The Wasserstein gradient flow of the Sinkhorn divergence between Gaussian distributions

Mathis Hardion

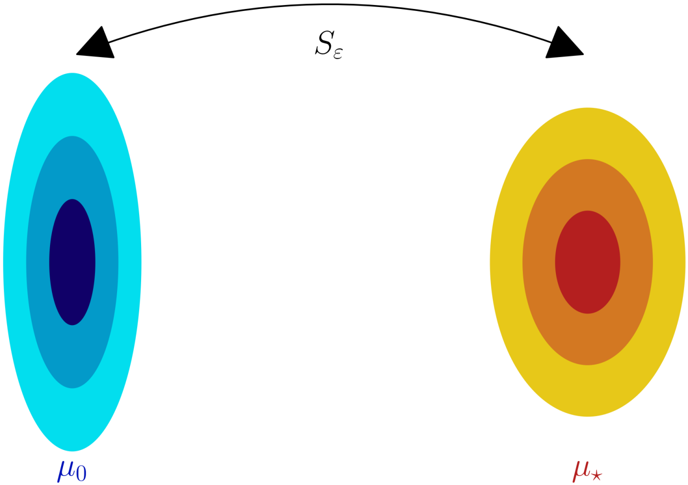
February 25, 2026

Université Gustave Eiffel

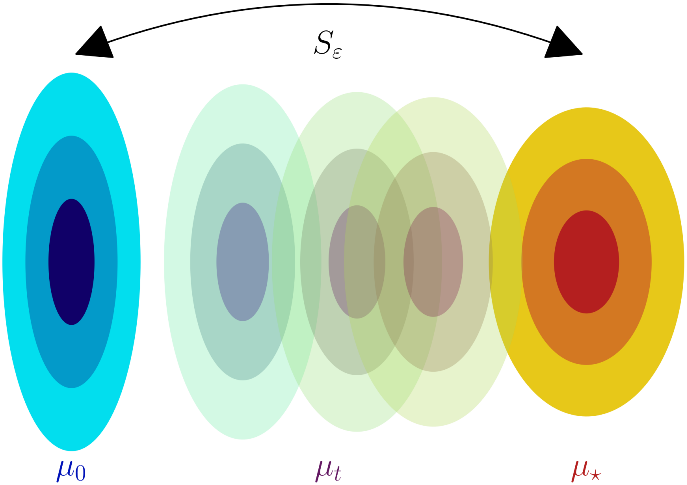
Gaspard - Monge · Laboratoire d'Informatique

Introduction

Introduction

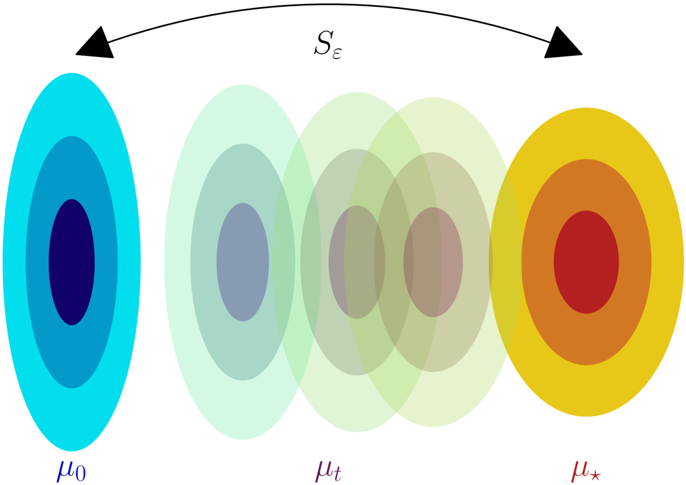


Introduction



$(\mu_t)_t$ described by a PDE.

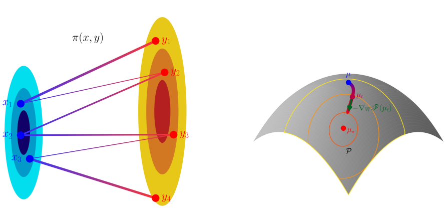
Introduction



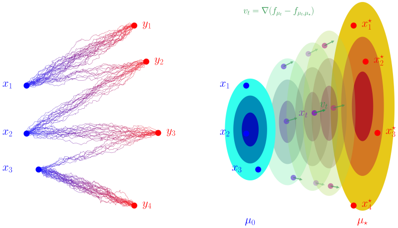
$(\mu_t)_t$ described by a PDE.

Goals: Well-posedness of that PDE, convergence criterion.

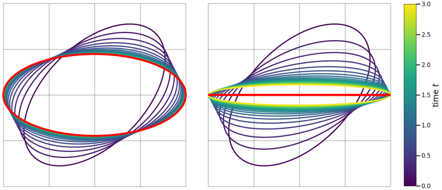
Plan



1. Optimal transport and gradient flows

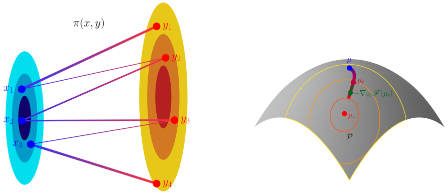


2. The Sinkhorn divergence and its flow

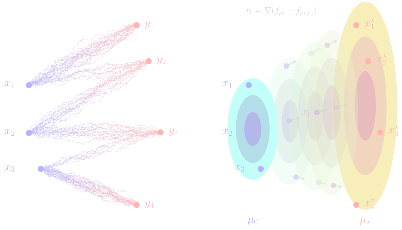


3. Main results

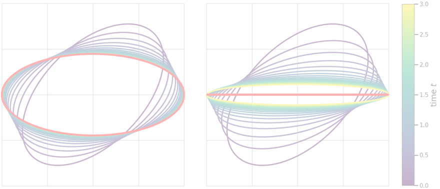
Plan



1. Optimal transport and gradient flows

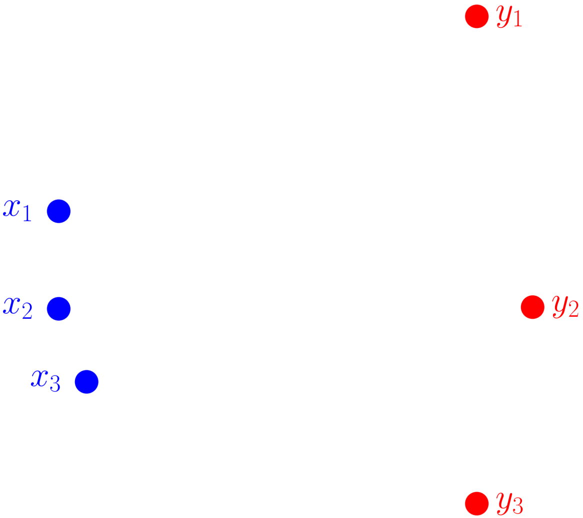


2. The Sinkhorn divergence and its flow

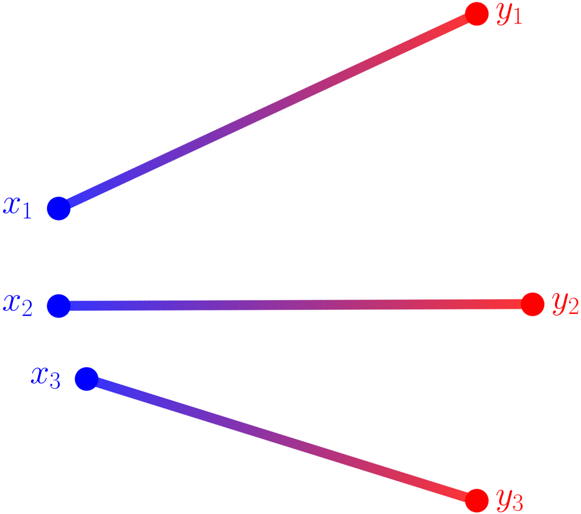


3. Main results

The Monge assignment problem

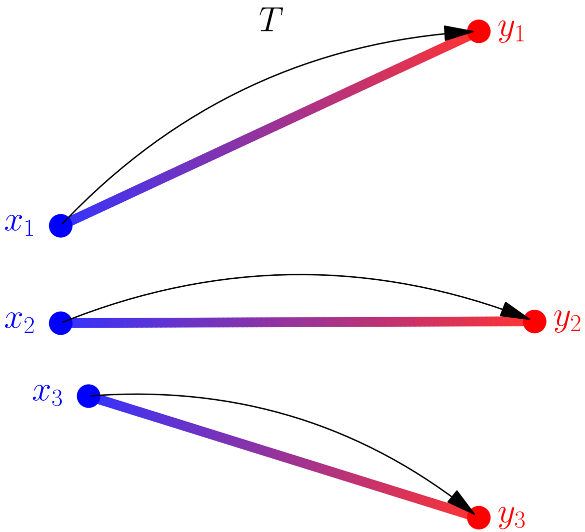


The Monge assignment problem



$$\min_{\sigma \in \mathfrak{S}_3} \sum_{i=1}^{3} \frac{1}{3} \|x_i - y_{\sigma(i)}\|^2$$

The Monge assignment problem



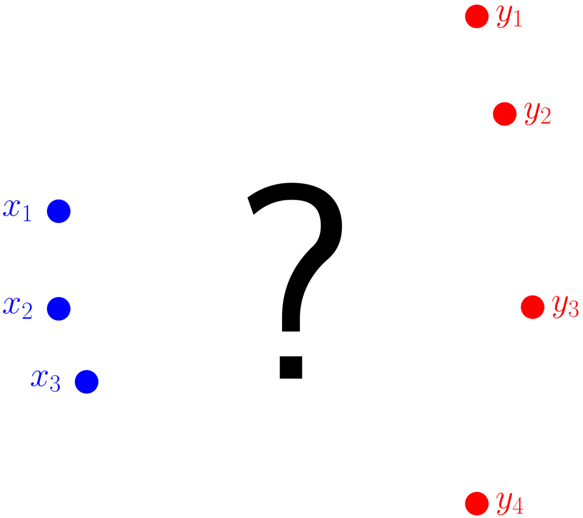
$$\min_{\sigma \in \mathfrak{S}_3} \sum_{i=1}^{3} \frac{1}{3} \|x_i - y_{\sigma(i)}\|^2$$

$$\mu := \frac{1}{3} \sum_{i=1}^{3} \delta_{x_i}$$

$$\nu := \frac{1}{3} \sum_{i=1}^{3} \delta_{y_i}$$

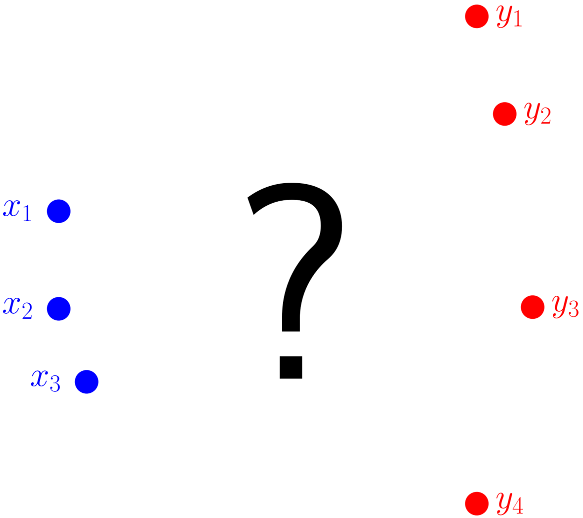
$$\min_{T_\# \mu = \nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

The Monge assignment problem



$$\min_{T_{\#}\mu=\nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

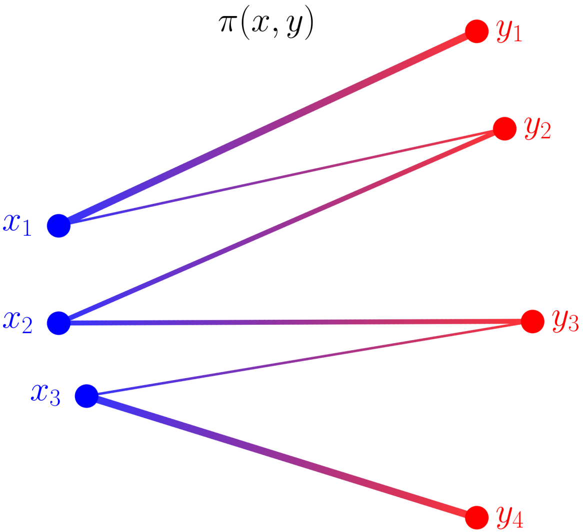
The Monge assignment problem



$$\min_{T_\# \mu = \nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

May be empty and is non-convex in general !

Kantorovitch relaxation



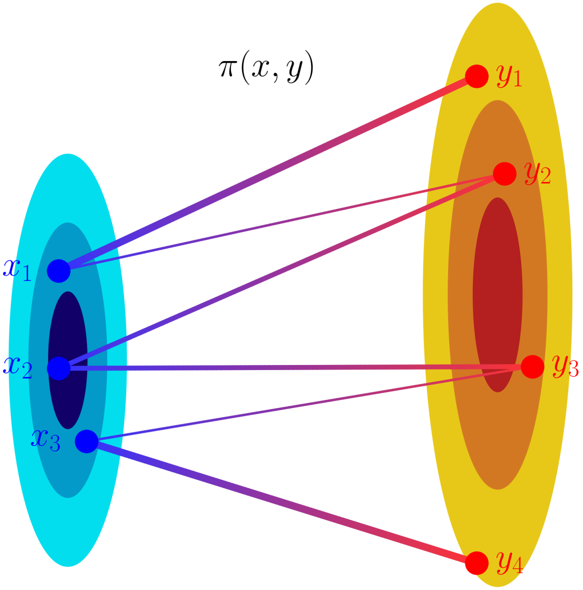
$$\min_{T_\# \mu = \nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

$\downarrow$

$$W_2(\mu, \nu)^2 := \min_{\pi \in \Pi(\mu, \nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y)$$

↳ Set of couplings

Kantorovitch relaxation



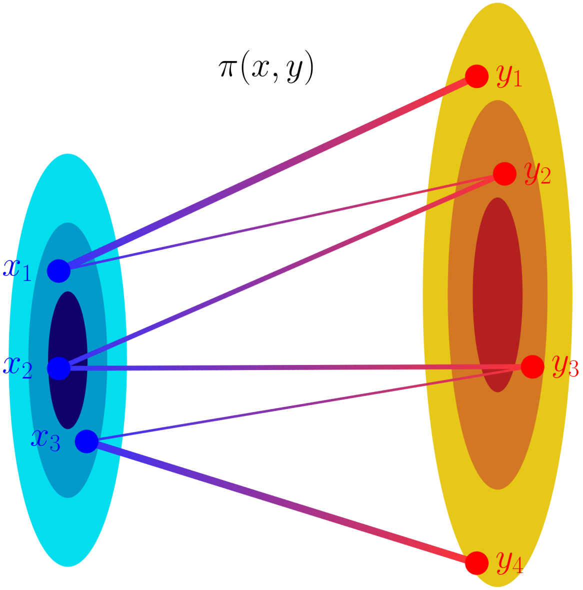
$$\min_{T_\# \mu = \nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

$$\downarrow$$

$$W_2(\mu, \nu)^2 := \min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y)$$

↳ Set of couplings

Kantorovich relaxation



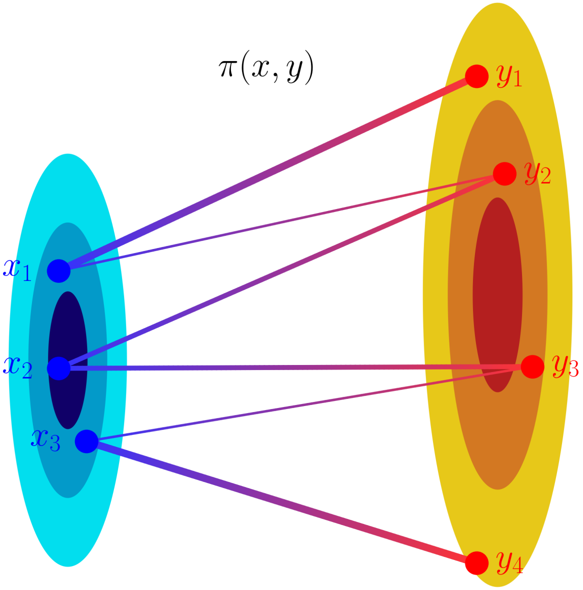
$$\min_{T_{\#}\mu=\nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

$$\Downarrow$$

$$W_2(\mu, \nu)^2 := \min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y)$$

$$= \max_{\varphi \oplus \psi \leq c} \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$$

Kantorovitch relaxation



$$\min_{T_\# \mu = \nu} \int \|x - T(x)\|^2 \mathrm{d}\mu(x)$$

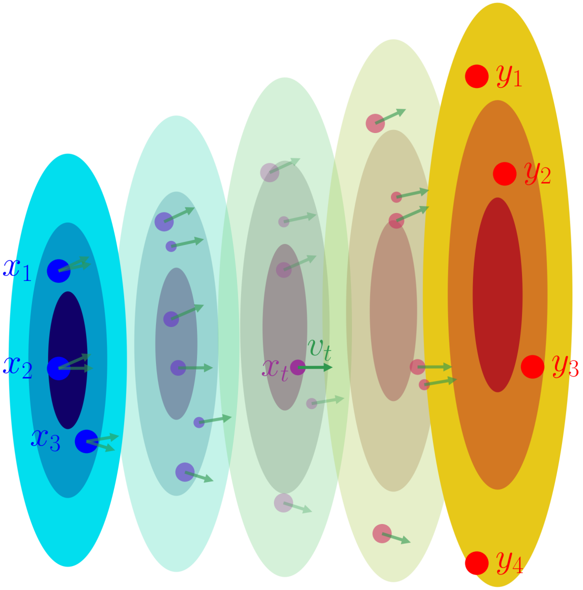
$$\Downarrow$$

$$W_2(\mu, \nu)^2 := \min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y)$$

$$= \max_{\varphi \oplus \psi \le c} \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$$

Theorem (Brenier). *When μ has a density, the Monge and Kantorovitch problems are equivalent: there is a unique Monge map $T = \mathrm{Id} - \frac{1}{2}\nabla\varphi$ and the optimal Kantorovitch plan is $(\mathrm{Id}, T)_\# \mu$.*

Benamou-Brenier dynamical formulation

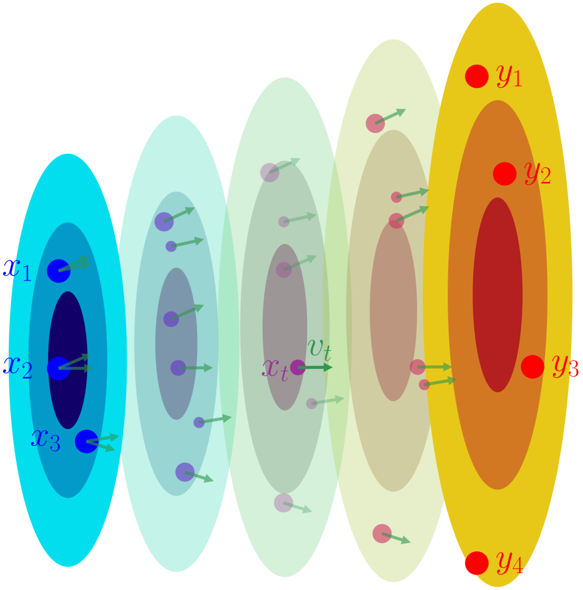


$$W_2(\mu,\nu)^2 = \min \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t$$

over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu.$

That is: particles follow $\dot{x}_t = v_t(x_t)$,
μ_t is the aggregate.

Benamou-Brenier dynamical formulation



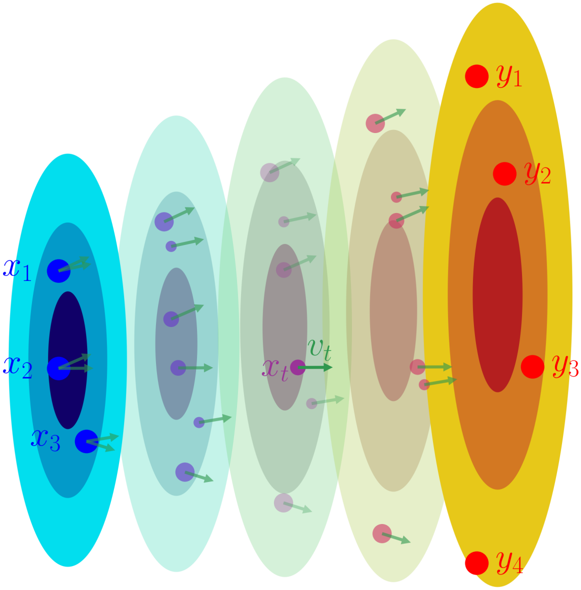
$$W_2(\mu,\nu)^2 = \min \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t$$

over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0$, $\mu_0 = \mu$, $\mu_1 = \nu$.

That is: particles follow $\dot{x}_t = v_t(x_t)$,
μ_t is the aggregate.

If T exists, $v_t = T - \mathrm{Id}$ for all t.

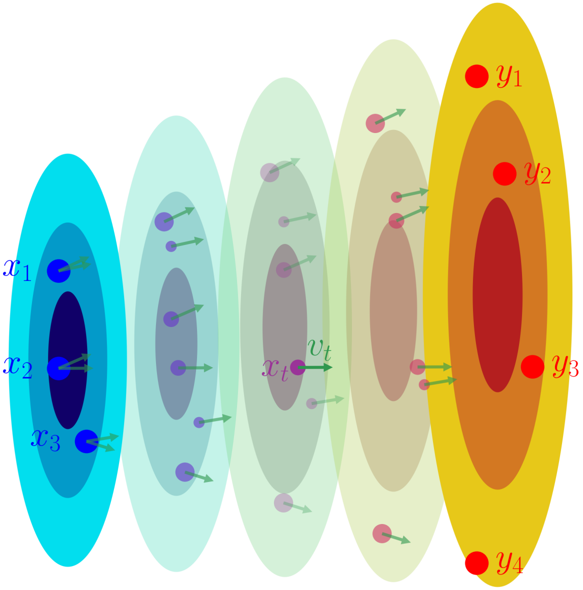
Benamou-Brenier dynamical formulation



$$W_2(\mu, \nu)^2 = \min \int_0^1 \|v_t\|^2_{L^2_{\mu_t}} \, \mathrm{d}t$$

over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0$, $\mu_0 = \mu$, $\mu_1 = \nu$.

Benamou-Brenier dynamical formulation

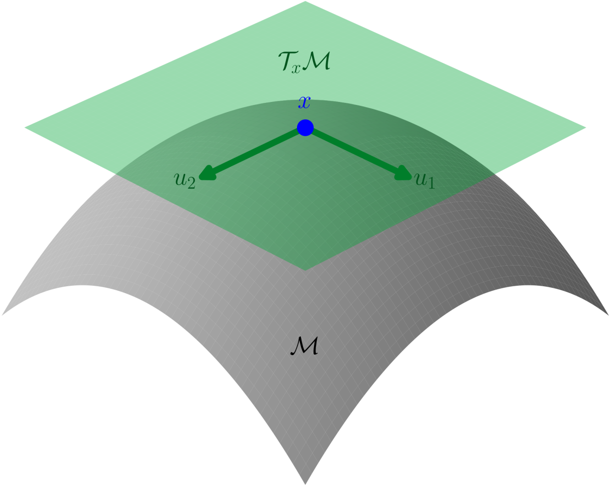


$$W_2(\mu, \nu)^2 = \min \int_0^1 \|v_t\|^2_{L^2_{\mu_t}} \, \mathrm{d}t$$

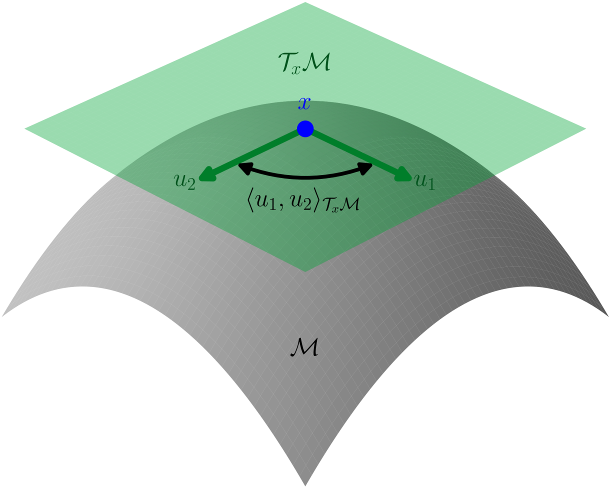
over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0$, $\mu_0 = \mu$, $\mu_1 = \nu$.

Looks like a formula from Riemannian geometry...

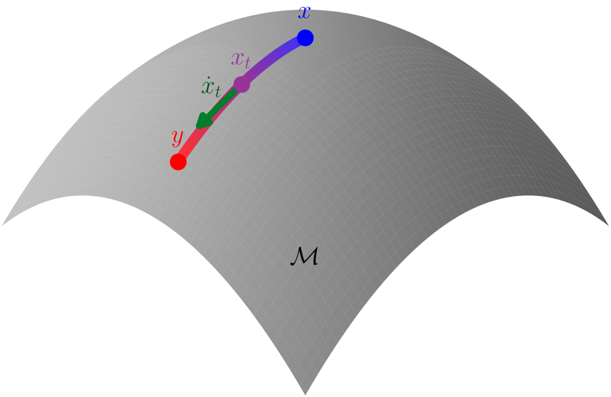
Riemannian geometry



Riemannian geometry



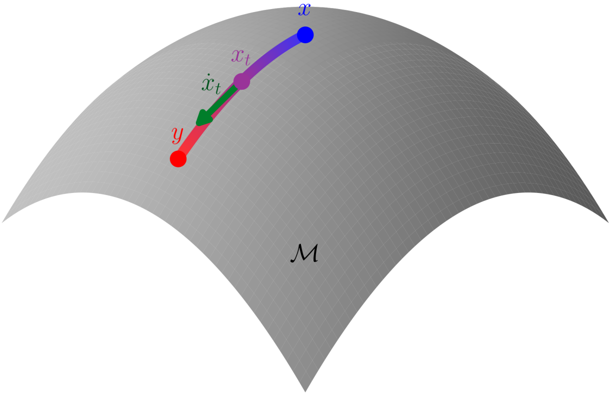
Riemannian geometry



$$\mathsf{d}(x, y)^2 = \min \int_0^1 \|\dot{x}_t\|^2_{\mathcal{T}_{x_t}\mathcal{M}} \mathrm{d}t$$

over $(x_t)_t \subset \mathcal{M}$ such that $x_0 = x$, $x_1 = y$

Riemannian geometry



$$\mathsf{d}(x, y)^2 = \min \int_0^1 \|\dot{x}_t\|^2_{\mathcal{T}_{x_t}\mathcal{M}} \mathrm{d}t$$

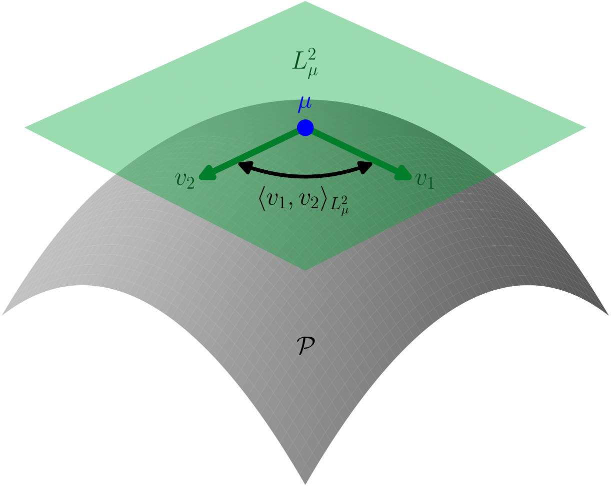
over $(x_t)_t \subset \mathcal{M}$ such that $x_0 = x$, $x_1 = y$

$\Updownarrow$

$$W_2(\mu, \nu)^2 = \min \int_0^1 \|v_t\|^2_{L^2_{\mu_t}} \mathrm{d}t$$

over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0$, $\mu_0 = \mu$, $\mu_1 = \nu$.

Riemannian geometry



$$\mathsf{d}(x,y)^2 = \min \int_0^1 \|\dot{x}_t\|^2_{\mathcal{T}_{x_t}\mathcal{M}} \mathrm{d}t$$

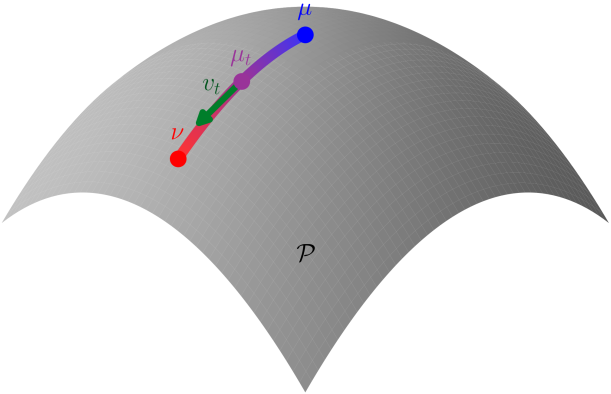
over $(x_t)_t \subset \mathcal{M}$ such that $x_0 = x$, $x_1 = y$

$$\Updownarrow$$

$$W_2(\mu,\nu)^2 = \min \int_0^1 \|v_t\|^2_{L^2_{\mu_t}} \mathrm{d}t$$

over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0$, $\mu_0 = \mu$, $\mu_1 = \nu$.

Riemannian geometry



$$\mathsf{d}(x, y)^2 = \min \int_0^1 \|\dot{x}_t\|^2_{\mathcal{T}_{x_t}\mathcal{M}} \mathrm{d}t$$

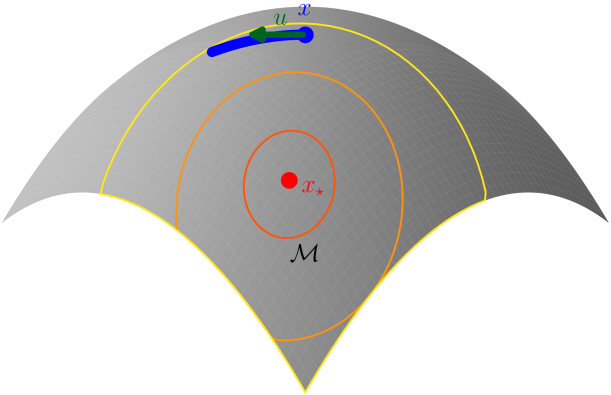
over $(x_t)_t \subset \mathcal{M}$ such that $x_0 = x$, $x_1 = y$

$\Updownarrow$

$$W_2(\mu, \nu)^2 = \min \int_0^1 \|v_t\|^2_{L^2_{\mu_t}} \mathrm{d}t$$

over $\dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0$, $\mu_0 = \mu$, $\mu_1 = \nu$.

Gradient flows

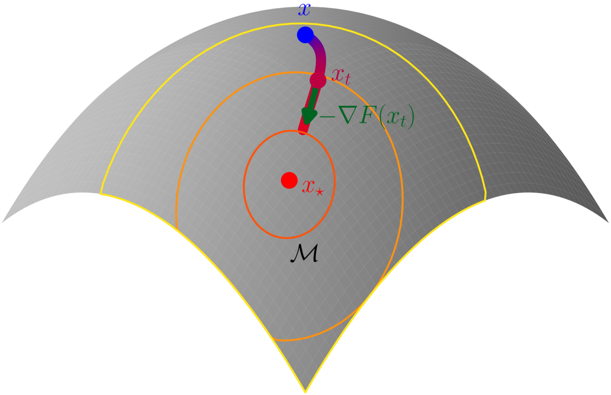


For $F : \mathcal{M} \to \mathbb{R}$,

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} F(x_s^u) = \langle \nabla F(x), u \rangle_{\mathcal{T}_x \mathcal{M}}$$

with $x_0^u = x$ and $\dot{x}_0^u = u$

Gradient flows



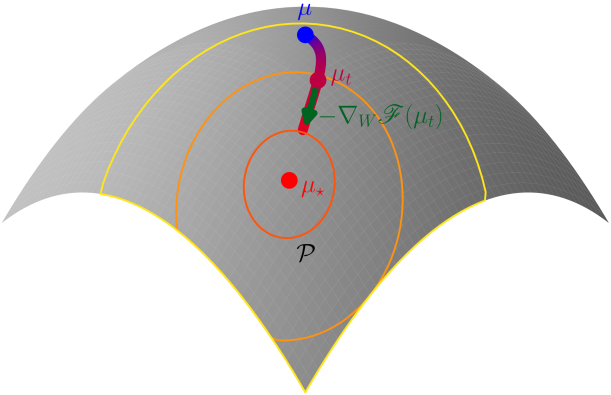
For $F : \mathcal{M} \to \mathbb{R}$,

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} F(x_s^u) = \langle \nabla F(x), u \rangle_{\mathcal{T}_x\mathcal{M}}$$

with $x_0^u = x$ and $\dot{x}_0^u = u$

Gradient flow: $\dot{x}_t = -\nabla F(x_t)$

Gradient flows



For $F : \mathcal{M} \to \mathbb{R}$,

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} F(x_s^u) = \langle \nabla F(x), u \rangle_{\mathcal{T}_x \mathcal{M}}$$

with $x_0^u = x$ and $\dot{x}_0^u = u$

Gradient flow: $\dot{x}_t = -\nabla F(x_t)$

$$\Updownarrow$$

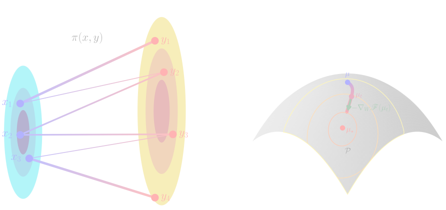
For $\mathscr{F} : \mathcal{P} \to \mathbb{R}$,

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=0} \mathscr{F}(\mu_s^v) = \langle \nabla_W \mathscr{F}(\mu), v \rangle_{L^2_\mu}$$

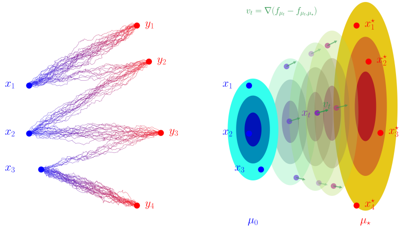
with $\mu_0^v = \mu$ and $\dot{\mu}_0^v = v$

Gradient flow: $\begin{array}{c} \dot{\mu}_t + \mathrm{div}(\mu_t v_t) = 0, \\ v_t = -\nabla_W \mathscr{F}(\mu_t) \end{array}$

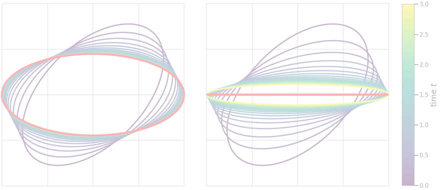
Plan



1. Optimal transport and gradient flows

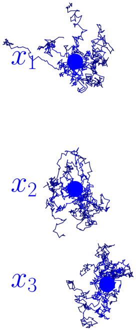


2. The Sinkhorn divergence and its flow

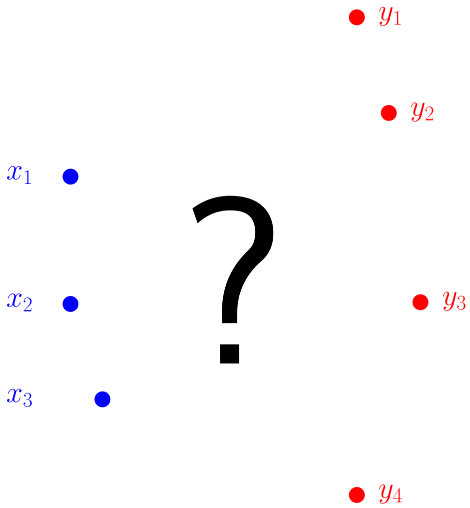


3. Main results

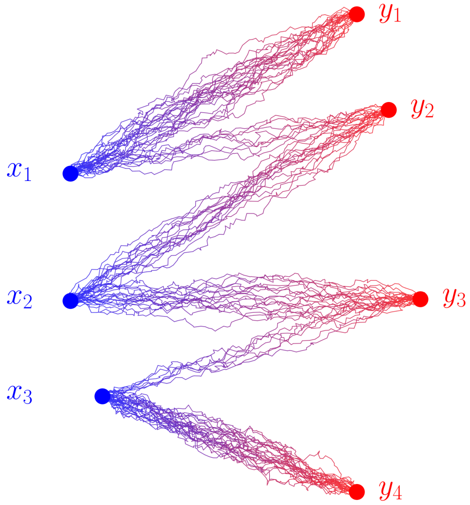
Entropic optimal transport



Entropic optimal transport



Entropic optimal transport

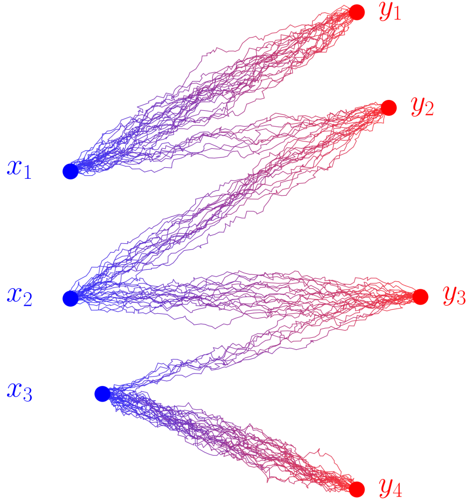


Dynamic Schrödinger problem:

$$\min_{P_0=\mu, P_1=\nu} \mathrm{KL}(P|R)$$

where P is a distribution of paths, R is the Brownian motion of diffusivity $\varepsilon > 0$.

Entropic optimal transport



Dynamic Schrödinger problem:

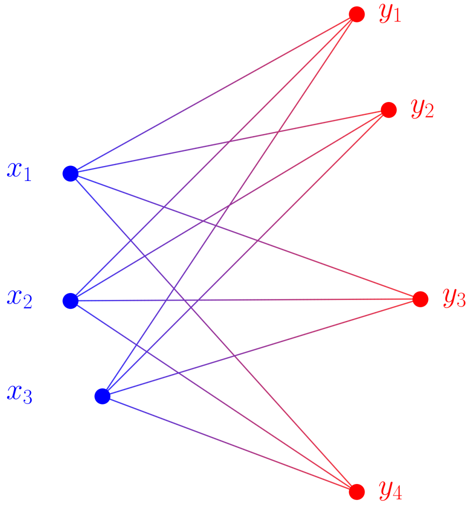
$$\min_{P_0=\mu, P_1=\nu} \mathrm{KL}(P|R)$$

where P is a distribution of paths,
R is the Brownian motion of
diffusivity $\varepsilon > 0$.

Entropic optimal transport problem:

$$\min_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu \otimes \nu)$$

Entropic optimal transport



Dynamic Schrödinger problem:

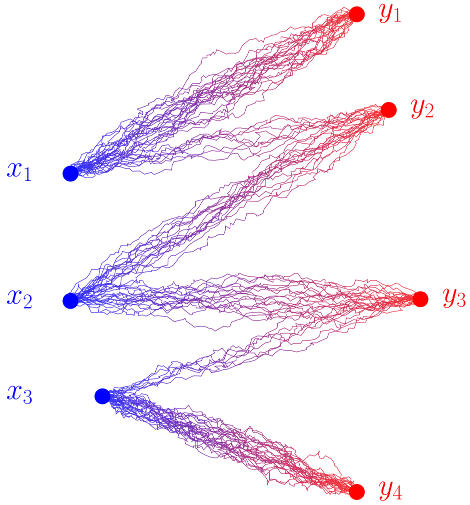
$$\min_{P_0=\mu, P_1=\nu} \mathrm{KL}(P|R)$$

where P is a distribution of paths,
R is the Brownian motion of
diffusivity $\varepsilon > 0$.

Entropic optimal transport problem:

$$\min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi | \mu \otimes \nu)$$

Entropic optimal transport



Dynamic Schrödinger problem:

$$\min_{P_0=\mu, P_1=\nu} \mathrm{KL}(P|R)$$

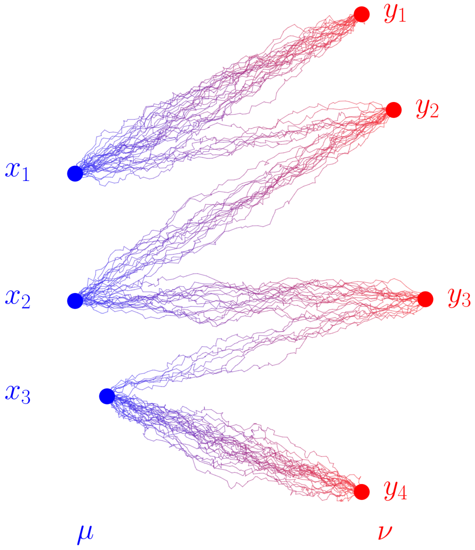
where P is a distribution of paths, R is the Brownian motion of diffusivity $\varepsilon > 0$.

Entropic optimal transport problem:

$$\min_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu \otimes \nu)$$

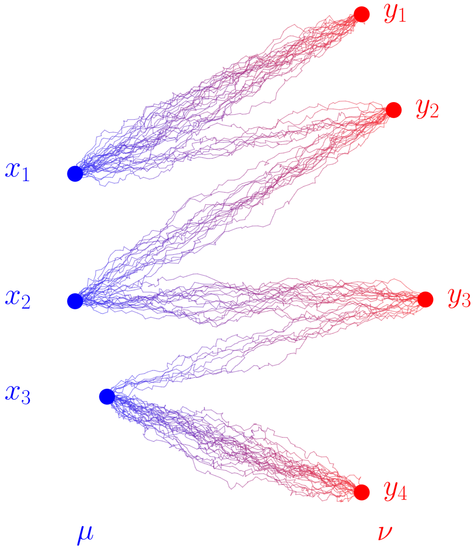
$\pi(x,y)$ is the mass that goes from x to y in the Schrödinger bridge.

Dual problem and barycentric map



$$\mathrm{OT}_\varepsilon(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi | \mu \otimes \nu)$$

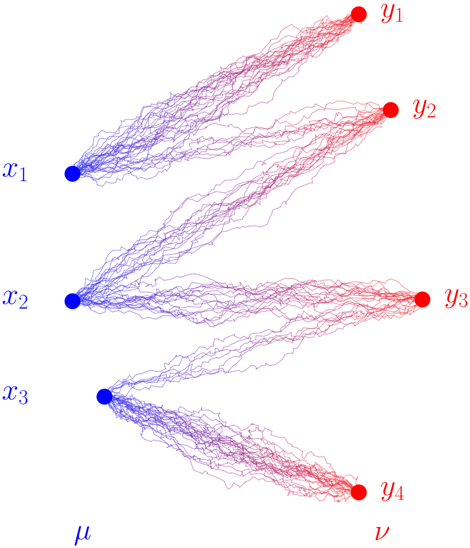
Dual problem and barycentric map



$$\mathrm{OT}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$

$$= \max_{f,g} \int f\mathrm{d}\mu + \int g\mathrm{d}\nu - \varepsilon\langle \mu\otimes\nu, e^{\frac{1}{\varepsilon}(f\oplus g - c)} - 1\rangle$$

Dual problem and barycentric map

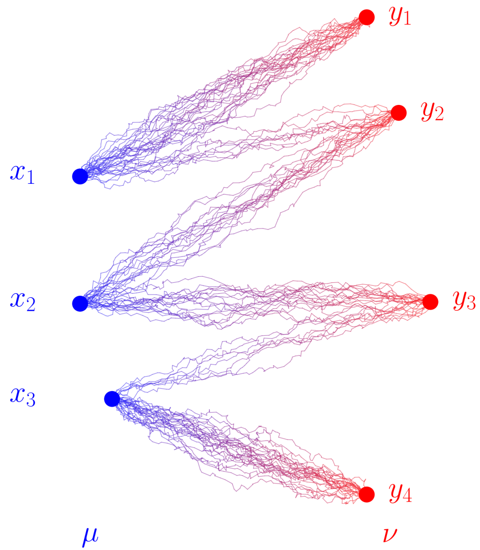


$$\mathrm{OT}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$

$$= \max_{f,g} \int f\mathrm{d}\mu + \int g\mathrm{d}\nu - \varepsilon\langle \mu\otimes\nu, e^{\frac{1}{\varepsilon}(f\oplus g-c)} - 1\rangle$$

Recall: $W_2^2(\mu,\nu) = \max_{\varphi\oplus\psi\le c} \int \varphi\mathrm{d}\mu + \int \psi\mathrm{d}\nu$

Dual problem and barycentric map



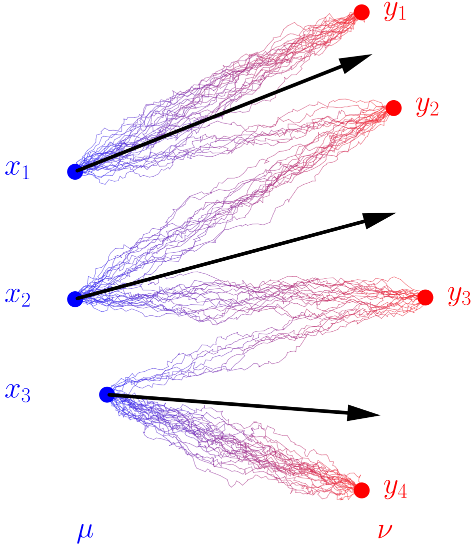
$$\mathrm{OT}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$

$$= \max_{f,g} \int f\mathrm{d}\mu + \int g\mathrm{d}\nu - \varepsilon\langle \mu\otimes\nu, e^{\frac{1}{\varepsilon}(f\oplus g-c)} - 1\rangle$$

Recall: $W_2^2(\mu,\nu) = \max_{\varphi\oplus\psi\le c} \int \varphi\mathrm{d}\mu + \int \psi\mathrm{d}\nu$

$$T = \mathrm{Id} - \tfrac{1}{2}\nabla\varphi$$

Dual problem and barycentric map



$$\mathrm{OT}_{\varepsilon}(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$

$$= \max_{f,g} \int f \mathrm{d}\mu + \int g \mathrm{d}\nu - \varepsilon\langle \mu\otimes\nu, e^{\frac{1}{\varepsilon}(f\oplus g - c)} - 1\rangle$$

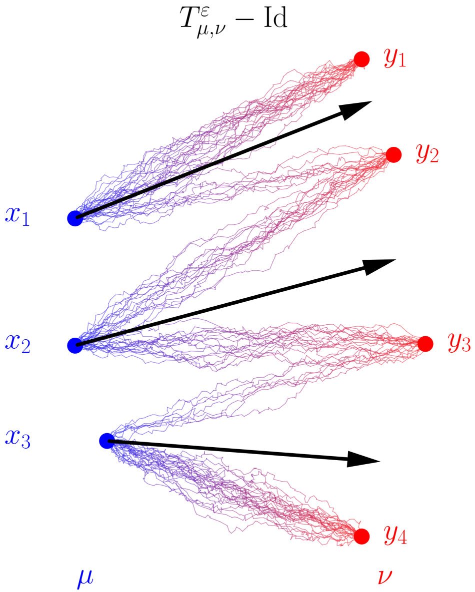
Recall: $W_2^2(\mu,\nu) = \max_{\varphi\oplus\psi\leq c} \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$

$$T = \mathrm{Id} - \tfrac{1}{2}\nabla\varphi$$

Writing $f_{\mu,\nu}, g_{\mu,\nu}$ the Schrödinger potentials,

$$T^{\varepsilon}_{\mu,\nu} := \mathrm{Id} - \tfrac{1}{2}\nabla f_{\mu,\nu}$$

Dual problem and barycentric map



$$\mathrm{OT}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$

$$= \max_{f,g} \int f \mathrm{d}\mu + \int g \mathrm{d}\nu - \varepsilon\langle \mu\otimes\nu, e^{\frac{1}{\varepsilon}(f\oplus g - c)} - 1\rangle$$

Recall: $W_2^2(\mu,\nu) = \max_{\varphi\oplus\psi\leq c} \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$

$$T = \mathrm{Id} - \tfrac{1}{2}\nabla\varphi$$

Writing $f_{\mu,\nu}, g_{\mu,\nu}$ the Schrödinger potentials,

$$T^{\varepsilon}_{\mu,\nu} := \mathrm{Id} - \tfrac{1}{2}\nabla f_{\mu,\nu}$$

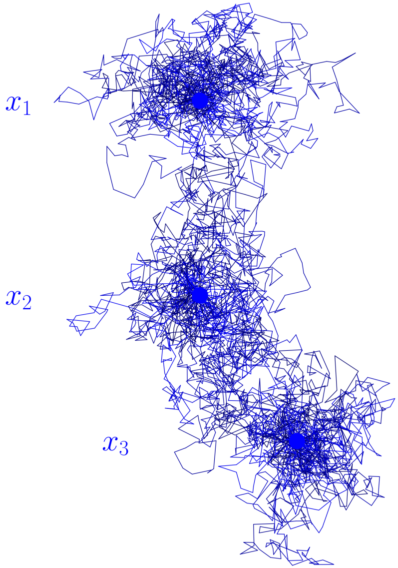
$$x \mapsto \int y \underbrace{p_{\mu,\nu}(x,y)}_{\propto \pi(x,y)} \mathrm{d}\nu(y)$$

The Sinkhorn divergence

Advantages of OT_ε:

- Computed efficiently (Sinkhorn's algorithm)
- Smooth
- Retains the geometric flavour of W_2

The Sinkhorn divergence

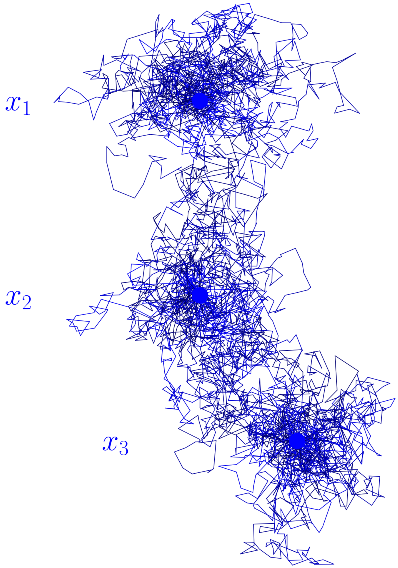


Advantages of OT_ε:

- Computed efficiently (Sinkhorn's algorithm)
- Smooth
- Retains the geometric flavour of W_2

Main drawback: biased ! $\mathrm{OT}_\varepsilon(\mu, \mu) \neq 0$ and even $\arg\min_\nu \mathrm{OT}_\varepsilon(\mu, \nu) \neq \mu$ (shrinks)

The Sinkhorn divergence



Advantages of OT_ε:

- Computed efficiently (Sinkhorn's algorithm)
- Smooth
- Retains the geometric flavour of W_2

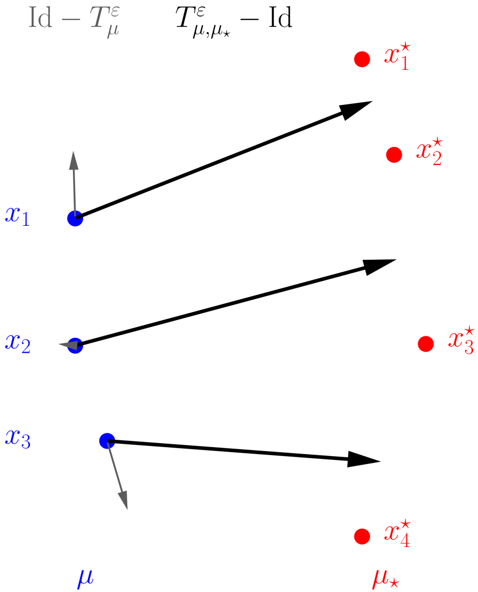
Main drawback: biased ! $\mathrm{OT}_\varepsilon(\mu,\mu) \neq 0$ and even $\arg\min_\nu \mathrm{OT}_\varepsilon(\mu,\nu) \neq \mu$ (shrinks)

↓

Solution: simply subtract bias !

$$S_\varepsilon(\mu,\nu) := \mathrm{OT}_\varepsilon(\mu,\nu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu)$$

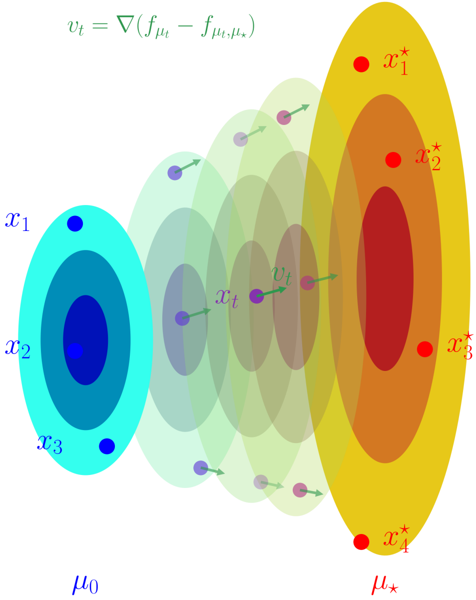
Wasserstein gradient flow of S_ε



We *should* have

$$-\nabla_W^1 S_\varepsilon(\mu, \mu_\star) = \nabla(f_\mu - f_{\mu,\mu_\star})$$
$$= 2(T_{\mu,\mu_\star}^\varepsilon - \mathrm{Id} + \mathrm{Id} - T_\mu^\varepsilon)$$

Wasserstein gradient flow of S_ε



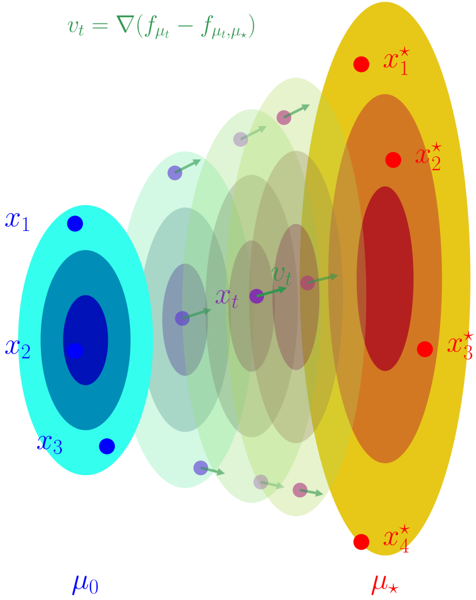
We *should* have

$$-\nabla^1_W S_\varepsilon(\mu, \mu_\star) = \nabla(f_\mu - f_{\mu,\mu_\star})$$
$$= 2(T^\varepsilon_{\mu,\mu_\star} - \mathrm{Id} + \mathrm{Id} - T^\varepsilon_\mu)$$

The gradient flow *should* be described by

$$\dot{\mu}_t + \mathrm{div}(\mu_t \nabla(f_{\mu_t} - f_{\mu_t,\mu_\star})) = 0$$

Wasserstein gradient flow of S_ε



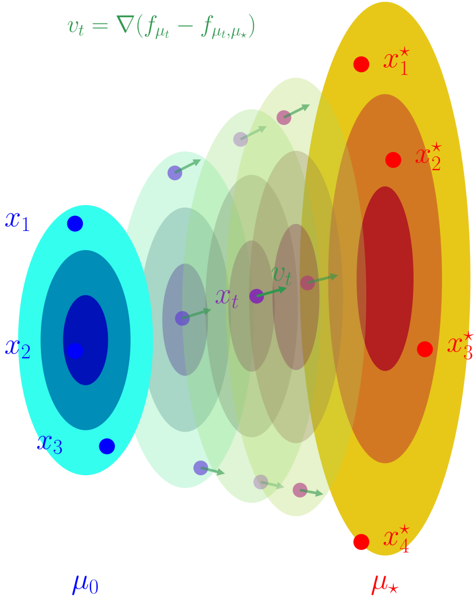
We *should* have

$$-\nabla_W^1 S_\varepsilon(\mu,\mu_\star) = \nabla(f_\mu - f_{\mu,\mu_\star}) \\ = 2(T^\varepsilon_{\mu,\mu_\star} - \mathrm{Id} + \mathrm{Id} - T^\varepsilon_\mu)$$

The gradient flow *should* be described by
$$\dot\mu_t + \mathrm{div}(\mu_t \nabla(f_{\mu_t} - f_{\mu_t,\mu_\star})) = 0$$

Working assumptions: $\mu_0, \mu_\star$ Gaussian

Wasserstein gradient flow of S_ε



We *should* have

$$-\nabla_W^1 S_\varepsilon(\mu,\mu_\star) = \nabla(f_\mu - f_{\mu,\mu_\star}) = 2(T^\varepsilon_{\mu,\mu_\star} - \mathrm{Id} + \mathrm{Id} - T^\varepsilon_\mu)$$

The gradient flow *should* be described by

$$\dot\mu_t + \mathrm{div}(\mu_t \nabla(f_{\mu_t} - f_{\mu_t,\mu_\star})) = 0$$

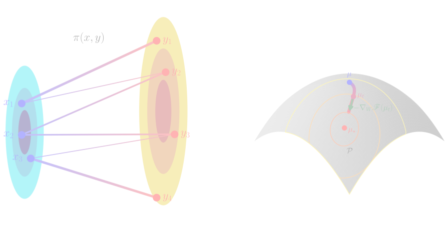
Working assumptions: $\mu_0, \mu_\star$ Gaussian

For $\mu = \mathcal{N}(m,\Sigma), \nu = \mathcal{N}(n,\Gamma)$, we have

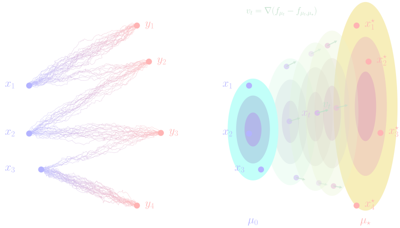
$$S_\varepsilon(\mu,\nu) = \|m-n\|^2 + B_\varepsilon(\Sigma,\Gamma)$$

$\longrightarrow$ we can consider centered measures.

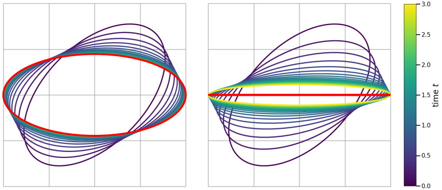
Plan



1. Optimal transport and gradient flows



2. The Sinkhorn divergence and its flow



3. Main results

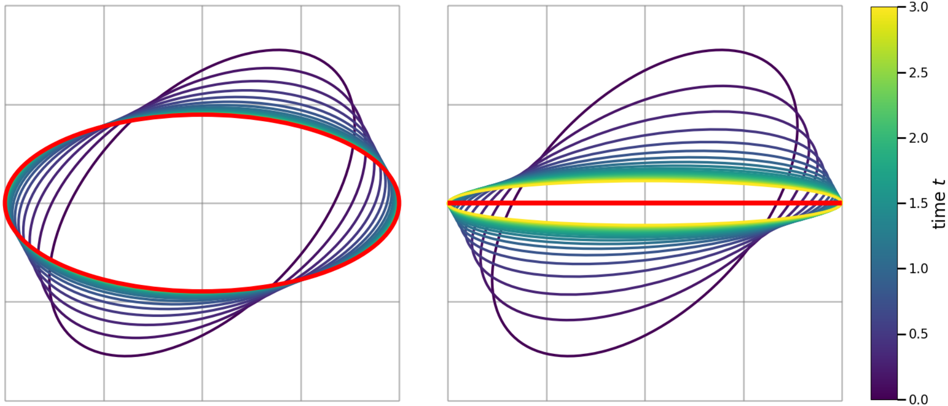
Well-posedness

Theorem (MH, T. Lacombe (2026)). *μ_0, $\mu_\star$ Gaussian (can be singular).*
There exists a solution to

$$\dot{\mu}_t + \operatorname{div}(\mu_t \nabla (f_{\mu_t} - f_{\mu_t,\mu_\star})) = 0$$

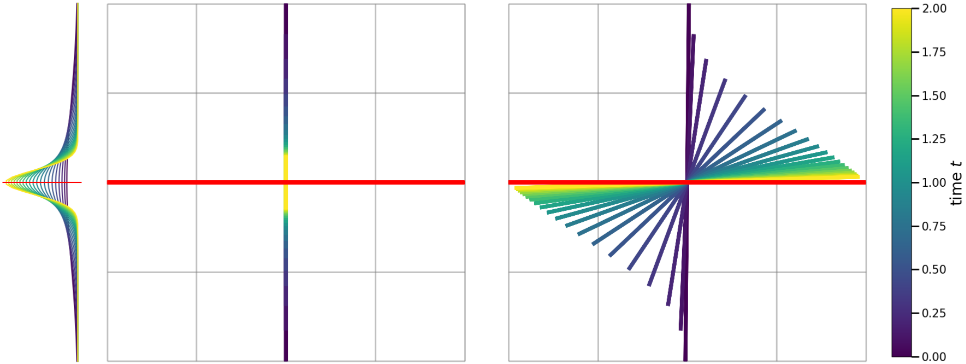
which stays Gaussian.
It is unique among Gaussians and in a larger class $\mathcal{R} = \{\exp(-V), \alpha_V I \preceq \nabla^2 V \preceq \beta_V I\}$.
It is a Wasserstein gradient flow.

Convergence



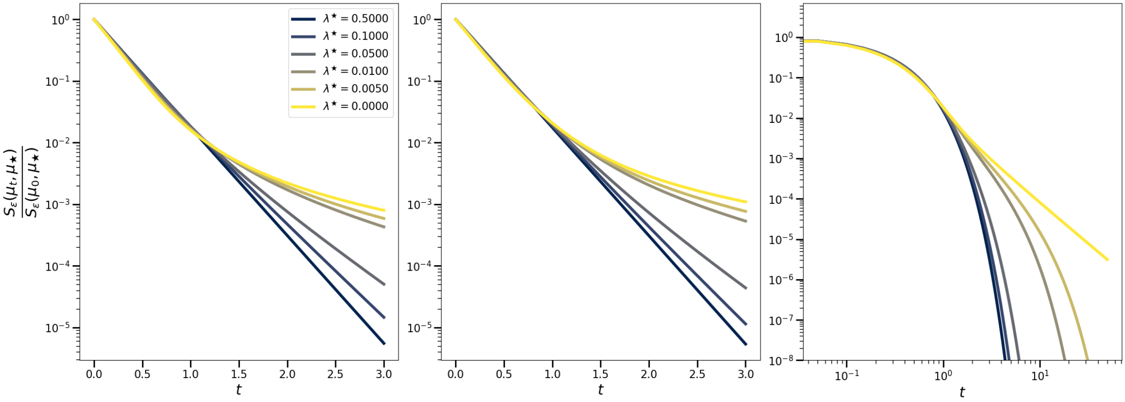
Theorem (MH, T. Lacombe). *If μ_0 is non-singular, $\mu_t \xrightarrow[t\to\infty]{} \mu_\star$.*

Convergence



Theorem (MH, T. Lacombe). μ_0 *is singular* $\iff \forall t, \mu_t$ *also singular.*
$\Sigma_t \xrightarrow[t\to\infty]{} \Sigma_\infty = P\mathrm{diag}((\lambda_i)_i)P^T$ *where* $\Sigma_\star = P\mathrm{diag}((\lambda_i^\star)_i)P^T$ *and* $\lambda_i \in \{0, \lambda_i^\star\}$

Convergence



Theorem (MH, T. Lacombe). *If Σ_0 and $\Sigma_\star$ commute, convergence holds iff $\mathrm{supp}(\mu_\star) \subset \mathrm{supp}(\mu_0)$, in $O(e^{-Ct})$ if equality and $O(\frac{1}{t})$ otherwise.*

Conclusion

What we saw:

- The Wasserstein space and its geometry, gradient flows
- Entropic optimal transport, the Sinkhorn divergence and its flow
- First convergence properties for Gaussians

Next steps:

- Particle case study
- More general results: convergence criterion related to existence and uniqueness of Monge maps ?

Thank you for your attention !

Appendix

Convergence rate as function of ε