-1}\hat{\mathbf{x}}.$$

Expanding the $\mathbf{b}_\lambda$ term and re-arranging, we have:

$$\lambda\mathbf{b}_\lambda^\top H^\top R^{-1}\mathbf{z} + \mathbf{b}_\lambda^\top P^{-1}\hat{\mathbf{x}} = -\frac{1}{2}\boldsymbol{\mu}_\lambda^\top H^\top R^{-1}H\boldsymbol{\mu}_\lambda + \boldsymbol{\mu}_\lambda^\top H^\top R^{-1}\mathbf{z}.$$

The term $\mathrm{tr}(A_\lambda)$ has the following expression:

$$\mathrm{tr}(A_\lambda) = -\frac{1}{2}\mathbb{E}_{p(\mathbf{x}|\mathbf{z};\lambda)}\left[\mathbf{x}^\top H^\top R^{-1}H\mathbf{x}\right] + \frac{1}{2}\boldsymbol{\mu}_\lambda^\top H^\top R^{-1}H\boldsymbol{\mu}_\lambda.$$

By the definition of $\log(p(\mathbf{z};\lambda))$, we have

$$\frac{\partial\log(p(\mathbf{z};\lambda))}{\partial\lambda} + \frac{1}{2}\log(|2\pi R|) + \frac{1}{2}\mathbf{z}^\top R^{-1}\mathbf{z} = \boldsymbol{\mu}_\lambda^\top H^\top R^{-1}\mathbf{z} - \frac{1}{2}\mathbb{E}_{p(\mathbf{x}|\mathbf{z};\lambda)}\left[\mathbf{x}^\top H^\top R^{-1}H\mathbf{x}\right].$$

Finally, the result is obtained by combining the equations above with the transient parameters $\boldsymbol{\mu}_\lambda$ and Σ_λ given in (6). ■

2.2 Variational Inference

Variational inference (VI) methods approximate the posterior in (1) using a PDF $q(\mathbf{x})$ with an explicit expression (Bishop, 2006, Ch. 10), termed *variational density*. To perform the approximation, VI minimizes the Kullback-Leibler (KL) divergence between the true posterior $p(\mathbf{x}|\mathbf{z})$ and the variational density $q(\mathbf{x})$:

$$D_{KL}\left(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z})\right) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \, \mathrm{d}\mathbf{x}.$$

Formally, given the statistical manifold $\mathcal{P}$ of probability measures with smooth positive densities,

$$\mathcal{P} = \left\{ p(\mathbf{x}) \in C^{\infty} : \int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1, p(\mathbf{x}) > 0 \right\}, \quad (8)$$

we seek a variational density $q(\mathbf{x}) \in \mathcal{P}$ that solves the optimization problem:

$$\min_{q(\mathbf{x}) \in \mathcal{P}} D_{KL}(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z})). \quad (9)$$

If $p(\mathbf{x}|\mathbf{z}) \in \mathcal{P}$, then the optimal variational density $q^*(\mathbf{x})$ is given by $q^*(\mathbf{x}) = p(\mathbf{x}|\mathbf{z})$. Conversely, if $p(\mathbf{x}|\mathbf{z}) \notin \mathcal{P}$, the optimal variational density $q^*(\mathbf{x})$ satisfies $D_{KL}(q^*(\mathbf{x})\|p(\mathbf{x}|\mathbf{z})) \leq D_{KL}(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z}))$, for all $q(\mathbf{x}) \in \mathcal{P}$.

Optimizing directly over $\mathcal{P}$ is challenging due to its infinite-dimensional nature. VI methods usually select a parametric family of variational densities $q(\mathbf{x};\boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$ from an admissible parameter set Θ. Popular choices of variational densities include multivariate Gaussian (Opper and Archambeau, 2009) and Gaussian mixture (Lin et al., 2019a). This makes the optimization problem in (9) finite dimensional:

$$\min_{\boldsymbol{\theta} \in \Theta} D_{KL}(q(\mathbf{x};\boldsymbol{\theta})\|p(\mathbf{x}|\mathbf{z})). \quad (10)$$

We distinguish between discrete-time and continuous-time VI techniques. Discrete-time VI performs gradient descent on D_{KL} to update (the parameters of) the variational density, while continuous-time VI uses gradient flow.

2.2.1 Discrete-Time Formulation: Gradient Descent and Mean Field

A common approach to optimize the parameters in (10) is to use gradient descent:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \beta_t \frac{\partial D_{KL}\left(q(\mathbf{x};\boldsymbol{\theta})\|p(\mathbf{x}|\mathbf{z})\right)}{\partial \boldsymbol{\theta}^{\top}} \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \\ &= \boldsymbol{\theta}_t - \beta_t \mathbb{E}_{q(\mathbf{x};\boldsymbol{\theta})} \left[\frac{\partial \log(q(\mathbf{x};\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^{\top}} \left(\log\left(\frac{q(\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{z})} \right) + 1 \right) \right] \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}, \end{aligned} \quad (11)$$

where $\beta_t > 0$ is the step size. Starting with initial parameters $\boldsymbol{\theta}_0$, gradient descent updates $\boldsymbol{\theta}_t$ in the direction of the negative gradient of the KL divergence. However, for a non-conjugate prior and likelihood pair, the KL gradient is challenging to obtain due to the expectation in (11). Stochastic gradient methods (Hoffman et al., 2013; Ranganath et al., 2014) have been proposed to overcome this challenge. Stochastic gradient methods

use samples to approximate the expectation over $q(\mathbf{x};\boldsymbol{\theta})$ in (11). The KL divergence is convex with respect to the variational density $q(\mathbf{x};\boldsymbol{\theta})$ but it is not necessarily convex with respect to the density parameters $\boldsymbol{\theta}$. As a result, local convergence is to be expected. The convergence analysis of the general gradient descent method is available in Curry (1944).

When the dimension of $\mathbf{x}$ is high, using a fully parameterized variational density is computationally challenging. As an alternative to the parametric approach, we can use a mean-field variational density, where the elements in $\mathbf{x}$ are mutually independent and each is governed by a component of the mean-field variational density. Formally,

$$q(\mathbf{x};\boldsymbol{\theta}) = \prod_{j=1}^{n} q_j(x_j;\boldsymbol{\theta}_j),$$

where x_j denotes the jth component of $\mathbf{x}$ and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, ..., \boldsymbol{\theta}_n^\top]$. Due to the mutual independence of the variational density components, we can write the KL divergence as follows (Blei et al., 2017):

$$\begin{aligned} D_{KL}\left(q(\mathbf{x};\boldsymbol{\theta})\|p(\mathbf{x}|\mathbf{z})\right) = &\int q(\mathbf{x};\boldsymbol{\theta})\log(q(\mathbf{x};\boldsymbol{\theta}))\,\mathrm{d}\mathbf{x} + \log(p(\mathbf{z})) \\ &- \int q_j(x_j;\boldsymbol{\theta}_j)\int q_{-j}(\mathbf{x}_{-j};\boldsymbol{\theta}_{-j})\log(p(x_j,\mathbf{x}_{-j},\mathbf{z}))\,\mathrm{d}\mathbf{x}_{-j}\,\mathrm{d}x_j, \end{aligned}$$

where $\mathbf{x}_{-j} \in \mathbb{R}^{n-1}$ represents the part of $\mathbf{x}$ remaining after removing the jth component, and $q_{-j}(\mathbf{x}_{-j};\boldsymbol{\theta}_{-j}) = \prod_{l\neq j} q_l(x_l;\boldsymbol{\theta}_l)$ denotes the corresponding variational density of $\mathbf{x}_{-j}$. By consolidating terms that do not depend on the jth component of the variational density into a constant, we obtain:

$$\begin{aligned} &D_{KL}\left(q(\mathbf{x};\boldsymbol{\theta})\|p(\mathbf{x}|\mathbf{z})\right) \propto D_{KL}\left(q_j(x_j;\boldsymbol{\theta}_j)\|\exp\left(\int q_{-j}(\mathbf{x}_{-j};\boldsymbol{\theta}_{-j})\log(p(x_j,\mathbf{x}_{-j},\mathbf{z}))\,\mathrm{d}\mathbf{x}_{-j}\right)\right) \\ &= \int q_j(x_j;\boldsymbol{\theta}_j)\log(q_j(x_j;\boldsymbol{\theta}_j))\,\mathrm{d}x_j - \int q_j(x_j;\boldsymbol{\theta}_j)\int q_{-j}(\mathbf{x}_{-j};\boldsymbol{\theta}_{-j})\log(p(x_j,\mathbf{x}_{-j},\mathbf{z}))\,\mathrm{d}\mathbf{x}_{-j}\,\mathrm{d}x_j. \end{aligned}$$

As a result, the optimal jth component of the mean-field variational density minimizing the KL divergence satisfies the following condition:

$$q_j(x_j;\boldsymbol{\theta}_j) \propto \exp\Big(\int q_{-j}(\mathbf{x}_{-j};\boldsymbol{\theta}_{-j})\log(p(x_j,\mathbf{x}_{-j},\mathbf{z}))\,\mathrm{d}\mathbf{x}_{-j}\Big). \quad (12)$$

The coordinate ascent variation inference (CAVI) algorithm (Bishop, 2006) updates each component of the mean-field variation density using (12), which is a common approach to the mean-field variational inference. The work of Blei et al. (2017) provides a concise introduction with illustrative examples of such methods.

2.2.2 Continuous-Time Formulation: Fischer-Rao Gradient Flow

Next, we describe an alternative method for solving the optimization problem (9) formulated by the VI method. Following the exposition in Chen et al. (2023a), we consider a specific geometry over the space of probability densities and present a continuous-time gradient flow approach as an alternative to the usual discrete-time gradient descent method. At a point

$q \in \mathcal{P}$, consider the associated tangent space $T_q\mathcal{P}$ of the probability space $\mathcal{P}$ in (8). Note that

$$T_q\mathcal{P} \subseteq \Big\{\sigma \in C^\infty : \int \sigma(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0\Big\}.$$

The cotangent space $T_q^*\mathcal{P}$ is the dual of $T_q\mathcal{P}$. We can introduce a bilinear map $\langle\cdot,\cdot\rangle$ as the duality pairing $T_q^*\mathcal{P} \times T_q\mathcal{P} \to \mathbb{R}$. For any $\psi \in T_q^*\mathcal{P}$ and $\sigma \in T_q\mathcal{P}$, the duality pairing between $T_q^*\mathcal{P}$ and $T_q\mathcal{P}$ can be identified in terms of L^2 integration as $\langle\psi,\sigma\rangle = \int \psi\sigma\,\mathrm{d}\mathbf{x}$. The most important element in $T_q^*\mathcal{P}$ we consider is the first variation of the KL divergence $\frac{\delta D_{KL}(q||p)}{\delta q}$,

$$\left\langle \frac{\delta D_{KL}(q||p)}{\delta q}, \sigma \right\rangle = \lim_{\epsilon\to 0} \frac{D_{KL}(q+\epsilon\sigma||p) - D_{KL}(q||p)}{\epsilon}, \quad (13)$$

for $\sigma \in T_q\mathcal{P}$. Given a metric tensor at q, denoted by $M(q) : T_q\mathcal{P} \to T_q^*\mathcal{P}$, we can express the Riemannian metric $g_q : T_q\mathcal{P} \times T_q\mathcal{P} \to \mathbb{R}$ as $g_q(\sigma_1,\sigma_2) = \langle M(q)\sigma_1, \sigma_2\rangle$. Consider the Fisher-Rao Riemannian metric (Amari, 2016):

$$g_q^{\mathrm{FR}}(\sigma_1,\sigma_2) = \int \frac{\sigma_1(\mathbf{x})\sigma_2(\mathbf{x})}{q(\mathbf{x})}\,\mathrm{d}\mathbf{x}, \quad \text{for } \sigma_1,\sigma_2 \in T_q\mathcal{P}. \quad (14)$$

The choice of elements in $T_q^*\mathcal{P}$ is not unique under L^2 integration, since $\langle\psi,\sigma\rangle = \langle\psi + c,\sigma\rangle$ for all $\psi \in T_q^*\mathcal{P},\ \sigma \in T_q\mathcal{P}$ and any constant c. However, a unique representation can be obtained by requiring $\int q\psi\,\mathrm{d}\mathbf{x} = 0$, and in that case, the metric tensor associated with the Fisher-Rao Riemannian metric g_q^{FR} is the Fisher-Rao metric tensor $M^{\mathrm{FR}}(q)$, which satisfies the following equation (Chen et al., 2023b):

$$M^{\mathrm{FR}}(q)^{-1}\psi = q\psi \in T_q\mathcal{P}, \quad \forall\psi \in T_q^*\mathcal{P}. \quad (15)$$

The gradient of the KL divergence under the Fisher-Rao Riemannian metric, denoted by $\nabla_q^{\mathrm{FR}} D_{KL}$, is defined according to the following condition:

$$g_q^{\mathrm{FR}}\left(\nabla_q^{\mathrm{FR}} D_{KL}, \sigma\right) = \left\langle \frac{\delta D_{KL}(q||p)}{\delta q}, \sigma \right\rangle, \quad \forall\sigma \in T_q\mathcal{P}.$$

Proposition 2 (Gradient of KL Divergence) *The Fisher-Rao Riemannian gradient of the KL divergence $D_{KL}(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z}))$ is given by*

$$\nabla_q^{\mathrm{FR}} D_{KL}(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z})) = q(\mathbf{x})\left(\log\left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})}\right) - \mathbb{E}_{q(\mathbf{x})}\left[\log\left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})}\right)\right]\right), \quad (16)$$

where $p(\mathbf{x}|\mathbf{z})$ is the Bayes' posterior given by (1).

Proof The Fréchet derivative of the KL divergence $D_{KL}(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z}))$ is given by

$$\lim_{\epsilon\to 0} \frac{D_{KL}(q(\mathbf{x}) + \epsilon\sigma(\mathbf{x})||p(\mathbf{x}|\mathbf{z})) - D_{KL}(q(\mathbf{x})||p(\mathbf{x}|\mathbf{z}))}{\epsilon} = \int \sigma(\mathbf{x})\left(1 + \log\left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})}\right)\right)\mathrm{d}\mathbf{x}.$$

Substituting this expression in (13) and using that $\int \sigma(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0$ (since $\sigma \in T_q\mathcal{P}$), we have

$$\int \frac{\delta D_{KL}(q(\mathbf{x})\|p(\mathbf{x}|\mathbf{z}))}{\delta q(\mathbf{x})}\sigma(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int \sigma(\mathbf{x})\log\left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})}\right)\mathrm{d}\mathbf{x}.$$

However, the first variation of the KL divergence is not uniquely determined because

$$\int \sigma(\mathbf{x}) \left(\log \left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \right) + c \right) \mathrm{d}\mathbf{x} = \int \sigma(\mathbf{x}) \log \left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \right) \mathrm{d}\mathbf{x}, \quad \forall c \in \mathbb{R}.$$

Based on our discussion above, the first variation of the KL divergence can be uniquely identified by further requiring that

$$\int q(\mathbf{x}) \left(\log \left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \right) + c \right) \mathrm{d}\mathbf{x} = 0,$$

which leads to the selection $c = -\mathbb{E}_{q(\mathbf{x})} \left[\log \left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \right) \right]$. Therefore, the first variation of the KL divergence is given by

$$\frac{\delta D_{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{z}))}{\delta q(\mathbf{x})} = \log \left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \right) - \mathbb{E}_{q(\mathbf{x})} \left[\log \left(\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{z})} \right) \right]. \quad (17)$$

Using (15), the Fisher-Rao gradient of the KL divergence $D_{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{z}))$ is then given by

$$\nabla_q^{\mathrm{FR}} D_{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{z})) = q(\mathbf{x}) \frac{\delta D_{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{z}))}{\delta q(\mathbf{x})}.$$

Substituting (17) into the equation above, we get the desired result. ■

The Fisher-Rao gradient flow in the space of probability measures $\mathcal{P}$ takes the following form:

$$\frac{\partial q(\mathbf{x}; t)}{\partial t} = -\nabla_q^{\mathrm{FR}} D_{KL}(q(\mathbf{x}; t) \| p(\mathbf{x}|\mathbf{z})). \quad (18)$$

Further, we consider the case where the variational density is parameterized by $\boldsymbol{\theta}$, and denote the space of $\boldsymbol{\theta}$-parameterized positive probability densities as:

$$\mathcal{P}_{\boldsymbol{\theta}} = \{ p(\mathbf{x}; \boldsymbol{\theta}) \in \mathcal{P} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^l \} \subset \mathcal{P}$$

The basis of $T_{p(\mathbf{x};\boldsymbol{\theta})} \mathcal{P}_{\boldsymbol{\theta}}$ is given by

$$\left\{ \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_2}, \ldots, \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_l} \right\},$$

where θ_i denotes the ith element of $\boldsymbol{\theta}$. As a result, the Fisher-Rao metric tensor $M^{FR}(\boldsymbol{\theta}) : \Theta \to \mathbb{R}^{l \times l}$ under parametrization $\boldsymbol{\theta}$ is given as follows (Nielsen, 2020):

$$M^{FR}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})} \left[\frac{\partial \log(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^\top} \frac{\partial \log(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right], \quad (19)$$

which is identical to the Fisher Information Matrix (FIM), denoted $\mathcal{I}(\boldsymbol{\theta}) = M^{FR}(\boldsymbol{\theta})$. Restricting the variational density to the space of $\boldsymbol{\theta}$-parameterized densities, we consider the finite-dimensional constrained optimization problem (10). Given the metric tensor (19), the Fisher-Rao parameter flow is given by

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = -\mathcal{I}^{-1}(\boldsymbol{\theta}_t) \nabla_{\boldsymbol{\theta}_t} D_{KL}(q(\mathbf{x}; \boldsymbol{\theta}_t) \| p(\mathbf{x}|\mathbf{z})). \quad (20)$$

Notice that the functional space Fisher-Rao gradient flow (18) and the parameter space Fisher-Rao parameter flow (20) can be interpreted as the Riemannian gradient flow induced by the Fisher-Rao Riemannian metric (14).

3 Transient Density as a Solution to Fisher-Rao Gradient Flow

We aim to identify a VI formulation that yields the transient density (4) as its solution, thus establishing a connection between the transient density (4) in the particle flow formulation and the Fisher-Rao gradient flow (18) in the VI formulation. In order to do this, we must address the following challenges.

Time parameterizations: As shown in Sec. 2.1.2, the particle flow particle filter is derived using a particular parameterization of Bayes' rule, which introduces a pseudo-time parameter λ. However, the Fisher-Rao gradient flow (18) derived in Sec. 2.2.2 does not possess such a pseudo-time parameter. In fact, we have that the transient density trajectory satisfies $p(\mathbf{x}|\mathbf{z};\lambda) \to p(\mathbf{x}|\mathbf{z})$ as $\lambda \to 1$, while the variational density trajectory defined by the Fisher-Rao gradient flow satisfies $q(\mathbf{x};t) \to p(\mathbf{x}|\mathbf{z})$ as $t \to \infty$.

Initialization: Another key difference is that the transient density trajectory $p(\mathbf{x}|\mathbf{z};\lambda)$ defines a transformation from the prior to the Bayes' posterior, while the variational density trajectory defines a transformation from any density to the Bayes' posterior.

As our result below shows, we can obtain the transient density as a solution to the Fisher-Rao gradient flow by initializing the variational density with the prior and introducing a time scaling function. The time scaling ensures that the time rate of change of the variational density matches the pseudo-time rate of change of the transient density, which is then used to derive the particle dynamics function governing the EDH flow.

Theorem 3 (Transient Density as a Solution to Fisher-Rao Gradient Flow) *The transient density* (4) *with pseudo-time scaling function* $\lambda(t) = 1 - \exp(-t)$ *is a solution to the Fisher-Rao gradient flow* (18) *of the KL divergence with* $q(\mathbf{x};0) = p(\mathbf{x})$.

Proof We have to verify that $q^*(\mathbf{x};\lambda(t)) = p(\mathbf{z}|\mathbf{x})^{\lambda(t)}p(\mathbf{x})/p(\mathbf{z};\lambda(t))$ satisfies

$$\frac{\partial q^*(\mathbf{x};\lambda(t))}{\partial t} = -\nabla^{\mathrm{FR}}_{q^*} D_{KL}(q^*(\mathbf{x};\lambda(t))\|p(\mathbf{x}|\mathbf{z})). \quad (21)$$

The derivative of $q^*(\mathbf{x};\lambda(t))$ with respect to time t in the left-hand side above can be written as:

$$\frac{\partial q^*(\mathbf{x};\lambda(t))}{\partial t} = (1-\lambda(t))q^*(\mathbf{x};\lambda(t))\left(\log\left(p(\mathbf{z}|\mathbf{x})\right) - \frac{1}{p(\mathbf{z};\lambda(t))}\frac{\partial p(\mathbf{z};\lambda(t))}{\partial \lambda(t)}\right). \quad (22)$$

By the definition of $p(\mathbf{z};\lambda(t))$ in (4), we have:

$$\frac{\partial p(\mathbf{z};\lambda(t))}{\partial \lambda(t)} = \int p(\mathbf{z}|\mathbf{x})^{\lambda(t)}p(\mathbf{x})\log\left(p(\mathbf{z}|\mathbf{x})\right)\mathrm{d}\mathbf{x}.$$

As a result, it holds that

$$\frac{1}{p(\mathbf{z};\lambda(t))}\frac{\partial p(\mathbf{z};\lambda(t))}{\partial \lambda(t)} = \int \frac{p(\mathbf{z}|\mathbf{x})^{\lambda(t)}p(\mathbf{x})}{p(\mathbf{z};\lambda(t))}\log\left(p(\mathbf{z}|\mathbf{x})\right)\mathrm{d}\mathbf{x} = \mathbb{E}_{q^*(\mathbf{x};\lambda(t))}\left[\log(p(\mathbf{z}|\mathbf{x}))\right].$$

Substituting into (22), we obtain

$$\frac{\partial q^*(\mathbf{x};\lambda(t))}{\partial t} = (1-\lambda(t))q^*(\mathbf{x};\lambda(t))\left(\log\left(p(\mathbf{z}|\mathbf{x})\right) - \mathbb{E}_{q^*(\mathbf{x};\lambda(t))}\left[\log(p(\mathbf{z}|\mathbf{x}))\right]\right).$$

On the other hand, regarding the right-hand side of (21), using the definition of $q^*(\mathbf{x};\lambda(t))$ and (1), we obtain

$$\log\left(\frac{q^*(\mathbf{x};\lambda(t))}{p(\mathbf{x}|\mathbf{z})}\right) = (\lambda(t)-1)\log(p(\mathbf{z}|\mathbf{x})) + \log\left(\frac{p(\mathbf{z})}{p(\mathbf{z};\lambda(t))}\right).$$

Substituting this expression into (16), the negative Fisher-Rao gradient of the KL divergence can be expressed as

$$-\nabla^{\mathrm{FR}}_{q^*} D_{KL}(q^*(\mathbf{x};\lambda(t))\|p(\mathbf{x}|\mathbf{z})) = (1-\lambda(t))q^*(\mathbf{x};\lambda(t))\left(\log(p(\mathbf{z}|\mathbf{x})) - \mathbb{E}_{q^*(\mathbf{x};\lambda(t))}\left[\log(p(\mathbf{z}|\mathbf{x}))\right]\right),$$

which matches the expression for $\frac{\partial q^*(\mathbf{x};\lambda(t))}{\partial t}$. ■

Theorem 3 shows that a Fisher-Rao particle flow can be derived by finding a particle dynamics function $\boldsymbol{\phi}$ such that the following equation holds:

$$\nabla_{\mathbf{x}} \cdot (q(\mathbf{x};t)\boldsymbol{\phi}(\mathbf{x},t)) = \nabla^{FR}_q D_{KL}(q(\mathbf{x};t)\|p(\mathbf{x}|\mathbf{z})),$$

with the initial variational density set to the prior $q(\mathbf{x};0) = p(\mathbf{x})$. Obtaining a closed-form expression for this particle dynamics function is challenging in general. To alleviate this, we can restrict the variational density to have a specific parametric form. In the next section, we restrict the variational density to a single Gaussian and show that, under linear Gaussian assumptions, the corresponding particle dynamics function is a time-scaled version of the particle dynamics function governing the EDH flow.

4 Gaussian Fisher-Rao Flows

This section focuses on the case where the variational density is selected as Gaussian and establishes a connection between the Gaussian Fisher-Rao gradient flow and the particle dynamics in the particle flow. Since Gaussian densities can be specified by mean $\boldsymbol{\mu}$ and covariance Σ parameters, instead of working in the space of Gaussian densities, we work in the space of parameters:

$$\mathcal{A} \coloneqq \left\{\left(\boldsymbol{\mu}, \mathrm{vec}(\Sigma^{-1})\right) : \boldsymbol{\mu} \in \mathbb{R}^n, \Sigma^{-1} \in \mathbb{R}^{n\times n}, \Sigma \succ 0\right\},$$

where $\mathrm{vec}(\cdot)$ is the vectorization operator that converts a matrix to a vector by stacking its columns. This parametrization results in the same update for the inverse covariance matrix as the half-vectorization parametrization, which accounts for the symmetry of the inverse covariance matrix (Barfoot, 2020). For $\boldsymbol{\alpha} \in \mathcal{A}$, the inverse FIM (19) evaluates to:

$$\mathcal{I}^{-1}(\boldsymbol{\alpha}) = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & 2\left(\Sigma^{-1} \otimes \Sigma^{-1}\right) \end{bmatrix}, \quad (23)$$

where $\otimes$ denotes the Kronecker product. To simplify the notation, define:

$$V(\mathbf{x};\boldsymbol{\alpha}) = \log(q(\mathbf{x};\boldsymbol{\alpha})) - \log(p(\mathbf{x},\mathbf{z})), \quad (24)$$

where $q(\mathbf{x};\boldsymbol{\alpha})$ is the variational density and $p(\mathbf{x},\mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x})$ is the joint density. The derivative of the KL divergence with respect to the Gaussian parameters is given by (Opper and Archambeau, 2009):

$$\begin{aligned}\frac{\partial D_{KL}(q(\mathbf{x};\boldsymbol{\alpha})\|p(\mathbf{x}|\mathbf{z}))}{\partial \boldsymbol{\mu}^\top} &= \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})}\left[\frac{\partial V(\mathbf{x};\boldsymbol{\alpha})}{\partial \mathbf{x}^\top}\right], \\ \frac{\partial D_{KL}(q(\mathbf{x};\boldsymbol{\alpha})\|p(\mathbf{x}|\mathbf{z}))}{\partial \Sigma^{-1}} &= -\frac{1}{2}\Sigma\mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha})}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha})}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right]\Sigma, \end{aligned} \quad (25)$$

with $V(\mathbf{x};\boldsymbol{\alpha})$ defined in (24). Inserting (23) and (25) into (20), and using the fact $\mathrm{vec}(ABC) = (C^\top \otimes A)\,\mathrm{vec}(B)$, the Gaussian Fisher-Rao parameter flow takes the form:

$$\frac{\mathrm{d}\boldsymbol{\mu}_t}{\mathrm{d}t} = -\Sigma_t \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top}\right], \quad \frac{\mathrm{d}\Sigma_t^{-1}}{\mathrm{d}t} = \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right], \quad (26)$$

where $q(\mathbf{x};\boldsymbol{\alpha}_t) = p_{\mathcal{N}}(\mathbf{x};\mu_t,\Sigma_t)$. As a result, the Gaussian Fisher-Rao parameter flow (26) defines a time rate of change of the variational density $q(\mathbf{x};\boldsymbol{\alpha}_t)$, which can be captured using the Liouville equation. This observation is formally stated in the following result.

Lemma 4 (Gaussian Fisher-Rao Particle Flow) *The time rate of change of the variational density $q(\mathbf{x};\boldsymbol{\alpha}_t)$ induced by the Gaussian Fisher-Rao parameter flow* (26) *is captured by the Liouville equation:*

$$\frac{\mathrm{d}q(\mathbf{x};\boldsymbol{\alpha}_t)}{\mathrm{d}t} = -\nabla_{\mathbf{x}} \cdot \big(q(\mathbf{x};\boldsymbol{\alpha}_t)\tilde{\phi}(\mathbf{x},t)\big), \quad (27)$$

with particle dynamics function $\tilde{\phi}(\mathbf{x},t) = \tilde{A}_t\mathbf{x} + \tilde{\mathbf{b}}_t$, where

$$\tilde{A}_t = -\frac{1}{2}\Sigma_t \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right], \qquad \tilde{\mathbf{b}}_t = -\Sigma_t \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top}\right] - \tilde{A}_t\boldsymbol{\mu}_t. \quad (28)$$

Proof Using the chain rule and (26), we can write:

$$\begin{aligned}&\frac{\mathrm{d}q(\mathbf{x};\boldsymbol{\alpha}_t)}{\mathrm{d}t} = \frac{\partial q(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\mu}_t}\frac{\mathrm{d}\boldsymbol{\mu}_t}{\mathrm{d}t} + \mathrm{tr}\left(\frac{\partial q(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \Sigma_t^{-1}}\frac{\mathrm{d}\Sigma_t^{-1}}{\mathrm{d}t}\right) \\ &= q(\mathbf{x};\boldsymbol{\alpha}_t)(\boldsymbol{\mu}_t - \mathbf{x})^\top \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top}\right] + \frac{1}{2}q(\mathbf{x};\boldsymbol{\alpha}_t)\,\mathrm{tr}\left(\Sigma_t \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right]\right) \\ &\quad - \frac{1}{2}q(\mathbf{x};\boldsymbol{\alpha}_t)(\mathbf{x} - \boldsymbol{\mu}_t)^\top \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right](\mathbf{x} - \boldsymbol{\mu}_t), \end{aligned} \quad (29)$$

where we have employed Jacobi's formula (Petersen et al., 2008) to write the derivatives of the Gaussian density. Substituting the particle dynamics function defined by (28) into

(27), we obtain

$$
\begin{aligned}
&-\nabla_{\mathbf{x}} \cdot \left(q(\mathbf{x};\boldsymbol{\alpha}_t)(\tilde{A}_t\mathbf{x}+\tilde{\mathbf{b}}_t)\right) = -q(\mathbf{x};\boldsymbol{\alpha}_t)\,\mathrm{tr}(\tilde{A}_t) + \frac{\partial q(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\mu}_t}(\tilde{A}_t\mathbf{x}+\tilde{\mathbf{b}}_t) \\
&= \frac{1}{2}q(\mathbf{x};\boldsymbol{\alpha}_t)\,\mathrm{tr}\left(\Sigma_t \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right]\right) + q(\mathbf{x};\boldsymbol{\alpha}_t)(\mathbf{x}-\boldsymbol{\mu}_t)^\top \Sigma_t^{-1}\tilde{A}_t(\mathbf{x}-\boldsymbol{\mu}_t) \\
&\quad - q(\mathbf{x};\boldsymbol{\alpha}_t)(\mathbf{x}-\boldsymbol{\mu}_t)^\top \mathbb{E}_{q(\mathbf{x};\boldsymbol{\alpha}_t)}\left[\frac{\partial V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top}\right].
\end{aligned}
$$

Substituting the value of $\tilde{A}_t$ in this expression, we see that it matches (29). ■

According to Theorem 3, the particle dynamics function described in (28), which governs the Gaussian Fisher-Rao particle flow, must correspond to a time-scaled version of the particle dynamics function in (7) that governs the EDH flow. This equivalence holds under the linear Gaussian assumptions. This observation motivates our forthcoming discussion.

Under Assumption 1 (linear Gaussian assumption), we have:

$$
\frac{\partial V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top} = \Sigma_t^{-1}(\boldsymbol{\mu}_t - \mathbf{x}) + \Sigma_p^{-1}(\mathbf{x}-\boldsymbol{\mu}_p), \quad \frac{\partial^2 V(\mathbf{x};\boldsymbol{\alpha}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}} = -\Sigma_t^{-1} + \Sigma_p^{-1},
$$

where $\boldsymbol{\mu}_p = \hat{\mathbf{x}} + PH^\top(R+HPH^\top)^{-1}(\mathbf{z}-H\hat{\mathbf{x}})$ and $\Sigma_p^{-1} = P^{-1} + H^\top R^{-1} H$ denote the posterior mean and the inverse of the posterior covariance, respectively. As a result, the Gaussian Fisher-Rao parameter flow (26) becomes:

$$
\frac{\mathrm{d}\boldsymbol{\mu}_t}{\mathrm{d}t} = -\Sigma_t\Sigma_p^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_p), \qquad \frac{\mathrm{d}\Sigma_t^{-1}}{\mathrm{d}t} = \Sigma_p^{-1} - \Sigma_t^{-1}. \tag{30}
$$

Also, the particle dynamics function from Lemma 4, which describes the Gaussian Fisher-Rao particle flow, is determined by,

$$
\tilde{A}_t = -\frac{1}{2}\Sigma_t\left(\Sigma_p^{-1} - \Sigma_t^{-1}\right), \qquad \tilde{\mathbf{b}}_t = \Sigma_t\Sigma_p^{-1}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right) - \tilde{A}_t\boldsymbol{\mu}_t. \tag{31}
$$

Based on Theorem 3, the transient parameters (6) describing the transient density should be a time-scaled solution to the Gaussian Fisher-Rao parameter flow (30) under linear Gaussian assumptions, indicating a connection between the particle dynamics functions. This intuition is formalized in the following result.

Theorem 5 (EDH Flow as Fisher-Rao Particle Flow) *Under Assumption 1, the particle dynamics function $\tilde{\phi}(\mathbf{x},t) = \tilde{A}_t\mathbf{x} + \tilde{\mathbf{b}}_t$ determined by* (31)*, which governs the Gaussian Fisher-Rao particle flow, is a time-scaled version of the EDH flow particle dynamics $\phi(\mathbf{x},\lambda) = A_\lambda\mathbf{x}+\mathbf{b}_\lambda$ determined by* (7)*, with time scaling function given by $\lambda(t) = 1-\exp(-t)$.*

Proof Under Assumption 1, a solution to the Gaussian Fisher-Rao parameter flow (30) is given as follows:

$$
\boldsymbol{\mu}_t = \Sigma_t(P^{-1}\hat{\mathbf{x}} + \lambda(t)H^\top R^{-1}\mathbf{z}), \qquad \Sigma_t^{-1} = P^{-1} + \lambda(t)H^\top R^{-1}H. \tag{32}
$$

This can be verified by observing that

$$\begin{aligned}\frac{\mathrm{d}\Sigma_t^{-1}}{\mathrm{d}t} &= (1-\lambda(t))H^\top R^{-1}H = \Sigma_p^{-1} - \Sigma_t^{-1}\\ \frac{\mathrm{d}\boldsymbol{\mu}_t}{\mathrm{d}t} &= \frac{\mathrm{d}\Sigma_t}{\mathrm{d}t}\Sigma_t^{-1}\boldsymbol{\mu}_t + (1-\lambda(t))\Sigma_t H^\top R^{-1}\mathbf{z} = -\Sigma_t\Sigma_p^{-1}\boldsymbol{\mu}_t + \boldsymbol{\mu}_t + (1-\lambda(t))\Sigma_t H^\top R^{-1}\mathbf{z}\\ &= -\Sigma_t\Sigma_p^{-1}\boldsymbol{\mu}_t + \Sigma_t(P^{-1}\hat{\mathbf{x}} + H^\top R^{-1}\mathbf{z}) = -\Sigma_t\Sigma_p^{-1}\boldsymbol{\mu}_t + \Sigma_t\Sigma_p^{-1}\boldsymbol{\mu}_p,\end{aligned}$$

where the equality $P^{-1}\hat{\mathbf{x}} + H^\top R^{-1}\mathbf{z} = \Sigma_p^{-1}\boldsymbol{\mu}_p$ can be obtained by applying the Woodbury matrix identity (Petersen et al., 2008). Using the closed-form expressions for $\boldsymbol{\mu}_t$ and Σ_t in equation (32), we can write $\tilde{A}_t$ and $\tilde{\mathbf{b}}_t$ in (31) as follows:

$$\begin{aligned}\tilde{A}_t &= \frac{1}{2}\Sigma_t\left(-\Sigma_p^{-1} + \Sigma_t^{-1}\right) = -\frac{1}{2}(1-\lambda(t))\Sigma_t H^\top R^{-1}H,\\ \tilde{\mathbf{b}}_t &= \Sigma_t\Sigma_p^{-1}\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t\right) - \tilde{A}_t\boldsymbol{\mu}_t = -\Sigma_t\Sigma_p^{-1}\boldsymbol{\mu}_t + \boldsymbol{\mu}_t + (1-\lambda(t))\Sigma_t H^\top R^{-1}\mathbf{z} - \tilde{A}_t\boldsymbol{\mu}_t\\ &= \tilde{A}_t\boldsymbol{\mu}_t + (1-\lambda(t))\Sigma_t H^\top R^{-1}\mathbf{z}.\end{aligned} \tag{33}$$

Next, we rewrite the term $\mathbf{b}_\lambda$ in (7) such that it shares a similar structure to the term $\tilde{\mathbf{b}}_t$ in (33). First, observe that

$$\lambda(R + \lambda HPH^\top)^{-1}HPH^\top = I - (R + \lambda HPH^\top)^{-1}R. \tag{34}$$

Using this equation, we deduce that

$$\begin{aligned}(I + 2\lambda A_\lambda)PH^\top R^{-1} &= PH^\top R^{-1} - \lambda PH^\top(R + \lambda HPH^\top)^{-1}HPH^\top R^{-1}\\ &= PH^\top R^{-1} - PH^\top(I - (R + \lambda HPH^\top)^{-1}R)R^{-1}\\ &= PH^\top(R + \lambda HPH^\top)^{-1},\end{aligned} \tag{35}$$

which in turn implies that

$$\lambda A_\lambda PH^\top R^{-1} = \frac{1}{2}\left(PH^\top(R + \lambda HPH^\top)^{-1} - PH^\top R^{-1}\right). \tag{36}$$

Now, we employ (34) and the definitions of $\boldsymbol{\mu}_\lambda$ and Σ_λ in (6), to write

$$\boldsymbol{\mu}_\lambda = \hat{\mathbf{x}} + \lambda PH^\top(R + \lambda HPH^\top)^{-1}(\mathbf{z} - H\hat{\mathbf{x}}).$$

Using this equation and the definition of A_λ, we can express

$$\begin{aligned}A_\lambda\boldsymbol{\mu}_\lambda &= (I + 2\lambda A_\lambda)A_\lambda\hat{\mathbf{x}} + \lambda A_\lambda PH^\top(R + \lambda HPH^\top)^{-1}\mathbf{z}\\ &= (I + 2\lambda A_\lambda)A_\lambda\hat{\mathbf{x}} - \frac{1}{2}\underbrace{PH^\top(R + \lambda HPH^\top)^{-1}}_{(35)}\underbrace{\lambda HPH^\top(R + \lambda HPH^\top)^{-1}}_{(34)}\mathbf{z}\\ &= (I + 2\lambda A_\lambda)\Big(A_\lambda\hat{\mathbf{x}} - \frac{1}{2}(PH^\top R^{-1} - PH^\top(R + \lambda HPH^\top)^{-1})\mathbf{z}\Big).\end{aligned}$$

Using this expression, the difference between $\mathbf{b}_\lambda$ and $A_\lambda\boldsymbol{\mu}_\lambda$ can be expressed as follows:

$$\begin{aligned}\mathbf{b}_\lambda - A_\lambda\boldsymbol{\mu}_\lambda &= (I + 2\lambda A_\lambda)(A_\lambda\hat{\mathbf{x}} + (I + \lambda A_\lambda)PH^\top R^{-1}\mathbf{z}) - A_\lambda\boldsymbol{\mu}_\lambda\\ &= (I + 2\lambda A_\lambda)\Big(\tfrac{3}{2}PH^\top R^{-1} + \underbrace{\lambda A_\lambda PH^\top R^{-1}}_{(36)} - \tfrac{1}{2}PH^\top(R + \lambda HPH^\top)^{-1}\Big)\mathbf{z}\\ &= (I + 2\lambda A_\lambda)PH^\top R^{-1}\mathbf{z} = PH^\top(R + \lambda HPH^\top)^{-1}\mathbf{z},\end{aligned}$$

where the last equality is obtained using (35). As a result, we can rewrite the $\mathbf{b}_\lambda$ term in (7) as follows:

$$\mathbf{b}_\lambda = PH^\top(R+\lambda HPH^\top)^{-1}\mathbf{z} + A_\lambda \boldsymbol{\mu}_\lambda.$$

Replacing λ with $\lambda(t)$ in the definition of $\boldsymbol{\mu}_\lambda$ and Σ_λ in (6) and utilizing the Woodbury formula (Petersen et al., 2008), we have:

$$\begin{aligned}
\Sigma_{\lambda(t)} &= P - \lambda(t)PH^\top(R+\lambda(t)HPH^\top)^{-1}HP = \left(P^{-1}+\lambda(t)H^\top R^{-1}H\right)^{-1} = \Sigma_t \\
\boldsymbol{\mu}_{\lambda(t)} &= \Sigma_{\lambda(t)}(P^{-1}\hat{\mathbf{x}} + \lambda(t)H^\top R^{-1}\mathbf{z}) = \boldsymbol{\mu}_t.
\end{aligned}$$

Using the expression (32) of Σ_t^{-1}, we can obtain the following identity using (34):

$$\begin{aligned}
\Sigma_t H^\top R^{-1} &= \left(P^{-1}+\lambda(t)H^\top R^{-1}H\right)^{-1}H^\top R^{-1} \\
&= PH^\top R^{-1} - PH^\top \underbrace{\lambda(t)(R+\lambda(t)HPH^\top)^{-1}HPH^\top}_{(34)} R^{-1} \\
&= PH^\top(R+\lambda(t)HPH^\top)^{-1}.
\end{aligned} \tag{37}$$

Finally, by applying (37), we have:

$$A_{\lambda(t)} = -\frac{1}{2}\Sigma_t H^\top R^{-1}H = \frac{1}{1-\lambda(t)}\tilde{A}_t, \quad \mathbf{b}_{\lambda(t)} = A_{\lambda(t)}\boldsymbol{\mu}_t + \Sigma_t H^\top R^{-1}\mathbf{z} = \frac{1}{1-\lambda(t)}\tilde{\mathbf{b}}_t.$$

■

The proof of Theorem 5 reveals the fact that by rewriting (31) using the closed-form expressions for $\boldsymbol{\mu}_t$ and Σ_t in (32), the particle dynamics function determined by (31) shares the same expression as (7) up to an appropriate time scaling coefficient.

Simulation results comparing the Gaussian Fisher-Rao particle flow and the EDH flow are presented in Figure 1. We use the following parameters in the simulation:

$$\hat{\mathbf{x}} = \begin{bmatrix}0\\0\end{bmatrix}, \quad P = \begin{bmatrix}1.5 & 0.5\\0.5 & 5.5\end{bmatrix}, \quad H = \begin{bmatrix}1 & 1.5\\0.2 & 2\end{bmatrix}, \quad R = \begin{bmatrix}0.2 & 0.1\\0.1 & 0.2\end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix}-1.18\\4.12\end{bmatrix},$$

where $\mathbf{x}^*$ denotes true state.

A single Gaussian approximation to the posterior is often times insufficient due to its single-modal nature, especially when the observation model is nonlinear (which potentially results in a multimodal posterior). This motivates our discussion in the next section, where the variational density follows a Gaussian mixture density.

5 Approximated Gaussian Mixture Fisher-Rao Flows

This section focuses on the case where the variational density is selected as a Gaussian mixture density. Computing the Fisher information matrix associated with a Gaussian mixture density is costly, however. For better efficiency, we adopt a diagonal approximation of the FIM proposed in Lin et al. (2019a) and establish a connection between the approximated

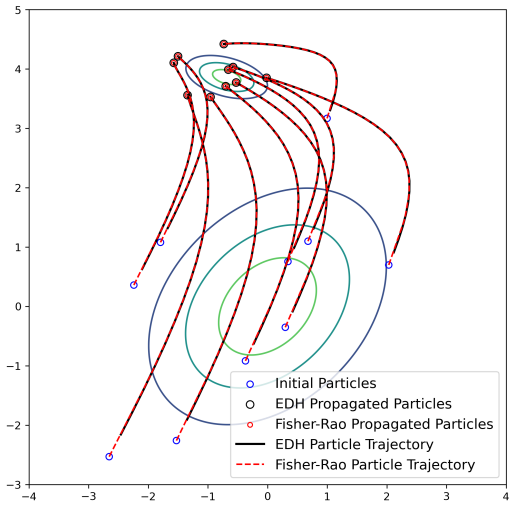

Figure 1: Comparison of the EDH flow and the Gaussian Fisher-Rao flow under linear Gaussian assumptions. We propagate 10 randomly selected particles through both flows. The trajectories of the particles are identical, verifying the results stated in Theorem 5.

Gaussian mixture Fisher-Rao gradient flow and the particle dynamics in the particle flow. A Gaussian mixture density with K components can be expressed as follows:

$$q(\mathbf{x}) = \sum_{k=1}^{K} q(\mathbf{x}|\omega = k)q(\omega = k),$$

where $q(\mathbf{x}|\omega = k) = p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}^{(k)}, \Sigma^{(k)})$ and $q(\omega = k) = \pi^{(k)}$ is a multinomial PDF such that $\sum_{k=1}^{K} \pi^{(k)} = 1$. As noted by Lin et al. (2019a), employing a natural parameterization of the Gaussian mixture density and an approximated FIM simplifies the derivation of an approximation of the Fisher-Rao parameter flow (20). Furthermore, the natural parameterization allows the component weight parameters to be expressed in real-valued log-odds, which eliminates the need for re-normalization to ensure $\sum_{k=1}^{K} \pi^{(k)} = 1$. Let $\boldsymbol{\eta}$ denote the natural parameters of the Gaussian mixture density:

$$\eta_{\omega}^{(k)} = \log\left(\frac{\pi^{(k)}}{\pi^{(K)}}\right),\ \boldsymbol{\eta}_{x}^{(k)} = \left(\boldsymbol{\gamma}^{(k)},\ \Gamma^{(k)}\right),\ \boldsymbol{\gamma}^{(k)} = [\Sigma^{(k)}]^{-1}\boldsymbol{\mu}^{(k)}, \Gamma^{(k)} = -\frac{1}{2}[\Sigma^{(k)}]^{-1}, \quad (38)$$

where $\boldsymbol{\eta}_{x}^{(k)}$ denotes the natural parameters of the kth Gaussian mixture component and $\eta_{\omega}^{(k)}$ denotes the natural parameter of the kth component weight. Notice that we have $\eta_{\omega}^{(K)} = 0$ and thus we set $\pi^{(K)} = 1 - \sum_{k=1}^{K-1} \pi^{(k)}$ in order to ensure $\sum_{k=1}^{K} \pi^{(k)} = 1$. The component weights from their natural parameterization are recovered via:

$$\pi^{(k)} = \frac{\exp(\eta^{(k)})}{\sum_{j=1}^{K} \exp(\eta^{(j)})}.$$

However, the FIM $\mathcal{I}(\boldsymbol{\eta})$ defined in (19) of the Gaussian mixture variational density $q(\mathbf{x})$ is difficult to compute. We adopt the block-diagonal approximation $\tilde{\mathcal{I}}(\boldsymbol{\eta})$ of the FIM proposed in Lin et al. (2019a):

$$\tilde{\mathcal{I}}(\boldsymbol{\eta}) = \text{diag}\left(\mathcal{I}(\boldsymbol{\eta}_x^{(1)}, \eta_\omega^{(1)}), ..., \mathcal{I}(\boldsymbol{\eta}_x^{(K)}, \eta_\omega^{(K)})\right), \quad (39)$$

where $\mathcal{I}(\boldsymbol{\eta}_x^{(k)}, \eta_\omega^{(k)})$ is the FIM of the kth joint Gaussian mixture component $q(\mathbf{x}|\omega = k)q(\omega = k)$. Using the approximated FIM, we can define the approximated Gaussian mixture Fisher-Rao parameter flow as follows:

$$\frac{\mathrm{d}\boldsymbol{\eta}_t}{\mathrm{d}t} = -\tilde{\mathcal{I}}^{-1}(\boldsymbol{\eta}_t)\nabla_{\boldsymbol{\eta}_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})). \quad (40)$$

To ease notation, define:

$$V(\mathbf{x};\boldsymbol{\eta}) = \log(q(\mathbf{x};\boldsymbol{\eta})) - \log(p(\mathbf{x},\mathbf{z})), \quad (41)$$

where $q(\mathbf{x};\boldsymbol{\eta})$ is the variational density and $p(\mathbf{x},\mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x})$ is the joint density. Due to the block-diagonal structure of the approximated FIM, the approximated Gaussian mixture Fisher-Rao parameter flow can be computed for the individual components, which is justified in the following proposition.

Proposition 6 (Approximated Gaussian Mixture Fisher-Rao Parameter Flow) *Consider the approximated Gaussian mixture Fisher-Rao parameter flow* (40)*. The component-wise approximated Gaussian mixture Fisher-Rao parameter flow for each Gaussian mixture component $q(\mathbf{x}|\omega = k)q(\omega = k)$ takes the following form:*

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{\gamma}_t^{(k)}}{\mathrm{d}t} &= -\mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)}\left[\frac{1}{2}\frac{\partial^2 V(\mathbf{x};\boldsymbol{\eta}_t)}{\partial\mathbf{x}^\top\partial\mathbf{x}}[\Gamma_t^{(k)}]^{-1}\boldsymbol{\gamma}_t^{(k)} + \frac{\partial V(\mathbf{x};\boldsymbol{\eta}_t)}{\partial\mathbf{x}^\top}\right], \\
\frac{\mathrm{d}\Gamma_t^{(k)}}{\mathrm{d}t} &= -\frac{1}{2}\mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)}\left[\frac{\partial^2 V(\mathbf{x};\boldsymbol{\eta}_t)}{\partial\mathbf{x}^\top\partial\mathbf{x}}\right], \\
\frac{\mathrm{d}[\eta_\omega^{(k)}]_t}{\mathrm{d}t} &= \mathbb{E}_{q(\mathbf{x}|\omega=K;\boldsymbol{\eta}_t)}\left[V(\mathbf{x};\boldsymbol{\eta}_t)\right] - \mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)}\left[V(\mathbf{x};\boldsymbol{\eta}_t)\right].
\end{aligned} \quad (42)$$

Proof According to the definition of the approximated FIM (39), the approximated Gaussian mixture Fisher-Rao parameter flow can be decomposed into K components with each component taking the following form:

$$\frac{\mathrm{d}([\boldsymbol{\eta}_x^{(k)}]_t, [\eta_\omega^{(k)}]_t)}{\mathrm{d}t} = -\mathcal{I}^{-1}([\boldsymbol{\eta}_x^{(k)}]_t, [\eta_\omega^{(k)}]_t)\nabla_{([\boldsymbol{\eta}_x^{(k)}]_t, [\eta_\omega^{(k)}]_t)} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})).$$

According to Lin et al. (2019a, Lemma 2), the FIM $\mathcal{I}(\boldsymbol{\eta}_x^{(k)}, \eta_\omega^{(k)})$ of the kth joint Gaussian mixture component $q(\mathbf{x}|\omega = k)q(\omega = k)$ takes the following form:

$$\mathcal{I}(\boldsymbol{\eta}_x^{(k)}, \eta_\omega^{(k)}) = \text{diag}(\pi^{(k)}\mathcal{I}(\boldsymbol{\eta}_x^{(k)}), \mathcal{I}(\eta_\omega^{(k)})).$$

As a result, we can further decompose (40) as follows:

$$\begin{aligned}\frac{\mathrm{d}[\boldsymbol{\eta}_x^{(k)}]_t}{\mathrm{d}t} &= -\frac{1}{\pi^{(k)}}\mathcal{I}^{-1}([\boldsymbol{\eta}_x^{(k)}]_t)\nabla_{[\boldsymbol{\eta}_x^{(k)}]_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})),\\ \frac{\mathrm{d}[\eta_\omega^{(k)}]_t}{\mathrm{d}t} &= -\mathcal{I}^{-1}([\eta_\omega^{(k)}]_t)\nabla_{[\eta_\omega^{(k)}]_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})),\end{aligned} \tag{43}$$

where $\mathcal{I}([\boldsymbol{\eta}_x^{(k)}]_t)$ is the FIM of the kth Gaussian mixture component $q(\mathbf{x}|\omega=k)$ and $\mathcal{I}([\eta_\omega^{(k)}]_t)$ is the FIM of the kth component weight $q(\omega=k)$. Also, according to Lin et al. (2019a, Theorem 3), the following equations hold:

$$\begin{aligned}\nabla_{[\mathbf{m}_x^{(k)}]_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})) &= \mathcal{I}^{-1}([\boldsymbol{\eta}_x^{(k)}]_t)\nabla_{[\boldsymbol{\eta}_x^{(k)}]_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})),\\ \nabla_{[m_\omega^{(k)}]_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})) &= \mathcal{I}^{-1}([\eta_\omega^{(k)}]_t)\nabla_{[\eta_\omega^{(k)}]_t} D_{KL}(q(\mathbf{x};\boldsymbol{\eta}_t)\|p(\mathbf{x}|\mathbf{z})),\end{aligned} \tag{44}$$

where $\mathbf{m}_x^{(k)} = \left(\boldsymbol{\mu}^{(k)}, \boldsymbol{\mu}^{(k)}[\boldsymbol{\mu}^{(k)}]^\top + \Sigma^{(k)}\right)$ and $m_\omega^{(k)} = \pi^{(k)}$ denotes the expectation parameters of the kth Gaussian mixture component $q(\mathbf{x}|\omega=k)$ and the kth component weight $q(\omega=k)$, respectively. Using the chain rule, we can derive the following expression:

$$\nabla_{\mathbf{m}_x^{(k)}} D_{KL} = \Big(\left(\nabla_{\boldsymbol{\mu}^{(k)}} D_{KL} - 2[\nabla_{\Sigma^{(k)}} D_{KL}]\boldsymbol{\mu}^{(k)}\right), [\nabla_{\Sigma^{(k)}} D_{KL}]\Big). \tag{45}$$

The gradient of the KL divergence with respect to $\boldsymbol{\mu}^{(k)}$ and $\Sigma^{(k)}$ can be expressed in terms of the gradient and Hessian of $V(\mathbf{x})$ defined in (41) by using Bonnet's and Price's theorem, cf. Bonnet (1964); Price (1958); Lin et al. (2019b):

$$\nabla_{\boldsymbol{\mu}^{(k)}} D_{KL} = \mathbb{E}_{q(\mathbf{x}|\omega=k)}\left[\pi^{(k)}\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top}\right], \quad \nabla_{\Sigma^{(k)}} D_{KL} = \frac{1}{2}\mathbb{E}_{q(\mathbf{x}|\omega=k)}\left[\pi^{(k)}\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right]. \tag{46}$$

Substituting (46) into (45), we have:

$$\nabla_{\mathbf{m}_x^{(k)}} D_{KL} = \Big(\pi^{(k)}\mathbb{E}_{q(\mathbf{x}|\omega=k)}\left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top} - \frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}}\boldsymbol{\mu}^{(k)}\right], \frac{\pi^{(k)}}{2}\mathbb{E}_{q(\mathbf{x}|\omega=k)}\left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right]\Big). \tag{47}$$

The gradient of the variational density with respect to the expectation parameter of the kth component weight takes the following form:

$$\nabla_{m_\omega^{(k)}} q(\mathbf{x};\eta) = q(\mathbf{x}|\omega=k) - q(\mathbf{x}|\omega=K),$$

where the second term appears due to the fact $\pi^{(K)} = 1 - \sum_{k=1}^{K-1}\pi^{(k)}$. Utilizing the fact above, the gradient of the KL divergence with respect to the expectation parameter of the kth component weight takes the following form:

$$\begin{aligned}\nabla_{m_\omega^{(k)}} D_{KL} &= \int V(\mathbf{x})\nabla_{m_\omega^{(k)}} q(\mathbf{x};\eta)\,\mathrm{d}\mathbf{x} + \mathbb{E}_{q(\mathbf{x};\eta)}\left[\nabla_{m_\omega^{(k)}}\log(q(\mathbf{x};\eta))\right]\\ &= \mathbb{E}_{q(\mathbf{x}|\omega=k)}\left[V(\mathbf{x})\right] - \mathbb{E}_{q(\mathbf{x}|\omega=K)}\left[V(\mathbf{x})\right] + \int q(\mathbf{x}|\omega=k) - q(\mathbf{x}|\omega=K)\,\mathrm{d}\mathbf{x}\\ &= \mathbb{E}_{q(\mathbf{x}|\omega=k)}\left[V(\mathbf{x})\right] - \mathbb{E}_{q(\mathbf{x}|\omega=K)}\left[V(\mathbf{x})\right].\end{aligned} \tag{48}$$

The desired results can be obtained by combining (47), (48), (44) and (43). ■

The approximated Gaussian mixture Fisher-Rao parameter flow (42) defines a time rate of change of the kth Gaussian mixture component. The corresponding particle flow for the kth Gaussian mixture component can be obtained by finding a particle dynamics function $\tilde{\phi}_k(\mathbf{x}, t)$ such that the following equation holds:

$$\frac{\partial q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t)}{\partial \boldsymbol{\gamma}_t^{(k)}} \frac{\mathrm{d}\boldsymbol{\gamma}_t^{(k)}}{\mathrm{d}t} + \mathrm{tr}\left(\frac{\partial q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t)}{\partial \Gamma_t^{(k)}} \frac{\mathrm{d}\Gamma_t^{(k)}}{\mathrm{d}t} \right) = -\nabla_{\mathbf{x}} \cdot \big(q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t) \tilde{\phi}_k(\mathbf{x}, t) \big).$$

The closed-from expression for this particle dynamics function $\tilde{\phi}_k(\mathbf{x}, t)$ is given in the following result.

Proposition 7 (Approximated Gaussian Mixture Fisher-Rao Particle Flow) *The time rate of change of the kth Gaussian mixture component $q(\mathbf{x}|\omega = k)$ induced by the approximated Gaussian mixture Fisher-Rao parameter flow* (42) *is captured by the Liouville equation:*

$$\frac{\mathrm{d}q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t)}{\mathrm{d}t} = -\nabla_{\mathbf{x}} \cdot \big(q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t) \tilde{\phi}_k(\mathbf{x}, t) \big),$$

with particle dynamics function $\tilde{\phi}_k(\mathbf{x}, t) = \tilde{A}_t^{(k)} \mathbf{x} + \tilde{\mathbf{b}}_t^{(k)}$, where

$$\tilde{A}_t^{(k)} = \frac{1}{4} [\Gamma_t^{(k)}]^{-1} \mathbb{E}_{q(\mathbf{x}|\omega = k; t)} \left[\frac{\partial^2 V(\mathbf{x}; \boldsymbol{\eta}_t)}{\partial \mathbf{x}^\top \partial \mathbf{x}} \right],$$
$$\tilde{\mathbf{b}}_t^{(k)} = \frac{1}{2} [\Gamma_t^{(k)}]^{-1} \mathbb{E}_{q(\mathbf{x}|\omega = k; t)} \left[\frac{\partial V(\mathbf{x}; \boldsymbol{\eta}_t)}{\partial \mathbf{x}^\top} \right] + \frac{1}{2} \tilde{A}_t^{(k)} [\Gamma_t^{(k)}]^{-1} \boldsymbol{\gamma}_t^{(k)},$$

with $V(\mathbf{x}; \boldsymbol{\eta}_t)$ given in (41)*, $\Gamma^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ are natural parameters of the kth Gaussian mixture components given in* (38)*.*

Proof The gradient of a Gaussian PDF $p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ with respect to its natural parameters (38) is given by:

$$\frac{\partial p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)}{\partial \boldsymbol{\gamma}} = p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \big(\mathbf{x}^\top + \frac{1}{2} \boldsymbol{\gamma}^\top \Gamma^{-1} \big),$$
$$\frac{\partial p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)}{\partial \Gamma} = p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \big(\mathbf{x}\mathbf{x}^\top - \frac{1}{4} \Gamma^{-1} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top \Gamma^{-1} + \frac{1}{2} \Gamma^{-1} \big).$$

The time rate of change of the kth Gaussian mixture component $q(\mathbf{x}|\omega = k)$ induced by the approximated Gaussian mixture Fisher-Rao parameter flow (42) is given by:

$$\frac{\partial q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t)}{\partial \boldsymbol{\gamma}_t^{(k)}} \frac{\mathrm{d}\boldsymbol{\gamma}_t^{(k)}}{\mathrm{d}t} + \mathrm{tr}\left(\frac{\partial q(\mathbf{x}|\omega = k; \boldsymbol{\eta}_t)}{\partial \Gamma_t^{(k)}} \frac{\mathrm{d}\Gamma_t^{(k)}}{\mathrm{d}t} \right),$$

where $\frac{\mathrm{d}\boldsymbol{\gamma}_t^{(k)}}{\mathrm{d}t}$ and $\frac{\mathrm{d}\Gamma_t^{(k)}}{\mathrm{d}t}$ are given in (42). With these expressions, the proof now proceeds analogously to the proof of Lemma 4. ■

Remark 8 *Using the definition* (38) *of the natural parameters and the chain rule, one can express the approximated Gaussian mixture Fisher-Rao parameter flow in conventional parameters as follows:*

$$\frac{\mathrm{d}\boldsymbol{\mu}_t^{(k)}}{\mathrm{d}t} = -\Sigma_t^{(k)} \mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)} \left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top} \right], \quad \frac{\mathrm{d}(\Sigma_t^{(k)})^{-1}}{\mathrm{d}t} = \mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)} \left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}} \right], \tag{49}$$

as well as the particle dynamics function $\tilde{\phi}_k(\mathbf{x}, t)$ *governing the approximated Gaussian mixture Fisher-Rao particle flow*

$$\begin{aligned} \tilde{A}_t^{(k)} &= -\frac{1}{2} \Sigma_t^{(k)} \mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)} \left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}} \right], \\ \tilde{\mathbf{b}}_t^{(k)} &= \Sigma_t^{(k)} \mathbb{E}_{q(\mathbf{x}|\omega=k;\boldsymbol{\eta}_t)} \left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top} \right] - \tilde{A}_t^{(k)} \boldsymbol{\mu}_t^{(k)}. \end{aligned} \tag{50}$$

The approximated Gaussian mixture Fisher-Rao particle flow derived in this section can capture multimodal behavior in the posterior density. We find that the approximated Gaussian mixture Fisher-Rao flows in (49) and (50) share a similar form as the Gaussian Fisher-Rao flows, cf. (26) and (28), which indicates that the approximated Gaussian mixture Fisher-Rao particle flow can be interpreted as a weighted mixture of Gaussian Fisher-Rao particle flows. Note that all these flows require the evaluation of the expectation of the gradient and Hessian of the function $V(\mathbf{x}; \boldsymbol{\eta})$ defined in (41), which in the case of Gaussian mixtures is not readily available. In the following section, we show that one can utilize the particle flow method to obtain these expectations in an efficient way.

6 Derivative- and Inverse-Free Formulation of the Fisher-Rao Flows

In this section, we derive several identities associated with the particle flow method and show how they can make the computation of the flow parameters more efficient. To start, we notice the component-wise approximated Gaussian mixture Fisher-Rao parameter flow (49) takes the same form as the Gaussian Fisher-Rao parameter flow (26). To make the discussion clear, we consider the parameter flow of each Gaussian mixture component:

$$\frac{\mathrm{d}\bar{\boldsymbol{\mu}}_t}{\mathrm{d}t} = -\bar{\Sigma}_t \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)} \left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top} \right], \quad \frac{\mathrm{d}\bar{\Sigma}_t^{-1}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)} \left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}} \right], \tag{51}$$

where $V(\mathbf{x})$ is defined in (24). Similarly, we consider the particle dynamics function governing the particle flow:

$$\bar{\phi}(\mathbf{x}, t) = \bar{A}_t \mathbf{x} + \bar{\mathbf{b}}_t, \tag{52}$$

where $\bar{A}_t = -\frac{1}{2}\bar{\Sigma}_t \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)} \left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}} \right]$, $\bar{\mathbf{b}}_t = -\bar{\Sigma}_t \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)} \left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top} \right] - \bar{A}_t \bar{\boldsymbol{\mu}}_t$. We first find a derivative-free expression for the expectation terms in (51) and (52), and then compare several particle-based approximation methods to compute the resulting derivative-free expectation terms.

To avoid computing derivatives of $V(\mathbf{x})$, we apply Stein's lemma (Stein, 1981), yielding the following expressions:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[\frac{\partial V(\mathbf{x})}{\partial \mathbf{x}^\top}\right] &= \bar{\Sigma}_t^{-1}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)V(\mathbf{x})\right], \\
\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}}\right] &= \bar{\Sigma}_t^{-1}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)^\top V(\mathbf{x})\right]\bar{\Sigma}_t^{-1} \\
&\quad - \bar{\Sigma}_t^{-1}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[V(\mathbf{x})\right].
\end{aligned} \tag{53}$$

Using the identities above, we obtain a derivative-free parameter flow:

$$\begin{aligned}
\frac{\mathrm{d}\bar{\boldsymbol{\mu}}_t}{\mathrm{d}t} &= -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)V(\mathbf{x})\right], \\
\frac{\mathrm{d}\bar{\Sigma}_t^{-1}}{\mathrm{d}t} &= \bar{\Sigma}_t^{-1}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)^\top V(\mathbf{x})\right]\bar{\Sigma}_t^{-1} - \bar{\Sigma}_t^{-1}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[V(\mathbf{x})\right].
\end{aligned} \tag{54}$$

Similarly, we can obtain derivative-free expressions for $\bar{A}_t$ and $\bar{\mathbf{b}}_t$ describing the particle flow:

$$\begin{aligned}
\bar{A}_t &= -\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)^\top V(\mathbf{x})\right]\bar{\Sigma}_t^{-1} + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[V(\mathbf{x})\right] \\
\bar{\mathbf{b}}_t &= -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[(\mathbf{x}-\bar{\boldsymbol{\mu}}_t)V(\mathbf{x})\right] - \bar{A}_t\bar{\boldsymbol{\mu}}_t.
\end{aligned} \tag{55}$$

Note that the derivative-free flow expressions in (54) and (55) only depend on $\bar{\Sigma}_t^{-1}$. By propagating $\bar{\Sigma}_t^{-1}$ instead of $\bar{\Sigma}_t$, we can also avoid inefficient matrix inverse calculations in the flow expressions.

The computation of the expectation terms in (54) and (55) analytically is generally not possible and often times numerical approximation is sought. Since the expectation is with respect to a Gaussian density, there are various accurate and efficient approximation techniques, such as linearization of the terms (Anderson and Moore, 2005) and Monte-Carlo integration using Unscented or Gauss-Hermite particles (Julier et al., 1995; Särkkä and Svensson, 2023). We use a particle-based approximation for the expectation terms:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\bar{\boldsymbol{\mu}}_t,\bar{\Sigma}_t)}\left[f(\mathbf{x})\right] \approx \sum_{i=1}^{N} w_i f(\mathbf{x}_t^{(i)}),$$

where w_i are weights, $\mathbf{x}_t^{(i)} \in \mathbb{R}^n$ are particles and $f(\mathbf{x})$ is an arbitrary function. The particles and associated weights can be generated using the Gauss-Hermite cubature method (Särkkä and Svensson, 2023). We first generate p Gauss-Hermite particles $\{\xi_i\}_{i=1}^p$ of dimension one corresponding to $\mathcal{N}(0,1)$ by computing the roots of the Gauss-Hermite polynomial of order p:

$$h_p(\xi) = (-1)^p \exp(\xi^2/2)\frac{\mathrm{d}^p \exp(-\xi^2/2)}{\mathrm{d}\xi^p},$$

with the corresponding weights $\{{}^{(1)}w_i\}_{i=1}^p$ obtained as follows:

$${}^{(1)}w_i = \frac{p!}{(p h_{p-1}(\xi_i))^2},$$

where we use the left superscript in ${}^{(1)}w_i$ to indicate that these weights correspond to the one-dimensional Gauss-Hermite quadrature rule. The Gauss-Hermite particles $\{\boldsymbol{\xi}_i\}_{i=1}^{p^n}$ of dimension n correspond to $\mathcal{N}(\mathbf{0}, I)$ and are generated as the Cartesian product of the one-dimensional Gauss-Hermite particles:

$$\{\boldsymbol{\xi}_i\}_{i=1}^{p^n} = \{(\xi_{i_1}, \xi_{i_2}, ..., \xi_{i_n}), \quad i_1, i_2, ..., i_n \in \{1, 2, ..., p\}\},$$

with the corresponding weights $\{w_i\}_{i=1}^{p^n}$ obtained as follows:

$$\{w_i\}_{i=1}^{p^n} = \{{}^{(1)}w_{i_1}^{(1)} w_{i_2} ... {}^{(1)}w_{i_n}, \quad i_1, i_2, ..., i_n \in \{1, 2, ..., p\}\}.$$

Finally, Gauss-Hermite particles $\{\mathbf{x}_i\}_{i=1}^{p^n}$ of dimension n corresponding to $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$ are obtained as follows:

$$\mathbf{x}_t^{(i)} = \bar{L}_t \boldsymbol{\xi}_i + \bar{\boldsymbol{\mu}}_t, (56)$$

where $\bar{L}_t$ is a matrix square root that satisfies $\bar{\Sigma}_t = \bar{L}_t \bar{L}_t^\top$ and the superscript i in $\mathbf{x}_t^{(i)}$ denotes the ith particle. The weights $\{w_i\}_{i=1}^{p^n}$ remain unchanged.

Although we only need to generate Gauss-Hermite particles $\{\boldsymbol{\xi}_i\}_{i=1}^{p^n}$ and corresponding weights $\{w_i\}_{i=1}^{p^n}$ for a standard normal distribution once, we need to scale and translate the particles using $\bar{L}_t$ and $\bar{\boldsymbol{\mu}}_t$ over time. However, we can avoid the scaling and translation to obtain the Gauss-Hermite particle corresponding to $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$ by propagating the Gauss-Hermite particles using the particle flow described in (2) with pseudo dynamics given by (52). In order to establish this, we need to introduce the following auxiliary result.

Theorem 9 (Mahalanobis Distance is Invariant) *Define the Mahalanobis distance as:*

$$D_{\mathcal{M}}(\mathbf{x}, \boldsymbol{\mu}, \Sigma) \coloneqq (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

Consider a Gaussian distribution $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$ with its parameters evolving according to (51)*. Let $\mathbf{x}_0$ be a particle that evolves according to* (2) *with the particle dynamics function given by* (52)*. Then,*

$$D_{\mathcal{M}}(\mathbf{x}_t, \bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t) = D_{\mathcal{M}}(\mathbf{x}_0, \bar{\boldsymbol{\mu}}_{t=0}, \bar{\Sigma}_{t=0}), \quad \forall t > 0.$$

Proof According to the chain rule, we have:

$$\begin{aligned}
&\frac{\mathrm{d}D_{\mathcal{M}}(\mathbf{x}_t, \bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)}{\mathrm{d}t} = 2(\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} - \frac{\mathrm{d}\bar{\boldsymbol{\mu}}_t}{\mathrm{d}t})^\top \bar{\Sigma}_t^{-1} (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t) + (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t)^\top \frac{\mathrm{d}\bar{\Sigma}_t^{-1}}{\mathrm{d}t} (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t) \\
&= 2(\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t)^\top \bar{A}_t^\top \bar{\Sigma}_t^{-1} (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t) + (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t)^\top \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)} \left[\frac{\partial^2 V(\mathbf{x})}{\partial \mathbf{x}^\top \partial \mathbf{x}} \right] (\mathbf{x}_t - \bar{\boldsymbol{\mu}}_t) = 0,
\end{aligned}$$

indicating that the time rate of change of the Mahalanobis distance is zero. ■

We are ready to prove that Gauss-Hermite particles remain Gauss-Hermite when propagated along the Fisher-Rao particle flow.

Theorem 10 (Fisher-Rao Particle Flow as Gauss-Hermite Transform) *Consider a Gaussian distribution $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$ with parameters evolving according to* (51)*. Let $\mathbf{x}_0$ be a Gauss-Hermite particle obtained from $\mathcal{N}(\bar{\boldsymbol{\mu}}_{t=0}, \bar{\Sigma}_{t=0})$ which evolves according to* (2) *with particle dynamics function given by* (52)*. Then, $\mathbf{x}_t$ is a Gauss-Hermite particle corresponding to $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$.*

Proof Let $\boldsymbol{\xi}$ be arbitrary, consider the particle $\boldsymbol{\xi}_t = \bar{L}_t\boldsymbol{\xi} + \bar{\boldsymbol{\mu}}_t$, where $\bar{\boldsymbol{\mu}}_t$ evolves according to (51) and $\bar{L}_t$ evolves according to the following equation:

$$\frac{\mathrm{d}\bar{L}_t}{\mathrm{d}t} = \bar{A}_t\bar{L}_t, (57)$$

with $\bar{A}_t$ given in (52). We first show $\boldsymbol{\xi}_t$ is the solution to the particle flow (2), where the particle dynamics function is given by (52). Next, we show that if we have $\bar{L}_{t=0}$ satisfy $\bar{\Sigma}_{t=0} = \bar{L}_{t=0}\bar{L}_{t=0}^\top$ and the evolution of $\bar{L}_t$ satisfies (57), then $\bar{\Sigma}_t = \bar{L}_t\bar{L}_t^\top$.

According to the chain rule, the evolution of the particle $\boldsymbol{\xi}_t$ satisfies:

$$\frac{\mathrm{d}\boldsymbol{\xi}_t}{\mathrm{d}t} = \frac{\mathrm{d}\bar{L}_t}{\mathrm{d}t}\boldsymbol{\xi} + \frac{\mathrm{d}\bar{\boldsymbol{\mu}}_t}{\mathrm{d}t} = \bar{A}_t\bar{L}_t\boldsymbol{\xi} + \bar{A}_t\bar{\boldsymbol{\mu}}_t + \bar{\mathbf{b}}_t = \bar{A}_t\boldsymbol{\xi}_t + \bar{\mathbf{b}}_t,$$

which is exactly the same as the particle flow (2) with the particle dynamics given in (52). This justifies our first claim. According to Theorem 9, we have:

$$D_{\mathcal{M}}(\boldsymbol{\xi}_t, \bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t) = \boldsymbol{\xi}^\top\bar{L}_t^\top\Sigma_t^{-1}\bar{L}_t\boldsymbol{\xi} = \boldsymbol{\xi}^\top\boldsymbol{\xi}.$$

Since $\bar{L}_t^\top\Sigma_t^{-1}\bar{L}_t$ is symmetric, the following equation holds:

$$\bar{L}_t^\top\Sigma_t^{-1}\bar{L}_t = I,$$

indicating that $\bar{\Sigma}_t^{-1} = [\bar{L}_t^{-1}]^\top\bar{L}_t^{-1}$ and $\bar{\Sigma}_t = \bar{L}_t\bar{L}_t^\top$. This justifies our second claim. Hence, a Gauss-Hermite particle corresponding to $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$ can be obtained by solving the particle flow (2) with the particle dynamics function given by (52). ■

This result shows that, instead of obtaining Gauss-Hermite particles corresponding to $\mathcal{N}(\bar{\boldsymbol{\mu}}_t, \bar{\Sigma}_t)$, we can propagate Gauss-Hermite particles according to the particle flow in (2) with the particle dynamics given in (52), starting from the Gauss-Hermite particles corresponding to $\mathcal{N}(\bar{\boldsymbol{\mu}}_{t=0}, \bar{\Sigma}_{t=0})$.

Remark 11 (Propagating Covariance in Square Root Form) *According to the proof of Theorem 10, we have that if $\bar{L}_{t=0}$ satisfies $\bar{\Sigma}_{t=0} = \bar{L}_{t=0}\bar{L}_{t=0}^\top$ and the evolution of $\bar{L}_t$ satisfies* (57)*, then $\bar{\Sigma}_t = \bar{L}_t\bar{L}_t^\top$. In this regard, we can propagate the inverse covariance matrix in square-root form as:*

$$\frac{\mathrm{d}\bar{L}_t^{-1}}{\mathrm{d}t} = -\bar{L}_t^{-1}\bar{A}_t,$$

where $\bar{L}_{t=0}$ satisfies $\bar{\Sigma}_{t=0} = \bar{L}_{t=0}\bar{L}_{t=0}^\top$. This flow enforces that the inverse covariance matrix stays symmetric and positive definite during propagation. Note that the matrix square root decomposition is not unique and, hence, depending on the initialization of this flow, the evolution of the matrix square root will differ. Our proposed method is a special case of the one in Morf et al. (1977), while other solutions exist, such as the one proposed in Särkkä (2007), which considers the lower triangular structure of the covariance square root matrix.

In this section, we demonstrated several important identities satisfied by the particle flow and its underlying variational density. Corollary 10 demonstrates that Gauss-Hermite

particles, generated by a specific covariance square root, evolve according to the particle flow with the particle dynamics function defined by (52). Consequently, evaluating the Gaussian probability density function at these particle locations amounts to rescaling the density evaluated at the particles' initial positions. However, this result should be used with caution when dealing with Gaussian mixture variational distribution since it only applies to particles associated with their respective Gaussian mixture components. In the next section, we present numerical results to demonstrate the accuracy and efficiency of the Fisher-Rao particle flow method. We all also evaluate the expectation terms in (51) and (52) using the methods discussed in this section and compare their accuracy.

7 Evaluation

In this section, we first evaluate our Fisher-Rao flows under two low-dimension scenarios. In the first scenario, the prior density is a Gaussian mixture, and the likelihood function is Gaussian with a linear observation model, leading to a multimodal posterior density. In the second scenario, the prior density is a single Gaussian, and the likelihood function is a Gaussian density with a nonlinear observation model, leading to a posterior density that is not Gaussian. We compare the approximated Gaussian mixture Fisher-Rao particle flow with the Gaussian sum particle flow (Zhang and Meyer, 2024) and the Wasserstein gradient flow (Lambert et al., 2022). We also demonstrate numerically that using Stein's gradient and Hessian (53) with Gauss-Hermite particles (56) leads to efficient and stable approximation of the expectation terms (51). Finally, we also evaluate our Fisher-Rao flows in high-dimension scenarios, adopting the posterior generated in the context of Bayesian logistic regression with dimensions 50 and 100.

7.1 Gaussian Mixture Prior Case

In this case, we have the prior follows an equally weighted Gaussian mixture distribution with four components:

$$p(\mathbf{x}) = \sum_{i=1}^{4} p_{\mathcal{N}}(\mathbf{x}; \hat{\mathbf{x}}^{(i)}, P),$$

where:

$$\hat{\mathbf{x}}^{(i)} = \begin{bmatrix} \pm 5 \\ \pm 5 \end{bmatrix}, \quad P = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}.$$

The likelihood function also follows a Gaussian density $\mathcal{N}(\mathbf{z}; H\mathbf{x}, R)$, with

$$H = \begin{bmatrix} 2 & -0.2 \\ 0.3 & 2.5 \end{bmatrix}, \quad R = \begin{bmatrix} 170 & 64 \\ 64 & 230 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} 2.67 \\ 1.67 \end{bmatrix},$$

where $\mathbf{x}^*$ denotes the true state. In this case, we set a large observation covariance to clearly separate the four different modals of the posterior distribution. For the approximated Gaussian mixture Fisher-Rao flows, each mean parameter is sampled from one of the prior Gaussian mixture components $\boldsymbol{\mu}_{t=0}^{(k)} \sim \mathcal{N}(\hat{\mathbf{x}}^{(i)}, P)$. The initial variance parameters are set to $\Sigma_{t=0}^{(k)} = 3P$, and initial weights are set to $\omega_{t=0}^{(k)} = 1/K$, where K denotes the total number of Gaussian mixture components employed in our approach. Each Gaussian

mixture component holds 16 Gauss-Hermite particles generated according to (56). The Wasserstein gradient flow method uses the same initialization approach as the approximated Gaussian mixture Fisher-Rao flows. For the Gaussian sum particle flow method, each Gaussian particle flow component is initialized with 1000 randomly sampled particles from $\mathcal{N}(\boldsymbol{\mu}^{(k)}, P)$, where $\boldsymbol{\mu}^{(k)}$ is sampled from one of the prior Gaussian mixture components $\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\hat{\mathbf{x}}^{(i)}, P)$. This initialization method is employed for the Gaussian sum particle flow to ensure consistency with the other two methods, as it utilizes particles to compute the empirical mean during propagation. Figure 2 shows the results when a single Gaussian density is used to approximate the posterior density. Figure 3 shows simulation results with 20 Gaussian mixture components. Figure 4 shows a comparison of the approximation accuracy of the expectation terms.

7.2 Nonlinear Observation Case

In this case, we have the prior follow a single Gaussian distribution $\mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, P)$, with

$$\hat{\mathbf{x}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad P = \begin{bmatrix} 5.5 & -1.5 \\ -1.5 & 5.5 \end{bmatrix}.$$

The likelihood function also follows a Gaussian distribution $\mathcal{N}(\mathbf{z}; H(\mathbf{x}), R)$, with

$$H(\mathbf{x}) = \|\mathbf{x}\|, \quad R = 2, \quad \mathbf{x}^* = \begin{bmatrix} 4.7 \\ -3.1 \end{bmatrix},$$

where $\mathbf{x}^*$ denotes the true state used to generate observations. For our approximated Gaussian mixture Fisher-Rao flows, each initial mean parameter is sampled from the prior $\boldsymbol{\mu}_{t=0}^{(k)} \sim \mathcal{N}(\hat{\mathbf{x}}, P)$, the initial variance parameters are set to $\Sigma_{t=0}^{(k)} = 3P$ and initial weights are set to $\omega_{t=0}^{(k)} = 1/K$, where K denotes the total number of Gaussian mixture components employed in our approach.. The Wasserstein gradient flow method uses the same initialization approach as the Gaussian mixture approximated Fisher-Rao flows. For the Gaussian sum particle flow method, each Gaussian particle flow component is initialized with 1000 randomly sampled particles from $\mathcal{N}(\boldsymbol{\mu}^{(k)}, P)$, where $\boldsymbol{\mu}^{(k)}$ is sampled from the prior density $\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\hat{\mathbf{x}}, P)$. This initialization method is employed for the Gaussian sum particle flow to ensure consistency with the other two methods, as it utilizes particles to compute the empirical mean during propagation. Figure 5 shows the case where a single Gaussian density is used to approximate the posterior distribution. Figure 6 shows simulation results with 20 Gaussian mixture components. Finally, Figure 7 shows a comparison of the approximation accuracy of the expectation terms.

7.3 Bayesian Logistic Regression

In this case, we use an unnormalized posterior generated in the context of Bayesian logistic regression associated with a two-class dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ provided by Lambert et al. (2022). The probability of the binary label $y_i \in \{0, 1\}$ given $\mathbf{x}_i \in \mathbb{R}^n$ and parameter $\mathbf{z} \in \mathbb{R}^n$ is defined by the following Bernoulli density:

$$p(y_i|\mathbf{x}_i; \mathbf{z}) = \sigma(\mathbf{x}_i^\top \mathbf{z})^{y_i}(1 - \sigma(\mathbf{x}_i^\top \mathbf{z}))^{1-y_i},$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic function. The unnormalized posterior associated with data $\mathcal{D}$ can be specified as follows:

$$p(\mathbf{z}|\mathcal{D}) = \Pi_{i=1}^{N} p(y_i|\mathbf{x}_i; \mathbf{z}).$$

For our Gaussian Fisher-Rao flows, the initial mean parameter is sampled from a Gaussian density $\boldsymbol{\mu}_{t=0} \sim \mathcal{N}(\mathbf{0}, 5I)$, and the initial variance parameter is set to $\Sigma_{t=0} = 5I$. The Wasserstein gradient flow method uses the same initialization approach as the Gaussian Fisher-Rao flows. For our approximated Gaussian mixture Fisher-Rao flows, each initial mean parameter is sampled from the prior $\boldsymbol{\mu}_{t=0}^{(k)} \sim \mathcal{N}(\mathbf{0}, I)$, the initial variance parameters are set to $\Sigma_{t=0}^{(k)} = I$ and initial weights are set to $\omega_{t=0}^{(k)} = 1/K$, where K denotes the total number of Gaussian mixture components employed in our approach. Due to high dimensionality, we report the approximated evidence lower bound (ELBO) for each method instead of the approximated KL divergence.

When simulating the Fisher-Rao particle flows, it is more efficient to only propagate the mean and the covariance parameter and recover the particles using Theorem 10 when necessary. Figure 8 shows that using the recovered particles achieves identical trajectories compared to the propagated particles. Moreover, for all test cases, both the Gaussian Fisher-Rao particle flow and the approximated Gaussian mixture particle flow achieve similar ELBO after 100 iterations, indicating that it is sufficient to use a single Gaussian to approximate the posterior density. For all test cases, the Fisher-Rao particle flows outperform the Wasserstein gradient flow and achieve a better convergence rate.

8 Conclusions

We have developed a variational formulation of the particle flow particle filter. We have shown that the transient density used to derive the particle flow particle filter is a time-scaled solution to the Fisher-Rao gradient flow. Based on this observation, we have first derived the Gaussian Fisher-Rao particle flow, which reduces to the Exact Daum and Huang (EDH) flow under linear Gaussian assumptions. Next, we have adopted a Gaussian mixture variational density to further enhance its expressive power and derive the corresponding approximated Gaussian mixture Fisher-Rao particle flow. We have also discussed various implementation strategies that make the computation of the flow parameters more efficient and ensure stable propagation of the Fisher-Rao flows. In a series of simulations, we illustrate the equivalence between the Gaussian Fisher-Rao particle flow and the EDH flow under linear Gaussian assumptions, how the approximated Gaussian mixture Fisher-Rao particle flow captures the multimodal behavior of the posterior, and the good performance of the Fisher-Rao particle flows in a Bayesian logistic regression task. Future work will explore the application of the Fisher-Rao flows to robot state estimation tasks and the extension of the results to Lie groups, exploiting the geometry of the configuration space, particularly the special Euclidean manifold.

Acknowledgments and Disclosure of Funding

The authors declare no competing interests. We gratefully acknowledge support from ONR Award N00014-23-1-2353 and NSF FRR CAREER 2045945.

References

Johan Alenlöv, Arnoud Doucet, and Fredrik Lindsten. Pseudo-Marginal Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22(141):1–45, 2021.

Daniel Alspach and Harold Sorenson. Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.

Shun-ichi Amari. *Information Geometry and its Applications*, volume 194. Springer, 2016.

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel A. J. van Gerven, and Eric Maris. Wasserstein Variational Inference. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Brian DO Anderson and John B Moore. *Optimal Filtering*. Courier Corporation, 2005.

David Barber and Christopher Bishop. Ensemble Learning for Multi-Layer Networks. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

Timothy D Barfoot. Multivariate Gaussian Variational Inference by Natural Gradient Descent. *arXiv preprint arXiv:2001.10025*, 2020.

Thomas Bayes. LII. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a Letter to John Canton, A.M.F.R.S. *Philosophical transactions of the Royal Society of London*, 53:370–418, 1763.

Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for Hierarchical Models. In *Current Trends in Bayesian Methodology with Applications*, pages 119–142. Chapman and Hall/CRC, 2015.

Peter Bickel, Bo Li, and Thomas Bengtsson. Sharp Failure Rates for the Bootstrap Particle Filter in High Dimensions. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3, pages 318–330. Institute of Mathematical Statistics, 2008.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Georges Bonnet. Transformations des Signaux Aléatoires a Travers Les Systemes Non Linéaires Sans Mémoire. In *Annales des Télécommunications*, volume 19, pages 203–220. Springer, 1964.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance. *arXiv preprint arXiv:2302.11024*, 2023a.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Sampling via Gradient Flows in the Space of Probability Measures. *arXiv preprint arXiv:2310.03597*, 2023b.

David F Crouse and Codie Lewis. Consideration of Particle Flow Filter Implementations and Biases. *Naval Research Laboratory Memo*, pages 1–17, 2019.

Haskell B Curry. The Method of Steepest Descent for Non-Linear Minimization Problems. *Quarterly of Applied Mathematics*, 2(3):258–261, 1944.

Fred Daum and Jim Huang. Nonlinear Filters with Log-Homotopy. In *SPIE Signal and Data Processing of Small Targets*, volume 6699, pages 423–437, 2007.

Fred Daum and Jim Huang. Nonlinear Filters with Particle Flow. In *SPIE Signal and Data Processing of Small Targets*, volume 7445, pages 315–323, 2009.

Fred Daum, Jim Huang, and Arjang Noushin. Exact Particle Flow for Nonlinear Filters. In *SPIE Signal Processing, Sensor Fusion, and Target Recognition*, volume 7697, pages 92–110, 2010.

Arnaud Doucet, Will Sussman Grathwohl, Alexander G. D. G. Matthews, and Heiko Strathmann. Score-Based Diffusion meets Annealed Importance Sampling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=9cU2iW3bz0`.

Arnaud Ducet, Adam M Johansen, et al. A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later. *Handbook of Nonlinear Filtering*, 12(656-704):3, 2009.

Tomas Geffner and Justin Domke. Langevin Diffusion Variational Inference. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 576–593. PMLR, 25–27 Apr 2023.

Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel Approach to Nonlinear/non-Gaussian Bayesian State Estimation. In *IEE Proceedings F Radar and Signal Processing*, volume 140, pages 107–113, 1993.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(40):1303–1347, 2013.

Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Efficient Derivative-Free Bayesian Inference for Large-Scale Inverse Problems. *Inverse Problems*, 38(12):125006, oct 2022.

Andrew H Jazwinski. *Stochastic Processes and Filtering Theory*. Courier Corporation, 2007.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1999.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker-Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Simon J Julier, Jeffrey K Uhlmann, and Hugh F Durrant-Whyte. A new Approach for Filtering Nonlinear Systems. In *American Control Conference*, volume 3, pages 1628–1632 vol.3, 1995.

Rudolph Emil Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian Learning Rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

Genshiro Kitagawa. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational Inference via Wasserstein Gradient Flows. In *Advances in Neural Information Processing Systems*, volume 35, pages 14434–14447, 2022.

Yunpeng Li and Mark Coates. Particle Filtering with Invertible Particle Flow. *IEEE Transactions on Signal Processing*, 65(15):4102–4116, 2017.

Yunpeng Li, Soumyasundar Pal, and Mark J. Coates. Invertible Particle-Flow-Based Sequential MCMC with Extension to Gaussian Mixture Noise Models. *IEEE Transactions on Signal Processing*, 67(9):2499–2512, 2019.

Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-Family Approximations. In *PMLR International Conference on Machine Learning*, pages 3992–4002, 2019a.

Wu Lin, Mohammad Emtiyaz Khan, and Mark Schmidt. Stein's Lemma for the Reparameterization Trick with Exponential Family Mixtures. *arXiv preprint arXiv:1910.13398*, 2019b.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Martin Morf, Bernard Levy, and Thomas Kailath. Square-Root Algorithms for the Continuous-Time Linear Least Squares Estimation Problem. In *IEEE Conference on Decision and Control*, pages 944–947, 1977.

Radford Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10), 2020.

Manfred Opper and Cédric Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, 21:786–792, 2009.

Soumyasundar Pal and Mark Coates. Gaussian Sum Particle Flow Filter. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 1–5, 2017.

Soumyasundar Pal and Mark Coates. Particle Flow Particle Filter for Gaussian Mixture Noise Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4249–4253, 2018.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The Matrix Cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Robert Price. A Useful Theorem for Nonlinear Devices Having Gaussian Inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 814–822, 2014.

Simo Särkkä. On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems. *IEEE Transactions on Automatic Control*, 52(9):1631–1641, 2007.

Simo Särkkä and Lennart Svensson. *Bayesian Filtering and Smoothing*, volume 17. Cambridge University Press, 2023.

Matthias Seeger. Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

Martin J Wainwright, Michael I Jordan, et al. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Kari C. Ward and Kyle J. DeMars. Information-Based Particle Flow With Convergence Control. *IEEE Transactions on Aerospace and Electronic Systems*, 58(2):1377–1390, 2022.

Andre Wibisono, Varun Jog, and Po-Ling Loh. Information and Estimation in Fokker-Planck Channels. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2673–2677, 2017.

Wenyu Zhang and Florian Meyer. Multisensor Multiobject Tracking with Improved Sampling Efficiency. *IEEE Transactions on Signal Processing*, 72:2036–2053, 2024.

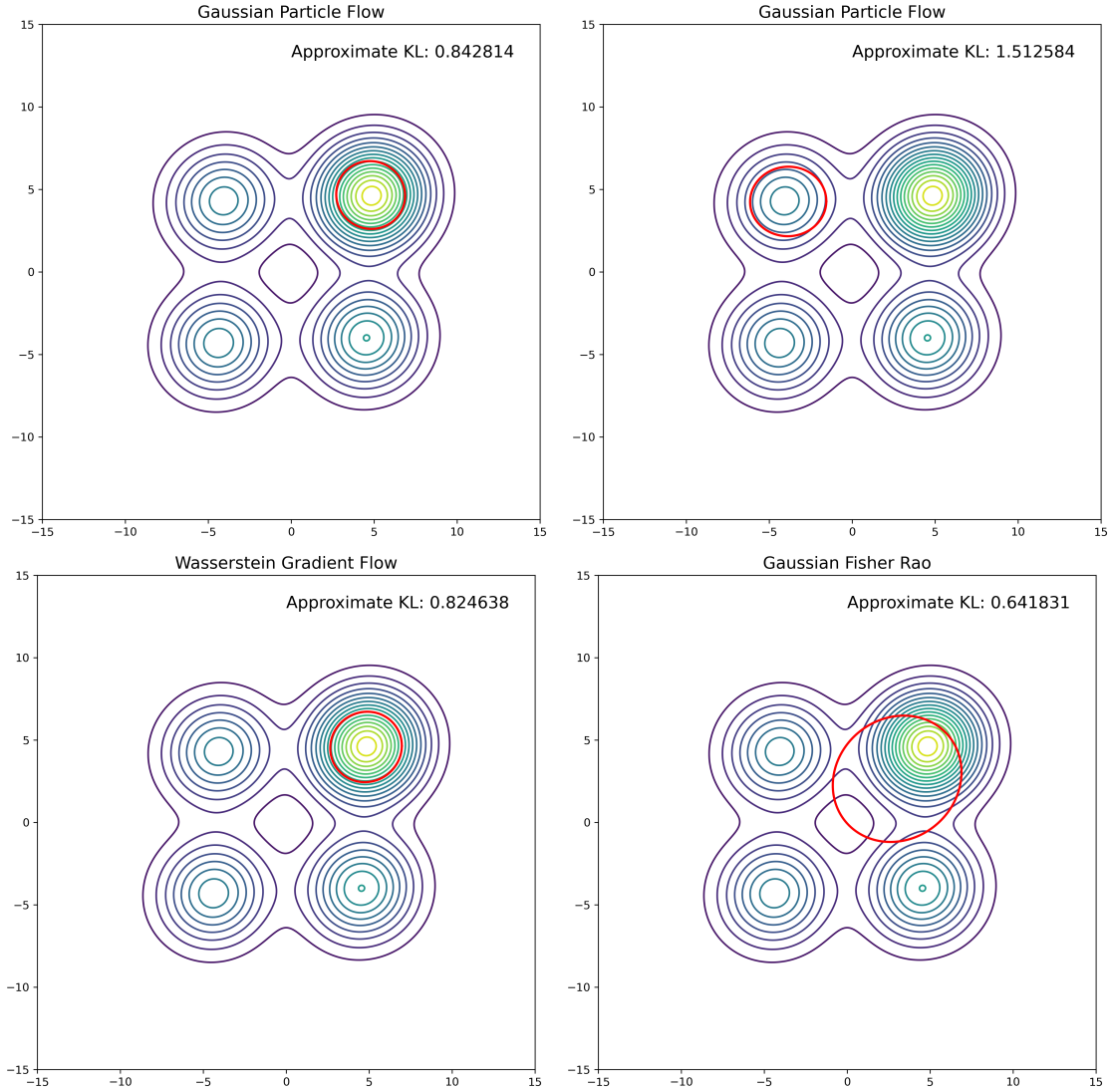

Figure 2: Comparison of the Gaussian approximated Fisher-Rao particle flow with the Gaussian particle flow and the Wasserstein gradient flow for the Gaussian mixture prior case. For each method, we use a Gaussian mixture with one component to approximate the posterior density. The covariance contour corresponding to one Mahalanobis distance is overlaid on the reference contour. The top two figures display results generated by the Gaussian particle flow with different initializations. The top left figure shows the result with particles generated from $\mathcal{N}(\hat{\mathbf{x}}^{(1)}, P)$, while the top right figure shows the result with particles generated from $\mathcal{N}(\hat{\mathbf{x}}^{(2)}, P)$. The Gaussian approximated Fisher-Rao particle flow achieves the lowest approximated KL divergence. In contrast, the Wasserstein gradient flow method converges to the modal with the largest component weight. Additionally, the Gaussian particle flow method is sensitive to initialization, and its convergence behavior is highly dependent on the initial conditions.

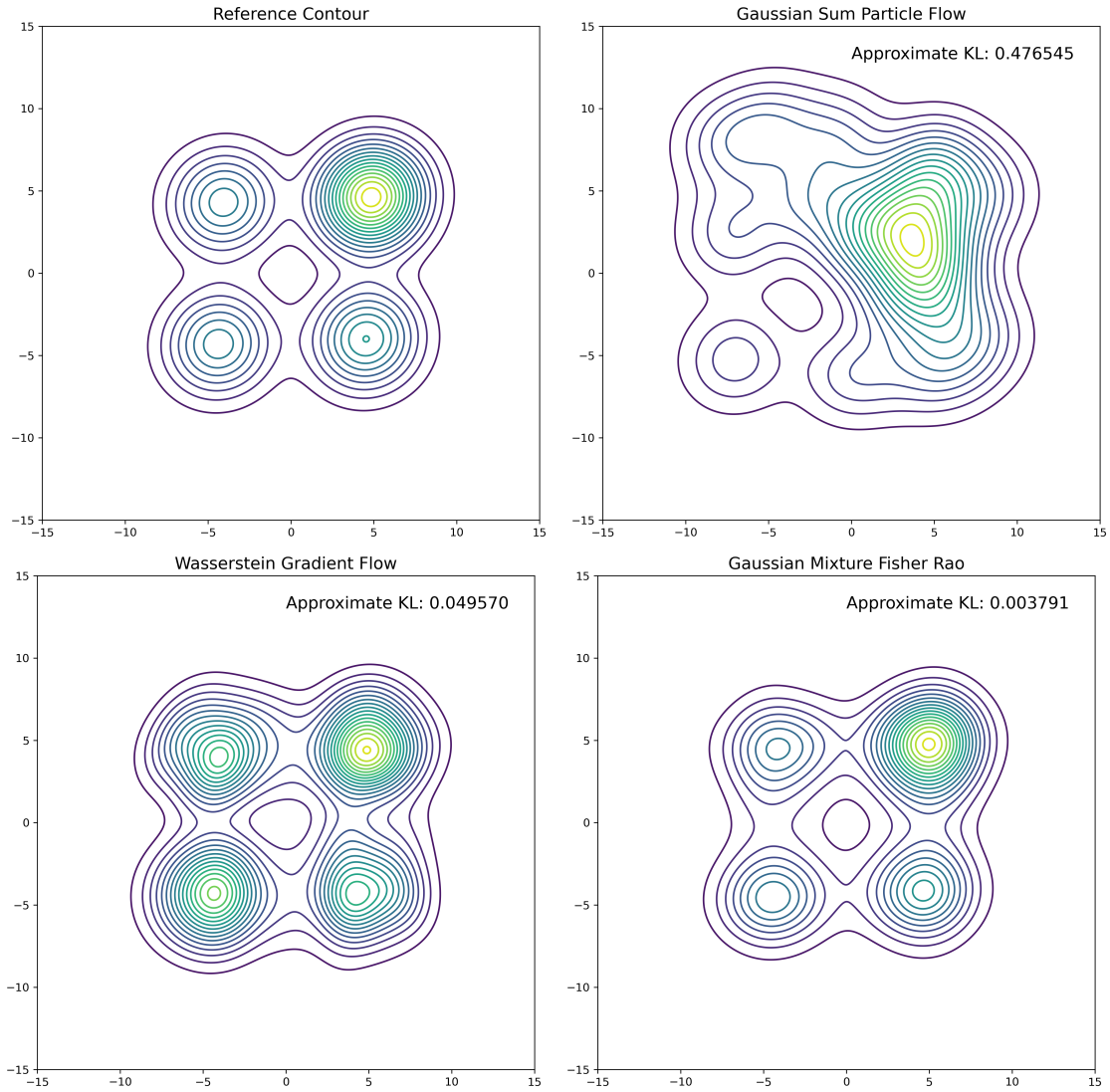

Figure 3: Comparison of the Gaussian mixture approximated Fisher-Rao particle flow with the Gaussian sum particle flow and the Wasserstein gradient flow for the Gaussian mixture prior case. For each method, we use a Gaussian mixture with 20 component to approximate the posterior density. Our Gaussian mixture approximated Fisher-Rao particle flow successfully captures the four unequally weighted modals of the posterior density, while the Gaussian sum particle flow fails to clearly separate these modals. The Wasserstein gradient flow achieves results comparable to ours but assigns incorrect component weights.

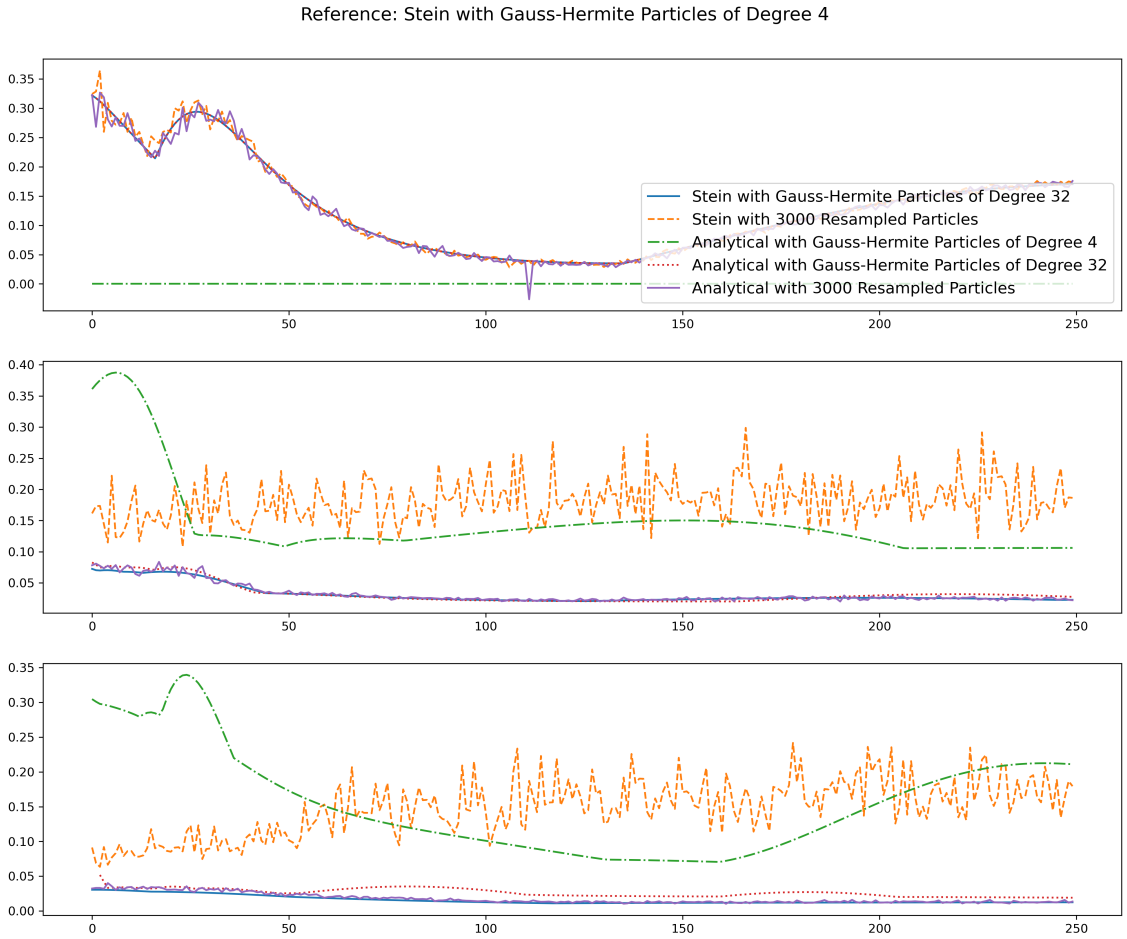

Figure 4: The comparison of expectation evaluations for the Gaussian mixture prior case is illustrated. The top plot displays the difference in the expected $V(\mathbf{x})$ function, the middle plot shows the difference in the expected gradient of $V(\mathbf{x})$, and the bottom plot depicts the difference in the expected Hessian $V(\mathbf{x})$. The maximum difference between the comparing methods and the reference method across all Gaussian mixture components is reported. The reference method uses Stein's gradient and Hessian with Gauss-Hermite particles of degree 4, resulting in stable propagation.

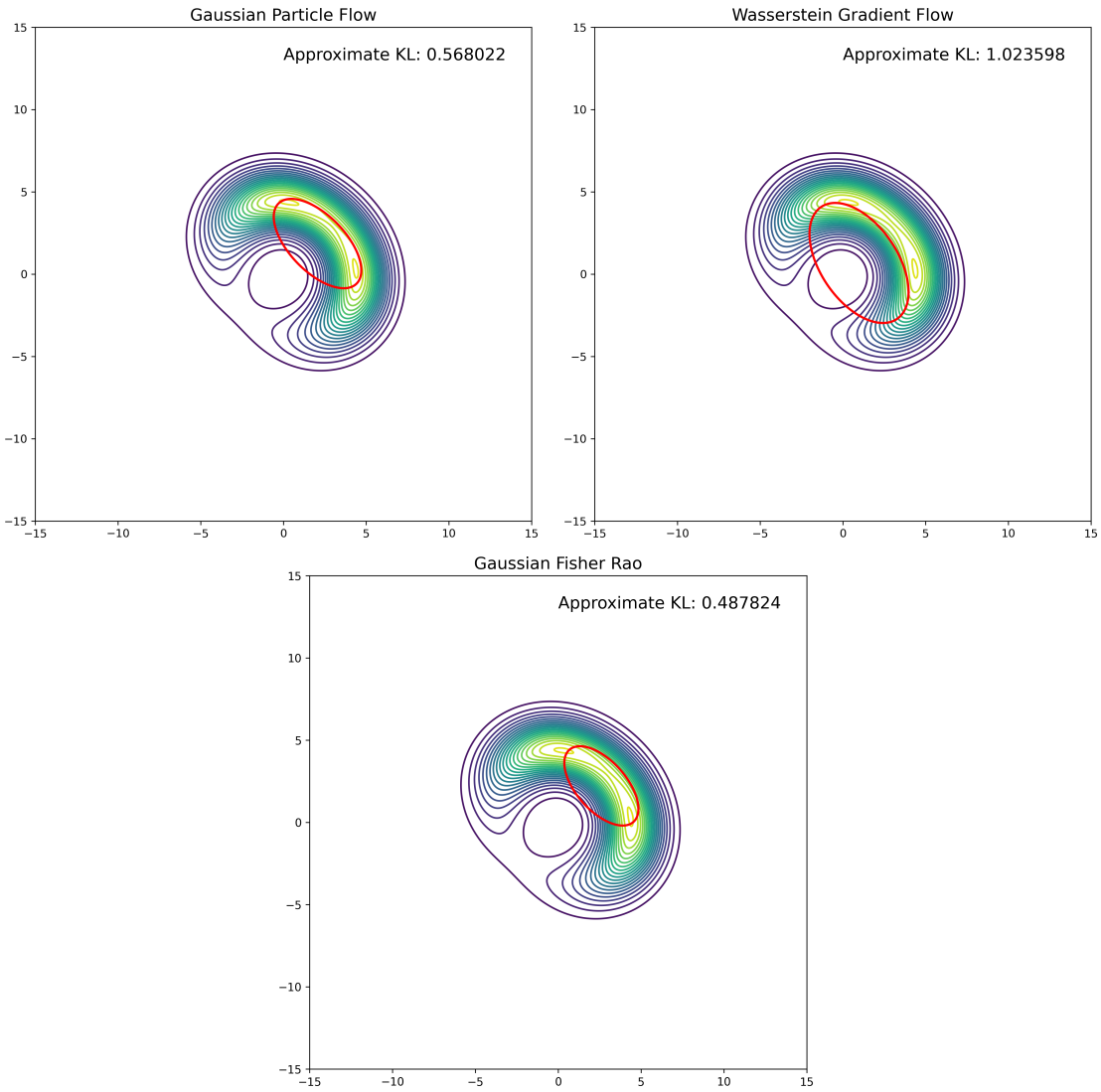

Figure 5: Comparison of the Gaussian approximated Fisher-Rao particle flow with the Gaussian particle flow and the Wasserstein gradient flow for the nonlinear observation model case. For each method, we use a single Gaussian to approximate the posterior density. The covariance contour corresponding to one Mahalanobis distance is overlaid on the reference contour. The Gaussian approximated Fisher-Rao particle flow provides the most accurate Gaussian approximation of the posterior density, resulting in the lowest approximated KL divergence. In contrast, the Gaussian particle flow method produces an approximation that is slightly misaligned with the region of highest posterior probability, leading to a higher approximated KL divergence. The Wasserstein gradient flow method fails to accurately capture the region of the highest posterior probability, resulting in the highest approximated KL divergence.

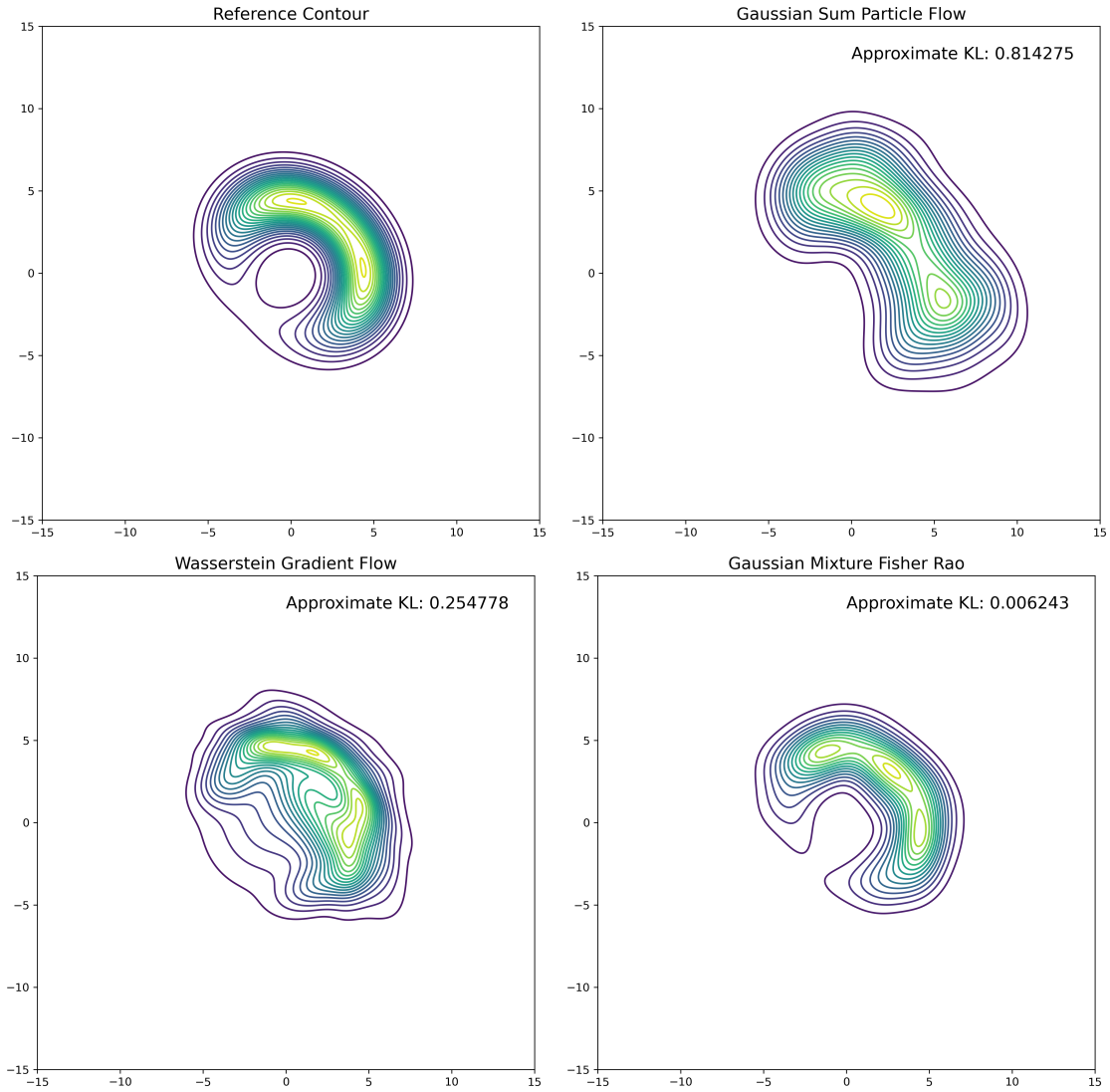

Figure 6: Comparison of the Gaussian mixture approximated Fisher-Rao particle flow with the Gaussian sum particle flow and the Wasserstein gradient flow for the nonlinear observation model case. For each method, we use a Gaussian mixture with 20 components to approximate the posterior density. For each method, we plot the contour generated by the variational Gaussian mixture density. The Gaussian mixture approximated Fisher-Rao particle flow captures the banana-shape posterior and yields the lowest approximated KL divergence. While the Wasserstein gradient flow does not capture the shape of the posterior well, yielding a slightly higher approximated KL divergence. On the other hand, the Gaussian sum particle flow fails to capture the shape of the posterior, yielding the highest approximated KL divergence.

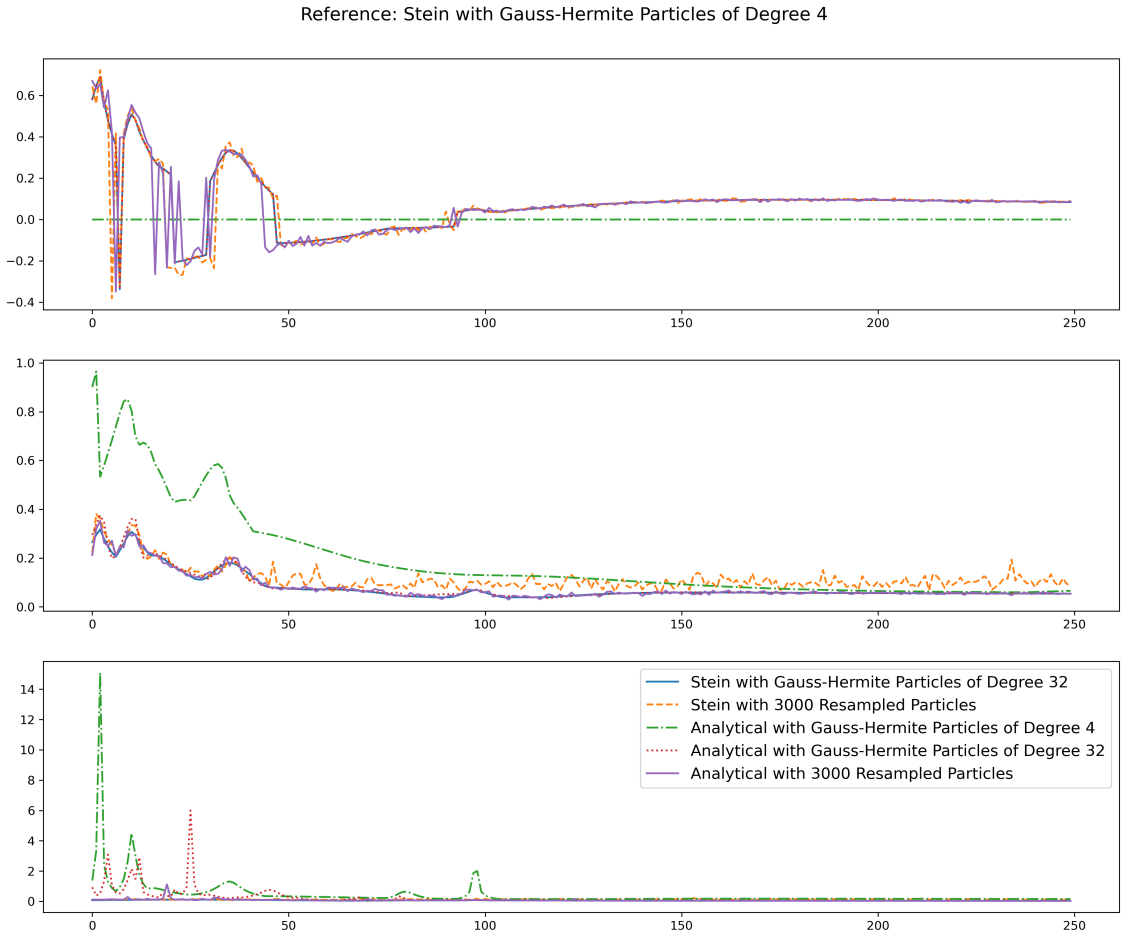

Figure 7: The comparison of expectation evaluations for the nonlinear observation model case is illustrated. The top plot displays the difference in the expected $V(\mathbf{x})$ function, the middle plot shows the difference in the expected gradient of $V(\mathbf{x})$, and the bottom plot depicts the difference in the expected Hessian $V(\mathbf{x})$. The maximum difference between the comparing methods and the reference method across all Gaussian mixture components is reported. The reference method uses Stein's gradient and Hessian with Gauss-Hermite particles of degree 4, resulting in stable propagation.

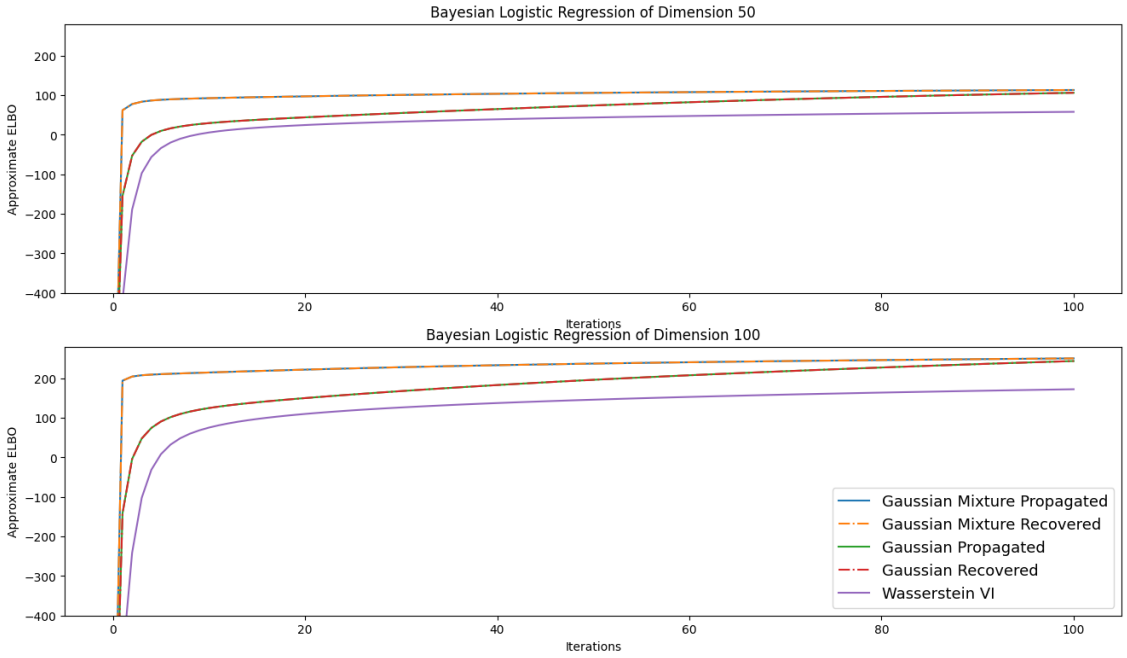

Figure 8: Comparison of the Fisher-Rao particle flows with the Wasserstein gradient flow for the Bayesian logistic regression task with different dimensions. The approximated ELBO is reported for each method. It can be shown that using the recovered particles from Theorem 10 yields performance identical to that of the propagated particles. Additionally, we observe that approximating the posterior with a single Gaussian is sufficient for this task, as it achieves nearly identical results to a Gaussian mixture with 5 components. Our Fisher-Rao particle flows outperform the Wasserstein gradient flow.