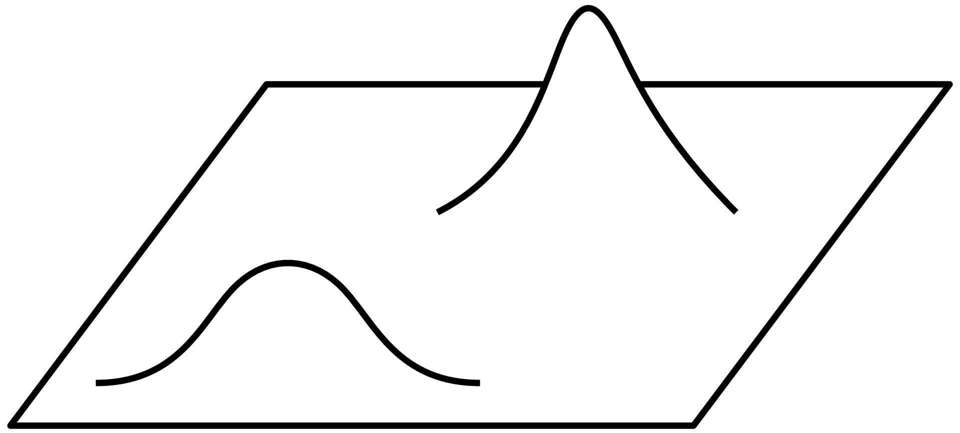
From Wasserstein Spaces to Score Matching

Luca Raffo

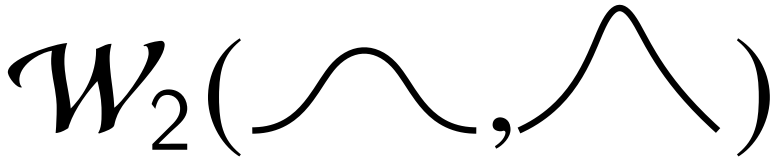
Wasserstein spaces

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

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

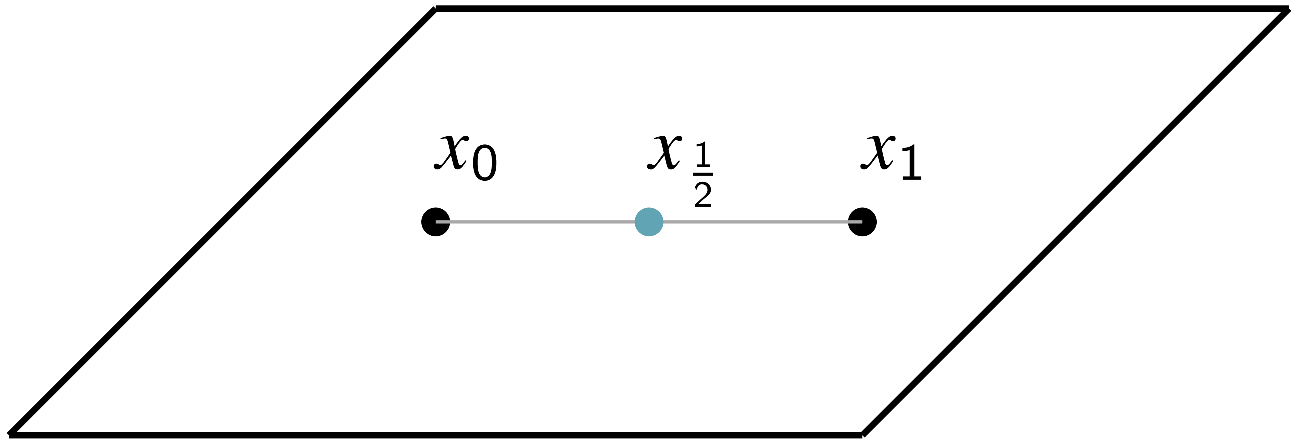


define a distance $\mathcal{W}_2$ between probability measures.

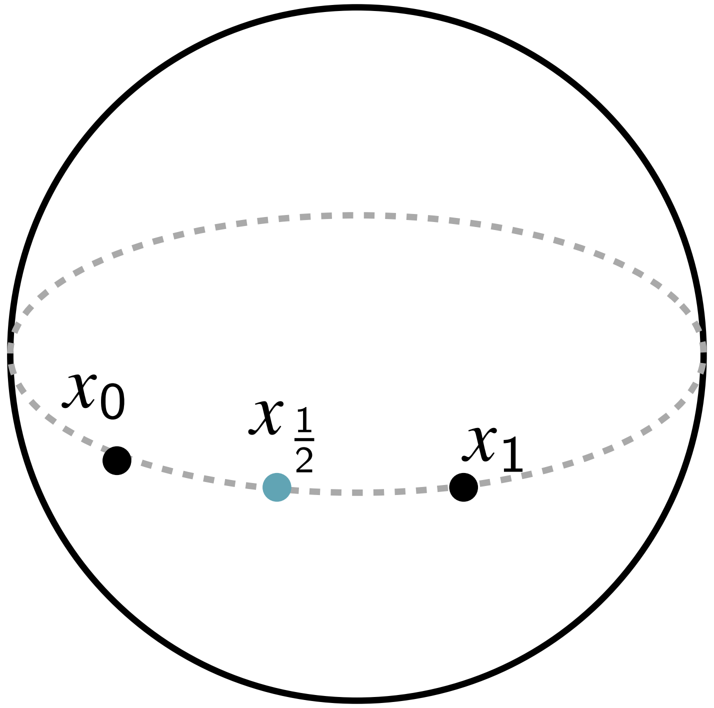


Wasserstein spaces

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



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



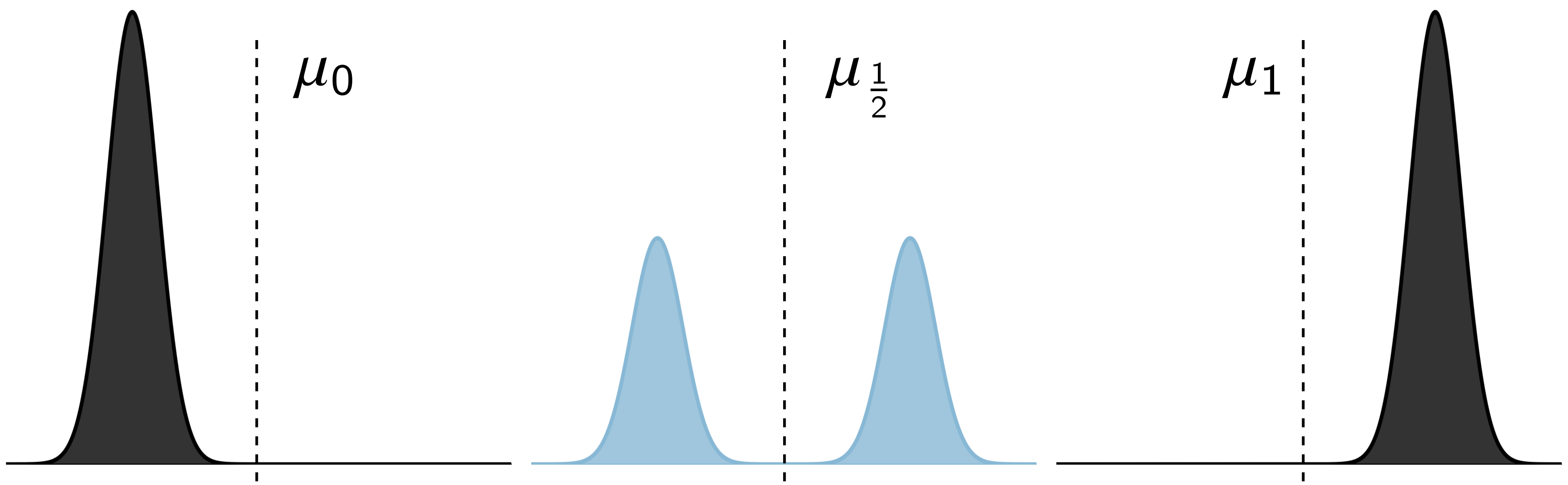
Wasserstein spaces

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

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



Wasserstein spaces

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

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

fact. it is actually a distance. not trivial.

if the optimal coupling is induced by a map T, i.e. $\gamma = (Id, T)_{\#}\mu_0$, then

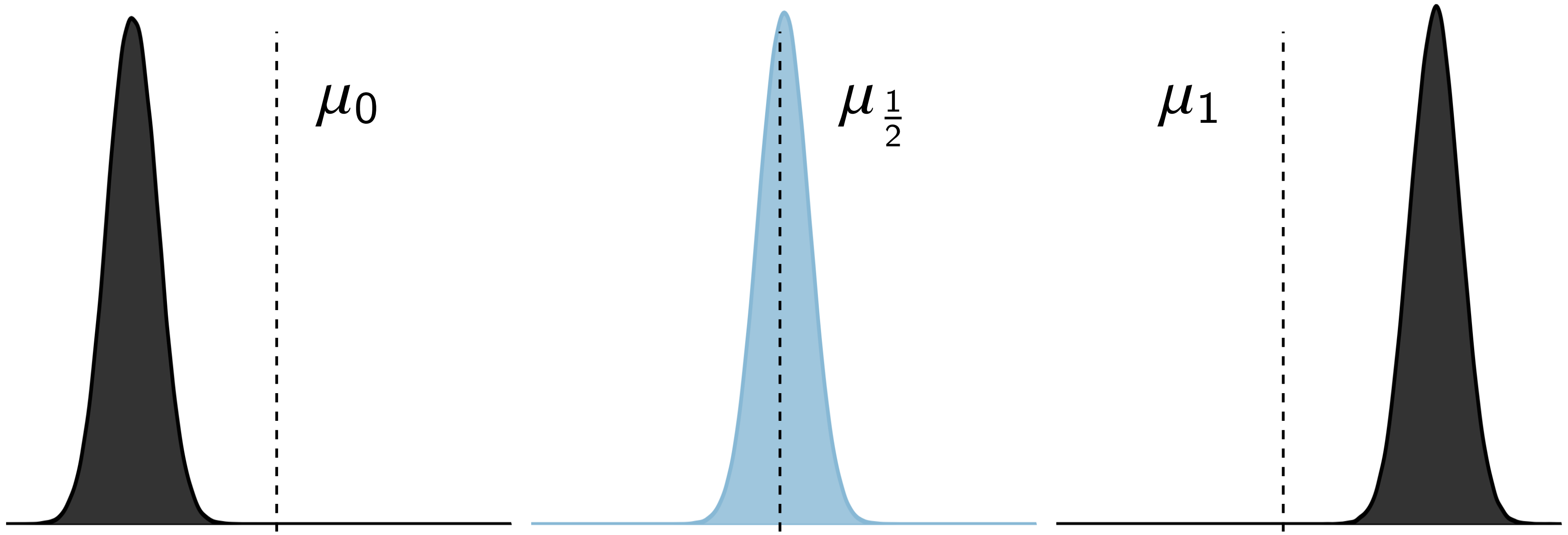
$$\mathcal{W}(\mu_0, \mu_1) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|T_{\mu_0 \to \mu_1}(x) - x\|_2^2 \, d\mu_0(x) \right)^{\frac{1}{2}} .$$

Wasserstein spaces

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

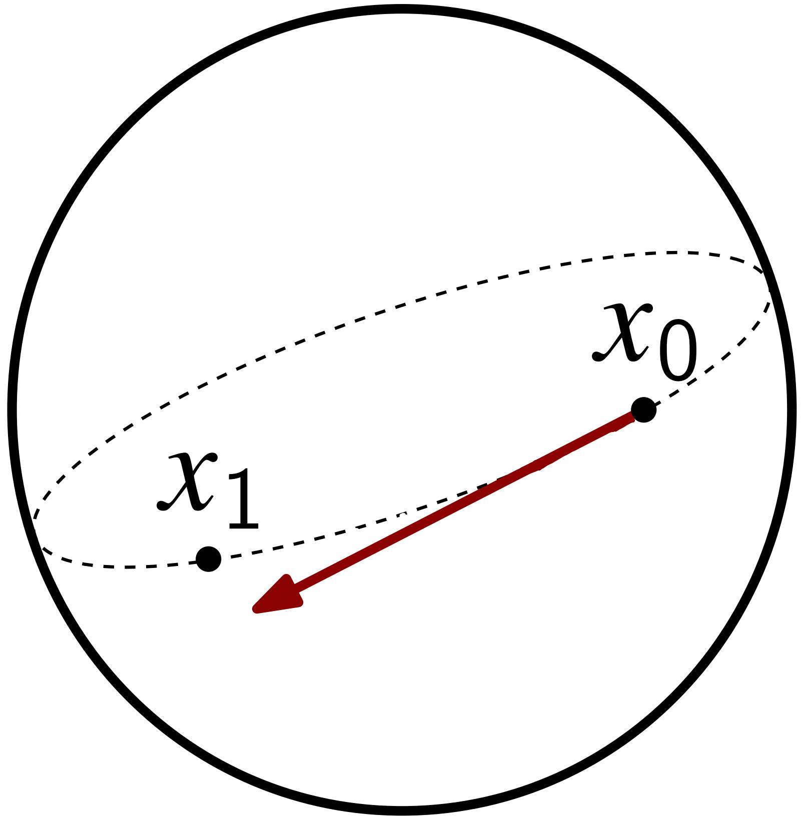
proof idea. the geodesic is $\mu_t := T_t \# \mu_0$.

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



Wasserstein spaces

we want to lift the idea of tangent spaces.



[Ambrosio, Gigli, Savaré]

$$\mathcal{T}_{\mu_0}\mathcal{P}_2 = \overline{\{\nabla\phi |\ \phi : \mathbb{R}^d \to \mathbb{R} \text{ is regular enough }\}}^{L_2(\mu_0)}$$

we get for free a inner product on $\mathcal{T}_{\mu_0}\mathcal{P}_2$

$$\langle \nabla f, \nabla g\rangle_{\mu_0} = \int_{\mathbb{R}^d} \nabla f(x) \nabla g(x)\, d\mu_0(x),\ \nabla f, \nabla g \in L_2(\mu_0).$$

Wasserstein spaces

if $X_0 \sim \mu_0$ with density g_0, and I evolve X_0 via $\dot{X}_t = v_t(X_t)$ for a vector field v_t, then the density g_t of μ_t satisfies the continuity equation $\partial_t g_t + \nabla \cdot (g_t v_t) = 0$.

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

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

Wasserstein spaces

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

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

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

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

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

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

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

we call $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_0) = \nabla[\delta\mathcal{F}(\mu_0)]$ the *Wasserstein gradient.*

we call $\partial_t g_t - \nabla \cdot (\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) g_t) = 0$ the *Wasserstein gradient flow.*

Wasserstein spaces

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

$$\nabla_{\mathcal{W}_2} \mathcal{V}(\mu) = \nabla V.$$

density of μ

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

$$\nabla_{\mathcal{W}_2} Ent(\mu) = \nabla \log g.$$

Wasserstein spaces

about convexity guarantees...

a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $f((1-t)x_0 + tx_1) \leq (1-t)f(x_0) + tf(x_1)$.

a functional $\mathcal{F} : \mathcal{P}_2 \to \mathbb{R}$ is geodesically convex if for a geodesic μ_t,

$$\mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1).$$

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

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

Wasserstein spaces

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

geodesically convex.

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

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

Wasserstein spaces

Theorem. if μ_t follows the Wasserstein gradient flow induced by $\mathcal{F}(\cdot)$,

$$\text{then } \frac{d}{dt}\mathcal{F}(\mu_t) = -\langle \nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t), \nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\rangle_{L_2(\mu_t)}$$

$$= -\int_{\mathbb{R}^d} ||\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)(x)||^2 \, d\mu_t(x).$$

Towards inference

In inference problems I have an empirical measure μ_g,

which we assume to have density g, and we have samples from it.

We fix a family $(\pi_\theta)_{\theta\in\Theta}$, with density $\frac{1}{Z}e^{-V_\theta}$,

and we look for the θ that best approximates μ_g given the samples.

Idea. I fix $\mu_0 = \mu_g$. Let μ_t be the Wasserstein gradient flow of $\mathcal{D}_{KL}(\cdot||\pi_\theta)$.

If $\mu_g \approx \pi_\theta$, I expect $|\frac{d}{dt}\mathcal{D}_{KL}(\mu_t||\pi_\theta)|_{t=0}| \approx 0$.

from the previous theorem, this is equivalent to $\int_{\mathbb{R}^d} ||\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\mu_0||\pi_\theta)||^2\, d\mu_0 \approx 0$.

where $\mu_0 = \mu_g$, $\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\mu_g\|\pi)|_\mu = \nabla V + \nabla \log g$.

Towards inference

$$\theta^* = \arg\min_\theta \int_{\mathbb{R}^d} ||\nabla_{\mathcal{W}_2} \mathcal{D}_{KL}(\mu_0||\pi_\theta)||^2 \, d\mu_0.$$

$$= \arg\min_\theta \int_{\mathbb{R}^d} ||\nabla V + \nabla \log g||^2 \, d\mu_g$$

$$= \arg\min_\theta \int_{\mathbb{R}^d} \tfrac{1}{2}||\nabla V_\theta||^2 - \Delta V_\theta \, d\mu_g$$

integration by parts

$$= \mathbb{E}_g[\tfrac{1}{2}||\nabla V_\theta||^2 - \Delta V_\theta]$$

$$\approx \mathbb{E}_{\mathcal{D}}[\tfrac{1}{2}||\nabla V_\theta||^2 - \Delta V_\theta]$$

this is exactly

score matching

Fokker Planck

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

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

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

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

So score matching can be summarized as

$$\theta^* = \arg\min_\theta |\tfrac{d}{dt}\mathcal{D}_{KL}(\mu_t||\pi_\theta)|_{t=0}|,$$

where μ_t follows the Fokker-Planck with potential V_θ and is started at $\mu_0 = \mu_g$

Discrete Wasserstein Spaces

In 2011, Maas, extended the notion of Wasserstein distance to discrete spaces.

We have a discrete space $\mathcal{X}$

a measure π_θ on $\mathcal{X}$

a transition kernel $K_\theta(\cdot,\cdot)$ in equilibrium with π_θ

$\mathcal{P}$ the family of densities on π_θ

He defines $\mathcal{W}(\rho_0||\rho_1), \rho_0, \rho_1 \in \mathcal{P}$ using the Benamou-Brenier characterization of $\mathcal{W}_2$.

He proves that the heat flow $\partial_t \rho_t = (K - I)\rho_t$ is the Wasserstein-Maas gradient flow of the entropy

$$Ent(\rho) = \sum_{x\in\mathcal{X}} \pi_\theta(x)\rho(x)\log\rho(x)$$

$$= \mathcal{D}_{KL}(\mu||\pi_\theta),\ d\mu = \rho d\pi_\theta$$

Discrete Wasserstein Spaces

Here the role of convexity is played by the Ricci curvature of K_θ...

more obscure, no theorems (to my knowledge)

If we try the same approach of the continuous setting, i.e. we aim at $\theta^* = \arg\min_\theta |\mathcal{D}_{KL}(\mu_t||\pi_\theta)|_{t=0}|$

where μ_t follows the Wasserstein gradient flow of the KL (a.k.a. the heat flow), started at μ_g

we end up with $\theta^* = \arg\min_\theta \sum_{x,y\in\mathcal{X}}[\frac{\log\mu(x)}{\log\mu(y)} - \frac{\log\mu(y)}{\log\mu(x)} + \frac{H_\theta(y)}{H_\theta(x)} - \frac{H_\theta(x)}{H_\theta(y)}]K_\theta(x,y)\mu(x)$

It doesn't seems to be usable in practice

Minimum Probability Flow does something similar: $\theta^* = \arg\min_\theta |\mathcal{D}_{KL}(\mu_0||\mu_t)|_{t=0}|$

where μ_t again follows the heat flow towards π_θ

Directions

If v_θ is strongly convex, do we learn faster? By how much

If $K_\theta(\cdot,\cdot)$ has a big Ricci curvature, do we learn faster? By how much?

If instead of $\mathcal{D}_{KL}$ we choose another functional, and we minimize
the Wasserstein gradient of the functional, when μ_t is following the Wasserstein gradient flow
do we get another inference method?

Can we design an inference method based on the
Wasserstein-Maas gradient of the KL divergence?