Gradient Flows in Wasserstein Spaces

Variational Inference and Sampling

Student. Luca Raffo

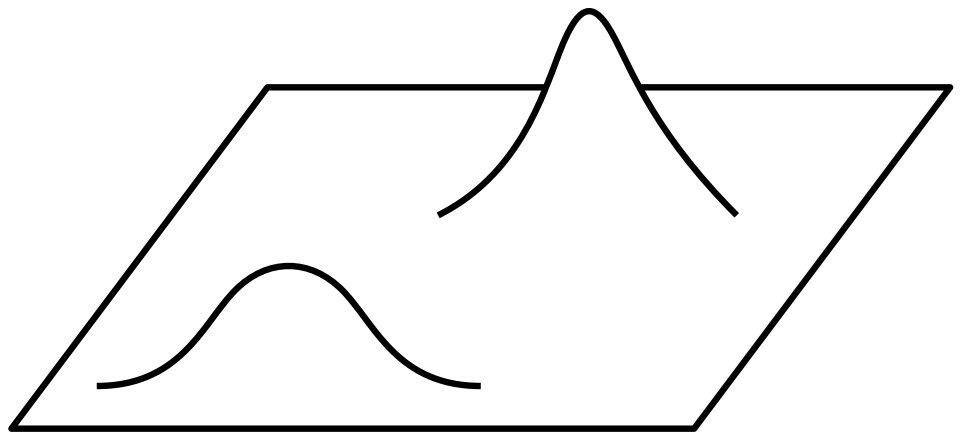
Prof. Victor M. Panaretos

Dr. Leonardo V. Santoro

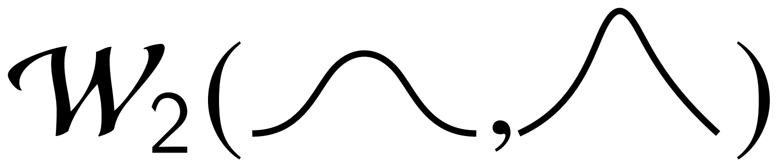
introduction and motivation.

consider the Euclidean space $\mathbb{R}^d$.

look at the family of probability measures $\mathcal{P}_2$ on $\mathbb{R}^d$.



define a distance $\mathcal{W}_2$ between probability measures.



introduction and motivation.

the metric space $(\mathcal{P}_2, \mathcal{W}_2)$ has nice geometric properties.

$\downarrow$

we can study this **geometry.**

$\downarrow$

get profound understanding of **evolutions of measures.**

$\downarrow$

get theoretical guarantees for **variational inference** and **sampling**.

OTTO

plan

1. preliminaries. metric geometry, ~~Monge and Kantorovich problems.~~

2. Wasserstein spaces. pseudo-Riemannian geometry, evolution of measures, ~~first variations~~, Wasserstein gradient flows.

3. variational inference. KL divergence, geodesic convexity, hints on the JKO scheme.

4. particles variational inference. SVGD.

5. sampling. Langevin diffusion as a gradient flow.

6. EXTRA.

preliminaries. metric geometry.

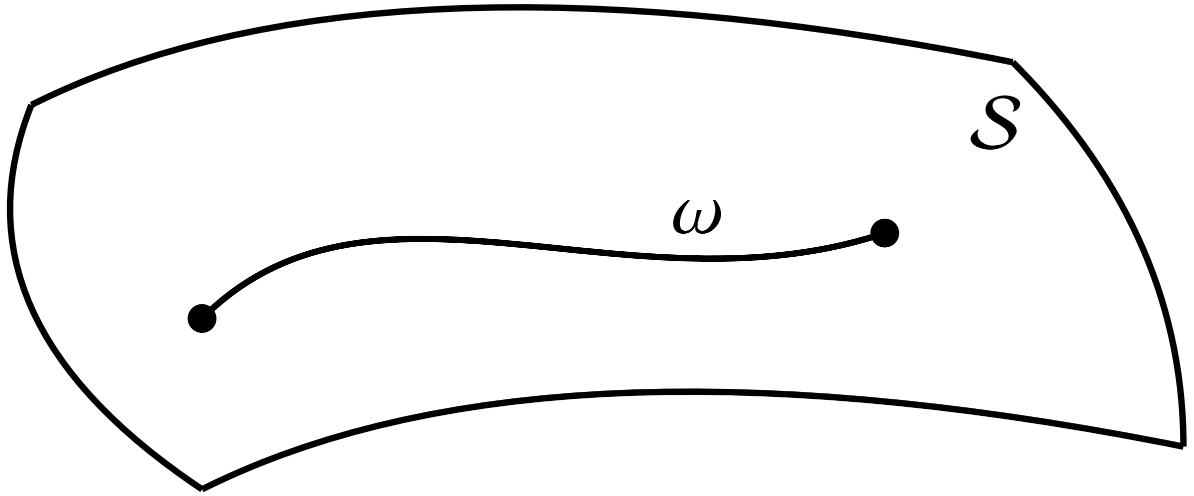
abstraction of key ideas from differential geometry.

fix any $\boxed{\text{metric space}}$ $(\mathcal{S}, d)$.

positive definiteness
symmetry
triangle inequality

we can define *paths* $\omega : I \subseteq \mathbb{R} \to \mathcal{S}$.

and *lenghts* $L(\omega)$.



preliminaries. metric geometry.

fix $x_0, x_1 \in \mathcal{S}$.

a path $\omega : [0,1] \to \mathcal{S}$, with $\omega(0) = x_0$ and $\omega(1) = x_1$ is a *geodesic* if $d(x_0, x_1) = L(\omega)$.

any ω can be reparametrized to be a *constant speed geodesic.*

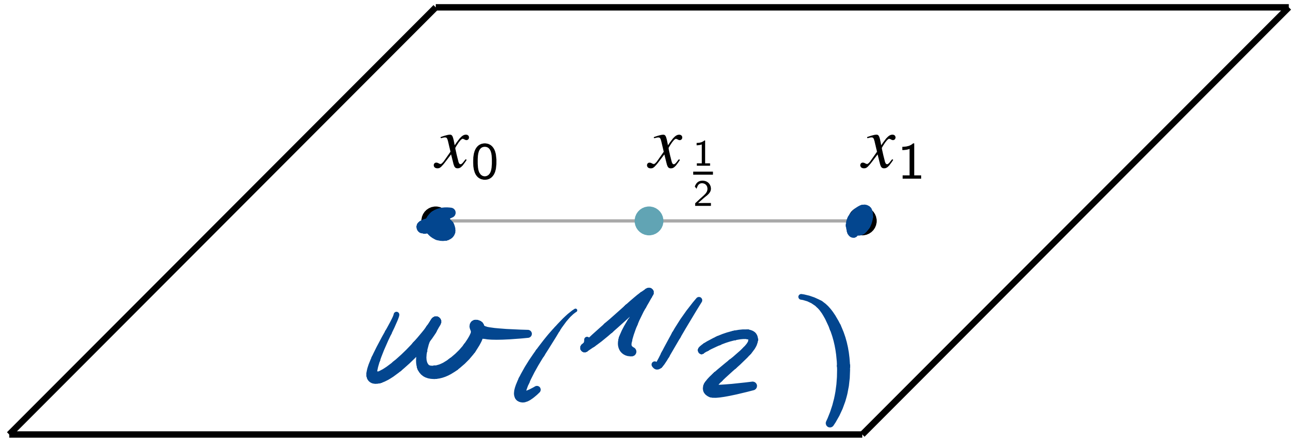
$(\mathcal{S}, d)$ is said to be a *geodesic space* if for any given x_0, x_1 we can exhibit a geodesic.

equivalently, a constant speed geodesic.

we can define **midpoints.**

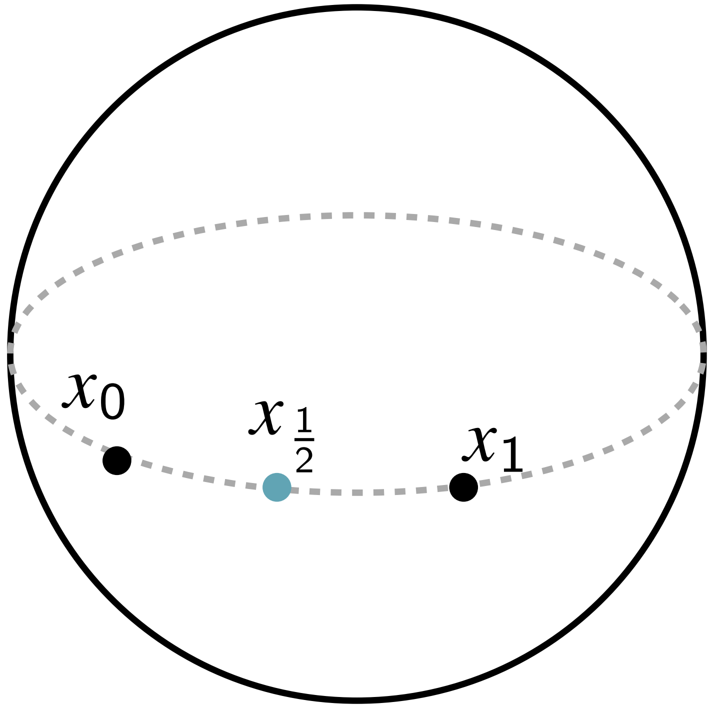
preliminaries. metric geometry.

$(\mathcal{S}, d) = (\mathbb{R}^2, d_{\|\cdot\|_2})$ is a geodesic space.



$(\mathcal{S}, d) = (\mathbb{S}^2, \boxed{d_r})$ is a geodesic space.

$d_r(x, y) = \arccos(x \cdot y)$



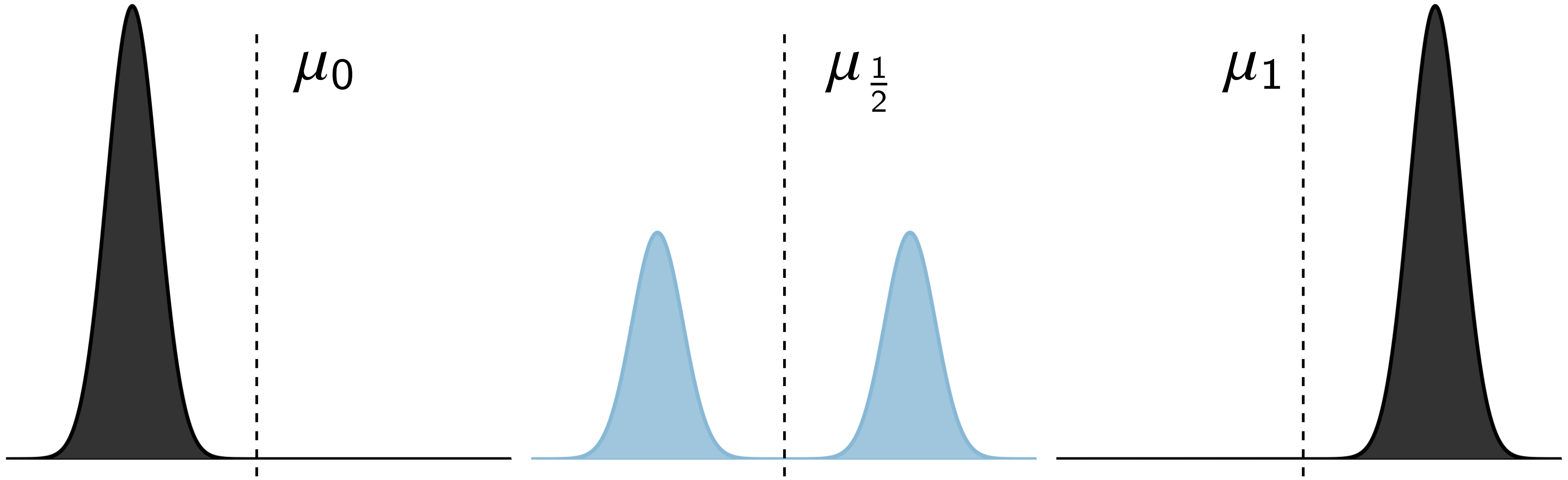
preliminaries. metric geometry.

let us move to spaces of measures.

proposition. the space $(\mathcal{P}_2^{ac}(\lambda), L_2(\lambda))$ is a geodesic space.

proof idea. the constant speed geodesic between μ_0 and μ_1 is

$$\mu_t := h(t)\, d\lambda = [(1-t)g_0 + tg_1]\, d\lambda.$$



~~**preliminaries.** Monge and Kantorovich.~~

given $\mu_0 \in \mathcal{P}_2$ and $T : \mathbb{R}^d \to \mathbb{R}^d$, we define the *push forward measure* as

$$T\#\mu_0 := \mu_0(T^{-1}(A)), \text{ for any } A \subseteq \mathbb{R}^d.$$

we define the *canonical projections* π_X and π_Y such that $\pi_X(x, y) = x$ and $\pi_Y(x, y) = y$.

given $\mu_0, \mu_1 \in \mathcal{P}_2$, we define the set of *couplings* as

$$\Gamma(\mu_0, \mu_1) = \{\gamma \in \mathcal{P}_2 \times \mathcal{P}_2 : \pi_X\#\gamma = \mu_0, \pi_Y\#\gamma = \mu_1\}.$$

preliminaries. ~~Monge and Kantorovich.~~

given $\mu_0, \mu_1 \in \mathcal{P}_2$ and $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define

$$(MP) = \inf_T \left\{ \int_{\mathbb{R}^d} c(T(x), x)\, d\mu_0(x) : T\#\mu_0 = \mu_1 \right\}$$

and its relaxation

$$(KP) = \inf_{\gamma \in \mathcal{P} \times \mathcal{P}} \left\{ \int_{\mathbb{R}^d} c(x, y)\, d\gamma(x, y) : \gamma \in \Gamma(\mu_0, \mu_1) \right\}.$$

any transport map T between μ_0 and μ_1 induces a coupling: $\gamma_T := (id, T)\#\mu_0$.

fact. if $\mu_0 \in \mathcal{P}_2^{ac}(\lambda)$, there exists ϕ convex such that $T = \nabla\phi$ is the unique optimizer in (MP).

(and $(id, T)\#\mu_0$ is the unique optimizer in (KP)).

Wasserstein spaces. geometry.

given $\mu_0, \mu_1 \in \mathcal{P}_2$, we define their *Wasserstein distance* as

$$\mathcal{W}_2(\mu_0, \mu_1) := \min_\gamma \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, d\gamma(x, y) \,|\, \gamma \in \Gamma(\mu_0, \mu_1) \right)^{\frac{1}{2}} .$$

fact. it is actually a distance. not trivial.

if the optimal coupling is induced by T, $\bar{\gamma} = (id, T)\#\mu_0$

$$\mathcal{W}_2(\mu_0, \mu_1) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|T_{\mu_0 \to \mu_1}(x) - x\|_2^2 \, d\mu_0(x) \right)^{\frac{1}{2}} .$$

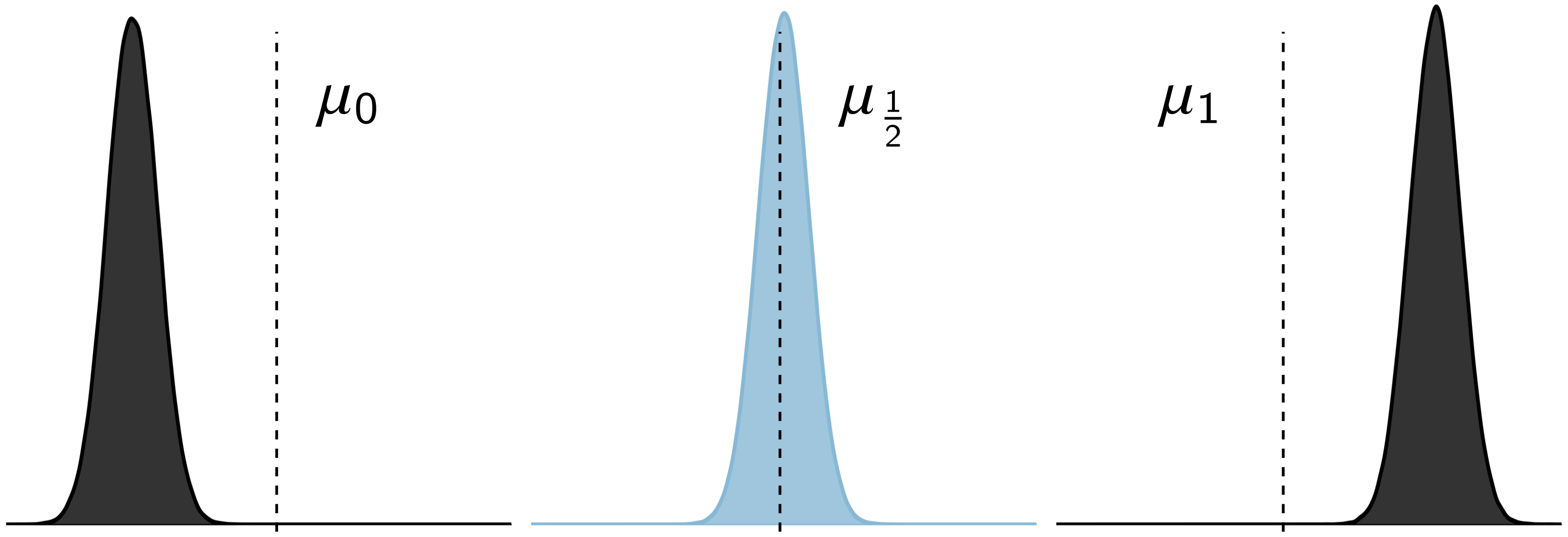
Wasserstein spaces. geometry.

fact. $(\mathcal{P}_2, \mathcal{W}_2)$ is a geodesic space.

δ_{x_0} δ_{x_1}

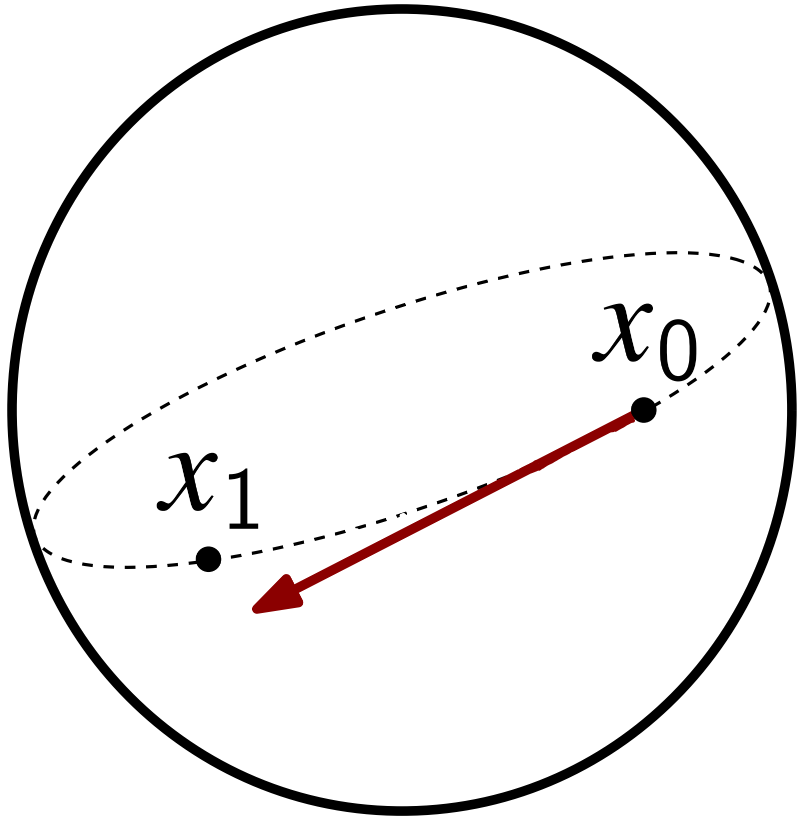
proof idea. the geodesic is $\mu_t := T_t \# \mu_0$.

$$T_t(x) = (1-t)x + tT_{\mu_0 \to \mu_1}(x)$$



Wasserstein spaces. geometry.

we want to lift the idea of tangent spaces.



fix μ_0. as μ_1 varies, consider the geodesic map

$$T_t(x) = (1-t)x + tT_{\mu_0 \to \mu_1}(x)$$

$$\mathcal{T}_{\mu_0}\mathcal{P}_2 := \overline{\{\eta(T_{\mu_0\to\mu_1} - id) : \mu_1 \in \mathcal{P}_2, \eta > 0\}}^{L_2(\mu_0)}$$

AGS'08

$$= \overline{\{\nabla\phi |\ \phi : \mathbb{R}^d \to \mathbb{R} \text{ is a test function }\}}^{L_2(\mu_0)}$$

we get for free a inner product on $\mathcal{T}_{\mu_0}\mathcal{P}_2$: $\langle f, g\rangle_{\mu_0} = \int_{\mathbb{R}^d} f(x)g(x)\, d\mu_0(x)$, $f, g \in L_2(\mu_0)$.

Wasserstein spaces. evolution of measures.

if $X_0 \sim \mu_0$, and I evolve X_0 via $\dot{X}_t = v_t(X_t)$ for a vector field v_t, then $Law(X_t)$ satisfies the continuity equation $\partial_t \mu_t + \langle \nabla, (\mu_t v_t) \rangle_2 = 0$.

if we have densities this is $\partial_t g_t + \langle \nabla, (g_t v_t) \rangle_2 = 0$.

ODE ⟶ PDE

given a regular flow μ_t, we can find the most *economical* vector field v_t that induces it i.e. that minimizes $\|v_t\|_{L_2(\mu_t)}$ for all t, moreover $v_t \in \mathcal{T}_{\mu_0}\mathcal{P}_2$ and can be written as

$$\boxed{v_t = \lim_{\delta \to 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - id}{\delta}}$$

Wasserstein spaces. first variations.

a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable if $f(x+h) - f(x) = \boxed{h[\delta f(x)]} + o(h)$

a functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$ has bounded first variation if $\mathcal{F}(\mu + \epsilon\chi) - \mathcal{F}(\mu) = \epsilon \boxed{[\delta\mathcal{F}(\mu)]}(\chi) + o(\epsilon)$

bounded linear functional

by Kantorovich-Rubinstein duality, $\mathcal{F}(\mu + \epsilon\chi) - \mathcal{F}(\mu) = \epsilon \int_{\mathbb{R}^d} \boxed{[\delta\mathcal{F}(\mu)]}\, d\chi + o(\epsilon)$.

continuous bounded function

Wasserstein spaces. flows.

$\mathcal{F}$

take μ_t a regular flow. we can expand $\boxed{\mu_t = \mu_0 + t\partial_t \mu_t + o(t).}$

$\lim_{\delta \to 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - id}{\delta}$

$$\boxed{\lim_{t\to 0} \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu_0)}{t}} = \int_{\mathbb{R}^d} [\delta \mathcal{F}(\mu_0)]\, d(\partial_t \mu_t)$$

$$= \int_{\mathbb{R}^d} \langle (\nabla[\delta\mathcal{F}(\mu_0)])(x), \boxed{v_t}(x)\rangle_2 \, d\mu_t(x) = \boxed{\langle (\boxed{\nabla[\delta\mathcal{F}(\mu_0)]}), v_t\rangle_{L_2(\mu_t)}}.$$

$\in \mathcal{T}_{\mu_0}\mathcal{P}_2$

we call $\nabla_{W_2}\mathcal{F}(\mu_0) = \nabla[\delta\mathcal{F}(\mu_0)]$ the *Wasserstein gradient.*

we call $\partial_t \mu_t - \langle \nabla, (\nabla_{W_2}\mathcal{F}(\mu_t)\mu_t)\rangle_2 = 0$ the *Wasserstein gradient flow.*

Wasserstein spaces. flows.

for the *potential energy* $\mathcal{V}(\mu) := \int_{\mathbb{R}^d} V\, d\mu \longrightarrow \partial_t \mathcal{V}(\mu_t) = \int_{\mathbb{R}^d} V\, d(\partial_t \mu_t).$

$$\nabla_{W_2} \mathcal{V}(\mu) = \nabla V.$$

density of μ

for the *entropy functional* $Ent(\mu) := \int_{\mathbb{R}^d} \boxed{g} \log(g)\, d\lambda \longrightarrow \partial_t Ent(\mu_t) = \int_{\mathbb{R}^d} \partial_t g_t (\log(g_t) + 1)\, d\lambda.$

$$\boxed{\nabla_{W_2} Ent(\mu) = \nabla \log g.}$$

variational inference. KL divergence.

$\mathcal{F}(\cdot) = D_{KL}(\cdot \| \pi)$

fix $\pi \in \mathcal{P}_2$, with density $\boxed{\frac{1}{Z} f = \frac{1}{Z} e^{-V}}$

I cannot evaluate Z, so I want to find $\mu^* = \arg\min_{\mu \in Q} \mathcal{D}_{KL}(\mu \| \pi) \geq 0$
with Q convex and computationally feasible.

I approach this problem via Wasserstein gradient flows.

I need convexity.

variational inference. convexity.

about convexity guarantees...

a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-t)x_0 + tx_1) \le (1-t)f(x_0) + tf(x_1)$.
$(-\frac{\alpha}{2}t(1-t)\|x_0 - x_1\|_2^2)$

a functional $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R}$ is geodesically convex if for a geodesic μ_t,
$\mathcal{F}(\mu_t) \le (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1)$. $(-\frac{\alpha}{2}t(1-t)\|\mu_0 - \mu_1\|_{\mathcal{W}_2}^2)$

fact. if $\mathcal{F}$ is (strongly) geodesically convex and $Q \subseteq \mathcal{P}_2^{ac}(\lambda)$ is convex, then the Wasserstein gradient flow of $\mathcal{F}$ started in Q lies in Q and converges exponentially fast towards

$$\mu^* = \arg\min_{\mu \in Q} \mathcal{F}(\mu).$$

PL

variational inference. convexity.

$$\mathcal{D}_{KL}(\mu\|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{g}{f}\right) g \, d\lambda = Ent(\mu) + \mathcal{V}(\mu) - \log Z.$$

geodesically convex.

(strongly) geodesically convex.
(provided V is strongly convex).

$\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot\|\pi)|_\mu = \nabla V + \nabla \log g$ does not require us to compute Z.

variational inference. JKO.

is there a *scheme* that produces approximately a Wasserstein gradient flow?

for functions $f : \mathbb{R}^d \to \mathbb{R}$, $\frac{dx}{dt} = -\nabla F(x)$ leads to $x_{k+1} = x_k - \tau \nabla F(x_{k+1})$.

$$x_{k+1}^{\tau} = \text{arg min}_{x \in \mathbb{R}^d} \left\{ \frac{1}{2\tau} \|x - x_k^{\tau}\|^2 + F(x) \right\}$$

accordingly, $\mu_{k+1}^{\tau} = \text{arg min}_{\mu \in \mathcal{P}_2} \left\{ \frac{1}{2\tau} \mathcal{W}_2^2(\mu, \mu_k^{\tau}) + \mathcal{F}(\mu) \right\}$
leads to the Wasserstein gradient flow in the limit as $\tau \to 0$.

(18 SINKORN

if $\mathcal{F}(\cdot) = \mathcal{D}_{KL}(\cdot \| \pi)$, we can use f instead of $\frac{1}{Z} f$.

variational inference. particles v.i.

LIU ET AL. '19

if I have samples $X_0, \ldots, X_N$ from μ, I can evolve them via $\dot{X}_i^t = -\nabla_{\mathcal{W}_2} \mathcal{D}_{KL}(\cdot\|\pi)|_{\mu_t}(X_i^t)$.

if $\mathcal{Q} = \mathcal{Q}_N$ the family of discrete measures (N particles), I could in principle evolve particles and track them to get a perfect description of WGF.

problem. $\nabla_{\mathcal{W}_2} \mathcal{D}_{KL}(\cdot\|\pi)|_{\mu}$ is not defined if $\mu \in \mathcal{Q}_N$.

can we do something similar? is there some ϕ_t^* such that $\dot{X}_i^t = \phi_t^*(X_i^t)$ is close to $\dot{X}_i^t = -\nabla_{\mathcal{W}_2} \mathcal{D}_{KL}(\cdot\|\pi)|_{\mu_t}(X_i^t)$?

variational inference. particles v.i.

fix a RKHS $\mathcal{H}$.

fix $\epsilon > 0$. define $T_\epsilon(x) := x + \epsilon\phi(x), \phi \in \mathcal{H}^d$.

let $\mu_\epsilon := (T_\epsilon)\#\mu$.

then, $\frac{d}{d\epsilon}\mathcal{D}_{KL}(\mu_\epsilon\|\pi)|_{\epsilon=0} = -\mathbb{E}_{X\sim\mu}[\langle\nabla \log f(X), \phi(X)\rangle_2 + \langle\nabla, \phi(X)\rangle_2]$.

formal abuse. we think μ_ϵ as abs. cont. in LHS.
RHS is well defined even for μ discrete.

variational inference. particles v.i.

we want $\phi^* = \arg\min_{\phi\in\mathcal{H}^d, \|\phi\|_{\mathcal{H}^d}\leq 1} -\mathbb{E}_{X\sim\mu}[\langle\nabla \log f(X), \phi(X)\rangle_2 + \langle\nabla, \phi(X)\rangle_2]$.

$$= \arg\min_{\phi\in\mathcal{H}^d} \left\{\frac{1}{N}\sum_{i=1}^N [\langle V(x_i), \phi(x_i)\rangle_2 - \langle\nabla, \phi(x_i)\rangle_2] + \lambda\|\phi\|^2_{\mathcal{H}^d}\right\}.$$

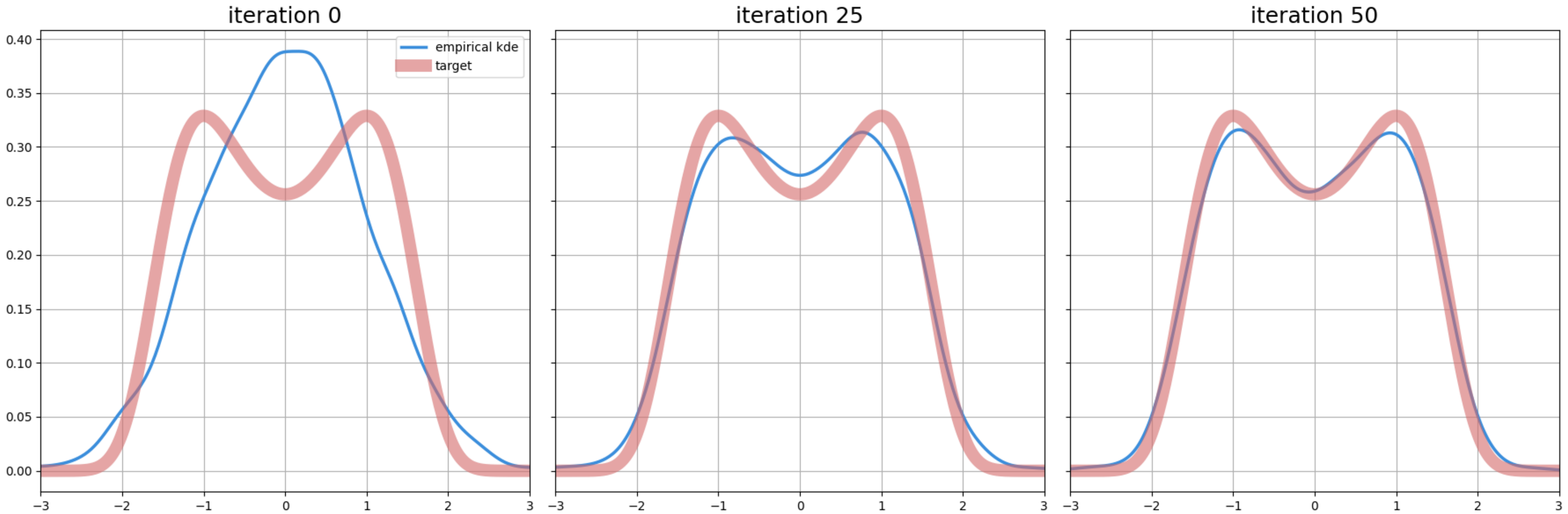
from RKHS theory, $\phi^*(x) = \frac{1}{N}\sum_{i=1}^N [K(x_j, x)(-\nabla V(x_i)) + \nabla_{x_j} K(x_j, x)]$.

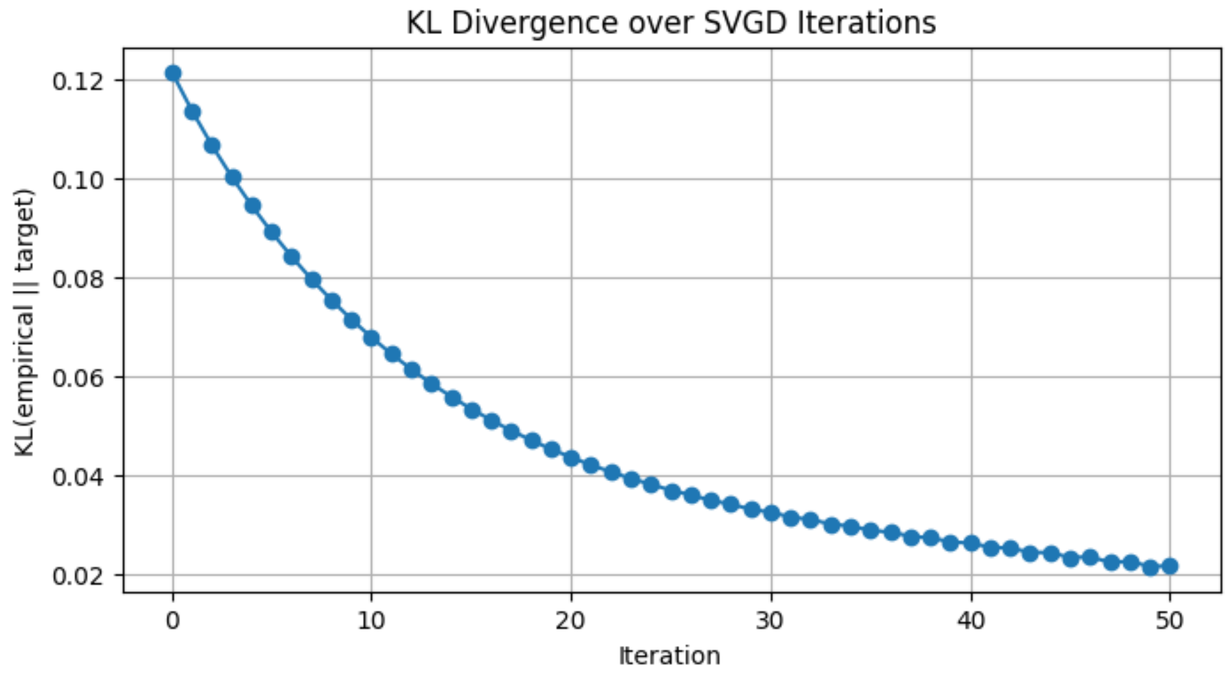
to get something similar to WGF, I evolve particles via $\dot{X}_i^t = \phi_t^*(X_i^t)$.

variational inference. particles v.i.

we start with N samples from a gaussian.

we want to move towards π whose density is $f(x) \propto e^{-V(x)}$, $V(x) = \frac{x^4}{4} - \frac{x^2}{2}$.

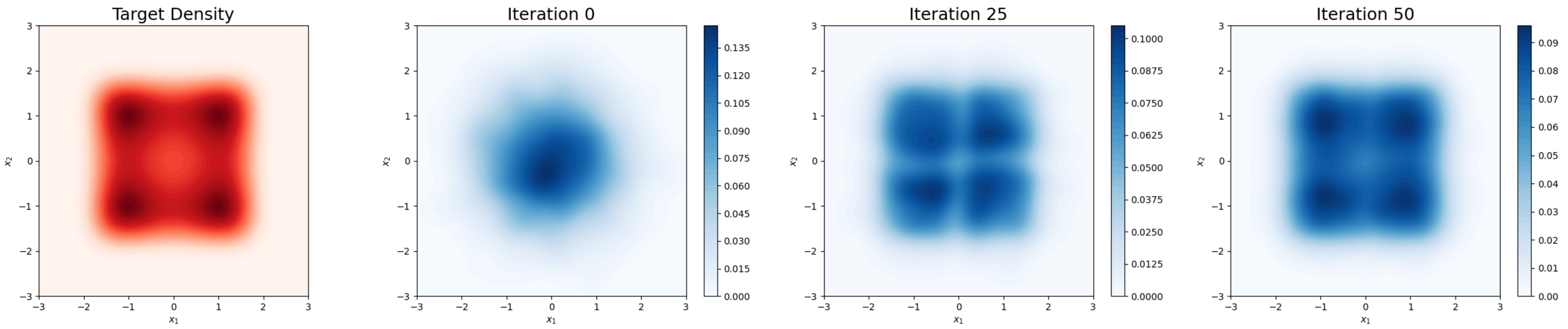


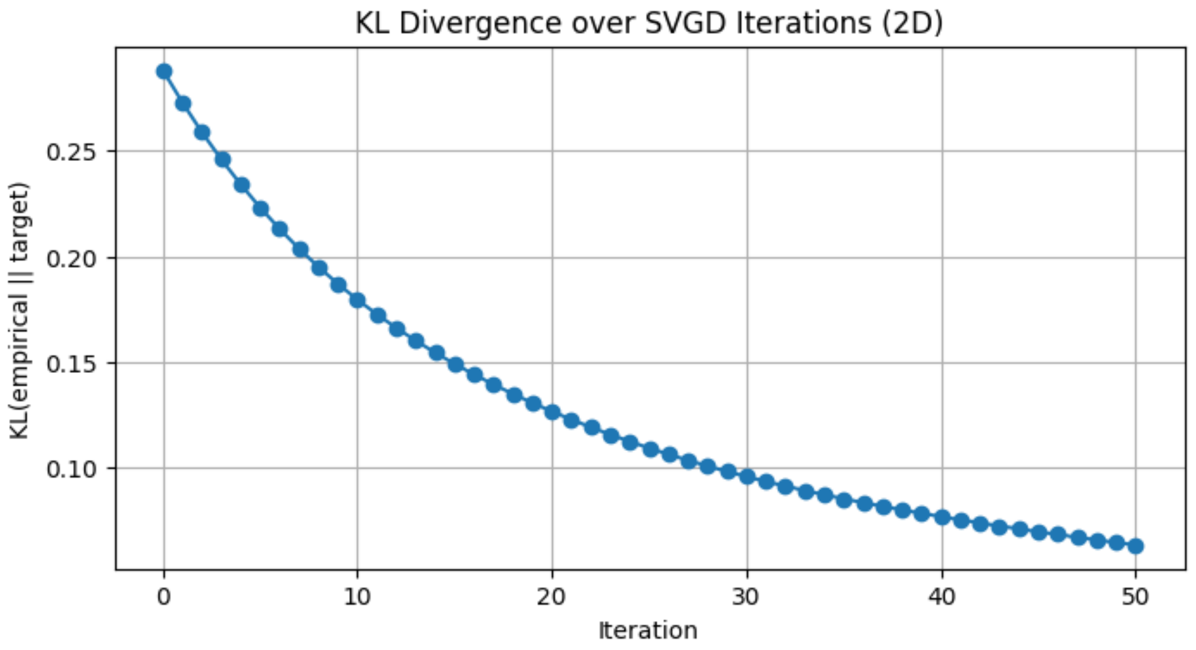


we use a RBF kernel, the median trick for the bandwith,
adagrad for the evolution and KDE for visualizations.

variational inference. particles v.i.

we want to move towards π whose density is $f(x) \propto e^{-V(x)}$, $V(x) = \frac{\|x\|^4}{4} - \frac{\|x\|^2}{2}$.





sampling. Fokker Planck.

let us look at our usual Wasserstein gradient $-\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot\|\pi)|_{\mu_t} = -\nabla \log g_t - \nabla V$.

its flow is $\partial_t g_t = \langle \nabla, g_t(\nabla \log g_t + \nabla V)\rangle_2$.

$$= \langle \nabla, g_t \nabla \log g_t + g_t \nabla V\rangle_2.$$

$$= \Delta g_t + \langle \nabla, (g_t \nabla V)\rangle_2. \longrightarrow \text{Fokker Planck.}$$

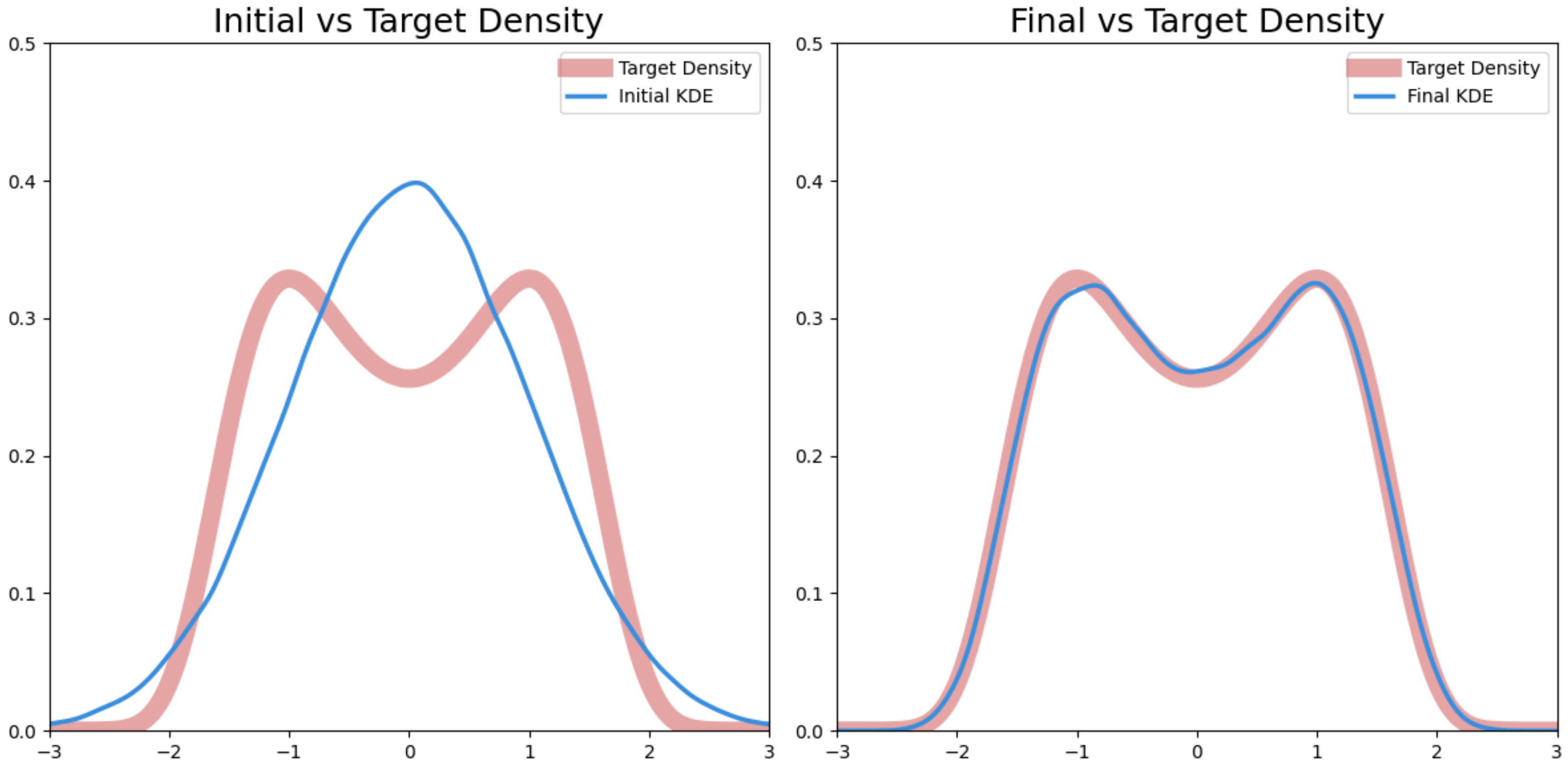
the density of the Lagenvin diffusion satisfies the Fokker Planck. $dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$.

we have a different way to evolve a discrete measure to π!

sampling. Fokker Planck.

we start with N samples from a gaussian.

we want to move towards π whose density is $f(x) \propto e^{-V(x)}$, $V(x) = \frac{x^4}{4} - \frac{x^2}{2}$.

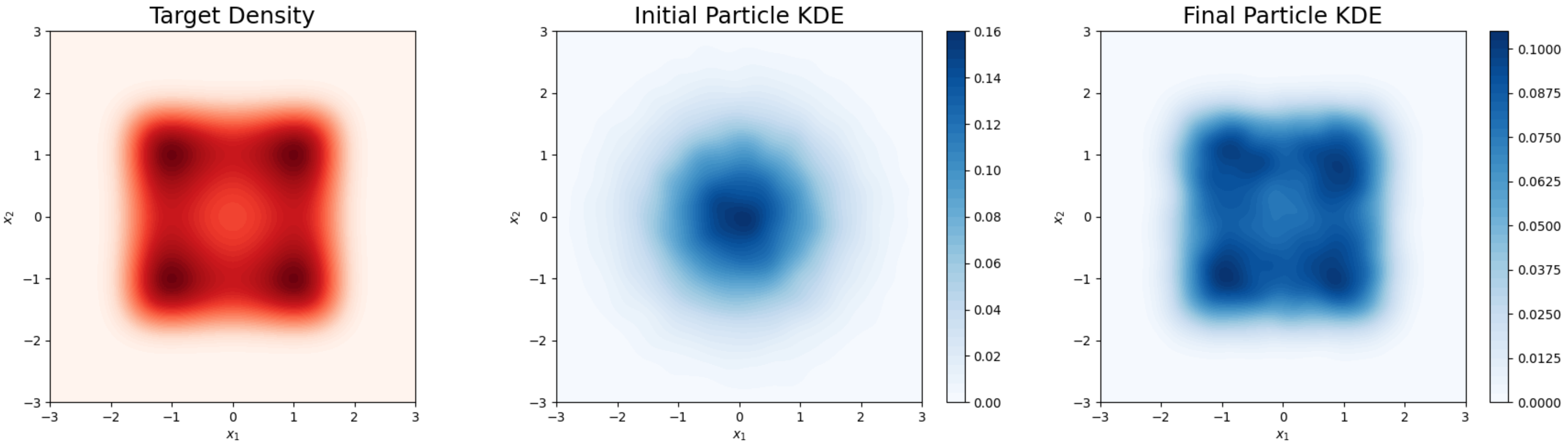


we use Euler-Maruyama scheme, and KDE for visualizations.

sampling. Fokker Planck.

we start with N samples from a gaussian.

we want to move towards π whose density is $f(x) \propto e^{-V(x)}$, $V(x) = \frac{\|x\|^4}{4} - \frac{\|x\|^2}{2}$.



we use Euler-Maruyama scheme, and KDE for visualizations.

extra. score matching.

suppose we have access to $X_1, \ldots, X_N \overset{iid}{\sim} \mu$, with $d\mu = g\, d\lambda$ unknown.

we want to find our best guess $\pi_\theta \approx \mu$, where $d\pi_\theta = f_\theta\, d\lambda$, and $f \propto e^{-V_\theta}$.

idea. if $\pi_\theta \approx \mu$, then the WGF of $\mathcal{D}_{KL}(\cdot \| \pi_\theta)|_\mu$ is almost stationary.

it is reasonable to ask that the most economic underlying flow on particles is small.

I search for
$$\begin{aligned}\theta^* &= \arg\min_\theta \mathbb{E}_\mu[\| -\nabla_{\mathcal{W}_2} \mathcal{D}_{KL}(\cdot\|\pi_\theta)|_\mu\|^2] \\ &= \arg\min_\theta \mathbb{E}_\mu[\| -V_\theta - \nabla \log g\|^2] \\ &= \arg\min_\theta \mathbb{E}_\mu[\|\nabla \log f_\theta - \nabla \log g\|^2]\end{aligned}$$

STARTING POINT OF SCORE MATCHING