Sampling Via Gradient Flows In The Space of Probability Measures

Andrew Stuart, Caltech

Collaborators

**Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance**


Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, Andrew M. Stuart
Outline

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Outline

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Goal

The Sampling Problem

\( V : \mathbb{R}^d \rightarrow \mathbb{R} \). Draw (approximate) samples from

\[ \rho^*(\theta) \propto \exp\left(-V(\theta)\right) \]


Unifying Framework

Ingredients For Gradient Flows

- \( \mathcal{P} \) (Probability Densities on \( \mathbb{R}^d \))
- \( \mathcal{E} : \mathcal{P} \to \mathbb{R}^+, \mathcal{E}(\rho^*) = 0 \) (Energy Functional)
- \( T_\rho \mathcal{P} \) : signed measures integrating to 0 (Tangent Space)
- \( g_\rho : T_\rho \mathcal{P} \times T_\rho \mathcal{P} \to \mathbb{R}_+, \ g_\rho(\sigma_1, \sigma_2) = \langle M(\rho)\sigma_1, \sigma_2 \rangle_{L^2} \) (Metric)
- \( \frac{\delta \mathcal{E}}{\delta \rho} \) (First Variation)

The Gradient Flow in \( \mathcal{P} \)

\[
\frac{\partial \rho_t}{\partial t} = -\nabla g \mathcal{E}(\rho_t) = -M(\rho_t)^{-1} \frac{\delta \mathcal{E}}{\delta \rho} (\rho_t)
\]

Key Identity

\[
\frac{d}{dt} \mathcal{E}(\rho_t) = \left\langle \frac{\delta \mathcal{E}}{\delta \rho} (\rho_t), \frac{\partial \rho_t}{\partial t} \right\rangle = -\left\langle M(\rho_t) \frac{\partial \rho_t}{\partial t}, \frac{\partial \rho_t}{\partial t} \right\rangle \leq 0
\]


At Our Disposal: Energy Functional, Metric.
Optimization and Variational Bayes

Variational Bayes

Kullback–Leibler (KL) Divergence:

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

- \( \mathcal{E} : \mathcal{P} \to \mathbb{R}^+ \), \( \mathcal{E}(\rho^*) = 0 \)
- \( \rho^* = \arg\min_{\rho \in \mathcal{P}} \mathcal{E}(\rho) \)
- \( \mathcal{P}_a : \) parameterized subset of probability density functions on \( \mathbb{R}^d \), \( a \in \mathbb{R}^p \)
- \( \rho_{a^*} = \arg\min_{\rho \in \mathcal{P}_a} \mathcal{E}(\rho) \)

Example: Gaussian Variational Bayes

- \( \mathcal{G} : \) all Gaussian probability measures on \( \mathbb{R}^d \)
- \( a = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}_{\text{sym}, \geq 0} \)


Sampling and MCMC

Mean Field Models

- Evolution (often stochastic) for $\theta_t$; evolution depends on $\rho_t = \text{Law}(\theta_t)$
- Approximate the evolution with interacting particle system

Example: Fokker-Planck and Langevin Equations

KL for energy: $\mathcal{E} = \text{KL}[\rho\|\rho^*]$; Wasserstein–2 for metric; then:

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

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


Outline

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Choice of $\mathcal{E}$

### Recap: Gradient Flow

$$\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathcal{E}}{\delta \rho}(\rho_t; \rho^*)$$

### Energy: Kullback-Leibler

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

$$\frac{\delta \mathcal{E}}{\delta \rho}(\rho; \rho^*) = \log \rho - \log \rho^* + \text{constant}$$

- $\mathcal{E}(\rho; c\rho^*) = \mathcal{E}(\rho; \rho^*) - \log(c)$
- Hence first variation independent of normalization
- Hence gradient flow independent of normalization
### Choice of $\mathcal{E}$: Kullback–Leibler (KL) is Special

#### $f$-divergence

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

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

#### Examples

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

#### Theorem

Chen, Huang, Huang, Reich, AMS [8] (2023)

KL is the only $f$-divergence whose first variation leads to a gradient flow which is independent of the normalization constant of $\rho^*$
Outline

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Choice of Metric

Recap: Gradient Flow

\[
\frac{\partial \rho_t}{\partial t} = -M(\rho_t)^{-1} \frac{\delta \mathcal{E}}{\delta \rho}(\rho_t)
\]

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

\[
\frac{\delta \mathcal{E}}{\delta \rho} = \log \rho - \log \rho^* + \text{constant}
\]

Review paper: Trillos, Hosseini, Sanz-Alonso [27] (2023)


Others will be employed, detailed and cited in what follows
## Two Metrics

### Wasserstein Metric  

**Metric:**  
\[ M(\rho)\psi^{-1} = -\nabla_\theta \cdot (\rho \nabla_\theta \psi) \in T_\rho \mathcal{P} \]

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

**Mean Field Model:**  
\[ d\theta_t = \nabla_\theta \log \rho^*(\theta_t) dt + \sqrt{2} dW_t \]

### Fisher-Rao Metric  

**Metric:**  
\[ M(\rho)\psi^{-1} = \rho(\psi - E_\rho[\psi]) \in T_\rho \mathcal{P} \]

**Flow:**  
\[ \frac{\partial \rho_t}{\partial t} = \rho_t (\log \rho^* - \log \rho_t) - \rho_t E_\rho_t [\log \rho^* - \log \rho_t] \]

**Mean Field Model:**  
Discuss later

---


Fisher-Rao Flow: Invariance Under Diffeomorphisms

**Pushforward**

Given diffeomorphism \( \varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d \)
- \( \tilde{\rho}_t = \varphi \# \rho_t \) is the transformed distribution at time \( t \)
- \( \tilde{\rho}^* = \varphi \# \rho^* \) is the transformed target distribution

**Proposition**

Fisher-Rao gradient flow is invariant under any diffeomorphism:

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

\[
\frac{\partial \tilde{\rho}_t}{\partial t} = \tilde{\rho}_t \left( \log \tilde{\rho}^* - \log \tilde{\rho}_t \right) - \tilde{\rho}_t \mathbb{E}_{\tilde{\rho}_t} \left[ \log \tilde{\rho}^* - \log \tilde{\rho}_t \right]
\]
Consequence of Invariance of Fisher-Rao Gradient Flow

Proposition
For any density $\rho^*$ for which there exists a $\varphi$ such that

$$\tilde{\rho}^* = \varphi \# \rho^* = \mathcal{N}(0, I)$$

it follows that

$$\text{KL}[\rho_t \| \rho^*] = \text{KL}[\tilde{\rho}_t \| \tilde{\rho}^*]$$
Consequence of Invariance of Fisher-Rao Gradient Flow

Proposition

For any density $\rho^*$ for which there exists a $\varphi$ such that

$$\tilde{\rho}^* = \varphi \# \rho^* = \mathcal{N}(0, I)$$

it follows that

$$KL[\rho_t \| \rho^*] = KL[\tilde{\rho}_t \| \tilde{\rho}^*]$$

Theorem

Lu, Slepčev, Wang [18] (2022); Chen, Huang, Huang, Reich, AMS [8] (2023)

Assume

- $\exists K > 0$:
  
  $$e^{-K(1 + |\theta|^2)} \leq \frac{\rho_0(\theta)}{\rho^*(\theta)} \leq e^{K(1 + |\theta|^2)}$$

- $\exists B > 0$ bounding first and second moments of $\rho_0, \rho^*$

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

$$KL[\rho_t \| \rho^*] \leq (2 + B + eB)Ke^{-t}$$
Mean-Field ODE  
Chen, Huang, Huang, Reich, AMS [8] (2023)

\[
\frac{d \theta_t}{dt} = -\nabla_\theta F(\theta; \rho_t, \rho^*)|_{\theta=\theta_t} \\
-\nabla_\theta \cdot \left( \rho(\theta)\nabla_\theta F(\theta; \rho, \rho^*) \right) = \rho(\theta) \mathbb{E}_\rho \left( \log \rho^* - \log \rho \right) - \rho(\theta) \left( \log \rho^*(\theta) - \log \rho(\theta) \right)
\]

Particle approximation: \( \{\theta_{t,\ell}\}_{\ell=1}^N \)

Birth-Death Process  

\[
\Omega^\ell_t = \log \left( \frac{1}{N} \sum_{j=1}^N K(\theta_{t,\ell} - \theta_{t,j})/\rho^*(\theta_{t,\ell}) \right), \quad K \approx \delta
\]

\[
\Lambda^i_t = \Omega^i_t - \frac{1}{N} \sum_{\ell=1}^N \Omega^\ell_t \quad \text{Particle } i \text{ birth-death rate}
\]

- Both face significant obstacles in order to implement
Outline

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Invariance Revisited


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.

**Affine Invariance**

Given an affine transformation $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$

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

Flow is **affine invariant** if $\tilde{\rho}_t$ satisfies same equation as $\rho_t$ with $\rho^*$ replaced by $\tilde{\rho}^*$


Examples

**Fisher-Rao Gradient Flow**
The Fisher-Rao gradient flow is affine invariant

The Kalman-Wasserstein gradient flow is affine invariant.

\[
\text{Covariance: } C(\rho) = \text{Cov}(\rho) \\
\text{Metric: } M(\rho)^{-1}\psi = -\nabla_\theta \cdot (\rho C(\rho) \nabla_\theta \psi) \in T_\rho \mathcal{P} \\
\text{Flow: } \frac{\partial \rho_t}{\partial t} = -\nabla_\theta \cdot (\rho_t C(\rho_t) \nabla_\theta \log \rho^*) + \nabla_\theta \cdot (C(\rho_t) \nabla_\theta \rho_t) \\
\text{Mean Field Model: } d\theta_t = C(\rho_t) \nabla_\theta \log \rho^*(\theta_t) dt + \sqrt{2C(\rho_t)} dW_t
\]

Kalman-Wasserstein metric first identified: Reich and Cotter [22] (2015)
Numerical Example Illustrating Affine Invariance

Experimental Set-Up

- **2D Rosenbrock potential:**
  \[ V(\theta) = \frac{\lambda}{20} (\theta_2 - \theta_1^2)^2 + \frac{1}{20} (1 - \theta_1)^2 \]
  for \( \theta = (\theta_1, \theta_2) \) and \( \lambda = 10^{-k}, \ k = 0, 1, 2 \)

- **Goal:** sample \( \rho^* \propto \exp(-V(\theta)) \)

- **Method 1:** Wasserstein using noninteracting Langvein, \( 10^3 \) particles.

- **Method 2:** Kalman-Wasserstein using interacting Langevin, \( 10^3 \) particles

- **Configuration:** Integrate to \( t = 15 \), initialized from
  \[ \theta_0 \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \right) \]
Numerical Example Illustrating Affine Invariance

Figure: $10^3$ particles at $t = 15$ from Langevin (top row) and affine invariant Langevin (bottom row). Grey lines represent the contour of the true posterior.
Outline

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Gradient Descent for Variational Bayes

Ingredients For Gradient Flows

- $\mathcal{E} : \mathcal{P} \to \mathbb{R}_+$, $\mathcal{E}(\rho^*) = 0$ (Energy Functional)
- $\mathcal{P}_a \subset \mathcal{P}$, $a \in \mathbb{R}^p$, $\rho(a) \in \mathcal{P}_a$ (Candidate Density)
- $g_{\rho(a)}(\nabla_a \rho(a) \cdot \sigma_1, \nabla_a \rho(a) \cdot \sigma_2) = \langle M(a)\sigma_1, \sigma_2 \rangle_{\mathbb{R}^p}$ (Metric)

The Gradient Flow in $\mathcal{P}$

$$\frac{d}{dt} a_t = -M(a_t)^{-1} \frac{\partial}{\partial a} \mathcal{E}(\rho_a) \bigg|_{a=a_t}$$

Example: Gaussian Variational Bayes

- $\mathcal{G}$ : all Gaussian probability measures on $\mathbb{R}^d$
- $\mathcal{G} = \mathcal{P}_a$, $a = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}_{\text{sym, } \geq 0}$
Identifying The Gradient Flow: Gaussian Case

Chen, Huang, Huang, Reich, AMS [8] (2023)

**Theorem**  Chen, Huang, Huang, Reich, AMS [8] (2023)

- Moment closure gives the gradient flow

**Consequence**

- Consider a gradient flow in \( P \):
  \[
  \frac{\partial \rho_t(\theta)}{\partial t} = \sigma_t(\theta, \rho_t)
  \]

- Then mean and covariance evolve according to
  \[
  \frac{dm_t}{dt} = \int \sigma_t(\theta, \rho_t)\theta d\theta, \quad \frac{dC_t}{dt} = \int \sigma_t(\theta, \rho_t)(\theta - m_t)(\theta - m_t)^T d\theta
  \]

- Closure: to obtain gradient flow in \( P_a \) use \( \rho_t = \rho_a_t = \mathcal{N}(m_t, C_t) \)


Identifying The Gradient Flow: Gaussian Case

### Gaussian Approximate Fisher-Rao Gradient Flows

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

- Stein integration by parts used in derivation.
- Same as natural gradient flow applied to solve

\[
\min_{m,C} \text{KL}[\mathcal{N}(m, C) \| \rho^*]
\]

- Fisher information matrix is used for preconditioning

Convergence Rates

Theorem: Gaussian Targets

If \( \rho^* = \mathcal{N}(m^*, C^*) \), and \( C_0 = \lambda_0 I, \lambda_0 > 0 \), then

\[
\|m_t - m^*\|_2 = \Theta(e^{-t}), \quad \|C_t - C^*\|_2 = \Theta(e^{-t})
\]

Theorem: General Targets

Assume \( \alpha I \preceq -\nabla_\theta \nabla_\theta \log \rho^* \preceq \beta I \), and \( \lambda_{0,\min} I \preceq C_0 \preceq \lambda_{0,\max} I \), then

\[
\text{KL}[\rho_{a_t} \| \rho^*] \leq e^{-tK} \text{KL}[\rho_{a_0} \| \rho^*] + (1 - e^{-tK}) \text{KL}[\rho_{a_*} \| \rho^*]
\]

where \( a_t = (m_t, C_t) \), \( \rho_{a_t} = \mathcal{N}(m_t, C_t) \), \( K = \alpha \min\{1/\beta, \lambda_{0,\min}\} \) and

\[
a_* = \arg\min_{m,C} \text{KL}[\mathcal{N}(m, C) \| \rho^*]
\]

See also: Lambert, Chewi, Bach, Bonnabel, Rigollet [14] (2022)
2D convex potential:

\[ V(\theta) = \frac{1}{20} (\sqrt{\lambda} \theta_1 - \theta_2)^2 + \frac{1}{20} (\theta_2)^4 \]

for \( \theta = (\theta_1, \theta_2) \) and \( \lambda = 10^{-k}, \ k = 0, 1, 2 \)

Goal: sample \( \rho^* \propto \exp(-V(\theta)) \)

Method 1: Gaussian approximation of Fisher-Rao GF

Method 2: Gaussian approximation of Wasserstein GF

Method 3: Gaussian approximation of vanilla GF

Configuration: Integrate to \( t = 15 \) initialized from the Gaussian

\[ \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}) \]
Numerical Examples

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

Unifying Framework

Choice of Energy Functional

Choice of Metric

Affine Invariant Metrics

Gaussian Variational Bayes

Conclusions
Summary

Gradient Flows for Sampling  
Chen, Huang, Huang, Reich, AMS [8] (2023)

- **Energy Functional**: KL divergence
  - invariant to normalization consts
  - unique property among all $f$ divergences

- **Fisher-Rao Metric**:
  - invariant to any diffeomorphism of the parameters
  - unique property among all metrics on probability space
  - uniform exponential convergence
  - implementing mean field models is difficult

- **Affine Invariance**:
  - uniform exponential convergence for Gaussian target
  - examples: affine invariant Wasserstein, Stein metrics
  - implementation of mean field models is straightforward

- **Numerics**:
  - demonstrate benefits of affine invariance for mean field
  - demonstrate benefits of Fisher-Rao metric for variational Bayes
Gradient flows for sampling: mean-field models, Gaussian approximations and affine invariance

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, Andrew M. Stuart
References I

Natural gradient works efficiently in learning.


*Gradient flows: in Metric Spaces and in the Space of Probability Measures.*

*Information Geometry*, volume 64.

Uniqueness of the Fisher–Rao metric on the space of smooth densities.

*Handbook of Markov chain Monte Carlo.*
References II

Statistical decision rules and optimal inference. 

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

Sequential Monte Carlo samplers. 

Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. 

Ensemble samplers with affine invariance. 


Birth-death dynamics for sampling: Global convergence, approximations and their asymptotics. 

*Information Theory, Inference and Learning Algorithms.* 

*Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. 

Information and the accuracy attainable in the estimation of statistical parameters. 

*Probabilistic forecasting and Bayesian data assimilation.* 
References V

[23] H. Risken.  
_Fokker-Planck Equation._  
Springer, 1996.

Optimal scaling for various metropolis–hastings algorithms.  

Exponential convergence of Langevin distributions and their discrete approximations.  

On unscented Kalman filtering for state estimation of continuous-time nonlinear systems.  

From optimization to sampling through gradient flows.  

_Optimal transport: old and new_, volume 338.  
References VI

Graphical models, exponential families, and variational inference.

Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem.