

# Gradient Flows for Sampling

Energy Functionals, Invariance and Gaussian Approximation

Yifan Chen

Applied and Computational Mathematics, Caltech

*Illinois Institute of Technology, July 2023*

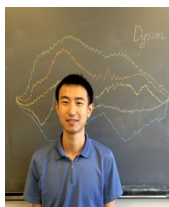
# The Paper

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

Gradient flows for sampling:  
Mean-field models, Gaussian approximations and affine invariance



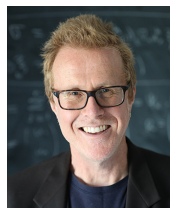
Daniel Huang  
Caltech



Jiaoyang Huang  
University of  
Pennsylvania



Sebastian Reich  
University of  
Potsdam



Andrew Stuart  
Caltech

Link: <https://arxiv.org/abs/2302.11024>.

# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows
- 3 On Choosing Energy Functionals
- 4 On Choosing Metrics
- 5 On Gaussian Approximation
- 6 Conclusions

# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows
- 3 On Choosing Energy Functionals
- 4 On Choosing Metrics
- 5 On Gaussian Approximation
- 6 Conclusions

# Context

## The sampling problem

Goal: draw (approximate) samples from

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

# Context

## The sampling problem

Goal: draw (approximate) samples from

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

*Set-up: **assuming**  $V(\theta)$  **available**, in contrast to generative modeling*

# Context

## The sampling problem

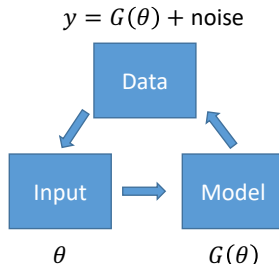
Goal: draw (approximate) samples from

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

Set-up: **assuming**  $V(\theta)$  **available**, in contrast to generative modeling

Many applications in

- Uncertainty quantification
- Bayes inverse problems
- Filtering
- Active learning
- ...



# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows**
- 3 On Choosing Energy Functionals
- 4 On Choosing Metrics
- 5 On Gaussian Approximation
- 6 Conclusions



# Methodology

## Dynamics for sampling

Idea: construct a **dynamics of**  $\rho_t$  that gradually converges to

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

*Note: for simplicity we consider continuous-time*

## Dynamics for sampling

Idea: construct a **dynamics of**  $\rho_t$  that gradually converges to

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

*Note: for simplicity we consider continuous-time*

- **Finite time dynamics**  $\rho_1 = \rho^*$ , from a given  $\rho_0$  (e.g. prior)
  - Sequential Monte Carlo, e.g.,  $\rho_t \propto \exp(-tV(\theta))$ , ...

## Dynamics for sampling

Idea: construct a **dynamics of  $\rho_t$**  that gradually converges to

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

*Note: for simplicity we consider continuous-time*

- **Finite time dynamics**  $\rho_1 = \rho^*$ , from a given  $\rho_0$  (e.g. prior)
  - Sequential Monte Carlo, e.g.,  $\rho_t \propto \exp(-tV(\theta))$ , ...
- **Infinite time dynamics**  $\rho_\infty = \rho^*$ , from arbitrary  $\rho_0$ 
  - MCMC, Langevin's dynamics, ...

## Dynamics for sampling

Idea: construct a **dynamics of  $\rho_t$**  that gradually converges to

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

*Note: for simplicity we consider continuous-time*

- **Finite time dynamics**  $\rho_1 = \rho^*$ , from a given  $\rho_0$  (e.g. prior)
  - Sequential Monte Carlo, e.g.,  $\rho_t \propto \exp(-tV(\theta))$ , ...
- **Infinite time dynamics**  $\rho_\infty = \rho^*$ , from arbitrary  $\rho_0$ 
  - MCMC, Langevin's dynamics, ...

The focus of this talk: **Infinite time dynamics**

# Dynamics through Gradient Flows (GFs)

## Gradient flow dynamics for sampling

Idea: construct a **gradient flow dynamics of**  $\rho_t$  that converges to

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

*Namely, dynamics comes from gradient based optimization methods*

# Dynamics through Gradient Flows (GFs)

## Gradient flow dynamics for sampling

Idea: construct a **gradient flow dynamics of  $\rho_t$**  that converges to

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

*Namely, dynamics comes from gradient based optimization methods*

- Langevin's dynamics and Wasserstein GFs  
[Jordan, Kinderlehrer, Otto 1998], ...
- Stein variational GD and Stein variational GFs  
[Liu, Wang 2016], [Liu 2017], ...
- Interaction between optimization and sampling  
[Wibisono 2018], ...
- A recent review paper  
[Trillos, Hosseini, Sanz-Alonso 2023]
- ...

# Gradient Flows

## Ingredients in gradient flows

Formally: ( $\mathcal{P}$  is the space of probability densities)

- **An energy functional** to minimize

$$\mathcal{E} : \mathcal{P} \rightarrow \mathbb{R}$$

- **A metric** for descent direction

$$g_\rho : T_\rho \mathcal{P} \times T_\rho \mathcal{P} \rightarrow \mathbb{R}, \quad g_\rho(\sigma_1, \sigma_2) = \langle M(\rho)\sigma_1, \sigma_2 \rangle_{L^2}$$

$$\implies \text{Flow: } \frac{\partial \rho_t}{\partial t} = -\nabla_g \mathcal{E}(\rho_t) = -M(\rho_t)^{-1} \frac{\delta \mathcal{E}}{\delta \rho} \Big|_{\rho=\rho_t}$$

- $T_\rho \mathcal{P}$  (tangent space) is the space of measures integrated to 0
- $\frac{\delta \mathcal{E}}{\delta \rho}$  is the first variation of  $\mathcal{E}$  at  $\rho$
- $M(\rho_t)^{-1}$  can be understood as a **preconditioner**

# Sampling through Numerical Approximation of GFs

## Gradient flow equation

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

**Numerical approximations** of GFs lead to sampling methods

- Particle methods such as SDEs

$$d\theta_t = f(\theta_t; \rho_t, \rho^*)dt + h(\theta_t; \rho_t, \rho^*)dW_t$$

e.g., Langevin's dynamics  $d\theta_t = \nabla_{\theta} \log \rho^*(\theta_t)dt + \sqrt{2}dW_t$

- Parametric approximations such as Gaussian approximation  
e.g., Gaussian variational inference, Kalman filters, ...



# The Focus of this Talk

## The question:

Any guiding principles for designing  $\mathcal{E}$  and  $M(\rho)$ ?

# The Focus of this Talk

## The question:

Any guiding principles for designing  $\mathcal{E}$  and  $M(\rho)$ ?

We approach the question through the **perspective of invariance**

- In **energy functionals**: invariance to normalization consts
- In **metrics**: invariance to transformation of the space

We then discuss **numerics** of the resulting flow

# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows
- 3 On Choosing Energy Functionals**
- 4 On Choosing Metrics
- 5 On Gaussian Approximation
- 6 Conclusions

# On Choosing the Energy Functionals

## Recap: Gradient flow equation

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

# On Choosing the Energy Functionals

## Recap: Gradient flow equation

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

- Most popular choice of  $\mathcal{E}(\rho)$ : Kullback–Leibler divergence

$$\mathcal{E}(\rho; \rho^\star) = \text{KL}[\rho \parallel \rho^\star] = \int \rho \log \left( \frac{\rho}{\rho^\star} \right) d\theta$$

# On Choosing the Energy Functionals

## Recap: Gradient flow equation

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

- Most popular choice of  $\mathcal{E}(\rho)$ : Kullback–Leibler divergence

$$\mathcal{E}(\rho; \rho^\star) = \text{KL}[\rho \| \rho^\star] = \int \rho \log \left( \frac{\rho}{\rho^\star} \right) d\theta$$

- Property:  $\mathcal{E}(\rho; c\rho^\star) = \mathcal{E}(\rho; \rho^\star) - \log c$  for any  $c \in \mathbb{R}_+$

# On Choosing the Energy Functionals

## Recap: Gradient flow equation

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

- Most popular choice of  $\mathcal{E}(\rho)$ : Kullback–Leibler divergence

$$\mathcal{E}(\rho; \rho^\star) = \text{KL}[\rho \| \rho^\star] = \int \rho \log \left( \frac{\rho}{\rho^\star} \right) d\theta$$

- Property:  $\mathcal{E}(\rho; c\rho^\star) = \mathcal{E}(\rho; \rho^\star) - \log c$  for any  $c \in \mathbb{R}_+$   
 $\Rightarrow$  first variation  $\frac{\delta \mathcal{E}}{\delta \rho}$  is independent of  $c$

# On Choosing the Energy Functionals

## Recap: Gradient flow equation

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

- Most popular choice of  $\mathcal{E}(\rho)$ : Kullback–Leibler divergence

$$\mathcal{E}(\rho; \rho^\star) = \text{KL}[\rho \| \rho^\star] = \int \rho \log \left( \frac{\rho}{\rho^\star} \right) d\theta$$

- Property:  $\mathcal{E}(\rho; c\rho^\star) = \mathcal{E}(\rho; \rho^\star) - \log c$  for any  $c \in \mathbb{R}_+$ 
  - $\Rightarrow$  first variation  $\frac{\delta \mathcal{E}}{\delta \rho}$  is independent of  $c$
  - $\Rightarrow$  the gradient flow equation is independent of  $c$



# On Choosing the Energy Functionals

## Recap: Gradient flow equation

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

- Most popular choice of  $\mathcal{E}(\rho)$ : Kullback–Leibler divergence

$$\mathcal{E}(\rho; \rho^\star) = \text{KL}[\rho \parallel \rho^\star] = \int \rho \log \left( \frac{\rho}{\rho^\star} \right) d\theta$$

- Property:  $\mathcal{E}(\rho; c\rho^\star) = \mathcal{E}(\rho; \rho^\star) - \log c$  for any  $c \in \mathbb{R}_+$ 
  - $\Rightarrow$  first variation  $\frac{\delta \mathcal{E}}{\delta \rho}$  is independent of  $c$
  - $\Rightarrow$  the gradient flow equation is independent of  $c$

**Implication:** no need to worry about **normalization consts** of  $\rho^\star$

### The question

Any other choices of  $\mathcal{E}$  that have such invariance property?

### The question

Any other choices of  $\mathcal{E}$  that have such invariance property?

The answer is **NO** among a large class of  $\mathcal{E}$

# KL Divergence is Special

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

Among all  $f$ -divergence with continuously differentiable  $f$ , KL divergence is the only one, up to scaling, whose first variation is invariant to the normalization const of  $\rho^*$

# KL Divergence is Special

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

Among all  **$f$ -divergence** with continuously differentiable  $f$ , KL divergence is the only one, up to scaling, whose first variation is **invariant to the normalization const of  $\rho^\star$**

- $f$ -divergence: for  $f(0) = 1$  and  $f$  convex

$$D_f[\rho \parallel \rho^\star] = \int \rho^\star f\left(\frac{\rho}{\rho^\star}\right) d\theta$$

- Kullback–Leibler divergence:  $f(x) = x \log x$
- $\chi^2$  divergence:  $f(x) = (x - 1)^2$
- Hellinger distance:  $f(x) = (\sqrt{x} - 1)^2$
- ...

# KL Divergence is Special

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

Among all  **$f$ -divergence** with continuously differentiable  $f$ , KL divergence is the only one, up to scaling, whose first variation is **invariant to the normalization const of  $\rho^*$**

- $f$ -divergence: for  $f(0) = 1$  and  $f$  convex

$$D_f[\rho \parallel \rho^*] = \int \rho^* f\left(\frac{\rho}{\rho^*}\right) d\theta$$

- Kullback–Leibler divergence:  $f(x) = x \log x$
- $\chi^2$  divergence:  $f(x) = (x - 1)^2$
- Hellinger distance:  $f(x) = (\sqrt{x} - 1)^2$
- ...

Use KL divergence from now on

# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows
- 3 On Choosing Energy Functionals
- 4 On Choosing Metrics**
- 5 On Gaussian Approximation
- 6 Conclusions

# Two Metrics

## Wasserstein metric [Jordan, Kinderlehrer, Otto 1998]

$$\text{Metric: } M(\rho)^{-1}\psi = -\nabla \cdot (\rho \nabla \psi)$$

$$\text{Flow: } \frac{\partial \rho_t}{\partial t} = -\nabla_{\theta} \cdot (\rho_t \nabla_{\theta} \log \rho^*) + \nabla \cdot (\nabla \rho_t)$$

$$\text{SDEs: } d\theta_t = \nabla_{\theta} \log \rho^* dt + \sqrt{2} dW_t$$

- Optimal transport [Villani 2003, 2008]

## Fisher-Rao metric [Rao 1945]

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

$$\text{Flow: } \frac{\partial \rho_t}{\partial t} = \rho_t(\log \rho^* - \log \rho_t) - \rho_t \mathbb{E}_{\rho_t}[\log \rho^* - \log \rho_t]$$

- Information geometry [Amari 2016], [Ay, Jost, Lê, Schwachhöfer, 2017]



# Two Metrics

## Wasserstein metric [Jordan, Kinderlehrer, Otto 1998]

Metric:  $M(\rho)^{-1}\psi = -\nabla \cdot (\rho \nabla \psi)$

Flow:  $\frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot (\rho_t \nabla_\theta \log \rho^*) + \nabla \cdot (\nabla \rho_t)$

SDEs:  $d\theta_t = \nabla_\theta \log \rho^* dt + \sqrt{2}dW_t$

- Optimal transport [Villani 2003, 2008]

## Fisher-Rao metric [Rao 1945]

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

Flow:  $\frac{\partial \rho_t}{\partial t} = \rho_t(\log \rho^* - \log \rho_t) - \rho_t \mathbb{E}_{\rho_t}[\log \rho^* - \log \rho_t]$

- Information geometry [Amari 2016], [Ay, Jost, Lê, Schwachhöfer, 2017]

# Convergence Property of Wasserstein Gradient Flow

## Theorem [Markowich, Villani 2000]

Assume  $\exists \lambda > 0$  such that

$$-D^2 \log \rho^*(\cdot) \succeq \lambda I$$

Then, for all  $t \geq 0$ ,

$$\text{KL}[\rho_t \| \rho^*] \leq \text{KL}[\rho_0 \| \rho^*] e^{-2\lambda t}$$

Rate of exponential convergence depends on  $\rho^*$

# Two Metrics

## Wasserstein metric [Jordan, Kinderlehrer, Otto 1998]

$$\text{Metric: } M(\rho)^{-1}\psi = -\nabla \cdot (\rho \nabla \psi)$$

$$\text{Flow: } \frac{\partial \rho_t}{\partial t} = -\nabla_{\theta} \cdot (\rho_t \nabla_{\theta} \log \rho^*) + \nabla \cdot (\nabla \rho_t)$$

$$\text{SDEs: } d\theta_t = \nabla_{\theta} \log \rho^* dt + \sqrt{2} dW_t$$

- Optimal transport [Villani 2003, 2008]

## Fisher-Rao metric [Rao 1945]

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

$$\text{Flow: } \frac{\partial \rho_t}{\partial t} = \rho_t(\log \rho^* - \log \rho_t) - \rho_t \mathbb{E}_{\rho_t}[\log \rho^* - \log \rho_t]$$

- Information geometry [Amari 2016], [Ay, Jost, Lê, Schwachhöfer, 2017]

# A Closer Look at Fisher-Rao

## Fisher-Rao gradient flow

$$\frac{\partial \rho_t}{\partial t} = \rho_t (\log \rho^\star - \log \rho_t) - \rho_t \mathbb{E}_{\rho_t} [\log \rho^\star - \log \rho_t]$$

**Apply transformation** of any **diffeomorphism**  $\varphi : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}^{d_\theta}$

- $\tilde{\rho}_t = \varphi \# \rho_t$  is the transformed distribution at time  $t$
- $\tilde{\rho}^\star = \varphi \# \rho^\star$  is the transformed target distribution

Then, the form of the flow equation remains **invariant**

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

# Why Care About Invariance?

## Implication of invariance

Convergence rates of the gradient flow are **the same** for  
general  $\rho^*$  and **Gaussian**  $\rho^*$

# Why Care About Invariance?

## Implication of invariance

Convergence rates of the gradient flow are **the same** for general  $\rho^\star$  and **Gaussian**  $\rho^\star$

- Assume there exists a diffeomorphism  $\varphi$  such that

$$\tilde{\rho}^\star = \varphi \# \rho^\star = \text{Gaussian}$$

- Recall the property of the KL divergence

$$\text{KL}[\rho_t \| \rho^\star] = \text{KL}[\varphi \# \rho_t \| \varphi \# \rho^\star] = \text{KL}[\tilde{\rho}_t \| \tilde{\rho}^\star]$$

Thus, a general  $\rho^\star$  problem  $\sim$  a simpler **Gaussian**  $\rho^\star$  problem

# Theoretical Results of Fisher-Rao

## Convergence of Fisher-Rao gradient flows

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

Let  $\rho_t$  satisfy the Fisher-Rao gradient flow. Assume

- there exist constants  $K, B > 0$  such that  $\rho_0$  satisfies

$$e^{-K(1+|\theta|^2)} \leq \rho_0(\theta)/\rho^\star(\theta) \leq e^{K(1+|\theta|^2)}$$

- the second moments of  $\rho_0, \rho^\star$  are both bounded by  $B$

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

$$\text{KL}[\rho_t \parallel \rho^\star] \leq (2 + B + eB)Ke^{-t}$$

See also: [Lu, Slepčev, Wang 2022]

# Theoretical Results of Fisher-Rao

## Convergence of Fisher-Rao gradient flows

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

Let  $\rho_t$  satisfy the Fisher-Rao gradient flow. Assume

- there exist constants  $K, B > 0$  such that  $\rho_0$  satisfies

$$e^{-K(1+|\theta|^2)} \leq \rho_0(\theta)/\rho^\star(\theta) \leq e^{K(1+|\theta|^2)}$$

- the second moments of  $\rho_0, \rho^\star$  are both bounded by  $B$

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

$$\text{KL}[\rho_t \parallel \rho^\star] \leq (2 + B + eB)Ke^{-t}$$

See also: [Lu, Slepčev, Wang 2022]

**“Unconditional” uniform exponential convergence**

- In sharp contrast to **Wasserstein gradient flows** whose convergence rates depend on  $\rho^\star$



# Further Thoughts

## Simulating the Fisher-Rao GF is not as straightforward

- Birth-death dynamics, Wasserstein-Fisher-Rao gradient flow  
[Lu, Lu, Nolen 2019], [Lu, Slepčev, Wang 2022]
- Gaussian approximation [Chen, Huang, Huang, Reich, Stuart 2023]  
Derivative-free Kalman method [Huang, Huang, Reich, Stuart 2022]

*We will talk about it later ...*

# Further Thoughts

## Simulating the Fisher-Rao GF is not as straightforward

- Birth-death dynamics, Wasserstein-Fisher-Rao gradient flow  
[Lu, Lu, Nolen 2019], [Lu, Slepčev, Wang 2022]
- Gaussian approximation [Chen, Huang, Huang, Reich, Stuart 2023]  
Derivative-free Kalman method [Huang, Huang, Reich, Stuart 2022]

*We will talk about it later ...*

*At first, let's ask a basic question*

### The question:

Any other choices of metric having such invariance property?

# Further Thoughts

## Simulating the Fisher-Rao GF is not as straightforward

- Birth-death dynamics, Wasserstein-Fisher-Rao gradient flow [Lu, Lu, Nolen 2019], [Lu, Slepčev, Wang 2022]
- Gaussian approximation [Chen, Huang, Huang, Reich, Stuart 2023]  
Derivative-free Kalman method [Huang, Huang, Reich, Stuart 2022]

*We will talk about it later ...*

*At first, let's ask a basic question*

### The question:

Any other choices of metric having such invariance property?

The answer is again, **NO**

# Fisher-Rao Metric is 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.

# Fisher-Rao Metric is 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.

*No other alternatives if we ask for **diffeomorphism invariance**!*

# Fisher-Rao Metric is 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.

*No other alternatives if we ask for **diffeomorphism invariance**!*

*But can ask for a relaxed **affine invariance***

- Affine invariant MCMC [Goodman, Weare 2010]
- Kalman-Wasserstein gradient flows  
[Garbuno-Inigo, Hoffmann, Li, Stuart 2020]
- Other affine invariant gradient flow examples in our paper
  - e.g., affine invariant Stein gradient flow

# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows
- 3 On Choosing Energy Functionals
- 4 On Choosing Metrics
- 5 On Gaussian Approximation**
- 6 Conclusions

# Numerical Approximation of the Fisher-Rao Gradient Flow

## Related literature:

- Birth-death dynamics, Wasserstein-Fisher-Rao gradient flow  
[Lu, Lu, Nolen 2019], [Lu, Slepčev, Wang 2022]
- **Gaussian approximation** [Chen, Huang, Huang, Reich, Stuart 2023]  
Derivative-free Kalman method [Huang, Huang, Reich, Stuart 2022]

The focus of this talk: **Gaussian approximation**



# Gaussian Approximation by Moment Closures

## The general procedures:

- Consider any dynamics in the **density space**

$$\frac{\partial \rho_t(\theta)}{\partial t} = \sigma_t(\theta, \rho_t)$$

- Write down the dynamics of the **mean and covariance**

$$\begin{aligned}\frac{dm_t}{dt} &= \int \sigma_t(\theta, \rho_t) \theta d\theta \\ \frac{dC_t}{dt} &= \int \sigma_t(\theta, \rho_t) (\theta - m_t)(\theta - m_t)^T d\theta\end{aligned}$$

- Closure: replace  $\rho_t$  in the above RHS by  $\rho_{a_t} = \mathcal{N}(m_t, C_t)$   
Notation:  $a_t = (m_t, C_t)$

References: Moment closure in variational Kalman filtering [Särkkä, 2007], and in Wasserstein gradient flow [Lambert, Chewi, Bach, Bonnabel, Rigollet 2022]

# Gaussian Approximation by Moment Closures

## Gaussian approximate Fisher-Rao gradient flow

$$\frac{dm_t}{dt} = C_t \mathbb{E}_{\rho_{a_t}} [\nabla_{\theta} \log \rho^*],$$

$$\frac{dC_t}{dt} = C_t + C_t \mathbb{E}_{\rho_{a_t}} [\nabla_{\theta} \nabla_{\theta} \log \rho^*] C_t$$

# Gaussian Approximation by Moment Closures

## Gaussian approximate Fisher-Rao gradient flow

$$\begin{aligned}\frac{dm_t}{dt} &= C_t \mathbb{E}_{\rho_{a_t}} [\nabla_{\theta} \log \rho^*], \\ \frac{dC_t}{dt} &= C_t + C_t \mathbb{E}_{\rho_{a_t}} [\nabla_{\theta} \nabla_{\theta} \log \rho^*] C_t\end{aligned}$$

- Derived using Stein's lemma
- Equivalent to **natural gradient** flow [Amari 1998] for

Gaussian variational inference:  $\min_{m, C} \text{KL}[\mathcal{N}(m, C) \parallel \rho^*]$

Key: Fisher information matrix is used for preconditioning

$$\frac{d}{dt}(m_t, C_t) = -\text{FI}(m_t, C_t)^{-1} \nabla_{m_t, C_t} \text{KL}$$

# Convergence Guarantee [Chen, Huang, Huang, Reich, Stuart 2023]

## Gaussian target

If  $\rho^\star = \mathcal{N}(m_\star, C_\star)$ , and  $C_0 = \lambda_0 I$ ,  $\lambda_0 > 0$ , then

$$\|m_t - m_\star\|_2 = \mathcal{O}(e^{-t}), \quad \|C_t - C_\star\|_2 = \mathcal{O}(e^{-t})$$

- Same story, due to **invariance** property

# Convergence Guarantee [Chen, Huang, Huang, Reich, Stuart 2023]

## Logconcave target

Assume

- $\alpha I \preceq -\nabla_{\theta} \nabla_{\theta} \log \rho^* \preceq \beta I$
- $\lambda_{0,\min} I \preceq C_0 \preceq \lambda_{0,\max} I$

Then

$$\text{KL}[\rho_{a_t} \|\rho^*] - \text{KL}[\rho_{a_{\star}} \|\rho^*] \leq e^{-Kt} (\text{KL}[\rho_{a_0} \|\rho^*] - \text{KL}[\rho_{a_{\star}} \|\rho^*])$$

where

- $a_t = (m_t, C_t), \rho_{a_t} = \mathcal{N}(m_t, C_t)$
- $a_{\star} = \operatorname{argmin}_a \text{KL}[\rho_a \|\rho^*]$
- $K = \alpha \min\{1/\beta, \lambda_{0,\min}\}$

- See also the proof for the case of Wasserstein gradient flow in Gaussian variational inference for logconcave target

[Lambert, Chewi, Bach, Bonnabel, Rigollet 2022]

# Local Convergence Rates: Linearized Analysis

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

Assume  $\alpha I \preceq -\nabla_{\theta} \nabla_{\theta} \log \rho^* \preceq \beta I$ . For  $N_{\theta} = 1$ , let  $\lambda_{\star, \max} < 0$  denote the largest eigenvalue of the linearized Jacobian matrix of the flow around  $a_{\star}$ . Then we have

$$-\lambda_{\star, \max} \geq \frac{1}{(7 + \frac{4}{\sqrt{\pi}})(1 + \log(\frac{\beta}{\alpha}))}$$

Moreover, the bound is sharp: it is possible to construct a sequence of triplets  $\rho_n^*$ ,  $\alpha_n$  and  $\beta_n$ , where  $\lim_{n \rightarrow \infty} \frac{\beta_n}{\alpha_n} = \infty$ , such that, if we let  $\lambda_{\star, \max, n}$  denote the corresponding largest eigenvalues of the linearized Jacobian matrix for the  $n$ -th triple, then, it holds that

$$-\lambda_{\star, \max, n} = \mathcal{O}\left(1 / \log \frac{\beta_n}{\alpha_n}\right)$$

Convergence rates only depend on  $\log(\text{condition number})$

# Numerical Examples

- **2D Convex Potential:**  $\theta = (\theta^{(1)}, \theta^{(2)})$

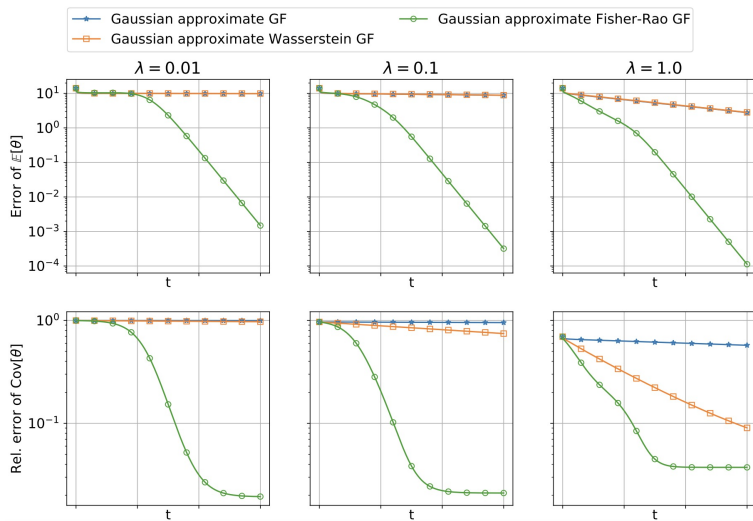
$$V(\theta) = \frac{(\sqrt{\lambda}\theta^{(1)} - \theta^{(2)})^2}{20} + \frac{(\theta^{(2)})^4}{20} \quad \text{with } \lambda = 0.01, 0.1, 1$$

- **Method:** Gaussian approximation of Fisher-Rao GF, Wasserstein GF and vallina GF
- **Configuration:** we initialize the Gaussian at

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}\right)$$

We integrate the mean and covariance dynamics to  $t = 15$

# Numerical Examples



**Figure:**  $x$  axis is from  $t = 0$  to 15. Convergence rate of Gaussian approximate Fisher-Rao gradient flows not influenced by values of  $\lambda$



# Outline

- 1 The Sampling Problem
- 2 The Methodology: Dynamics and Gradient Flows
- 3 On Choosing Energy Functionals
- 4 On Choosing Metrics
- 5 On Gaussian Approximation
- 6 Conclusions**

# Summary

## Gradient flows for sampling

- **Energy functional:** KL divergence is special
  - Invariance to normalization consts
- **Metric:** Fisher-Rao metric is special
  - Invariance to any diffeomorphism of the parameter space  
⇒ unconditional uniform exponential convergence
  - Relaxed to affine invariance and many constructions
- **Gaussian approximation via moment closures**
  - Equivalent to natural gradient in Gaussian VI
  - Convergence guarantee for Gaussian and logconcave targets
- **Further directions**
  - Prove optimal convergence rates in Gaussian VI
  - Gaussian mixture approximations for multimodal targets
  - Derivative free approximations via Kalman's methodology

# Thank You

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

Gradient flows for sampling:  
Mean-field models, Gaussian approximations and affine invariance

Link: <https://arxiv.org/abs/2302.11024>.