The geometry of Sinkhorn divergences

Hugo Lavenant

Bocconi University

Workshop “Variational Analysis, Models and Methods in Measure Spaces”

Marseille (France), April 30, 2024

Joint work with

Jonas Luckhardt

UNIVERSITÄT GÖTTINGEN

Gilles Mordant

UNIVERSITÄT GÖTTINGEN

Bernhard Schmitzer

UNIVERSITÄT GÖTTINGEN

Luca Tamanini

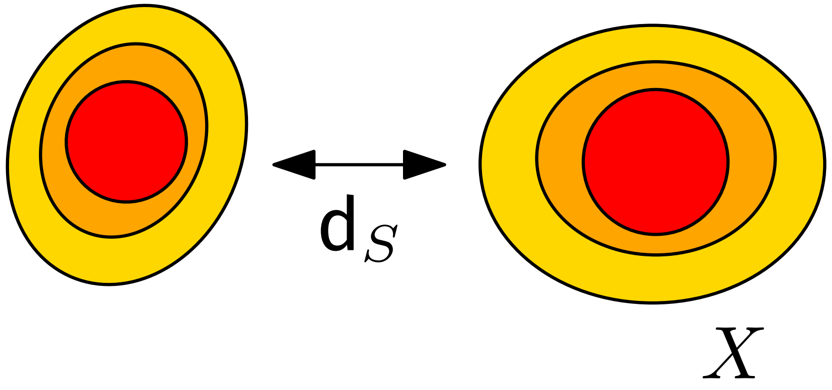
UNIVERSITAS CATHOLICA SACRI CORDIS IESU MEDIOLANI

Preprint on arxiv hopefully soon!

What I will present

$\mathcal{P}(X)$ probability distributions over (X, d) compact metric space.

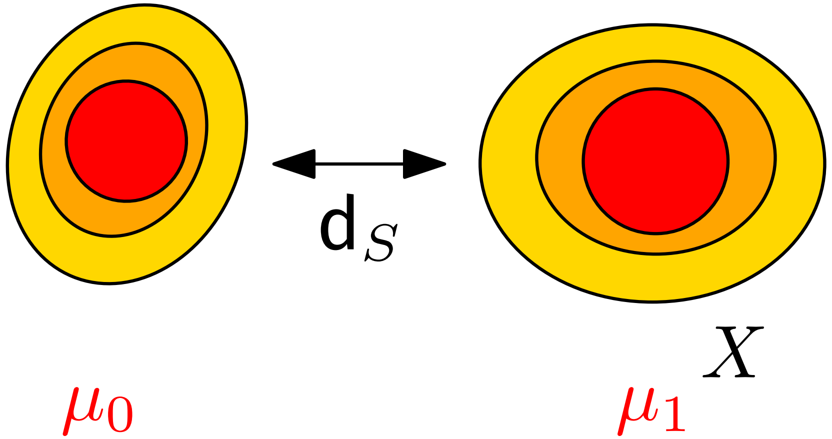
We propose d_S a new distance over $\mathcal{P}(X)$:



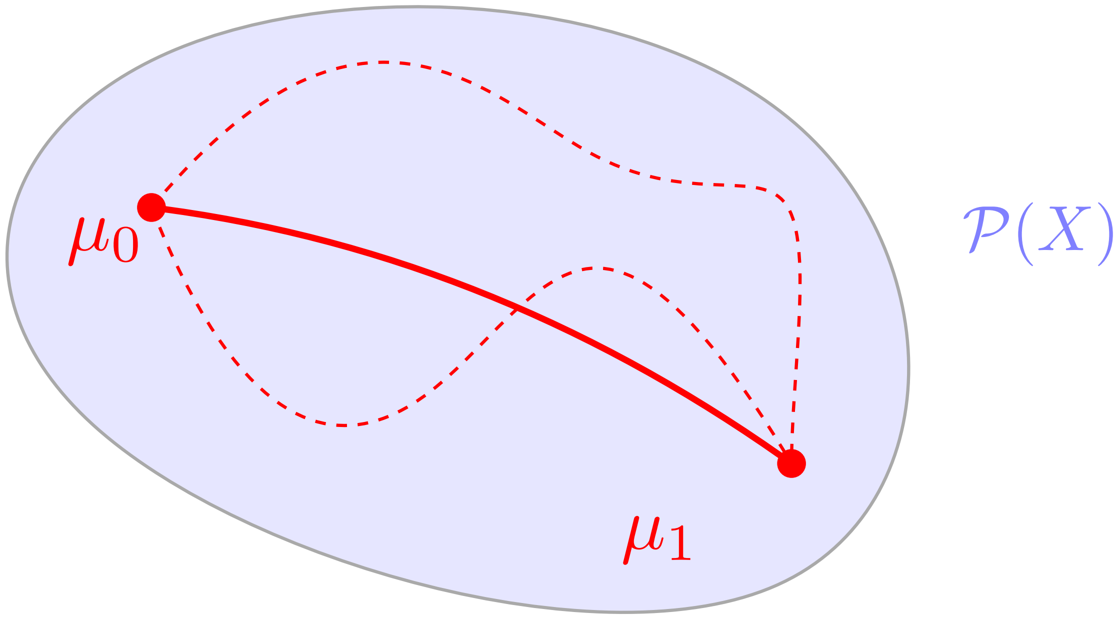
What I will present

$\mathcal{P}(X)$ probability distributions over (X, d) compact metric space.

We propose d_S a new distance over $\mathcal{P}(X)$:

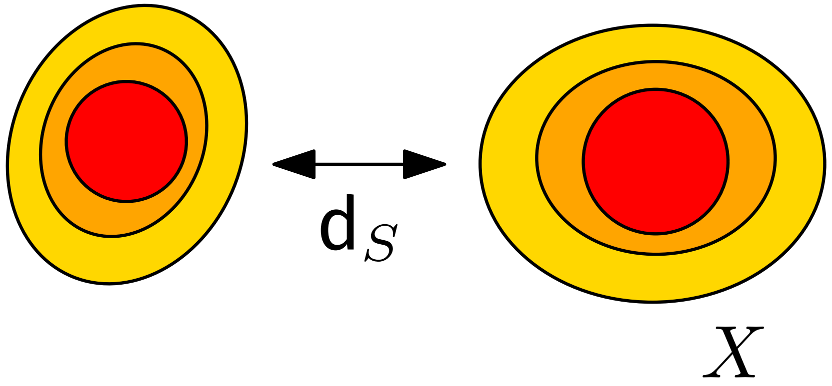


1. It is a "Riemannian" metric.



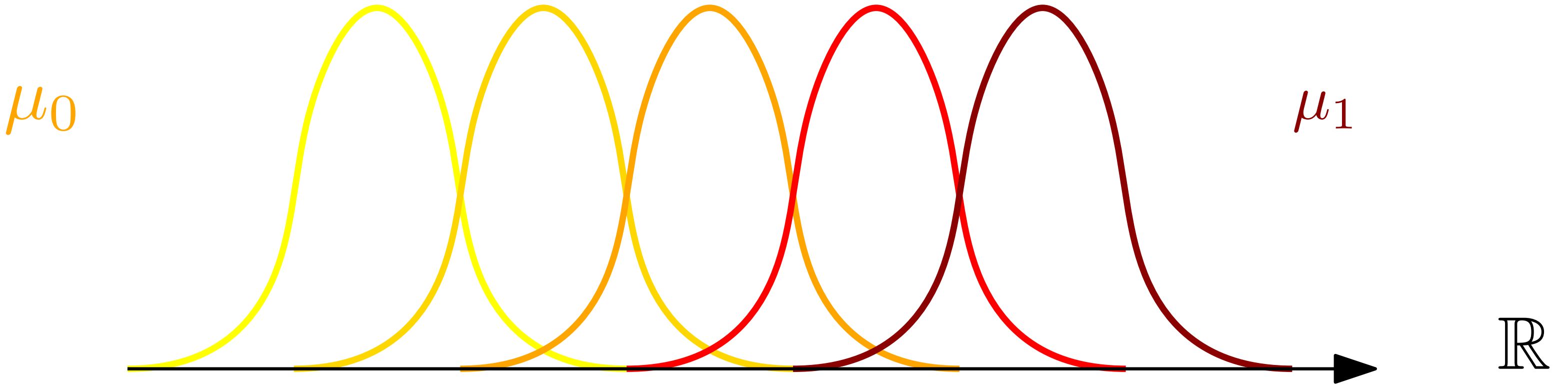
What I will present

$\mathcal{P}(X)$ probability distributions over (X, d) compact metric space.



We propose d_S a new distance over $\mathcal{P}(X)$:

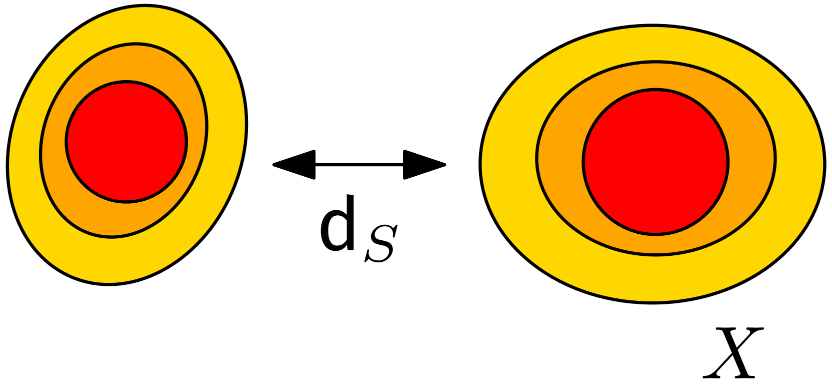
1. It is a "Riemannian" metric.
2. Translation are geodesics for this metric.



What I will present

$\mathcal{P}(X)$ probability distributions over (X, d) compact metric space.

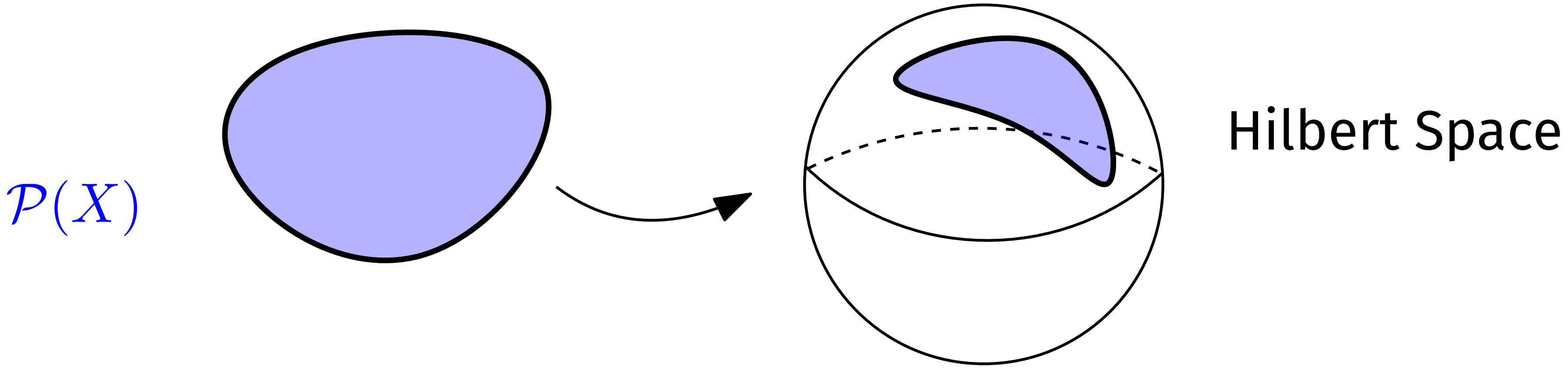
We propose d_S a new distance over $\mathcal{P}(X)$:



1. It is a “Riemannian” metric.

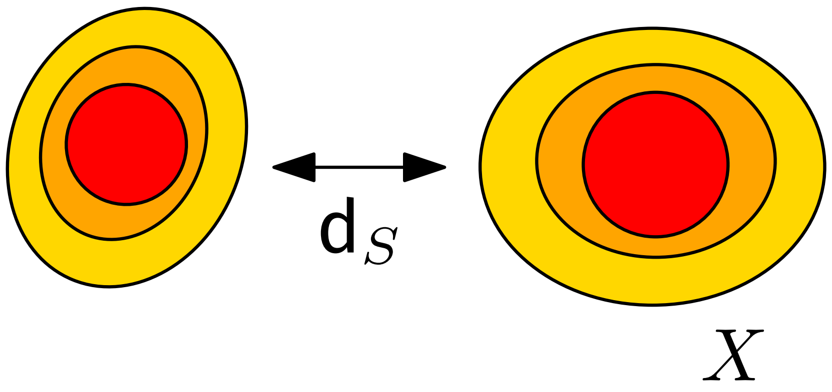
2. Translation are geodesics for this metric.

3. The metric tensor is “smooth” and (X, d_S) embeds bi-Lipschitzly in a (RK) Hilbert space.



What I will present

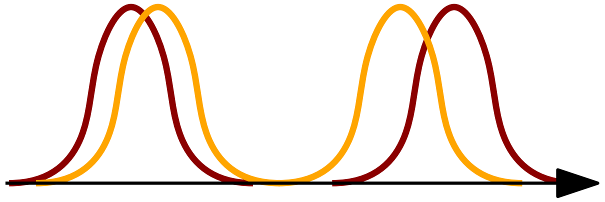
$\mathcal{P}(X)$ probability distributions over (X, d) compact metric space.



We propose d$_S$ a new distance over $\mathcal{P}(X)$:

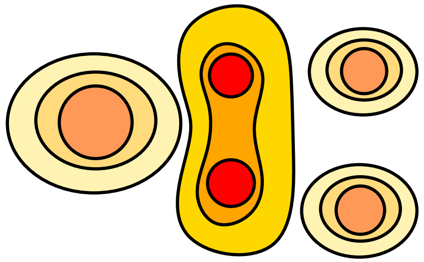
1. It is a “Riemannian” metric.
2. Translation are geodesics for this metric.
3. The metric tensor is “smooth” and (X, d_S) embeds bi-Lipschitzly in a (RK) Hilbert space.

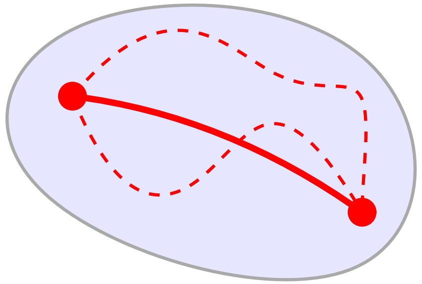
Idea: construct a Riemannian distance out of entropic optimal transport.



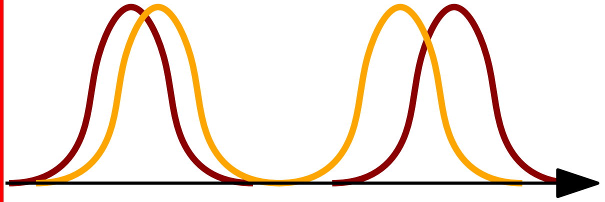
1 - Optimal transport and its geometry

2 - Entropic optimal transport and Sinkhorn divergences



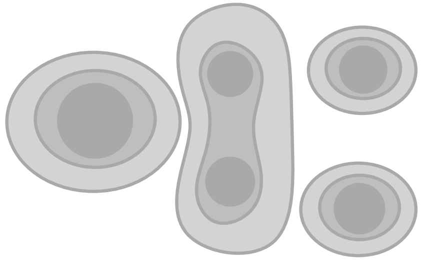


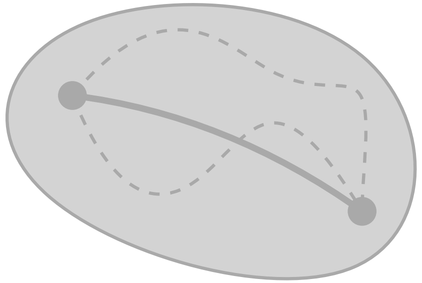
3 - Building a Riemannian geometry out of Sinkhorn divergences



1 - Optimal transport and its geometry

2 - Entropic optimal transport and Sinkhorn divergences





3 - Building a Riemannian geometry out of Sinkhorn divergences

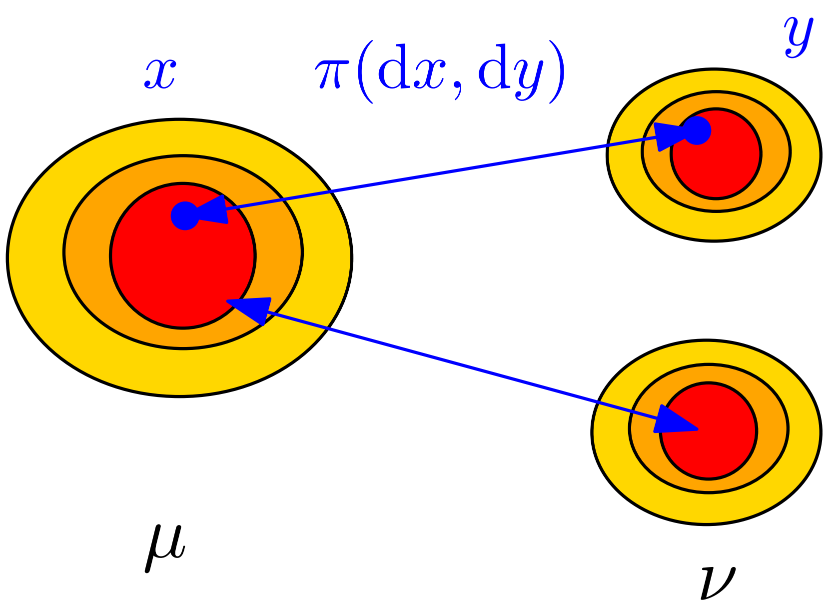
Quadratic optimal transport

(X, d) compact metric space.

Definition

$$\mathrm{OT}(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \iint_{X\times X} d(x,y)^2 \, \mathrm{d}\pi(x,y)$$

Subset of $\mathcal{P}(X \times X)$, coupling between μ and ν

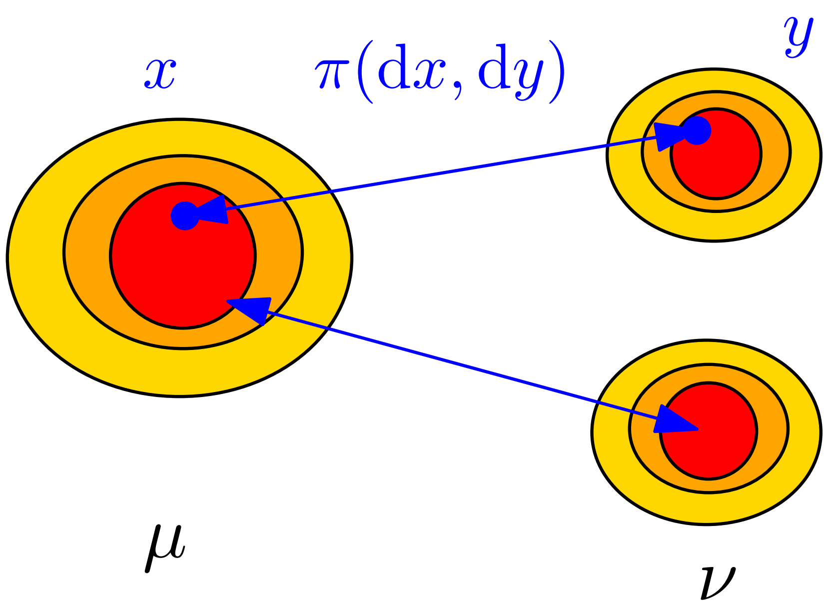


Quadratic optimal transport

(X, d) compact metric space.

Definition

$$\mathrm{OT}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \iint_{X \times X} d(x, y)^2 \, \mathrm{d}\pi(x, y)$$

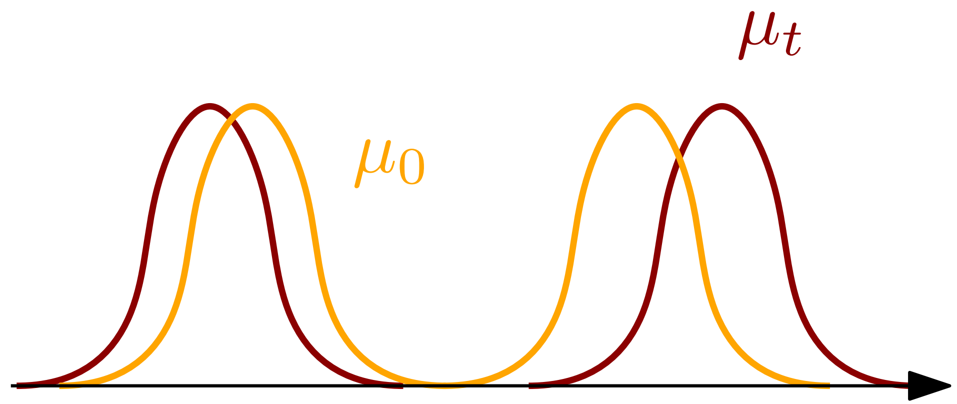


Theorem. OT is the square of a distance on $\mathcal{P}(X)$ metrizing the weak convergence.

The linearization of optimal transport

On $\mathbb{R}^d$, what happens to $\mathrm{OT}(\mu, \nu)$ if $\mu \simeq \nu$?

$\rightsquigarrow$ $(\mu_t)_t$ curve in $\mathcal{P}(\mathbb{R}^d)$, we look at $\mathrm{OT}(\mu_0, \mu_t)$.

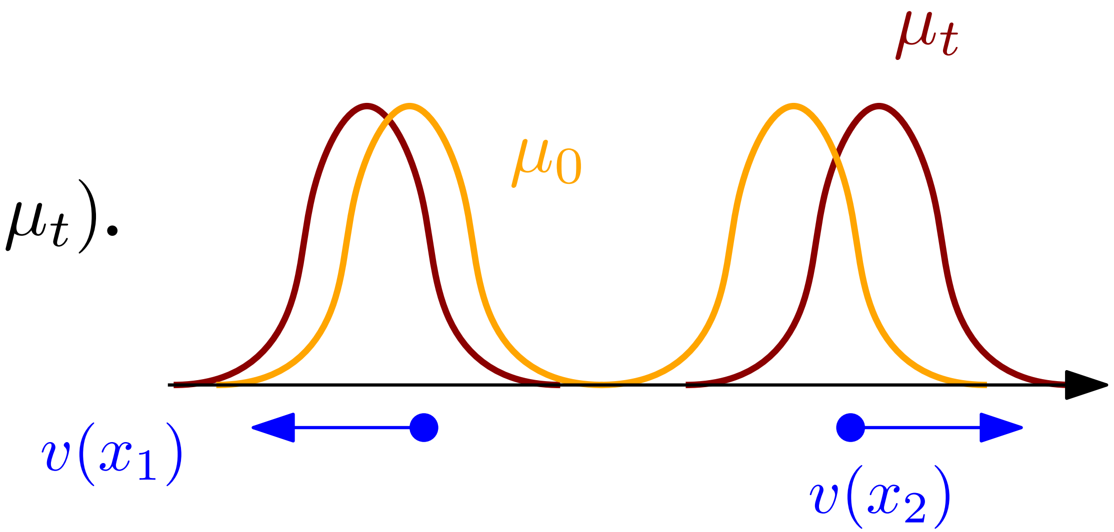


Ambrosio, Gigli & Savaré (2008). Gradient flows: in metric spaces and in the space of probability measures.

The linearization of optimal transport

On $\mathbb{R}^d$, what happens to $\mathrm{OT}(\mu,\nu)$ if $\mu \simeq \nu$?

$\rightsquigarrow$ $(\mu_t)_t$ curve in $\mathcal{P}(\mathbb{R}^d)$, we look at $\mathrm{OT}(\mu_0,\mu_t)$.



Theorem. $\mathrm{OT}(\mu_0,\mu_t) \sim t^2 \left(\min_v \int_{\mathbb{R}^d} |v(x)|^2 \, \mathrm{d}\mu_0(x) \right),$

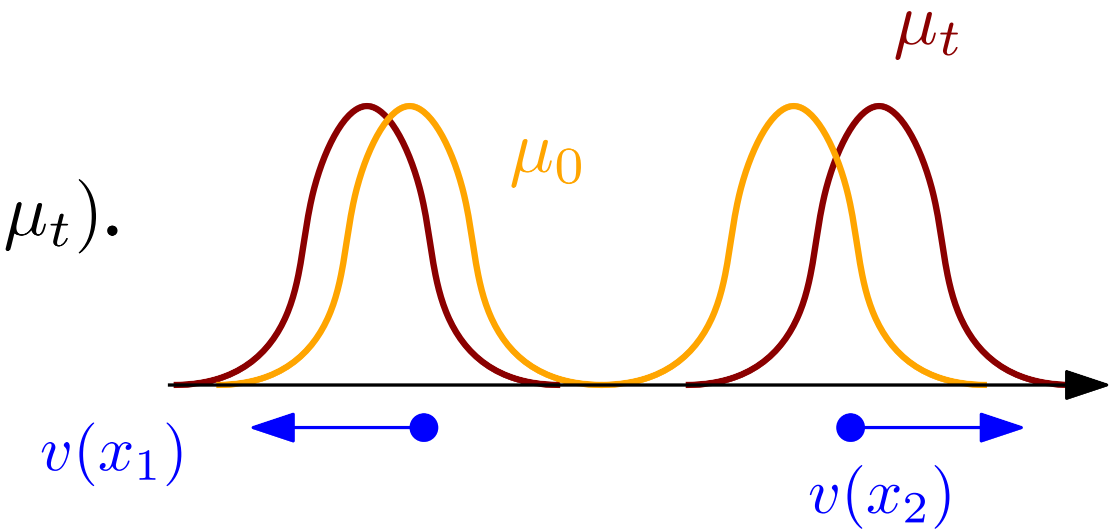
where $v : \mathbb{R}^d \to \mathbb{R}^d$ such that $\left.\dfrac{\partial \mu}{\partial t}\right|_{t=0} = -\mathrm{div}(\mu_0 v)$.

Ambrosio, Gigli & Savaré (2008). Gradient flows: in metric spaces and in the space of probability measures.

The linearization of optimal transport

On $\mathbb{R}^d$, what happens to $\mathrm{OT}(\mu, \nu)$ if $\mu \simeq \nu$?

$\rightsquigarrow$ $(\mu_t)_t$ curve in $\mathcal{P}(\mathbb{R}^d)$, we look at $\mathrm{OT}(\mu_0, \mu_t)$.



Theorem. $\mathrm{OT}(\mu_0, \mu_t) \sim t^2 \left(\min_v \int_{\mathbb{R}^d} |v(x)|^2 \, \mathrm{d}\mu_0(x) \right),$

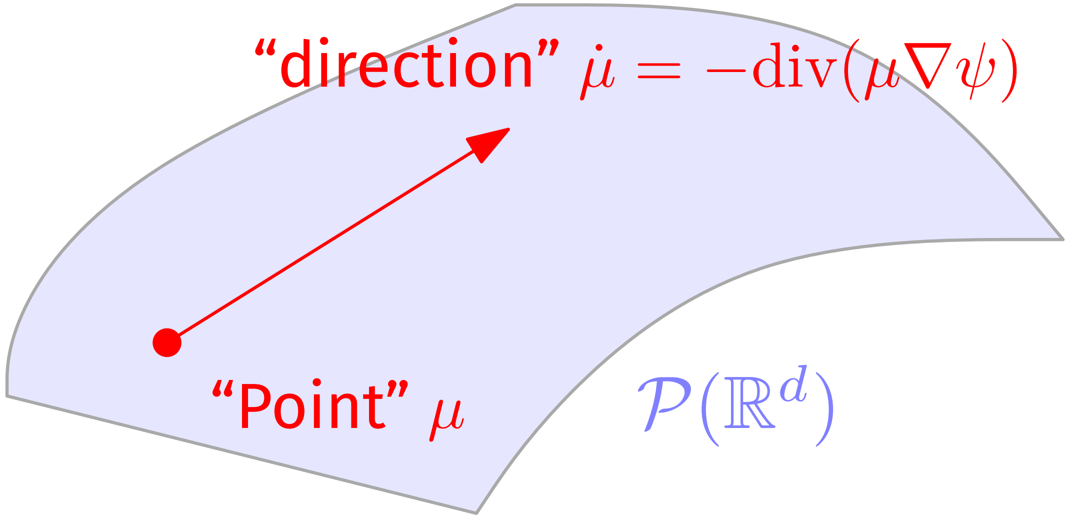
where $v : \mathbb{R}^d \to \mathbb{R}^d$ such that $\left.\frac{\partial \mu}{\partial t}\right|_{t=0} = -\mathrm{div}(\mu_0 v).$

elliptic equation in ψ

Optimal v is $\nabla\psi$, obtained by solving $-\mathrm{div}(\mu_0 \nabla \psi) = \dot{\mu}_0.$

Ambrosio, Gigli & Savaré (2008). Gradient flows: in metric spaces and in the space of probability measures.

The metric tensor and the geometry of optimal transport

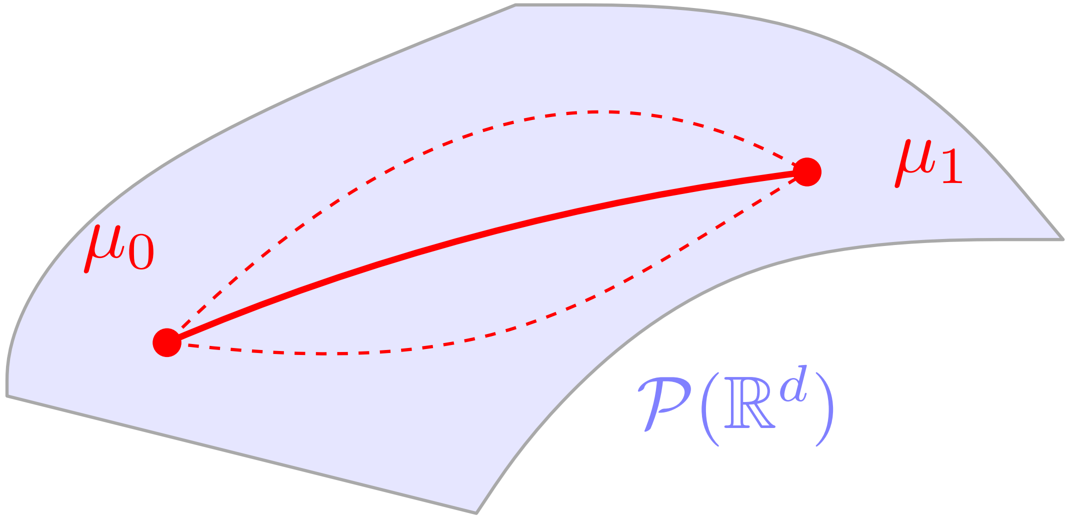


Metric tensor:

$$\mathbf{g}_\mu^{\mathrm{OT}}(\dot\mu, \dot\mu) = \int_X |\nabla \psi|^2 \, \mathrm{d}\mu.$$

Otto (2001). The geometry of dissipative evolution equations: the porous medium equation.

The metric tensor and the geometry of optimal transport



Metric tensor:

$$\mathbf{g}^{\mathrm{OT}}_{\mu}(\dot{\mu}, \dot{\mu}) = \int_X |\nabla \psi|^2 \,\mathrm{d}\mu.$$

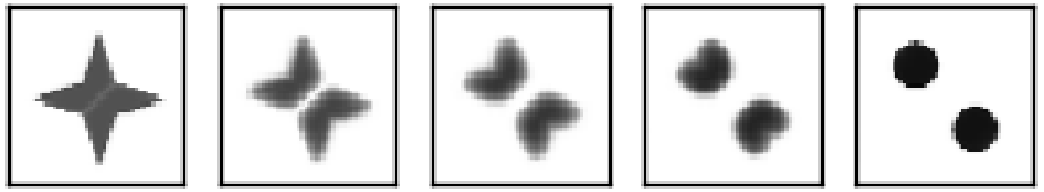
Theorem (Benamou and Brenier, 2000):

$$\mathrm{OT}(\mu_0, \mu_1) = \min_{(\mu_t)_t} \int_0^1 \mathbf{g}^{\mathrm{OT}}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)\,\mathrm{d}t$$

with μ_0, μ_1 fixed.

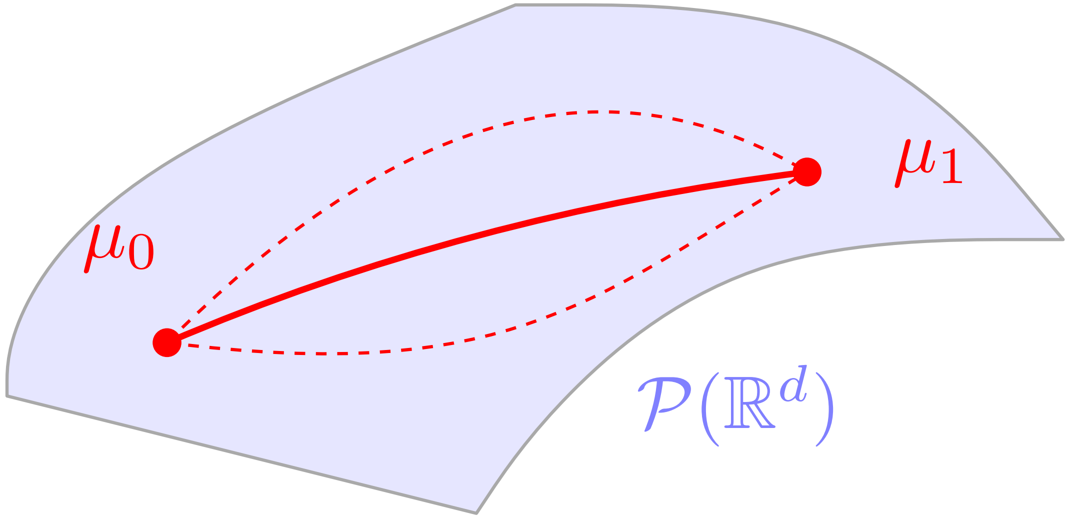
Minimizers are **geodesics**.

μ_0 μ_0



Example geodesic

The metric tensor and the geometry of optimal transport



Metric tensor:

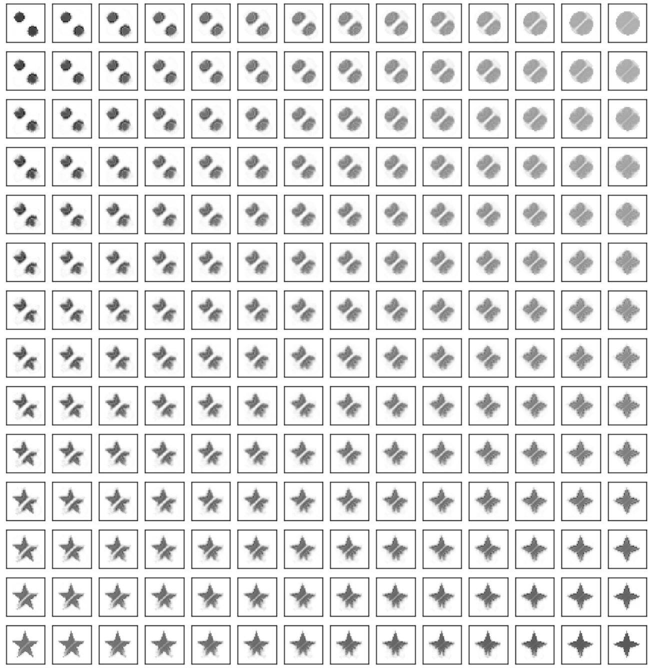
$$\mathbf{g}^{\mathrm{OT}}_{\mu}(\dot{\mu}, \dot{\mu}) = \int_X |\nabla \psi|^2 \,\mathrm{d}\mu.$$

Theorem (Benamou and Brenier, 2000):

$$\mathrm{OT}(\mu_0, \mu_1) = \min_{(\mu_t)_t} \int_0^1 \mathbf{g}^{\mathrm{OT}}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)\,\mathrm{d}t$$

with μ_0, μ_1 fixed.

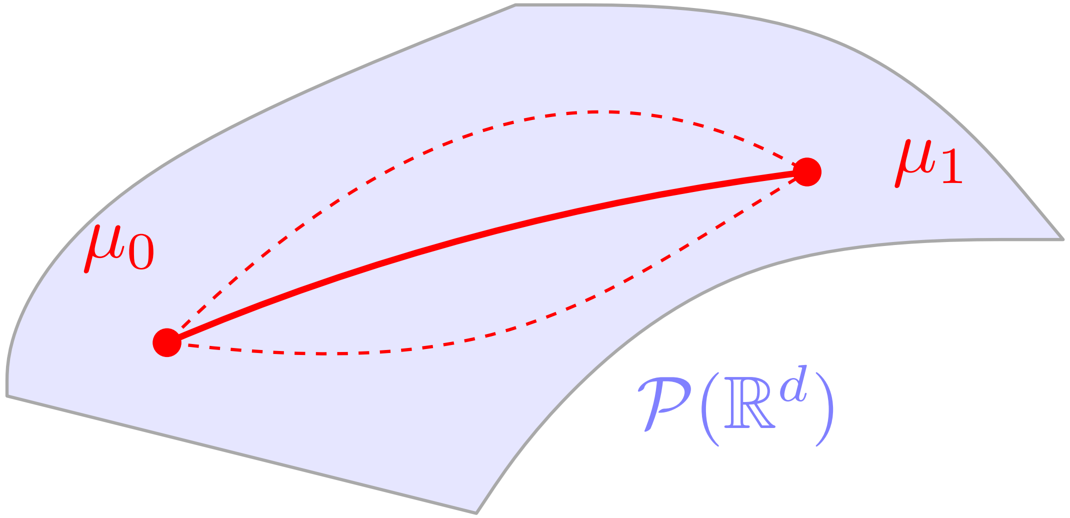
Minimizers are **geodesics**.



Example harmonic map

Lavenant (2019). Harmonic mappings valued in the Wasserstein space.

The metric tensor and the geometry of optimal transport



Metric tensor:

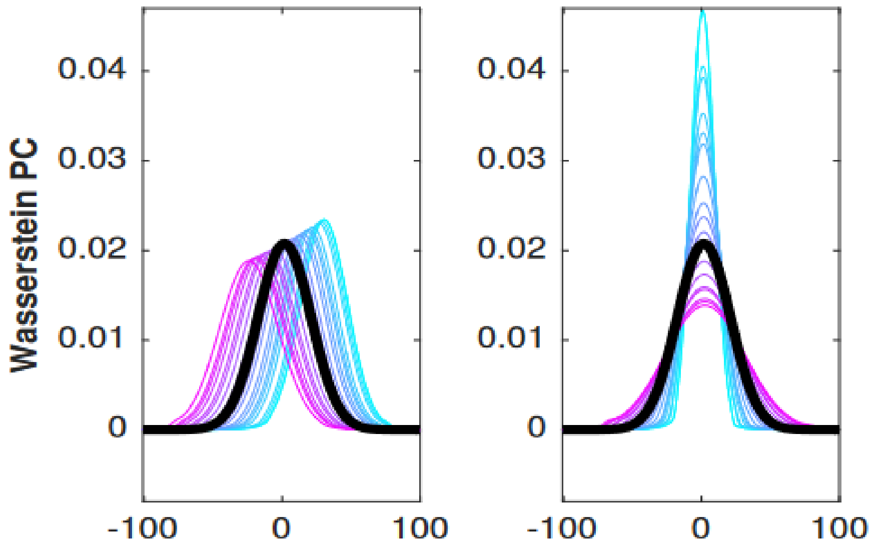
$$\mathbf{g}^{\mathrm{OT}}_{\mu}(\dot{\mu}, \dot{\mu}) = \int_X |\nabla \psi|^2 \,\mathrm{d}\mu.$$

Theorem (Benamou and Brenier, 2000):

$$\mathrm{OT}(\mu_0, \mu_1) = \min_{(\mu_t)_t} \int_0^1 \mathbf{g}^{\mathrm{OT}}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)\,\mathrm{d}t$$

with μ_0, μ_1 fixed.

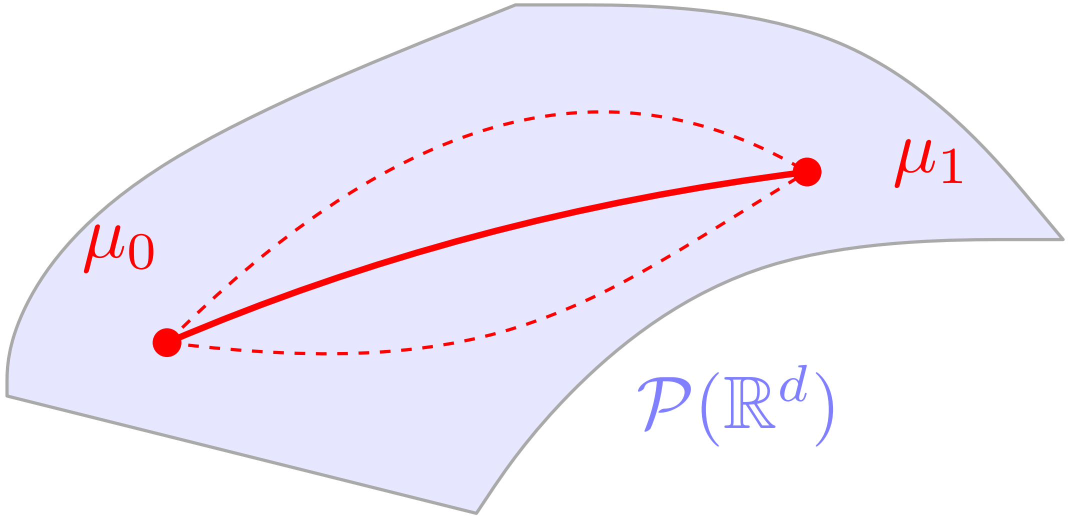
Minimizers are **geodesics**.



Example: Wasserstein PCA

Cazelles, Seguy, Bigot, Cuturi & Papadakis (2017). Log-PCA versus Geodesic PCA of histograms in the Wasserstein space.

The metric tensor and the geometry of optimal transport



Metric tensor:

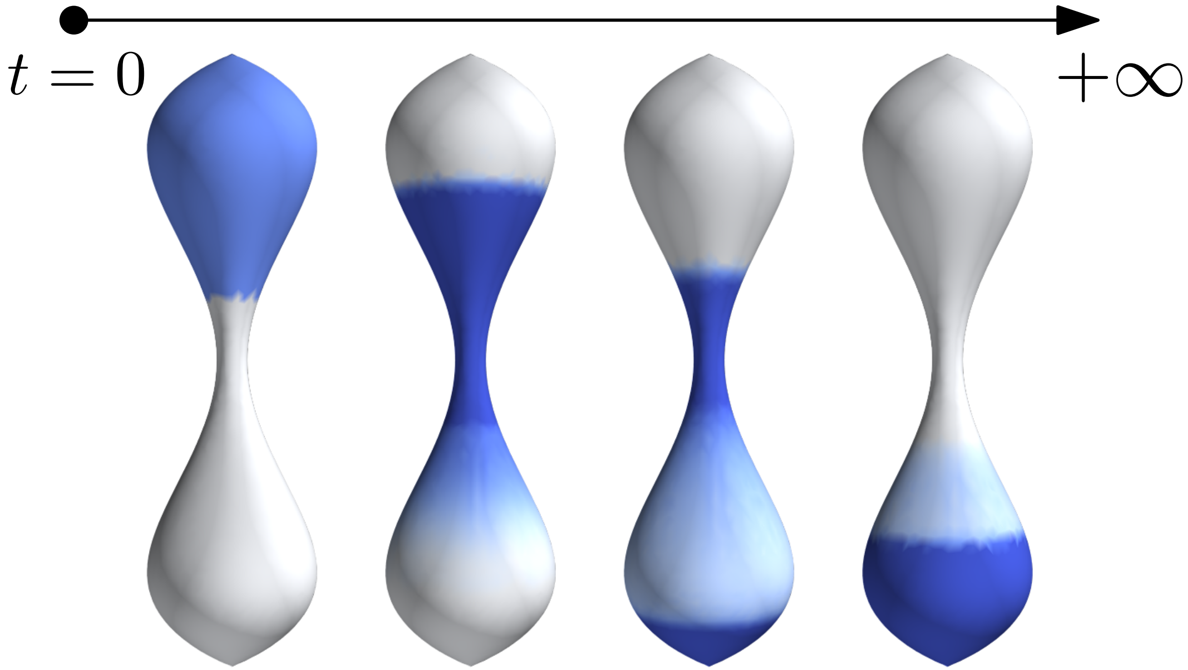
$$\mathbf{g}^{\mathrm{OT}}_{\mu}(\dot\mu, \dot\mu) = \int_X |\nabla \psi|^2 \, \mathrm{d}\mu.$$

Theorem (Benamou and Brenier, 2000):

$$\mathrm{OT}(\mu_0, \mu_1) = \min_{(\mu_t)_t} \int_0^1 \mathbf{g}^{\mathrm{OT}}_{\mu_t}(\dot\mu_t, \dot\mu_t)\, \mathrm{d}t$$

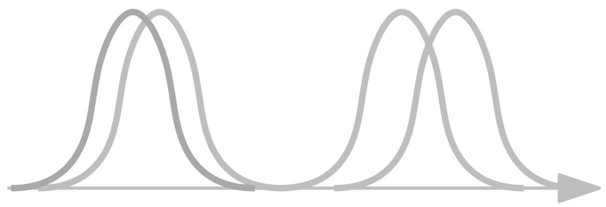
with μ_0, μ_1 fixed.

Minimizers are **geodesics**.



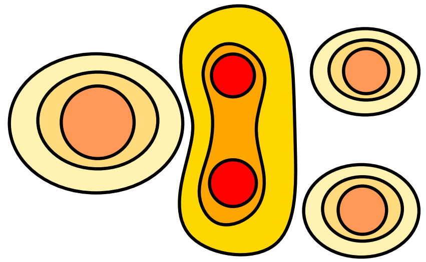
Example gradient flow

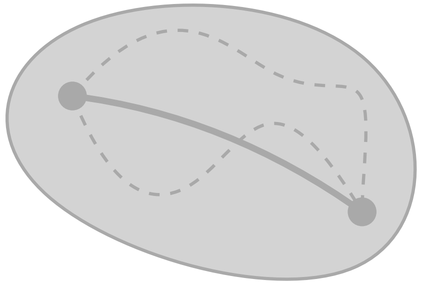
Jordan, Kinderlehrer & Otto (1998). The variational formulation of the Fokker–Planck equation.
Ambrosio, Gigli & Savaré (2008). Gradient flows: in metric spaces and in the space of probability measures.



1 - Optimal transport and its geometry

2 - Entropic optimal transport and Sinkhorn divergences





3 - Building a Riemannian geometry out of Sinkhorn divergences

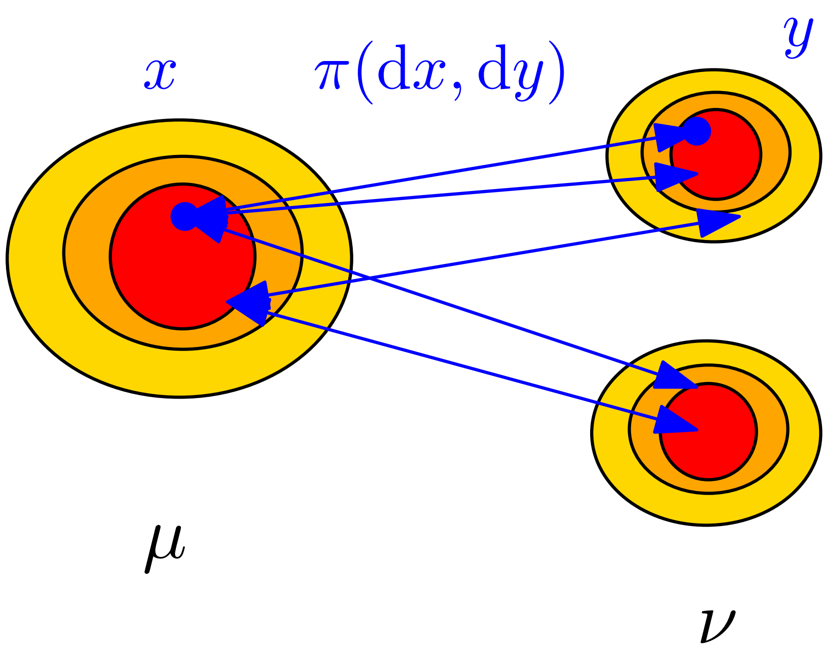
Entropic optimal transport

(X, d) compact metric space with symmetric cost function c, and $\varepsilon > 0$.

Definition

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} c(x,y)\,\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$

$$\mathrm{KL}(\alpha|\beta) = \int \log(\mathrm{d}\alpha/\mathrm{d}\beta)\mathrm{d}\alpha.$$



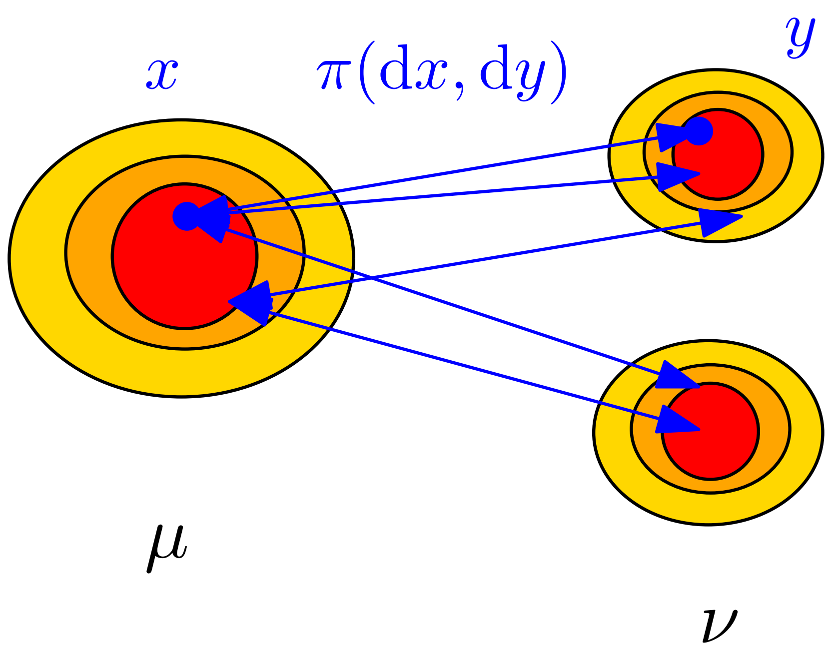
Peyré & Cuturi (2019). Computational optimal transport.

Entropic optimal transport

(X, d) compact metric space with symmetric cost function c, and $\varepsilon > 0$.

Definition

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} c(x,y)\,\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu)$$



Why?

1. easier to compute (**Sinkhorn algorithm**),
2. better statistical complexity,
3. smoother dependence in (μ,ν).

Peyré & Cuturi (2019). Computational optimal transport.

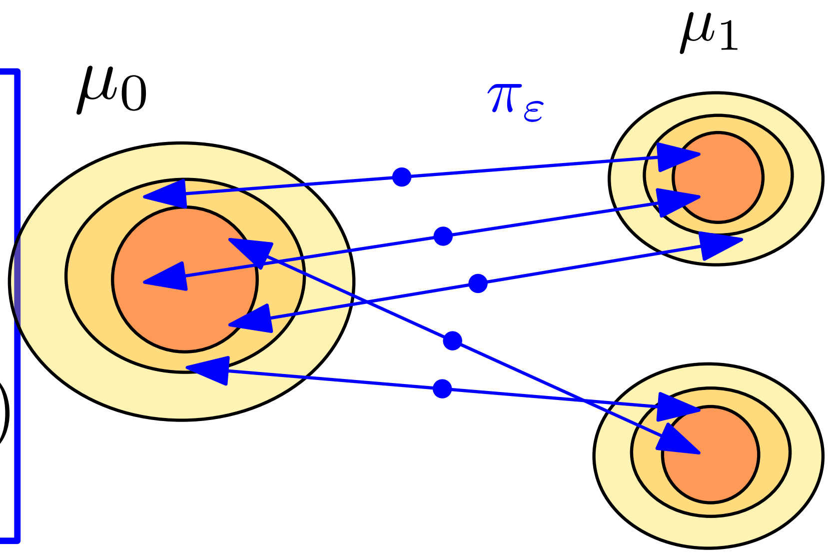
Schrödinger bridges as generalizations of geodesics?

Take c the quadratic cost on $\mathbb{R}^d$.

Select the entropic optimal coupling π_ε and define $(\mu_t)_t$ **Schrödinger bridge** between μ_0 and μ_1:

$$\mu_t = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{N}\left((1-t)x + ty, \frac{t(1-t)\varepsilon}{2} \right) \, \mathrm{d}\pi_\varepsilon(x,y)$$

Gaussian distribution



Léonard (2013). A survey of the Schrödinger problem and some of its connections with optimal transport.

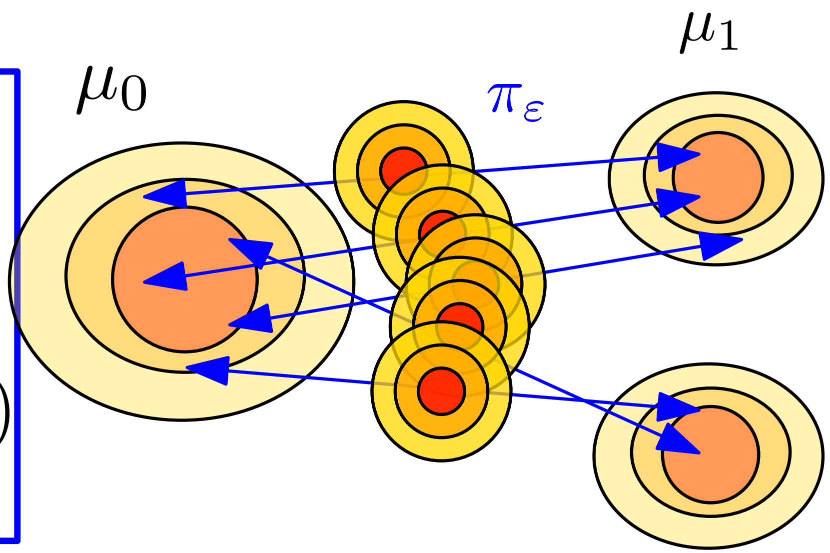
Schrödinger bridges as generalizations of geodesics?

Take c the quadratic cost on $\mathbb{R}^d$.

Select the entropic optimal coupling π_ε and define $(\mu_t)_t$ **Schrödinger bridge** between μ_0 and μ_1:

$$\mu_t = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{N}\left((1-t)x + ty, \frac{t(1-t)\varepsilon}{2} \right) \mathrm{d}\pi_\varepsilon(x,y)$$

Gaussian distribution



Léonard (2013). A survey of the Schrödinger problem and some of its connections with optimal transport.

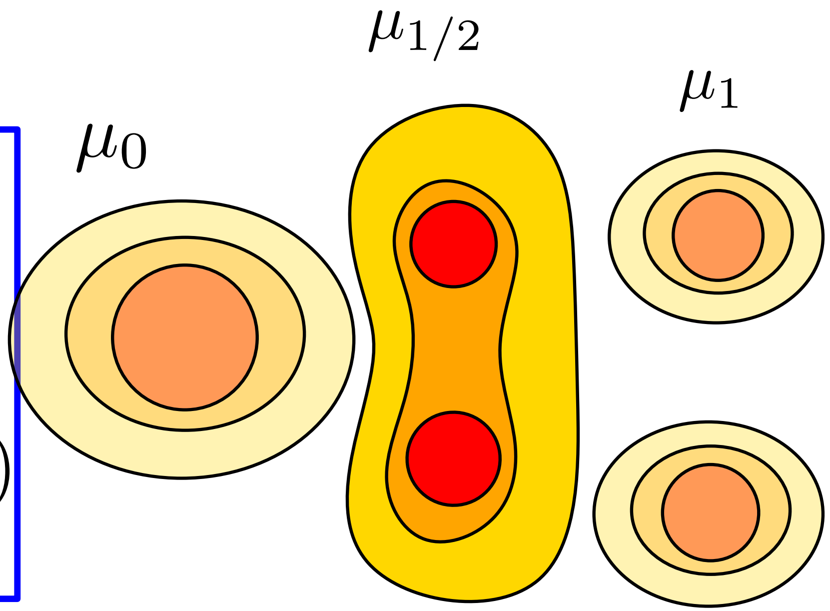
Schrödinger bridges as generalizations of geodesics?

Take c the quadratic cost on $\mathbb{R}^d$.

Select the entropic optimal coupling π_ε and define $(\mu_t)_t$ **Schrödinger bridge** between μ_0 and μ_1:

$$\mu_t = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{N}\left((1-t)x + ty, \frac{t(1-t)\varepsilon}{2}\right) \, \mathrm{d}\pi_\varepsilon(x, y)$$

Gaussian distribution



Léonard (2013). A survey of the Schrödinger problem and some of its connections with optimal transport.

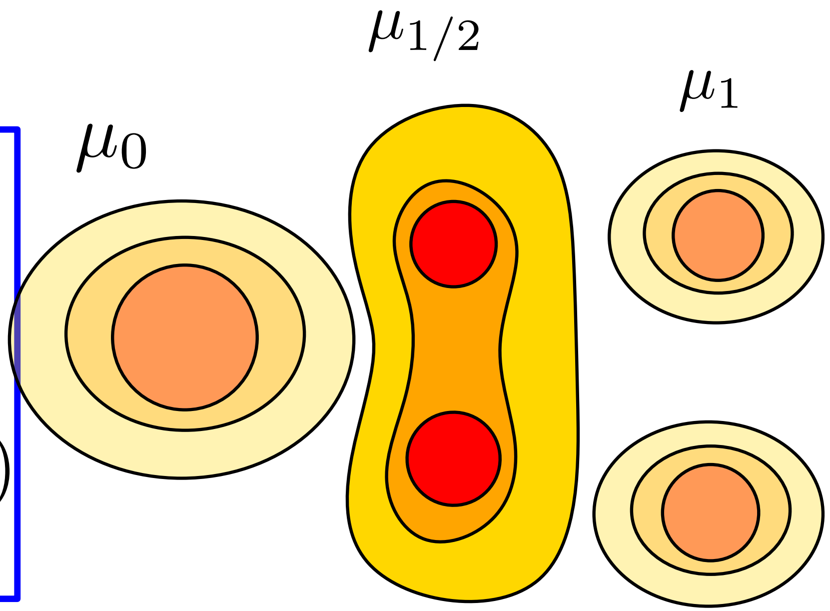
Schrödinger bridges as generalizations of geodesics?

Take c the quadratic cost on $\mathbb{R}^d$.

Select the entropic optimal coupling π_ε and define $(\mu_t)_t$ **Schrödinger bridge** between μ_0 and μ_1:

$$\mu_t = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{N}\left((1-t)x + ty, \frac{t(1-t)\varepsilon}{2} \right) \mathrm{d}\pi_\varepsilon(x,y)$$

Gaussian distribution



- Interpolates between μ_0 and μ_1, converges to OT geodesic as $\varepsilon \to 0$.

Léonard (2013). A survey of the Schrödinger problem and some of its connections with optimal transport.

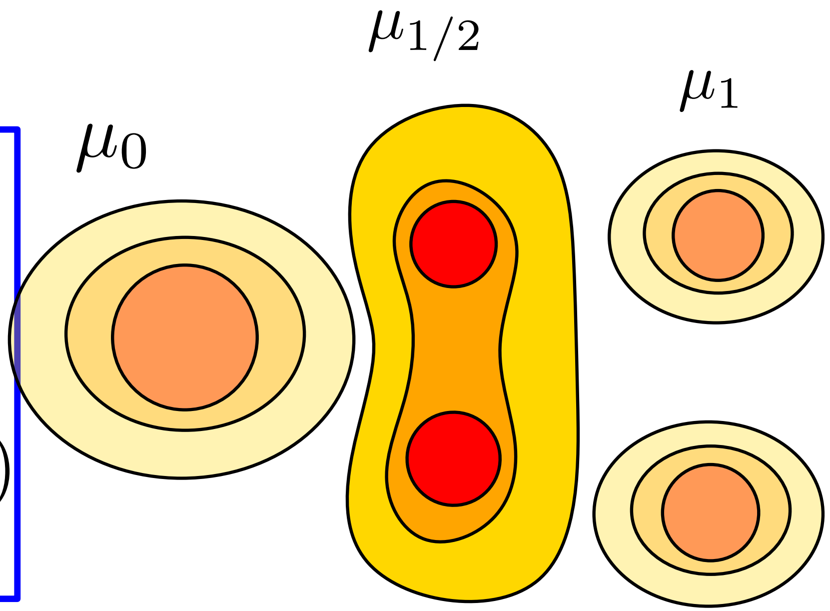
Schrödinger bridges as generalizations of geodesics?

Take c the quadratic cost on $\mathbb{R}^d$.

Select the entropic optimal coupling π_ε and define $(\mu_t)_t$ **Schrödinger bridge** between μ_0 and μ_1:

$$\mu_t = \iint_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{N}\left((1-t)x + ty, \frac{t(1-t)\varepsilon}{2}\right) \mathrm{d}\pi_\varepsilon(x,y)$$

Gaussian distribution



- Interpolates between μ_0 and μ_1, converges to OT geodesic as $\varepsilon \to 0$.
- **But** the bridge between μ and itself is **not** $\mu_t = \mu$ for all t.
- **But** the temporal rescaling of a ε-bridge by τ is a $\tau\varepsilon$-bridge.

Léonard (2013). A survey of the Schrödinger problem and some of its connections with optimal transport.

Sinkhorn divergence as a distance?

As $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$ generically, **debias** by defining

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

Genevay, Peyré, & Cuturi (2018). Learning generative models with Sinkhorn divergences.

Sinkhorn divergence as a distance?

As $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$ generically, **debias** by defining

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

Theorem (Feydy et al., 2019). Assume $\exp(-c/\varepsilon)$ positive definite universal kernel.

1. $S_\varepsilon(\mu,\nu) \geq 0$ with equality iff $\mu = \nu$.
2. $S_\varepsilon(\mu_n,\mu) \to 0$ iff $\mu_n \to \mu$ weakly.
3. S_ε convex in each of its inputs.

Assumption until the end of the talk

Genevay, Peyré, & Cuturi (2018). Learning generative models with Sinkhorn divergences.
Feydy, Séjourné, Vialard, Amari, Trouvé & Peyré (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences.

Sinkhorn divergence as a distance?

As $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$ generically, **debias** by defining

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

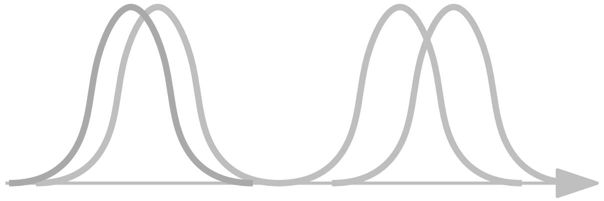
Theorem (Feydy et al., 2019). Assume $\exp(-c/\varepsilon)$ positive definite universal kernel.

1. $S_\varepsilon(\mu,\nu) \geq 0$ with equality iff $\mu = \nu$.
2. $S_\varepsilon(\mu_n,\mu) \to 0$ iff $\mu_n \to \mu$ weakly.
3. S_ε convex in each of its inputs.

Assumption until the end of the talk

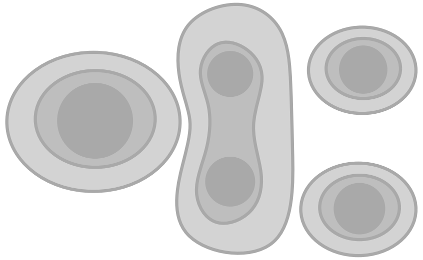
But $\sqrt{S_\varepsilon}$ does not satisfy the triangle inequality.

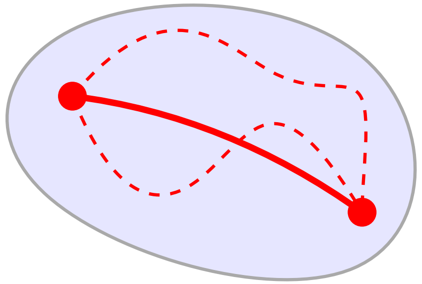
Genevay, Peyré, & Cuturi (2018). Learning generative models with Sinkhorn divergences.
Feydy, Séjourné, Vialard, Amari, Trouvé & Peyré (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences.



1 - Optimal transport and its geometry

2 - Entropic optimal transport and Sinkhorn divergences





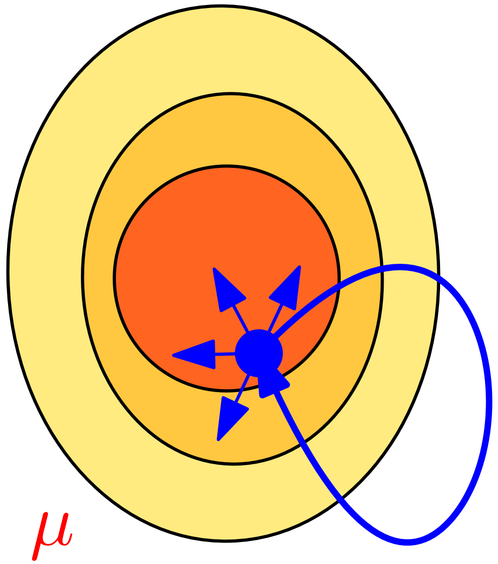
3 - Building a Riemannian geometry out of Sinkhorn divergences

1. Define $\mathbf{g}_\mu(\dot{\mu}, \dot{\mu})$ by $S_\varepsilon(\mu_0, \mu_t) \sim t^2 \mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)$.
2. Define $\mathrm{d}_S(\mu_0, \mu_1)^2 = \inf \int_0^1 \mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)\,\mathrm{d}t$.

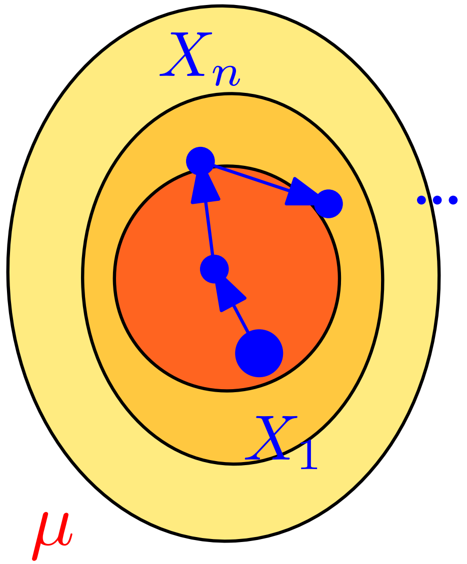
Understanding $\mathrm{OT}_\varepsilon(\mu, \mu)$

With $f_\mu : X \to \mathbb{R}$ Schrödinger potential, π_ε entropic optimal plan between μ and μ is:

$$\mathrm{d}\pi_\varepsilon(x, y) = \exp\left(\frac{f_\mu(x) + f_\mu(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\mu(y)$$



Understanding $\mathrm{OT}_\varepsilon(\mu, \mu)$



With $f_\mu : X \to \mathbb{R}$ Schrödinger potential, π_ε entropic optimal plan between μ and μ is:

$$\mathrm{d}\pi_\varepsilon(x, y) = \exp\left(\frac{f_\mu(x) + f_\mu(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\mu(y)$$

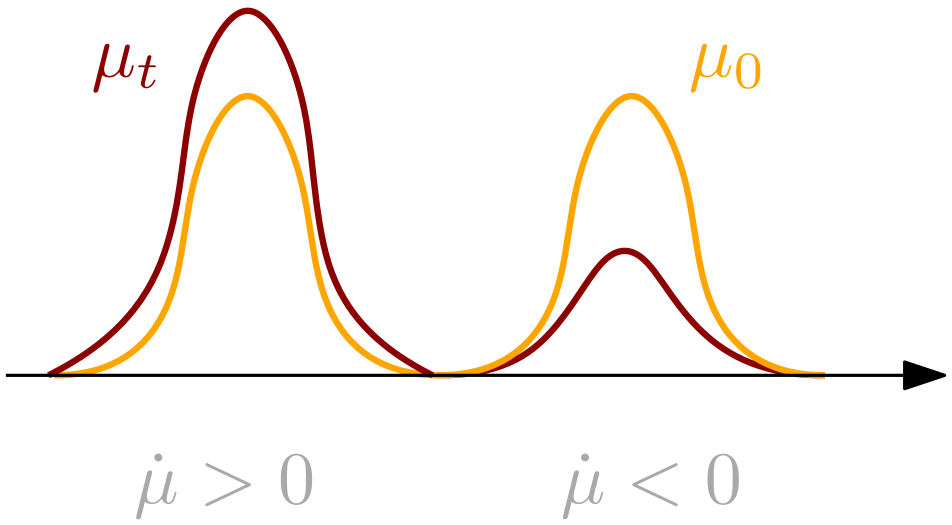
Interpretation. Take $X_1, \ldots, X_n, \ldots$ Markov chain with $(X_n, X_{n+1}) \sim \pi_\varepsilon$. Then:

1. Invariant distribution μ,
2. Reversible Markov chain,
3. Transition probability close to $\exp(-c(x, y)/\varepsilon)$.

Gaussian kernel if c quadratic

The Hessian of the Sinkhorn divergence

$\mu_t = \mu + t\dot{\mu}$, with $\dot{\mu}$ signed measure with zero mass.

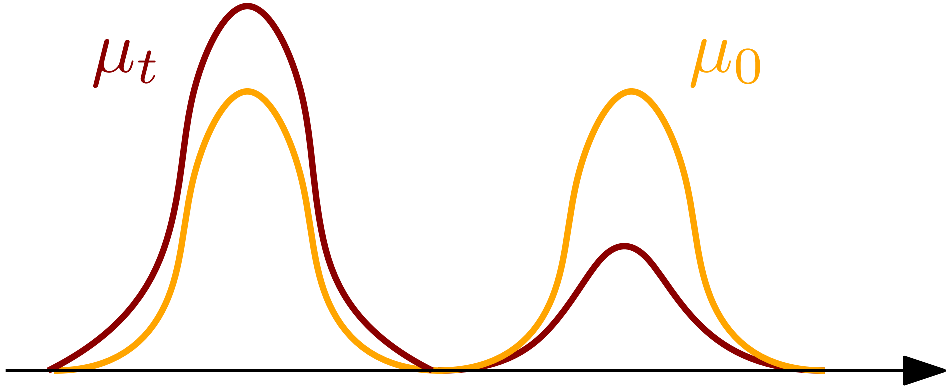


The Hessian of the Sinkhorn divergence

$\mu_t = \mu + t\dot{\mu}$, with $\dot{\mu}$ signed measure with zero mass.

Theorem.

$$S_\varepsilon(\mu_0, \mu_t) \sim t^2 \frac{\varepsilon}{2} \langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \rangle.$$



Where $k_\mu(x, y) = \exp((f_\mu(x) + f_\mu(y) - c(x, y))/\varepsilon)$ and:

$$K_\mu(\phi)(x) = \int_X k_\mu(x, y)\phi(y)\, \mathrm{d}\mu(y),$$

$$H_\mu[\sigma](x) = \int_X k_\mu(x, y)\, \mathrm{d}\sigma(y).$$

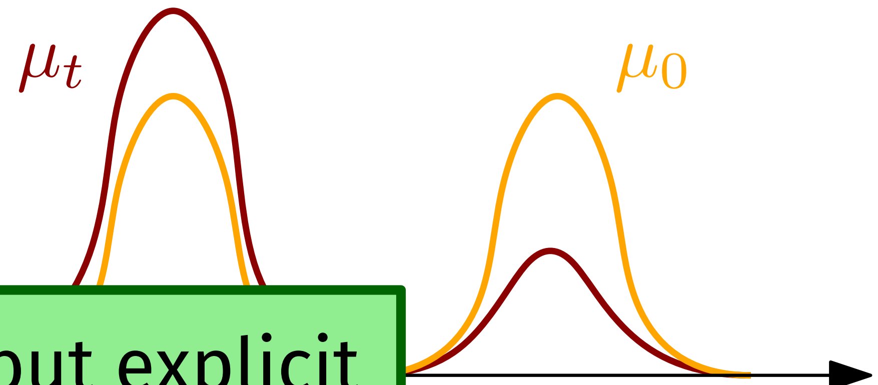
$(\mathrm{Id} - K_\mu^2)/\varepsilon \sim$ Laplacian

The Hessian of the Sinkhorn divergence

$\mu_t = \mu + t\dot{\mu}$, with $\dot{\mu}$ signed measure with zero mass.

Theorem.

$$S_\varepsilon(\mu_0, \mu_t) \sim t^2 \frac{\varepsilon}{2} \langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \rangle.$$



Main message: heavy but explicit and interpretable formula!

Where $k_\mu(x, y) = \exp((f_\mu(x$

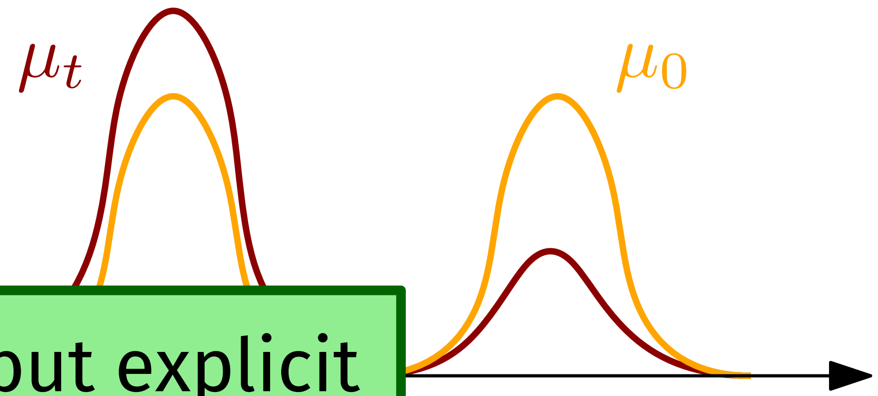
$$K_\mu(\phi)(x) = \int_X k_\mu(x,y)\phi(y)\,\mathrm{d}\mu(y),$$

$$H_\mu[\sigma](x) = \int_X k_\mu(x,y)\,\mathrm{d}\sigma(y).$$

$(\mathrm{Id} - K_\mu^2)/\varepsilon \sim$ Laplacian

The Hessian of the Sinkhorn divergence

$\mu_t = \mu + t\dot{\mu}$, with $\dot{\mu}$ signed measure with zero mass.



Theorem.

$$S_\varepsilon(\mu_0, \mu_t) \sim t^2 \frac{\varepsilon}{2} \langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \rangle.$$

Main message: heavy but explicit and interpretable formula!

Where $k_\mu(x, y) = \exp((f_\mu(x$

$$K_\mu(\phi)(x) = \int_X k_\mu(x, y)\phi(y)\,\mathrm{d}\mu(y),$$

$$H_\mu[\sigma](x) = \int_X k_\mu(x, y)\,\mathrm{d}\sigma(y).$$

$(\mathrm{Id} - K_\mu^2)/\varepsilon \sim$ Laplacian

Same formula

Definition. $\mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) = \frac{\varepsilon}{2} \langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \rangle.$

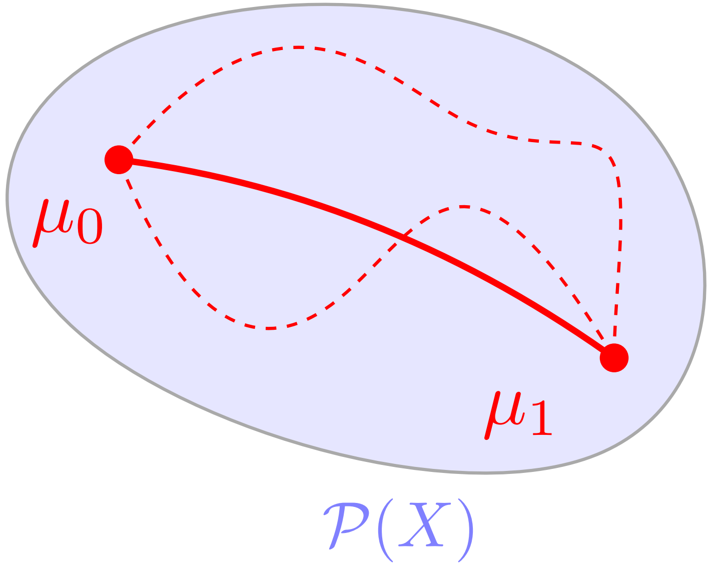
Definition of the distance and main results

Recall $\mathbf{g}_\mu(\dot\mu, \dot\mu) = \frac{\varepsilon}{2}\langle \dot\mu, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot\mu]\rangle.$

Definition. Given μ_0, μ_1:

$$\mathrm{d}_S(\mu_0, \mu_1)^2 = \inf \int_0^1 \mathbf{g}_\mu(\dot\mu_t, \dot\mu_t)\,\mathrm{d}t$$

where infimum over (μ_t) on a class of path to be specified later.



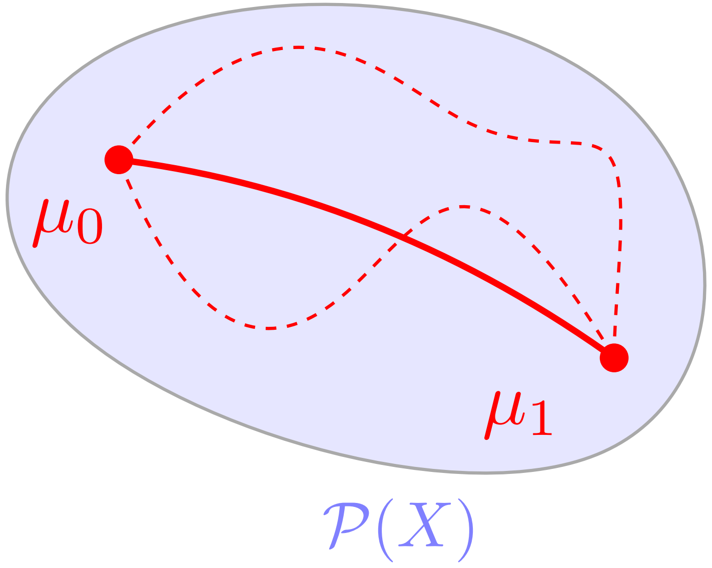
Definition of the distance and main results

Recall $\mathbf{g}_\mu(\dot\mu, \dot\mu) = \frac{\varepsilon}{2}\langle \dot\mu, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot\mu]\rangle.$

Definition. Given μ_0, μ_1:

$$\mathrm{d}_S(\mu_0, \mu_1)^2 = \inf \int_0^1 \mathbf{g}_\mu(\dot\mu_t, \dot\mu_t)\,\mathrm{d}t$$

where infimum over (μ_t) on a class of path to be specified later.



Theorem. d_S is a distance over $\mathcal{P}(X)$ **metrizing weak convergence of measures,** and the infimum in the definition is reached (**geodesics exist**).

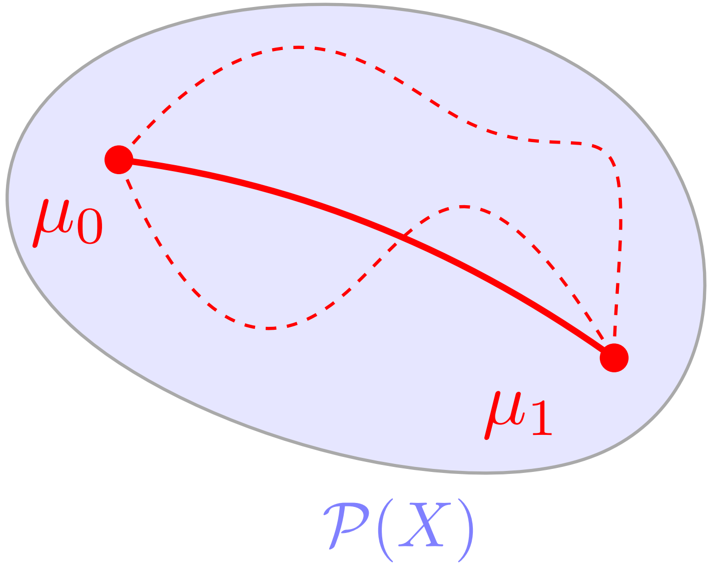
Definition of the distance and main results

Recall $\mathbf{g}_\mu(\dot\mu,\dot\mu) = \frac{\varepsilon}{2}\langle \dot\mu, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot\mu]\rangle.$

Definition. Given μ_0, μ_1:

$$\mathrm{d}_S(\mu_0,\mu_1)^2 = \inf \int_0^1 \mathbf{g}_\mu(\dot\mu_t,\dot\mu_t)\,\mathrm{d}t$$

where infimum over (μ_t) on a class of path to be specified later.



Theorem. d_S is a distance over $\mathcal{P}(X)$ **metrizing weak convergence of measures**, and the infimum in the definition is reached (**geodesics exist**).

Next slides: elements of the proof (and of functional analysis!).

Reminder on Reproducing Kernel Hilbert Spaces (RKHS)

Fix $k : X \times X \to \mathbb{R}$ positive definite.

Definition. $\mathcal{H}_k$ Hilbert space of functions $X \to \mathbb{R}$: start with

$$\operatorname{span}\{k(\cdot, x) \ : \ x \in X\}$$

with $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k} = k(x, y)$. Then take completion.

k positive definite if this defines dot product

(k universal $\Leftrightarrow \mathcal{H}_k$ dense in $C(X)$)

Paulsen & Raghupathi (2016). An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.

Reminder on Reproducing Kernel Hilbert Spaces (RKHS)

Fix $k : X \times X \to \mathbb{R}$ positive definite.

Definition. $\mathcal{H}_k$ Hilbert space of functions $X \to \mathbb{R}$: start with

$$\operatorname{span}\{k(\cdot, x) \;:\; x \in X\}$$

with $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k} = k(x, y)$. Then take completion.

Remark. $\mathcal{H}_k$ Hilbert space of functions on X such that $\phi \mapsto \phi(x)$ is continuous for any x, and this characterizes a RKHS.

Paulsen & Raghupathi (2016). An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.

Reminder on Reproducing Kernel Hilbert Spaces (RKHS)

Fix $k : X \times X \to \mathbb{R}$ positive definite.

Definition. $\mathcal{H}_k$ Hilbert space of functions $X \to \mathbb{R}$: start with

$$\operatorname{span}\{k(\cdot, x) \ : \ x \in X\}$$

with $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k} = k(x, y)$. Then take completion.

Remark. $\mathcal{H}_k$ Hilbert space of functions on X such that $\phi \mapsto \phi(x)$ is continuous for any x, and this characterizes a RKHS.

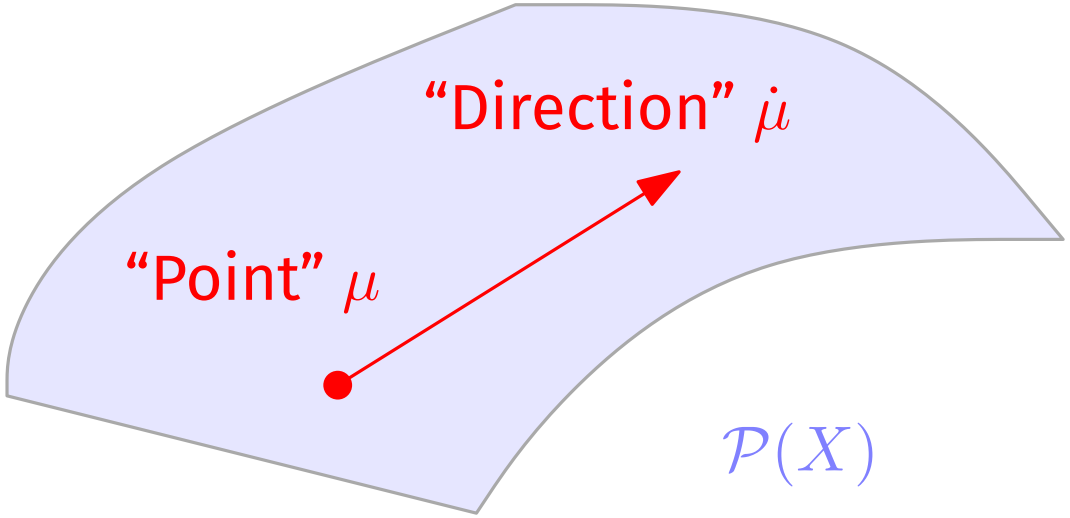
In our case:

Typically smooth functions!

- $k = \exp(-c/\varepsilon)$, space $\mathcal{H}_c$.
- $k = k_\mu = \exp((f_\mu \oplus f_\mu - c)/\varepsilon)$, space $\mathcal{H}_\mu$.

Paulsen & Raghupathi (2016). An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.

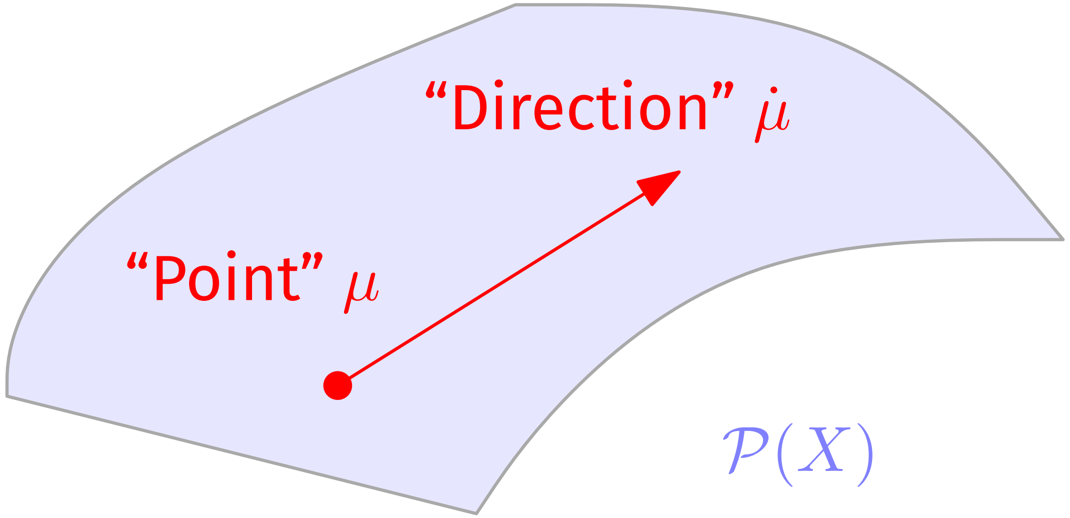
The tangent space



Recall:

- $\mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) = \frac{\varepsilon}{2} \langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \rangle$
  quadratic form in $\dot{\mu}$
- $\mathcal{H}_\mu$ RKHS with kernel k_μ.

The tangent space



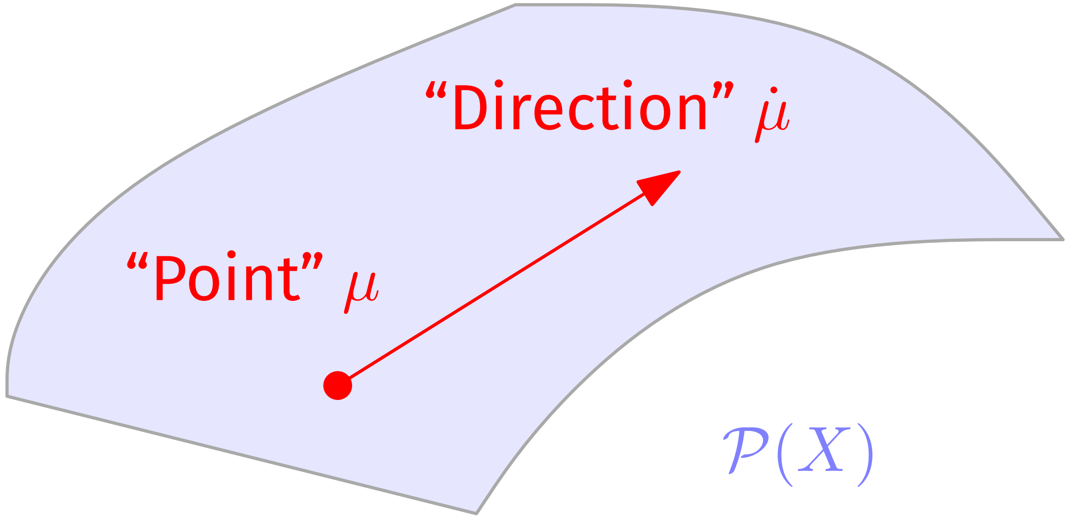
Recall:

- $\mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) = \frac{\varepsilon}{2}\langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}]\rangle$
  quadratic form in $\dot{\mu}$
- $\mathcal{H}_\mu$ RKHS with kernel k_μ.

Theorem. The completion of signed measures with zero mass with respect to $\mathbf{g}_\mu$ is $\mathcal{H}^*_{\mu,0}$ the space of linear forms σ on $\mathcal{H}_\mu$ with $\langle \sigma, 1\rangle = 0$.

That is, we want $\left| \frac{\mathrm{d}}{\mathrm{d}t} \int \phi \, \mathrm{d}\mu_t \right| \leq C \|\phi\|_{\mathcal{H}_\mu}$ for any $\phi \in \mathcal{H}_\mu$.

The tangent space



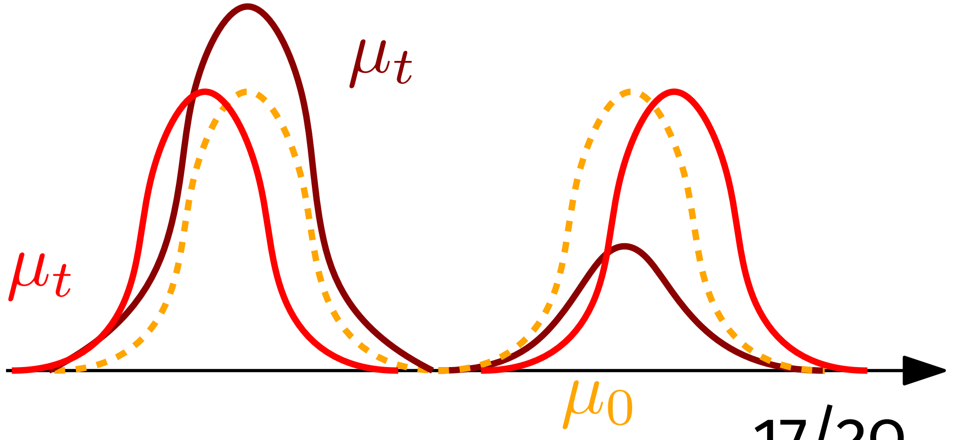
Recall:

- $\mathbf{g}_\mu(\dot\mu, \dot\mu) = \frac{\varepsilon}{2}\langle \dot\mu, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot\mu]\rangle$

  quadratic form in $\dot\mu$
- $\mathcal{H}_\mu$ RKHS with kernel k_μ.

Theorem. The completion of signed measures with zero mass with respect to $\mathbf{g}_\mu$ is $\mathcal{H}^*_{\mu,0}$ the space of linear forms σ on $\mathcal{H}_\mu$ with $\langle \sigma, 1\rangle = 0$.

If c quadratic cost, both $\dot\mu$ signed measure (“**vertical**”) and $\dot\mu = -\mathrm{div}(\mu v)$ (“**horizontal**”) are in the tangent space $\mathcal{H}^*_{\mu,0}$.



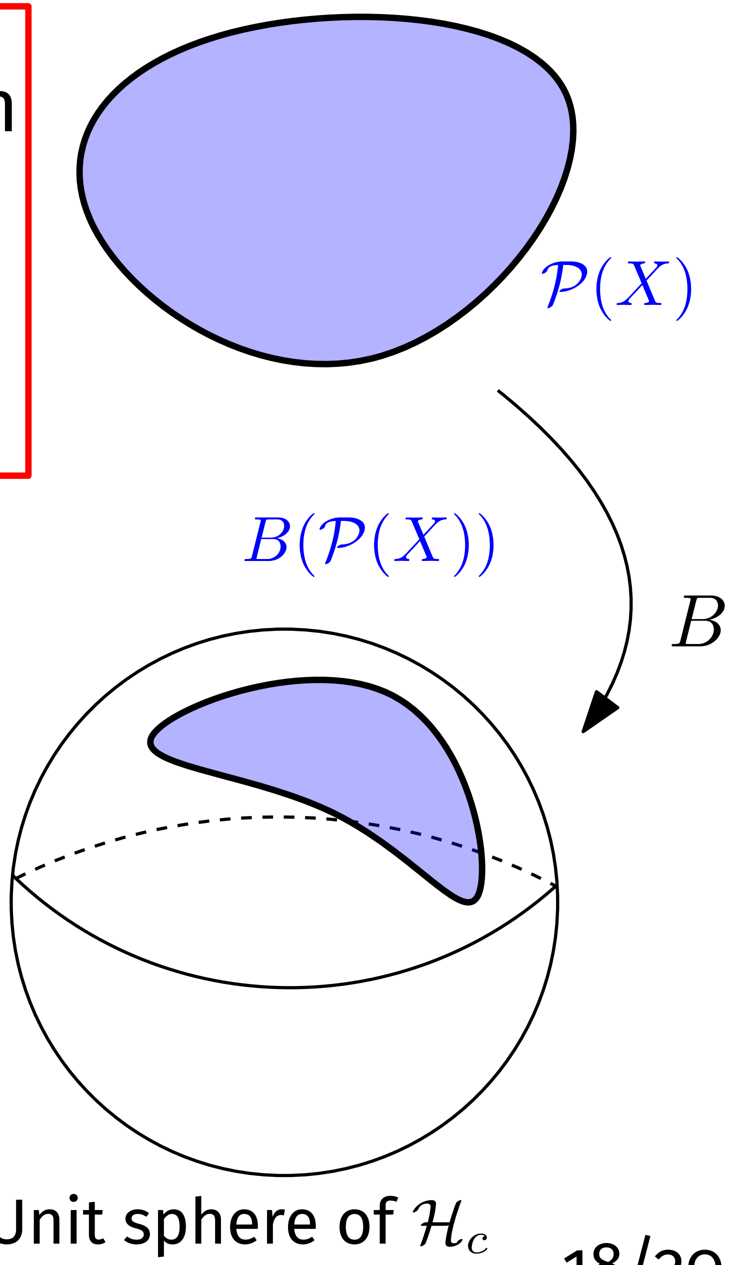
A useful change of variable

Define:

$$\beta = B(\mu) = \exp\left(-\frac{f_\mu}{\varepsilon}\right)$$

where $f_\mu : X \to \mathbb{R}$ self Schrödinger potential.

Theorem. The map B is an homeomorphism onto its image, included in unit sphere of $\mathcal{H}_c$.



(Change of variable suggested by Feydy et al, Séjourné et al)

Feydy, Séjourné, Vialard, Amari, Trouvé & Peyré (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences.
Séjourné, Feydy Vialard, Trouvé & Peyré (2019). Sinkhorn divergences for unbalanced optimal transport.

A useful change of variable

Define:

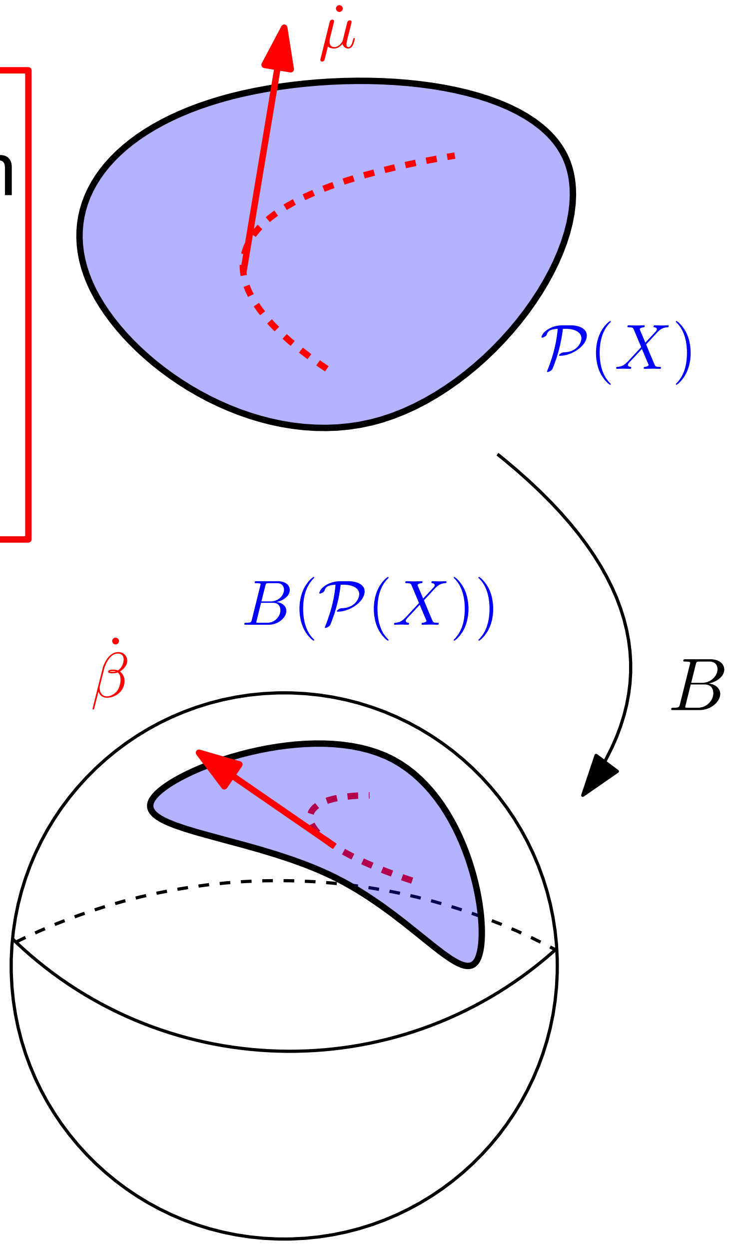
$$\beta = B(\mu) = \exp\left(-\frac{f_\mu}{\varepsilon}\right)$$

where $f_\mu : X \to \mathbb{R}$ self Schrödinger potential.

Theorem. The map B is an homeomorphism onto its image, included in unit sphere of $\mathcal{H}_c$.

Theorem. We have $\mathbf{g}_{\mu_t}(\dot\mu_t, \dot\mu_t) = \tilde{\mathbf{g}}_{\mu_t}(\dot\beta_t, \dot\beta_t)$ and:

- $(\mu, \dot\beta) \mapsto \tilde{\mathbf{g}}_\mu(\dot\beta, \dot\beta)$ jointly continuous,
- $\tilde{\mathbf{g}}_\mu(\dot\beta, \dot\beta) \asymp \|\dot\beta\|^2_{\mathcal{H}_c}$ uniformly in μ (but not in ε).



Unit sphere of $\mathcal{H}_c$

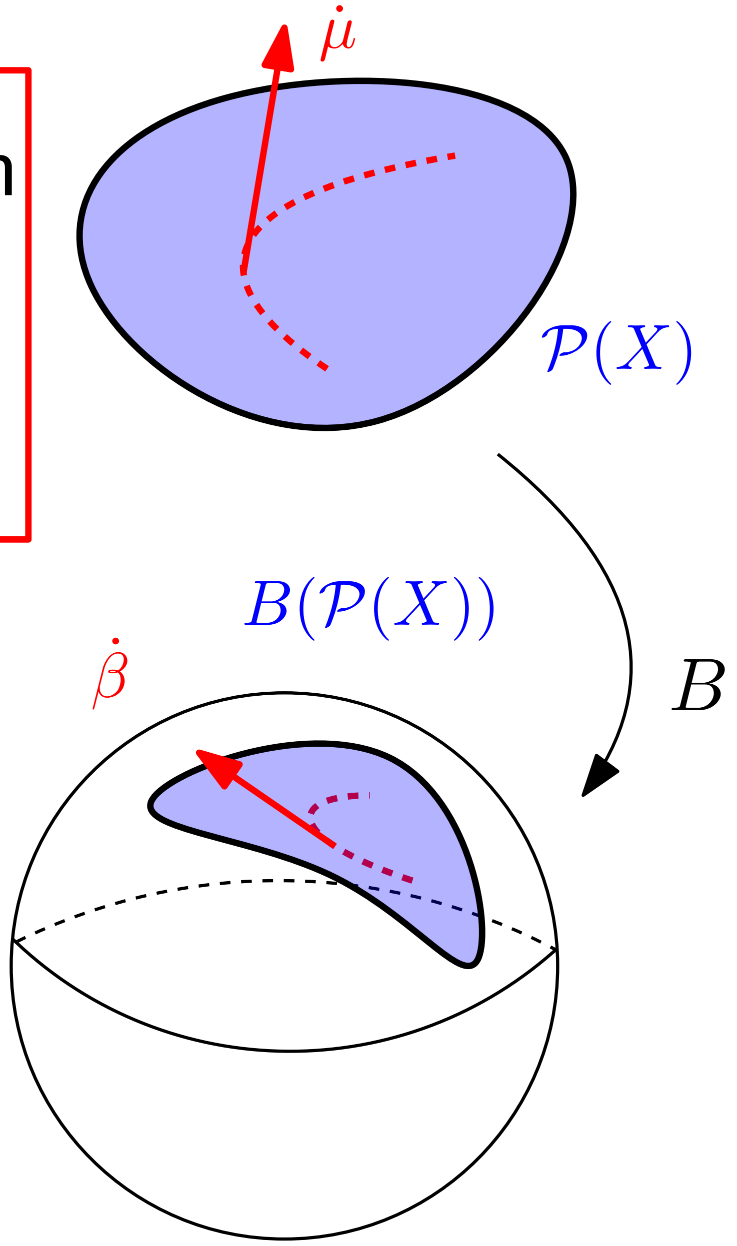
A useful change of variable

Define:

$$\beta = B(\mu) = \exp\left(-\frac{f_\mu}{\varepsilon}\right)$$

where $f_\mu : X \to \mathbb{R}$ self Schrödinger potential.

Theorem. The map B is an homeomorphism onto its image, included in unit sphere of $\mathcal{H}_c$.



Theorem. We have $\mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t) = \tilde{\mathbf{g}}_{\mu_t}(\dot{\beta}_t, \dot{\beta}_t)$ and:

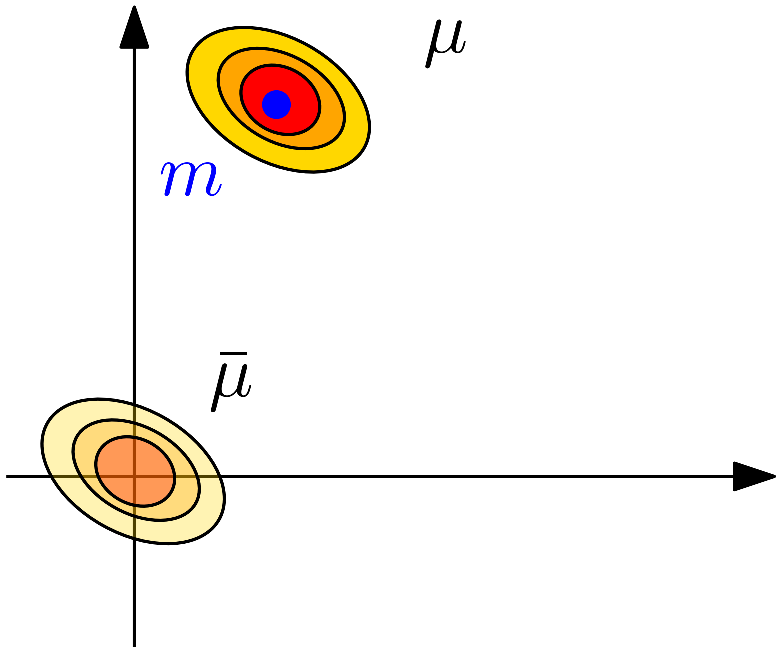
- $(\mu, \dot{\beta}) \mapsto \tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta})$ jointly continuous,
- $\tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta}) \asymp \|\dot{\beta}\|^2_{\mathcal{H}_c}$ uniformly in μ (but not in ε).

Consequence. Admissible paths: $(\beta_t)_t$ H^1 valued in $\mathcal{H}_c$,

$$c_\varepsilon \|\beta_1 - \beta_0\|_{\mathcal{H}_c} \leq \mathbf{d}_S(\mu_0, \mu_1) \leq C_\varepsilon \|\beta_1 - \beta_0\|_{\mathcal{H}_c}.$$

The quadratic case

Previous results hold for **any** compact space X if $\exp(-c/\varepsilon)$ positive definite universal kernel.

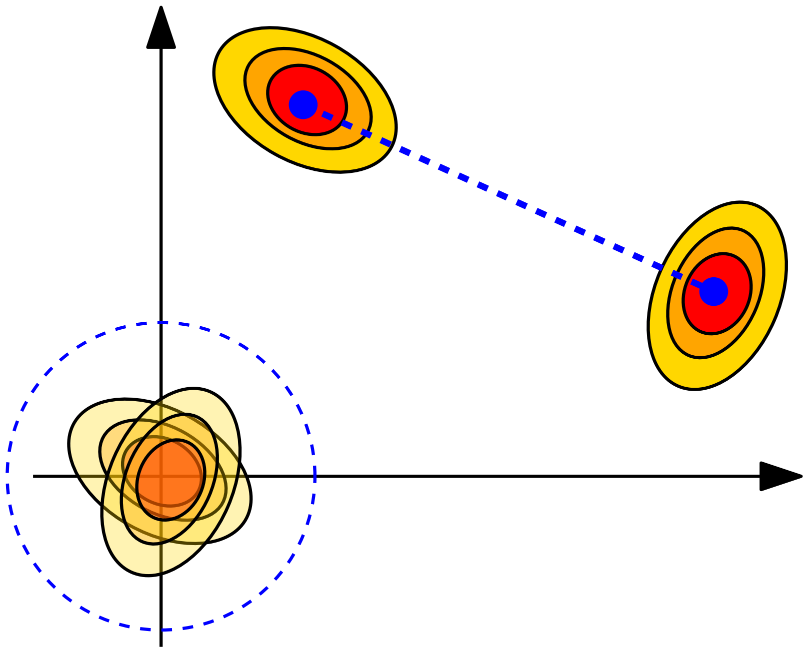


Now $X \subset \mathbb{R}^d$ and $c(x, y) = |x - y|^2$.

If $\mu \in \mathcal{P}(X)$, m barycenter and $\bar{\mu}$ centered part.

The quadratic case

Previous results hold for **any** compact space X if $\exp(-c/\varepsilon)$ positive definite universal kernel.



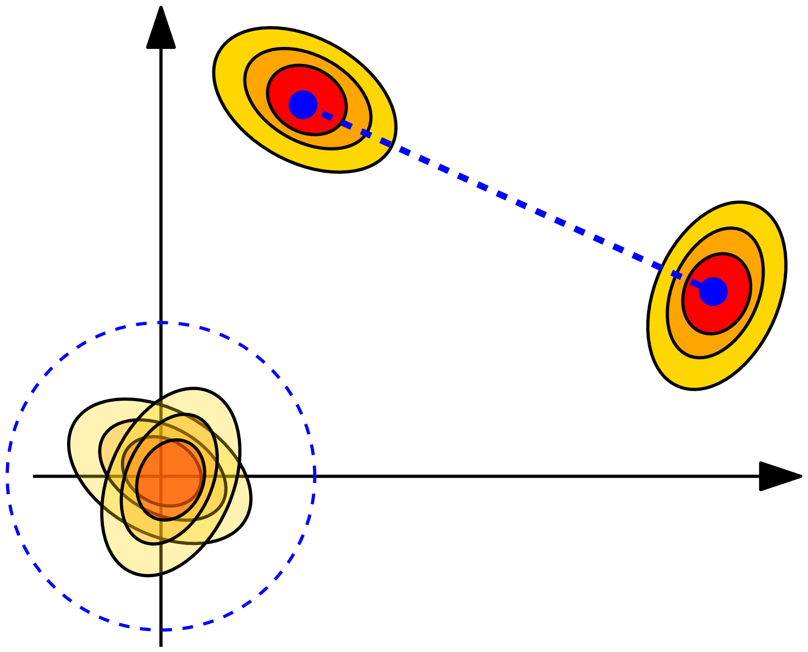
Now $X \subset \mathbb{R}^d$ and $c(x, y) = |x - y|^2$.

If $\mu \in \mathcal{P}(X)$, m barycenter and $\bar{\mu}$ centered part.

Reminder: $\mathrm{OT}(\mu_0, \mu_1) = |m_1 - m_0|^2 + \mathrm{OT}(\bar{\mu}_0, \bar{\mu}_1)$.

The quadratic case

Previous results hold for **any** compact space X if $\exp(-c/\varepsilon)$ positive definite universal kernel.



Now $X \subset \mathbb{R}^d$ and $c(x,y) = |x-y|^2$.

If $\mu \in \mathcal{P}(X)$, m barycenter and $\bar{\mu}$ centered part.

Reminder: $\mathrm{OT}(\mu_0, \mu_1) = |m_1 - m_0|^2 + \mathrm{OT}(\bar{\mu}_0, \bar{\mu}_1)$.

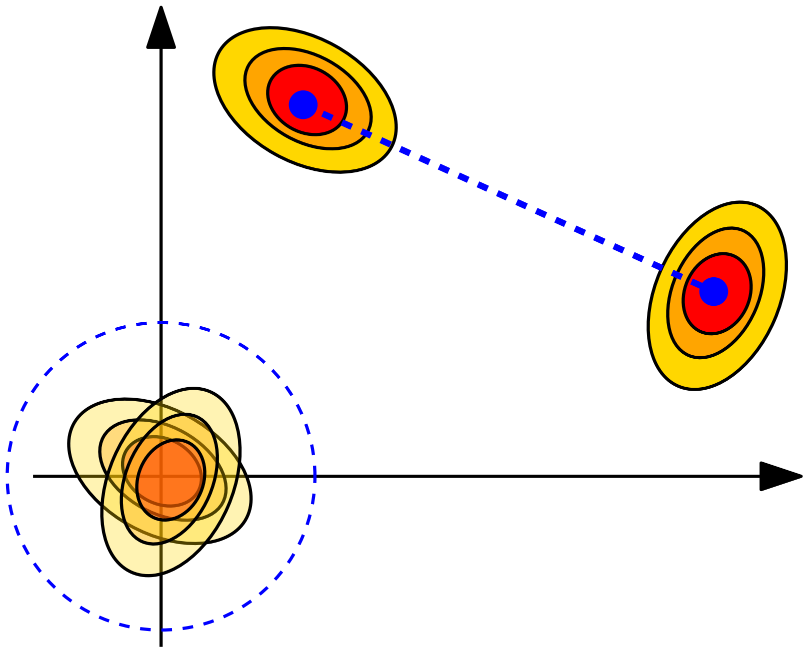
Theorem. $\mathsf{d}_S(\mu_0, \mu_1)^2 = |m_1 - m_0|^2 + \mathsf{d}_S(\bar{\mu}_0, \bar{\mu}_1)^2$.

The quadratic case

Previous results hold for **any** compact space X if $\exp(-c/\varepsilon)$ positive definite universal kernel.

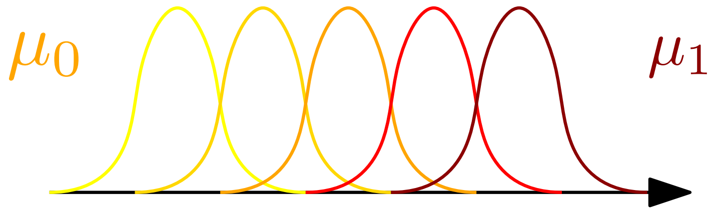
Now $X \subset \mathbb{R}^d$ and $c(x,y) = |x-y|^2$.

If $\mu \in \mathcal{P}(X)$, m barycenter and $\bar{\mu}$ centered part.



Reminder: $\mathrm{OT}(\mu_0,\mu_1) = |m_1 - m_0|^2 + \mathrm{OT}(\bar{\mu}_0,\bar{\mu}_1)$.

Theorem. $\mathrm{d}_S(\mu_0,\mu_1)^2 = |m_1 - m_0|^2 + \mathrm{d}_S(\bar{\mu}_0,\bar{\mu}_1)^2$.



Consequence: constant-speed translations are geodesics.

Conclusion and open questions

What I have not presented

- Explicit formula for Gaussians and the “two points” space.
- Example showing the Sinkhorn divergence is **not** jointly convex.

Open questions and future directions

- Limit $\varepsilon \to 0$ towards optimal transport.
- Numerical approximation of the distance?
- Gradient flows with respect to d_S (ongoing work with Mathis Hardion).

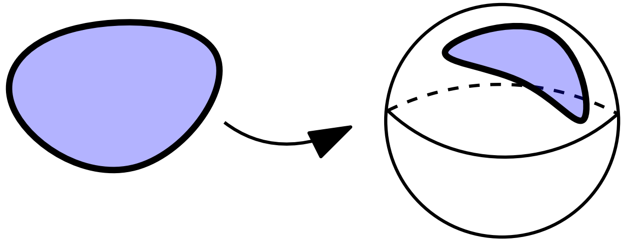
Conclusion and open questions

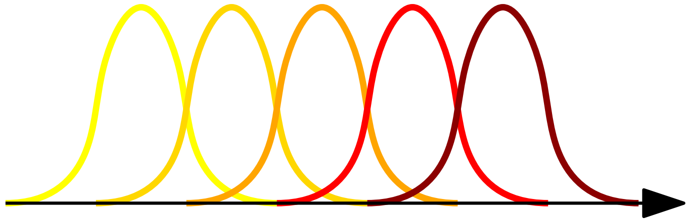
What I have not presented

- Explicit formula for Gaussians and the “two points” space.
- Example showing the Sinkhorn divergence is **not** jointly convex.

Open questions and future directions

- Limit $\varepsilon \to 0$ towards optimal transport.
- Numerical approximation of the distance?
- Gradient flows with respect to d_S (ongoing work with Mathis Hardion).





Thank you for your attention