Kernel Approximation of Wasserstein and Fisher-Rao Gradient flows

Jia-Jie Zhu

Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

Weierstraß-Institut für
Angewandte Analysis und Stochastik

SPP2298

December 15th, 2024
Department of Mathematics, University of British Columbia
Vancouver, B.C., Canada

This talk is based on the joint works with Alexander Mielke

Z. Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective. arXiv preprint

Gladin-Dvurechensky-Mielke-**Z**. Interaction-Force Transport Gradient Flows. *NeurIPS 2024*

Z-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows. arXiv preprint

Z-Mielke. Approximation, Kernelization, and Entropy-Dissipation of Gradient Flows: from Wasserstein to Fisher-Rao. technical report

Direct optimization over measures

$$\min_{\mu \in A \subset \mathcal{M}^+} F(\mu)$$

e.g., $F(\mu) = \mathrm{D_{KL}}(\mu|\pi)$.

Direct optimization over measures

$$\min_{\mu \in A \subset \mathcal{M}^+} F(\mu)$$

e.g., $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.

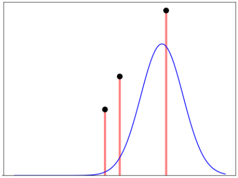
$\mathcal{P}$ probability measure "simplex"

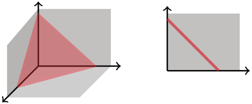
$$\text{continuous:} \quad \mathrm{d}\mu(x) = \rho(x)\ \mathrm{d}x, \int \rho = 1, \rho \geq 0$$

$$\text{discrete:} \ \mu = \sum_{i \in I} p_i \delta_{x_i}\ \mathbf{1}^\top p = 1, p \geq 0$$

$\mathcal{M}^+$ non-negative measure "cone"

$$\rho \geq 0, p \geq 0$$





bottom figure: A. Mielke

Direct optimization over measures

$$\min_{\mu \in A \subset \mathcal{M}^+} F(\mu)$$

e.g., $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.

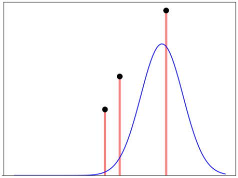
$\mathcal{P}$ probability measure "simplex"

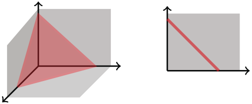
$$\text{continuous: } \mathrm{d}\mu(x) = \rho(x)\ \mathrm{d}x, \int \rho = 1, \rho \geq 0$$

$$\text{discrete: } \mu = \sum_{i \in I} p_i \delta_{x_i}\ \mathbf{1}^\top p = 1, p \geq 0$$

$\mathcal{M}^+$ non-negative measure "cone"

$$\rho \geq 0, p \geq 0$$





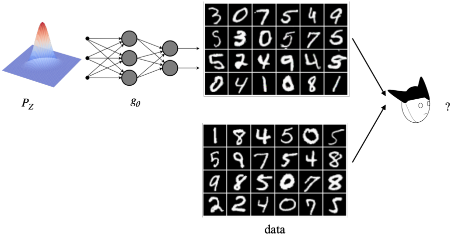
bottom figure: A. Mielke

We must respect the geometry of **(probability) measures**.

Motivation

Deep generative models

Previouis view of DGM (static) [Goodfellow et al. 2014...]



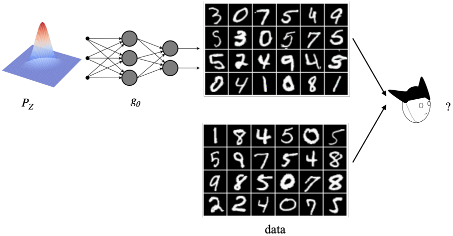
$$\min_{\theta} \mathrm{D}(\pi_{\mathrm{data}} | g_{\theta\#} P_Z)$$

New view of DGM (dynamic) Simulate an S/O/PDE [Chen et al. 2018, Song et al. 2021]

$$\dot{X}_t = -\nabla \xi_t(X_t), \text{ for some learned } \nabla \xi_t, \text{ e.g. NN}$$

Deep generative models

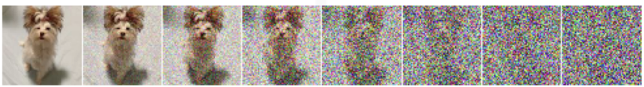
Previouis view of DGM (static) [Goodfellow et al. 2014...]



$$\min_\theta \mathrm{D}(\pi_{\mathrm{data}} | {g_\theta}_\# P_Z)$$

New view of DGM (dynamic) Simulate an S/O/PDE [Chen et al. 2018, Song et al. 2021]

$$\dot{X}_t = -\nabla \xi_t(X_t), \text{ for some learned } \nabla \xi_t, \text{ e.g. NN}$$



Perspective: flow and evolution of prob. measures

Robust learning under distribution shifts [Z. et al. AISTATS 2021, AISTATS 2023, ...]

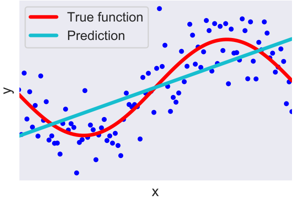
$$\text{Empirical risk minimization} \min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell(\theta, [x_i, y_i])$$

$x_i, y_i \sim P_0$: data sample. θ: learning parameter e.g. DNN weights.

Robust learning under distribution shifts [Z. et al. AISTATS 2021, AISTATS 2023, ...]

Empirical risk minimization $\min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell(\theta, [x_i, y_i])$

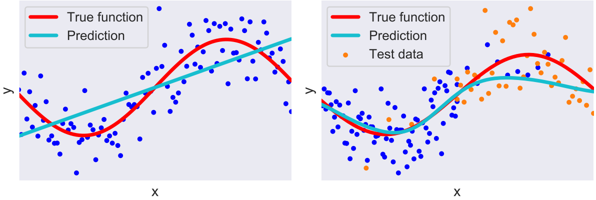
$x_i, y_i \sim P_0$: data sample. θ: learning parameter e.g. DNN weights.



Robust learning under distribution shifts [Z. et al. AISTATS 2021, AISTATS 2023, ...]

$$\text{Empirical risk minimization} \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta, [x_i, y_i])$$

$x_i, y_i \sim P_0$: data sample. θ: learning parameter e.g. DNN weights.

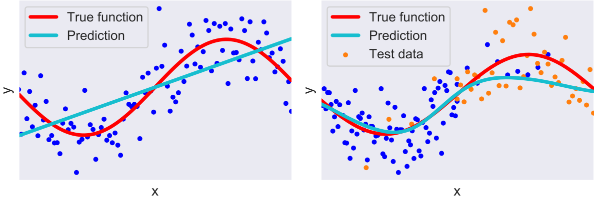


What if the test data **distribution shifts** from P_0?

Robust learning under distribution shifts [Z. et al. AISTATS 2021, AISTATS 2023, ...]

Empirical risk minimization $\min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell(\theta, [x_i, y_i])$

$x_i, y_i \sim P_0$: data sample. θ: learning parameter e.g. DNN weights.



What if the test data **distribution shifts** from P_0?

Distributionally robust optimization (DRO) $\min_\theta \sup_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathbb{E}_\mu \ell(\theta, [X, Y])$

$\mathcal{A} = \left\{ \mu \in \mathcal{P} \middle| \ \mathrm{D}(\mu | \hat{P}_N) \leq \epsilon \right\}$

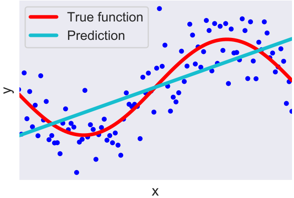
$\hat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$: empirical dist.

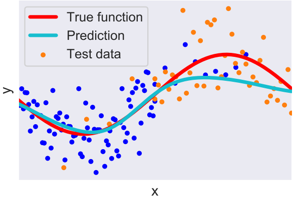
D: divergence between measures

Robust learning under distribution shifts [Z. et al. AISTATS 2021, AISTATS 2023, ...]

Empirical risk minimization $\min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell(\theta, [x_i, y_i])$

$x_i, y_i \sim P_0$: data sample. θ: learning parameter e.g. DNN weights.





What if the test data **distribution shifts** from P_0?

Distributionally robust optimization (DRO) $\min_\theta \sup_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathbb{E}_\mu \ell(\theta, [X, Y])$

$\mathcal{A} = \left\{ \mu \in \mathcal{P} \middle| \, \mathrm{D}(\mu | \hat{P}_N) \leq \epsilon \right\}$

$\hat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$: empirical dist.

D: divergence between measures

Wasserstein DRO [Esfahani & Kuhn 2018; Sinha et al. 2017] loss ℓ: (p/w) quadratic, logistic, etc.

Kernel DRO [Z. et al. AISTATS 2021, 22...] for general nonlinear loss in ML

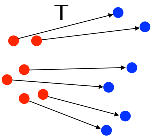
Gradient Flows of Probability Measures

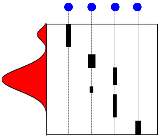
Kantorovich-Wasserstein distance and optimal transport

"Euclidean distance" between probability measures

p-th order **Kantorovich-Wasserstein distance** between (probability) measures μ_0, μ_1 on $X \subset \mathbb{R}^d$ with p finite moments is defined through the Kantorovich problem

$$W_p^p(\mu_0, \mu_1) := \min \left\{ \int |x_0 - x_1|^p \, \mathrm{d}\Pi \;\middle|\; \pi_\#^{(1)}\Pi = \mu_0, \pi_\#^{(2)}\Pi = \mu_1 \right\}$$





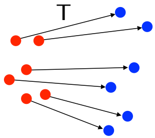
[Peyré and Cuturi, 2019]

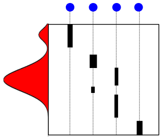
Kantorovich-Wasserstein distance and optimal transport

"Euclidean distance" between probability measures

p-th order **Kantorovich-Wasserstein distance** between (probability) measures μ_0, μ_1 on $X \subset \mathbb{R}^d$ with p finite moments is defined through the Kantorovich problem

$$W_p^p(\mu_0, \mu_1) := \min \left\{ \int |x_0 - x_1|^p \, \mathrm{d}\Pi \;\middle|\; \pi^{(1)}_{\#}\Pi = \mu_0, \pi^{(2)}_{\#}\Pi = \mu_1 \right\}$$





[Peyré and Cuturi, 2019]

Dynamic formulation: Benamou-Brenier

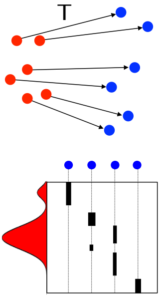
$$W_2^2(\mu_0, \mu_1) = \min \left\{ \int_0^1 \|\nabla \xi\|^2_{L^2_\mu} \, \mathrm{d}t \;\middle|\; \partial_t \mu = -\mathrm{div}\,(\mu \nabla \xi), \mu(0) = \mu_0, \mu(1) = \mu_1 \right\}$$

Kantorovich-Wasserstein distance and optimal transport

"Euclidean distance" between probability measures

p-th order **Kantorovich-Wasserstein distance** between (probability) measures μ_0, μ_1 on $X \subset \mathbb{R}^d$ with p finite moments is defined through the Kantorovich problem

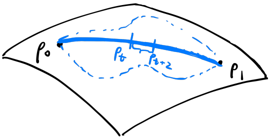
$$W_p^p(\mu_0, \mu_1) := \min \left\{ \int |x_0 - x_1|^p \, \mathrm{d}\Pi \;\middle|\; \pi^{(1)}_{\#}\Pi = \mu_0, \pi^{(2)}_{\#}\Pi = \mu_1 \right\}$$



[Peyré and Cuturi, 2019]

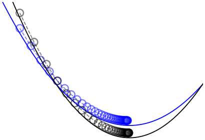
Dynamic formulation: Benamou-Brenier

$$W_2^2(\mu_0, \mu_1) = \min \left\{ \int_0^1 \|\nabla \xi\|^2_{L^2_\mu} \, \mathrm{d}t \;\middle|\; \partial_t \mu = -\mathrm{div}\,(\mu \nabla \xi), \mu(0) = \mu_0, \mu(1) = \mu_1 \right\}$$

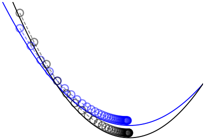


From gradient descent to gradient flow

Optimization problem in $\mathbb{R}^d$: $\min_{x \in \mathbb{R}^d} F(x)$



From gradient descent to gradient flow

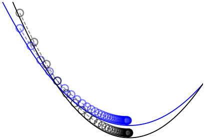


Optimization problem in $\mathbb{R}^d$: $\min_{x \in \mathbb{R}^d} F(x)$

Gradient descent $\quad x_{k+1} = x_k - \tau \cdot \nabla F(x_k)^\top$

Prox. step (implicit)/ JKO $\quad x_{k+1} \in \operatorname*{argmin}_x \left(F(x) + \frac{1}{2\tau} \|x - x_k\|^2 \right)$

From gradient descent to gradient flow



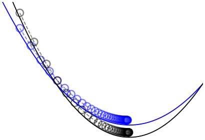
Optimization problem in $\mathbb{R}^d$: $\min_{x \in \mathbb{R}^d} F(x)$

Gradient descent $\quad x_{k+1} = x_k - \tau \cdot \nabla F(x_k)^\top$

Prox. step (implicit)/ JKO $\quad x_{k+1} \in \operatorname{argmin}_x \left(F(x) + \frac{1}{2\tau} \|x - x_k\|^2 \right)$

$\tau \to 0$ continuous time: ODE $\quad \dot{x}(t) = -\nabla F(x(t))^\top$

From gradient descent to gradient flow



Optimization problem in $\mathbb{R}^d$: $\min_{x \in \mathbb{R}^d} F(x)$

$$\text{Gradient descent} \quad x_{k+1} = x_k - \tau \cdot \nabla F(x_k)^\top$$

$$\text{Prox. step (implicit)/ JKO} \quad x_{k+1} \in \operatorname*{argmin}_x \left(F(x) + \frac{1}{2\tau} \|x - x_k\|^2 \right)$$

$$\tau \to 0 \text{ continuous time: ODE} \quad \dot{x}(t) = -\nabla F(x(t))^\top$$

is the *gradient-flow equation* of the energy $F(x)$ in the **space** $\mathbb{R}^d$ with the Eulidean geometry described by $\|x\|^2$.

Otto-Wasserstein gradient flow

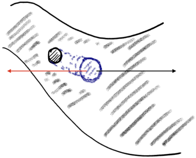
Generalizing the Euclidean geometry $\left(\mathbb{R}^d, \|\cdot\|\right)$ to $(\mathcal{P}, W_2)$

Otto-Wasserstein gradient flow

Generalizing the Euclidean geometry $\left(\mathbb{R}^d, \|\cdot\|\right)$ to $(\mathcal{P}, W_2)$

Otto-Wasserstein gradient flow [Otto, 1996, 2001]

$$\mu^{k+1} \in \operatorname*{argmin}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} {W_2}^2(\mu, \mu^k) \qquad \text{(JKO)}$$



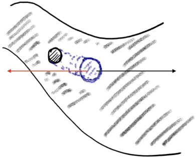
Continuous-time ($\tau \to 0$) **gradient flow equation**

$$\partial_t \mu = -\mathrm{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\right)$$

Otto-Wasserstein gradient flow

Generalizing the Euclidean geometry $\left(\mathbb{R}^d, \|\cdot\|\right)$ to $(\mathcal{P}, W_2)$
Otto-Wasserstein gradient flow [Otto, 1996, 2001]

$$\mu^{k+1} \in \underset{\mu \in \mathcal{P}}{\operatorname{argmin}}\, F(\mu) + \frac{1}{2\tau} {W_2}^2(\mu, \mu^k) \qquad \text{(JKO)}$$



Continuous-time ($\tau \to 0$) **gradient flow equation**

$$\partial_t \mu = -\mathrm{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\right)$$

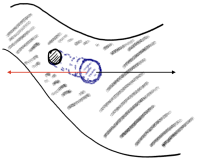
PDE has a **gradient structure**:

$$\begin{cases} \textbf{Measure Space}: & \mathcal{P} \text{ or } \mathcal{M}^+ \\ \text{Energy functional}: & F \text{ (e.g. KL)} \\ \text{Dissipation Geometry}: & W_2 \text{ or He} \end{cases}$$

Otto-Wasserstein gradient flow

Generalizing the Euclidean geometry $\left(\mathbb{R}^d, \|\cdot\|\right)$ to $(\mathcal{P}, W_2)$
Otto-Wasserstein gradient flow [Otto, 1996, 2001]

$$\mu^{k+1} \in \operatorname*{argmin}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} W_2{}^2(\mu, \mu^k) \qquad \text{(JKO)}$$



Continuous-time ($\tau \to 0$) **gradient flow equation**

$$\partial_t \mu = -\mathrm{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\right)$$

PDE has a **gradient structure**:
$$\begin{cases} \textbf{Measure Space}: & \mathcal{P} \text{ or } \mathcal{M}^+ \\ \text{Energy functional}: & F \text{ (e.g. KL)} \\ \text{Dissipation Geometry}: & W_2 \text{ or He} \end{cases}$$

> The merit of the right gradient flow formulation of a dissipative evolution equation is that it **separates energetics and kinetics**: The energetics endow the state space with a functional, the kinetics endow the **state space** with a (Riemannian) geometry via the metric tensor. [Otto 2001]

Information divergence and Hellinger (~~Fisher-Rao~~) distance

φ-divergence energy [Csiszar 1967] $\quad \mathrm{D}_\varphi(\mu|\nu) := \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu$

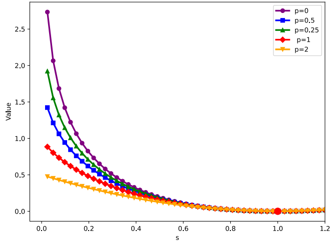
"p-relative entropy"

$$\varphi_p(s) := \frac{1}{p(p-1)}\left(s^p - p(s-1) - 1\right)$$

$$p = 2 : \chi^2, \quad p = \frac{1}{2} : \text{Hellinger}$$

$$p \to 1 : \text{KL}, \ \varphi_1(s) := \varphi_{\text{KL}} = s \log s - s + 1$$

$$p \to 0 : \text{rev. KL}, \ \varphi_0(s) := s - 1 - \log s$$



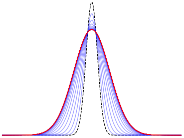
Hellinger distance over $\mathcal{M}^+$

$$\text{He}^2(\mu_0, \mu_1) = 4 \cdot \int \left(\sqrt{\mu_0} - \sqrt{\mu_1}\right)^2$$

Hellinger (~~Fisher–Rao~~) distance over $\mathcal{M}^+$

$$\mathrm{He}^2(\mu_0, \mu_1) =$$

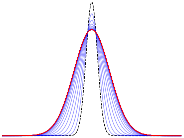
$$4 \cdot \int \left(\sqrt{\frac{\delta \mu_0}{\delta \gamma}} - \sqrt{\frac{\delta \mu_1}{\delta \gamma}} \right)^2 \mathrm{d}\gamma, \mu_0, \mu_1 << \gamma$$



Hellinger (~~Fisher-Rao~~) distance over $\mathcal{M}^+$

$$\mathsf{He}^2(\mu_0, \mu_1) =$$

$$4 \cdot \int \left(\sqrt{\frac{\delta \mu_0}{\delta \gamma}} - \sqrt{\frac{\delta \mu_1}{\delta \gamma}} \right)^2 \mathrm{d}\gamma, \mu_0, \mu_1 << \gamma$$

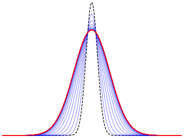


Dynamic formulation: Benamou-Brenier

$$\mathsf{He}^2(\mu_0, \nu_0) = \min_{\mu, \xi_t} \left\{ \int_0^1 \|\xi_t\|^2_{L^2_\mu} \, \mathrm{d}t \;\middle|\; \dot{\mu} = -\mu \cdot \xi_t, \; \mu(0) = \mu_0, \; \mu(1) = \mu_1 \right\}.$$

Hellinger (~~Fisher-Rao~~) distance over $\mathcal{M}^+$

$$\mathrm{He}^2(\mu_0, \mu_1) = 4 \cdot \int \left(\sqrt{\frac{\delta \mu_0}{\delta \gamma}} - \sqrt{\frac{\delta \mu_1}{\delta \gamma}} \right)^2 \mathrm{d}\gamma, \mu_0, \mu_1 << \gamma$$



Dynamic formulation: Benamou-Brenier

$$\mathrm{He}^2(\mu_0, \nu_0) = \min_{\mu, \xi_t} \left\{ \int_0^1 \|\xi_t\|^2_{L^2_\mu} \, \mathrm{d}t \;\middle|\; \dot{\mu} = -\mu \cdot \xi_t,\ \mu(0) = \mu_0,\ \mu(1) = \mu_1 \right\}.$$

Geodesic curves

$$\begin{cases} \dot{\mu} = -\xi \mu, \\ \dot{\xi} = -\frac{1}{2} |\xi|^2. \end{cases}$$

Dual Kantorovich type form: (see [Z and Mielke, 2024] for details)

$$\frac{1}{2} \mathrm{He}^2(\mu_0, \mu_1) = \sup_{(2+\phi)(2-\psi)=4} \left\{ \int \psi \, \mathrm{d}\mu_1 - \int \phi \, \mathrm{d}\mu_0 \right\}.$$

Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

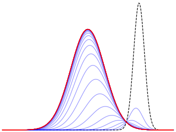
Birth-death process $2\mathrm{H}_2\mathrm{O} \rightleftharpoons 2\mathrm{H}_2 + 1\mathrm{O}_2$

Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

Birth-death process $2\mathrm{H_2O} \rightleftharpoons 2\mathrm{H_2} + 1\mathrm{O_2}$

Hellinger gradient flows $(\mathcal{M}^+, F, \mathrm{He})$



$$\min_{\mu \in \mathcal{M}^+} F(\mu) + \frac{1}{2\tau} \mathrm{He}^2(\mu, \mu^k)$$

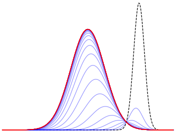
continuous-time $\tau \to 0 : \dot{\mu} = -\mu \cdot \frac{\delta F}{\delta \mu} [\mu]$

Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

Birth-death process $2\mathrm{H_2O} \rightleftharpoons 2\mathrm{H_2} + 1\mathrm{O_2}$

Hellinger gradient flows $(\mathcal{M}^+, F, \mathsf{He})$



$$\min_{\mu \in \mathcal{M}^+} F(\mu) + \frac{1}{2\tau} \mathsf{He}^2(\mu, \mu^k)$$

continuous-time $\tau \to 0$: $\dot{\mu} = -\mu \cdot \frac{\delta F}{\delta \mu}[\mu]$

Example

- [Z and Mielke, 2024] Convergence analysis of KL-inference

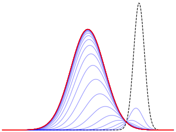
$$\min_{\mu \in \mathcal{M}^+} \mathrm{D_{KL}}(\mu | \pi) \text{ in } (\mathcal{P}, \mathsf{He})$$

Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

Birth-death process $2\mathrm{H_2O} \rightleftharpoons 2\mathrm{H_2} + 1\mathrm{O_2}$

Hellinger gradient flows $(\mathcal{M}^+, F, \mathrm{He})$



$$\min_{\mu \in \mathcal{M}^+} F(\mu) + \frac{1}{2\tau} \mathrm{He}^2(\mu, \mu^k)$$

continuous-time $\tau \to 0 : \dot{\mu} = -\mu \cdot \frac{\delta F}{\delta \mu}[\mu]$

Example

- [Z and Mielke, 2024] Convergence analysis of KL-inference

$$\min_{\mu \in \mathcal{M}^+} \mathrm{D_{KL}}(\mu|\pi) \text{ in } (\mathcal{P}, \mathrm{He})$$

- variational inference via natural gradient: (spherical) Hellinger metric tensor gives the Fisher information matrix [Amari, 1998, Khan and Nielsen, 2018]
- entropic mirror descent in optimization [Nemirovskij and Yudin, 1983, Beck and Teboulle, 2003]

Inference via interacting particle systems: Langevin MC

Goal: to sample from $\pi(x) = \frac{1}{\int e^{-V(x)}\,\mathrm{d}x} e^{-V(x)}$

Inference via interacting particle systems: Langevin MC

Goal: to sample from $\pi(x) = \frac{1}{\int e^{-V(x)}\,\mathrm{d}x} e^{-V(x)}$

Langevin SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck PDE

$$\partial_t \mu = \Delta \mu + \mathrm{div}\,(\mu \nabla V)$$

Inference via interacting particle systems: Langevin MC

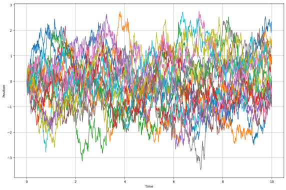
Goal: to sample from $\pi(x) = \dfrac{1}{\int e^{-V(x)}\, \mathrm{d}x} e^{-V(x)}$

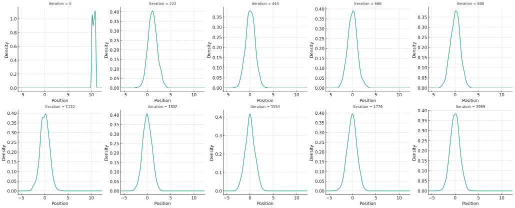
Langevin SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck PDE

$$\partial_t \mu = \Delta \mu + \mathrm{div}\,(\mu \nabla V)$$





Inference via interacting particle systems: Langevin MC

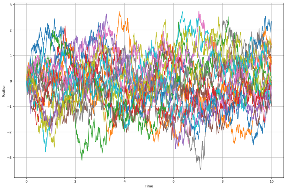
Goal: to sample from $\pi(x) = \frac{1}{\int e^{-V(x)}\, \mathrm{d}x} e^{-V(x)}$

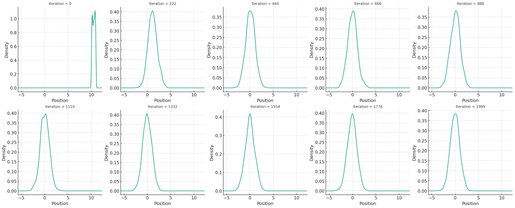
Langevin SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck PDE

$$\partial_t \mu = \Delta \mu + \mathrm{div}\,(\mu \nabla V)$$





GF & OPT perspective

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi) \text{ in } (\mathcal{P}, W_2)$$

Inference via interacting particle systems: Langevin MC

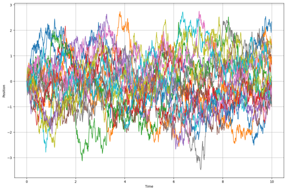
Goal: to sample from $\pi(x) = \frac{1}{\int e^{-V(x)}\, \mathrm{d}x} e^{-V(x)}$

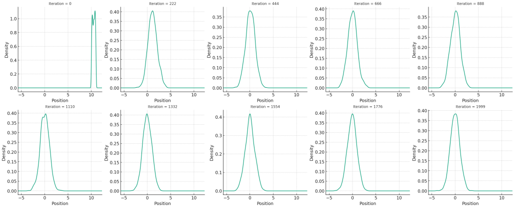
Langevin SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck PDE

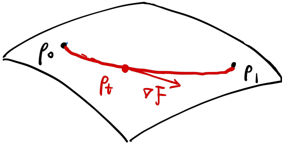
$$\partial_t \mu = \Delta \mu + \mathrm{div}\,(\mu \nabla V)$$





GF & OPT perspective

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi) \text{ in } (\mathcal{P}, W_2)$$



Inference via interacting particle systems: Langevin MC

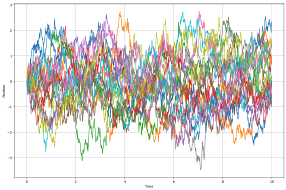
Goal: to sample from $\pi(x) = \frac{1}{\int e^{-V(x)}\,\mathrm{d}x} e^{-V(x)}$

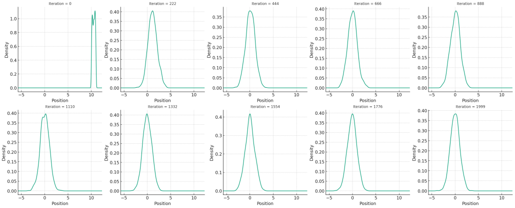
Langevin SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck PDE

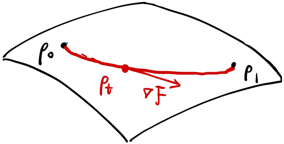
$$\partial_t \mu = \Delta \mu + \mathrm{div}\,(\mu \nabla V)$$





GF & OPT perspective

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi) \text{ in } (\mathcal{P}, W_2)$$



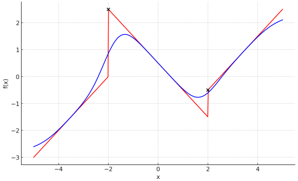
In a series of papers jointly with A. Mielke, we provide rigorous analysis of various gradient flows beyond the W_2 setting of [Bakry and Émery, 1985] e.g. log-Sobolev.

Kernel Approximation

Kernel methods and MMD

$\mathcal{H}$ is the **reproducing kernel Hilbert space** (RKHS), which satisfies

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$$

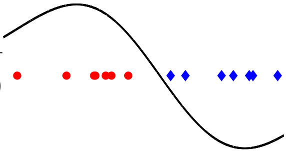


Integral operator $\mathcal{K}_\rho : L^2(\rho) \to L^2(\rho)$:

$$\mathcal{K}_\rho g(x) := \int k(x, x') \, g(x') \, d\rho(x')$$

Maximum-mean discrepancy (MMD) [Gretton et al., 2012]

$$\begin{aligned} \mathrm{MMD}(\mu_0, \mu_1) :=& \left\| \int k(x, \cdot) d\mu_0 - \int k(x, \cdot) d\mu_1 \right\|_{\mathcal{H}} \\ =& \sqrt{\int \int k(x, x') \; \mathrm{d}(\mu_0 - \mu_1)(x) \; \mathrm{d}(\mu_0 - \mu_1)(x')} \\ =& \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f \, d(\mu_0 - \mu_1) \end{aligned}$$



MMD as dekernelized Hellinger distance

The "MMD paper" [Gretton et al., 2012] has now $> 5k$ citations.
Dyanmic formulation of MMD: "straight line" geodesics

$$\mathrm{MMD}^2(\mu,\nu) = \min \left\{ \int_0^1 \|\xi_t\|^2_{\mathcal{H}} \, \mathrm{d}t \;\middle|\; \dot{u} = -\mathcal{K}^{-1}\xi_t, u(0) = \mu, u(1) = \nu \right\}.$$

MMD as dekernelized Hellinger distance

The "MMD paper" [Gretton et al., 2012] has now $> 5k$ citations.
Dyanmic formulation of MMD: "straight line" geodesics

$$\mathrm{MMD}^2(\mu,\nu) = \min\left\{ \int_0^1 \|\xi_t\|_{\mathcal{H}}^2 \,\mathrm{d}t \;\middle|\; \dot{u} = -\mathcal{K}^{-1}\xi_t, u(0) = \mu, u(1) = \nu \right\}.$$

The integral operator
$\mathcal{K}_\rho := g(x) := \int k\left(x, x'\right) g\left(x'\right) d\rho\left(x'\right), \quad g \in L^2_\rho, L^2(\rho) \to L^2(\rho)$ is compact, positive, self-adjoint, and nuclear.

MMD as dekernelized Hellinger distance

The "MMD paper" [Gretton et al., 2012] has now $> 5k$ citations.
Dyanmic formulation of MMD: "straight line" geodesics

$$\mathrm{MMD}^2(\mu,\nu) = \min \left\{ \int_0^1 \|\xi_t\|_{\mathcal{H}}^2 \, \mathrm{d}t \;\middle|\; \dot{u} = -\mathcal{K}^{-1}\xi_t, u(0) = \mu, u(1) = \nu \right\}.$$

The integral operator
$\mathcal{K}_\rho := g(x) := \int k\left(x, x'\right) g\left(x'\right) d\rho\left(x'\right), \quad g \in L^2_\rho, L^2(\rho) \to L^2(\rho)$ is compact, positive, self-adjoint, and nuclear.

Theorem (MMD = de-kernelized Hellinger)

The dynamic formulation of the kernelized squared MMD coincides with that of the squared Hellinger distance

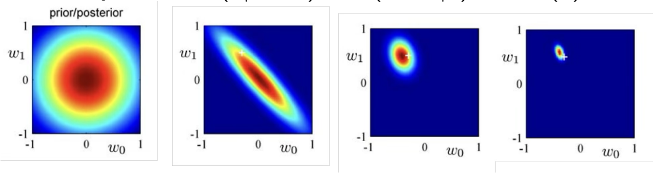
The Riemannian metric tensors are related by $\mathbb{G}_{\mathrm{MMD}} = \mathcal{K}_\mu \circ \mathbb{G}_{\mathrm{He}}(\mu)$.

Statistical Inference via Gradient Flows

Bayesian inference and probabilistic ML

Infer posterior distribution π of the model parameters θ given data,

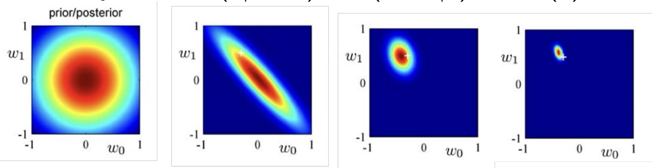
$$\text{Bayes rule: } \pi(\theta|\mathrm{Data}) \propto P(\mathrm{Data}|\theta) \cdot \mathrm{Prior}(\theta)$$



Bayesian inference and probabilistic ML

Infer posterior distribution π of the model parameters θ given data,

$$\text{Bayes rule: } \pi(\theta|\mathrm{Data}) \propto P(\mathrm{Data}|\theta) \cdot \mathrm{Prior}(\theta)$$



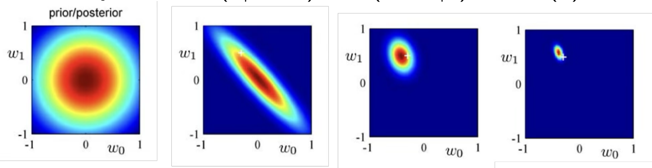
In practice, the exact π is intractable: **approximate inference** [Jordan et al., 1999, Wainwright and Jordan, 2008]

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D_{KL}}(\mu|\pi(\theta|\mathrm{Data})).$$

Bayesian inference and probabilistic ML

Infer posterior distribution π of the model parameters θ given data,

$$\text{Bayes rule: } \pi(\theta|\mathrm{Data}) \propto P(\mathrm{Data}|\theta) \cdot \mathrm{Prior}(\theta)$$



In practice, the exact π is intractable: **approximate inference** [Jordan et al., 1999, Wainwright and Jordan, 2008]

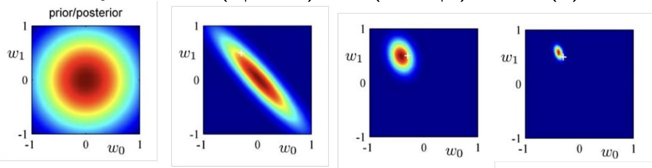
$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu | \pi(\theta|\mathrm{Data})).$$

Gaussian **variational inference**: $\mu \in \mathcal{N}^d$; also Laplace approx.

Bayesian inference and probabilistic ML

Infer posterior distribution π of the model parameters θ given data,

$$\text{Bayes rule: } \pi(\theta|\mathrm{Data}) \propto P(\mathrm{Data}|\theta) \cdot \mathrm{Prior}(\theta)$$



In practice, the exact π is intractable: **approximate inference** [Jordan et al., 1999, Wainwright and Jordan, 2008]

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D_{KL}}(\mu|\pi(\theta|\mathrm{Data})).$$

Gaussian **variational inference**: $\mu \in \mathcal{N}^d$; also Laplace approx.

Sampling / MCMC: generate samples $\theta^i \sim \pi, \; \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^i} \to \pi$

figure: [Bishop 2006]

Inference with forward and reverse KL

$$\min_{\mu \in \mathcal{N}^d \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{vs.} \quad \min_{\mu \in \mathcal{N}^d \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$$

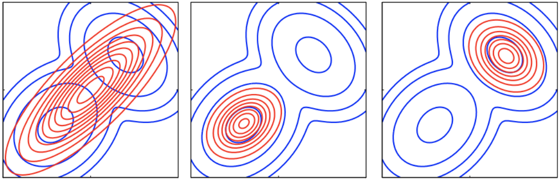
forward / inclusive	reverse / exclusive
mode-covering	mode-seeking

Inference with forward and reverse KL

$$\min_{\mu \in \mathcal{N}^d \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{vs.} \quad \min_{\mu \in \mathcal{N}^d \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$$

forward / inclusive — reverse / exclusive

mode-covering — mode-seeking



[Bishop 2006]

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathrm{D_{KL}}(\pi | \mu).$$

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Wasserstein gradient flow [Z. 2024]: $\dot{\mu} = \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Wasserstein gradient flow [Z. 2024]: $\dot{\mu} = \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$

Not implementable due to $\nabla(1 - \mathrm{d}\pi / \mathrm{d}\mu)$.

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Wasserstein gradient flow [Z. 2024]: $\dot{\mu} = \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$

Not implementable due to $\nabla(1 - \mathrm{d}\pi/\mathrm{d}\mu)$.

Kernel approx. [Z. 2024; Gladin et al. & Z. NeurIPS 2024]:

$$\dot{\mu} = \mathrm{div}\left(\mu \nabla \int k(z, \cdot) \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(z)\right) \mathrm{d}\mu(z)\right)$$

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Wasserstein gradient flow [Z. 2024]: $\dot{\mu} = \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$

Not implementable due to $\nabla(1 - \mathrm{d}\pi / \mathrm{d}\mu)$.

Kernel approx. [Z. 2024; Gladin et al. & Z. NeurIPS 2024]:

$$\dot{\mu} = \mathrm{div}\left(\mu \nabla \int k(z, \cdot)\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(z)\right) \mathrm{d}\mu(z)\right)$$

Theorem. The PDE has gradient structure:

$$\begin{cases} \text{Energy functional}: & \frac{1}{2}\,\mathsf{MMD}^2(\cdot, \pi) \\ \text{Geometry}: & \text{Wasserstein} \end{cases}$$

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

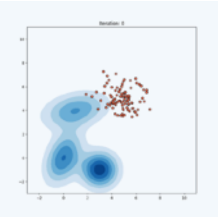
Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Wasserstein gradient flow [Z. 2024]: $\dot{\mu} = \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$

Not implementable due to $\nabla(1 - \mathrm{d}\pi/\mathrm{d}\mu)$.

Kernel approx. [Z. 2024; Gladin et al. & Z. NeurIPS 2024]:

$$\dot{\mu} = \mathrm{div}\left(\mu \nabla \int k(z, \cdot)\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(z)\right) \mathrm{d}\mu(z)\right)$$



Theorem. The PDE has gradient structure:

$$\begin{cases} \text{Energy functional}: & \frac{1}{2}\mathsf{MMD}^2(\cdot, \pi) \\ \text{Geometry}: & \text{Wasserstein} \end{cases}$$

$\mathsf{MMD}^2(\mu, \pi) = \mathbb{E}_{x,y\sim\mu} k(x, y) + \mathbb{E}_{x,y\sim\pi} k(x, y) - 2\mathbb{E}_{x\sim\mu, y\sim\pi} k(x, y)$

Gradient Flows in the Unbalanced Transport Geometry: Wasserstein-Fisher-Rao

Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

Best of both worlds

Wasserstein: transport/diffusion Fisher-Rao / Hellinger: growth/birth-death

H-K / W-FR [Chizat et al., 2018, 2019, Liero et al., 2018, Kondratyev et al., 2016] gradient flow:reaction-diffusion eqn.

Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

Best of both worlds

Wasserstein: transport/diffusion   Fisher-Rao / Hellinger: growth/birth-death

H-K / W-FR [Chizat et al., 2018, 2019, Liero et al., 2018, Kondratyev et al., 2016] gradient flow:reaction-diffusion eqn.

$$\dot{\mu} = -\alpha \mathbb{G}_W^{-1}(\mu) \frac{\delta F}{\delta \mu}[\mu]$$

Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

Best of both worlds

Wasserstein: transport/diffusion   Fisher-Rao / Hellinger: growth/birth-death

H-K / W-FR [Chizat et al., 2018, 2019, Liero et al., 2018, Kondratyev et al., 2016] gradient flow: reaction-diffusion eqn.

$$\dot{\mu} = -\alpha \mathbb{G}_W^{-1}(\mu) \frac{\delta F}{\delta \mu}[\mu] - \beta \mathbb{G}_{\mathsf{He}}^{-1}(\mu) \frac{\delta F}{\delta \mu}[\mu]$$

Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

Best of both worlds

Wasserstein: transport/diffusion   Fisher-Rao / Hellinger: growth/birth-death

H-K / W-FR [Chizat et al., 2018, 2019, Liero et al., 2018, Kondratyev et al., 2016] gradient flow: reaction-diffusion eqn.

$$\dot{\mu} = -\alpha \mathbb{G}_W^{-1}(\mu) \frac{\delta F}{\delta \mu}[\mu] - \beta \mathbb{G}_{\mathrm{He}}^{-1}(\mu) \frac{\delta F}{\delta \mu}[\mu] = \alpha \operatorname{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}\right) - \beta \mu \frac{\delta F}{\delta \mu}$$

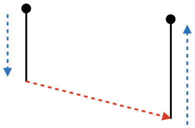
Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

Best of both worlds

Wasserstein: transport/diffusion   Fisher-Rao / Hellinger: growth/birth-death

H-K / W-FR [Chizat et al., 2018, 2019, Liero et al., 2018, Kondratyev et al., 2016] gradient flow:reaction-diffusion eqn.

$$\dot{\mu} = -\alpha \mathbb{G}_W^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] - \beta \mathbb{G}_{\mathsf{He}}^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] = \alpha \mathrm{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}\right) - \beta \mu \frac{\delta F}{\delta \mu}$$



Unbalanced transport gradient flows of forward (incl.) KL inference [Gladin et al. & Z. NeurIPS 2024; Z. 2024]

[Z. 2024] precise connection between the gradient flows of:

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D_{KL}}(\pi|\mu) \quad \text{and} \quad \min_{\mu \in A \subset \mathcal{P}} \mathsf{MMD}^2(\mu, \pi)$$

[Arbel et al. 2019] studied the latter without guarantee of convergence; [Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

Unbalanced transport gradient flows of forward (incl.) KL inference [Gladin et al. & Z. NeurIPS 2024; Z. 2024]

[Z. 2024] precise connection between the gradient flows of:

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{and} \quad \min_{\mu \in A \subset \mathcal{P}} \mathrm{MMD}^2(\mu, \pi)$$

[Arbel et al. 2019] studied the latter without guarantee of convergence; [Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

[Z. 2024] Wasserstein-Fisher-Rao flow: reaction-diffusion eq:

$$\dot{\mu} = \underbrace{\alpha \cdot \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)}_{\text{Wasserstein: transport}}$$

Unbalanced transport gradient flows of forward (incl.) KL inference [Gladin et al. & Z. NeurIPS 2024; Z. 2024]

[Z. 2024] precise connection between the gradient flows of:

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{and} \quad \min_{\mu \in A \subset \mathcal{P}} \mathrm{MMD}^2(\mu, \pi)$$

[Arbel et al. 2019] studied the latter without guarantee of convergence; [Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

[Z. 2024] Wasserstein-Fisher-Rao flow: reaction-diffusion eq:

$$\dot{\mu} = \underbrace{\alpha \cdot \operatorname{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)}_{\text{Wasserstein: transport}} - \underbrace{\beta \cdot \mu \cdot \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)}_{\text{Fisher-Rao: birth-death}}$$

Unbalanced transport gradient flows of forward (incl.) KL inference [Gladin et al. & Z. NeurIPS 2024; Z. 2024]

[Z. 2024] precise connection between the gradient flows of:

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{and} \quad \min_{\mu \in A \subset \mathcal{P}} \mathrm{MMD}^2(\mu, \pi)$$

[Arbel et al. 2019] studied the latter without guarantee of convergence; [Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

[Z. 2024] Wasserstein-Fisher-Rao flow: reaction-diffusion eq:

$$\dot{\mu} = \underbrace{\alpha \cdot \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)}_{\text{Wasserstein: transport}} - \underbrace{\beta \cdot \mu \cdot \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)}_{\text{Fisher-Rao: birth-death}}$$

Interaction-force transport [Z. 2024, Gladin et al. & Z. NeurIPS 2024] with global convergence guarantee

Unbalanced transport gradient flows of forward (incl.) KL inference [Gladin et al. & Z. NeurIPS 2024; Z. 2024]

[Z. 2024] precise connection between the gradient flows of:

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi | \mu) \quad \text{and} \quad \min_{\mu \in A \subset \mathcal{P}} \mathrm{MMD}^2(\mu, \pi)$$

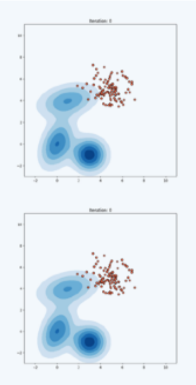
[Arbel et al. 2019] studied the latter without guarantee of convergence; [Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

[Z. 2024] Wasserstein-Fisher-Rao flow: reaction-diffusion eq:

$$\dot{\mu} = \underbrace{\alpha \cdot \operatorname{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)}_{\text{Wasserstein: transport}} \underbrace{- \beta \cdot \mu \cdot \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)}_{\text{Fisher-Rao: birth-death}}$$

Interaction-force transport [Z. 2024, Gladin et al. & Z. NeurIPS 2024] with global convergence guarantee

$$\dot{\mu} = \alpha \cdot \operatorname{div}\left(\mu \nabla \int k(x, \cdot)\, \mathrm{d}\left(\mu - \pi\right)(x)\right) - \beta \cdot (\mu - \pi)$$



JKO splitting scheme

The PDE

$$\dot{\mu} = \alpha \cdot \mathsf{div}\left(\mu \nabla \int k(x,\cdot)\,\mathrm{d}\,(\mu - \pi)\,(x)\right) - \beta \cdot (\mu - \pi)$$

can be simulated using the JKO scheme

$$\mu^{\ell+\frac{1}{2}} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu^\ell), \qquad \text{(Wasserstein step)}$$

$$\mu^{\ell+1} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\eta} \mathsf{MMD}^2(\mu, \mu^{\ell+\frac{1}{2}}), \quad \text{(MMD step)}$$

for $F(\mu) = \frac{1}{2}\,\mathsf{MMD}^2(\mu, \pi)$.

Insight on the variational principle: kernel methods vs information geometry

Theorem [Z, 2024] Suppose the kernel k is bounded and integrally strictly positive definite. Then, the solutions of the following variational problems coincide:

$$\min_{\mu\in\mathcal{P}} \frac{1}{2}\,\mathsf{MMD}^2(\mu,\pi) + \frac{1}{2\eta}\,\mathsf{MMD}^2(\mu,\mu^l).$$

$$\operatorname*{argmin}_{\mu\in\mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) + \frac{1}{\eta}\,\mathrm{D}_{\mathrm{KL}}(\mu^l|\mu).$$

Thank you!

References

- **Z**. Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective. arXiv preprint
- Gladin-Dvurechensky-Mielke-**Z**. Interaction-Force Transport Gradient Flows. *NeurIPS 2024*
- **Z**-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows. arXiv preprint
- **Z**-Mielke. Approximation, Kernelization, and Entropy-Dissipation of Gradient Flows: from Wasserstein to Fisher-Rao. technical report
- **Z**-Jitkrittum-Diehl-Schölkopf. Kernel Distributionally Robust Optimization. *AISTATS 2021*.

For more information, see my website: https://jj-zhu.github.io/ ; PhD position (Berlin) available

Appendix

Analysis of interaction-force transport gradient flows

Theorem [Gladin et al. & Z, 2024] Suppose $F = \frac{1}{2}\,\mathrm{MMD}^2(\cdot, \pi)$. Then the following functional inequality holds globally $\forall \mu \in \mathcal{P}$:

$$\alpha \left\| \nabla \frac{\delta F}{\delta \mu}[\mu] \right\|^2_{L^2_\mu} + \beta \left\| \frac{\delta F}{\delta \mu}[\mu] - \frac{\int \mathcal{K}^{-1} \frac{\delta F}{\delta \mu}[\mu]}{\int \mathcal{K}^{-1} 1} \right\|^2_{\mathcal{H}} \geq c \cdot \left(F(\mu(t)) - \inf_{\mu} F(\mu) \right)$$

with a constant $c \geq 2\beta > 0$. Consequently, the solution of the IFT gradient flow satisfies

$$\mathrm{MMD}(\mu_t, \pi) \leq e^{-\beta t} \cdot \mathrm{MMD}(\mu_0, \pi).$$

In addition, KL decays globally under the usual Bakry-Émery/LSI.