Fast, Multimodal, Derivative-Free Bayes Inference

with Fisher-Rao Gradient Flows

Yifan Chen

Courant Institute, New York University

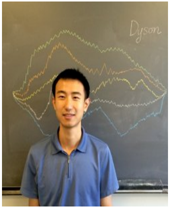
Applied Mathematics Colloquium, Columbia University

Relevant Papers

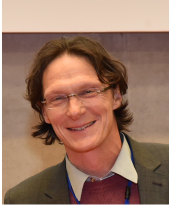
[Chen, Huang, Huang, Reich, Stuart 2023, 2024]

1. Sampling via gradient flows in the space of probability measures. `https://arxiv.org/abs/2310.03597`
2. Efficient, multimodal, and derivative-free Bayesian inference with Fisher-Rao gradient flows. `https://arxiv.org/abs/2406.17263`

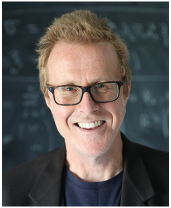
Daniel Huang
Peking University



Jiaoyang Huang
University of Pennsylvania



Sebastian Reich
University of Potsdam



Andrew Stuart
Caltech

Context

The sampling problem

Goal: draw (approximate) samples from

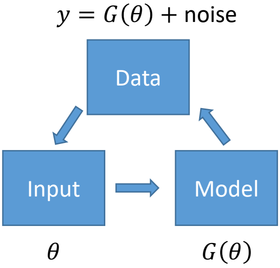
$$\rho^\star(\theta) \propto \exp(-V(\theta))$$

Set-up: $V(\theta)$ **available**, *versus samples in generative modeling*

Many applications in

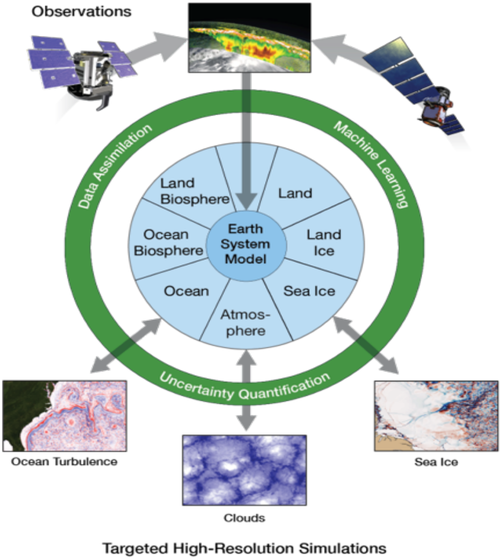
- Statistical physics
- Bayes inverse problems

  $\rho^\star(\theta) = \rho_{\text{post}}(\theta) \propto \rho(y|\theta)\rho_{\text{prior}}(\theta)$
- ...



One Particular Motivation: Climate Science

Next generation earth system model



Challenges

Bayes inverse problem under Gaussian priors and noises:

$$\rho_{\text{post}}(\theta) \propto \rho(y|\theta)\rho_{\text{prior}}(\theta) \propto \exp(-\Phi_R(\theta, y))$$

where $\Phi_R(\theta, y) = \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - G(\theta))\|^2 + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2$

1. Evaluating G is expensive: require large scale PDE solvers
2. Posterior distribution $\rho_{\text{post}}(\theta)$ can have multiple modes
3. Gradient of Φ_R may not available or even feasible

Challenges

Bayes inverse problem under Gaussian priors and noises:

$$\rho_{\text{post}}(\theta) \propto \rho(y|\theta)\rho_{\text{prior}}(\theta) \propto \exp(-\Phi_R(\theta, y))$$

where $\Phi_R(\theta, y) = \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - G(\theta))\|^2 + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2$

1. Evaluating G is expensive: require large scale PDE solvers
2. Posterior distribution $\rho_{\text{post}}(\theta)$ can have multiple modes
3. Gradient of Φ_R may not available or even feasible

Ask for fast, multimodal, and derivative-free Bayes sampler

Typical Sampling Approaches

Common structures of many sampling algorithms

1. Design a dynamics of ρ_t converging to (approximate) ρ_{post}
2. Develop a "numerical scheme" that implements the dynamics

Typical Sampling Approaches

Common structures of many sampling algorithms

1. Design a dynamics of ρ_t converging to (approximate) ρ_{post}
2. Develop a "numerical scheme" that implements the dynamics

- **Sequential Monte Carlo (SMC)**
  - Finite time dynamics such as $\rho_t \propto \rho_{\mathrm{prior}}^{1-t} \rho_{\mathrm{post}}^t$
  - E.g., implemented via importance sampling or ensembles
- **Markov Chain Monte Carlo (MCMC)**
  - Infinite time dynamics with $\rho_\infty = \rho_{\mathrm{post}}$
  - E.g., implemented via Markov chains or ensembles
- **Variational inference (VI), Kalman filter, ...**
  - Dynamics in a parametric family of distributions $\rho_t \in \mathcal{P}_\theta$
  - E.g., implemented via update of parameters or ensembles

MCMC: [Brooks, Galin, Jones, Meng, 2011], ...
SMC: [Del Moral, Doucet, Jasra, 2006], ...
Variational inference: [Mackay 2008], [Wainright, Jordan 2008], ...

Towards Fast, Multimodal, Derivative-Free Sampler?

Common structures of many sampling algorithms

1. Design a dynamics of ρ_t converging to (approximate) ρ_{post}
2. Develop a "numerical scheme" that implements the dynamics

- Dynamics of ρ_t needs to converge fast
  - Typical MCMC needs $O(10^4)$ runs
  - Many dynamics converges slowly in the case of multiple modes

- Dynamics amenable to derivative free numerical approximation
  - Small number of forward map evaluations in each iteration
  - Vanilla SMC may suffer from weight collapse

Towards Fast, Multimodal, Derivative-Free Sampler?

Common structures of many sampling algorithms

1. Design a dynamics of ρ_t converging to (approximate) ρ_{post}
2. Develop a "numerical scheme" that implements the dynamics

- Dynamics of ρ_t needs to converge fast
  - Typical MCMC needs $O(10^4)$ runs
  - Many dynamics converges slowly in the case of multiple modes

- Dynamics amenable to derivative free numerical approximation
  - Small number of forward map evaluations in each iteration
  - Vanilla SMC may suffer from weight collapse

Our proposal of algorithms

Fisher-Rao gradient flow w/ Gaussian mixture + Kalman approx.

Towards Fast, Multimodal, Derivative-Free Sampler

1. Fisher-Rao Gradient Flow for Efficiency
2. Gaussian Mixture + Kalman for Multimodal and Derivative-Free
3. Theoretical Insights
4. Numerical Demonstrations

Towards Efficient, Multimodal, Derivative-Free Sampler

1 Fisher-Rao Gradient Flow for Efficiency

2 Gaussian Mixture + Kalman for Multimodal and Derivative-Free

3 Theoretical Insights

4 Numerical Demonstrations

Fisher-Rao Gradient Flow

Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho_{\text{post}} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\text{post}} - \log \rho_t]$$

Fisher-Rao Gradient Flow

Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho_{\rm post} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\rm post} - \log \rho_t]$$

- **KL divergence**

$$\mathcal{E}(\rho) = \mathrm{KL}[\rho \| \rho_{\rm post}] = \int \rho \log\Big(\frac{\rho}{\rho_{\rm post}}\Big)\,\mathrm{d}\theta$$

- **Fisher-Rao metric tensor**

$$M(\rho)^{-1}\psi = \rho(\psi - \mathbb{E}_\rho[\psi])$$

- **The gradient flow equation**

$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1}\frac{\delta \mathcal{E}}{\delta \rho}\Big|_{\rho=\rho_t} = -M(\rho_t)^{-1}(\log \rho_t - \log \rho_{\rm post})$$

Information geometry [Amari 2016], [Ay, Jost, Lê, Schwachhöfer, 2017]
See also: Wasserstein gradient flow, Stein variational gradient flow, ...

Properties of Fisher-Rao Gradient Flow

Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho_{\text{post}} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\text{post}} - \log \rho_t]$$

Property (1): Apply any diffeomorphism $\varphi : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_\theta}$

- $\tilde{\rho}_t = \varphi\#\rho_t$ is the transformed distribution at time t
- $\tilde{\rho}_{\text{post}} = \varphi\#\rho_{\text{post}}$ is the transformed target distribution

Then, the form of the flow equation remains invariant

$$\frac{\partial \tilde{\rho}_t}{\partial t} = \tilde{\rho}_t\big(\log \tilde{\rho}_{\text{post}} - \log \tilde{\rho}_t\big) - \tilde{\rho}_t \mathbb{E}_{\tilde{\rho}_t}[\log \tilde{\rho}_{\text{post}} - \log \tilde{\rho}_t]$$

Properties of Fisher-Rao Gradient Flow

Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho_{\text{post}} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\text{post}} - \log \rho_t]$$

Property (1): Apply any diffeomorphism $\varphi : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_\theta}$

- $\tilde{\rho}_t = \varphi \# \rho_t$ is the transformed distribution at time t
- $\tilde{\rho}_{\text{post}} = \varphi \# \rho_{\text{post}}$ is the transformed target distribution

Then, the form of the flow equation remains invariant

$$\frac{\partial \tilde{\rho}_t}{\partial t} = \tilde{\rho}_t\big(\log \tilde{\rho}_{\text{post}} - \log \tilde{\rho}_t\big) - \tilde{\rho}_t \mathbb{E}_{\tilde{\rho}_t}[\log \tilde{\rho}_{\text{post}} - \log \tilde{\rho}_t]$$

Note: Invariance is useful for fast convergence of dynamics

- Affine invariant MCMC [Goodman, Weare 2010]
- Preconditioned Langevin, Kalman-Wasserstein gradient flow [Reich Cotter 2015], [Leimkuhler, Matthews, Weare 2018], [Garbuno-Inigo, Hoffmann, Li, Stuart 2020]

Convergence of Fisher-Rao gradient flows of KL divergence

[Chen, Huang, Huang, Reich, Stuart 2023]

Let ρ_t satisfy the Fisher-Rao gradient flow. Assume

- there exist constants $K, B > 0$ such that ρ_0 satisfies

$$e^{-K(1+|\theta|^2)} \leq \rho_0(\theta)/\rho_{\text{post}}(\theta) \leq e^{K(1+|\theta|^2)}$$

- the second moments of $\rho_0, \rho_{\text{post}}$ are both bounded by B

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

$$\text{KL}[\rho_t \| \rho_{\text{post}}] \leq (2 + B + eB)Ke^{-t}$$

See also: **[Lu, Slepčev, Wang 2022], [Domingo-Enrich, Pooladian 2023]**

Convergence of Fisher-Rao gradient flows of KL divergence

[Chen, Huang, Huang, Reich, Stuart 2023]

Let ρ_t satisfy the Fisher-Rao gradient flow. Assume

- there exist constants $K, B > 0$ such that ρ_0 satisfies

$$e^{-K(1+|\theta|^2)} \le \rho_0(\theta)/\rho_{\text{post}}(\theta) \le e^{K(1+|\theta|^2)}$$

- the second moments of $\rho_0, \rho_{\text{post}}$ are both bounded by B

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

$$\mathrm{KL}[\rho_t \| \rho_{\text{post}}] \le (2 + B + eB)Ke^{-t}$$

See also: **[Lu, Slepčev, Wang 2022], [Domingo-Enrich, Pooladian 2023]**

"Unconditional" uniform exponential convergence

- In sharp contrast to Wasserstein gradient flows and Langvin dynamics whose convergence rates depend on ρ_{post} (e.g., log-concavity, or log-Sobolev constants)

  [Jordan, Kinderlehrer, Otto 1998], [Villani 2003, 2008], ...

Properties of Fisher-Rao Gradient Flow

Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho_{\text{post}} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\text{post}} - \log \rho_t]$$

Property (2): independent of the normalization consts of ρ_{post}

- Useful for the numerical implementation of the dynamics
- No need to worry about the approximation of the normalization constant

Properties (1) (2) Are Special

Unique property of Fisher-Rao metric

[Cencov 2000], [Ay, Jost, Lê, Schwachhöfer 2015], [Bauer, Bruveris, Michor 2016]

The Fisher-Rao metric is the only Riemannian metric on smooth positive densities (up to scaling) that is invariant under any diffeomorphism of the parameter space

Unique property of KL divergence

[Chen, Huang, Huang, Reich, Stuart 2023]

Among all f-divergence with continuously differentiable f, KL divergence is the only one, up to scaling, whose induced gradient flow under any metric is invariant to the normalization consts of ρ_{post}

Fisher-Rao gradient flow is special in the context of sampling

Exploration-Exploitation Scheme for Fisher-Rao GFs

Continuous Fisher-Rao gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \rho_t\big(\log \rho_{\text{post}} - \log \rho_t\big) - \rho_t \mathbb{E}_{\rho_t}[\log \rho_{\text{post}} - \log \rho_t]$$

Discrete scheme via operator splitting

$$\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t} \quad \text{(exploration)}$$
$$\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\rho_{\text{post}}(\theta)^{\Delta t} \quad \text{(exploitation)}$$

- Exploration steps connected to tempering/annealing
- Fixed point interpretation [Huang, Huang, Reich, Stuart 2022]
- Mirror descent interpretation [Chopin, Crucinio, Korba 2023]
- Compared to dynamics in SMC: additional exploration step
- Compared to dynamics in MCMC: exponential convergence
    - unconditional convergence also holds in the discrete level

Towards Efficient, Multimodal, Derivative-Free Sampler

1 Fisher-Rao Gradient Flow for Efficiency

2 Gaussian Mixture + Kalman for Multimodal and Derivative-Free

3 Theoretical Insights

4 Numerical Demonstrations

Numerical Approximation of Fisher-Rao Gradient Flow

Particle methods (i.e. Diracs ansatz)

- Birth-death dynamics [Lu, Lu, Nolen 2019], [Lu, Slepčev, Wang 2022]
- Ensemble MCMC [Lindsey, Weare, Zhang 2021]

Need ways to move the support of the particles to explore the space and choices of smoothing kernels. Challenging in high dim space.

Our focus: parametric approximation (full support ansatz)

- Gaussian and mixture approximations
- Kalman methodology for derivative-free updates

Gaussian Approximation by Direct Projection

Gaussian approximate Fisher-Rao gradient flow

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

- Project the dynamics into Gaussian space
- Can also be obtained by moment closures
- Equivalent to natural gradient flow [Amari 1998] for Gaussian VI
- Gradient is needed (can be avoided by using Stein's lemma, but numerically we found it not very stable)

Gaussian Approximation by Kalman's Methodology

Discrete scheme of Fisher-Rao gradient flow

$$\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t} \quad \text{(exploration)}$$

$$\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\rho_{\text{post}}(\theta)^{\Delta t} \quad \text{(exploitation)}$$

- Current approximation $\rho_n(\theta) = \mathcal{N}(\theta; m_n, C_n)$
- **Prediction step**: $\hat{\rho}_{n+1}(\theta) = \mathcal{N}(\theta; m_n, \frac{1}{1-\Delta t}C_n)$

Gaussian Approximation by Kalman's Methodology

Discrete scheme of Fisher-Rao gradient flow

$$\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t} \quad \text{(exploration)}$$
$$\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\rho_{\text{post}}(\theta)^{\Delta t} \quad \text{(exploitation)}$$

- Current approximation $\rho_n(\theta) = \mathcal{N}(\theta; m_n, C_n)$
- **Prediction step**: $\hat{\rho}_{n+1}(\theta) = \mathcal{N}(\theta; m_n, \frac{1}{1-\Delta t}C_n)$
- **Analysis step**: $\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta) \exp(-\Delta t \Phi_R(\theta, y))$
  where $\Phi_R(\theta, y) = \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - G(\theta))\|^2 + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2$

Gaussian Approximation by Kalman's Methodology

Discrete scheme of Fisher-Rao gradient flow

$$\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t} \quad \text{(exploration)}$$
$$\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\rho_{\text{post}}(\theta)^{\Delta t} \quad \text{(exploitation)}$$

- Current approximation $\rho_n(\theta) = \mathcal{N}(\theta; m_n, C_n)$
- **Prediction step**: $\hat{\rho}_{n+1}(\theta) = \mathcal{N}(\theta; m_n, \frac{1}{1-\Delta t}C_n)$
- **Analysis step**: $\rho_{n+1}(\theta) \propto \hat{\rho}_{n+1}(\theta)\exp(-\Delta t \Phi_R(\theta, y))$
  where $\Phi_R(\theta, y) = \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - G(\theta))\|^2 + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2$
- Consider $x = F(\theta) + \nu$ with $\theta \sim \hat{\rho}_{n+1}, \nu \sim \mathcal{N}(0, \frac{\Sigma_\nu}{\Delta t})$

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

Then $\rho(\theta|x) = \dfrac{\rho(\theta)\rho(x|\theta)}{\rho(x)} \propto \rho(\theta)\exp(-\Delta t \Phi_R(\theta)) = \rho_{n+1}(\theta)$

Kalman Filter Type Approximation

- Gaussian moment closure of **joint states and observations**

$$\rho^{\mathrm{G}}(\theta, x) \sim \mathcal{N}\Big(\begin{bmatrix} \widehat{m}_{n+1} \\ \hat{x}_{n+1} \end{bmatrix}, \begin{bmatrix} \widehat{C}_{n+1} & \widehat{C}^{\theta x}_{n+1} \\ \widehat{C}^{\theta x^T}_{n+1} & \widehat{C}^{xx}_{n+1} \end{bmatrix} \Big)$$

w/ $\hat{x}_{n+1} = \mathbb{E}[F(\theta)], \widehat{C}^{\theta x}_{n+1} = \mathrm{Cov}[\theta, F(\theta)], \widehat{C}^{xx}_{n+1} = \mathrm{Cov}[F(\theta)] + \frac{\Sigma_\nu}{\Delta t}$
these integrals are approximated by quadratures

Kalman Filter Type Approximation

- Gaussian moment closure of joint states and observations

$$\rho^{\mathrm{G}}(\theta, x) \sim \mathcal{N}\Big(\begin{bmatrix} \widehat{m}_{n+1} \\ \hat{x}_{n+1} \end{bmatrix}, \begin{bmatrix} \widehat{C}_{n+1} & \widehat{C}^{\theta x}_{n+1} \\ \widehat{C}^{\theta x^T}_{n+1} & \widehat{C}^{xx}_{n+1} \end{bmatrix} \Big)$$

w/ $\hat{x}_{n+1} = \mathbb{E}[F(\theta)], \widehat{C}^{\theta x}_{n+1} = \mathrm{Cov}[\theta, F(\theta)], \widehat{C}^{xx}_{n+1} = \mathrm{Cov}[F(\theta)] + \frac{\Sigma_\nu}{\Delta t}$
these integrals are approximated by quadratures

- Gaussian conditional approximations

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

$$m_{n+1} = \widehat{m}_{n+1} + \widehat{C}^{\theta x}_{n+1}(\widehat{C}^{xx}_{n+1})^{-1}(x - \hat{x}_{n+1})$$
$$C_{n+1} = \widehat{C}_{n+1} - \widehat{C}^{\theta x}_{n+1}(\widehat{C}^{xx}_{n+1})^{-1}(\widehat{C}^{\theta x}_{n+1})^T$$

which is derivative free

EnKF, EKI: [Evensen 1994], [Iglesias, Law, Stuart 2013], ...

UKF, UKI: [Julier, Uhlmann, and Durrant-Whyte 1994], [Wan, Van Der Merwe 2000], [Huang, Huang, Reich, Stuart 2022], ...

Gaussian Mixtures with Kalman's Methodology

The Gaussian mixture ansatz

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

Prediction step:

- $\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t} \propto \sum_{k=1}^{K} w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})\rho_n(\theta)^{-\Delta t}$

Gaussian Mixtures with Kalman's Methodology

The Gaussian mixture ansatz

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

Prediction step:

- $\hat{\rho}_{n+1}(\theta) \propto \rho_n(\theta)^{1-\Delta t} \propto \sum_{k=1}^{K} w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})\rho_n(\theta)^{-\Delta t}$

Gaussian moment closure for each component

- $w_{n,k}\mathcal{N}(\theta; m_{n,k}, C_{n,k})\rho_n(\theta) \approx \hat{w}_{n+1,k}\mathcal{N}(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k})$ achieved by numerical quadratures
- Normalize weights $\hat{w}_{n+1,k}$ to sum to 1
- Then $\hat{\rho}_{n+1}(\theta) \approx \sum_{k=1}^{K} \hat{w}_{n+1,k}\mathcal{N}(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k})$

Gaussian Mixtures with Kalman's Methodology

Analysis step:

$$
\begin{aligned}
\rho_{n+1}(\theta) &\propto \hat{\rho}_{n+1}(\theta)\rho_{\mathrm{post}}(\theta)^{\Delta t} \\
&\approx \sum_{k=1}^{K} \hat{w}_{n+1,k}\mathcal{N}(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k})\rho_{\mathrm{post}}(\theta)^{\Delta t}
\end{aligned}
$$

Gaussian Mixtures with Kalman's Methodology

Analysis step:

$$\begin{aligned}\rho_{n+1}(\theta) &\propto \hat{\rho}_{n+1}(\theta)\rho_{\text{post}}(\theta)^{\Delta t} \\ &\approx \sum_{k=1}^{K} \hat{w}_{n+1,k}\mathcal{N}(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k})\rho_{\text{post}}(\theta)^{\Delta t}\end{aligned}$$

Kalman filter type approx. for each component

$$\hat{w}_{n+1,k}\mathcal{N}(\theta; \widehat{m}_{n+1,k}, \widehat{C}_{n+1,k})\rho_{\text{post}}(\theta)^{\Delta t} \approx w_{n+1,k}\mathcal{N}(\theta; m_{n+1,k}, C_{n+1,k})$$

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

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

Different to many Gaussian mixture Kalman filter and sequential Monte Carlo approach, the algorithm here has an exploration component

Towards Efficient, Multimodal, Derivative-Free Sampler

1 Fisher-Rao Gradient Flow for Efficiency

2 Gaussian Mixture + Kalman for Multimodal and Derivative-Free

3 Theoretical Insights

4 Numerical Demonstrations

Continuous limit of Fisher-Rao with Gaussian mixture + Kalman

$$
\begin{aligned}
\dot{m}_{t,k} =& - C_{t,k} \int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \nabla_\theta \log \rho_t \mathrm{d}\theta + \widehat{C}^{\theta x}_{t,k} \Sigma_\nu^{-1} (x - \hat{x}_{t,k}) \\
\dot{C}_{t,k} =& - C_{t,k} \left(\int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \nabla_\theta \nabla_\theta \log \rho_t \mathrm{d}\theta \right) C_{t,k} \\
& - \widehat{C}^{\theta x}_{t,k} \Sigma_\nu^{-1} {\widehat{C}^{\theta x}_{t,k}}^T \\
\dot{w}_{t,k} =& - w_{t,k} \int (\mathcal{N}(\theta; m_{t,k}, C_{t,k}) - \rho_t)(\log \rho_t - \log \rho_{\mathrm{post}}) \mathrm{d}\theta
\end{aligned}
$$

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

$$
\hat{x}_{t,k} = \mathbb{E}[F(\theta)], \ \widehat{C}^{\theta x}_{t,k} = \mathrm{Cov}[\theta, F(\theta)], \ \text{with} \ \ \theta \sim \mathcal{N}(m_{t,k}, C_{t,k})
$$

Continuous limit of Fisher-Rao with Gaussian mixture + Kalman

$$
\begin{aligned}
\dot{m}_{t,k} &= -C_{t,k}\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\nabla_\theta \log \rho_t \mathrm{d}\theta + \widehat{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}(x - \hat{x}_{t,k}) \\
\dot{C}_{t,k} &= -C_{t,k}\left(\int \mathcal{N}(\theta; m_{t,k}, C_{t,k})\nabla_\theta\nabla_\theta \log \rho_t \mathrm{d}\theta\right)C_{t,k} \\
&\quad - \widehat{C}^{\theta x}_{t,k}\Sigma_\nu^{-1}\widehat{C}^{\theta x\,T}_{t,k} \\
\dot{w}_{t,k} &= -w_{t,k}\int (\mathcal{N}(\theta; m_{t,k}, C_{t,k}) - \rho_t)(\log \rho_t - \log \rho_{\mathrm{post}})\mathrm{d}\theta
\end{aligned}
$$

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

$$
\hat{x}_{t,k} = \mathbb{E}[F(\theta)],\ \widehat{C}^{\theta x}_{t,k} = \mathrm{Cov}[\theta, F(\theta)],\ \text{with}\ \ \theta \sim \mathcal{N}(m_{t,k}, C_{t,k})
$$

- Without red terms, entropy always increases, i.e., exploration

$$
\frac{\mathrm{d}}{\mathrm{d}t}\int -\rho_t \log \rho_t \geq 0
$$

- Red terms depend on posterior information

Gradient flow of KL divergence with respect to GMM parameters

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

$$\dot{C}_{t,k} = -C_{t,k} \Big(\int \mathcal{N}(\theta; m_{t,k}, C_{t,k}) \big(\nabla_\theta \nabla_\theta \log \rho_t - \nabla_\theta \nabla_\theta \log \rho_{\mathrm{post}} \big) \mathrm{d}\theta \Big) C_{t,k}$$

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

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

- Let $a = \{w_k, m_k, C_k : 1 \leq k \leq K\}$

  $$\frac{\mathrm{d}a}{\mathrm{d}t} = -(\tilde{\mathrm{FI}}(a))^{-1} \nabla_a \mathrm{KL}[\sum_{k=1}^K w_k \mathcal{N}(m_k, C_k) \| \rho_{\mathrm{post}}]$$

  $\tilde{\mathrm{FI}}(a)$: diagonal approximations of Fisher information matrix

- Thus, our method replaces red terms involving derivatives of ρ_{post} by $\widehat{C}_{t,k}^{\theta x} \Sigma_\nu^{-1} (x - \hat{x}_{t,k})$,$\widehat{C}_{t,k}^{\theta x} \Sigma_\nu^{-1} \widehat{C}_{t,k}^{\theta x^T}$
- This derivative free approx. is exact for Gaussian posterior

  Statistical linearization [Calvello, Reich, Stuart 2022]

Implications and Properties of The Algorithm

- **Gradient flow structure** regarding the KL divergence

$$\mathrm{KL}[\rho \| \rho_{\mathrm{post}}] = \int \rho \log \rho - \int \rho \log \rho_{\mathrm{post}}$$

- Mode repulsion and exploration effects due to entropy term
- Fast exploitation of Gaussian-like modes

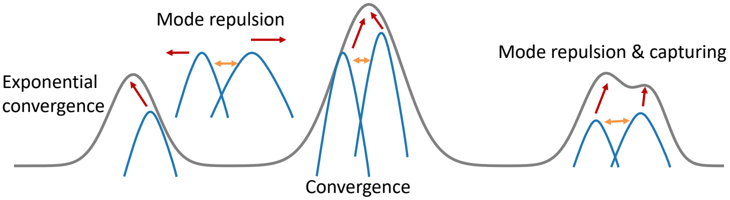


Figure: Conceptual properties of our algorithm

Towards Efficient, Multimodal, Derivative-Free Sampler

1 Fisher-Rao Gradient Flow for Efficiency

2 Gaussian Mixture + Kalman for Multimodal and Derivative-Free

3 Theoretical Insights

4 Numerical Demonstrations

Algorithm Complexity Analysis

Setting: number of mixtures: K; number of iterations: N

- **Prediction step**: exploration, without evaluating forward map
- **Analysis step**: Gaussian integration for moment closures can be achieved by quadrature, e.g., by unscented transformation, require $(2d_\theta + 1)K$ forward map evaluation per step

Algorithmic complexity

- Number of forward map evaluation $(2d_\theta + 1)KN$
  In each iteration, $(2d_\theta + 1)K$ forward evaluations in parallel
- Arithmetic complexity: $O(d_\theta^3 KN)$
- In our experiments: $N = O(10)$ suffices to work
- K selected by the user

Numerical Study

We present two experimental results

1. One-dim bimodal synthetic problem
2. 128-dim bimodal problem in Navier Stokes equations

We use $\Delta t = 0.5$, and run $N = 30$ iterations

We term our algorithm GMKI (Gaussian mixture Kalman inversion)

One-dimensional Bimodal Problem

Consider the 1D inverse problem

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

The prior is $\rho_{\text{prior}} \sim \mathcal{N}(3, 2^2)$.

Different noise levels:

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

where the overlap between these two modes becomes larger, when the noise level increases

One-dimensional Bimodal Problem: Case A

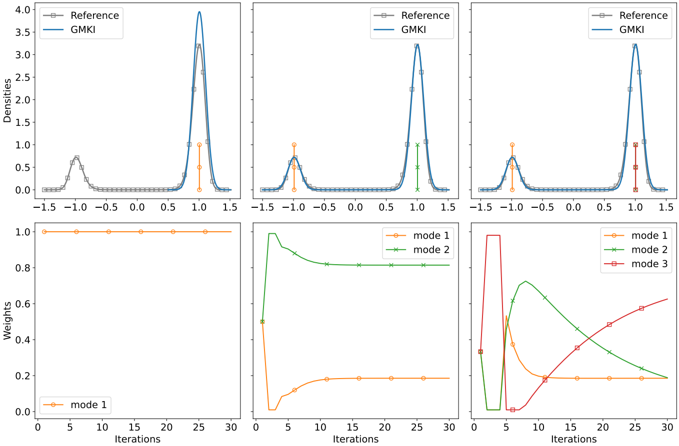


Figure: One-dimensional bimodal problem with $\Sigma_\eta = 0.2^2$. Top row: posterior distributions estimated by random walk MCMC (black bins) and GMKI (blue lines) at the 30-th iteration obtained by 1-modal GMKI, 2-modal GMKI and 3-modal GMKI (from left to right); Mean estimation of each mode is marked. Bottom row: weight estimations obtained by 1-modal GMKI, 2-modal GMKI and 3-modal GMKI

One-dimensional Bimodal Problem: Case B

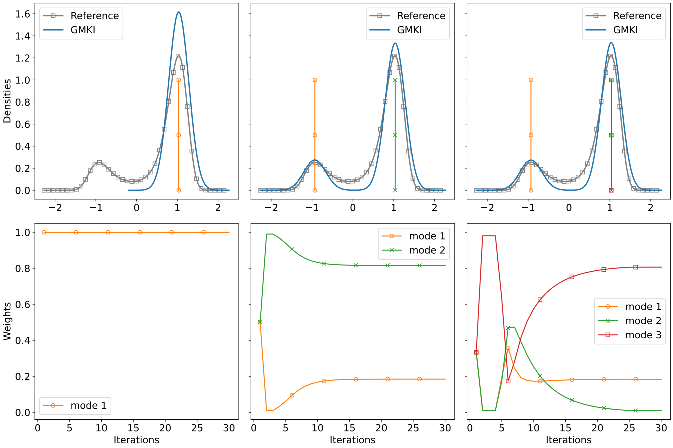


Figure: One-dimensional bimodal problem with $\Sigma_\eta = 0.5^2$. Top row: posterior distributions estimated by random walk MCMC (black bins) and GMKI (blue lines) at the 30-th iteration obtained by 1-modal GMKI, 2-modal GMKI and 3-modal GMKI (from left to right); Mean estimation of each mode is marked. Bottom row: weight estimations obtained by 1-modal GMKI, 2-modal GMKI and 3-modal GMKI

One-dimensional Bimodal Problem: Case C

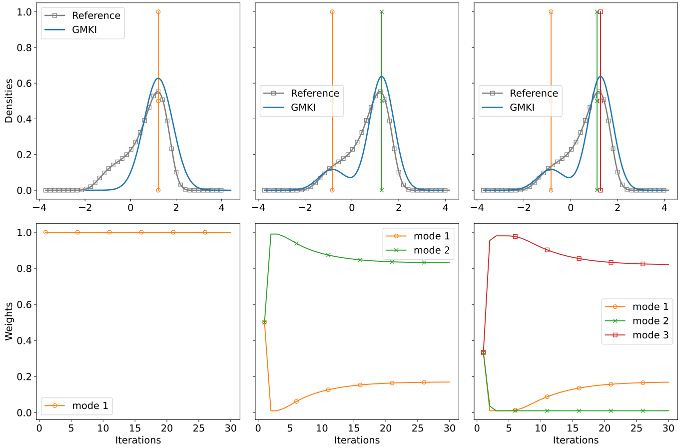


Figure: One-dimensional bimodal problem with $\Sigma_\eta = 1.5^2$. Top row: posterior distributions estimated by random walk MCMC (black bins) and GMKI (blue lines) at the 30-th iteration obtained by 1-modal GMKI, 2-modal GMKI and 3-modal GMKI (from left to right); Mean estimation of each mode is marked. Bottom row: weight estimations obtained by 1-modal GMKI, 2-modal GMKI and 3-modal GMKI

High-dimensional Bimodal Problem

Consider 2d NSE on a periodic domain $D = [0, 2\pi] \times [0, 2\pi]$

$$\frac{\partial \omega}{\partial t} + (v \cdot \nabla)\omega - \nu \Delta \omega = \nabla \times f$$

- Viscosity $\nu = 0.01$
- Non-zero mean background velocity $v_b = [0, 2\pi]$
- $f(x_1, x_2) = [0, \cos(4x_1)]$
- **Goal:** learn initial vorticity based on observed vorticity at some observation points at later times $T = 0.25, 0.5$
- Gaussian process prior on initial vorcitity (we keep the first 128 Karhunen-Loève expansion coefficients and use data to learn these coefficients $\theta \in \mathbb{R}^{128}$)

Multimodal Setting: Symmetry in Observations

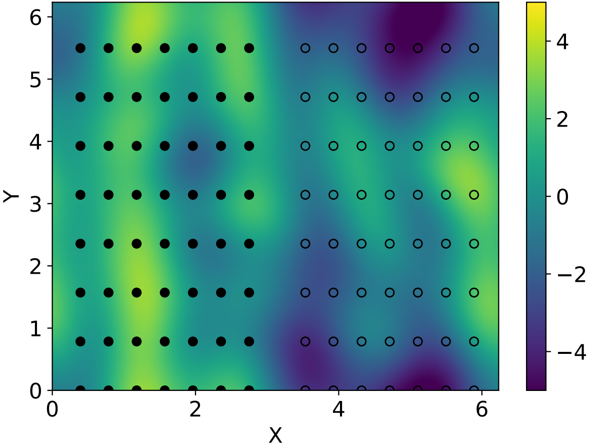


Figure: Vorticity observations $\omega([x_1, x_2]) - \omega([2\pi - x_1, x_2])$ at 56 equidistant points (solid black dots)

Results for Learning Initial Vorticity in 2D NSE: $K = 3$

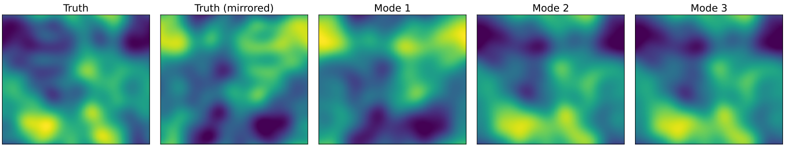


Figure: The true vorticity field, and these modes obtained by GMKI

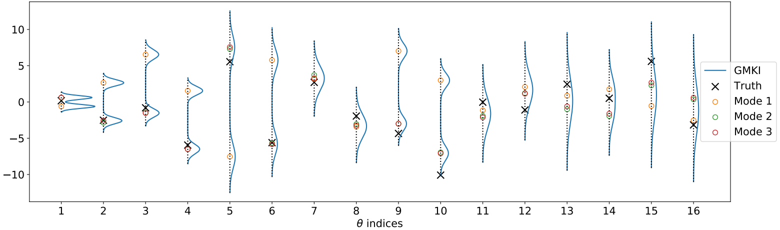


Figure: The truth KL expansion coefficients θ_i (black crosses), and mean estimations of θ_i for each modes (circles) and the associated marginal distributions obtained GMKI at the 30th iteration

Summary

Towards fast, multimodal, derivative-free Bayes sampler

- **Dynamics:** Fisher-Rao gradient flow of KL divergence
  - Unconditional exponential convergence
  - A special gradient flow for sampling
  - Connections to SMC, MCMC, annealing/tempering
- **Approximations:** Gaussian mixture + Kalman methods
  - Gaussian moment closures in joint state and observations
  - Gradient flow structure in GMM parameter space
  - Mode repulsion and fast convergence for each mode
- **Future works**: theoretical analysis and refined approximations

Thank You!