Two New Machine Learning Applications of Gradient Flows for Sampling

J.J. (Jia-Jie) Zhu

Weierstrass Institute, Berlin

(✈️ KTH Royal Institute of Technology, Stockholm)

Weierstraß-Institut für Angewandte Analysis und Stochastik

(upcoming: I'm hiring!)

more info

SPP2298

Berlin Mathematics Research Center

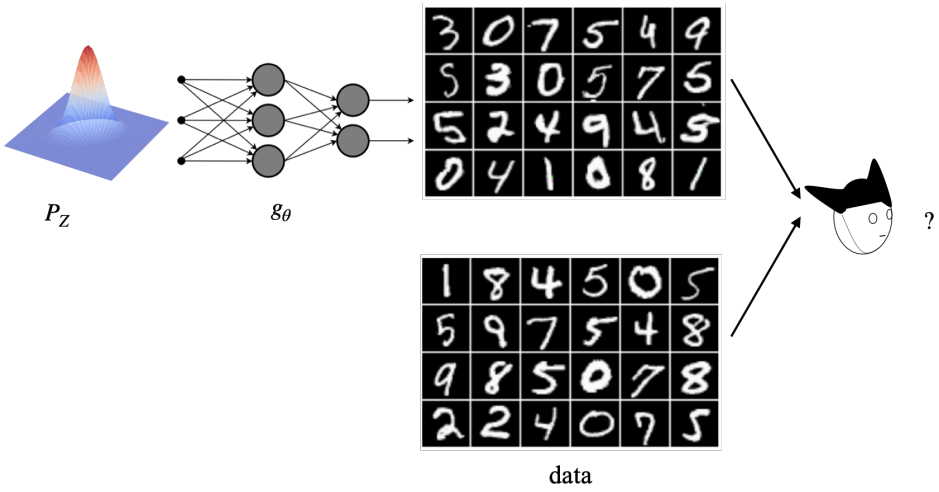
Deutsche Forschungsgemeinschaft

September 18th, 2025

Gradient flows face-to-face 5 in Granada, Spain

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]

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

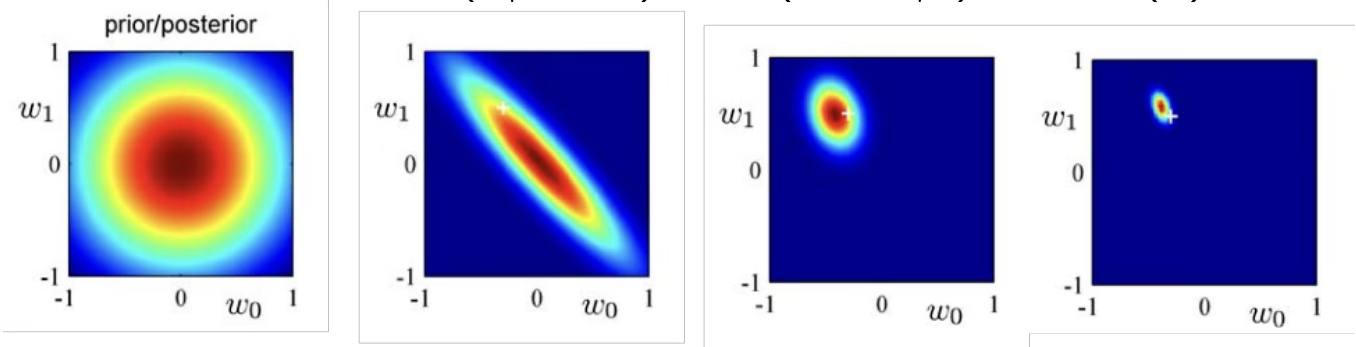


Perspective: flow and evolution of prob. measures

Bayesian inference and probabilistic ML

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

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



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

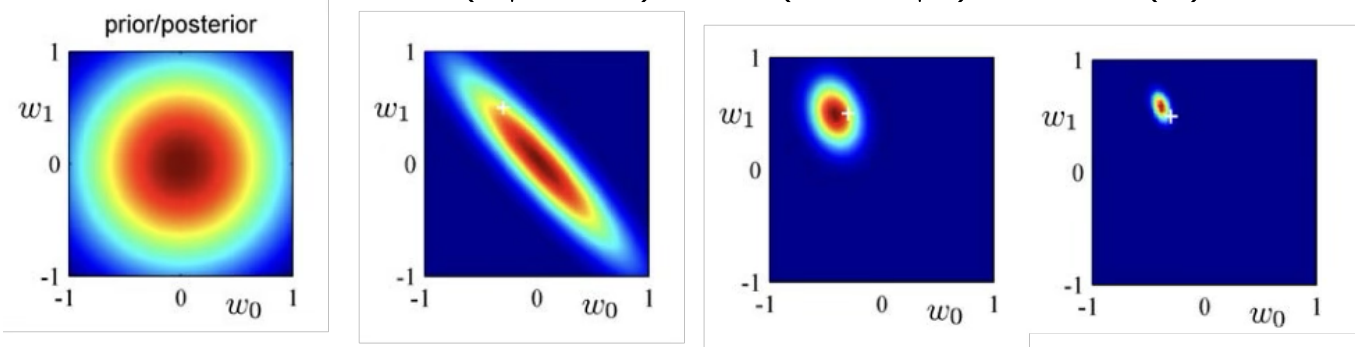
$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi(\theta|\text{Data})).$$

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,

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



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

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

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

Recall Oliver's talk:

$$\dot{\eta} = -\alpha \mathbb{G}_{\mathsf{W}}^{\mathsf{red}\,-1}(\eta) \nabla_\eta F(\mu(\eta)) - \beta \mathbb{G}_{\mathsf{He}}^{\mathsf{red}\,-1}(\eta) \nabla_\eta F(\mu(\eta))$$

where $\eta = [m, \Sigma]^\top$

Goal: General-purpose optimization over (probability) measures

Goal: General-purpose optimization over (probability) measures

Type of applications of modeling a measure π

Goal: General-purpose optimization over (probability) measures

Type of applications of modeling a measure π

Input (access to the target)

- samples:

  $y_i \sim \pi$ (target distribution)

- *score function:*

  $V(x) = -\log \pi(x), \quad \pi(x) \propto e^{-V(x)},$

Goal: General-purpose optimization over (probability) measures

Type of applications of modeling a measure π

Input (access to the target)

- samples:

  $y_i \sim \pi \quad$ (target distribution)

- *score function:*

  $$V(x) \;=\; -\log \pi(x), \quad \pi(x) \propto e^{-V(x)},$$

Output (model)

- *Sampling / Generative models:*
  Produce new draws

  $$x_i \sim \pi$$

- *Inference:*
  Approximate π by a parametric family,
  e.g.

  $$\pi(x) \approx \mathcal{N}\big(x|\, \hat{m}, \hat{S}\big)$$

Goal: General-purpose optimization over (probability) measures

Type of applications of modeling a measure π

Input (access to the target)

- samples:

  $y_i \sim \pi$ (target distribution)

- *score function:*

  $$V(x) = -\log \pi(x), \quad \pi(x) \propto e^{-V(x)},$$

Output (model)

- *Sampling / Generative models:*
  Produce new draws

  $$x_i \sim \pi$$

- *Inference:*
  Approximate π by a parametric family,
  e.g.

  $$\pi(x) \approx \mathcal{N}\left(x \middle| \hat{m}, \hat{S}\right)$$

Goal. statistical computation
as optimization over (prob.) measures

$$\pi = \operatorname*{argmin}_{\mu \in A \subset \mathcal{P}} F(\mu)$$

Goal: General-purpose optimization over (probability) measures

Type of applications of modeling a measure π

Input (access to the target)

- samples:

  $y_i \sim \pi$ (target distribution)

- *score function:*

  $$V(x) = -\log \pi(x), \quad \pi(x) \propto e^{-V(x)},$$

Output (model)

- *Sampling / Generative models:*
  Produce new draws

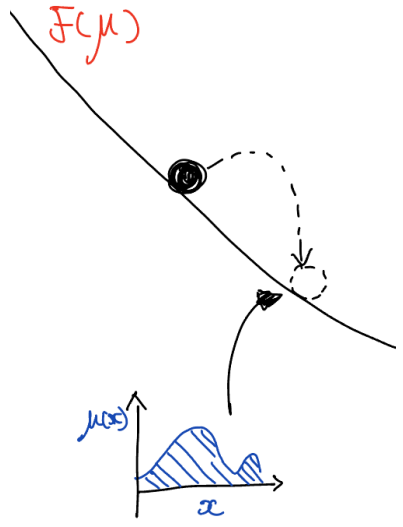
  $$x_i \sim \pi$$

- *Inference:*
  Approximate π by a parametric family,
  e.g.

  $$\pi(x) \approx \mathcal{N}\left(x \middle| \hat{m}, \hat{S}\right)$$

Goal. statistical computation
as optimization over (prob.) measures

$$\pi = \operatorname*{argmin}_{\mu \in A \subset \mathcal{P}} F(\mu)$$



Goal: General-purpose optimization over (probability) measures

Type of applications of modeling a measure π

Input (access to the target)

- samples:

  $y_i \sim \pi$ (target distribution)

- *score function:*

  $$V(x) = -\log \pi(x), \quad \pi(x) \propto e^{-V(x)},$$

Output (model)

- *Sampling / Generative models:*
  Produce new draws

  $$x_i \sim \pi$$

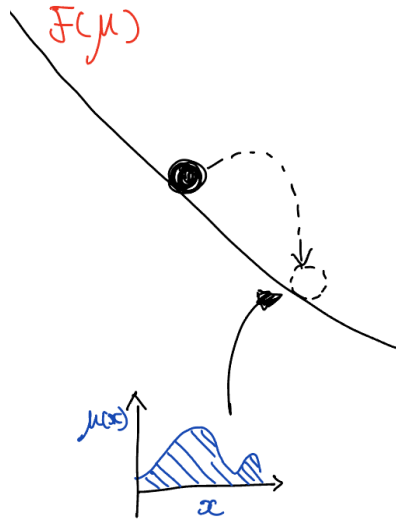
- *Inference:*
  Approximate π by a parametric family,
  e.g.

  $$\pi(x) \approx \mathcal{N}\left(x \middle| \hat{m}, \hat{S}\right)$$

Goal. statistical computation
as optimization over (prob.) measures

$$\pi = \operatorname*{argmin}_{\mu \in A \subset \mathcal{P}} F(\mu)$$

$$F = \mathrm{D}_\varphi := \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right) \mathrm{d}\pi$$



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

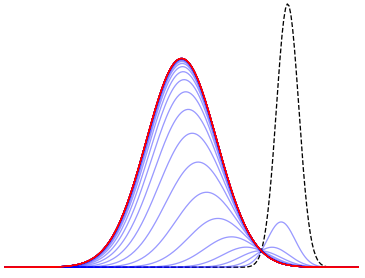
Wasserstein/diffusion: mass-preserving

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

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

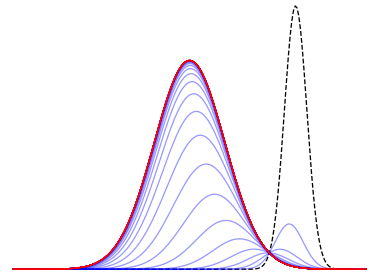
Wasserstein/diffusion: mass-preserving

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

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

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

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



Example. $F(\mu) := \mathrm{KL}(\mu|\pi)$

Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

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

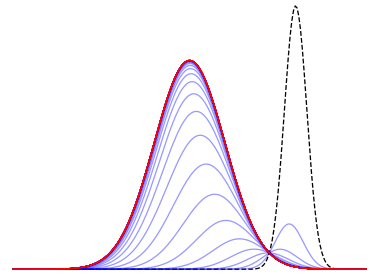
Wasserstein/diffusion: mass-preserving

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

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

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

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



Example. $F(\mu) := \mathrm{KL}(\mu|\pi)$

$$\nexists c > 0 : \left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu\|\pi),\ \forall \mu \in \mathcal{M}^+$$

Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

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

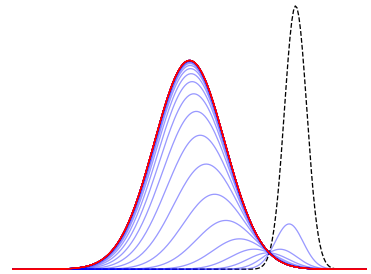
Wasserstein/diffusion: mass-preserving

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

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

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

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



Example. $F(\mu) := \mathrm{KL}(\mu|\pi)$ no global PL / gradient dominance!

$$\nexists c > 0 : \left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu\|\pi),\ \forall \mu \in \mathcal{M}^+$$

Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

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

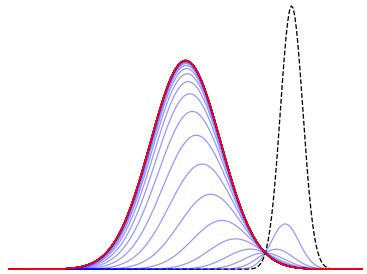
Wasserstein/diffusion: mass-preserving

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

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

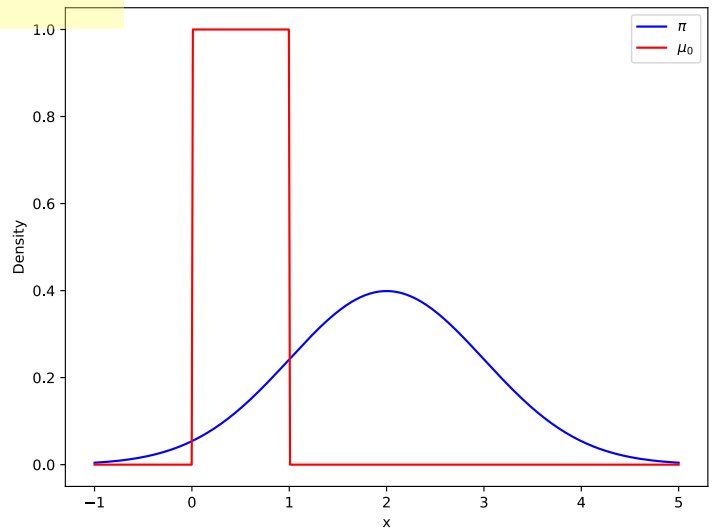
$$\min_{\mu\in\mathcal{M}^+} F(\mu) + \frac{1}{2\tau}\mathrm{He}^2(\mu, \mu^k)$$

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



Example. $F(\mu) := \mathrm{KL}(\mu|\pi)$ no global PL / gradient dominance!

$$\nexists c > 0 : \left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu\|\pi),\ \forall \mu \in \mathcal{M}^+$$



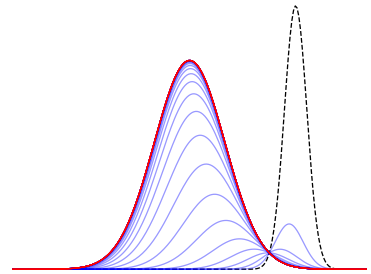
Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

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

Wasserstein/diffusion: mass-preserving

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

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

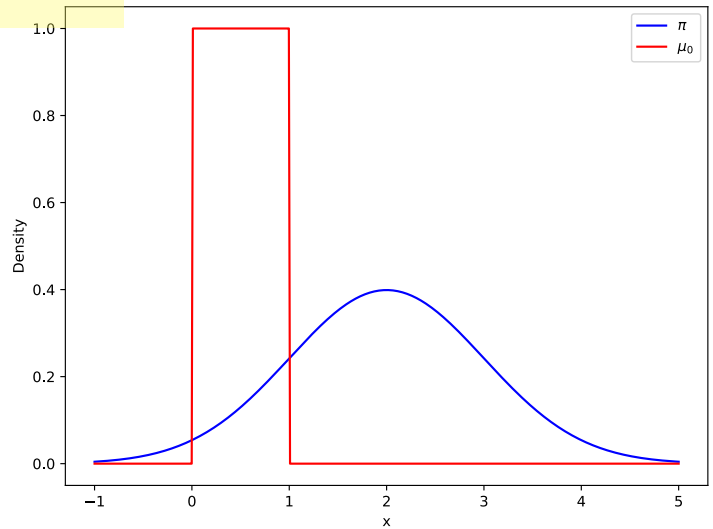


$$\min_{\mu\in\mathcal{M}^+} F(\mu) + \frac{1}{2\tau}\mathrm{He}^2(\mu, \mu^k)$$

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

Example. $F(\mu) := \mathrm{KL}(\mu|\pi)$ no global PL / gradient dominance!

$$\nexists c > 0 : \left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu\|\pi), \ \forall \mu \in \mathcal{M}^+$$



ODE solution: $\mu(t,x) = \pi(x)\left(\frac{\mathrm{d}\mu(0,\cdot)}{\mathrm{d}\pi}(x)\right)^{\mathrm{e}^{-t}}$

Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

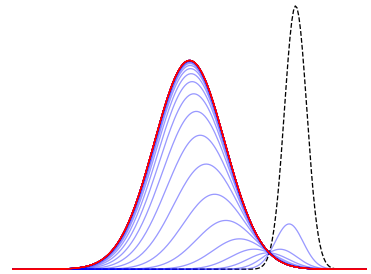
Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

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

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

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



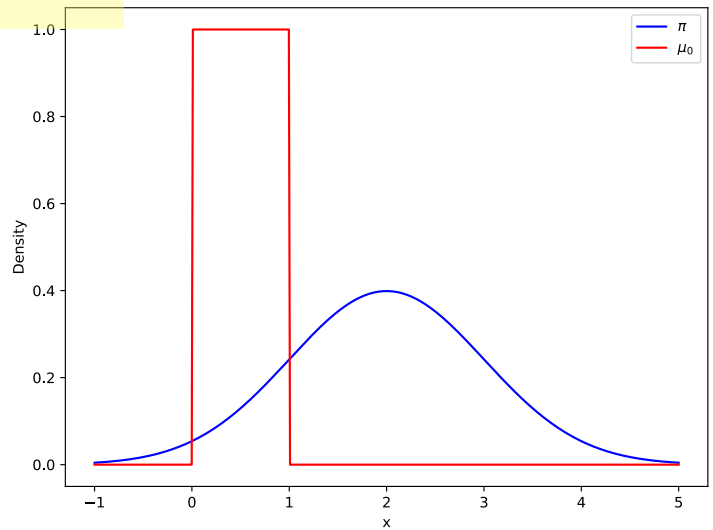
$$\text{continuous-time } \tau \to 0 : \dot{\mu} = -\mu \cdot \frac{\delta F}{\delta \mu}[\mu]$$

Example. $F(\mu) := \mathrm{KL}(\mu|\pi)$ no global PL / gradient dominance!

$$\nexists c > 0 : \left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|^2_{L^2_\mu} \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu\|\pi), \ \forall \mu \in \mathcal{M}^+$$

ODE solution: $\mu(t,x) = \pi(x)\left(\frac{\mathrm{d}\mu(0,\cdot)}{\mathrm{d}\pi}(x)\right)^{\mathrm{e}^{-t}}$

KL is non-convex in Hell.



Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

Recent Advances on the analysis of Fisher-Rao flows

The gradient system $(\mathcal{P}, \mathrm{D}_{\varphi_p}(\cdot|\pi), \mathsf{FR})$ generates the flow equation

$$\dot{\rho} = -\rho\left(\varphi_p'(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}) - \int \varphi_p'(\frac{\mathrm{d}\rho}{\mathrm{d}\pi})\,\mathrm{d}\rho\right)$$

Recall the φ_p-divergence

$$\mathrm{D}_{\varphi_p}(\rho|\pi) = \int \varphi_p(\frac{\mathrm{d}\rho}{\mathrm{d}\pi})\,\mathrm{d}\rho, \quad \varphi_p(s) := \frac{1}{p(p-1)}\left(s^p - p(s-1) - 1\right)$$

- **Carrillo et al. 2024:**
  - No geodesic convexity or PL (Polyak-Łojasiewicz) condition for the functional $\mathrm{KL}(\cdot|\pi)$, even if π is Gaussian
  - Sufficient condition: φ_p-divergence satisfies the PL inequality for $p \leq 0$.
- **Mielke & Zhu 2025:**
  - Power threshold $p \leq \frac{1}{2}$ further refines Carrillo et al. results. ($p = \frac{1}{2}$ recovers Fisher-Rao distance.)
  - Explicit Lyapunov functions $\Longrightarrow$ dual gradient dominance of Carrillo et al. 2024.

Carrillo et al. Fisher-Rao gradient flow: geod. convexity and functional ineq.
Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals.

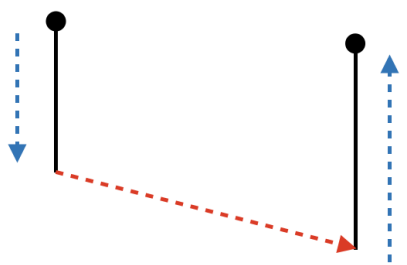
Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

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.

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



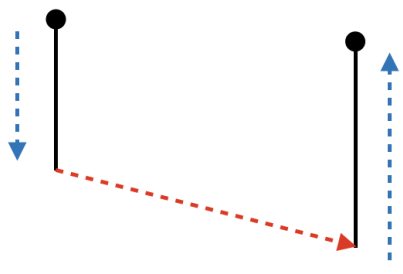
Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

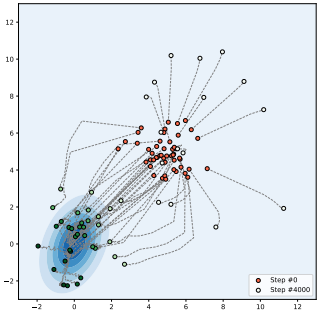
Best of both worlds

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.

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





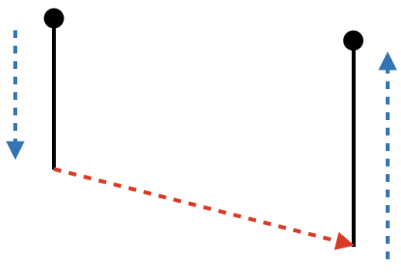
Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

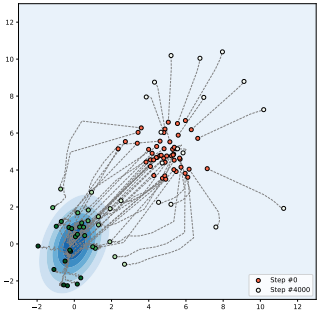
Best of both worlds

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.

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





Wasserstein WFR

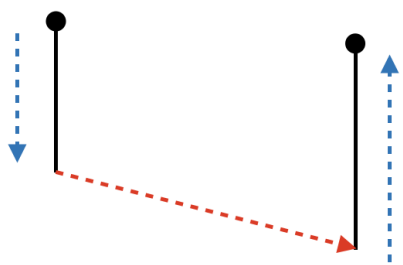
Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

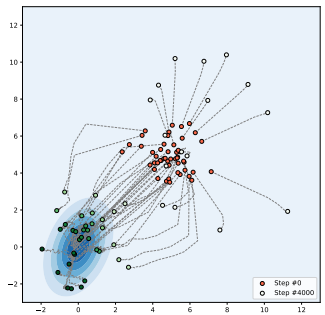
Best of both worlds

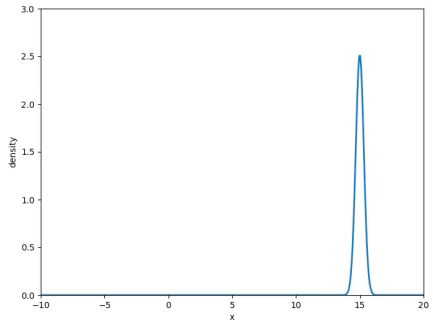
Wasserstein: transport/diffusion Fisher-Rao / Hellinger: growth/birth-death

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

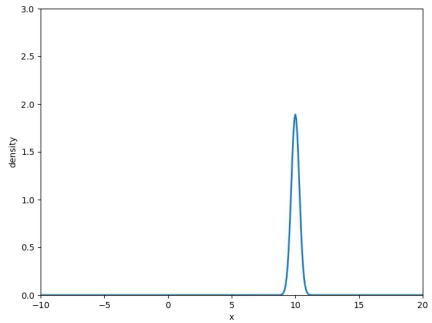
$$\dot{\mu} = -\alpha \mathbb{G}_W^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] - \beta \mathbb{G}_{\text{He}}^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] = \alpha \text{div}\,(\mu \nabla \frac{\delta F}{\delta \mu}) - \beta \mu \frac{\delta F}{\delta \mu}$$







Wasserstein



WFR

Liero-Mielke-Tse-**Z.** Evolution of Gaussians in the HK-Boltzmann gradient flow.

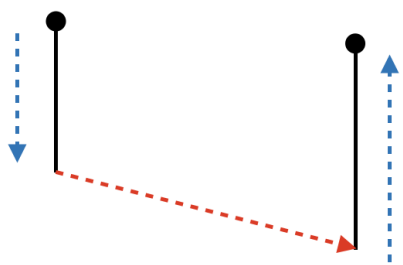
Unbalanced transport: Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao

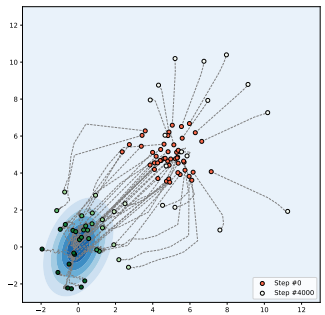
Best of both worlds

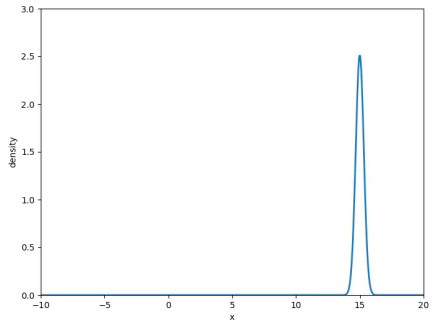
Wasserstein: transport/diffusion Fisher-Rao / Hellinger: growth/birth-death

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

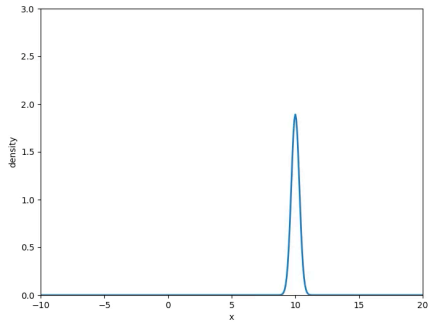
$$\dot{\mu} = -\alpha \mathbb{G}_W^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] - \beta \mathbb{G}_{\text{He}}^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] = \alpha \text{div}\,(\mu \nabla \frac{\delta F}{\delta \mu}) - \beta \mu \frac{\delta F}{\delta \mu}$$







Wasserstein



WFR

If $F(\mu) := \text{KL}(\mu|\pi)$, can we establish decay of $F(\mu(t))$ over time?

Liero-Mielke-Tse-**Z.** Evolution of Gaussians in the HK-Boltzmann gradient flow.

Gradient Dominance/PL of Gradient Flows of Entropy Functional

Gradient Dominance/PL of Gradient Flows of Entropy Functional

$\min_{\mu} \mathrm{D}_{\varphi_p}(\mu|\pi)$ in $(\mathcal{P}$ or $\mathcal{M}^+, \mathrm{D})$

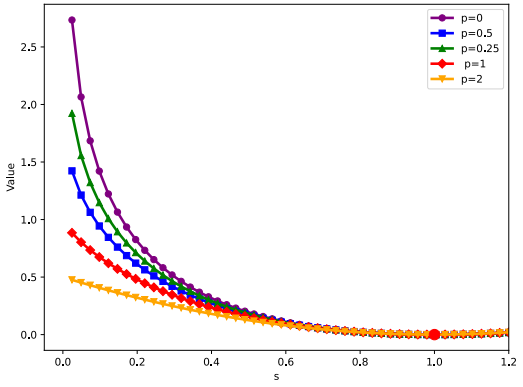
Gradient Dominance/PL of Gradient Flows of Entropy Functional

$$\min_{\mu} \mathrm{D}_{\varphi_p}(\mu|\pi) \text{ in } (\mathcal{P} \text{ or } \mathcal{M}^+, \mathrm{D}) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P} \text{ or } \mathcal{M}^+} \mathrm{D}_{\varphi_p}(\mu|\pi) + \frac{1}{2\tau}\mathrm{D}^2(\mu, \mu_k)$$

Gradient Dominance/PL of Gradient Flows of Entropy Functional

$$\min_{\mu} \mathrm{D}_{\varphi_p}(\mu|\pi) \text{ in } (\mathcal{P} \text{ or } \mathcal{M}^+, \mathrm{D}) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P} \text{ or } \mathcal{M}^+} \mathrm{D}_{\varphi_p}(\mu|\pi) + \frac{1}{2\tau}\mathrm{D}^2(\mu, \mu_k)$$

$$\mathrm{D}_{\varphi_p}(\mu|\nu) := \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu \quad \varphi_p(s) := \frac{1}{p(p-1)}\left(s^p - p(s-1) - 1\right)$$





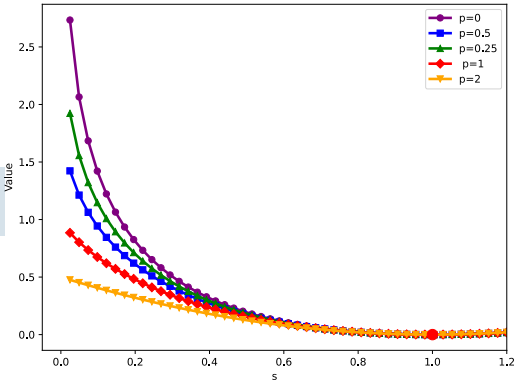
based on joint works
with Alexander Mielke

Gradient Dominance/PL of Gradient Flows of Entropy Functional

$$\min_{\mu} \mathrm{D}_{\varphi_p}(\mu|\pi) \text{ in } (\mathcal{P} \text{ or } \mathcal{M}^+, \mathrm{D}) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P} \text{ or } \mathcal{M}^+} \mathrm{D}_{\varphi_p}(\mu|\pi) + \frac{1}{2\tau}\mathrm{D}^2(\mu, \mu_k)$$

$$\mathrm{D}_{\varphi_p}(\mu|\nu) := \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu \quad \varphi_p(s) := \frac{1}{p(p-1)}\left(s^p - p(s-1) - 1\right)$$

Gradient-flow geometry	Global exp. decay, & fcn. ineq. for φ_p-divergence
Otto-Wasserstein on $\mathcal{P}$	• $\Omega \subsetneq \mathbb{R}^d$ bounded Lipschitz, $p \geq 1 - \frac{1}{d} \implies$ Ł with $c_* > 0$ • $\Omega = \mathbb{R}^d, p \in [1, 2]$ and (BE) $\implies$ Ł with $c_* = 2c_{\mathrm{BE}}$
Otto-Wasserstein on $\mathcal{M}^+$ (Prop. 5.2, 5.3)	$\nexists\, c_* > 0$ for Ł; see (LSI-$\mathcal{M}^+$)
Hellinger on $\mathcal{M}^+$ (Prop. 3.7)	$p \in (-\infty, \frac{1}{2}] \iff$ Ł with $c_* = \frac{1}{1-p}$
Spherical Hellinger on $\mathcal{P}$ (Thm. 4.1)	$p \in (-\infty, \frac{1}{2}] \iff$ Ł with $c_* = M_p := \begin{cases} \frac{1}{1-p} & \text{for } p \leq \frac{1}{3} \\ \frac{p(7-12p)}{1-p} & \text{for } p \in [\frac{1}{3}, \frac{1}{2}] \end{cases}$
Hellinger-Kantorovich on $\mathcal{M}^+$ (Thm. 5.8)	• $p \in (-\infty, \frac{1}{2}] \implies$ Ł with $c_* = \frac{1}{1-p}$ • $p > \frac{1}{2} \implies$ there exists no with $c_* > 0$ • $p = 1$ and (LSI) $\implies$ No Ł; exp. decay is possible
Spherical Hellinger-Kantorovich on $\mathcal{P}$ (Thm. 4.4)	In general, decay rate $c_* = \max\{\alpha\, c_{\text{Ł-W}}\,,\ \beta\, M_p\}$ (see Thm. 4.4). Specifically: • $p \in (-\infty, \frac{1}{2}] \implies$ Ł with $c_* = \frac{1+2p}{1-p}$ • $\Omega \subsetneq \mathbb{R}^d$ bounded Lipschitz, $p \in (-\infty, 1/2] \cup [1 - \frac{1}{d}, \infty) \implies$ Ł with $c_* > 0$ • $\Omega = \mathbb{R}^d,\ p \in [1, 2]$ and (Ł-W) $\implies$ Ł with $c_* = 2c_{\text{Ł-W}}$





based on joint works with Alexander Mielke

Mielke, A., **Z.** (2025). Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals

Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$

Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

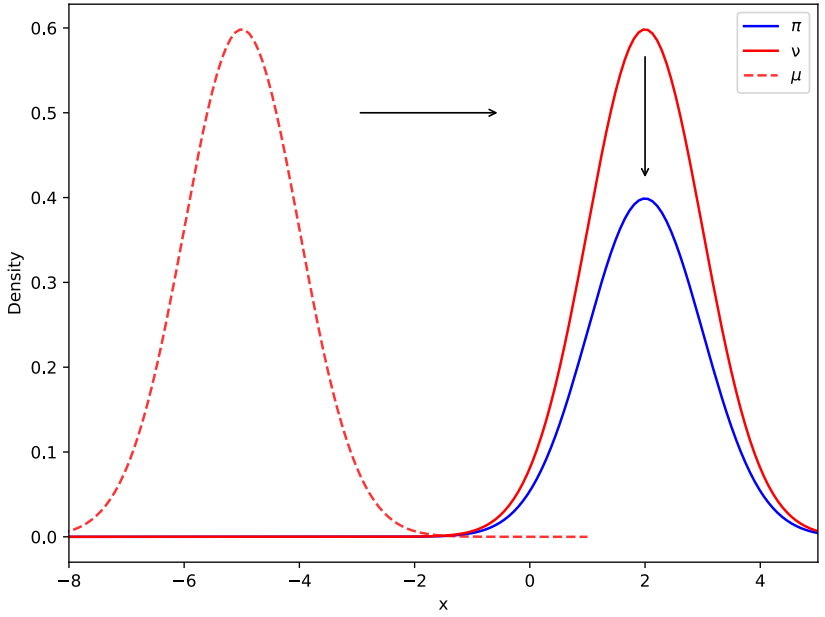
PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$

Observation 1. Even for the KL energy, no PL-inequality for HK/WFR gradient flows [Mielke-Z.'25, Kondratyev-Vorotnikov].
Transport cannot change mass.

Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$

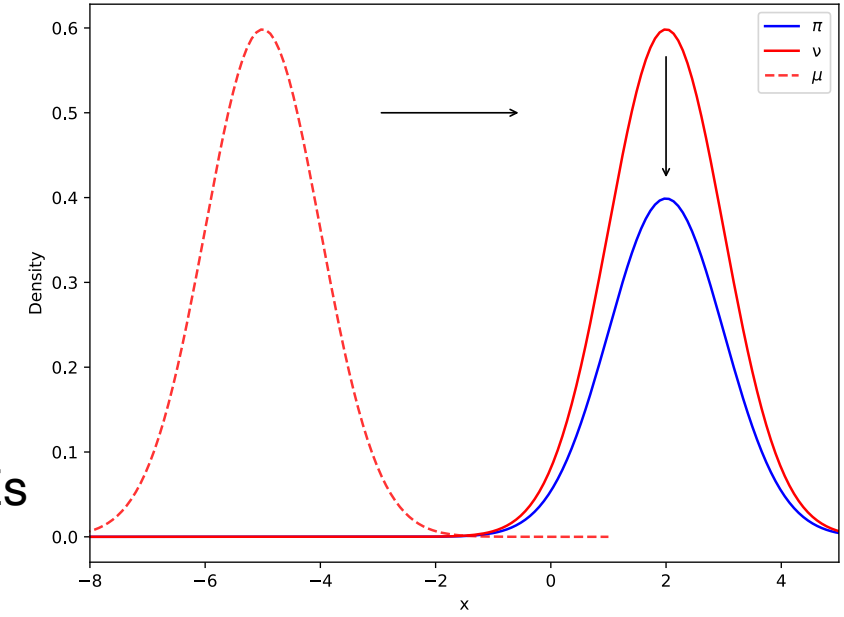
Observation 1. Even for the KL energy, no PL-inequality for HK/WFR gradient flows [Mielke-Z.'25, Kondratyev-Vorotnikov].
Transport cannot change mass.



Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$

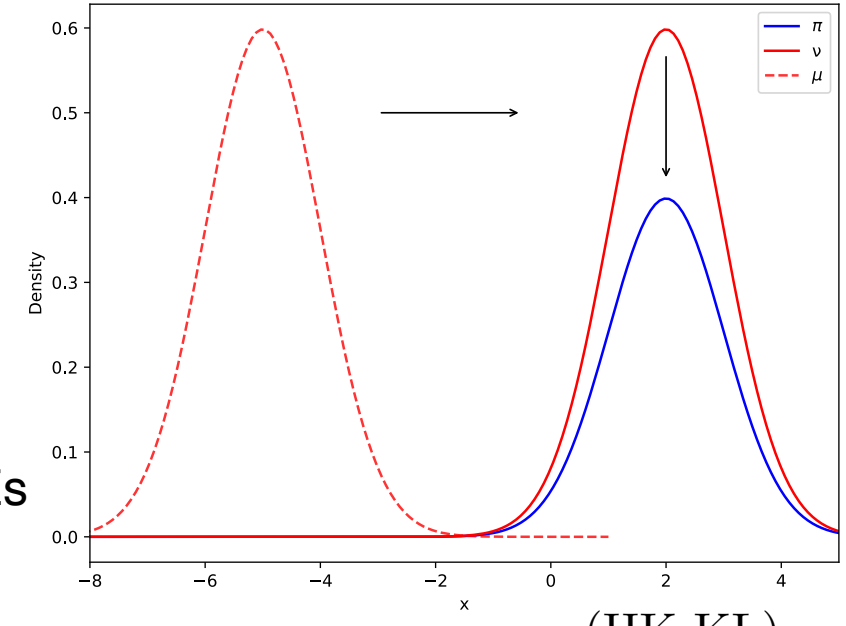
Observation 1. Even for the KL energy, no PL-inequality for HK/WFR gradient flows [Mielke-Z.'25, Kondratyev-Vorotnikov].
Transport cannot change mass.



Observation 2. the following solutions of the PDEs

Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$



Observation 1. Even for the KL energy, no PL-inequality for HK/WFR gradient flows [Mielke-Z.'25, Kondratyev-Vorotnikov].
Transport cannot change mass.

Observation 2. the following solutions of the PDEs

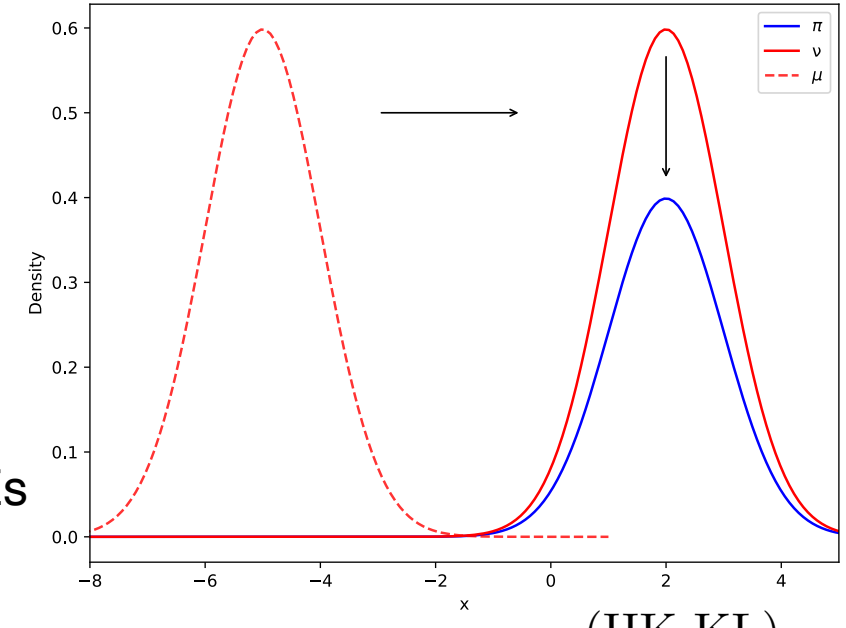
$$\dot{\mu} = \alpha \operatorname{div}\left(\nabla \mu + \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \nabla \pi\right) - \beta \mu \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right), \qquad \text{(HK-KL)}$$

$$\dot{\rho} = \alpha \operatorname{div}\left(\nabla \rho + \frac{\mathrm{d}\rho}{\mathrm{d}\pi} \nabla \pi\right) - \beta \rho \left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \int_{\mathbb{R}^d} \rho \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}x\right). \qquad \text{(SHK-KL)}$$

Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$

Observation 1. Even for the KL energy, no PL-inequality for HK/WFR gradient flows [Mielke-Z.'25, Kondratyev-Vorotnikov].
Transport cannot change mass.



Observation 2. the following solutions of the PDEs

$$\dot{\mu} = \alpha \operatorname{div}\left(\nabla \mu + \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \nabla \pi\right) - \beta \mu \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right), \qquad \text{(HK-KL)}$$

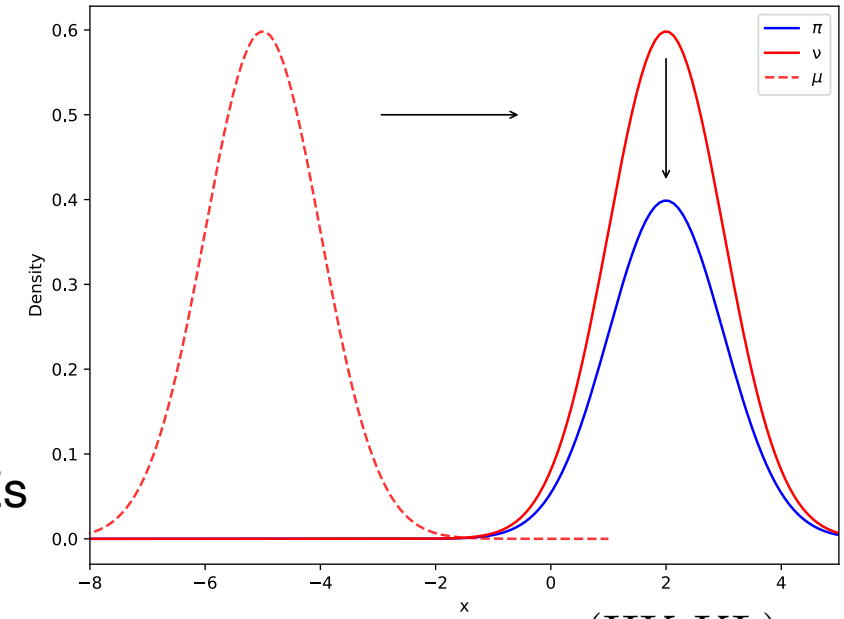
$$\dot{\rho} = \alpha \operatorname{div}\left(\nabla \rho + \frac{\mathrm{d}\rho}{\mathrm{d}\pi} \nabla \pi\right) - \beta \rho \left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \int_{\mathbb{R}^d} \rho \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}x\right). \qquad \text{(SHK-KL)}$$

correspond to each other by a mass-scaling: $\mu(t) = z(t)\rho(t)$

Observations of Hellinger-Kantorovich (a.k.a. WFR) flows of KL

PL: $$\int \left(\alpha \left| \nabla \frac{\delta F}{\delta \mu}[\mu] \right|^2 + \beta \left| \frac{\delta F}{\delta \mu}[\mu] \right|^2 \right) \mathrm{d}\mu \geq \beta c_* \left(F(\mu) - \inf_{\nu \in \mathcal{M}^+} F(\nu) \right)$$

Observation 1. Even for the KL energy, no PL-inequality for HK/WFR gradient flows [Mielke-Z.'25, Kondratyev-Vorotnikov].
Transport cannot change mass.



Observation 2. the following solutions of the PDEs

$$\dot{\mu} = \alpha \operatorname{div}\left(\nabla \mu + \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \nabla \pi\right) - \beta \mu \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right), \qquad \text{(HK-KL)}$$

$$\dot{\rho} = \alpha \operatorname{div}\left(\nabla \rho + \frac{\mathrm{d}\rho}{\mathrm{d}\pi} \nabla \pi\right) - \beta \rho \left(\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) - \int_{\mathbb{R}^d} \rho \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}x\right). \qquad \text{(SHK-KL)}$$

correspond to each other by a mass-scaling: $\mu(t) = z(t)\rho(t)$

$$\dot{z} = -\beta z \log z - \beta z \int_{\mathbb{R}^d} \rho \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}x. \qquad \text{(Mass-HK)}$$

[Mielke-Z. 2025]'s results for the HK/WFR gradient flows of KL

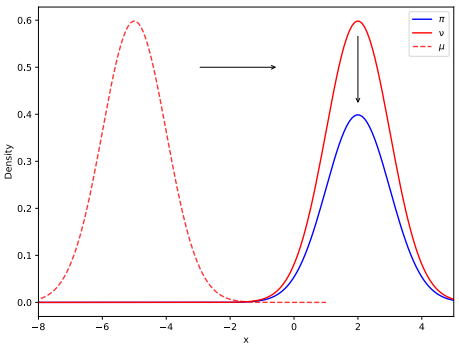
Proposition 5.7 (Shape and mass convergence) *The normalized probability measure* $\rho(t) = \frac{1}{z(t)}\mu(t)$ *(the shape) converges to the target* π_* *exponentially in KL divergence along the HK gradient flow, i.e.,*

$$\mathrm{D}_{KL}(\rho(t)|\pi_*) \le \mathrm{e}^{-\widehat{\alpha}t} H_0. \qquad \text{(shape convergence)}$$

The mass variable $z(t)$ *converges to the target mass* z_* *exponentially, i.e.,*

$$|z(t) - z_*| \le \max\{z_0, z_*\} \left|\log\left(\frac{z_0}{z_*}\right)\right| \mathrm{e}^{-\beta t} + H_0 \frac{\mathrm{e}^{-\widehat{\alpha}t} - \mathrm{e}^{-\beta t}}{\beta - \widehat{\alpha}}. \qquad \text{(mass convergence)}$$

Setting $H_0 = \mathrm{D}_{\mathrm{KL}}(\rho(0)|\pi_*)$ and $\widehat{\alpha} = \alpha c_{\mathrm{LSI}\text{-}\mathcal{P}}$



Theorem 5.8 (Convergence to equilibrium via shape-mass analysis) *The following convergence estimate in the Hellinger distance holds*

$$\mathsf{He}(\mu(t), \pi) \le \left(\frac{\max\{z_0^{1/2}, z_*^{1/2}\}}{2} H_0^{1/2} + z_*^{1/2} \left(g\left(\left(\frac{z_0}{z_*} \right)^{1/2} \right) + \frac{1}{\widehat{\alpha}} \right) \right) \mathrm{e}^{-\gamma t} \quad \textit{for } t > 0, \tag{5.7}$$

where $\gamma = \min\left\{\beta, \widehat{\alpha}/2\right\}$ *and* $g(a) = \max\{\log(1/a), a-1\} \ge 0$.

Application 1

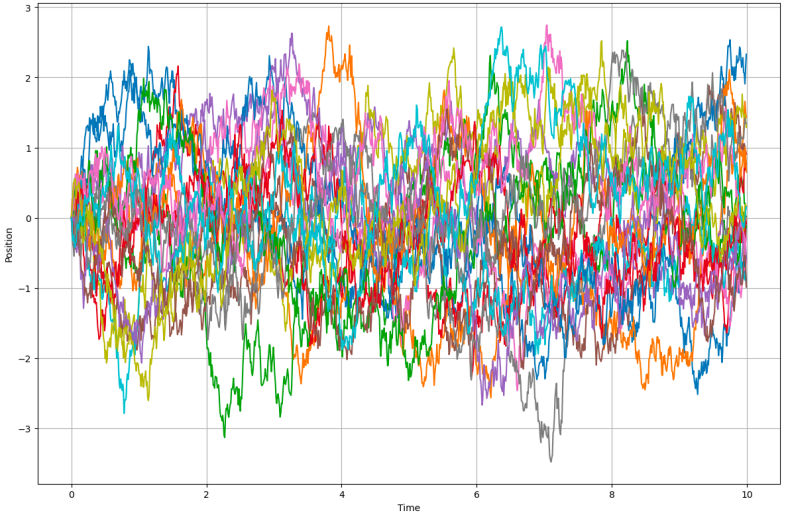
Sampling via particle systems

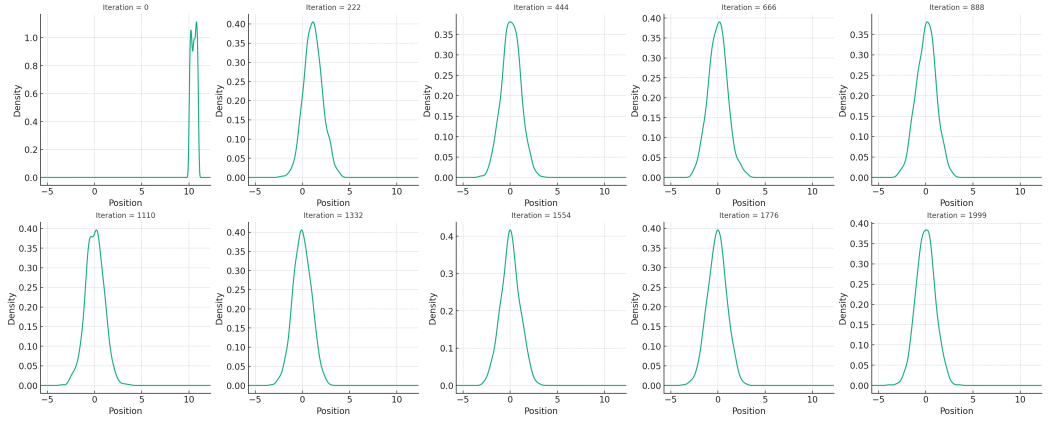
Sampling via particle systems

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

Sampling via particle systems

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





Langevin SDE

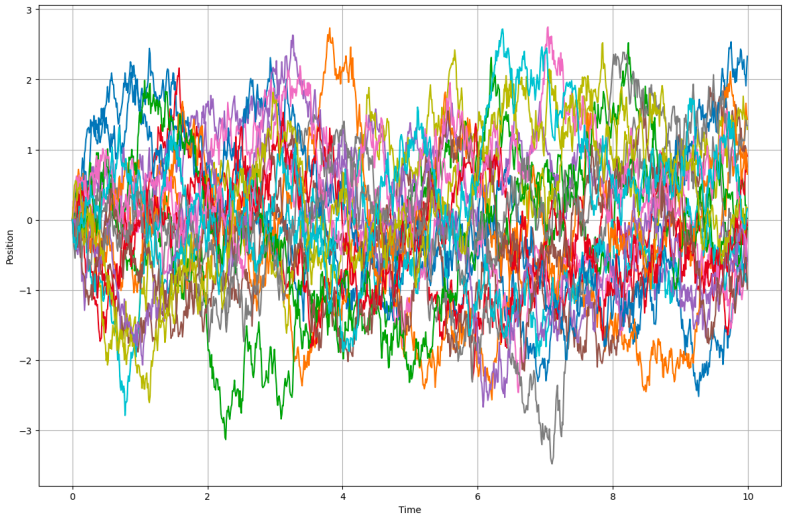
$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

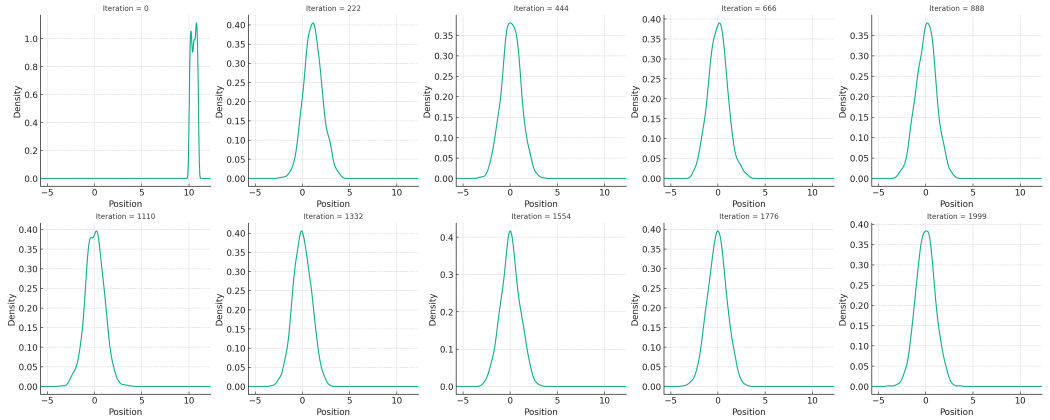
Fokker-Planck PDE

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

Sampling via particle systems

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





Langevin SDE

Fokker-Planck PDE

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

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

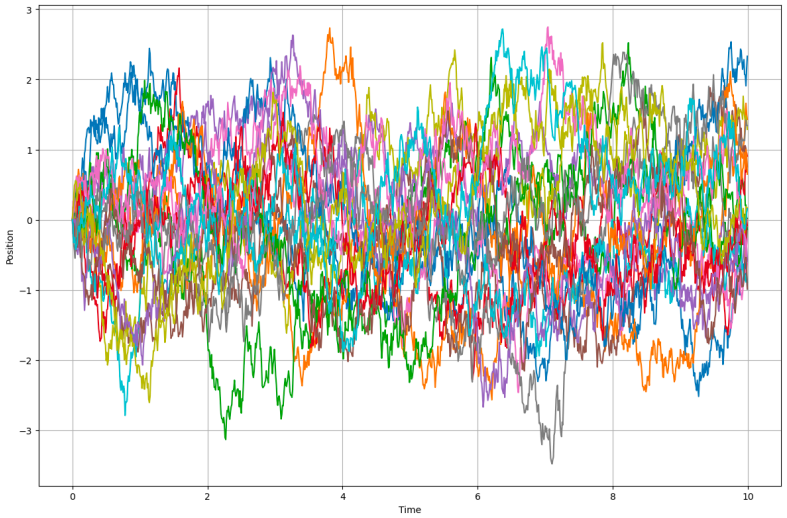
GF & OPT perspective

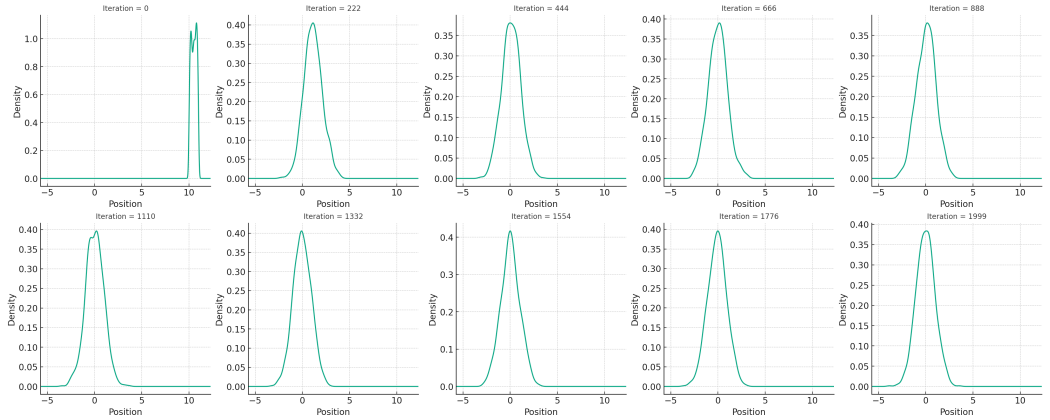
$$\mu_{k+1} \leftarrow \underset{\mu \subset \mathcal{P}}{\mathrm{argmin}}\, \mathrm{D}_{\mathrm{KL}}(\mu|\pi) + \frac{1}{2\tau} W_2^2(\mu, \mu_k)$$

(let $\tau \to 0$)

Sampling via particle systems

Goal: generating samples 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

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

GF & OPT perspective

$$\mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi) + \frac{1}{2\tau} W_2^2(\mu, \mu_k)$$

(let $\tau \to 0$)

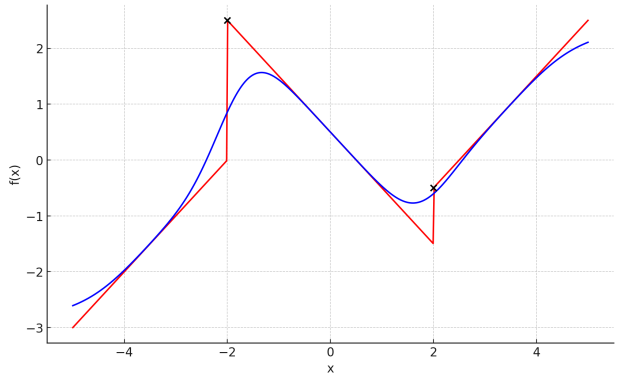
hope: for some $\alpha > 0$

$$\mathrm{D}_{\mathrm{KL}}(\mu_t|\pi) \le \mathrm{e}^{-\alpha t}\,\mathrm{D}_{\mathrm{KL}}(\mu_0|\pi)$$

Kernel methods and MMD

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

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

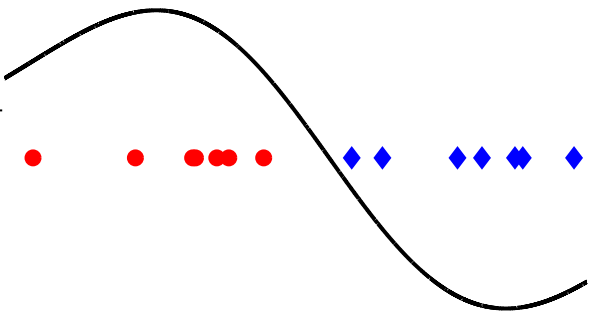


Integral operator $\mathcal{K}_\rho : L^2(\rho) \to L^2(\rho)$:

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

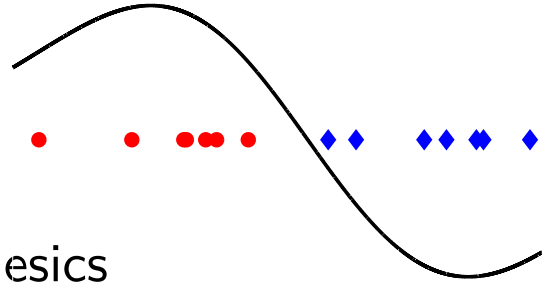
Maximum-mean discrepancy (MMD) [Gretton et al., 2012]

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



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

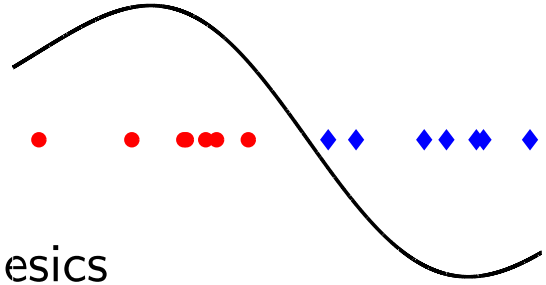
MMD as “de-kernelized” Hellinger distance



Dyanmic formulation of MMD: “straight line” geodesics

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

MMD as “de-kernelized” Hellinger distance



Dyanmic formulation of MMD: “straight line” geodesics

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

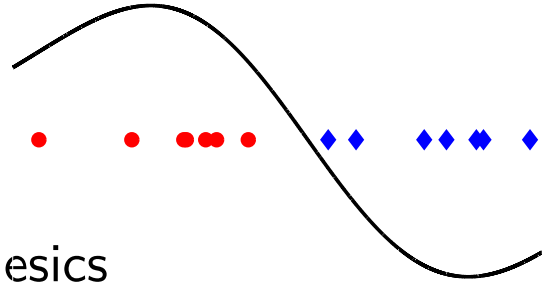
Theorem (MMD = de-kernelized Hellinger)

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

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

Z-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows; Gladin et al. & Z. 2024

MMD as “de-kernelized” Hellinger distance



Dyanmic formulation of MMD: “straight line” geodesics

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

Theorem (MMD = de-kernelized Hellinger)

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

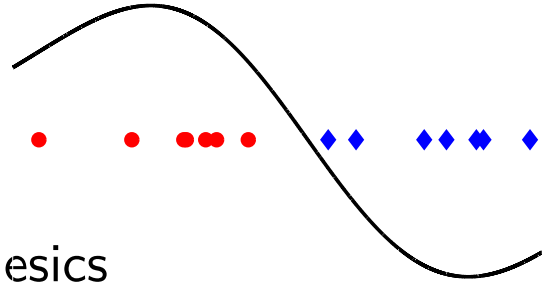
The Riemannian metric tensors are related by $\mathbb{G}_{\mathrm{MMD}} = \mathcal{K}_\mu \circ \mathbb{G}_{\mathrm{He}}(\mu)$.

recall:

$$\mathrm{He}^2(\mu,\nu) = \min\left\{ \int_0^1 \|\xi\|_{L^2_u}^2 \,\mathrm{d}t \;\middle|\; \dot{u} = -u\cdot\xi,\;\; u(0)=\mu,\;\; u(1)=\nu \right\}$$

Z-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows; Gladin et al. & Z. 2024

MMD as "de-kernelized" Hellinger distance



Dyanmic formulation of MMD: "straight line" geodesics

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

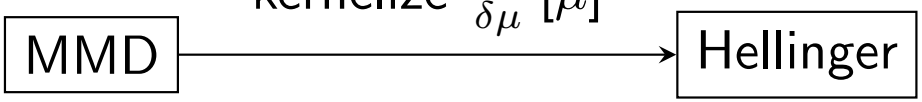
Theorem (MMD = de-kernelized Hellinger)

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

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

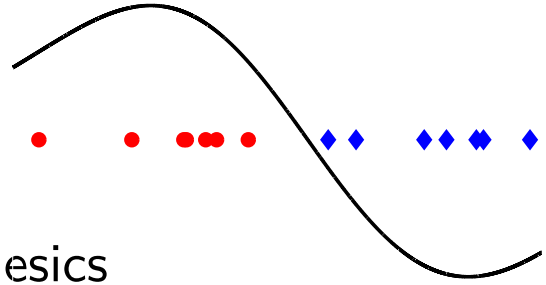
recall:

$$\mathrm{He}^2(\mu,\nu) = \min\left\{ \int_0^1 \|\xi\|_{L^2_u}^2 \,\mathrm{d}t \;\middle|\; \dot{u} = -u\cdot\xi,\ u(0)=\mu,\ u(1)=\nu \right\}$$



Z-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows; Gladin et al. & Z. 2024

MMD as "de-kernelized" Hellinger distance



Dyanmic formulation of MMD: "straight line" geodesics

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

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 $\mathbb{G}_{\mathrm{MMD}} = \mathcal{K}_\mu \circ \mathbb{G}_{\mathrm{He}}(\mu)$.

recall:

$$\mathrm{He}^2(\mu,\nu) = \min\left\{ \int_0^1 \|\xi\|_{L^2_u}^2 \,\mathrm{d}t \;\middle|\; \dot{u} = -u\cdot\xi,\;\; u(0)=\mu,\;\; u(1)=\nu \right\}$$

MMD —kernelize $\frac{\delta F}{\delta \mu}\,[\mu]$→ Hellinger

Wasserstein —kernelize $\nabla\frac{\delta F}{\delta \mu}\,[\mu]$→ Stein

Z-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows; Gladin et al. & Z. 2024

Forward (incl.) KL inference as kernelized Wasserstein flows

Forward (incl.) KL inference as kernelized Wasserstein flows

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

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

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

Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Otto-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

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

Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Otto-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

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

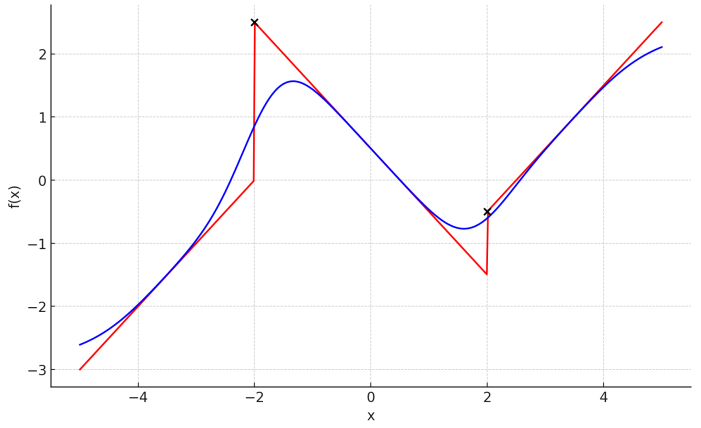
Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

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

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

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

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



Forward (incl.) KL inference as kernelized Wasserstein flows

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

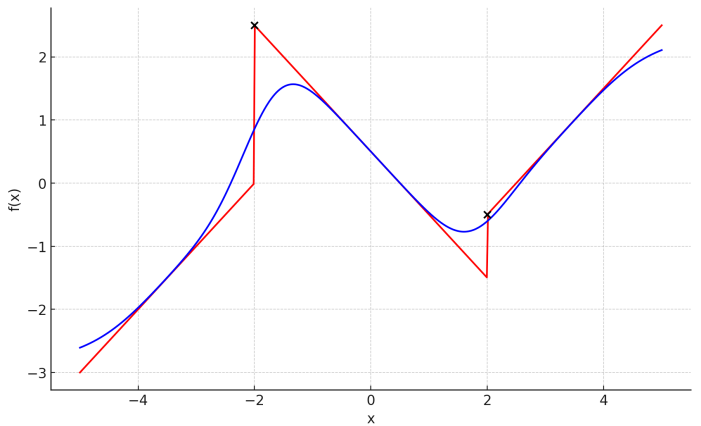
Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Otto-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:

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

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

Forward (incl.) KL inference as kernelized Wasserstein flows

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

Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

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

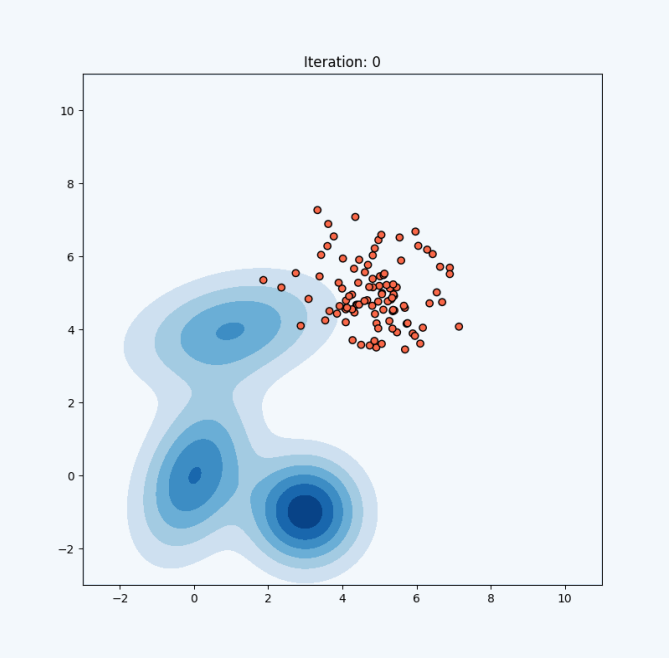
Not implementable due to $\nabla(1 - \mathrm{d}\pi/\mathrm{d}\mu)$.

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:

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



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

Unbalanced transport gradient flows of forward (incl.) KL inference

Unbalanced transport gradient flows of forward (incl.) KL inference

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

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

Unbalanced transport gradient flows of forward (incl.) KL inference

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

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

The latter is studied by [Arbel et al. 2019, Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

Unbalanced transport gradient flows of forward (incl.) KL inference

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

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

The latter is studied by [Arbel et al. 2019, Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

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

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

Unbalanced transport gradient flows of forward (incl.) KL inference

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

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

The latter is studied by [Arbel et al. 2019, 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**

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

Note: choosing the Stein kernel, we can also perform score-based inference
Z. Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective
Gladin et al. & **Z.** Interaction-Force Transport Gradient Flows

Unbalanced transport gradient flows of forward (incl.) KL inference

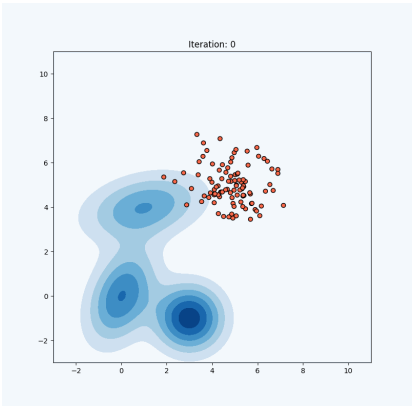
[Z. 2024]: precise connection between the gradient flows for:

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

The latter is studied by [Arbel et al. 2019, 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**

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

Note: choosing the Stein kernel, we can also perform score-based inference
Z. Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective
Gladin et al. & **Z.** Interaction-Force Transport Gradient Flows

Unbalanced transport gradient flows of forward (incl.) KL inference

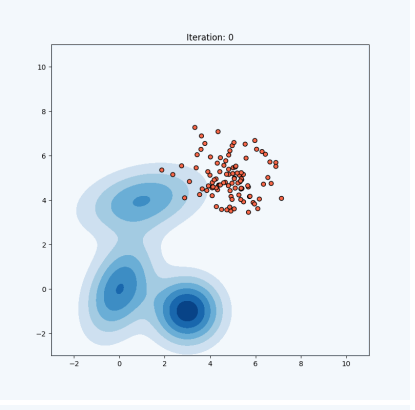
[Z. 2024]: precise connection between the gradient flows for:

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

The latter is studied by [Arbel et al. 2019, Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

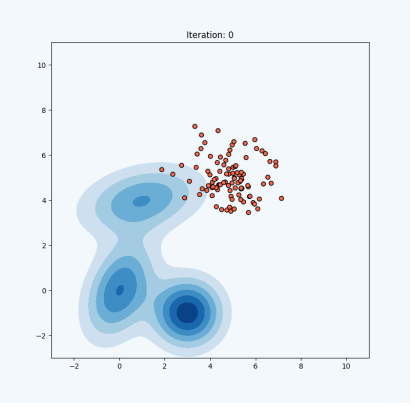
[Z. 2024] Wasserstein-Fisher-Rao flow: reaction-diffusion eq:

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



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

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



Note: choosing the Stein kernel, we can also perform score-based inference
Z. Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective
Gladin et al. & **Z.** Interaction-Force Transport Gradient Flows

Flow in the MMD space = kernelized Hellinger/Fisher-Rao

Minimize the MMD in the HK/WFR:

$$\min_{\mu \in \mathcal{P}} \left\{ F(\mu) := \mathrm{MMD}^2(\mu, \pi) \right\}$$

Flow in the MMD space = kernelized Hellinger/Fisher-Rao

Minimize the MMD in the HK/WFR:

$$\min_{\mu \in \mathcal{P}} \left\{ F(\mu) := \mathrm{MMD}^2(\mu, \pi) \right\}$$

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

Flow in the MMD space = kernelized Hellinger/Fisher-Rao

Minimize the MMD in the HK/WFR:

$$\min_{\mu\in\mathcal{P}}\left\{ F(\mu) := \mathrm{MMD}^2(\mu,\pi)\right\}$$

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

$$\overset{[\mathrm{ZM'24}]}{=} -\alpha\cdot \mathbb{G}_W^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] - \beta\cdot \mathbb{G}_{\mathrm{MMD}}^{-1}\frac{\delta F}{\delta \mu}[\mu]$$

Flow in the MMD space = kernelized Hellinger/Fisher-Rao

Minimize the MMD in the HK/WFR:

$$\min_{\mu\in\mathcal{P}}\left\{ F(\mu) := \mathrm{MMD}^2(\mu,\pi)\right\}$$

$$\begin{aligned}
\dot{\mu} &= -\alpha\cdot \mathbb{G}_W^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] - \beta\cdot \mathbb{G}_{\mathrm{He}}^{-1}(\mu)\circ \mathcal{K}_\mu^{-1}\frac{\delta F}{\delta \mu}[\mu] \\
&\overset{\text{[ZM'24]}}{=} -\alpha\cdot \mathbb{G}_W^{-1}(\mu)\frac{\delta F}{\delta \mu}[\mu] - \beta\cdot \mathbb{G}_{\mathrm{MMD}}^{-1}\frac{\delta F}{\delta \mu}[\mu] \\
&\overset{F=\mathrm{MMD}^2}{=} \alpha\cdot \mathrm{div}\left(\mu\int \nabla_2 k(x,\cdot)\,\mathrm{d}\,(\mu-\pi)\,(x)\right) - \beta\cdot(\mu-\pi)
\end{aligned}$$

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 $\forall \mu \in \mathcal{P}$:

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

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

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

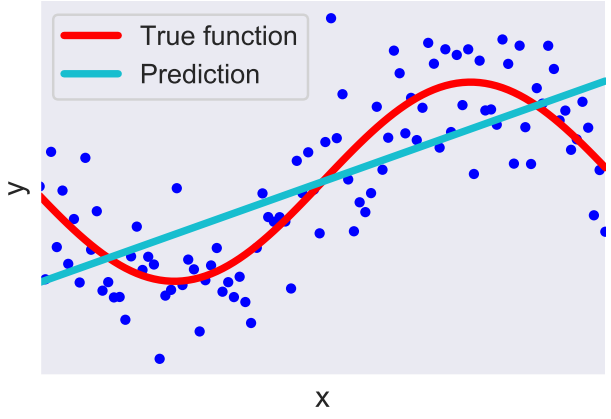
In addition, KL decays globally under the usual Bakry-Émery/LSI.

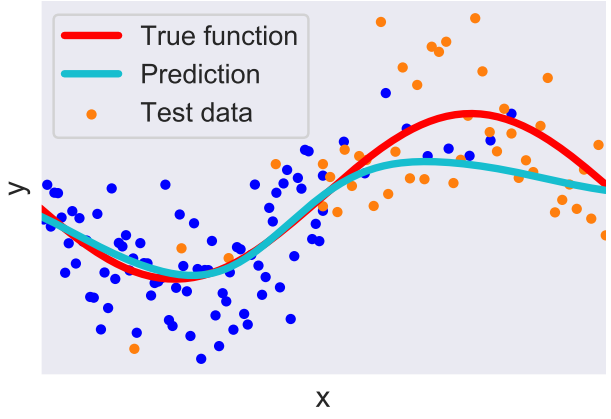
Application 2

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.



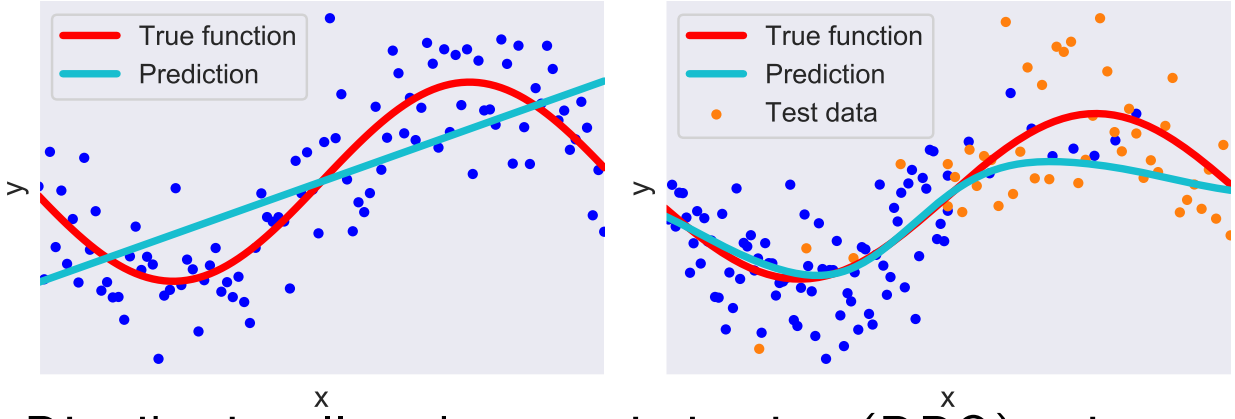


What if the test data **distribution shifts** from P_0?

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

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

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

D: divergence between measures

Wasserstein DRO: exploit c-transforms of special loss functions in ML [Esfahani-Kuhn et al., etc.]

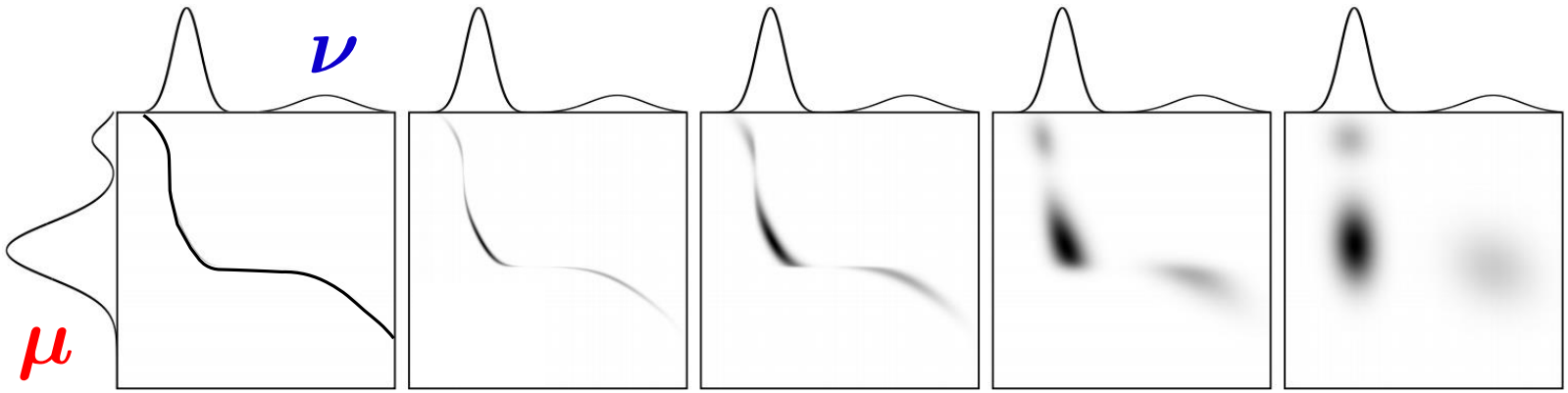
Joint entropy-regularization

$$\mathrm{OT}_\gamma\left(\Pi|\mu_0,\mu_1\right) := \left\{ \int c(x_0,x_1)\,\mathrm{d}\Pi\left(x_0,x_1\right) + \gamma\Psi(\Pi) \middle| \right.$$

$$\left. \pi_0(\mathrm{d}x_0) := \Pi\left(\mathrm{d}x_0,\Omega\right),\ \pi_1(\mathrm{d}x_1) := \Pi\left(\Omega,\mathrm{d}x_1\right) \right\},$$

$$W_\gamma^2(\mu_0,\mu_1) = \inf_{\Pi\in\mathcal{M}^+(\Omega\times\Omega)} \mathrm{OT}_\gamma\left(\Pi|\mu_0,\mu_1\right)$$

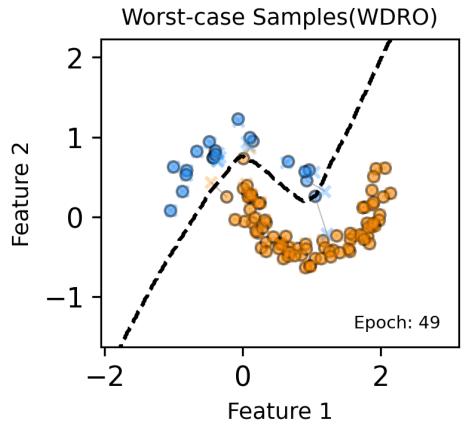
Modern rivival of optimal transport [Cuturi et al.]



Optimization problem of robust classification in entropic OT

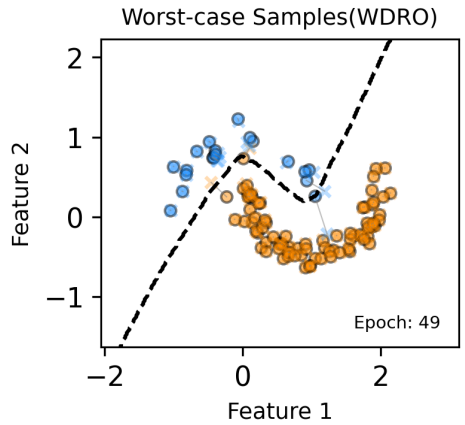
Optimization problem of robust classification in entropic OT

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_2^2(\rho, \widehat{\rho_N})$$

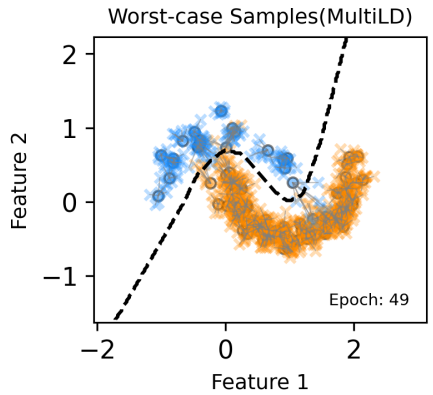


Optimization problem of robust classification in entropic OT

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_2^2(\rho, \widehat{\rho_N})$$

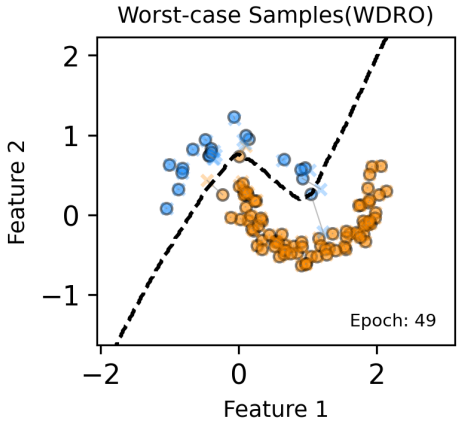


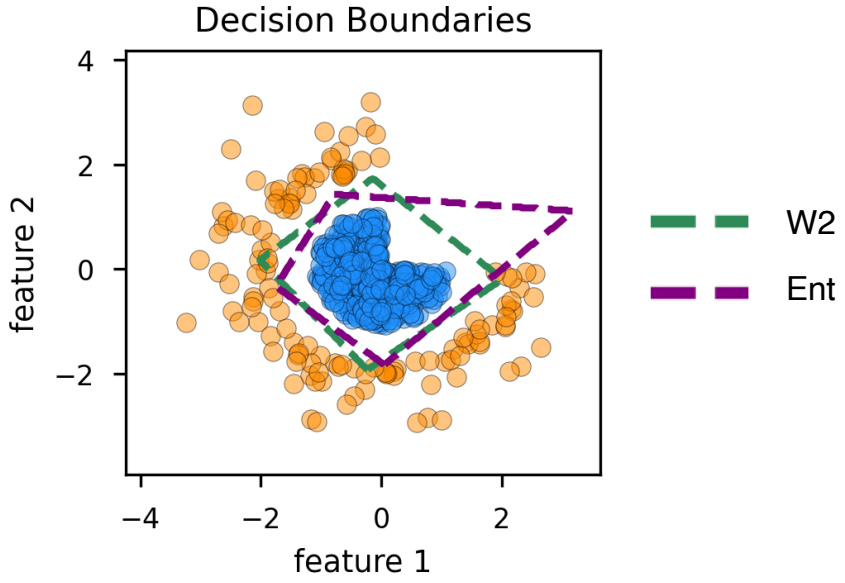
$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$



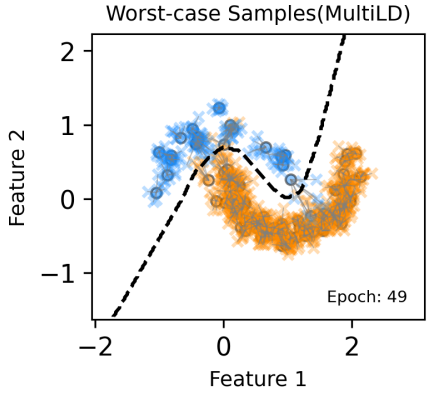
Optimization problem of robust classification in entropic OT

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_2^2(\rho, \widehat{\rho_N})$$

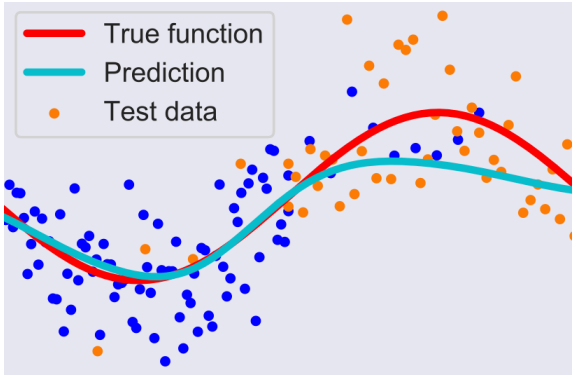




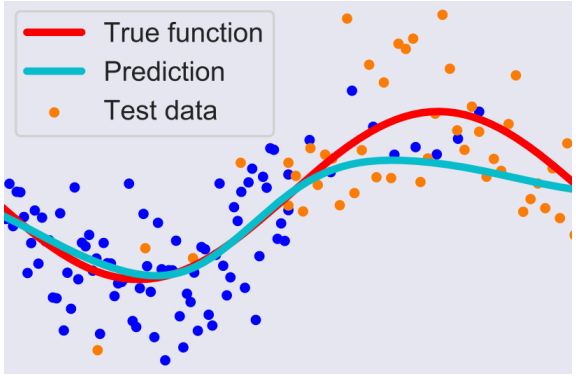
$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$



From empirical risk to worst-case risk minimization

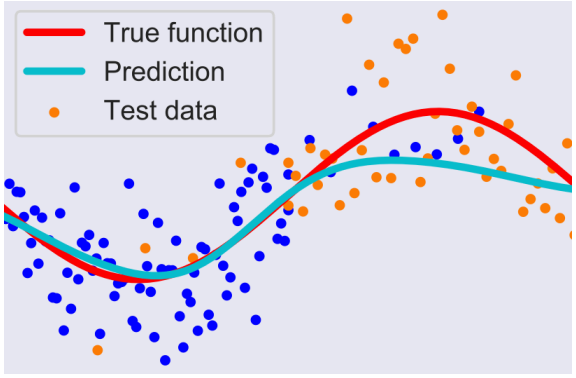


From empirical risk to worst-case risk minimization



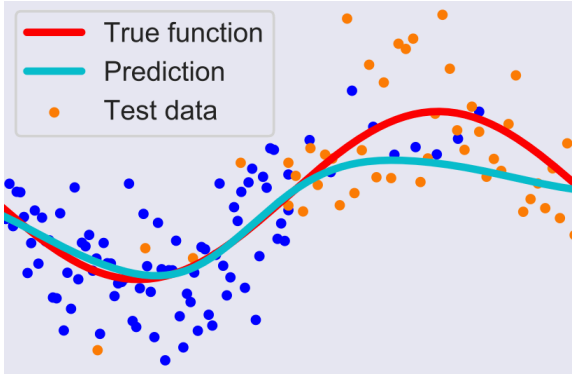
$$\min_{\theta} \int \ell(\theta, x)\, \mathrm{d}\widehat{\rho_N}(x)$$

From empirical risk to worst-case risk minimization



$$\min_{\theta} \int \ell(\theta, x)\, \mathrm{d}\widehat{\rho_N}(x) \quad \Rightarrow \quad \min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

From empirical risk to worst-case risk minimization

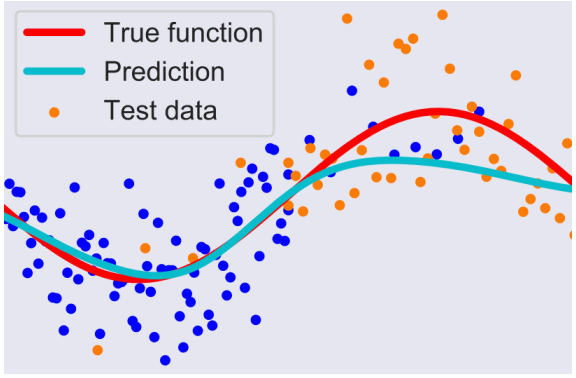


$$\min_{\theta} \int \ell(\theta, x)\, \mathrm{d}\widehat{\rho_N}(x) \quad \Longrightarrow \quad \min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

Idea: generate sample $y_i \sim \rho^*$, then

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta, y_i)$$

From empirical risk to worst-case risk minimization



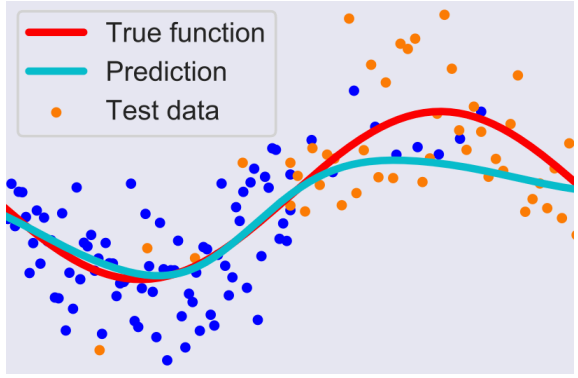
$$\min_\theta \int \ell(\theta, x)\, \mathrm{d}\widehat{\rho_N}(x) \quad \Longrightarrow \quad \min_\theta \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

Idea: generate sample $y_i \sim \rho^*$, then

How?

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \ell(\theta, y_i)$$

From empirical risk to worst-case risk minimization



JKO / proximal

$$\min_\theta \int \ell(\theta, x)\, \mathrm{d}\widehat{\rho_N}(x) \quad \Rightarrow \quad \min_\theta \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

Idea: generate sample $y_i \sim \rho^*$, then

How?

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \ell(\theta, y_i)$$

Generate a "worst-case" sample $y \sim \rho^*$ (The JKO$_\epsilon$-solution)

$$\rho^* \leftarrow \operatorname*{argmin}_{\rho \in \mathcal{P}} \int V \, d\rho \; + \; \tfrac{1}{2\tau} \, W_\epsilon^2(\rho, \rho_0)$$

Generate a "worst-case" sample $y \sim \rho^*$ (The JKO$_\epsilon$-solution)

$$\rho^* \leftarrow \operatorname*{argmin}_{\rho \in \mathcal{P}} \int V \, d\rho \; + \; \tfrac{1}{2\tau} \, W_\epsilon^2(\rho, \rho_0)$$

The sampling scheme:

1. sample $x \sim \rho_0$
2. sample $y \sim \rho_{Y|X=x}$

Generate a "worst-case" sample $y \sim \rho^*$ (The JKO$_\epsilon$-solution)

$$\rho^* \leftarrow \operatorname*{argmin}_{\rho \in \mathcal{P}} \int V \, d\rho \; + \; \tfrac{1}{2\tau} \, W_\epsilon^2(\rho, \rho_0)$$

The sampling scheme:

1. sample $x \sim \rho_0$
2. sample $y \sim \rho_{Y|X=x}$

The conditional distribution of $Y|X = x$ has a nice structure

Robustness via interacting/coupled gradient flows

Robustness via interacting/coupled gradient flows

View the optimization problem as a game

$$\min_{\theta} \max_{x} f(\theta, x)$$

descent-ascent interacting/coupled gradient flow

$$\dot{\theta} = -\nabla_{\theta} f(\theta, x)$$
$$\dot{x} = \nabla_{x} f(\theta, x)$$

[Lin et al, Conger et al, Xu et al., Dvurechensky & Z., . . .]

Robustness via interacting/coupled gradient flows

View the optimization problem as a game

$$\min_{\theta} \max_{x} f(\theta, x)$$

descent-ascent interacting/coupled gradient flow

$$\dot{\theta} = -\nabla_{\theta} f(\theta, x)$$
$$\dot{x} = \nabla_{x} f(\theta, x)$$

[Lin et al, Conger et al, Xu et al., Dvurechensky & Z., . . .]

❌ we will avoid doing this directly in the entropic OT geometry

A gradient-flow approach to ent-OT distributional robustness

General energy functional:

$$\rho_{Y|X=x} \in \operatorname*{argmin}_{\rho} \left\{ J(\rho) := \int V \,\mathrm{d}\rho + \frac{1}{2\tau} W_2^2(\rho, \delta_x) + \frac{\epsilon}{\tau} \int \rho \log \rho \right\}$$

A gradient-flow approach to ent-OT distributional robustness

General energy functional:

$$\rho_{Y|X=x} \in \operatorname*{argmin}_{\rho} \left\{ J(\rho) := \int V \,\mathrm{d}\rho + \frac{1}{2\tau} W_2^2(\rho, \delta_x) + \frac{\epsilon}{\tau} \int \rho \log \rho \right\}$$

$$J(\rho) = \mathsf{KL}(\rho \mid \tfrac{1}{Z} \exp(-\tfrac{\tau}{\epsilon} \widetilde{V})) \quad \text{for} \quad \widetilde{V}_{x,\tau}(y) := V(y) + \tfrac{1}{2\tau} c(y, x)$$

A gradient-flow approach to ent-OT distributional robustness

General energy functional:

$$\rho_{Y|X=x} \in \operatorname*{argmin}_{\rho} \left\{ J(\rho) := \int V \,\mathrm{d}\rho + \frac{1}{2\tau} W_2^2(\rho, \delta_x) + \frac{\epsilon}{\tau} \int \rho \log \rho \right\}$$

$$J(\rho) = \mathsf{KL}(\rho \mid \tfrac{1}{Z} \exp(-\tfrac{\tau}{\epsilon} \widetilde{V})) \quad \text{for} \quad \widetilde{V}_{x,\tau}(y) := V(y) + \tfrac{1}{2\tau} c(y, x)$$

consider the gradient system $(\mathcal{P}, J, \mathbb{K})$

$$\dot{\rho} = -\mathbb{K} DJ(\rho)$$

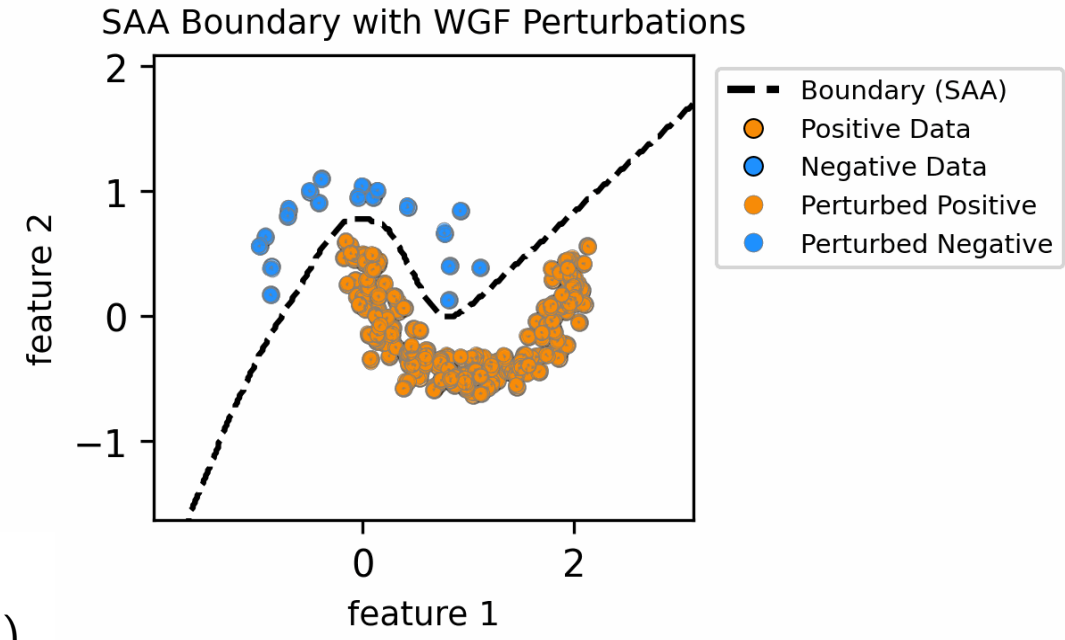
Xu-**Z**. working paper

Visualize samples of W or HK GF

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z) \, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

1. sample $x \sim \widehat{\rho}_N$
2. sample $y \sim \rho_{Y|X=x}$

Visualize samples of W or HK GF

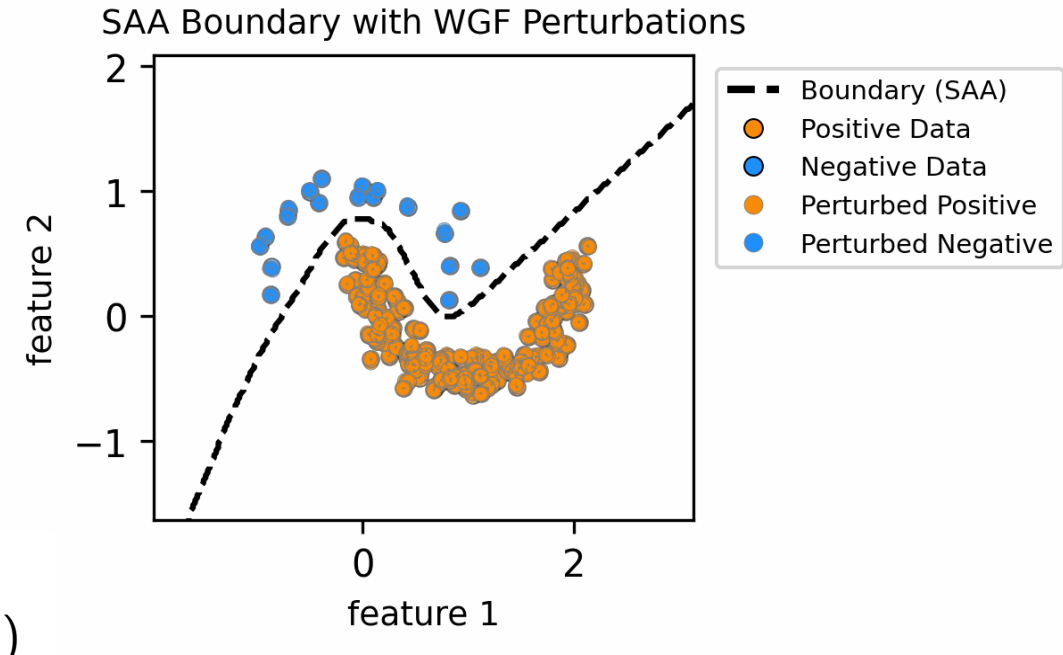


$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

1. sample $x \sim \widehat{\rho}_N$
2. sample $y \sim \rho_{Y|X=x}$

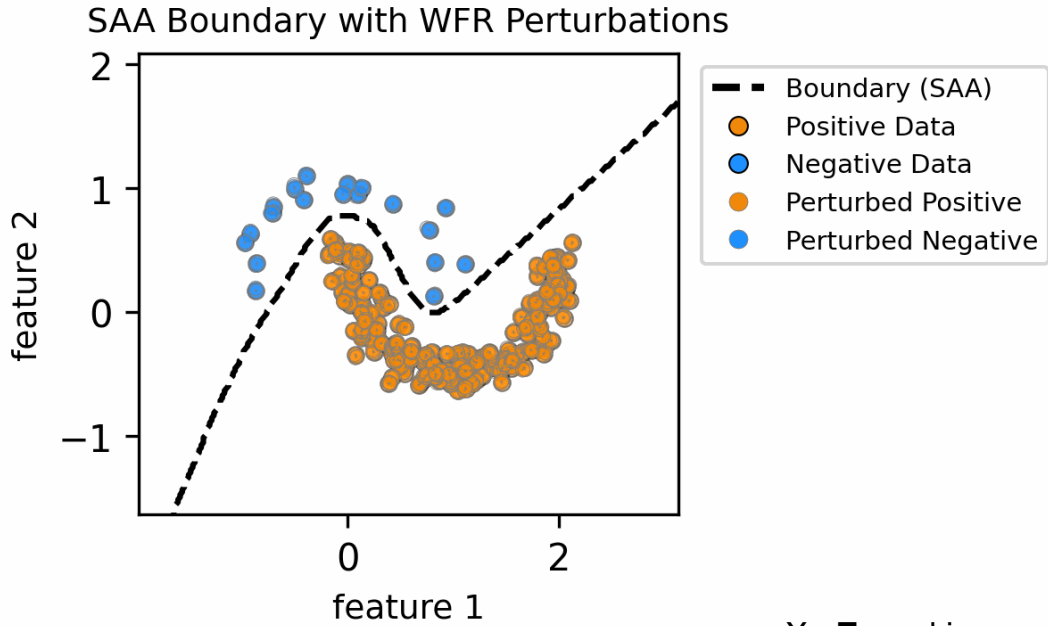
Xu-**Z**. working paper

Visualize samples of W or HK GF



$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

1. sample $x \sim \widehat{\rho}_N$
2. sample $y \sim \rho_{Y|X=x}$



Xu-**Z**. working paper

DRO via sampling

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

Transport from $X \sim \widehat{\rho_N}$ to $Y \sim \rho$

Algorithm. Iterate the following steps:

1. generate samples from desired distribution $y_i \sim \rho_{Y|X=x_i}$
   - sample $x_i \sim \widehat{\rho_N}$
   - sample $y_i \sim \rho_{Y|X=x_i} \propto \exp\left[\frac{\tau}{\epsilon}\ell(\theta^k, y) - \frac{1}{\epsilon}c(y, x_i)\right]$
2. gradient step on θ

$$\theta^{k+1} \leftarrow \theta^k - \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \ell(\theta^k, y_i)$$

DRO via sampling

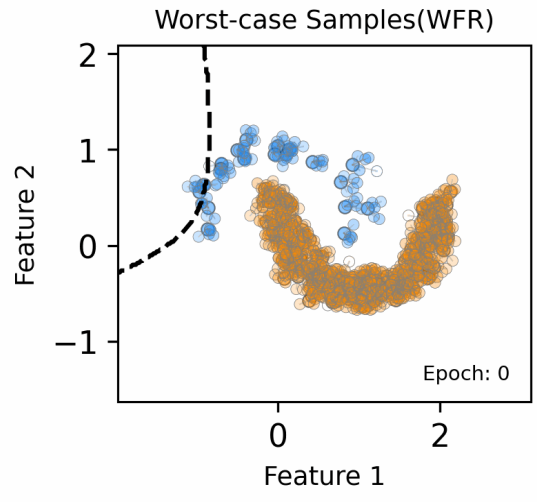
$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

Transport from $X \sim \widehat{\rho_N}$ to $Y \sim \rho$

Algorithm. Iterate the following steps:

1. generate samples from desired distribution $y_i \sim \rho_{Y|X}$:
   - sample $x_i \sim \widehat{\rho_N}$
   - sample $y_i \sim \rho_{Y|X=x_i} \propto \exp\left[\frac{\tau}{\epsilon}\ell(\theta^k, y) - \frac{1}{\epsilon}c(y, x_i)\right]$
2. gradient step on θ

$$\theta^{k+1} \leftarrow \theta^k - \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \ell(\theta^k, y_i)$$

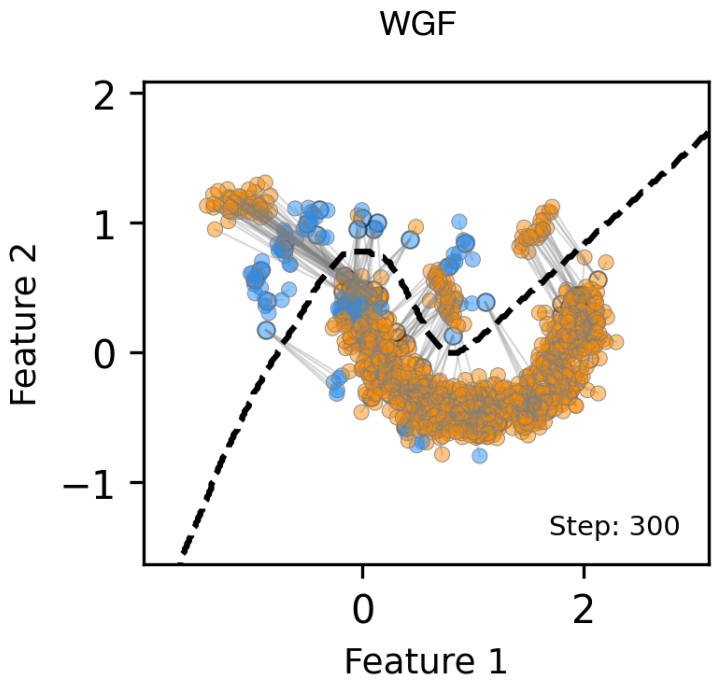


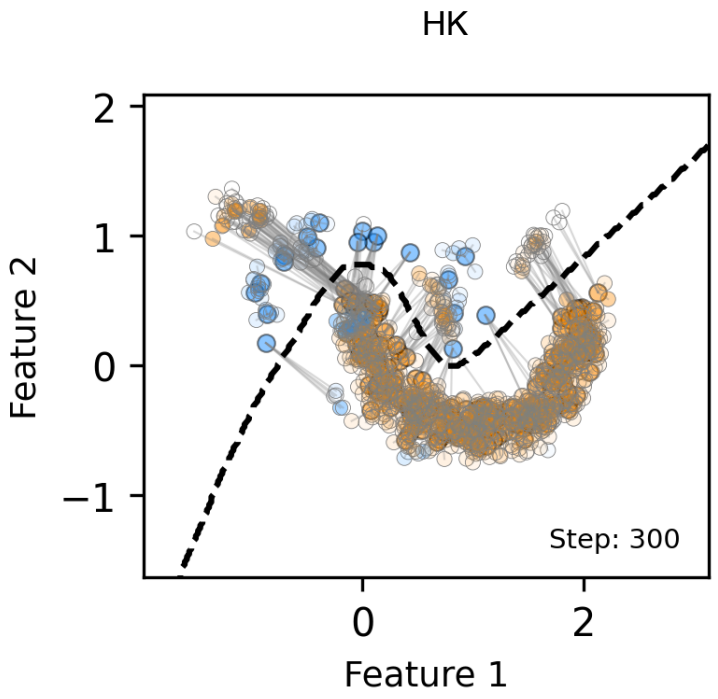


Xu-**Z**. working paper

Visualize the numerical results

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$

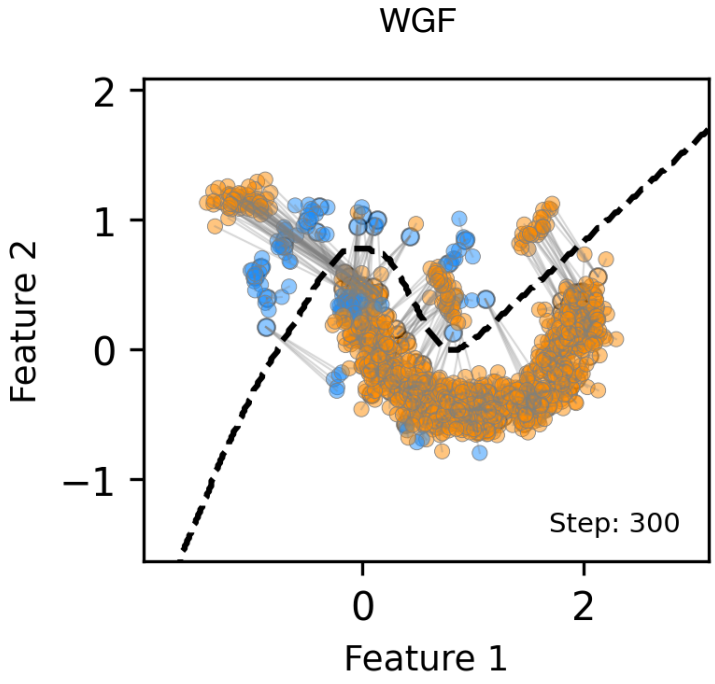


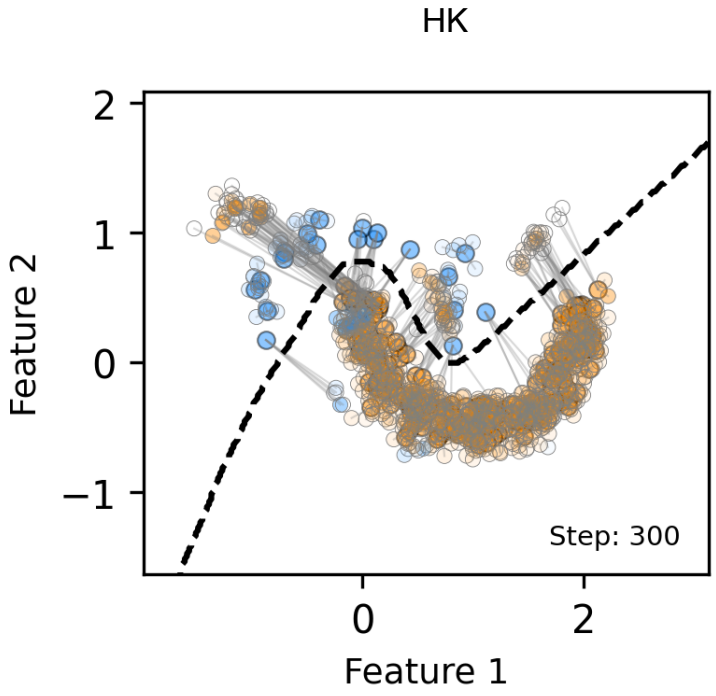


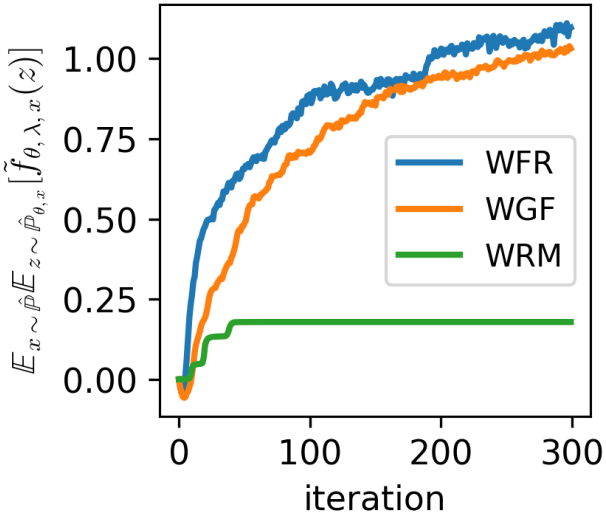
Xu-**Z**. working paper

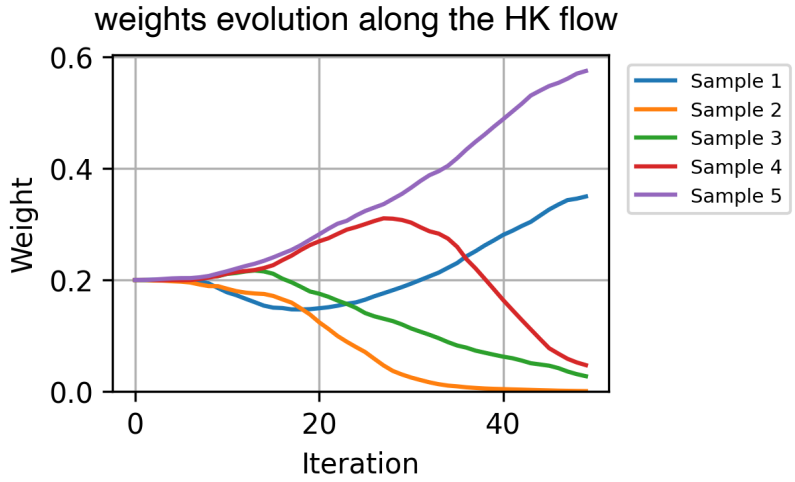
Visualize the numerical results

$$\min_{\theta} \max_{\rho \in \mathcal{P}} \int \ell(\theta, z)\, \mathrm{d}\rho(z) - \tfrac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$$









Xu-**Z**. working paper

Summary

Summary

$$\min_{\mu \in \mathcal{M}^+} F(\mu) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu} F(\mu|\pi) + \frac{1}{2\tau}\mathcal{R}(\mu, \mu_k) \qquad \dot{\mu} = -\mathbb{G}^{-1} DF(\mu)$$

Summary

$$\min_{\mu \in \mathcal{M}^+} F(\mu) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu} F(\mu|\pi) + \frac{1}{2\tau}\mathcal{R}(\mu, \mu_k) \qquad \dot{\mu} = -\mathbb{G}^{-1} DF(\mu)$$

PDE analysis ↔️ dynamics of optimization, sampling, inference

Summary

$$\min_{\mu \in \mathcal{M}^+} F(\mu) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu} F(\mu|\pi) + \frac{1}{2\tau}\mathcal{R}(\mu, \mu_k) \qquad \dot{\mu} = -\mathbb{G}^{-1} DF(\mu)$$

PDE analysis ↔️ dynamics of optimization, sampling, inference

Example applications:

variational inference (parametric)	sampling / generative model (nonparametric / interacting particle sys.)

Summary

$$\min_{\mu\in\mathcal{M}^+} F(\mu) \qquad \mu_{k+1} \leftarrow \operatorname*{argmin}_{\mu} F(\mu|\pi) + \frac{1}{2\tau}\mathcal{R}(\mu,\mu_k) \qquad \dot{\mu} = -\mathbb{G}^{-1}DF(\mu)$$

PDE analysis ↔️ dynamics of optimization, sampling, inference

Example applications:

variational inference (parametric)	sampling / generative model (nonparametric / interacting particle sys.)

This talk is based on:

Mielke-**Z**. Hellinger-Kantorovich Gradient Flows: Global Exponential Decay of Entropy Functionals. arXiv preprint

Liero-Mielke-Tse-**Z**. Evolution of Gaussians in the HK-Boltzmann gradient flow. CPAA special issue on “Transport Equations in Optimization, Sampling, and Control”

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

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

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

Xu-**Z**. Gradient Flow Sampler-based Distributionally Robust Optimization. working paper

For more information, see the website: https://jj-zhu.github.io/

I’m hiring! Get in touch

Appendix

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:

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

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

JKO splitting scheme

The PDE

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

can be simulated using the JKO scheme

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

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

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

Interaction-Force Transport flow implementation

We keep track of the particles by their location x_i^ℓ and their weight α_i^ℓ.

Location update. For every particle location $x_i^\ell \sim \mu^\ell$,

$$x_i^{\ell+1} = x_i^\ell - \tau \cdot \nabla \partial_\mu \, \mathrm{MMD}^2 \left[\mu^\ell\right] \left(x_i^\ell\right), \quad i = 1, \ldots, n,$$

where

$$\begin{aligned} \nabla \partial_\mu \, \mathrm{MMD}^2 \left[\rho\right](x) &= \int \nabla_x k(z, x) \, \mathrm{d} \left(\rho(z) - \pi(z)\right) \\ &\approx \sum_{j=1}^{n_\rho} \nabla_x k(z_j, x) - \sum_{j=1}^{n_\pi} \nabla_x k(y_j, x) \end{aligned}$$

where the samples are drawn according to $z_j \sim \rho$ and $y_j \sim \pi$.

Weight update.

$$\alpha_i^{\ell+1} \leftarrow \alpha_i^\ell \cdot \exp\left(-\eta \cdot \partial_\mu \, \mathrm{MMD}^2 \left[\mu^\ell\right] \left(x_i^{\ell+1}\right)\right)$$