PAPER

Efficient, multimodal, and derivative-free bayesian inference with Fisher–Rao gradient flows

IOP Publishing Inverse Problems

Inverse Problems **40** (2024) 125001 (39pp) https://doi.org/10.1088/1361-6420/ad847b

Efficient, multimodal, and derivative-free bayesian inference with Fisher–Rao gradient flows

Yifan Chen[1], **Daniel Zhengyu Huang**[2,*], **Jiaoyang Huang**[3], **Sebastian Reich**[4] **and Andrew M Stuart**[5]

[1] Courant Institute, New York University, New York, NY, United States of America
[2] Beijing International Center for Mathematical Research, Center for Machine Learning Research, Peking University, Beijing, People's Republic of China
[3] University of Pennsylvania, Philadelphia, PA, United States of America
[4] Universität Potsdam, Potsdam, Germany
[5] California Institute of Technology, Pasadena, CA, United States of America

E-mail: huangdz@bicmr.pku.edu.cn, yifan.chen@nyu.edu, huangjy@wharton.upenn.edu, sebastian.reich@uni-potsdam.de and astuart@caltech.edu

Received 27 June 2024; revised 3 October 2024
Accepted for publication 8 October 2024
Published 17 October 2024

Abstract

In this paper, we study efficient approximate sampling for probability distributions known up to normalization constants. We specifically focus on a problem class arising in Bayesian inference for large-scale inverse problems in science and engineering applications. The computational challenges we address with the proposed methodology are: (i) the need for repeated evaluations of expensive forward models; (ii) the potential existence of multiple modes; and (iii) the fact that gradient of, or adjoint solver for, the forward model might not be feasible. While existing Bayesian inference methods meet some of these challenges individually, we propose a framework that tackles all three systematically. Our approach builds upon the Fisher–Rao gradient flow in probability space, yielding a dynamical system for probability densities that converges towards the target distribution at a uniform exponential rate. This rapid convergence is advantageous for the computational burden outlined in (i). We apply Gaussian mixture approximations with operator splitting techniques to simulate the flow numerically; the resulting approximation can capture multiple modes thus addressing (ii). Furthermore, we employ the Kalman methodology to facilitate a derivative-free update of these Gaussian components and their respective weights, addressing the issue in (iii). The proposed methodology

* Author to whom any correspondence should be addressed.

results in an efficient derivative-free posterior approximation method, flexible enough to handle multi-modal distributions: Gaussian Mixture Kalman Inversion (GMKI). The effectiveness of GMKI is demonstrated both theoretically and numerically in several experiments with multimodal target distributions, including proof-of-concept and two-dimensional examples, as well as a large-scale application: recovering the Navier–Stokes initial condition from solution data at positive times.

Keywords: Bayesian inverse problems, sampling, derivative-free methods, multimodal, kalman methodology, fisher–rao gradient flow, gaussian mixtures

1. Introduction

In this paper, we introduce the posterior approximation method called Gaussian Mixture Kalman Inversion (GMKI), designed for solution of partial differential equation (PDE) inverse problems for which forward model evaluation is expensive, derivative/adjoint calculations cannot be used and multiple modes are present. In section 1.1 we give the context, followed in section 1.2 with details of our guiding motivations. Section 1.3 describes the key ingredients of the algorithm and section 1.4 the contributions. In section 1.5 we give a detailed literature review and in section 1.6 we describe the organization of the paper.

1.1. Context

Sampling a target probability distribution known up to normalization constants is a classical problem in science and engineering. In this paper, we focus specifically on targets resulting from Bayesian inverse problems [1, 2] involving recovery of unknown parameter $\theta \in \mathbb{R}^{N_\theta}$ from noisy observation $y \in \mathbb{R}^{N_y}$, through forward model

$$y = \mathcal{G}(\theta) + \eta. \tag{1}$$

Here, $\mathcal{G}$ denotes the forward mapping which, for the problems we focus on, is nonlinear and requires solution of a PDE to evaluate. The observational noise η is here assumed to be Gaussian: $\eta \sim \mathcal{N}(0, \Sigma_\eta)$. By assigning a Gaussian prior $\mathcal{N}(r_0, \Sigma_0)$ to the unknown θ, the Bayesian framework leads to the posterior distribution

$$\rho_{\text{post}}(\theta) \propto \exp(-\Phi_R(\theta)), \tag{2a}$$

$$\Phi_R(\theta) = \Phi(\theta) + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2, \tag{2b}$$

$$\Phi(\theta) = \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - \mathcal{G}(\theta))\|^2. \tag{2c}$$

Here Φ is the negative log likelihood. Minimization of Φ_R is a nonlinear least-squares problem which defines the maximum a posterior (MAP) point estimator for the Bayesian inverse problem. It is the goal of this paper to develop an efficient method for approximating ρ_{post} defined by Φ_R in the specific setting which we now outline.

1.2. Guiding motivations

We give more detail concerning the motivations behind the specific posterior approximation method developed here. Firstly we note that an appropriate unit of cost in solution of Bayesian

inverse problems is the evaluation of $\mathcal{G}$ as this will be required multiple times for methods such as Markov chain Monte Carlo (MCMC) [3] and SMC [4, 5]; when evaluation of $\mathcal{G}$ requires running large scale PDE solvers fast convergence is paramount. Secondly, we note that multiple modes, caused by multiple minimizers of Φ_R, cause many methods to become slow [6], expending multiple steps in one mode before moving to another [7, 8]; in addition, many Gaussian approximation based methods are unable to capture multiple modes. Nevertheless, exploring all these modes is necessary since missing one could lead to detrimental effects on engineering or science predictions; thirdly we note that the gradient of Φ_R may not be available or even feasible. This might be because the computational models are only given as a black box (e.g. in global climate model calibration [9, 10]), the numerical methods are not differentiable (e.g. in the embedded boundary method [11–14] and adaptive mesh refinement [15, 16]), or because of inherently discontinuous physics (i.e. in fracture [17] or cloud modeling [18, 19]). In this paper, we address these three challenges by combining, respectively, Fisher–Rao gradient flows, Gaussian mixture approximations, and Kalman methodology. The resulting posterior approximation method, GMKI, is fast due to the uniform exponential convergence of Fisher–Rao gradient flows, can capture multiple modes since Gaussian mixture approximations are employed, and is derivative-free thanks to the systematic Kalman methodology.

1.3. Key ingredients of GMKI

In sampling, it is widely accepted practice to construct a dynamical system for a density that gradually evolves to the posterior distribution, or its approximation, after a specified finite time or at infinite time. Numerical approximation of this dynamics, using either particle or parametric methods, leads to practical algorithms. These include sequential Monte Carlo (SMC, specified finite time) [4], and MCMC (infinite time) [3] that are commonly used in Bayesian inference. In recent years, gradient flows in the probability space have become a popular choice of dynamical systems [20–22]; their study presents the opportunity to profoundly influence our understanding and development of sampling algorithms.

In general, the convergence rates of different gradient flows can vary significantly. In this paper, we focus on the Fisher–Rao gradient flow of the Kullback–Leibler (KL) divergence [22–24]:

$$\frac{\partial \rho_t}{\partial t} = \rho_t \left(\log \rho_{\text{post}} - \log \rho_t\right) - \rho_t \mathbb{E}_{\rho_t}\left[\log \rho_{\text{post}} - \log \rho_t\right]. \tag{3}$$

The Fisher–Rao gradient flow converges to its steady state, ρ_{post}, exponentially fast, with a rate of $\mathcal{O}(e^{-t})$; see proposition 1, [22, theorem 4.1], and also [23–25]. This convergence rate is uniform and independent of ρ_{post}, in particular its log-Sobolev constant, which typically determines the convergence rates of other gradient flows, such as the Wasserstein gradient flow. It is worth noting that the log-Sobolev constant may behave poorly when the posterior distribution ρ_{post} is highly anisotropic or multimodal [7, 8]. Thus, we consider equation (3) as a desirable flow for sampling general distributions.

We introduce numerical approximations of equation (3) to construct practical algorithms. Particle methods represent the current density ρ_t by a (possibly weighted) sum of Dirac measures evaluated at an ensemble of particles. The flow equation (3) can then be realized as a birth-death dynamics of these particles [23, 24]. However, the birth-death rate depends on the density, so it is necessary to constantly reconstruct ρ_t from the empirical particle distribution. In [23, 24], kernel density estimators have been applied for the reconstruction, but their performance may be affected when the dimension of the problem becomes large. Moreover, birth-death dynamics alone cannot change the support of the distribution, so additional steps

need to be added to explore the space [23, 24, 26]; such exploration steps change the dynamics and may also lead to challenges in high dimensional problems.

Parametric methods, which reduce the gradient flow into some parametric density space, constitute another common choice of numerical approximation. One way to do this is to project the flow equation (3) into the Gaussian space [27–29], via a moment closure approach. The resulting system for the mean and covariance is given by [22]:

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = C_t \mathbb{E}_{\rho_{a_t}} \left[\nabla_\theta \log \rho_{\text{post}}\right] \qquad \frac{\mathrm{d}C_t}{\mathrm{d}t} = C_t + C_t \mathbb{E}_{\rho_{a_t}} \left[\nabla_\theta \nabla_\theta \log \rho_{\text{post}}\right] C_t, \tag{4}$$

where ρ_t in equation (3) is approximated by a Gaussian $\rho_{a_t} = \mathcal{N}(m_t, C_t)$ in equation (4); here $a_t = (m_t, C_t)$ is the unknown parameter. We note that one may also derive the above flow by natural gradient methods in variational inference [30–32]; see discussions in [22]. Theoretically, it has been shown in [22] that equation (4) converges exponentially fast to the best Gaussian approximation of ρ_{post} in the KL divergence sense, when ρ_{post} is log-concave. Therefore, by simulating equation (4), we get a Gaussian approximation of the posterior; this can be done through direct time integration or ensemble methods.

More generally, for multimodal problems, Gaussian mixture approximations have been studied in the literature under the variational inference framework [29, 33, 34]. These approaches require the evaluation of the gradient and sometimes even the Hessian matrix of $\log \rho_{\text{post}}$, as shown in equation (4), which are not directly feasible for the type of problems which are our focus in this paper.

On the other hand, Kalman methodology has emerged as an effective methodology for sampling for both filter and inverse problems [35–42]. Similar to the parametric methods discussed above, it relies on Gaussian approximations; however, it additionally utilizes the structure of the problem, i.e. the least-squares form of the posterior as described in equation (2). Notably, the Kalman methodology can lead to derivative-free algorithms such as the Ensemble Kalman Filter, Unscented Kalman Filter, and Ensemble Kalman Inversion (EKI), all defined in [42]. Moreover, the recent work on EKI and its variants in [43] can be interpreted as applying Kalman-type approximations to the Fisher–Rao gradient flow equation (3), although this gradient structure was not explicitly pointed out in the original paper. The effectiveness of this method has been demonstrated on large-scale inverse problems in science and engineering, with up to hundreds of dimensions. However, since only Gaussian approximations are used, the method may not be suitable for multimodal posterior distributions.

1.4. Contributions

The primary focus of this paper is to extend the Kalman methodology in [43] to Gaussian mixture approximations of the Fisher–Rao gradient flow. This leads to GMKI, a derivative-free posterior approximation method that converges fast and captures multiple modes for the challenging inversion problems studied here. We make the following contributions:

(i) We propose an operator splitting approach to integrate the Fisher–Rao gradient flow in time, which leads to an exploration step that explores the space freely and an exploitation step that harnesses the data and prior information. We prove the resulting exploration-exploitation scheme converges exponentially fast to the target distribution at the discrete time level (section 2).
(ii) We demonstrate a connection between the continuous time limit of the pre-existing algorithm in [43] and Gaussian variational inference (section 3).

(iii) We apply Gaussian mixture approximations to the exploration-exploitation scheme. We utilize the Kalman methodology to update the weights and locations of the mixtures. This leads to our derivative-free algorithm, GMKI, for sampling multimodal distributions (section 4).
(iv) We analyze GMKI by deriving the continuous time limit of the dynamics. Based on the continuous dynamics, we study its exploration effects, establish its affine invariant property, connect the methodology to variational inference with Gaussian mixtures, and investigate the convergence properties (section 5).
(v) We demonstrate, on one/two-dimensional model problems as well as a high-dimensional application (recovering the Navier–Stokes initial condition from solution data at positive times), that GMKI is able to capture multiple modes in approximately $\mathcal{O}(10)$ iterations, making it a promising approach for solving large scale Bayesian inverse problems. Our code is accessible online (section 6).

1.5. Literature review

The review of relevant literature concerns SMC and MCMC, variational inference, gradient flows and Kalman methodology.

1.5.1. SMC and MCMC. SMC [44] and MCMC [3] are common approaches used in Bayesian inference for sampling posteriors. They lead to dynamical systems of densities that progressively converge to the target distribution. For SMC, the dynamical system operates over finite time intervals so converges fast in the density level, but numerical approximations of the dynamical system can be challenging, with difficulties such as weight collapses. Such issues are more pronounced in the case of multimodal posteriors, requiring a substantial number of particles and a good initialization for SMC to succeed, due to its lack of exploration. Approximation of the finite-time dynamics in SMC via transport of measures has also been investigated [45–47]. The Fisher–Rao gradient flow used in this paper can be seen as an infinite time extension of SMC dynamics that allows efficient exploration while converging exponentially fast in the density level. MCMC approaches typically require $O(10^4)$ model runs, or more, for the type of PDE-based inversion arising in this paper; thus they are too costly. Moreover, most MCMC approaches are based on local moves and face significant challenges in the multimodal scenario.

1.5.2. Variational Inference. Variational inference [48–50] addresses the sampling problem equation (2) using optimization, typically with a lower computational cost compared to MCMC. The objective function, often chosen to be the KL divergence between the target distribution and a variational distribution, is minimized to get a closest approximate distribution within the variational distribution family. Gaussian distributions and Gaussian mixtures are often used as the variational distribution [33, 51–53]. The concept of natural gradients [28, 30–33] has been widely used to derive efficient optimization algorithms for variational inference. These algorithms typically require evaluations of gradient information for the log density. We also note that the Gaussian and Gaussian mixture ansatz has been used in conjunction with the Dirac–Frenkel variational principle to solve time-dependent PDEs of wave functions and probability densities [54, 55]. When the PDE is the Fisher–Rao gradient flow, these

methods can recover the parameter dynamics obtained by natural gradient flow in variational inference [56].

1.5.3. Fisher–Rao Gradient Flow. The Fisher–Rao gradient flow plays a key role in the design of sampling algorithms studied in this paper. There is a vast literature on the use of gradient flows of the KL divergence in the density space, employing different metric tensors, for sampling. We specifically focus on the Fisher–Rao metric, introduced by C.R. Rao [57], to derive the gradient flow equation (3), as it is the only metric, up to scaling, invariant under any diffeomorphism of the parameter space [58–60]. This invariance leads to a gradient flow converging at a rate independent of the target distribution. In practice, the Fisher–Rao gradient flow and its simulation by birth-death processes have been used in SMC samplers to reduce the variance of particle weights [4] and accelerate Langevin sampling [23, 24, 26] and statistical learning [61]. Kernel approximation of the flow has also been considered [46, 62]. Gaussian approximation of the Fisher–Rao gradient flow is studied in [27], with close connections to natural gradient methods in variational inference.

1.5.4. Kalman methodology. The Kalman methodology encompasses a general class of approaches for solving filtering and inverse problems. They are based on replacing the Bayesian inference step in a filter, which may be viewed as governed by a prior to posterior map, by an approximate transport map which is exact for Gaussians; inverse problems are solved by linking them to a filter. Ensemble Kalman methods give rise to derivative-free algorithms, and are appropriate for solving filtering and inverse problems in which the desired probability distribution is close to Gaussian [35, 43, 63–65].

Beyond Gaussian approximations, a strand of research has extended Kalman filters to operate on Gaussian mixtures [66–73]. These methods model both prior and posterior distributions using Gaussian mixture distributions, leveraging a componentwise application of the Kalman methodology for each Gaussian component. Various techniques, such as recluster analysis and resampling techniques [74–78], as well as localization techniques [79–81], have been developed to enhance the robustness of these approaches. Nevertheless, existing methods in this category are tailored to transform a Gaussian mixture prior into a Gaussian mixture posterior; they can be understood as a Gaussian mixture approximation of the dynamics in SMC. The resulting methods lack full exploration of the space of possible solutions. In contrast, GMKI incorporates gradient flows, resulting in theoretical advantages manifest in its analysis. In practice, GMKI's exploration component enables effective traversal of the solution space, leading to robust performance without weight collapse.

1.6. Organization

The paper is organized as follows. In section 2, we introduce the Fisher–Rao gradient flow and the exploration-exploitation scheme for discretizing the flow in time. In section 3, the Gaussian approximation approach for spatial approximation is reviewed. In section 4, the proposed GMKI approach is presented, which relies on the Gaussian mixture approximation and Kalman methodology. In section 5, the continuous time dynamics of the GMKI approach is derived and analyzed. In section 6, numerical experiments are provided. We make concluding remarks in section 7.

2. Fisher–Rao gradient flow

In the following two subsections, we (i) briefly describe the Fisher–Rao gradient flow in the time-continuous settings; and (ii) introduce our operator splitting approach for simulating this flow in practice.

2.1. Continuous flow

In this paper we focus on the gradient flow arising from using the KL divergence

$$\mathrm{KL}\left[\rho\|\rho_{\mathrm{post}}\right]=\int\rho\log\left(\frac{\rho}{\rho_{\mathrm{post}}}\right)\mathrm{d}\theta=\mathbb{E}_{\rho}\left[\log\rho-\log\rho_{\mathrm{post}}\right] \tag{5}$$

as the energy functional, along with the Fisher–Rao metric tensor $M^{\mathrm{FR}}(\rho)^{-1}\psi=\rho\psi$; the resulting gradient flow has the form (see [22, section 4.1])

$$\begin{aligned}\frac{\partial\rho_t}{\partial t}&=-M^{\mathrm{FR}}\left(\rho_t\right)^{-1}\frac{\delta\mathrm{KL}\left[\rho_t\|\rho_{\mathrm{post}}\right]}{\delta\rho}\\&=\rho_t\left(\log\rho_{\mathrm{post}}-\log\rho_t\right)-\rho_t\mathbb{E}_{\rho_t}\left[\log\rho_{\mathrm{post}}-\log\rho_t\right].\end{aligned} \tag{6}$$

We have the following uniform exponential convergence result for this flow ([22, theorem 4.1]; see also related results in [23–25, 82]):

Proposition 1. *Let ρ_t satisfy equation (6). Assume there exist constants $K,B>0$ such that the initial density ρ_0 satisfies*

$$e^{-K\left(1+|\theta|^2\right)}\leqslant\frac{\rho_0\left(\theta\right)}{\rho_{\mathrm{post}}\left(\theta\right)}\leqslant e^{K\left(1+|\theta|^2\right)}, \tag{7}$$

and $\rho_0,\rho_{\mathrm{post}}$ have bounded second moments

$$\int|\theta|^2\rho_0\left(\theta\right)\mathrm{d}\theta\leqslant B,\quad\int|\theta|^2\rho_{\mathrm{post}}\left(\theta\right)\mathrm{d}\theta\leqslant B. \tag{8}$$

Then, for any $t\geqslant\log\big((1+B)K\big)$,

$$KL\left[\rho_t\|\rho_{\mathrm{post}}\right]\leqslant\left(2+B+eB\right)Ke^{-t}. \tag{9}$$

Accurate numerical simulation of equation (6) thus has the potential to exhibit uniform exponential convergence across a wide range of targets ρ_{post}.

2.2. Time-stepping via operator splitting

As a first step towards the derivation of an algorithm, we apply operator splitting to equation (6). Abusing notation we let ρ_n denote our approximation of ρ_{t_n} at time $t=t_n=n\Delta t$, where Δt denotes the time step, and solve sequentially

$$\frac{\partial\hat{\rho}_t}{\partial t}=-\hat{\rho}_t\left(\log\hat{\rho}_t-\mathbb{E}_{\hat{\rho}_t}\left[\log\hat{\rho}_t\right]\right),\quad\hat{\rho}_{t_n}=\rho_n,\quad t_n\leqslant t\leqslant t_{n+1}, \tag{10}$$

and

$$\frac{\partial\check{\rho}_t}{\partial t}=\check{\rho}_t\left(\log\rho_{\mathrm{post}}-\mathbb{E}_{\check{\rho}_t}\left[\log\rho_{\mathrm{post}}\right]\right),\quad\check{\rho}_{t_n}=\hat{\rho}_{t_{n+1}},\quad t_n\leqslant t\leqslant t_{n+1}. \tag{11}$$

The map $\rho_n \mapsto \rho_{n+1}$ is then defined by setting $\rho_{n+1} = \check{\rho}_{t_{n+1}}$. Further abusing notation we write $\hat{\rho}_{t_{n+1}} = \hat{\rho}_{n+1}$ and $\check{\rho}_{t_{n+1}} = \check{\rho}_{n+1}$. Note that all of $\rho_n, \hat{\rho}_{n+1}$ and $\check{\rho}_{n+1}$ are functions of θ. Both equations (10) and (11) admit explicit solutions and we may write

$$\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{e^{-\Delta t}}, \tag{12a}$$

$$\check{\rho}_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\, \rho_{\text{post}}(\theta)^{\Delta t}. \tag{12b}$$

Furthermore, using the first order approximation $e^{-\Delta t} \approx 1 - \Delta t$ and the explicit formula (2) for ρ_{post}, we obtain the following time-stepping scheme:

$$\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t}, \tag{13a}$$

$$\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\, \rho_{\text{post}}(\theta)^{\Delta t} \propto \hat{\rho}_{n+1}(\theta)\, e^{-\Delta t \Phi_R(\theta)}. \tag{13b}$$

It is worth mentioning that the first order approximation corrects the bias introduced by the operator splitting, ensuring that ρ_{post} remains the fixed point of this time-stepping scheme. Moreover, the step (13*a*) and equation (10) can be interpreted as the Fisher–Rao gradient flow of the negative entropy term $\mathbb{E}_\rho[\log \rho]$, which tends to increase entropy by expanding the distribution to *explore* the state space. In contrast, equation (13*b*) multiplies the current distribution by the 'posterior function' to *exploit* the data and prior information, concentrating towards regions of high posterior density. It is worth noting that the exploration-exploitation concept distinguishes the present approach from SMC [4] and other homotopy based approaches for sampling [83], which instead rely on the updating rule

$$\rho_{n+1}(\theta) \propto \rho_n(\theta)\, e^{-\Delta t \Phi(\theta)}$$

to deform the prior into the posterior in one unit time. The iteration equation (13) is first proposed as the basis for sampling algorithms in [43] as a methodology to remedy the ensemble collapse of EKI in long time asymptotics; however, the connection to gradient flows is not pointed out. We note that the iteration equation (13) also connects to the tempering (or annealing) approaches that are commonly used in the Monte Carlo literature [84–86]. Finally equation (13) can also be interpreted as an entropic mirror descent algorithm in optimization [87].

The exploration-exploitation time-stepping scheme equation (13) inherits the convergence property of the continuous flow; see proposition 2. The proof can be found in appendix A.

Proposition 2. *Under the assumptions in proposition 1, let ρ_n solve equation (13), then for any $n \geqslant |\frac{\log((1+B)K)}{\log(1-\Delta t)}|$, it holds that*

$$\mathrm{KL}[\rho_n \| \rho_{post}] \leqslant (2 + B + eB) K (1 - \Delta t)^n. \tag{14}$$

3. Gaussian approximation and Kalman methodology

In this section, we discuss the Gaussian approximation of the scheme equation (13) through the Kalman methodology. In doing so we review the necessary techniques and pave the way for constructing our Gaussian mixture approximations in the next section.

In [43], the authors used Gaussian distributions to approximate the evolution of densities defined by equation (13). More precisely, assume $\rho_n = \mathcal{N}(m_n, C_n)$. Then, the first exploration step equation (13*a*) leads to

$$\hat{\rho}_{n+1} = \mathcal{N}\left(\widehat{m}_{n+1}, \widehat{C}_{n+1}\right) = \mathcal{N}\left(m_n, \frac{1}{1-\Delta t} C_n\right). \tag{15}$$

The distribution still remains Gaussian. However, the second exploitation step equation (13*b*) will map out of the space of Gaussian densities, unless Φ_R is quadratic. In [43], the Kalman methodology is employed to approximate equation (13*b*), which is similar to the analysis step in the Kalman filter. More precisely, the methodology starts with the following artificial inverse problem:

$$x = \mathcal{F}(\theta) + \nu, \tag{16}$$

where we have

$$x = \begin{bmatrix} y \\ r_0 \end{bmatrix} \quad \mathcal{F}(\theta) = \begin{bmatrix} \mathcal{G}(\theta) \\ \theta \end{bmatrix} \quad \Sigma_\nu = \begin{bmatrix} \Sigma_\eta & 0 \\ 0 & \Sigma_0 \end{bmatrix}. \tag{17}$$

Here, we set the prior on θ as $\hat{\rho}_{n+1}(\theta)$, and the observation noise $\nu \sim \mathcal{N}(0, \frac{1}{\Delta t}\Sigma_\nu)$. Following Bayes rule, the posterior distribution of the artificial inverse problem is

$$\rho(\theta|x) = \frac{\hat{\rho}_{n+1}(\theta)\,\rho(x|\theta)}{\rho(x)} \propto \hat{\rho}_{n+1}(\theta)\, e^{-\Delta t \Phi_R(\theta)} = \rho_{n+1}(\theta), \tag{18}$$

which matches the output of the step equation (13*b*). Here we used the fact that equation (2) can be rewritten as

$$\Phi_R(\theta) = \frac{1}{2}\|\Sigma_\nu^{-\frac{1}{2}}(x - \mathcal{F}(\theta))\|^2. \tag{19}$$

The Kalman methodology for approximating the posterior $\rho(\theta|x)$ may now be adopted. One first forms a Gaussian approximation of the joint distribution of θ and $\mathcal{F}(\theta)+\nu$, via standard moment matching, yielding

$$\rho^{\mathrm{G}}(\theta, \mathcal{F}(\theta)+\nu) \sim \mathcal{N}\left(\begin{bmatrix} \theta \\ x \end{bmatrix}; \begin{bmatrix} \widehat{m}_{n+1} \\ \widehat{x}_{n+1} \end{bmatrix}, \begin{bmatrix} \widehat{C}_{n+1} & \widehat{C}^{\theta x}_{n+1} \\ \widehat{C}^{\theta x^T}_{n+1} & \widehat{C}^{xx}_{n+1} \end{bmatrix}\right), \tag{20}$$

where $\widehat{m}_{n+1}, \widehat{C}_{n+1}$ are as specified previously, and

$$\begin{aligned} &\widehat{x}_{n+1} = \mathbb{E}[\mathcal{F}(\theta)], \quad \widehat{C}^{\theta x}_{n+1} = \mathrm{Cov}[\theta, \mathcal{F}(\theta)], \\ &\widehat{C}^{xx}_{n+1} = \mathrm{Cov}[\mathcal{F}(\theta)+\nu] = \mathrm{Cov}[\mathcal{F}(\theta)] + \frac{1}{\Delta t}\Sigma_\nu. \end{aligned} \tag{21}$$

In the above, the expectation and covariance are taken over $\theta \sim \hat{\rho}_{n+1}$. These integrals can be computed using Monte Carlo methods or quadrature rules.

Then, one can condition the joint Gaussian distribution equation (20) on the event $\mathcal{F}(\theta)+\nu = x$, to get a Gaussian approximation of the posterior $\rho(\theta|x)$. In detail, using the Gaussian conditioning formula, one obtains

$$\rho_{n+1}(\theta) \approx \rho^{\mathrm{G}}(\theta|\mathcal{F}(\theta)+\nu = x) = \mathcal{N}(\theta; m_{n+1}, C_{n+1}), \tag{22a}$$

where

$$m_{n+1} = \widehat{m}_{n+1} + \widehat{C}^{\theta x}_{n+1}\left(\widehat{C}^{xx}_{n+1}\right)^{-1}(x - \widehat{x}_{n+1}), \tag{23a}$$

$$C_{n+1} = \widehat{C}_{n+1} - \widehat{C}^{\theta x}_{n+1}\left(\widehat{C}^{xx}_{n+1}\right)^{-1}\left(\widehat{C}^{\theta x}_{n+1}\right)^T. \tag{23b}$$

Combining the two updates equations (15) and (23) leads to a Gaussian approximation scheme for solving the discrete Fisher–Rao gradient flow equation (13). The scheme is based

on the Kalman methodology and is derivative free. Some theoretical and numerical studies of this scheme can be found in [43].

Remark 1. We can connect the Gaussian approximation based on the Kalman methodology and the approximation equation (4) obtained by Gaussian variational inference. To do so we consider the continuous time limit of equation (23), calculated in [43, equation (A.2)], as

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = \widehat{C}_t^{\theta x}\Sigma_\nu^{-1}\left(x-\widehat{x}_t\right), \qquad \frac{\mathrm{d}C_t}{\mathrm{d}t} = C_t - \widehat{C}_t^{\theta x}\Sigma_\nu^{-1}\left(\widehat{C}_t^{\theta x}\right)^T, \tag{24}$$

where $\widehat{x}_t = \mathbb{E}_{\rho_t}[\mathcal{F}(\theta)]$, $\widehat{C}_t^{\theta x} = \mathbb{E}_{\rho_t}\left[\left(\theta - m\right)\otimes\left(\mathcal{F}(\theta) - \mathbb{E}\mathcal{F}(\theta)\right)\right]$ and the expectation is taken with respect to the distribution $\rho_t(\theta) = \mathcal{N}(\theta; m_t, C_t)$. We can view equation (24) as a derivative-free approximation of equation (4) through the *statistical linearization* [42, section 4.3.2] approach. More precisely, by Stein's identity which utilizes the integration by parts formula for Gaussian measures, we have the relation $\mathbb{E}_{\rho_t}[\nabla_\theta\mathcal{F}(\theta)] = C_t^{-1}\widehat{C}_t^{\theta x}$. Statistical linearization makes the approximation $\nabla_\theta\mathcal{F}(\theta) \approx C_t^{-1}\widehat{C}_t^{\theta x}$ for all θ; the approximation is exact when $\mathcal{F}$ is linear. Based on it, we can approximate the right hand side in the equation of the mean in equation (4) as follows:

$$C_t\mathbb{E}_{\rho_t}\left[\nabla_\theta \log\rho_{\mathrm{post}}\right] = C_t\mathbb{E}_{\rho_t}\left[\nabla_\theta\mathcal{F}\left(\theta\right)\Sigma_\nu^{-1}\left(x-\mathcal{F}\left(\theta\right)\right)\right] \approx \widehat{C}_t^{\theta x}\Sigma_\nu^{-1}\left(x-\widehat{x}_t\right), \tag{25}$$

where in the first identity we used the fact $\rho_{\mathrm{post}} \propto \exp(-\frac{1}{2}\|\Sigma_\nu^{-1/2}(\mathcal{F}(\theta)-x)\|^2)$, and we used the statistical linearization approximation in the last derivation.

The stochastic linearization essentially approximates $\nabla_\theta\mathcal{F}(\theta)$ by a constant vector (which is why the term *linearization* is used); under this approximation, the Hessian $\nabla_\theta\nabla_\theta\mathcal{F}(\theta)$ is zero[6]. Based on this fact, we obtain, for the equation of the covariance in equation (4), that

$$\begin{aligned} C_t - C_t\mathbb{E}_{\rho_t}\left[\nabla_\theta\nabla_\theta\log\rho_{\mathrm{post}}\right]C_t &\approx C_t - C_t\mathbb{E}_{\rho_t}\left[\nabla_\theta\mathcal{F}\left(\theta\right)\Sigma_\nu^{-1}\nabla_\theta\mathcal{F}\left(\theta\right)^T\right]C_t \\ &\approx C_t - \widehat{C}_t^{\theta x}\Sigma_\nu^{-1}\left(\widehat{C}_t^{\theta x}\right)^T. \end{aligned} \tag{26}$$

Thus, we recover equation (24). In this sense, we can understand the Kalman methodology in the continuous limit as applying a statistical linearization approximation to the dynamics obtained in variational inference.

4. GMKI

In this section, we study the use of Gaussian mixture models to approximate the evolution $\rho_n \mapsto \rho_{n+1}$ defined by equation (13). We assume that at step n the distribution has the form of a K-component Gaussian mixture:

$$\rho_n\left(\theta\right) = \sum_{k=1}^{K} w_{n,k}\mathcal{N}\left(\theta; m_{n,k}, C_{n,k}\right),$$

where $w_{n,k} \geqslant 0$, $\sum_{k=1}^{K} w_{n,k} = 1$. It remains to specify updates of the weights and Gaussian components, through the map defined by equation (13). In the two subsequent subsections, we consider steps (13*a*) and (13*b*) respectively.

[6] We can also interpret this step through the Gauss–Newton approximation.

4.1. The exploration step

The first step equation (13*a*), $\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t}$, is not closed in the space of K-component Gaussian mixtures. Thus, we choose to approximate $\hat{\rho}_{n+1}$ by a new Gaussian mixture model

$$\hat{\rho}_{n+1} \approx \hat{\rho}_{n+1}^{\mathrm{GM}} = \sum_{k=1}^{K} \hat{w}_{n+1,k} \mathcal{N}\left(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k}\right).$$

Note that $\rho_n(\theta)^{1-\Delta t} = \rho_n(\theta)^{-\Delta t}\rho_n(\theta)$. We will determine the parameters of the new Gaussian mixture model by applying Gaussian moment matching[7] to each component $\rho_n(\theta)^{-\Delta t} w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})$ in $\rho_n(\theta)^{-\Delta t}\rho_n(\theta)$. More specifically, we first rewrite the power of a Gaussian mixture as follows:

$$\begin{aligned}
\hat{\rho}_{n+1}(\theta) &\propto \left(\sum_{k=1}^{K} w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})\right)^{1-\Delta t} \\
&= \sum_{k=1}^{K}\left[w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})\left(\sum_{i=1}^{K} w_{n,i}\mathcal{N}(\theta; m_{n,i}, C_{n,i})\right)^{-\Delta t}\right] \\
&= \sum_{k=1}^{K}\left[f_{n,k}(\theta)\mathcal{N}\left(\theta; m_{n,k}, \frac{C_{n,k}}{1-\Delta t}\right)\right],
\end{aligned}$$

where

$$f_{n,k}(\theta) = \frac{(2\pi)^{\frac{\Delta t N_\theta}{2}}}{(1-\Delta t)^{\frac{N_\theta}{2}}} w_{n,k}^{1-\Delta t} \det(C_{n,k})^{\frac{\Delta t}{2}} \left(\frac{w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})}{\sum_{i=1}^{K} w_{n,i}\mathcal{N}(\theta; m_{n,i}, C_{n,i})}\right)^{\Delta t}.$$

We approximate each component above by a Gaussian distribution:

$$f_{n,k}(\theta)\mathcal{N}\left(\theta; m_{n,k}, \frac{C_{n,k}}{1-\Delta t}\right) \approx \hat{w}_{n+1,k}\mathcal{N}\left(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k}\right), \tag{27}$$

where we set

$$\hat{w}_{n+1,k} = \int f_{n,k}(\theta)\mathcal{N}\left(\theta; m_{n,k}, \frac{C_{n,k}}{1-\Delta t}\right)\mathrm{d}\theta, \tag{28a}$$

$$\widehat{m}_{n+1,k} = \frac{1}{\hat{w}_{n+1,k}}\int \theta f_{n,k}(\theta)\mathcal{N}\left(\theta; m_{n,k}, \frac{C_{n,k}}{1-\Delta t}\right)\mathrm{d}\theta, \tag{28b}$$

$$\widehat{C}_{n+1,k} = \frac{1}{\hat{w}_{n+1,k}}\int (\theta - \widehat{m}_{n+1,k})(\theta - \widehat{m}_{n+1,k})^T f_{n,k}(\theta)\mathcal{N}\left(\theta; m_{n,k}, \frac{C_{n,k}}{1-\Delta t}\right)\mathrm{d}\theta. \tag{28c}$$

Here, we determine $\hat{w}_{n+1,k}$, $\widehat{m}_{n+1,k}$, and $\widehat{C}_{n+1,k}$ by moment matching of both sides in equation (27). These integrals can be evaluated by Monte Carlo method or quadrature rules. Then, we normalize $\{\hat{w}_{n+1,k}\}_{k=1}^{k=K}$ so that their summation is 1. The updates (28) determine $\hat{\rho}_{n+1}^{\mathrm{GM}}$.

[7] Such approximation along with our exploitation step will connect our GMKI to Gaussian mixture variational inference (GMVI) and natural gradient; see remark 2.

4.2. The exploitation step

The second step equation (13*b*) leads to

$$\begin{aligned}\rho_{n+1}(\theta) &\propto \hat{\rho}_{n+1}^{\mathrm{GM}}(\theta)\, e^{-\Delta t \Phi_R(\theta)} \\ &\propto \sum_{k=1}^{K} \hat{w}_{n+1,k} \mathcal{N}\left(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k}\right) e^{-\Delta t \Phi_R(\theta)}.\end{aligned}$$

Now, the goal is to approximate the above ρ_{n+1} by a K-component Gaussian mixture $\sum_{k=1}^{K} w_{n+1,k}\mathcal{N}(\theta; m_{n+1,k}, C_{n+1,k})$. We adopt the Kalman methodology described in section 3, to update each Gaussian component individually such that

$$\hat{w}_{n+1,k} \mathcal{N}\left(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k}\right) e^{-\Delta t \Phi_R(\theta)} \approx w_{n+1,k} \mathcal{N}\left(\theta; m_{n+1,k}, C_{n+1,k}\right). \tag{29}$$

More precisely, following (23), for each $1 \leqslant k \leqslant K$, we obtain the mean and covariance updates as

$$\begin{aligned} m_{n+1,k} &= \widehat{m}_{n+1,k} + \widehat{C}^{\theta x}_{n+1,k}\left(\widehat{C}^{xx}_{n+1,k}\right)^{-1}\left(x - \widehat{x}_{n+1,k}\right), \\ C_{n+1,k} &= \widehat{C}_{n+1,k} - \widehat{C}^{\theta x}_{n+1,k}\left(\widehat{C}^{xx}_{n+1,k}\right)^{-1}\left(\widehat{C}^{\theta x}_{n+1,k}\right)^{T}, \end{aligned} \tag{30}$$

where

$$\widehat{x}_{n+1,k} = \mathbb{E}\left[\mathcal{F}(\theta)\right] \qquad \widehat{C}^{\theta x}_{n+1,k} = \mathrm{Cov}\left[\theta, \mathcal{F}(\theta)\right] \qquad \widehat{C}^{xx}_{n+1,k} = \mathrm{Cov}\left[\mathcal{F}(\theta)\right] + \frac{1}{\Delta t}\Sigma_\nu,$$

with $\theta \sim \mathcal{N}(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k})$. The weight $w_{n+1,k}$ is estimated by matching equation (29) via integration

$$w_{n+1,k} = \hat{w}_{n+1,k} \int \mathcal{N}\left(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k}\right) e^{-\Delta t \Phi_R(\theta)} \mathrm{d}\theta. \tag{31}$$

Equations (28), (30) and (31) define our GMKI algorithm, which leads to an iteration of Gaussian mixture approximations of equation (13) without using derivatives. As mentioned in section 3, the updates involve Gaussian integration equation (28), (30), (31), which can be approximated via Monte Carlo or quadrature rules. In this paper, we use the Monte Carlo method to approximate equation (28); these integrations do not require the forward evaluations so are inexpensive. Furthermore, we use the modified unscented transform detailed in [43, definition 1] to approximate equations (30) and (31), which requires $(2N_\theta + 1)K$ forward evaluations; these evaluations can be computed in parallel. The detailed algorithm is presented in appendix B.

5. Theoretical analysis

In this section, theoretical studies of our GMKI methodology are presented, through a continuous time analysis. In section 5.1, we discuss the exploration effect of the first step equation (28) of our GMKI. In section 5.2, we investigate the convergence properties of our GMKI method in scenarios where the posterior follows a Gaussian distribution, as well as in cases where it corresponds to a Gaussian mixture with well-separated Gaussian components. The analysis of the continuous time limit also allows us to connect GMKI with other variational inference approaches based on Gaussian mixtures. A schematic of properties of GMKI is also shown in

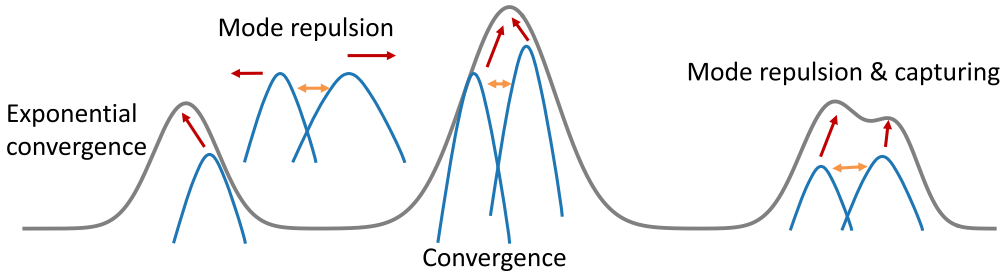

Figure 1. Schematic of properties of GMKI. The Grey curve represents the posterior distribution. Blue curves represent Gaussian components of the Gaussian mixture approximation. From left to right: Gaussian components can exhibit exponential convergence toward their respective Gaussian modes if these modes are well separated (see proposition 7); the repulsion between distinct Gaussian components in the iteration of GMKI helps explore the space and capture multiple modes (see section 5.1); when multiple Gaussian components converge towards a single Gaussian mode in the posterior distribution, they can provide a good approximation of the Gaussian mode (see proposition 6); GMKI can capture multiple modes even when these modes are intertwined (see numerical examples in section 6.1).

figure 1. In section 5.3, we discuss the affine invariance property of our GMKI. We summarize the conclusions of the theoretical studies in section 5.4.

5.1. Exploration effect

In the derivation of GMKI, the first step equation (28) is designed to approximate the exploration phase of the Fisher–Rao gradient flow equation (13*a*). In this subsection, we investigate whether the exploration effect still persists with the approximation made by GMKI. In fact, equation (28) tends to expand the distribution for exploration in the following two ways: (1) repulsion between Gaussian components; and (2) by an increase of the entropy. The repulsion effect can be understood through the following continuous time limit analysis. In this analysis we abuse notation, replacing the subscript n in the discrete iterations by t, which equals $n\Delta t$ in the continuous limit of the means, covariances and weights of the the Gaussian mixture.

Proposition 3. *The continuous time limit* $(\Delta t \to 0)$ *of the exploration step equation (28) is*

$$\dot{m}_{t,k} = -\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})(\theta - m_{t,k}) \log \rho_t(\theta)\, \mathrm{d}\theta, \tag{32a}$$

$$\dot{C}_{t,k} = -\int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \left((\theta - m_{t,k})(\theta - m_{t,k})^T - C_{t,k}\right) \log \rho_t(\theta)\, \mathrm{d}\theta, \tag{32b}$$

$$\dot{w}_{t,k} = -w_{t,k} \int [\mathcal{N}(\theta; m_{t,k}, C_{t,k}) - \rho_t(\theta)] \log \rho_t(\theta)\, \mathrm{d}\theta. \tag{32c}$$

Here $\rho_t(\theta) = \sum_{k=1}^K w_{t,k} \mathcal{N}(\theta; m_{t,k}, C_{t,k})$.

The proof is in C.1. The continuous time limit of the evolution equation of the mean equation (32*a*) suggests that, if $K \geqslant 2$, $m_{t,k}$ will move towards the direction where ρ_t is small.

Indeed, let us consider a specific scenario where $N_\theta = 1$, $K = 2$ and $m_{t,1} < m_{t,2}$, In this case, we have that

$$\begin{aligned}\dot{m}_{t,1} &= -\int (\theta - m_{t,1})\mathcal{N}(\theta; m_{t,1}, C_{t,1}) \log \rho_t(\theta)\, \mathrm{d}\theta \\ &= -\int_{\theta > m_{t,1} \cup \theta < m_{t,1}} (\theta - m_{t,1})\mathcal{N}(\theta; m_{t,1}, C_{t,1}) \log \rho_t(\theta)\, \mathrm{d}\theta \\ &= -\int_{\theta > m_{t,1}} (\theta - m_{t,1})\mathcal{N}(\theta; m_{t,1}, C_{t,1}) \left(\log \rho_t(\theta) - \log \rho_t(2m_{t,1} - \theta)\right) \mathrm{d}\theta \\ &< 0,\end{aligned} \tag{33}$$

where the third equality results from the change of variable $\theta \to 2m_{t,1} - \theta$. And the last inequality is due to the fact that

$$\log \rho_t(\theta) - \log \rho_t(2m_{t,1} - \theta) = \log \frac{w_{t,1}\mathcal{N}(\theta; m_{t,1}, C_{t,1}) + w_{t,2}\mathcal{N}(\theta; m_{t,2}, C_{t,2})}{w_{t,1}\mathcal{N}(\theta; m_{t,1}, C_{t,1}) + w_{t,2}\mathcal{N}(2m_{t,1} - \theta; m_{t,2}, C_{t,2})}.$$

The right hand side is non-negative, since when $\theta > m_{t,1}$, we have $|m_{t,2} - 2m_{t,1} + \theta| = (m_{t,2} - m_{t,1}) + (\theta - m_{t,1}) > |m_{t,2} - \theta|$ and hence $\mathcal{N}(2m_{t,1} - \theta; m_{t,2}, C_{t,2}) < \mathcal{N}(\theta; m_{t,2}, C_{t,2})$. Similarly, we can also establish that $\dot{m}_{t,2} > 0$. Hence these two Gaussian means are repulsed.

We can also understand the exploration effect through the increase of the entropy; see proposition 4. The proof can be found in C.2.

Proposition 4. *The entropy of the Gaussian mixture*

$$\rho_t(\theta) = \sum_{k=1}^{K} w_{t,k}\mathcal{N}(\theta; m_{t,k}, C_{t,k})$$

obtained from equation (32) is non-decreasing; indeed:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int -\rho_t \log \rho_t \mathrm{d}\theta = \sum_{k=1}^{K}\left(\frac{\dot{w}_{t,k}^2}{w_{t,k}} + w_{t,k}\dot{m}_{t,k}^T C_{t,k}^{-1}\dot{m}_{t,k} + \frac{w_{t,k}}{2} tr\left[\dot{C}_{t,k}^T C_{t,k}^{-1}\dot{C}_{t,k}C_{t,k}^{-1}\right]\right) \geqslant 0. \tag{34}$$

5.2. Convergence analysis

To provide insights for the convergence of GMKI, we consider its continuous limit in time. Similar to equation (32), the continuous time limit of our GMKI is given in proposition 5. The proof is in C.3.

Proposition 5. *The continuous time limit* $(\Delta t \to 0)$ *of the proposed GMKI defines the evolving Gaussian mixture measure*

$$\rho_t(\theta) = \sum_{k=1}^{K} w_{t,k}\mathcal{N}(\theta; m_{t,k}, C_{t,k})$$

where

$$\dot{m}_{t,k} = -C_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\nabla_\theta \log \rho_t(\theta)\, \mathrm{d}\theta + \widehat{C}_{t,k}^{\theta x}\Sigma_\nu^{-1}(x - \widehat{x}_{t,k}), \tag{35a}$$

$$\dot{C}_{t,k} = -C_{t,k}\left(\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\nabla_\theta \nabla_\theta \log \rho_t(\theta)\, \mathrm{d}\theta\right) C_{t,k} - \widehat{C}_{t,k}^{\theta x}\Sigma_\nu^{-1}\widehat{C}_{t,k}^T, \tag{35b}$$

$$\dot{w}_{t,k} = -w_{t,k}\int \left[\mathcal{N}(\theta; m_{t,k}, C_{t,k}) - \rho_t(\theta)\right]\left[\log \rho_t(\theta) + \Phi_R(\theta)\right] \mathrm{d}\theta. \tag{35c}$$

Here

$$\widehat{x}_{t,k} = \mathbb{E}\left[\mathcal{F}(\theta)\right], \qquad \widehat{C}^{\theta x}_{t,k} = \mathrm{Cov}\left[\theta, \mathcal{F}(\theta)\right], \quad \textit{with} \quad \theta \sim \mathcal{N}(m_{t,k}, C_{t,k}). \tag{36}$$

Remark 2. We can also connect the continuous limit in proposition 5 with the classical variational inference approach based on Gaussian mixtures. In fact, equation (35) can be obtained by combining natural gradient methods in the Gaussian mixture context and derivative-free Kalman approximation (similar to remark 1); see appendix C.8.

To gain insight into the convergence properties of the continuous flow, we first study equation (35) for the Gaussian posterior case; the proof can be found in C.4.

Proposition 6 (linear inverse problems). *Assume $\mathcal{G}(\theta) = G \cdot \theta$ is linear, and the posterior is Gaussian with the form*

$$\Phi_R(\theta) = \frac{1}{2}(\theta - m_{\mathrm{post}})^T C_{\mathrm{post}}^{-1}(\theta - m_{\mathrm{post}}).$$

Then, the KL divergence between the Gaussian mixture

$$\rho_t(\theta) = \sum_{k=1}^{K} w_{t,k}\mathcal{N}(\theta; m_{t,k}, C_{t,k})$$

obtained from equation (35) and ρ_{post} is non-increasing:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left[\rho_t \| \rho_{\mathrm{post}}\right] = -\sum_{k=1}^{K}\left(\frac{\dot{w}_{t,k}^2}{w_{t,k}} + w_{t,k}\dot{m}_{t,k}^T C_{\mathrm{post}}^{-1}\dot{m}_{t,k} + \frac{w_{t,k}}{2}tr\left[\dot{C}_{t,k}C_{t,k}^{-1}\dot{C}_{t,k}C_{t,k}^{-1}\right]\right) \leqslant 0. \tag{37}$$

Furthermore, the mean and Fisher information matrix of stationary points $\rho_\infty(\theta) = \sum_k w_{\infty,k}\mathcal{N}(\theta; m_{\infty,k}, C_{\infty,k})$ satisfy that

$$\sum_{k=1}^{K} w_{\infty,k}m_{\infty,k} = m_{\mathrm{post}}, \qquad FIM\left[\rho_\infty\right] = \int \frac{\nabla_\theta \rho_\infty \nabla_\theta \rho_\infty^T}{\rho_\infty}\mathrm{d}\theta = C_{\mathrm{post}}^{-1}. \tag{38}$$

Remark 3. Proposition 6 shows that, if the posterior is Gaussian, then the KL divergence of the GMKI is non-increasing in time. Furthermore, the Gaussian mixture converges to a distribution ρ_∞ from which the correct Gaussian statistics can be extracted. Nevertheless, from our current proof, it is not yet known whether ρ_∞ converge to ρ_{post}. We leave this question for future study.

Finally, we provide some formal analysis for the convergence of our GMKI in scenarios where the posterior distribution is close to Gaussian mixture with the same number of components, namely

$$\rho_{\mathrm{post}}(\theta) \propto \exp\left(-\Phi_R(\theta)\right) \underset{\sim}{\propto} \sum_{k=1}^{K} w_k^* \mathcal{N}(\theta; m_k^*, C_k^*) \quad \text{with} \inf_{1\leqslant k\leqslant K} w_k^* > 0.$$

For simplicity, we assume these Gaussian components are well separated. It is technical to give a precise definition of the well-separatedness of different Gaussian components; our argument here is purely formal and serves to provide insights for the behavior of GMKI. Suppose the kth Gaussian component $\mathcal{N}(\theta; m_k(0), C_k(0))$ in GMKI is close to its corresponding mode (e.g. the kth mode) of ρ_{post} while becoming well separated from other Gaussian components. In such case, we may simplify the continuous time limit equation (35) by neglecting the interaction between different Gaussian components. The simplified continuous time dynamics and its property are presented in proposition 7, with derivations in appendix C.5.

Proposition 7. *Consider the simplified continuous time dynamics*

$$\dot{m}_{t,k} = C_{t,k}\left(C_k^*\right)^{-1}\left(m_k^* - m_{t,k}\right), \tag{39a}$$

$$\dot{C}_{t,k} = C_{t,k} - C_{t,k}\left(C_k^*\right)^{-1} C_{t,k}, \tag{39b}$$

$$\dot{w}_{t,k} = w_{t,k}\left(\log w_k^* - \log w_{t,k} - \sum_{i=1}^{K} w_{t,i}\left(\log w_i^* - \log w_{t,i}\right)\right). \tag{39c}$$

Then, $m_{t,k}$, $C_{t,k}$, and $w_{t,k}$ will converge to m_k^*, C_k^*, and w_k^* exponentially fast.

The proof of proposition 7 can be found in appendix C.6.

5.3. Affine invariance

Sampling methods are said to be *affine invariant* if the algorithm is unchanged under any invertible affine mapping. It is a consequence of such a property that convergence rates across all Gaussians with positive covariance are the same. More generally affine invariant algorithms can be highly effective for highly anisotropic distributions [88, 89]. This effectiveness stems from their consistent behavior across all coordinate systems related by affine transformations. Specifically, the convergence properties of these methods can be understood by examining the optimal coordinate system, which minimizes anisotropy to the fullest extent, across all affine transformations. With this in mind, the following is of interest in relation to GMKI:

Proposition 8. *The continuous time limit in proposition 5 is affine invariant. Specifically, for any invertible affine mapping $\varphi : \theta \to \widetilde{\theta} = A\theta + b$, and define corresponding scalar, vector, and matrix transformations*

$$\widetilde{w}_{t,k} = w_{t,k} \quad \widetilde{m}_{t,k} = Am_{t,k} + b \quad \widetilde{C}_{t,k} = AC_{t,k}A^T,$$

and function transformations

$$\widetilde{\mathcal{F}}\left(\widetilde{\theta}\right) = \mathcal{F}\left(A^{-1}\left(\widetilde{\theta} - b\right)\right) \qquad \widetilde{\Phi}_R\left(\widetilde{\theta}\right) = \Phi_R\left(A^{-1}\left(\widetilde{\theta} - b\right)\right).$$

The evolution equations of $\widetilde{w}_{t,k}$, $\widetilde{m}_{t,k}$ and $\widetilde{C}_{t,k}$ remain the same, retaining the structure of the GMKI evolutions in equation (35).

The proof of proposition 8 can be found in appendix C.7.

5.4. Summary of theoretical analyses

Combining insights from propositions 3, 6 to 8, we anticipate that the repulsion between distinct Gaussian components will enable the proposed methodology GMKI to capture possible modes of the posterior distribution more effectively. Moreover, the algorithm will converge efficiently to modes of the posterior when these modes are well separated. Finally, we note that the affine invariance property of the GMKI shows that it will be effective for the approximation of certain highly anisotropic distributions.

6. Numerical study

In this section, we present numerical studies regarding the proposed GMKI algorithm. We focus on posterior distributions of unknown parameters or fields arising in inverse problems that may exhibit multiple modes. Three types of model problems are considered:

(i) A one-dimensional bimodal problem: we use this problem as a proof-of-concept example. Our result demonstrates that the convergence rate remains unchanged no matter how overlapped the two modes are. This implies the independence of the convergence rate regarding potential barriers.
(ii) A two-dimensional bimodal problem: we use this benchmark problem, introduced in [65, 90], to compare GMKI with other sampling methods such as the Langevin dynamics and birth death process [24]. Our result shows that GMKI is not only more accurate but also more cost-effective for this specific problem.
(iii) A high-dimensional bimodal problem: we consider the inverse problem of recovering the initial velocity field of the Navier–Stokes flow. The problem is designed to have a symmetry which induces two modes in the posterior. We show that GMKI can capture both modes efficiently; this indicates GMKI's potential for addressing multimodal problems in large-scale and high-dimensional applications.

Regarding the parameter of the GMKI algorithm (B), we will specify in detail the number of mixtures K in each problem. For all experiments, we use the time-step size $\Delta t = 0.5$ and adopt $J = 1000$ points for the Monte Carlo estimation of equation (28).

6.1. One-dimensional bimodal problem

We first consider the following 1D bimodal inverse problem, associated with a forward model

$$y = \mathcal{G}(\theta) + \eta \quad \text{with} \quad y = 1, \quad \mathcal{G}(\theta) = \theta^2.$$

We assume the prior is $\rho_{\text{prior}} \sim \mathcal{N}(3, 2^2)$ and consider different noise levels:

$$\begin{aligned} &\text{Case A:} \quad \eta \sim \mathcal{N}\left(0, 0.2^2\right); \\ &\text{Case B:} \quad \eta \sim \mathcal{N}\left(0, 0.5^2\right); \\ &\text{Case C:} \quad \eta \sim \mathcal{N}\left(0, 1.0^2\right); \\ &\text{Case D:} \quad \eta \sim \mathcal{N}\left(0, 1.5^2\right). \end{aligned}$$

Note that the overlap between these two modes is larger when the noise strength is larger. For case A, the two modes are well separated, and for case D, the two modes are nearly mingled.

We apply GMKI with $K = 1$, 2, and 3 modes, which are randomly initialized according to the prior distribution; we assign them equal weights. In each iteration, the GMKI algorithm requires 3, 6, and 9 forward evaluations, respectively. The reference posterior distribution is obtained by evaluating the unnormalized posterior on a uniform grid and then normalizing it.

The results for different cases are reported from figure 2. Each row first shows the reference posterior and posteriors approximated by GMKI at the 30th iteration, using different mode numbers $K = 1$, 2, 3 from left to right. And the fourth figure shows the convergence in terms of the total variation (TV) distance. When $K = 1$, we can only capture one mode, and this of course will not be weighted correctly since it will have weight one by construction; when $K = 2$ or 3, we can capture both modes. It is worth mentioning that GMKI converges in fewer than 30 iterations. The convergence behavior appears independent of the potential barrier. For case A, where the two modes are well separated, the approximated posteriors by GMKI with $K = 2$ or 3 match very well with the reference. This observation justifies our formal analysis in proposition 7. For cases B,C,D, we observe that GMKI is capable of capturing both modes, however, some discrepancy will arise in the region where the modes overlap. These discrepancies persist when increasing the mode number in GMKI from $K = 2$ to $K = 3$.

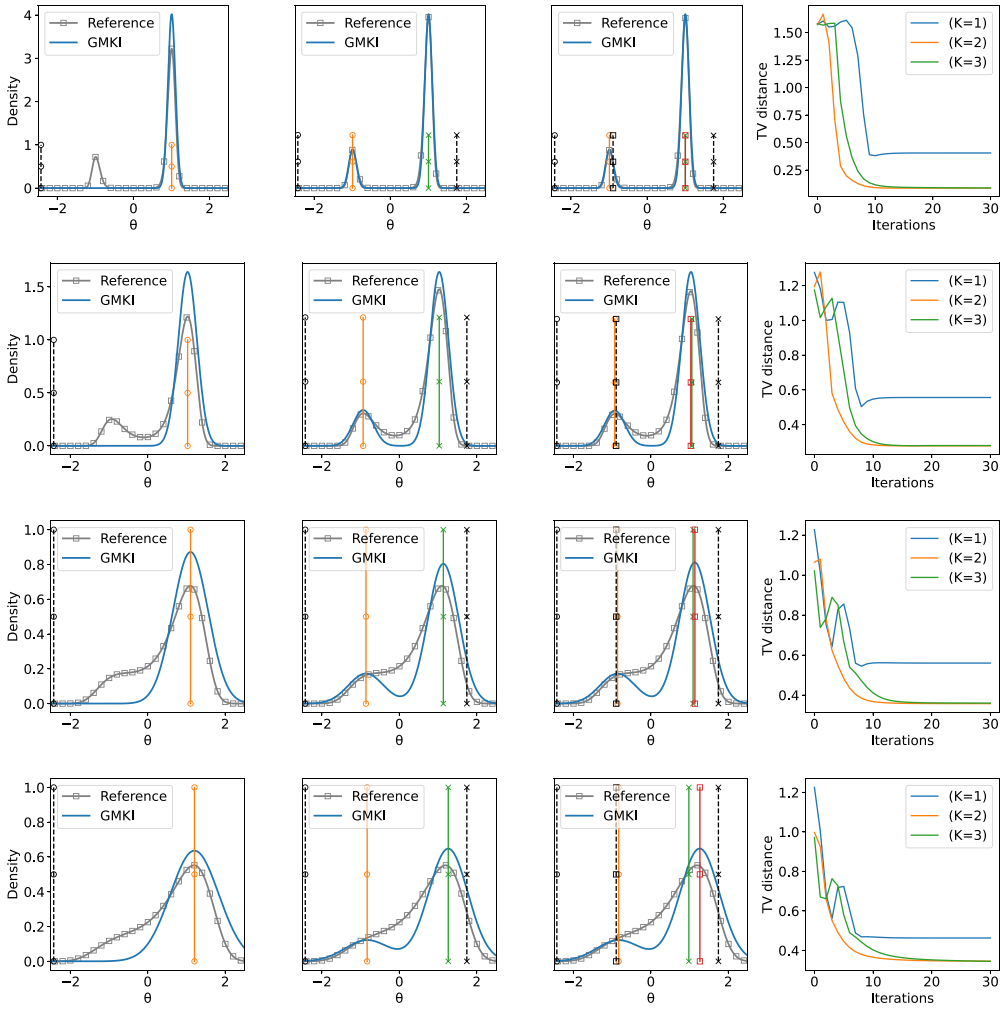

Figure 2. The one-dimensional bimodal problem with Σ_η values of 0.2^2 (top), 0.5^2 (top middle), 1.0^2 (bottom middle), and 1.5^2 (bottom). Each panel displays the reference posterior distribution (grey square lines) and posterior distributions estimated by the GMKI (blue lines) at the 30th iteration with mode number $K = 1, 2, 3$ (from left to right) with mean m_k (colored) and initial mean (black) of each Gaussian component marked. The fourth figure shows the total variation distance between the reference posterior distribution and the posterior distributions estimated by the GMKI with mode number $K = 1, 2, 3$.

6.2. *Two-dimensional bimodal problem*

In this subsection, we consider the 2D bimodal inverse problem from [90, 91], associated with the forward model

$$y = \mathcal{G}(\theta) + \eta \quad \text{with} \quad y = 4.2297, \quad \mathcal{G}(\theta) = \left(\theta_{(1)} - \theta_{(2)}\right)^2.$$

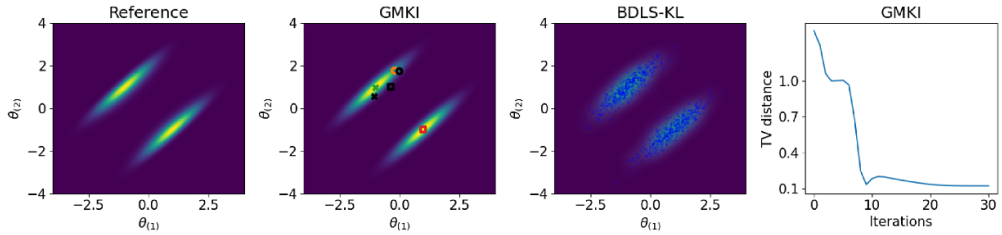

Figure 3. Two-dimensional bimodal problem with $\rho_{\text{prior}} \sim \mathcal{N}(0, I)$. From left to right: reference posterior distribution (left), posterior distributions estimated by 3-modal GMKI (middle left) at the 30th iteration (means m_k (colored) and initial means (black) are marked), BDLS-KL [24] (middle right) at the 1000th iteration, and total variation distance between the reference posterior distribution and the posterior distributions estimated by the GMKI (right).

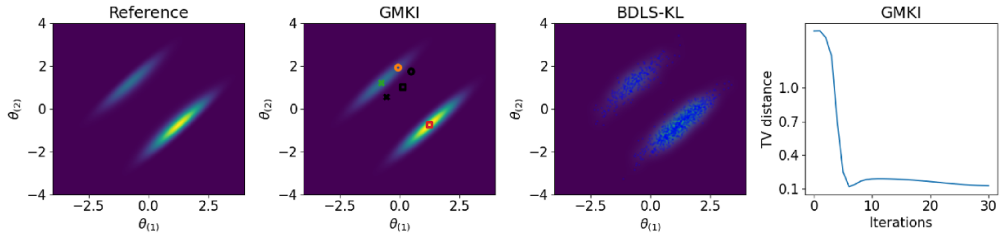

Figure 4. Two-dimensional bimodal problem with $\rho_{\text{prior}} \sim \mathcal{N}([0.5, 0]^T, I)$. From left to right: reference posterior distribution (left), posterior distributions estimated by 3-modal GMKI (middle left) at the 30th iteration (means m_k (colored) and initial means (black) are marked), BDLS-KL [24] (middle right) at the 1000th iteration, and total variation distance between the reference posterior distribution and the posterior distributions estimated by the GMKI (right).

Here $\theta = [\theta_{(1)}, \theta_{(2)}]^T$. We assume the noise distribution is $\eta \sim \mathcal{N}(0, I)$ and consider two different prior distributions:

$$\text{Case A:} \quad \rho_{\text{prior}} \sim \mathcal{N}(0, I);$$
$$\text{Case B:} \quad \rho_{\text{prior}} \sim \mathcal{N}\left([0.5, 0]^T, I\right).$$

For case A, the two modes are symmetric with respect to the line $\theta_{(1)} - \theta_{(2)} = 0$, while for case B, the two modes are not symmetric.

We apply GMKI with $K = 3$ modes, which are randomly initialized based on the prior distribution and we assign these components with equal weights. In each iteration, the algorithm requires $(2N_\theta + 1)K = 15$ forward evaluations. The reference posterior distribution is obtained by evaluating the unnormalized posterior on a uniform grid and then normalizing it. For both case A and case B, the estimated posterior distributions obtained by the GMKI are presented in figures 3 and 4. We observe a strong correspondence with the reference, where in GMKI, mode 1 converges to one target mode, while mode 2 and mode 3 converge to another target mode. Moreover, the evolution of the TV distance indicates rapid convergence.

As a comparison, the derivative-free affine-invariant Langevin dynamics (dfALDI) [64, 65] and derivative-free Bayesian inversion using multiscale dynamics [90] with 10^6 iterations

have been used for sampling the posterior distribution for case A. Their results are reported in [90, figure 5], where dfALDI fails while multiscale dynamics can capture both modes but with wrong weights. The local preconditioner, which involves employing local empirical covariance with distance-dependent weights as introduced in [91], gives rise to an alternative derivative-free affine-invariant Langevin dynamics sampling approach. That approach notably enhances the sampling results in these scenarios. Moreover, for both case A and case B, we apply the BDLS-KL algorithm proposed in [24], which is a gradient-based sampler relying on the birth-death dynamics with kernel density estimators to approximate the Wasserstein–Fisher–Rao gradient flow. For the implementation of BDLS-KL, we use $\Delta t = 10^{-2}$, $T = 10$, ensemble size $J = 10^3$, and the RBF kernel $k(x,x') = \exp(\frac{1}{h}\|x-x'\|^2)$ with the bandwidth $h = \text{med}^2/\log J$ adopted from [92]; here med^2 is the squared median of the pairwise Euclidean distance between the current particles. The results are shown in figures 3 and 4. For both cases, GMKI is not only more cost-effective but also more accurate compared to these existing approaches. Moreover, to study the behavior of GMKI in higher-modal problems, a two-dimensional four-modal problem is presented in appendix D, leading to a similar conclusion.

6.3. *High-dimensional bimodal problem: Navier Stokes problem*

Finally, we study the problem of recovering the initial vorticity field ω_0 of a fluid flow from measurements at later times. The flow is described by the 2D Navier–Stokes equation on a periodic domain $D = [0,2\pi] \times [0,2\pi]$, which can be written in the vorticity-streamfunction $\omega - \psi$ formulation:

$$\begin{aligned} &\frac{\partial \omega}{\partial t} + (v \cdot \nabla)\omega - \nu \Delta \omega = \nabla \times f, \\ &\omega = -\Delta \psi \qquad \frac{1}{4\pi^2}\int \psi = 0 \qquad v = \left[\frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1}\right]^T + v_b. \end{aligned} \tag{40}$$

Here v denotes the velocity vector, $\nu = 0.01$ denotes the viscosity, $v_b = [0,2\pi]^T$ denotes the non-zero mean background velocity, and $f(x) = [0, \cos(4x_{(1)})]^T$ denotes the external forcing.

The problem is built to be spatially symmetric with respect to $x_{(1)} = \pi$. The source of the fluid is chosen such that

$$\nabla \times f\left(\left[x_{(1)}, x_{(2)}\right]^T\right) = -\nabla \times f\left(\left[2\pi - x_{(1)}, x_{(2)}\right]^T\right).$$

The observations in the inverse problem are chosen as the difference of pointwise measurements of the vorticity value $\omega(\cdot)$

$$\omega\left(\left[x_{(1)}, x_{(2)}\right]^T\right) - \omega\left(\left[2\pi - x_{(1)}, x_{(2)}\right]^T\right)$$

at 56 equidistant points in the left domain (see figure 5), at $T = 0.25$ and $T = 0.5$, corrupted with observation error $\eta \sim \mathcal{N}(0, 0.1^2 I)$. Under this set-up, both $\omega_0([x_{(1)}, x_{(2)}]^T)$ and $-\omega_0([2\pi - x_{(1)}, x_{(2)}]^T)$ will lead to the same measurements. Thus the inverse problem will be at least bimodal.

We assume the prior of $\omega_0(x,\theta)$ is a Gaussian field with covariance operator $C = (-\Delta)^{-2}$, subject to periodic boundary conditions, on the space of mean zero functions. The corresponding KL expansion of the initial vorticity field is given by

$$\omega_0(x,\theta) = \sum_{l \in K} \theta^c_{(l)} \sqrt{\lambda_l}\psi^c_l(x) + \theta^s_{(l)}\sqrt{\lambda_l}\psi^s_l(x), \tag{41}$$

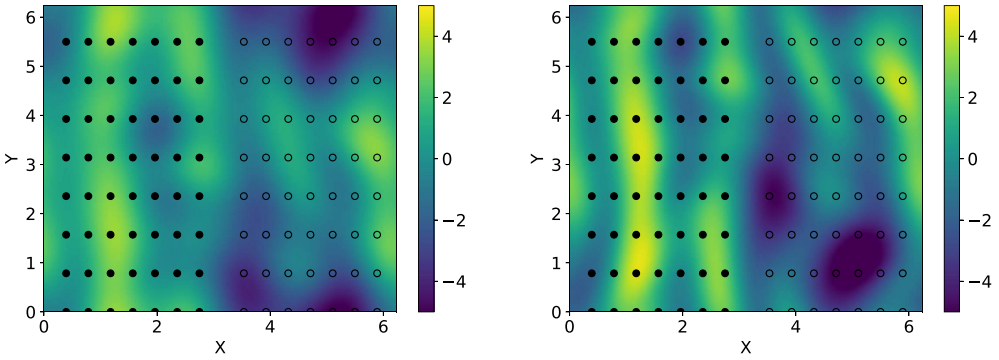

Figure 5. The vorticity field ω at $T=0.25$ and $T=0.5$ and observations $\omega([x_{(1)},x_{(2)}]^T)-\omega([2\pi-x_{(1)},x_{(2)}]^T)$ at 56 equidistant points (solid black dots). Their mirroring points are marked (empty black dots).

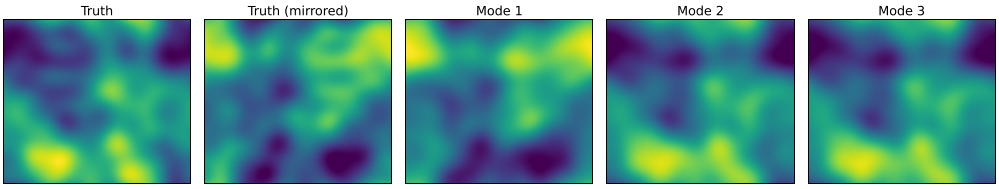

Figure 6. The true initial vorticity field $\omega_0(x;\theta_{\text{ref}})$, and recovered initial vorticity fields $\omega_0(x;m_k)$ obtained by GMKI.

where $\mathbb{L}=\{(l_x,l_y)|l_x+l_y>0 \text{or } (l_x+l_y=0 \text{and } l_x>0)\}$, and the eigenpairs are of the form

$$\psi_l^c(x)=\frac{\cos(l\cdot x)}{\sqrt{2}\pi} \quad \psi_l^s(x)=\frac{\sin(l\cdot x)}{\sqrt{2}\pi} \quad \lambda_l=\frac{1}{|l|^4},$$

and $\theta_{(l)}^c,\theta_{(l)}^s\sim\mathcal{N}(0,2\pi^2)$. The KL expansion equation (41) can be rewritten as a sum over positive integers rather than a lattice:

$$\omega_0(x,\theta)=\sum_{l=1}^{\infty}\theta_{(l)}\sqrt{\lambda_l}\psi_l(x), \tag{42}$$

where the eigenvalues λ_l are in descending order. We truncate the expansion to the first 128 terms and generate the true vorticity field $\omega_0(x;\theta_{\text{ref}})$ with $\theta_{\text{ref}}\in\mathbb{R}^{128}$; we aim to recover the parameter based on observation data.

We employ GMKI with $K=3$ modes, which are randomly initialized based on prior distribution with equal weights. Since we have 3 modes and $N_\theta=128$, in each iteration, we require $(2N_\theta+1)K=771$ forward evaluations. We depict the true initial vorticity field $\omega_0(x;\theta_{\text{ref}})$, its mirrored field (the mirroring of the velocity field induces the antisymmetry in the vorticity field) and the three recovered initial vorticity fields $\omega_0(x;m_k)$ obtained by GMKI at the 50th iteration in figure 6. Mode 1 captures the mirroring field of $\omega_0(x;\theta_{\text{ref}})$ and mode 2 and mode 3 capture $\omega_0(x;\theta_{\text{ref}})$. Figure 7 presents the relative errors of the vorticity field, the optimization errors $\Phi_R(m_{n,k})$, the Frobenius norm $\|C_{n,k}\|_F$ and the Gaussian mixture weights $w_{n,k}$ (from left to right). It shows that our GMKI converges in fewer than 50 iterations. Figure 8 displays the marginals of the estimated posterior distributions associated with the first 16 theta coefficients

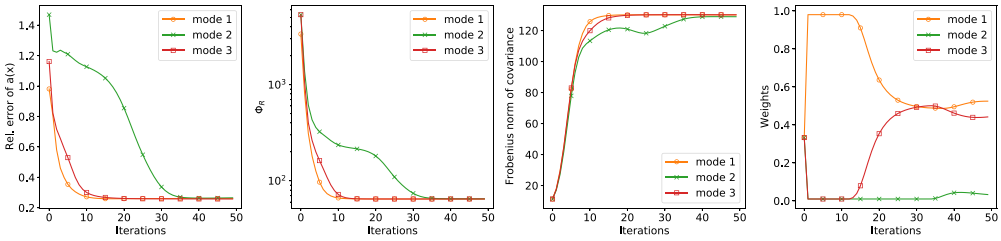

Figure 7. Navier–Stokes flow problem: the relative errors of the initial vorticity field, the optimization errors $\Phi_R(m_{n,k})$, the Frobenius norm $\|C_{n,k}\|_F$, and the Gaussian mixture weights $w_{n,k}$ (from left to right) for different modes.

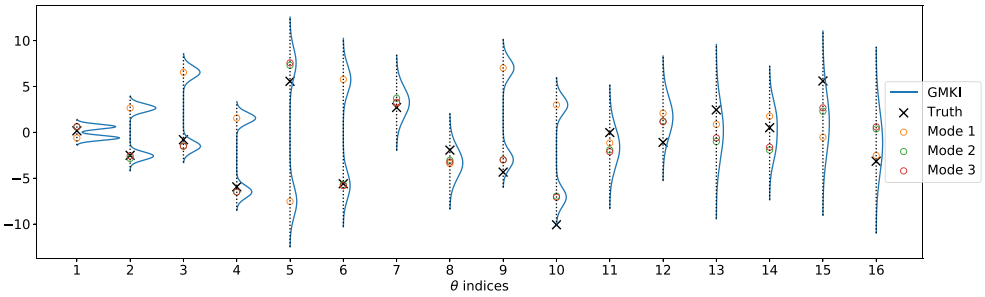

Figure 8. Navier–Stokes flow problem: the true Karhunen–Loeve expansion parameters $\theta_{(i)}$ (black crosses), and mean estimations of $\theta_{(i)}$ for each modes (circles) and the associated marginal distributions obtained GMKI at the 50th iteration.

obtained by GMKI. The marginal distributions exhibit clear bimodality, and the approximate posteriors have a high probability covering the true coefficients.

7. Conclusion

In this paper, we have presented a new framework for solving Bayesian inverse problems. The framework is based on the Fisher–Rao gradient flow. Within this framework, we introduce a novel approach, GMKI, which leverages Gaussian mixtures and Kalman's methodology for numerical approximations of the the flow. GMKI is particularly useful when the posterior distribution has multiple modes and when the derivative of the forward model is not available or computationally expensive.

We derive the continuous time dynamics of the GMKI, showing its connection to GMVI, and studying its exploration effects and convergence properties. Our numerical experiments showcase GMKI's capability in approximating posterior distributions with multiple modes. GMKI outperforms many existing Bayesian inference methods in terms of efficiency and accuracy.

There can be numerous avenues for future research. On the algorithmic side, it is of interest to refine the approximations employed by GMKI in regions where the mixture components overlap significantly; see the experimental results in section 6.1. Moreover, although the Kalman methodology achieves a derivative-free implementation, it may suffer from degeneracy issues when the modes of the distribution concentrate on a low dimensional manifold; for a demonstration see appendix E. Therefore, improving the derivative-free methodology in

such a scenario is important for enhancing GMKI. On the theoretical side, a thorough analysis of the convergence of GMKI for general target distribution could offer valuable insights for its practical application.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/Zhengyu-Huang/InverseProblems.jl/tree/master/Multimodal.

Acknowledgments

Y C is supported by the Courant Instructorship. D Z H is supported by the Fundamental Research Funds for the Central Universities and the high-performance computing platform of Peking University. J H is supported by NSF grant DMS-2331096 and the Sloan research fellowship. D Z H and A M S are supported by NSF award AGS1835860 and by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. A M S is supported by a Department of Defense Vannevar Bush Faculty Fellowship and by the SciAI Center, funded by the Office of Naval Research (ONR), under Grant Number N00014-23-1-2729. S R is supported by Deutsche Forschungsgemeinschaft (DFG)—Project-ID 318763901—SFB1294. We also thank the anonymous reviewers for their helpful suggestions.

Appendix A. Convergence of the exploration-exploitation scheme for the Fisher–Rao gradient flow

Proof of proposition 2. The solution of the Fisher–Rao gradient flow equation (3) has an analytical solution

$$\rho_t(\theta) = \frac{1}{Z_t}\rho_0(\theta)^{e^{-t}}\rho_{\text{post}}(\theta)^{1-e^{-t}}, \quad \frac{\rho_t(\theta)}{\rho_{\text{post}}(\theta)} = \frac{1}{Z_t}\left(\frac{\rho_0(\theta)}{\rho_{\text{post}}(\theta)}\right)^{e^{-t}}. \tag{A.1}$$

where Z_t is the normalization constant. The update equation (13) leads to

$$\rho_n(\theta) = \frac{1}{Z_n}\rho_0(\theta)^{(1-\Delta t)^n}\rho_{\text{post}}(\theta)^{1-(1-\Delta t)^n}, \quad \frac{\rho_n(\theta)}{\rho_{\text{post}}(\theta)} = \frac{1}{Z_n}\left(\frac{\rho_0(\theta)}{\rho_{\text{post}}(\theta)}\right)^{(1-\Delta t)^n}, \tag{A.2}$$

where Z_n is the normalization constant. By comparing equation (A.2) and equation (A.1), we have

$$\rho_n(\theta) = \rho_t(\theta) \quad \text{for} \quad t = -n\log(1-\Delta t).$$

Bringing $t = -n\log(1-\Delta t)$ into equation (9) leads to

$$\text{KL}[\rho_n\|\rho_{\text{post}}] = \text{KL}\left[\rho_{-n\log(1-\Delta t)}\|\rho\right] \leqslant (2+B+eB)K(1-\Delta t)^n, \tag{A.3}$$

when $-n\log(1-\Delta t) \geqslant \log((1+B)K)$. □

Appendix B. Gaussian mixture Kalman inversion algorithm

The detailed algorithm is presented in algorithm 1. There are four hyperparameters: the number of mixtures K, the time-step size Δt, the number of particles J for Monte Carlo integration, and the number of iterations N. Increasing K enhances the expressiveness of the Gaussian mixture model but also increases computational cost. Therefore, we recommend choosing K based on available computational resources and prior knowledge of the number of target modes. Once all modes are captured, there is no need to increase K further. A larger J improves the accuracy of integration approximations, but it also increases computational cost. In the experiments, we adopt $J = 1000$. The time step Δt lies within the range $(0, 1)$. While a larger Δt accelerates convergence, it may also cause numerical instability. However, we found that $\Delta t = 0.5$ strikes a good balance between stability and efficiency. All of these parameters can be chosen adaptively, which we plan to explore in future work. Moreover, in our implementation, we store $\log w$ instead of w, and hence w will never actually reach zero. Additionally, for efficiency, we avoid zero weights by setting a lower bound (e.g. 10^{-10}) for w.

Appendix C. Theoretical studies about GMKI

C.1. Proof of proposition 3

Update equation equation (28) and the normalization of weights $\{\hat{w}_{n+1,k}\}_k$ can be combined and rewritten as

$$\hat{w}_{n+1,k} = \frac{w_{n,k}\int \mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta}{\int \rho_n^{1-\Delta t}\mathrm{d}\theta}, \tag{C.1a}$$

$$\widehat{m}_{n+1,k} = \frac{\int \theta\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta}{\int \mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta}, \tag{C.1b}$$

$$\widehat{C}_{n+1,k} = \frac{\int (\theta - \widehat{m}_{n+1,k})(\theta - \widehat{m}_{n+1,k})^T\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta}{\int \mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta}. \tag{C.1c}$$

Now we derive the continuous time limit of the exploration step; we ignore the hat notation for simplicity.

$$\begin{aligned}\frac{\hat{w}_{n+1,k} - w_{n,k}}{\Delta t} &= w_{n,k}\frac{\int [\mathcal{N}(\theta; m_{n,k}, C_{n,k}) - \rho_n]\,\rho_n^{-\Delta t}\mathrm{d}\theta}{\Delta t\int \rho_n^{1-\Delta t}\mathrm{d}\theta} \\ &= w_{n,k}\frac{\int [\mathcal{N}(\theta; m_{n,k}, C_{n,k}) - \rho_n]\,(1 - \Delta t\log\rho_n)\,\mathrm{d}\theta}{\Delta t} + \mathcal{O}(\Delta t) \\ &= -w_{n,k}\int [\mathcal{N}(\theta; m_{n,k}, C_{n,k}) - \rho_n]\log\rho_n\mathrm{d}\theta + \mathcal{O}(\Delta t),\end{aligned} \tag{C.2}$$

where in the second identity, we used $\rho_n^{-\Delta t} = e^{-\Delta t\log\rho_n} = 1 - \Delta t\log\rho_n + \mathcal{O}(\Delta^2)$. Therefore the continuous limit can be written as

$$\dot{w}_{t,k} = -w_{t,k}\int [\mathcal{N}(\theta; m_{t,k}, C_{t,k}) - \rho_t(\theta)]\log\rho_t(\theta)\,\mathrm{d}\theta.$$

Algorithm 1. Gaussian mixture Kalman inversion.

Input: initial guess $\{w_{0,k}, m_{0,k}, C_{0,k}\}_{k=1}^{K}$, time-step size Δt, number of iterations N, number of particles J for Monte Carlo, forward model $\mathcal{F}$
Output: final solution $\{w_{N,k}, m_{N,k}, C_{N,k}\}_{k=1}^{K}$
for $n \leftarrow 0$ to $N-1$ **do**
  for $k \leftarrow 1$ to K **do**
    Sample $\{\theta^j\}_{j=1}^{J} \sim \mathcal{N}(\theta; m_{n,k}, \frac{C_{n,k}}{1-\Delta t})$, first step update

$$\hat{w}_{n+1,k} = \frac{1}{J}\sum_{j=1}^{J} f_{n,k}\left(\theta^j\right)$$
$$\widehat{m}_{n+1,k} = \frac{1}{\hat{w}_{n+1,k}J}\sum_{j=1}^{J} \theta^j f_{n,k}\left(\theta^j\right)$$
$$\widehat{C}_{n+1,k} = \frac{1}{\hat{w}_{n+1,k}(J-1)}\sum_{j=1}^{J}\left(\theta^j - \widehat{m}_{n+1,k}\right)\left(\theta^j - \widehat{m}_{n+1,k}\right)^T f_{n,k}\left(\theta^j\right)$$

  end for
  Normalize $\hat{w}_{n+1,k} := \frac{\hat{w}_{n+1,k}}{\sum_{k=1}^{K}\hat{w}_{n+1,k}}$
  for $k \leftarrow 1$ to K **do** ▷Apply modified unscented transform [43, equation (37)]
    Generate sigma-points ($a = \max\{\frac{1}{8}, \frac{1}{2N_\theta}\}$)

$$\theta_k^0 = \widehat{m}_{n+1,k}$$
$$\theta_k^j = \widehat{m}_{n+1,k} + \frac{1}{\sqrt{2a}}\left[\sqrt{\widehat{C}_{n+1,k}}\right]_j \quad (1 \leqslant j \leqslant N_\theta)$$
$$\theta_k^{j+N_\theta} = \widehat{m}_{n+1,k} - \frac{1}{\sqrt{2a}}\left[\sqrt{\widehat{C}_{n+1,k}}\right]_j \quad (1 \leqslant j \leqslant N_\theta)$$

    Approximate the mean and covariance

$$\widehat{x}_{n+1,k} := \mathcal{F}\left(\theta_k^0\right)$$
$$\widehat{C}_{n+1,k}^{\theta x} := \sum_{j=1}^{2N_\theta} a\left(\theta_k^j - \widehat{m}_{n+1,k}\right)\left(\mathcal{F}\left(\theta_k^j\right) - \widehat{x}_{n+1,k}\right)^T$$
$$\widehat{C}_{n+1,k}^{xx} := \sum_{j=1}^{2N_\theta} a\left(\mathcal{F}\left(\theta_k^j\right) - \widehat{x}_{n+1,k}\right)\left(\mathcal{F}\left(\theta_k^j\right) - \widehat{x}_{n+1,k}\right)^T + \frac{1}{\Delta t}\Sigma_\nu$$

    Second step update

$$m_{n+1,k} = \widehat{m}_{n+1,k} + \widehat{C}_{n+1,k}^{\theta x}\left(\widehat{C}_{n+1,k}^{xx}\right)^{-1}\left(x - \widehat{x}_{n+1,k}\right)$$
$$C_{n+1,k} = \widehat{C}_{n+1,k} - \widehat{C}_{n+1,k}^{\theta x}\left(\widehat{C}_{n+1,k}^{xx}\right)^{-1}\widehat{C}_{n+1,k}^{\theta x^T}$$
$$w_{n+1,k} = \hat{w}_{n+1,k}e^{-\Delta t\Phi_R\left(\theta_k^0\right)}$$

  end for
  Normalize $w_{n+1,k} := \frac{w_{n+1,k}}{\sum_{k=1}^{K} w_{n+1,k}}$
end for

Similarly, for $m_{n,k}$ we have

$$\begin{aligned}
\frac{\widehat{m}_{n+1,k} - m_{n,k}}{\Delta t} &= \frac{\int(\theta - m_{n,k})\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta}{\Delta t\int\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\rho_n^{-\Delta t}\mathrm{d}\theta} \\
&= \frac{\int(\theta - m_{n,k})\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,(1 - \Delta t\log\rho_n)\,\mathrm{d}\theta}{\Delta t} + \mathcal{O}(\Delta t) \\
&= -\int\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,(\theta - m_{n,k})\log\rho_n(\theta)\,\mathrm{d}\theta + \mathcal{O}(\Delta t) \\
&= -C_{n,k}\int\mathcal{N}(\theta; m_{n,k}, C_{n,k})\,\nabla_\theta\log\rho_n(\theta)\,\mathrm{d}\theta + \mathcal{O}(\Delta t),
\end{aligned} \tag{C.3}$$

where in the last identity, we used integration by parts; it is also known as the Stein's identity. Thus the continuous limit is

$$\dot{m}_{t,k} = -\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})(\theta - m_{t,k}) \log \rho_t(\theta) \, \mathrm{d}\theta.$$

Finally, for $C_{n,k}$, we have

$$\begin{aligned}
\frac{\widehat{C}_{n+1,k} - C_{n,k}}{\Delta t} &= \frac{\int \left[(\theta - \widehat{m}_{n+1,k})(\theta - \widehat{m}_{n+1,k})^T - C_{n,k}\right] \mathcal{N}(\theta; m_{n,k}, C_{n,k}) \rho_n^{-\Delta t} \mathrm{d}\theta}{\Delta t \int \mathcal{N}(\theta; m_{n,k}, C_{n,k}) \rho_n^{-\Delta t} \mathrm{d}\theta} \\
&= \frac{\int \left[(\theta - \widehat{m}_{n+1,k})(\theta - \widehat{m}_{n+1,k})^T - C_{n,k}\right] \mathcal{N}(\theta; m_{n,k}, C_{n,k})(1 - \Delta t \log \rho_n) \mathrm{d}\theta}{\Delta t \int \mathcal{N}(\theta; m_{n,k}, C_{n,k})(1 - \Delta t \log \rho_n) \mathrm{d}\theta} \\
&\quad + \mathcal{O}(\Delta t) \\
&= \frac{\int \left[(\theta - \widehat{m}_{n+1,k})(\theta - \widehat{m}_{n+1,k})^T - C_{n,k}\right] \mathcal{N}(\theta; m_{n,k}, C_{n,k})(1 - \Delta t \log \rho_n) \mathrm{d}\theta}{\Delta t} \\
&\quad + \mathcal{O}(\Delta t) \\
&= -\int \left[(\theta - m_{n,k})(\theta - m_{n,k})^T - C_{n,k}\right] \mathcal{N}(\theta; m_{n,k}, C_{n,k}) \log \rho_n \mathrm{d}\theta + \mathcal{O}(\Delta t) \\
&= -C_{n,k} \int \mathcal{N}(\theta; m_{n,k}, C_{n,k}) \nabla_\theta \nabla_\theta \log \rho_n \mathrm{d}\theta C_{n,k} + \mathcal{O}(\Delta t),
\end{aligned} \tag{C.4}$$

where in the fourth identity, we used equation (C.3) $\widehat{m}_{n+1,k} = m_{n,k} + \mathcal{O}(\Delta t)$. And in the last identity, we used integration by parts. Finally, this leads to the desired continuous limit stated in the proposition.

C.2. Proof of proposition 4

The evolution equation of the entropy of ρ_t is

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \int -\rho_t \log \rho_t \mathrm{d}\theta \\
&= -\int \frac{\mathrm{d}\rho_t}{\mathrm{d}t} \log \rho_t \mathrm{d}\theta \\
&= -\sum_k \dot{w}_{t,k} \int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \log \rho_t \mathrm{d}\theta \\
&\quad - \sum_k w_k \dot{m}_{t,k}^T C_{t,k}^{-1} \int (\theta - m_{t,k}) \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \log \rho_t \mathrm{d}\theta \\
&\quad - \sum_k \frac{w_{t,k}}{2} \mathrm{tr}\left[\dot{C}_{t,k}^T \int \left(C_{t,k}^{-1}(\theta - m_{t,k})(\theta - m_{t,k})^T C_{t,k}^{-1} - C_{t,k}^{-1}\right) \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \log \rho_t \mathrm{d}\theta\right] \\
&= \sum_k \left(\frac{\dot{w}_{t,k}^2}{w_{t,k}} + w_{t,k} \dot{m}_{t,k}^T C_{t,k}^{-1} \dot{m}_{t,k} + \frac{w_{t,k}}{2} \mathrm{tr}\left[\dot{C}_{t,k} C_{t,k}^{-1} \dot{C}_{t,k} C_{t,k}^{-1}\right]\right).
\end{aligned} \tag{C.5}$$

Here we have used the fact $\int \frac{\mathrm{d}\rho_t}{\mathrm{d}t} \mathrm{d}\theta = 0$ in the first identity. And in the second identity we used the fact that $\sum_k \dot{w}_{t,k} = 0$ and used the continuous time limit equations equation (32).

C.3. Proof of proposition 5

Combining proposition 3 and equation (30) leads to the following

$$\begin{aligned}\frac{m_{n+1,k}-m_{n,k}}{\Delta t} &= \frac{\widehat{m}_{n+1,k}-m_{n,k}}{\Delta t}+\frac{\widehat{C}^{\theta x}_{n+1,k}\left(\widehat{C}^{xx}_{n+1,k}\right)^{-1}\left(x-\widehat{x}_{n+1,k}\right)}{\Delta t}\\ &= -C_{n,k}\int \mathcal{N}\left(\theta;m_{n,k},C_{n,k}\right)\nabla_\theta \log\rho_n\left(\theta\right)\mathrm{d}\theta+\widehat{C}^{\theta x}_{n,k}\Sigma_\nu^{-1}\left(x-\widehat{x}_{n,k}\right)+\mathcal{O}\left(\Delta t\right).\\ &\quad\times\frac{C_{n+1,k}-C_{n,k}}{\Delta t}\\ &= \frac{\widehat{C}_{n+1,k}-C_{n,k}}{\Delta t}-\frac{\widehat{C}^{\theta x}_{n+1,k}\left(\widehat{C}^{xx}_{n+1,k}\right)^{-1}\widehat{C}^{\theta x^T}_{n+1,k}}{\Delta t}\\ &= -C_{n,k}\int \mathcal{N}\left(\theta;m_{n,k},C_{n,k}\right)\nabla_\theta\nabla_\theta \log\rho_n\left(\theta\right)\mathrm{d}\theta C_{n,k}-\widehat{C}^{\theta x}_{n,k}\Sigma_\nu^{-1}\widehat{C}^{\theta x^T}_{n,k}+\mathcal{O}\left(\Delta t\right).\end{aligned} \tag{C.6}$$

Using the above formula, we obtain the continuous limit. For the weights, we have

$$w_{n+1,k}=\frac{\hat{w}_{n+1,k}\int\mathcal{N}\left(\theta;\widehat{m}_{n+1,k},\widehat{C}_{n+1,k}\right)e^{-\Delta t\Phi_R(\theta)}\mathrm{d}\theta}{\int\hat{\rho}_{n+1}e^{-\Delta t\Phi_R(\theta)}\mathrm{d}\theta}. \tag{C.7a}$$

Therefore,

$$\begin{aligned}\frac{w_{n+1,k}-w_{n,k}}{\Delta t} &= \frac{w_{n+1,k}-\hat{w}_{n+1,k}}{\Delta t}+\frac{\hat{w}_{n+1,k}-w_{n,k}}{\Delta t}\\ &= \hat{w}_{n+1,k}\frac{\int\left[\mathcal{N}\left(\theta;\widehat{m}_{n+1,k},\widehat{C}_{n+1,k}\right)-\hat{\rho}_{n+1}\right]e^{-\Delta t\Phi_R}\mathrm{d}\theta}{\Delta t\int\hat{\rho}_{n+1}e^{-\Delta t\Phi_R}\mathrm{d}\theta}\\ &\quad+\frac{\hat{w}_{n+1,k}-w_{n,k}}{\Delta t}+\mathcal{O}\left(\Delta t\right)\\ &= \hat{w}_{n+1,k}\frac{\int\left[\mathcal{N}\left(\theta;\widehat{m}_{n+1,k},\widehat{C}_{n+1,k}\right)-\hat{\rho}_{n+1}\right]\left(1-\Delta t\Phi_R\right)\mathrm{d}\theta}{\Delta t}\\ &\quad+\frac{\hat{w}_{n+1,k}-w_{n,k}}{\Delta t}+\mathcal{O}\left(\Delta t\right)\\ &= -w_{n,k}\int\left[\mathcal{N}\left(\theta;m_{n,k},C_{n,k}\right)-\rho_n\right]\left(\log\rho_n+\Phi_R\right)\mathrm{d}\theta+\mathcal{O}\left(\Delta t\right),\end{aligned} \tag{C.8}$$

from which we readily obtain the continuous limit for the equation of the weights. Here in the last step we used the result in C.1.

C.4. Proof of proposition 6

Under the Gaussian posterior assumption (i.e. Φ_R is quadratic) and based on the formula in (25) and (26), we get

$$\widehat{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\left(x-\widehat{x}_{t,k}\right)=-C_{t,k}\int\mathcal{N}\left(\theta;m_{t,k},C_{t,k}\right)\nabla_\theta\Phi_R\mathrm{d}\theta, \tag{C.9}$$

$$\widehat{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\widehat{C}^{T}_{t,k}=C_{t,k}\left(\int\mathcal{N}\left(\theta;m_{t,k},C_{t,k}\right)\nabla_\theta\nabla_\theta\Phi_R\mathrm{d}\theta\right)C^{T}_{t,k}. \tag{C.10}$$

Bringing equation (C.9) into the continuous time dynamics equation (35) and using integration by parts (also known as Stein's identities) leads to

$$\begin{aligned}\dot{m}_{t,k} &= -C_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\,\nabla_\theta\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta \\ &= -\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\,(\theta - m_{t,k})\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta, \end{aligned} \tag{C.11a}$$

$$\begin{aligned}\dot{C}_{t,k} &= -\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\,\nabla_\theta\nabla_\theta\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta \\ &= -\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\left((\theta - m_{t,k})(\theta - m_{t,k})^T - C_{t,k}\right)\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta, \end{aligned} \tag{C.11b}$$

$$\dot{w}_{t,k} = -w_{t,k}\int \left[\mathcal{N}(\theta; m_{t,k}, C_{t,k}) - \rho_t\right]\left[\log\rho_t + \Phi_R\right]\mathrm{d}\theta. \tag{C.11c}$$

The evolution equation of the KL divergence between ρ_t and ρ_{post} is then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left[\rho_t\|\rho_{\text{post}}\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\int \rho_t \log\left(\frac{\rho_t}{\rho_{\text{post}}}\right)\mathrm{d}\theta = \int \frac{\mathrm{d}\rho_t}{\mathrm{d}t}\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta \\ &= \sum_k \dot{w}_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta \\ &\quad + \sum_k w_k \dot{m}_{t,k}^T C_{t,k}^{-1}\int (\theta - m_{t,k})\,\mathcal{N}(\theta; m_{t,k}, C_{t,k})\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta \\ &\quad + \sum_k \frac{w_{t,k}}{2}\mathrm{tr}\left[\dot{C}_{t,k}^T\int \left(C_{t,k}^{-1}(\theta - m_{t,k})(\theta - m_{t,k})^T C_{t,k}^{-1} - C_{t,k}^{-1}\right)\right. \\ &\quad \left.\times\mathcal{N}(\theta; m_{t,k}, C_{t,k})\left(\log\rho_t + \Phi_R\right)\mathrm{d}\theta\right] \\ &= -\sum_k\left(\frac{\dot{w}_{t,k}^2}{w_{t,k}} + w_{t,k}\dot{m}_{t,k}^T C_{t,k}^{-1}\dot{m}_{t,k} + \frac{w_{t,k}}{2}\mathrm{tr}\left[\dot{C}_{t,k}C_{t,k}^{-1}\dot{C}_{t,k}C_{t,k}^{-1}\right]\right). \end{aligned} \tag{C.12}$$

Here in the last identity, we used the equation of the continuous time dynamics in (C.11) to simplify the formula.

Consider any stationary point $\rho_\infty = \sum_k w_{k,\infty}\mathcal{N}(\theta; m_{k,\infty}, C_{k,\infty})$ with nonzero $w_{k,\infty}$. The stationary point condition for the mean $m_{k,\infty}$ (C.11*a*) is

$$\int \mathcal{N}(\theta; m_{k,\infty}, C_{k,\infty})\,\nabla_\theta \log\rho_\infty \mathrm{d}\theta = -C_{\text{post}}^{-1}\left(m_{k,\infty} - m_{\text{post}}\right), \tag{C.13}$$

where we used that $\int \mathcal{N}(\theta; m_{\infty,k}, C_{\infty,k})\nabla_\theta\Phi_R\mathrm{d}\theta = C_{\text{post}}^{-1}(m_{k,\infty} - m_{\text{post}})$. The stationary point condition for the covariance $C_{k,\infty}$ (C.11*b*) is

$$\int \mathcal{N}(\theta; m_{k,\infty}, C_{k,\infty})\,\nabla_\theta\nabla_\theta \log\rho_\infty \mathrm{d}\theta = -C_{\text{post}}^{-1}, \tag{C.14}$$

where we used that $\int \mathcal{N}(\theta; m_{\infty,k}, C_{\infty,k})\nabla_\theta\nabla_\theta\Phi_R\mathrm{d}\theta = C_{\text{post}}^{-1}$. Multiplying equations (C.13) and (C.14) by $w_{k,\infty}$ and summing the results yields

$$m_{\text{post}} - \sum_k w_{k,\infty} m_{k,\infty} = C_{\text{post}} \int \sum_k w_{k,\infty} \mathcal{N}\left(\theta; m_{k,\infty}, C_{k,\infty}\right) \nabla_\theta \log \rho_\infty \mathrm{d}\theta = 0,$$
$$C_{\text{post}}^{-1} = -\int \sum w_{k,\infty} \mathcal{N}\left(\theta; m_{k,\infty}, C_{k,\infty}\right) \nabla_\theta \nabla_\theta \log \rho_\infty \mathrm{d}\theta = \mathrm{FIM}\left[\rho_\infty\right]. \tag{C.15}$$

C.5. Derivation of the simplified continuous time dynamics equation (39)

In this section, we formally derive equation (39), assuming these Gaussian components in ρ_{post} are well separated. When θ is close to m_k^*, one may make the following approximation

$$\Phi_R\left(\theta\right) = \log \rho_{\text{post}}\left(\theta\right) \approx -\log \mathcal{N}\left(\theta, m_k^*, C_k^*\right) - \log\left(w_k^*\right). \tag{C.16}$$

Combining the definition of Φ_R in equation (19) with equation (C.16) leads to that

$$-\frac{1}{2}\left(x - \mathcal{F}\left(\theta\right)\right)^T \Sigma_\nu^{-1}\left(x - \mathcal{F}\left(\theta\right)\right) \approx -\frac{1}{2}\left(\theta - m_k^*\right)^T {C_k^*}^{-1}\left(\theta - m_k^*\right) + \text{constant}. \tag{C.17}$$

This implies that $\mathcal{F}(\theta)$ is approximately locally linear around $m = m_k^*$ with $\mathcal{F}(\theta) \approx F_k\theta + c$, such that $F_k^T \Sigma_\nu^{-1} F_k = {C_k^*}^{-1}$ and $C_k^* F_k^T \Sigma_\nu^{-1}(x - c) = m_k^*$. Based on the above derivation, when the kth component $\mathcal{N}(\theta; m_{t,k}, C_{t,k})$ in the Gaussian mixture approximation is concentrating around m_k^*, the expectation and covariance in the continuous time limit of the proposed GMKI (36) can be approximated as

$$\widehat{x}_{t,k} = \mathbb{E}\left[\mathcal{F}\left(\theta\right)\right] \approx \mathbb{E}\left[F_k\theta + c\right] = F_k m_{t,k} + c, \tag{C.18a}$$
$$\widehat{C}_{t,k}^{\theta x} = \mathrm{Cov}\left[\theta, \mathcal{F}\left(\theta\right)\right] \approx \mathbb{E}\left[\left(\theta - m_{t,k}\right) \otimes \left(F_k\left(\theta - m_{t,k}\right)\right)\right] = C_{t,k} F_k^T. \tag{C.18b}$$

Here the expectation are taken with respect to $\mathcal{N}(\theta; m_{t,k}, C_{t,k})$.

Now we will simplify equation (35) by neglecting the interaction between well separated Gaussian components to obtain equation (39). For the mean evolution equation (35*a*), we have

$$\begin{aligned}
\dot{m}_{t,k} &\approx -C_{t,k} \int \mathcal{N}\left(\theta; m_{t,k}, C_{t,k}\right) \nabla_\theta \log\left(w_k \mathcal{N}\left(\theta; m_{t,k}, C_{t,k}\right)\right) \mathrm{d}\theta + \widehat{C}_{t,k}^{\theta x} \Sigma_\nu^{-1}\left(x - \widehat{x}_{t,k}\right) \\
&= \widehat{C}_{t,k}^{\theta x} \Sigma_\nu^{-1}\left(x - \widehat{x}_{t,k}\right) \\
&\approx C_{t,k}\left(C_k^*\right)^{-1}\left(m_k^* - m_{t,k}\right).
\end{aligned} \tag{C.19}$$

The first approximation is obtained by substituting $\log \rho_t$ with $\log(w_k \mathcal{N}(\theta; m_{t,k}, C_{t,k}))$, due to the well separateness assumption; the resulting integral is zero so leads to the second identity. The third approximation is obtained by using equation (C.18) and the relation $F_k^T \Sigma_\nu^{-1} F_k = {C_k^*}^{-1}$ and $C_k^* F_k^T \Sigma_\nu^{-1}(x - c) = m_k^*$.

For the covariance evolution equation (35*b*), similarly we have

$$\begin{aligned}
\dot{C}_{t,k} &\approx -C_{t,k}\left(\int \mathcal{N}\left(\theta; m_{t,k}, C_{t,k}\right) \nabla_\theta \nabla_\theta \log\left(w_k \mathcal{N}\left(\theta; m_{t,k}, C_{t,k}\right)\right) \mathrm{d}\theta\right) C_{t,k} - \widehat{C}_{t,k}^{\theta x} \Sigma_\nu^{-1} \widehat{C}_{t,k}^T \\
&= C_{t,k} - \widehat{C}_{t,k}^{\theta x} \Sigma_\nu^{-1} \widehat{C}_{t,k}^T \\
&= C_{t,k} - C_{t,k}\left(C_k^*\right)^{-1} C_{t,k}.
\end{aligned} \tag{C.20}$$

The first approximation is obtained by substituting $\log \rho_t$ with $\log(w_k \mathcal{N}(\theta; m_{t,k}, C_{t,k}))$. The third approximation is obtained by using equation (C.18).

Finally, for the weight evolution equation (35*c*), using the formula $\rho_t(\theta) = \sum_i \int w_{t,i}\mathcal{N}(\theta; m_{t,i}, C_{t,i})$, we get

$$\begin{aligned}
\dot{w}_{t,k} &= -w_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \log \rho_t(\theta)\,\mathrm{d}\theta + w_{t,k}\sum_i \int w_{t,i}\mathcal{N}(\theta; m_{t,i}, C_{t,i}) \log \rho_t(\theta)\,\mathrm{d}\theta \\
&\quad - w_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\, \Phi_R(\theta)\,\mathrm{d}\theta + w_{t,k}\sum_i \int w_{t,i}\mathcal{N}(\theta; m_{t,i}, C_{t,i})\, \Phi_R(\theta)\,\mathrm{d}\theta \\
&\approx -w_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \log\left[w_{t,k}\mathcal{N}(\theta; m_{t,k}, C_{t,k})\right]\mathrm{d}\theta \\
&\quad + w_{t,k}\sum_i \int w_{t,i}\mathcal{N}(\theta; m_{t,i}, C_{t,i}) \log\left[w_{t,i}\mathcal{N}(\theta; m_{t,i}, C_{t,i})\right]\mathrm{d}\theta \\
&\quad + w_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \left(\log \mathcal{N}(\theta, m_k^*, C_k^*) + \log(w_k^*)\right)\mathrm{d}\theta \\
&\quad - w_{t,k}\sum_i \int w_{t,i}\mathcal{N}(\theta; m_{t,i}, C_{t,i}) \left(\log \mathcal{N}(\theta, m_i^*, C_i^*) + \log(w_i^*)\right)\mathrm{d}\theta \\
&= w_{t,k}\left(\log w_k^* - \log w_{t,k} - \sum_i w_{t,i}\left(\log w_i^* - \log w_{t,i}\right)\right).
\end{aligned} \tag{C.21}$$

The first approximation is obtained by substituting $\log \rho_t$ with $\log(w_k\mathcal{N}(\theta; m_{t,k}, C_{t,k}))$ and using equation (C.16) for approximating $\Phi_R(\theta)$. Combining equations (C.19)–C.21) leads to the simplified continuous time dynamics equation (39).

C.6. *Proof of proposition 7*

The mean, covariance and weight evolution equations (39*a*)–(39*c*) admit analytical solutions

$$m_{t,k} = m_k^* + e^{-t}\left(\left(1-e^{-t}\right)C_k^{*^{-1}} + e^{-t}C_k(0)^{-1}\right)^{-1} C_k(0)^{-1}\left(m_k(0) - m_k^*\right), \tag{C.22a}$$

$$C_{t,k}^{-1} = C_k^{*^{-1}} + e^{-t}\left(C_k(0)^{-1} - C_k^{*^{-1}}\right), \tag{C.22b}$$

$$w_k = \frac{w_k^*\left(\frac{w_k(0)}{w_k^*}\right)^{e^{-t}}}{\sum_i w_i^*\left(\frac{w_i(0)}{w_i^*}\right)^{e^{-t}}}. \tag{C.22c}$$

They will converge to m_k^*, C_k^* and w_k^* exponentially fast.

C.7. *Proof of proposition 8*

Consider any invertible affine mapping $\varphi : \theta \to \widetilde{\theta} = A\theta + b$, and define corresponding vector and matrix transformations

$$\widetilde{m}_{t,k} = Am_{t,k} + b \quad \widetilde{C}_{t,k} = AC_{t,k}A^T,$$

density transformations

$$\widetilde{\rho}\left(\widetilde{\theta}\right) = \varphi_\sharp\rho(\theta) = \rho\left(A^{-1}\left(\widetilde{\theta} - b\right)\right)|A|^{-1} \quad \mathcal{N}\left(\widetilde{\theta}; \widetilde{m}_{t,k}, \widetilde{C}_{t,k}\right) = \mathcal{N}(\theta; m_{t,k}, C_{t,k})\,|A|^{-1},$$

function transformations

$$\widetilde{\mathcal{F}}\left(\widetilde{\theta}\right)=\mathcal{F}\left(A^{-1}\left(\widetilde{\theta}-b\right)\right)\qquad\widetilde{\Phi}_R\left(\widetilde{\theta}\right)=\Phi_R\left(A^{-1}\left(\widetilde{\theta}-b\right)\right),$$

and their related expectation and covariance

$$\widetilde{x}_{t,k}=\mathbb{E}\left[\widetilde{\mathcal{F}}\left(\widetilde{\theta}\right)\right],\qquad\widetilde{C}^{\theta x}_{t,k}=\mathrm{Cov}\left[\widetilde{\theta},\widetilde{\mathcal{F}}\left(\widetilde{\theta}\right)\right],\quad\text{with}\quad\widetilde{\theta}\sim\mathcal{N}\left(\widetilde{m}_{t,k},\widetilde{C}_{t,k}\right),$$

then we have

$$\begin{aligned}&\nabla_{\widetilde{\theta}}\log\widetilde{\rho}\left(\widetilde{\theta}\right)=A^{-T}\nabla_\theta\log\rho\left(\theta\right)\qquad\nabla_{\widetilde{\theta}}\nabla_{\widetilde{\theta}}\log\widetilde{\rho}\left(\widetilde{\theta}\right)=A^{-T}\nabla_\theta\log\rho\left(\theta\right)A^{-1}\\&\widetilde{x}_{t,k}=\widehat{x}_{t,k}\qquad\widetilde{C}^{\theta x}_{t,k}=A\widehat{C}^{\theta x}_{t,k}.\end{aligned}\tag{C.23}$$

The evolution equations of $\widetilde{m}_{t,k}$, $\widetilde{C}_{t,k}$, $w_{t,k}$ in equation (35) can be rewritten as

$$\begin{aligned}\dot{\widetilde{m}}_{t,k}&=-AC_{t,k}\int\mathcal{N}\left(\theta;m_{t,k},C_{t,k}\right)\nabla_\theta\log\rho_t\left(\theta\right)\mathrm{d}\theta+A\widehat{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\left(x-\widehat{x}_{t,k}\right),\\&=-\widetilde{C}_{t,k}\int\mathcal{N}\left(\widetilde{\theta};\widetilde{m}_{t,k},\widetilde{C}_{t,k}\right)\nabla_{\widetilde{\theta}}\log\widetilde{\rho}_t\left(\widetilde{\theta}\right)\mathrm{d}\widetilde{\theta}+\widetilde{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\left(x-\widetilde{x}_{t,k}\right),\\\dot{\widetilde{C}}_{t,k}&=-AC_{t,k}\left(\int\mathcal{N}\left(\theta;m_{t,k},C_{t,k}\right)\nabla_\theta\nabla_\theta\log\rho_t\left(\theta\right)\mathrm{d}\theta\right)C_{t,k}A^T-A\widehat{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\widehat{C}^{\theta xT}_{t,k}A^T,\\&=-\widetilde{C}_{t,k}\left(\int\mathcal{N}\left(\widetilde{\theta};\widetilde{m}_{t,k},\widetilde{C}_{t,k}\right)\nabla_{\widetilde{\theta}}\nabla_{\widetilde{\theta}}\log\widetilde{\rho}_t\left(\widetilde{\theta}\right)\mathrm{d}\widetilde{\theta}\right)\widetilde{C}_{t,k}-\widetilde{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\widetilde{C}^{\theta x^T}_{t,k},\\\dot{w}_{t,k}&=-w_{t,k}\int\left[\mathcal{N}\left(\theta;m_{t,k},C_{t,k}\right)-\rho_t\left(\theta\right)\right]\left[\log\rho_t\left(\theta\right)+\Phi_R\left(\theta\right)\right]\mathrm{d}\theta\\&=-w_{t,k}\int\left[\mathcal{N}\left(\widetilde{\theta};\widetilde{m}_{t,k},\widetilde{C}_{t,k}\right)-\widetilde{\rho}_t\left(\widetilde{\theta}\right)\right]\left[\log\widetilde{\rho}_t\left(\widetilde{\theta}\right)+\widetilde{\Phi}_R\left(\widetilde{\theta}\right)\right]\mathrm{d}\widetilde{\theta}.\end{aligned}$$

Hence the continuous time limit equation equation (35) of GMKI is affine invariant.

C.8. Connections between the GMKI approach and GMVI

GMVI seeks to identify a minimizer of $\mathrm{KL}[\rho^{\mathrm{GM}}\|\rho_{\mathrm{post}}]$, where

$$\rho^{\mathrm{GM}}\left(\theta;a\right)=\sum_{k=1}^{K}w_k\mathcal{N}\left(\theta;m_k,C_k\right)$$

is a K-component Gaussian mixture, parameterized by their means, covariances and weights denoted by

$$a:=\left[m_1,\ldots,m_k,\ldots,m_K,C_1\ldots,C_k,\ldots,C_K,w_1,\ldots,w_k,\ldots,w_K\right].$$

The derivatives of the KL divergence with respect to a are

$$\frac{\partial\mathrm{KL}\left[\rho^{\mathrm{GM}}\left(\cdot;a\right)\|\rho_{\mathrm{post}}\right]}{\partial m_k}=w_k\int\mathcal{N}\left(\theta;m_k,C_k\right)\left(\nabla_\theta\log\rho^{\mathrm{GM}}-\nabla_\theta\log\rho_{\mathrm{post}}\right)\mathrm{d}\theta,\tag{C.24a}$$

$$\frac{\partial\mathrm{KL}\left[\rho^{\mathrm{GM}}\left(\cdot;a\right)\|\rho_{\mathrm{post}}\right]}{\partial C_k}=\frac{w_k}{2}\int\mathcal{N}\left(\theta;m_k,C_k\right)\left(\nabla_\theta\nabla_\theta\log\rho^{\mathrm{GM}}-\nabla_\theta\nabla_\theta\log\rho_{\mathrm{post}}\right)\mathrm{d}\theta,\tag{C.24b}$$

$$\frac{\partial \mathrm{KL}\left[\rho^{\mathrm{GM}}(\cdot;a)\,\|\rho_{\mathrm{post}}\right]}{\partial w_k} = \int \mathcal{N}(\theta;m_k,C_k)\left(\log\frac{\rho^{\mathrm{GM}}}{\rho_{\mathrm{post}}}+1\right)\mathrm{d}\theta. \tag{C.24c}$$

The algorithm of natural gradient descent uses the finite dimensional version of the Fisher–Rao metric tensor in the parameter space, also known as the Fisher information matrix with the form

$$\mathrm{FI}(a) = \int \frac{\nabla_a \rho^{\mathrm{GM}}(\theta;a)\otimes\nabla_a\rho^{\mathrm{GM}}(\theta;a)}{\rho^{\mathrm{GM}}(\theta;a)}\mathrm{d}\theta \tag{C.25}$$

as the preconditioner for gradient descent. Here, we write down its continuous limit, namely the natural gradient flow. To do so, we use the perspective of proximal point method and consider

$$a_{n+1} = \arg\min_a \mathrm{KL}\left[\rho^{\mathrm{GM}}(\cdot;a)\,\|\rho_{\mathrm{post}}\right] + \frac{1}{2\Delta t}\left\langle a-a_n, \mathrm{FI}(a_n)(a-a_n)\right\rangle, \tag{C.26a}$$

$$\text{subject to } \sum_{k=1}^{K} w_{n+1,k} = 1. \tag{C.26b}$$

By using the formula of derivatives in equation (C.24), the Lagrangian multiplier to handle the constraint in the above optimization, and taking $\Delta t \to 0$, we arrive at the following natural gradient flow

$$\begin{bmatrix}\dot m_k\\ \dot C_k\\ \dot w_k\end{bmatrix} = (\mathrm{FI}(a))^{-1}\begin{bmatrix} -w_k\int\mathcal{N}(\theta;m_k,C_k)\left(\nabla_\theta\log\rho^{\mathrm{GM}}-\nabla_\theta\log\rho_{\mathrm{post}}\right)\mathrm{d}\theta\\ -\frac{w_k}{2}\int\mathcal{N}(\theta;m_k,C_k)\left(\nabla_\theta\nabla_\theta\log\rho^{\mathrm{GM}}-\nabla_\theta\nabla_\theta\log\rho_{\mathrm{post}}\right)\mathrm{d}\theta\\ -\int\left(\mathcal{N}(\theta;m_k,C_k)-\rho^{\mathrm{GM}}\right)\log\left(\frac{\rho^{\mathrm{GM}}}{\rho_{\mathrm{post}}}\right)\mathrm{d}\theta\end{bmatrix}. \tag{C.27}$$

Computation of $\mathrm{FI}(a)$ is costly. For better efficiency, diagonal approximations of the Fisher information matrix have been used in the literature [33], which leads to

$$\begin{aligned}\mathrm{FI}(a) \approx \mathrm{diag}\big(&w_1C_1^{-1},\ldots,w_kC_k^{-1},\ldots,w_KC_K^{-1},w_1X_1,\ldots,w_kX_k,\ldots,w_KX_K,\\ &\times\frac{1}{w_1},\ldots,\frac{1}{w_k},\ldots,\frac{1}{w_K}\big).\end{aligned} \tag{C.28}$$

where each X_k is a 4th order tensor satisfying

$$X_kY = \frac{1}{4}C_k^{-1}\left(Y+Y^T\right)C_k^{-1},\quad \forall\, Y\in\mathbb{R}^{N_\theta\times N_\theta}. \tag{C.29}$$

Bringing the approximated Fisher information matrix equation (C.28) into the natural gradient flow equation (C.27) leads to the following equation:

$$\begin{aligned}\dot m_k &= -C_k\int\mathcal{N}(\theta;m_k,C_k)\left(\nabla_\theta\log\rho^{\mathrm{GM}}-\nabla_\theta\log\rho_{\mathrm{post}}\right)\mathrm{d}\theta,\\ \dot C_k &= -C_k\left(\int\mathcal{N}(\theta;m_k,C_k)\left(\nabla_\theta\nabla_\theta\log\rho^{\mathrm{GM}}-\nabla_\theta\nabla_\theta\log\rho_{\mathrm{post}}\right)\mathrm{d}\theta\right)C_k,\\ \dot w_k &= -w_k\int\left(\mathcal{N}(\theta;m_k,C_k)-\rho^{\mathrm{GM}}\right)\log\left(\frac{\rho^{\mathrm{GM}}}{\rho_{\mathrm{post}}}\right)\mathrm{d}\theta.\end{aligned} \tag{C.30}$$

This is the natural gradient flow with diagonal approximations of the Fisher information matrix; its discretization is the approximate natural gradient descent algorithm that has been used in the literature.

By comparing equation (C.30) with the continuous-time limit of our GMKI as presented in equation (35), we observe that our GMKI can be seen as a derivative-free approximation of the approximate natural gradient descent. The approximation is made through stochastic linearization, for $\nabla_\theta \log \rho_{\text{post}}$ and $\nabla_\theta \nabla_\theta \log \rho_{\text{post}}$, based on the Kalman methodology explained in remark 1.

For completeness, in the following part, we present a derivation of the diagonal approximations of the Fisher information matrix (C.28). Let $\mathcal{N}_k$ denote $\mathcal{N}(\theta; m_k, C_k)$ and $\delta_{k,i}$ be the indicator function which is zero if and only if $k = i$. We can get equation (C.28) by only keeping the diagonal blocks of $\text{FI}(a)$ and approximating the diagonals under the assumptions that different Gaussian components are well separated. More precisely, for the diagonal block regarding the weight $\{w_k\}$, we have

$$\int \frac{\nabla_{w_k}\rho^{\text{GM}} \otimes \nabla_{w_i}\rho^{\text{GM}}}{\rho^{\text{GM}}} \mathrm{d}\theta = \int \frac{\mathcal{N}_k \mathcal{N}_i}{\rho^{\text{GM}}} \mathrm{d}\theta \approx \delta_{k,i} \int \frac{\mathcal{N}_k \mathcal{N}_k}{w_k \mathcal{N}_k} \mathrm{d}\theta = \frac{\delta_{k,i}}{w_k}. \tag{C.31}$$

Here we substitute ρ^{GM} by $w_k \mathcal{N}_k$ during its integration with $\mathcal{N}_k$. We note that we will keep using this approximation multiple times in the following derivations.

For the diagonal block regarding the mean m_k, we have

$$\begin{aligned}
\int \frac{\nabla_{m_k}\rho^{\text{GM}} \otimes \nabla_{m_i}\rho^{\text{GM}}}{\rho^{\text{GM}}} \mathrm{d}\theta &= \int \frac{w_k w_i \mathcal{N}_k \mathcal{N}_i C_k^{-1} (\theta - m_k)(\theta - m_i)^T C_i^{-1}}{\rho^{\text{GM}}} \mathrm{d}\theta \\
&\approx \delta_{k,i} \int \frac{w_k^2 \mathcal{N}_k^2 C_k^{-1} (\theta - m_k)(\theta - m_k)^T C_k^{-1}}{w_k \mathcal{N}_k} \mathrm{d}\theta \\
&= \delta_{k,i} w_k C_k^{-1}.
\end{aligned} \tag{C.32}$$

For the diagonal block regarding the covariance C_k, we have

$$\begin{aligned}
&\int \frac{\nabla_{C_k}\rho^{\text{GM}} \otimes \nabla_{C_i}\rho^{\text{GM}}}{\rho^{\text{GM}}} \mathrm{d}\theta \\
&= \int \frac{w_k w_i \mathcal{N}_k \mathcal{N}_i \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right) \otimes \left(C_i^{-1}(\theta - m_i)(\theta - m_i)^T C_i^{-1} - C_i^{-1}\right)}{4\rho^{\text{GM}}} \mathrm{d}\theta \\
&\approx \delta_{k,i} \int \frac{w_k^2 \mathcal{N}_k^2 \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right) \otimes \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right)}{4 w_k \mathcal{N}_k} \mathrm{d}\theta \\
&= \frac{\delta_{k,i}}{4} \int w_k \mathcal{N}_k \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right) \otimes \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right) \mathrm{d}\theta.
\end{aligned} \tag{C.33}$$

It is worth noting that equation (C.33) is a 4th order tensor. To gain a more detailed understanding of this term, let us denote

$$\begin{aligned}
X_k :&= \frac{1}{4} \int \mathcal{N}_k \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right) \\
&\quad \otimes \left(C_k^{-1}(\theta - m_k)(\theta - m_k)^T C_k^{-1} - C_k^{-1}\right) \mathrm{d}\theta \\
&= \frac{1}{4} \int \mathcal{N}(y; 0, I) C_k^{-1/2} \left(yy^T - I\right) C_k^{-1/2} \otimes C_k^{-1/2} \left(yy^T - I\right) C_k^{-1/2} \mathrm{d}y \\
&\quad \text{where } y = C_k^{-1/2}(\theta - m_k).
\end{aligned} \tag{C.34}$$

We can show that X_k satisfies equation (C.29). To do so note that the (ij, lm) entry of X_k has the form

$$\begin{aligned} X_k[ij,lm] &= \frac{1}{4}\sum_{r,s,p,q} C_k^{-1/2}[i,r]\, C_k^{-1/2}[j,s]\, C_k^{-1/2}[l,p]\, C_k^{-1/2}[m,q] \\ &\quad \times \int (y_r y_s - \delta_{r,s})(y_p y_q - \delta_{p,q})\mathcal{N}(y;0,I)\,\mathrm{d}y \\ &= \frac{1}{4}\sum_{r,s,p,q} C_k^{-1/2}[i,r]\, C_k^{-1/2}[j,s]\, C_k^{-1/2}[l,p]\, C_k^{-1/2}[m,q]\,(\delta_{r,p}\delta_{s,q} + \delta_{r,q}\delta_{s,p}) \\ &= \frac{1}{4}\left(C_k^{-1}[i,l]\, C_k^{-1}[j,m] + C_k^{-1}[i,m]\, C_k^{-1}[j,l]\right). \end{aligned}$$

Therefore,

$$\begin{aligned} (X_k Y)_{ij} &= \sum_{l,m} X_k[ij,lm]\, Y[l,m] \\ &= \frac{1}{4}\sum_{l,m}\left(C_k^{-1}[i,l]\, Y[l,m]\, C_k^{-1}[j,m] + C_k^{-1}[i,m]\, Y[l,m]\, C_k^{-1}[j,l]\right) \\ &= \frac{1}{4}\left(C_k^{-1} Y C_k^{-1} + C_k^{-1} Y^T C_k^{-1}\right)_{ij}. \end{aligned} \tag{C.35}$$

The proof is complete.

Appendix D. Two-dimensional four-modal problem

In this subsection, we consider a 2D four-modal inverse problem, associated with the forward model

$$y = \mathcal{G}(\theta) + \eta \quad \text{with} \quad y = \begin{bmatrix} 4.2297 \\ 4.2297 \end{bmatrix}, \quad \mathcal{G}(\theta) = \begin{bmatrix} (\theta_{(1)} - \theta_{(2)})^2 \\ (\theta_{(1)} + \theta_{(2)})^2 \end{bmatrix}.$$

Here $\theta = [\theta_{(1)}, \theta_{(2)}]^T$. We assume the noise distribution is $\eta \sim \mathcal{N}(0, I)$ and consider the prior distribution $\rho_{\text{prior}} \sim \mathcal{N}([0.5, 0]^T, I)$. The reference posterior distribution has four modes with varying weights. We apply GMKI with $K = 3$ and 6 modes, randomly initialized based on the prior distribution, and assign equal weights to these components. The estimated posterior distributions obtained by GMKI at the 30th iteration, along with the convergence in terms of TV distance, are shown in figure D1. When $K = 3$ only three target modes are captured; when $K = 6$, all target modes are captured, and the approximation error becomes significantly smaller.

Appendix E. Limitation of the GMKI

There can be multimodal problems with many modes concentrating on a low dimensional manifold. GMKI may fail in such a case. To illustrate, we consider the posterior distribution $\exp(-\Phi_R(\theta))$ in $\mathbb{R}^2$, where $\Phi_R(\theta) = \frac{(1-\theta_{(1)}^2-\theta_{(2)}^2)^2}{2\sigma_\eta^2}$ and $\sigma_\eta = 0.3$. Clearly the mass is distributed along the unit circle, as depicted on the left side of figure E1. We sample this density with GMKI and GMVI, described in equation (C.30). Note that GMKI can be seen as a derivative free approximation of GMVI so this study is for the purpose of understanding the effect

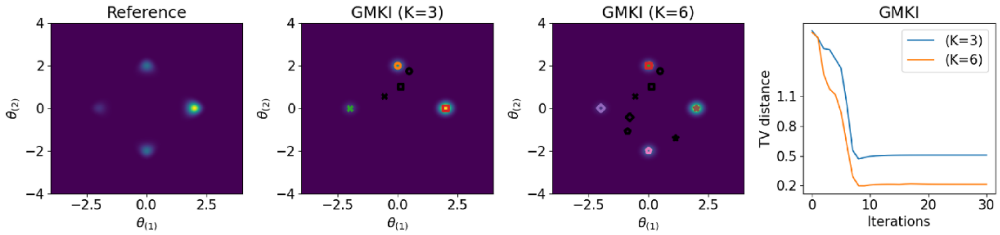

Figure D1. Two-dimensional four-modal problem. From left to right: reference posterior distribution (left), posterior distributions estimated by 3 and 6-modal GMKI (middle) at the 30th iteration (means m_k (colored) and initial means (black) are marked), and total variation distance between the reference posterior distribution and the posterior distributions estimated by the GMKI (right).

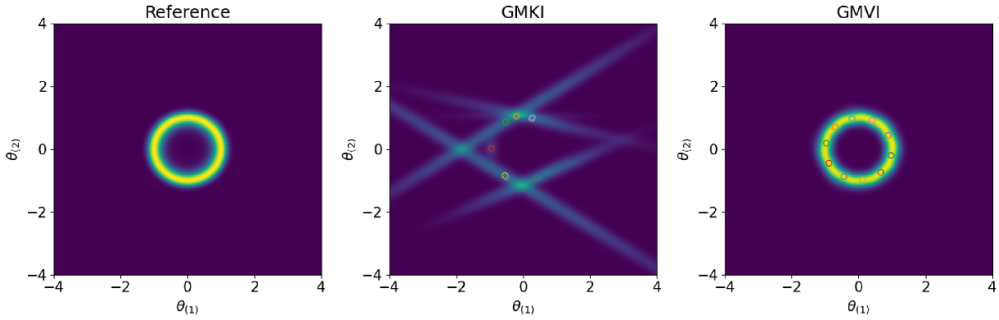

Figure E1. Circle shape posterior: reference posterior distribution (left), posterior distribution obtained by 10-modal GMKI (middle), posterior distribution obtained by 10-modal GMVI (right). Means m_k of Gaussian components are marked. In the middle, the color appears lighter because the distribution is concentrated along a thin, elongated line, resulting in a visually diluted effect.

of the derivative free approximation step. We and initialize both methods with 10 Gaussian components with means randomly sampled from $\mathcal{N}(0, I)$ and the same identity covariance I.

The sampling results obtained by GMKI at the 30th iteration are presented in the middle of figure E1. While the means of these Gaussian components migrate towards the unit circle, the covariance associated with each Gaussian component are elongated in the tangent direction. Consequently, the overall approximation of the target distribution is inaccurate. The covariance of these Gaussian components bear resemblance to those seen in the Laplace approximation. Indeed, the Laplace approximation at any maximum *a posteriori* (MAP) of the target distribution has the form $\mathcal{N}(\theta; m, H^{\dagger})$; here, $m = [m_{(1)}, m_{(2)}]$ lies on the unit circle and

$$H = -\nabla_\theta \nabla_\theta \Phi_R(\theta)\,|_{\theta=m} = \frac{4}{\sigma_\eta^2}\begin{bmatrix} m_{(1)}^{(2)} & m_{(1)}m_{(2)} \\ m_{(1)}m_{(2)} & m_{(2)}^2 \end{bmatrix}. \tag{E.1}$$

It is worth mentioning that H exhibits singularity, particularly along the tangent direction of the unit circle. Hence the Laplace approximation is degenerate and concentrates on the tangential line of the unit circle at m.

To further explore this issue, we turn to use the algorithm for GMVI. This approach requires the evaluation of the gradient and Hessian of $\log \rho_{\text{post}}$. We approximate these Gaussian integrations in equation (C.30) using the modified unscented transform, as detailed in [43, equation

(37)]. Moreover we employ a time-step of $\Delta t = 0.01$, which leads to a stable numerical scheme in our implementation. The outcome of GMVI at the 1000th iteration is depicted in the right of figure E1. The result matches the reference well. Since the main difference between GMVI equation (C.30) and the continuous time dynamics of GMKI is the derivative-free Kalman approximation for the gradient terms $\nabla_\theta \log \rho_{\text{post}}$ and $\nabla_\theta \nabla_\theta \log \rho_{\text{post}}$, we understand that the Kalman approximation step leads to the failure of GMKI for sampling the above distribution. It is the goal of future study to investigate other derivative free approximations that can circumvent this failure.

ORCID iDs

Yifan Chen https://orcid.org/0000-0001-5494-4435
Daniel Zhengyu Huang https://orcid.org/0000-0001-6072-9352

References

[1] Kaipio J and Somersalo E 2006 *Statistical and Computational Inverse Problems* vol 160 (Springer Science & Business Media)
[2] Stuart A M 2010 Inverse problems: a Bayesian perspective *Acta Numer.* **19** 451–559
[3] Brooks S, Gelman A, Jones G and Meng X-Li 2011 *Handbook of Markov Chain Monte Carlo* (CRC Press)
[4] Del Moral P, Doucet A and Jasra A 2006 Sequential Monte Carlo samplers *J. R. Stat. Soc.* B **68** 411–36
[5] Chopin N *et al* 2020 *An Introduction to Sequential Monte Carlo* vol 4 (Springer)
[6] Tebaldi C, Smith R L, Nychka D and Mearns L O 2005 Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles *J. Clim.* **18** 1524–40
[7] Gayrard V, Bovier A, Eckhoff M and Klein M 2004 Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times *J. Eur. Math. Soc.* **6** 399–424
[8] Gayrard V, Bovier A and Klein M 2005 Metastability in reversible diffusion processes ii: precise asymptotics for small eigenvalues *J. Eur. Math. Soc.* **7** 69–99
[9] Sen M K and Stoffa P L 2013 *Global Optimization Methods in Geophysical Inversion* (Cambridge University Press)
[10] Schneider T, Lan S, Stuart A and Teixeira J 2017 Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations *Geophys. Res. Lett.* **44** 12–396
[11] Peskin C S 1977 Numerical analysis of blood flow in the heart *J. Comput. Phys.* **25** 220–52
[12] Huang D Z, De Santis D and Farhat C 2018 A family of position-and orientation-independent embedded boundary methods for viscous flow and fluid–structure interaction problems *J. Comput. Phys.* **365** 74–104
[13] Huang D Z, Avery P, Farhat C, Rabinovitch J, Derkevorkian A and Peterson L D 2020 Modeling, simulation and validation of supersonic parachute inflation dynamics during Mars landing *AIAA Scitech 2020 Forum* p 0313
[14] Cao S and Zhengyu Huang D 2022 Bayesian calibration for large-scale fluid structure interaction problems under embedded/immersed boundary framework *Int. J. Numer. Methods Eng.* **123** 1791–812
[15] Berger M J *et al* 1989 Local adaptive mesh refinement for shock hydrodynamics *J. Comput. Phys.* **82** 64–84
[16] Borker R, Huang D, Grimberg S, Farhat C, Avery P and Rabinovitch J 2019 Mesh adaptation framework for embedded boundary methods for computational fluid dynamics and fluid-structure interaction *Int. J. Numer. Methods Fluids* **90** 389–424
[17] Moës N, Dolbow J and Belytschko T 1999 A finite element method for crack growth without remeshing *Int. J. Numer. Methods Eng.* **46** 131–50

[18] Tan Z, Kaul C M, Pressel K G, Cohen Y, Schneider T and ao Teixeira J 2018 An extended eddy-diffusivity mass-flux scheme for unified representation of subgrid-scale turbulence and convection *J. Adv. Modeling Earth Syst.* **10** 770–800
[19] Lopez-Gomez I, Christopoulos C D, Ludvig Ervik H, Dunbar O R, Cohen Y and Schneider T 2022 Training physics-based machine-learning parameterizations with gradient-free ensemble kalman methods *J. Adv. Modeling Earth Syst.* **14** e2022MS003105
[20] Garcia Trillos N and Sanz-Alonso D 2020 The Bayesian update: variational formulations and gradient flows *Bayesian Anal.* **15** 29–56
[21] Garcia Trillos N, Hosseini B and Sanz-Alonso D 2023 From optimization to sampling through gradient flows *Not. Am. Math. Soc.* **70** 1
[22] Chen Y, Zhengyu Huang D, Huang J, Reich S and Stuart A M 2023 Sampling via gradient flows in the space of probability measures (arXiv:2310.03597)
[23] Yulong L, Jianfeng L and Nolen J 2019 Accelerating langevin sampling with birth-death (arXiv:1905.09863)
[24] Yulong L, Slepčev D and Wang L 2022 Birth-death dynamics for sampling: global convergence, approximations and their asymptotics (arXiv:2211.00450)
[25] Domingo-Enrich C and Pooladian A-A 2023 An explicit expansion of the kullback-leibler divergence along its fisher-rao gradient flow (arXiv:2302.12229)
[26] Tan L and Jianfeng L 2023 Accelerate Langevin sampling with birth-death process and exploration component (arXiv:2305.05529)
[27] Chen Y, Zhengyu Huang D, Huang J, Reich S and Stuart A M 2023 Gradient flows for sampling: mean-field models, Gaussian approximations and affine invariance (arXiv:2302.11024)
[28] Chen Y and Wuchen Li 2020 Optimal transport natural gradient for statistical manifolds with continuous sample space *Inform. Geom.* **3** 1–32
[29] Lambert M, Chewi S, Bach F, Bonnabel S'ere and Rigollet P 2022 Variational inference via wasserstein gradient flows (arXiv:2205.15902)
[30] Amari S-I 1998 Natural gradient works efficiently in learning *Neural Comput.* **10** 251–76
[31] Martens J 2020 New insights and perspectives on the natural gradient method *J. Mach. Learn. Res.* **21** 5776–851 (available at: http://jmlr.org/papers/v21/17-678.html)
[32] Zhang G, Martens J and Grosse R B 2019 Fast convergence of natural gradient descent for over-parameterized neural networks *Advances in Neural Information Processing Systems* p 32 (available at: https://proceedings.neurips.cc/paper_files/paper/2019/file/1da546f25222c1ee710cf7e2f7a3ff0c-Paper.pdf)
[33] Lin W, Emtiyaz Khan M and Schmidt M 2019 Fast and simple natural-gradient variational inference with mixture of exponential-family approximations *Int. Conf. on Machine Learning* (PMLR) pp 3992–4002 (available at: https://proceedings.mlr.press/v97/lin19b.html)
[34] Huix T, Korba A, Durmus A and Moulines E 2024 Theoretical guarantees for variational inference with fixed-variance mixture of gaussians (arXiv:2406.04012)
[35] Chen Y and Oliver D S 2012 Ensemble randomized maximum likelihood method as an iterative ensemble smoother *Math. Geosci.* **44** 1–26
[36] Emerick A A and Reynolds A C 2013 Investigation of the sampling performance of ensemble-based methods with a simple reservoir model *Comput. Geosci.* **17** 325–50
[37] Iglesias M A, Law K J H and Stuart A M 2013 Ensemble Kalman methods for inverse problems *Inverse Problems* **29** 045001
[38] Pathiraja S and Reich S 2019 Discrete gradients for computational bayesian inference (arXiv:1903.00186)
[39] Chada N K, Stuart A M and Tong X T 2020 Tikhonov regularization within ensemble Kalman inversion *SIAM J. Numer. Anal.* **58** 1263–94
[40] Schneider T, Stuart A M and Jin-Long W 2020 Ensemble Kalman inversion for sparse learning of dynamical systems from time-averaged data (arXiv:2007.06175)
[41] Zhengyu Huang D, Schneider T and Stuart A M 2022 Iterated kalman methodology for inverse problems *J. Comput. Phys.* **463** 111262
[42] Calvello E, Reich S and Stuart A M 2022 Ensemble Kalman methods: a mean field perspective (arXiv:2209.11371)
[43] Zhengyu Huang D, Huang J, Reich S and Stuart A M 2022 Efficient derivative-free Bayesian inference for large-scale inverse problems *Inverse Problems* **38** 125006
[44] Doucet A *et al* 2009 A tutorial on particle filtering and smoothing: fifteen years later *Handbook of Nonlinear Filtering* **12** 3

[45] Klebanov I and John Sullivan T 2023 Transporting higher-order quadrature rules: Quasi-monte carlo points and sparse grids for mixture distributions (arXiv:2308.10081)
[46] Maurais A and Marzouk Y 2024 Sampling in unit time with kernel fisher-rao flow (arXiv:2401.03892)
[47] Nüsken N 2024 Stein transport for Bayesian inference (arXiv:2409.01464)
[48] Jordan M I, Ghahramani Z, Jaakkola T S and Saul L K 1999 An introduction to variational methods for graphical models *Mach. Learn.* **37** 183–233
[49] Wainwright M J *et al* 2008 Graphical models, exponential families and variational inference *Found. Trends Mach. Learn.* **1** 1–305
[50] Blei D M, Kucukelbir A and McAuliffe J D 2017 Variational inference: a review for statisticians *J. Am. Stat. Assoc.* **112** 859–77
[51] Quiroz M, Nott D J and Kohn R 2018 Gaussian variational approximation for high-dimensional state space models (arXiv:1801.07873)
[52] Khan M and Lin W 2017 Conjugate-computation variational inference: converting variational inference in non-conjugate models to inferences in conjugate models *Artificial Intelligence and Statistics* (PMLR) pp 878–87
[53] Galy-Fajou T, Perrone V and Opper M 2021 Flexible and efficient inference with particles for the variational Gaussian approximation *Entropy* **23** 990
[54] Lasser C and Lubich C 2020 Computing quantum dynamics in the semiclassical regime *Acta Numer.* **29** 229–401
[55] Anderson W and Farazmand M 2024 Fisher information and shape-morphing modes for solving the fokker–planck equation in higher dimensions *Appl. Math. Comput.* **467** 128489
[56] Zhang H, Chen Y, Vanden-Eijnden E and Peherstorfer B 2024 Sequential-in-time training of nonlinear parametrizations for solving time-dependent partial differential equations (arXiv:2404.01145)
[57] Rao C R 1992 Information and the accuracy attainable in the estimation of statistical parameters *Breakthroughs in Statistics* (Springer) pp 235–47
[58] Cencov N N 2000 *Statistical Decision Rules and Optimal Inference* vol 53 (American Mathematical Soc.)
[59] Nihat A, Jost J, Le H V and Schwachhöfer L 2015 Information geometry and sufficient statistics *Probab. Theory Relat. Fields* **162** 327–64
[60] Bauer M Bruveris M and Michor P W 2016 Uniqueness of the Fisher–Rao metric on the space of smooth densities *Bull. London Math. Soc.* **48** 499–506
[61] Yan Y, Wang K and Rigollet P 2023 Learning Gaussian mixtures using the wasserstein-fisher-rao gradient flow (arXiv:2301.01766)
[62] Wang L and Nüsken N 2024 Measure transport with kernel mean embeddings (arXiv:2401.12967)
[63] Ernst O G, Sprungk B and Starkloff H-J 2015 Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems *SIAM/ASA J. Uncertain. Quantification* **3** 823–51
[64] Garbuno-Inigo A, Hoffmann F, Wuchen Li and Stuart A M 2020 Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler *SIAM J. Appl. Dyn. Syst.* **19** 412–41
[65] Garbuno-Inigo A, Nüsken N and Reich S 2020 Affine invariant interacting Langevin dynamics for Bayesian inference *SIAM J. Appl. Dyn. Syst.* **19** 1633–58
[66] Alspach D and Sorenson H 1972 Nonlinear Bayesian estimation using Gaussian sum approximations *IEEE Trans. Autom. Control* **17** 439–48
[67] Ito K and Xiong K 2000 Gaussian filters for nonlinear filtering problems *IEEE Trans. Autom. Control* **45** 910–27
[68] Chen R and Liu J S 2000 Mixture kalman filters *J. R. Stat. Soc.* B **62** 493–508
[69] Reich S 2012 A Gaussian-mixture ensemble transform filter *Q. J. R. Meteorol. Soc.* **138** 222–33
[70] Ruoxia Li, Prasad V and Huang B 2016 Gaussian mixture model-based ensemble kalman filtering for state and parameter estimation for a PMMA process *Processes* **4** 9
[71] Fan R, Huang R and Diao R 2018 Gaussian mixture model-based ensemble kalman filter for machine parameter calibration *IEEE Trans. Energy Convers.* **33** 1597–9
[72] Grana D, Fjeldstad T and Omre H 2017 Bayesian Gaussian mixture linear inversion for geophysical inverse problems *Math. Geosci.* **49** 493–515
[73] Yuming B and Jiang L 2022 A residual-driven adaptive Gaussian mixture approximation for Bayesian inverse problems *J. Comput. Appl. Math.* **399** 113707

[74] Van Der Merwe R and Wan E 2003 Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models *2003 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2003 (Proc..(ICASSP'03)* vol 6 (IEEE) p VI–701
[75] Smith K W 2007 Cluster ensemble kalman filter *Tellus* A **59** 749–57
[76] Stordal A S, Karlsen H A, Nævdal G, Skaug H J and Vallès B 2011 Bridging the ensemble kalman filter and particle filters: the adaptive Gaussian mixture filter *Comput. Geosci.* **15** 293–305
[77] Hoteit I, Luo X and Pham D-T 2012 Particle Kalman filtering: a nonlinear Bayesian framework for ensemble kalman filters *Mon. Weather Rev.* **140** 528–42
[78] Frei M and Künsch H R 2013 Mixture ensemble kalman filters *Comput. Stat. Data Anal.* **58** 127–38
[79] Bengtsson T, Snyder C and Nychka D 2003 Toward a nonlinear ensemble filter for high-dimensional systems *J. Geophys. Res.* **108** 8775
[80] Sun A Y, Morris A P and Mohanty S 2009 Sequential updating of multimodal hydrogeologic parameter fields using localization and clustering techniques *Water Resour. Res.* **45** W07424
[81] Stordal A S, Karlsen H A, Nævdal G, Oliver D S and Skaug H J 2012 Filtering with state space localized kalman gain *Physica* D **241** 1123–35
[82] Carrillo J A, Chen Y, Zhengyu Huang D, Huang J and Wei D 2024 Fisher-rao gradient flow: geodesic convexity and functional inequalities (arXiv:2407.15693)
[83] Reich S 2011 A dynamical systems framework for intermittent data assimilation *BIT Numer. Math.* **51** 235–49
[84] Gelman A and Meng X-L 1998 Simulating normalizing constants: from importance sampling to bridge sampling to path sampling *Stat. Sci.* **13** 163–85
[85] Neal R M 2001 Annealed importance sampling *Stat. Comput.* **11** 125–39
[86] Chen H and Ying L 2024 Ensemble-based annealed importance sampling (arXiv:2401.15645)
[87] Chopin N, Crucinio F R and Korba A 2023 A connection between tempering and entropic mirror descent (arXiv:2310.11914)
[88] Goodman J and Weare J 2010 Ensemble samplers with affine invariance *Commun. Appl. Math. Comput. Sci.* **5** 65–80
[89] Foreman-Mackey D, Hogg D W, Lang D and Goodman J 2013 EMCEE: the MCMC hammer *Publ. Astron. Soc. Pac.* **125** 306
[90] Pavliotis G A, Stuart A M and Vaes U 2022 Derivative-free Bayesian inversion using multiscale dynamics *SIAM J. Appl. Dyn. Syst.* **21** 284–326
[91] Reich S and Weissmann S 2021 Fokker–Planck particle systems for Bayesian inference: computational approaches *SIAM/ASA J. Uncertain. Quantification* **9** 446–82
[92] Liu Q and Wang D 2016 Stein variational gradient descent: a general purpose Bayesian inference algorithm *Advances in Neural Information Processing Systems* p 29 (available at: https://proceedings.neurips.cc/paper_files/paper/2016/hash/b3ba8f1bee1238a2f37603d90b58898d-Abstract.html)