# Gradient flows in the geometry of Sinkhorn divergences

Hugo Lavenant

Bocconi University

Kantorovich Initiative Seminar

Vancouver (Canada), October 31, 2024

# Joint work with

Mathis Hardion

TELECOM Paris

IP PARIS

Jonas Luckhardt

UNIVERSITÄT GÖTTINGEN

Gilles Mordant

UNIVERSITÄT GÖTTINGEN

Bernhard Schmitzer

UNIVERSITÄT GÖTTINGEN

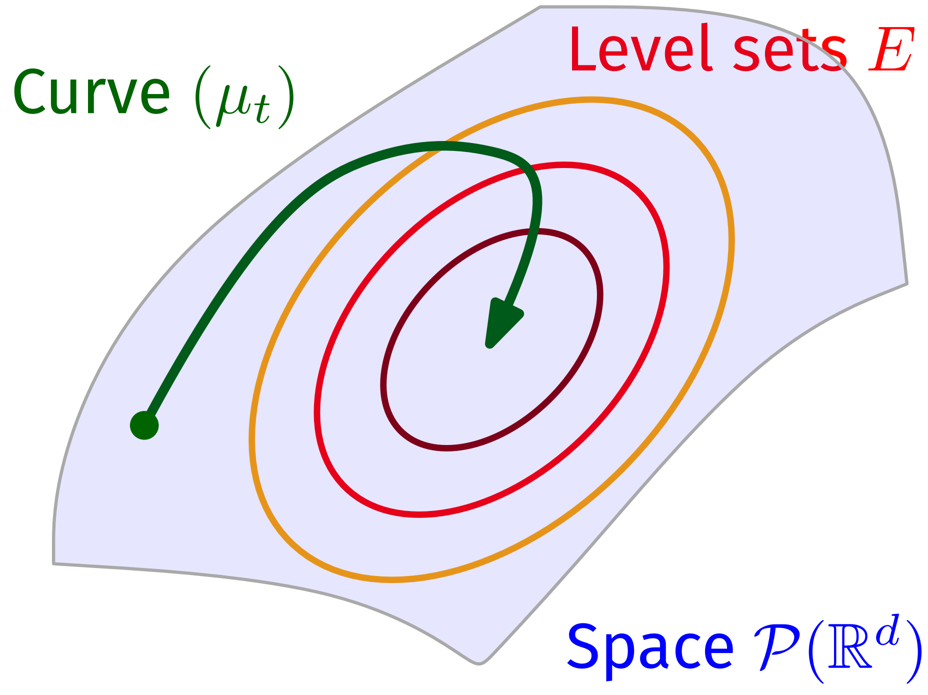
Luca Tamanini

UNIVERSITAS CATHOLICA SACRI CORDIS JESU MEDIOLANI

Lavenant, Luckhardt, Mordant, Schmitzer, Tamanini (2024). The Riemannian geometry of Sinkhorn divergences.
Hardion (2024). Master thesis: Gradient Flows in the Geometry of the Sinkhorn Divergence.

# Wasserstein gradient flows

$E : \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$ and $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ generate a curve $(\mu_t)_{t \geq 0}$ of **steepest descent** with respect to Wasserstein geometry.




## Examples

- $E(\mu) = \int V \, \mathrm{d}\mu$ gives the **transport equation**

$$\partial_t \mu = \mathrm{div}(\mu \nabla V).$$

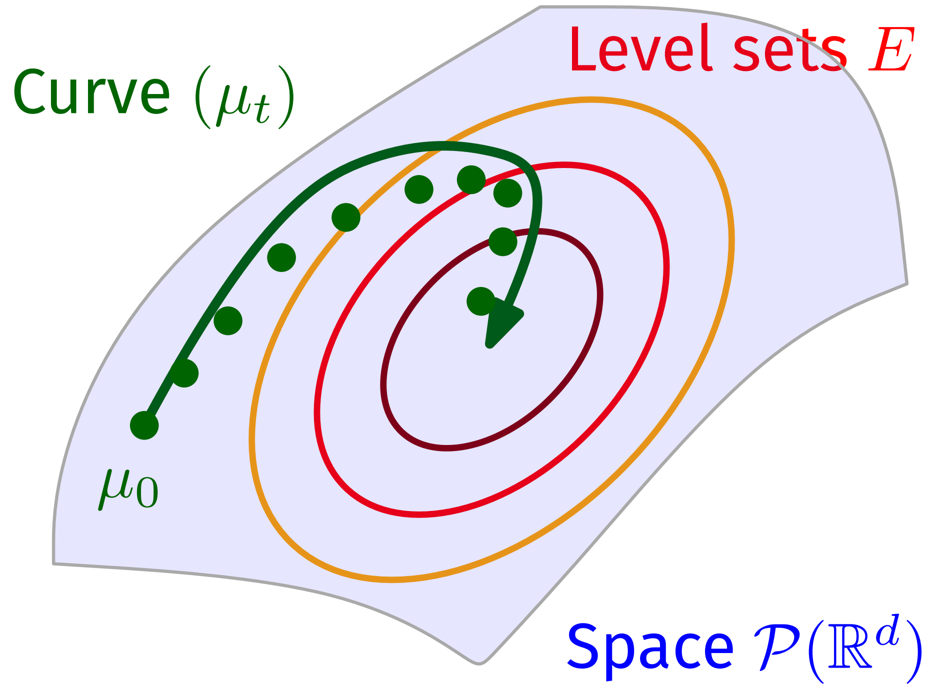
- $E(\mu) = \int \mu \log \mu$ gives the **heat equation**

$$\partial_t \mu = \Delta \mu.$$

Jordan, Kinderlehrer, Otto (1998). The variational formulation of the Fokker–Planck equation.

# Wasserstein gradient flows

$E : \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$ and $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ generate a curve $(\mu_t)_{t \geq 0}$ of **steepest descent** with respect to Wasserstein geometry.




**Recall**

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

Subset of $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, coupling between $\mu$ and $\nu$

**JKO/minimizing movement scheme.** For $\tau > 0$, define, for $k \geq 0$,

$$\mu_{k+1}^\tau \in \arg\min_{\mu} \; E(\mu) + \frac{\mathrm{OT}(\mu, \mu_k^\tau)}{2\tau}$$

Then $\mu_k^\tau \simeq \mu_{k\tau}$ as $\tau \to 0$.

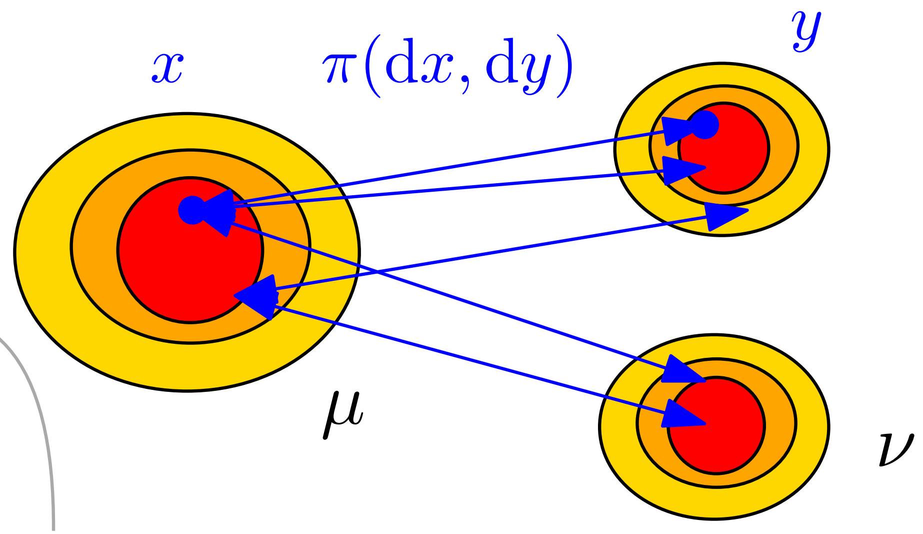
Jordan, Kinderlehrer, Otto (1998). The variational formulation of the Fokker–Planck equation.

# With entropic optimal transport?

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

**Definition**

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




Kullback-Leibler divergence, a.k.a relative entropy

**Why?**

1. easier to compute (**Sinkhorn algorithm**),
2. better statistical complexity,
3. smoother dependence in $(\mu, \nu)$.

Peyré & Cuturi (2018). Computational Optimal Transport.

## With entropic optimal transport?

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

**Definition**

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

If $c$ quadratic cost on $\mathbb{R}^d$, **Entropic JKO scheme:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}_\varepsilon(\mu,\mu_k^\tau)}{2\tau}$$

Peyré (2015). Entropic approximation of Wasserstein gradient flows.

## With entropic optimal transport?

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

**Definition**

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

**Recall**

$$\mathrm{OT}_\varepsilon = \mathrm{OT} + C_\varepsilon + \varepsilon[\text{Bias}] + \dots$$

If $c$ quadratic cost on $\mathbb{R}^d$, **Entropic JKO scheme:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}_\varepsilon(\mu,\mu_k^\tau)}{2\tau}$$

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}(\mu,\mu_k^\tau)}{2\tau} + \frac{\varepsilon}{2\tau}\text{Bias} + \dots$$

Conforti & Tamanini (2021). A formula for the time derivative of the entropic cost and applications.

## With entropic optimal transport?

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

**Definition**

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

**Recall**

$$\mathrm{OT}_\varepsilon = \mathrm{OT} + C_\varepsilon + \varepsilon[\text{Bias}] + \dots$$

If $c$ quadratic cost on $\mathbb{R}^d$, **Entropic JKO scheme:**

$$\mu^\tau_{k+1} \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}_\varepsilon(\mu,\mu^\tau_k)}{2\tau}$$

$$\mu^\tau_{k+1} \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}(\mu,\mu^\tau_k)}{2\tau} + \frac{\varepsilon}{2\tau}\text{Bias} + \dots$$

- $\varepsilon \ll \tau$: convergence to the Wasserstein GF of $E$.
- $\varepsilon \sim \tau$: convergence to a new flow.
- $\varepsilon \gg \tau$: the bias dominates, no evolution.

Conforti & Tamanini (2021). A formula for the time derivative of the entropic cost and applications.
Carlier, Duval, Peyré, Schmitzer (2017). Convergence of entropic schemes for optimal transport and gradient flows.
Baradat, Hraivoronska, Santambrogio (2024+). Using Sinkhorn in JKO adds diffusion in the limiting PDE.

## With entropic optimal transport?

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

**Definition**

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

**Recall**

$$\mathrm{OT}_\varepsilon = \mathrm{OT} + C_\varepsilon + \varepsilon[\text{Bias}] + \dots$$

If $c$ quadratic cost on $\mathbb{R}^d$, **Entropic JKO scheme:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}_\varepsilon(\mu, \mu_k^\tau)}{2\tau}$$

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{\mathrm{OT}(\mu, \mu_k^\tau)}{2\tau} + \frac{\varepsilon}{2\tau}\text{Bias} + \dots$$

- $\varepsilon \ll \tau$: convergence to the Wasserstein GF of $E$.
- $\varepsilon \sim \tau$: convergence to a new flow.
- $\varepsilon \gg \tau$: the bias dominates, no evolution.

**Today: I will keep $\varepsilon$ fixed.**

# Using Sinkhorn divergences

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

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

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

# Using Sinkhorn divergences

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

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

Assumption until the end of the talk

**Theorem** Assume $\exp(-c/\varepsilon)$ positive definite universal kernel.

1. $S_\varepsilon(\mu,\nu) \geq 0$ with equality iff $\mu = \nu$, and $S_\varepsilon$ "metrizes" weak convergence.
2. $S_\varepsilon$ convex in each of its inputs.

but $\sqrt{S_\varepsilon}$ not a distance

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.

## Using Sinkhorn divergences

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

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

Assumption until the end of the talk

**Theorem** Assume $\exp(-c/\varepsilon)$ positive definite universal kernel.

1. $S_\varepsilon(\mu,\nu) \geq 0$ with equality iff $\mu = \nu$, and $S_\varepsilon$ "metrizes" weak convergence.
2. $S_\varepsilon$ convex in each of its inputs.

but $\sqrt{S_\varepsilon}$ not a distance

**Sinkhorn JKO:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu,\mu_k^\tau)}{2\tau}$$

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.

## Using Sinkhorn divergences

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

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

Assumption until the end of the talk

**Theorem** Assume $\exp(-c/\varepsilon)$ positive definite universal kernel.

1. $S_\varepsilon(\mu,\nu) \geq 0$ with equality iff $\mu = \nu$, and $S_\varepsilon$ "metrizes" weak convergence.
2. $S_\varepsilon$ convex in each of its inputs.

but $\sqrt{S_\varepsilon}$ not a distance

**Sinkhorn JKO:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu,\mu_k^\tau)}{2\tau}$$

If $\quad S_\varepsilon(\mu_t,\mu_{t+\tau}) \sim \tau^2 \mathbf{g}_{\mu_t}(\dot\mu_t,\dot\mu_t)$,

we expect the equation when $\tau \to 0$:

$$\dot\mu_t \in \arg\min_\sigma \; DE(\mu_t)[\sigma] + \frac{\mathbf{g}_{\mu_t}(\sigma,\sigma)}{2}.$$

## Using Sinkhorn divergences

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

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

Assumption until the end of the talk

**Theorem** Assume $\exp(-c/\varepsilon)$ positive definite universal kernel.

1. $S_\varepsilon(\mu,\nu) \geq 0$ … onvergence.
2. $S_\varepsilon$ convex in …

**Metric tensor** $\mathbf{g}_\mu(\dot\mu,\dot\mu)$, defining the geometry of Sinkhorn divergences

**Sinkhorn JKO:**

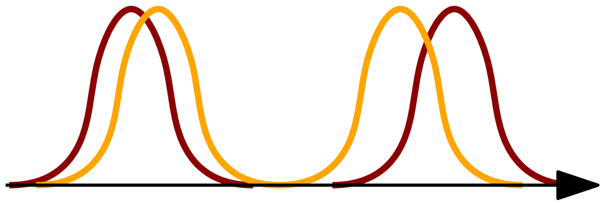
$$\mu^\tau_{k+1} \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu,\mu_k)}{2\tau}$$

… distance

If $\quad S_\varepsilon(\mu_t,\mu_{t+\tau}) \sim \tau^2 \mathbf{g}_{\mu_t}(\dot\mu_t,\dot\mu_t),$

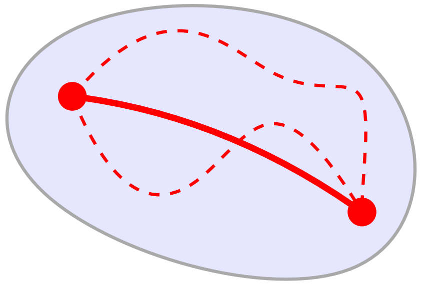
we expect the equation when $\tau \to 0$:

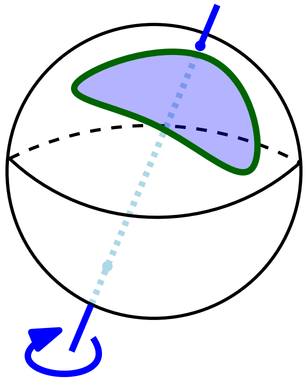
$$\dot\mu_t \in \arg\min_\sigma \; DE(\mu_t)[\sigma] + \frac{\mathbf{g}_{\mu_t}(\sigma,\sigma)}{2}.$$



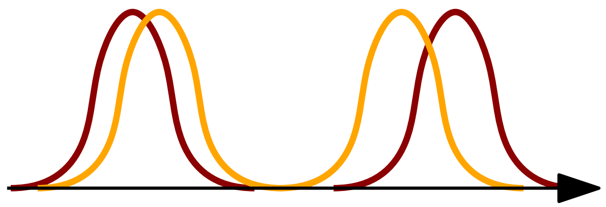
**1 - Optimal transport: metric tensor, geometry, gradient flows**

**2 - Building a Riemannian geometry out of Sinkhorn divergences**



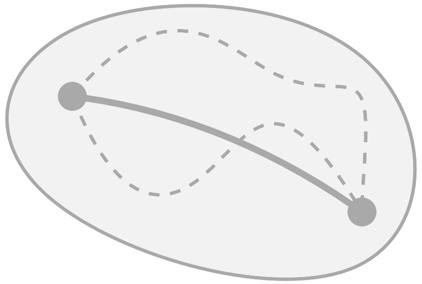


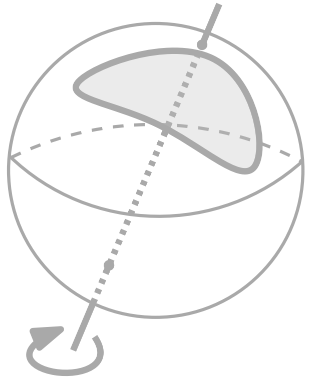
**3 - Gradient flows of potential energies for the Sinkhorn geometry**



**1 - Optimal transport: metric tensor, geometry, gradient flows**

**2 - Building a Riemannian geometry out of Sinkhorn divergences**



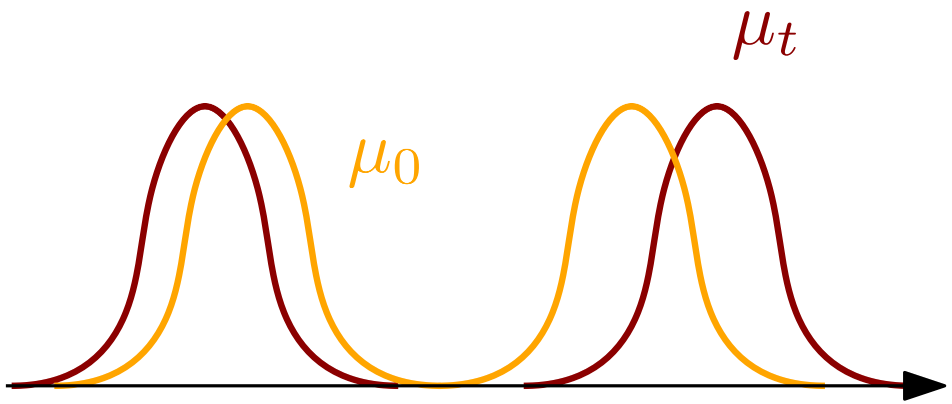


**3 - Gradient flows of potential energies for the Sinkhorn geometry**

# 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 $\mathrm{OT}(\mu_0, \mu_t)$.

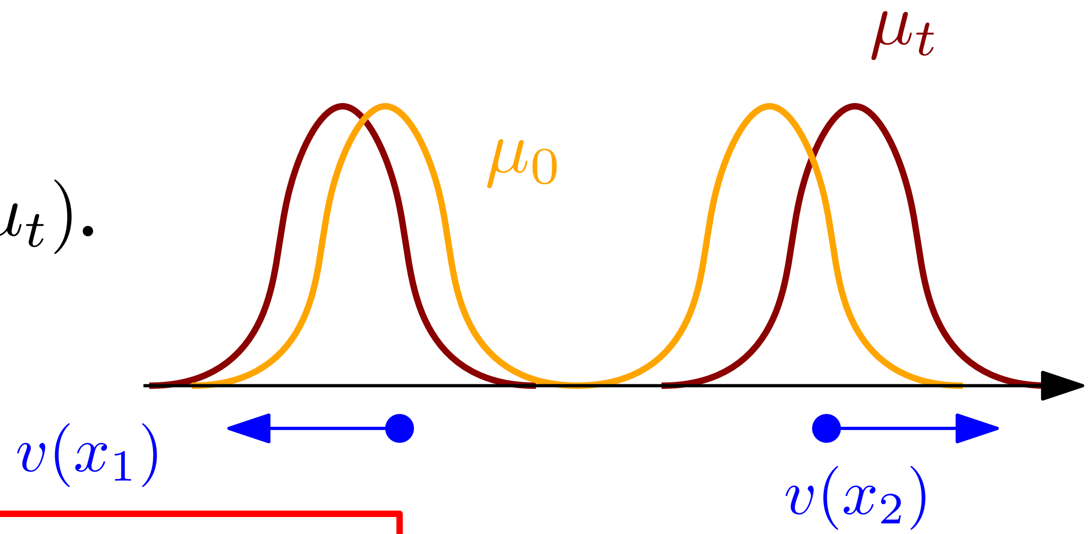



Ambrosio, Gigli & Savaré (2008). Gradient flows: in metric spaces and in the space of probability measures.

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

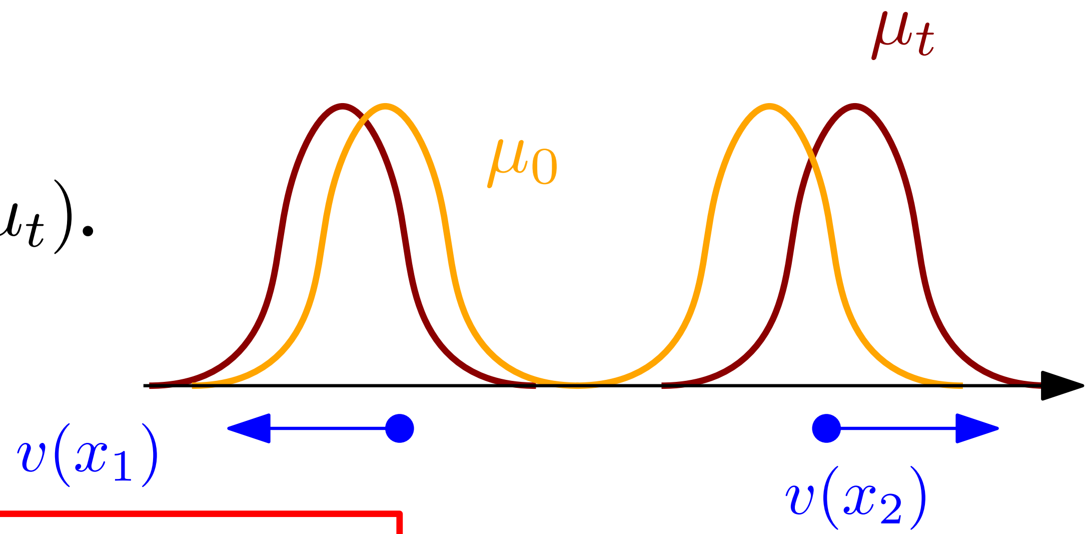
where $v : \mathbb{R}^d \to \mathbb{R}^d$ such that $\left.\dfrac{\partial \mu}{\partial t}\right|_{t=0} = -\mathrm{div}(\mu_0 v)$.

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

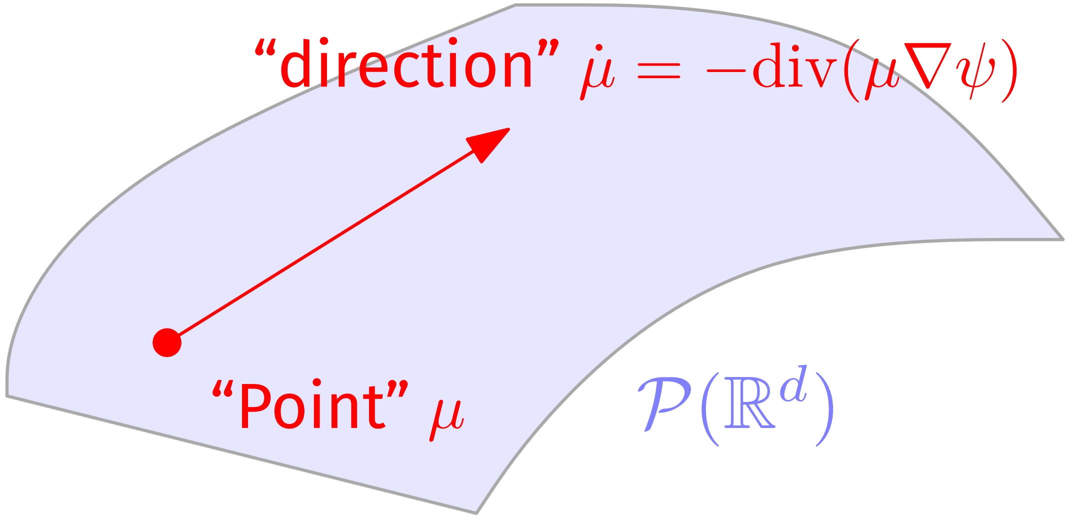
where $v : \mathbb{R}^d \to \mathbb{R}^d$ such that $\left.\dfrac{\partial \mu}{\partial t}\right|_{t=0} = -\mathrm{div}(\mu_0 v)$.

elliptic equation in $\psi$

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



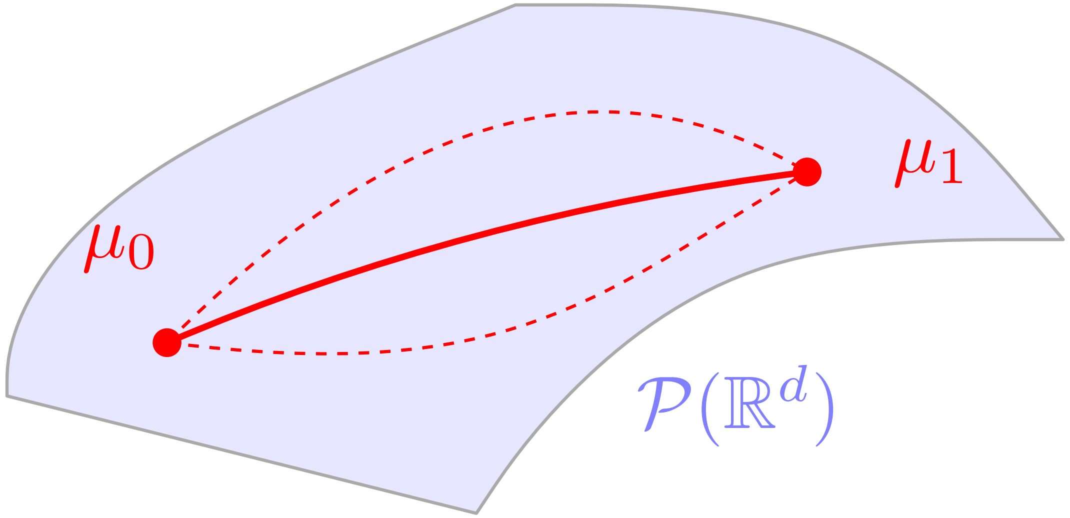

**Metric tensor:**

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

Quadratic form in $\dot\mu$, depending on $\mu$.

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

# The metric tensor and the geometry of optimal transport




Metric tensor:

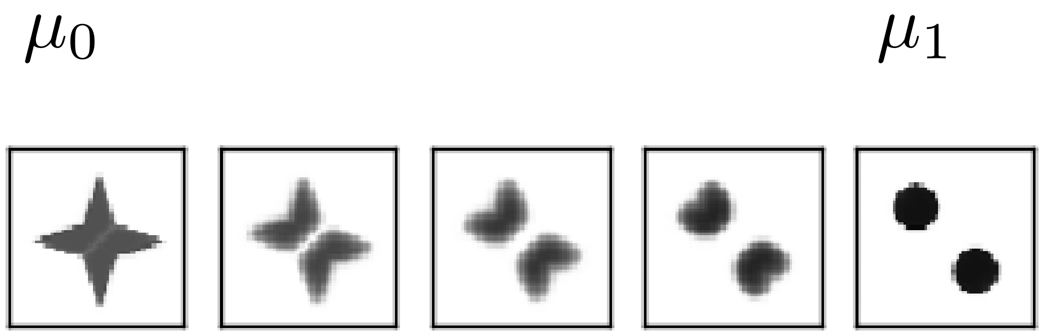
$$\mathbf{g}^{\mathrm{OT}}_{\mu}(\dot{\mu}, \dot{\mu}) = \int_X |\nabla\psi|^2 \, \mathrm{d}\mu.$$

**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 $\mu_0, \mu_1$ fixed.

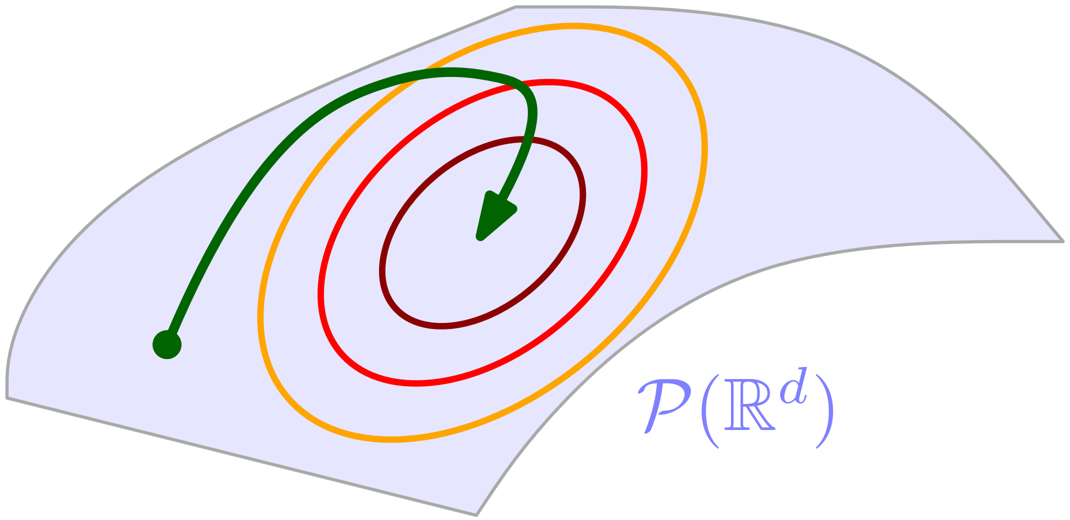
Minimizers are **geodesics**.




Example geodesic

Benamou and Brenier (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem.

# 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 $\mu_0, \mu_1$ fixed.

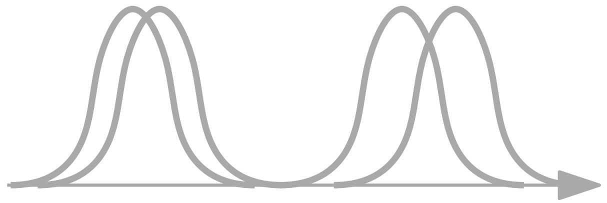
Minimizers are **geodesics**.

**Gradient flows:** JKO scheme

$$\mu^\tau_{k+1} \in \arg\min_{\mu} E(\mu) + \frac{\mathrm{OT}(\mu, \mu^\tau_k)}{2\tau},$$

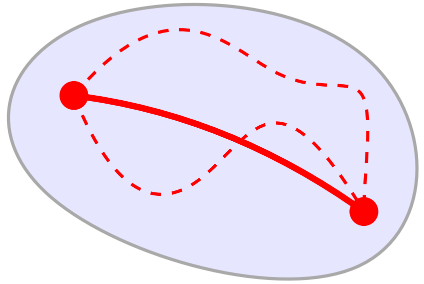
becomes with $\tau \to 0$

$$\dot\mu_t \in \arg\min_{\sigma} DE(\mu_t)[\sigma] + \frac{\mathbf{g}^{\mathrm{OT}}_{\mu_t}(\sigma, \sigma)}{2}.$$



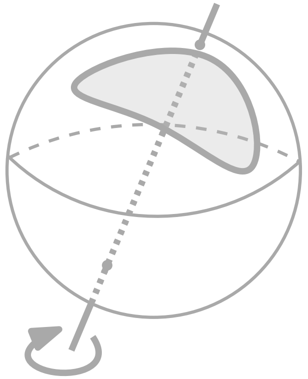
**1 - Optimal transport: metric tensor, geometry, gradient flows**

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

See also Park & Slepčev (2023). Geometry and analytic properties of the sliced Wasserstein space.

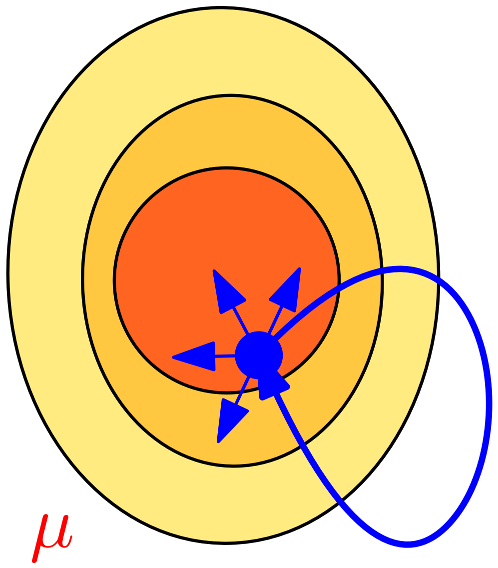


**3 - Gradient flows of potential energies for the Sinkhorn geometry**

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

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

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



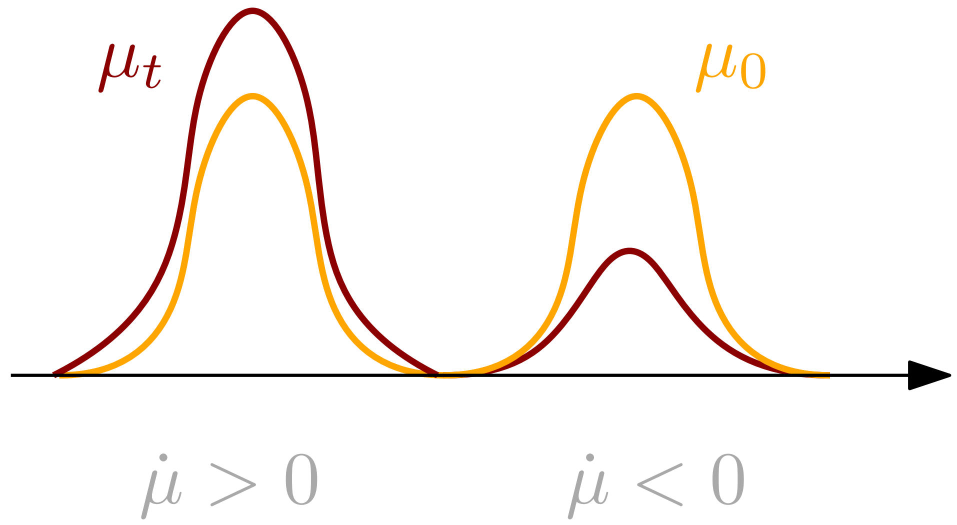

(Defines a reversible Markov chain with equilibirum measure $\mu$.)

**Definition:**

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

# The Hessian of the Sinkhorn divergence

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

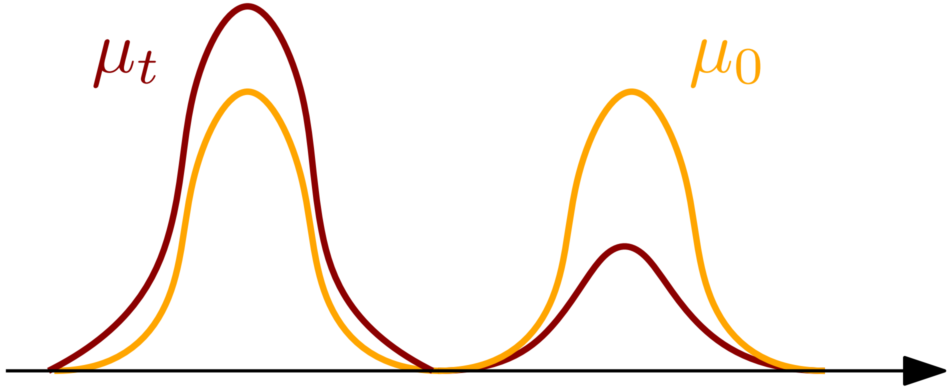


# The Hessian of the Sinkhorn divergence

$\mu_t = \mu + t\dot{\mu}$, 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:

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

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

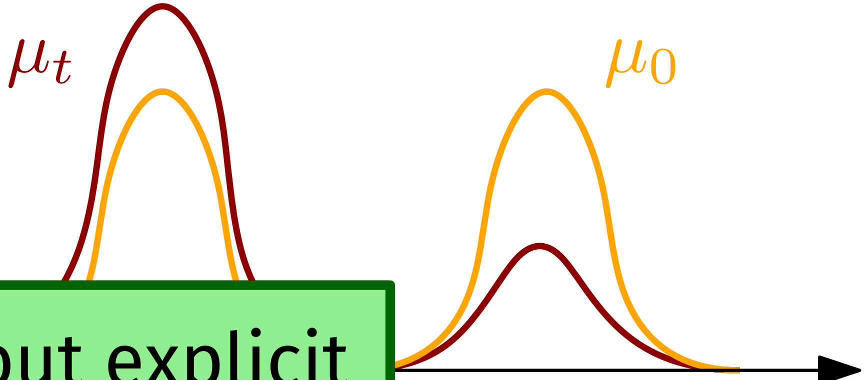
$(\mathrm{Id} - K_\mu^2)/\varepsilon \sim$ Laplacian

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

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

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

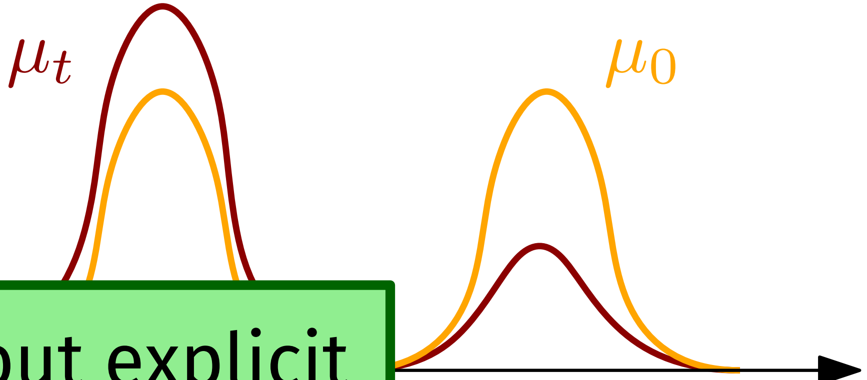
$(\mathrm{Id} - K_\mu^2)/\varepsilon \sim$ Laplacian

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

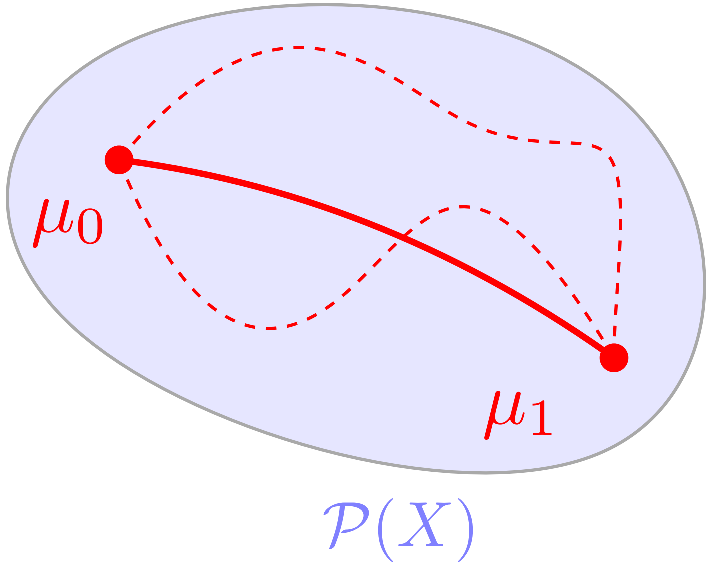
# Definition of the Riemannian distance and main results

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

**Definition.** Given $\mu_0, \mu_1$:

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

where infimum over $(\mu_t)$ on a class of path to be specified in the next slides.




Both "vertical" and "horizontal" are allowed!

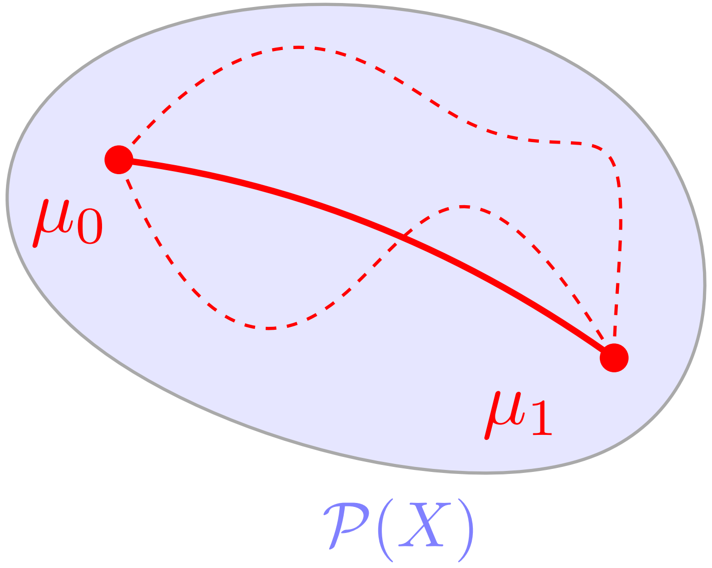
## Definition of the Riemannian distance and main results

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

**Definition.** Given $\mu_0, \mu_1$:

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

where infimum over $(\mu_t)$ on a class of path to be specified in the next slides.




Both "vertical" and "horizontal" are allowed!

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

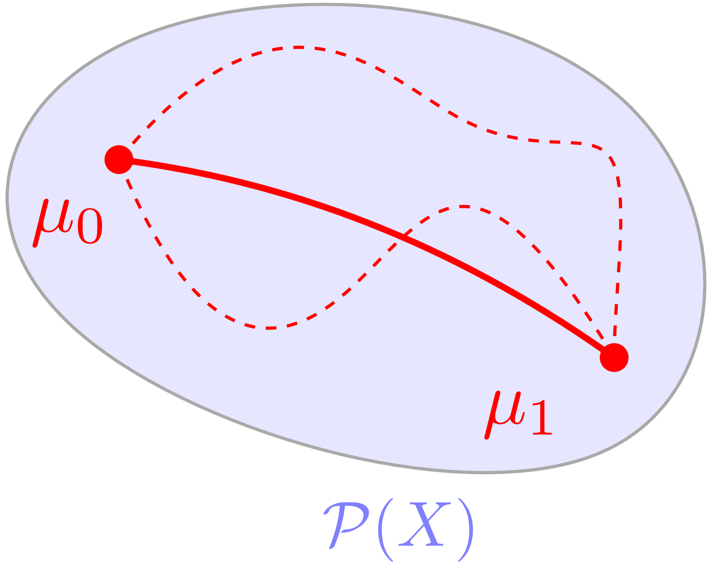
# Definition of the Riemannian distance and main results

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

**Definition.** Given $\mu_0, \mu_1$:

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

where infimum over $(\mu_t)$ on a class of path to be specified in the next slides.




Both "vertical" and "horizontal" are allowed!

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

**Elements of the proof**: next slides.

# 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:

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

Typically smooth functions!

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

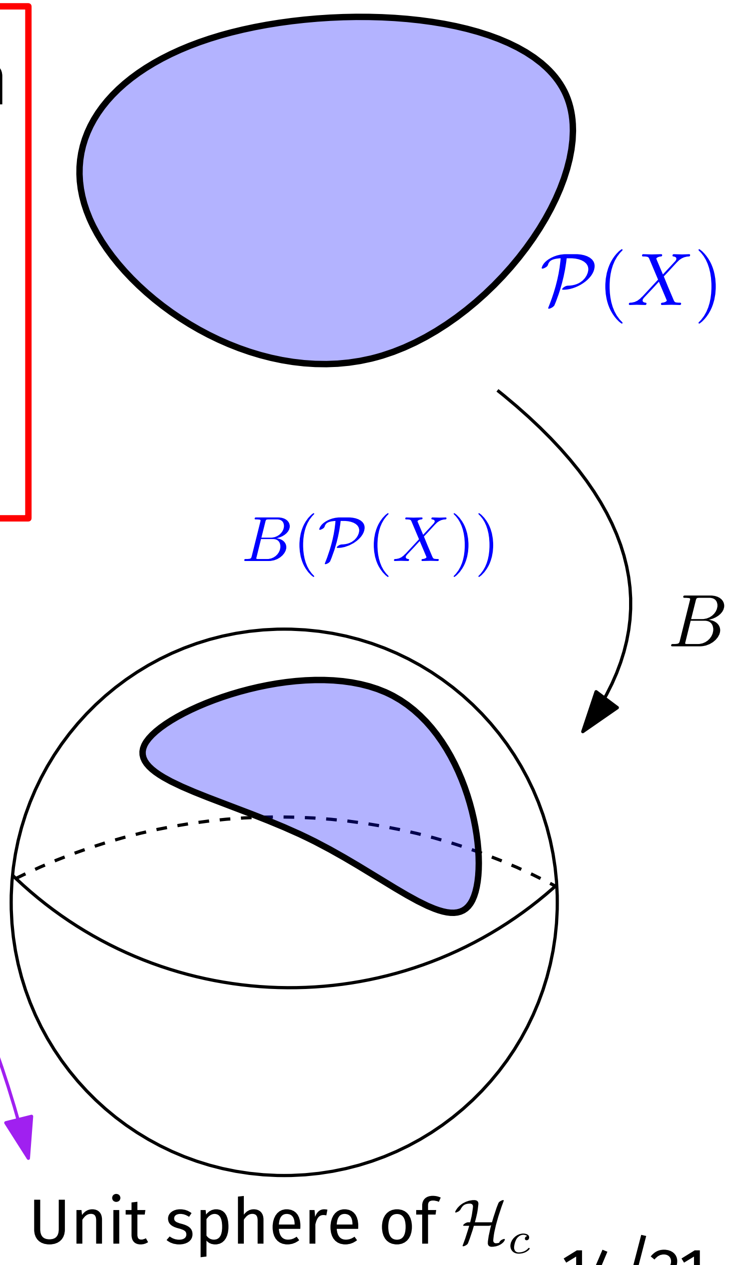
# A useful change of variable

**Define:**

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

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

**Theorem.** The map $B$ is an homeomorphism between $\mathcal{P}(X)$ and the intersection of a convex cone and the unit sphere of $\mathcal{H}_c$.




$\mathcal{H}_c$: Reproducing Kernel Hilbert Space built on $\exp(-c/\varepsilon)$.

(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:**

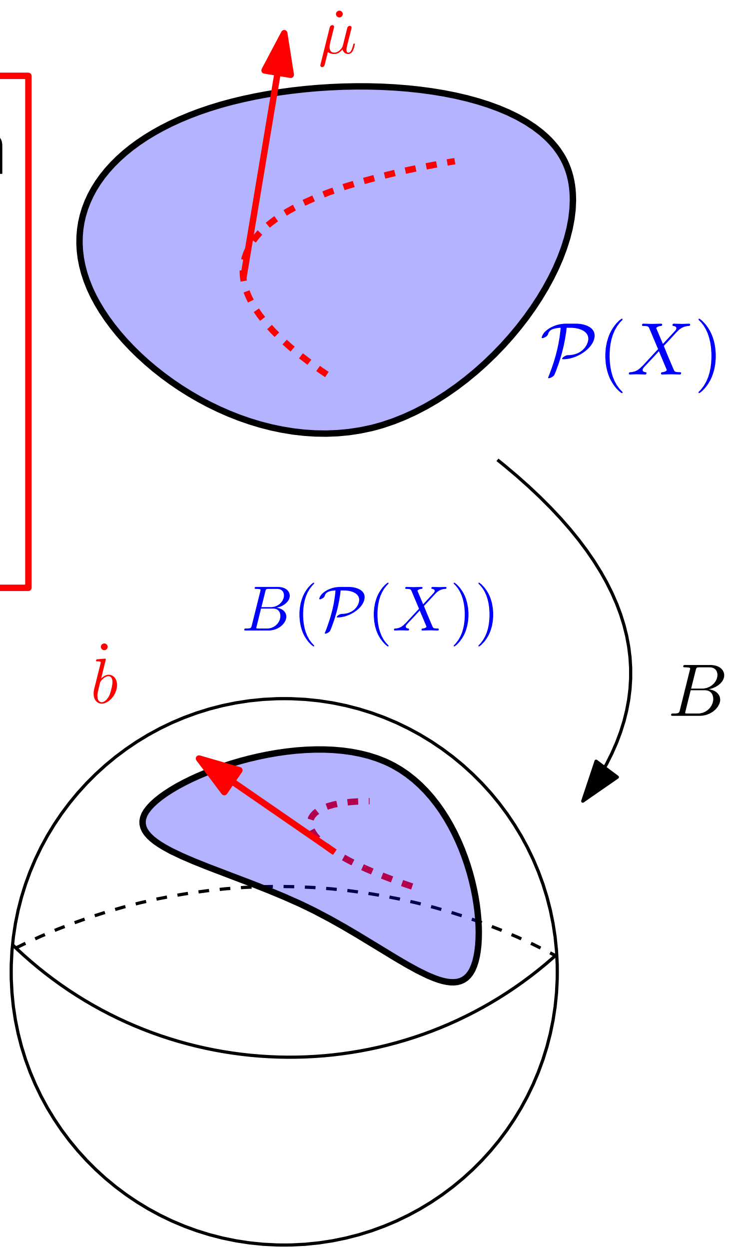
$$b = B(\mu) = \exp\left(-\frac{f_\mu}{\varepsilon}\right)$$

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

**Theorem.** The map $B$ is an homeomorphism between $\mathcal{P}(X)$ and the intersection of a convex cone and the unit sphere of $\mathcal{H}_c$.

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

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



## A useful change of variable

**Define:**

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

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

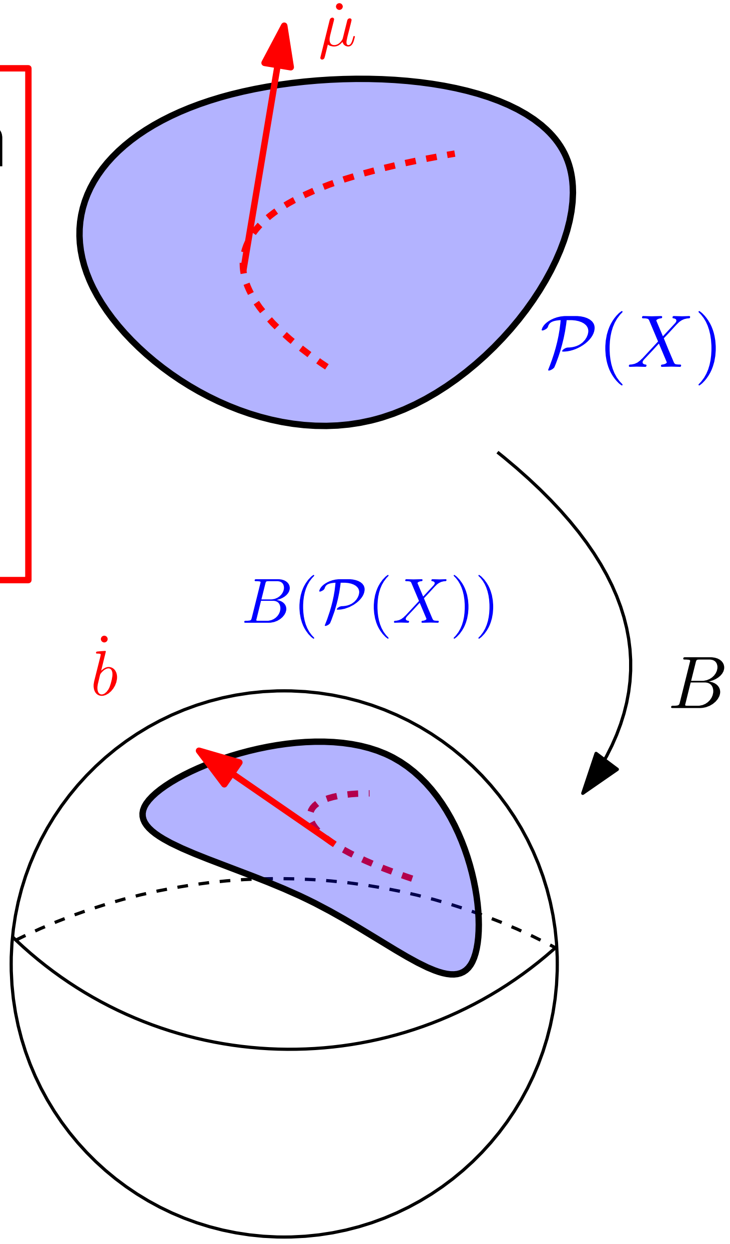
**Theorem.** The map $B$ is an homeomorphism between $\mathcal{P}(X)$ and the intersection of a convex cone and the unit sphere of $\mathcal{H}_c$.

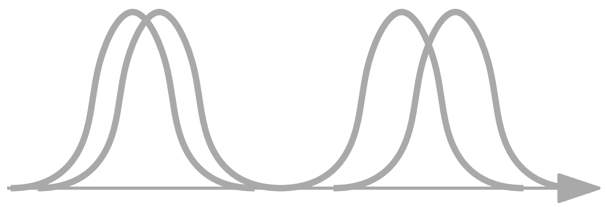
**Theorem.** We have $\mathbf{g}_{\mu_t}(\dot\mu_t, \dot\mu_t) = \tilde{\mathbf{g}}_{\mu_t}(\dot b_t, \dot b_t)$ and:

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

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

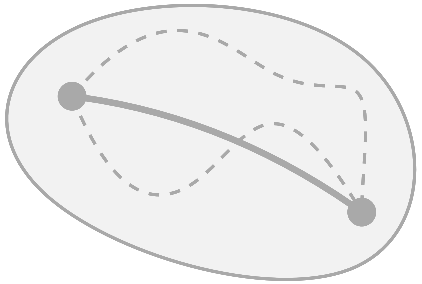
$$\mathsf{d}_S(\mu_0, \mu_1) \asymp \|b_1 - b_0\|_{\mathcal{H}_c}.$$

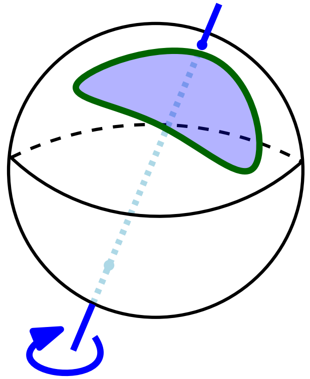




**1 - Optimal transport: metric tensor, geometry, gradient flows**

**2 - Building a Riemannian geometry out of Sinkhorn divergences**





**3 - Gradient flows of potential energies for the Sinkhorn geometry**

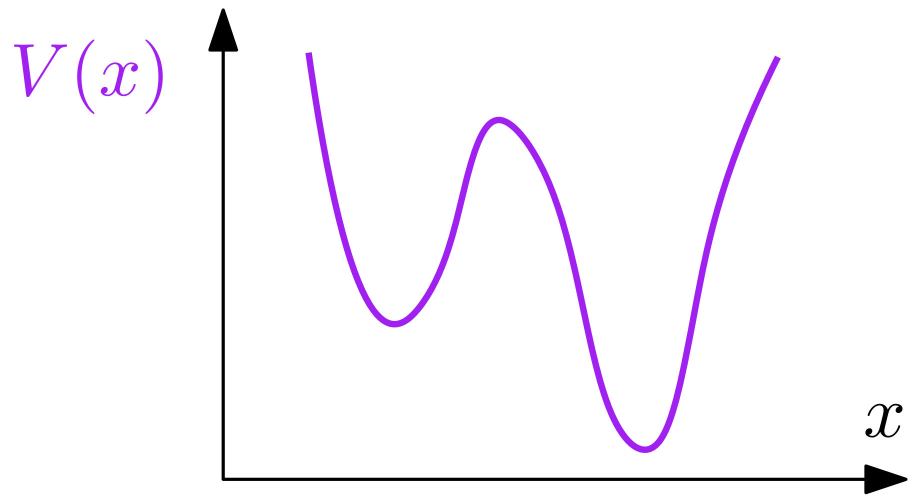
## Notations, object of study

$X$ compact metric, $\exp(-c/\varepsilon)$ p. d. and universal, and $V : X \to \mathbb{R}$ continuous.

**Sinkhorn JKO:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu, \mu_k^\tau)}{2\tau}.$$

with $S_\varepsilon$ Sinkhorn divergence and $E(\mu) = \int V \,\mathrm{d}\mu$ **potential energy**.



# Notations, object of study

$X$ compact metric, $\exp(-c/\varepsilon)$ p. d. and universal, and $V : X \to \mathbb{R}$ continuous.

**Sinkhorn JKO:**

$$\mu^\tau_{k+1} \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu, \mu^\tau_k)}{2\tau}.$$

with $S_\varepsilon$ Sinkhorn divergence and $E(\mu) = \int V \, \mathrm{d}\mu$ **potential energy**.

**Recall**

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

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

**Formal** limit when $\tau \to 0$:

Evolution equation, **Sinkhorn flow**

$$\frac{\varepsilon}{2}(\mathrm{Id} - K^2_{\mu_t})^{-1} H_{\mu_t}[\dot{\mu}_t] + V + p_t = \text{Cst.}$$

$p_t$ pressure: $p_t \leq 0$, and $p_t = 0$ on $\mathrm{supp}(\mu_t)$

## Notations, object of study

$X$ compact metric, $\exp(-c/\varepsilon)$ p. d. and universal, and $V : X \to \mathbb{R}$ continuous.

**Sinkhorn JKO:**

$$\mu_{k+1}^{\tau} \in \arg\min_{\mu} \; E(\mu) + \frac{S_\varepsilon(\mu, \mu_k^\tau)}{2\tau}.$$

with $S_\varepsilon$ Sinkhorn divergence and $E(\mu) = \int V \, \mathrm{d}\mu$ **potential energy**.

**Recall**

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

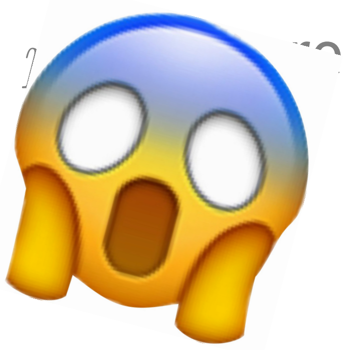
$$H_\mu[\sigma](x) = \int_X k_\mu(x,y)\,\mathrm{d}\sigma(y).$$

**Formal** limit when $\tau \to 0$:

Evolution equation, **Sinkhorn flow**

$$\frac{\varepsilon}{2}(\mathrm{Id} - K_{\mu_t}^2)^{-1} H_{\mu_t}[\dot{\mu}_t] + V + p_t = \text{Cst.}$$

: $p_t \leq 0$, and $p_t = 0$ on



So $\dot{\mu}_t = H_{\mu_t}^{-1}[\ldots]$ with $H_{\mu_t}$ "convolution".
**Non local** equation **of infinite order**.

# Sinkhorn flow in the Hilbert space $\mathcal{H}_c$

Recall $b_t = \exp(-f_{\mu_t}/\varepsilon) \in \mathcal{H}_c$ and $V : X \to \mathbb{R}$.

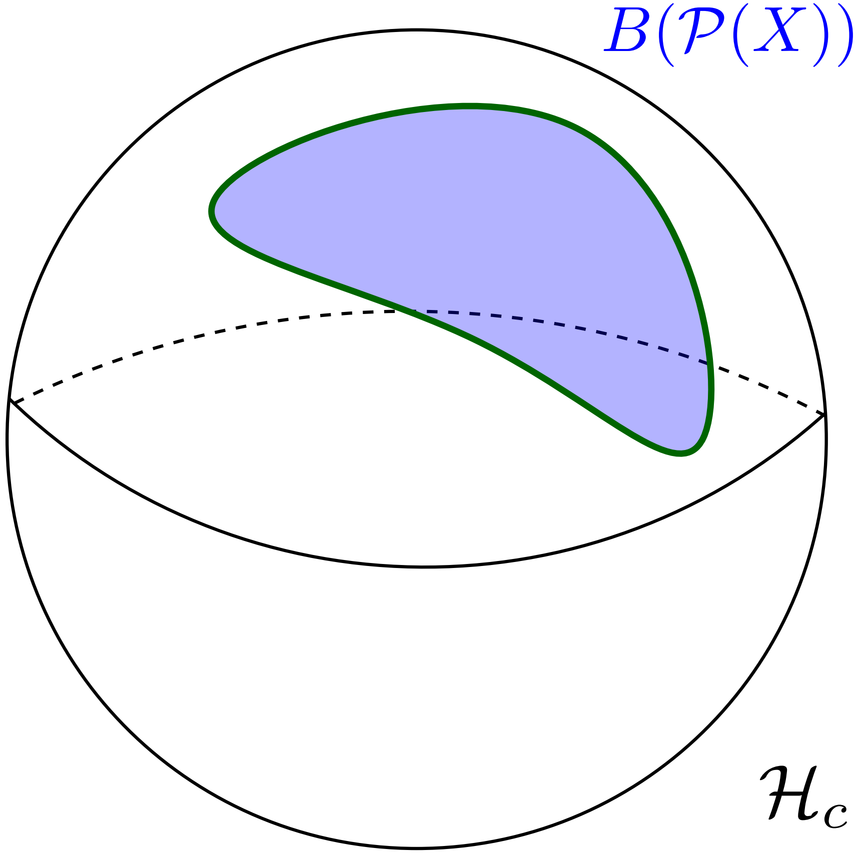
**Sinkhorn flow** in the $b$-variable.

$$\dot{b}_t + \frac{2}{\varepsilon}(V - V^*)\, b_t + p_t = 0.$$

$p_t \leq 0$, and $p_t = 0$ on $\operatorname{supp}(\mu_t)$

multiplication by $V$ in $\mathcal{H}_c$

$V^*$ Adjoint of multiplication by $V$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$



# Sinkhorn flow in the Hilbert space $\mathcal{H}_c$

Recall $b_t = \exp(-f_{\mu_t}/\varepsilon) \in \mathcal{H}_c$ and $V : X \to \mathbb{R}$.

**Sinkhorn flow in the $b$-variable.**

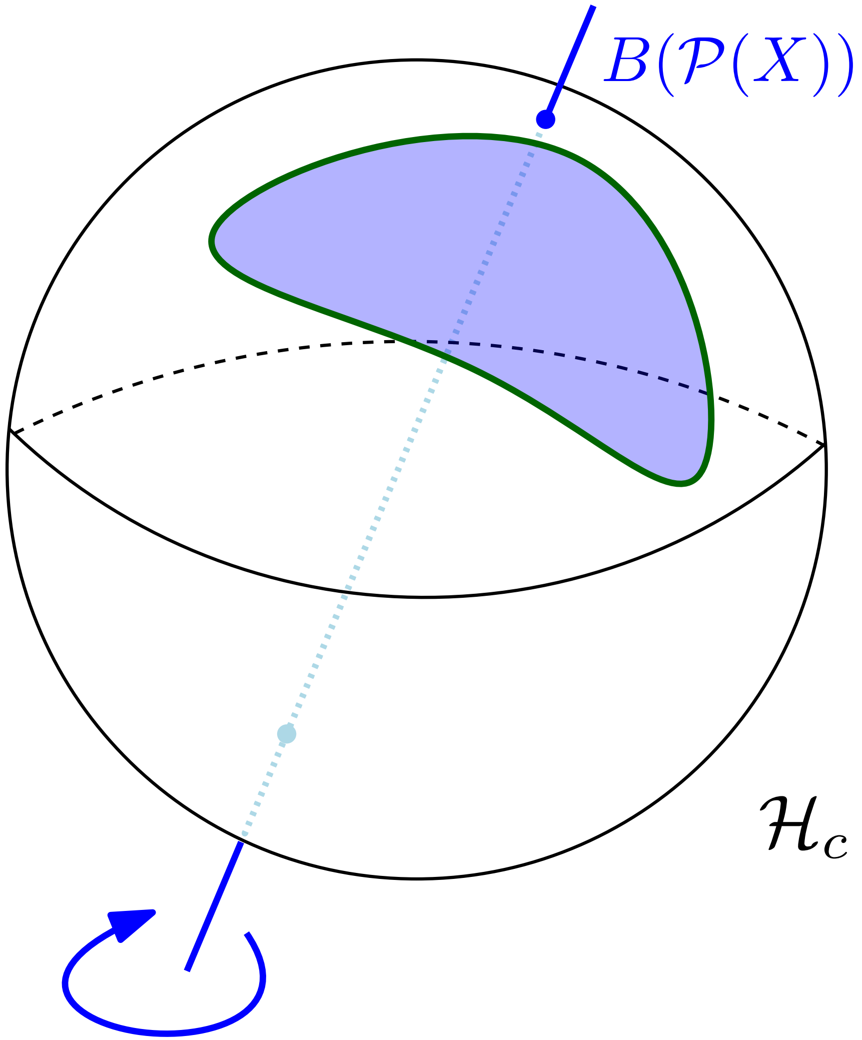
$$\dot{b}_t + \frac{2}{\varepsilon}(V - V^*)\, b_t + p_t = 0.$$

$p_t \leq 0$, and $p_t = 0$ on $\mathrm{supp}(\mu_t)$

multiplication by $V$ in $\mathcal{H}_c$

$V^*$ Adjoint of multiplication by $V$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$

$\frac{2}{\varepsilon}(V - V^*)$ skew-symmetric: generates group of unitary operators, but unbounded.




"Rotation" generated by $\frac{2}{\varepsilon}(V - V^*)$

# Sinkhorn flow in the Hilbert space $\mathcal{H}_c$

Recall $b_t = \exp(-f_{\mu_t}/\varepsilon) \in \mathcal{H}_c$ and $V : X \to \mathbb{R}$.

**Sinkhorn flow** in the $b$-variable.

$$\dot{b}_t + \frac{2}{\varepsilon}(V - V^*)\, b_t + p_t = 0.$$

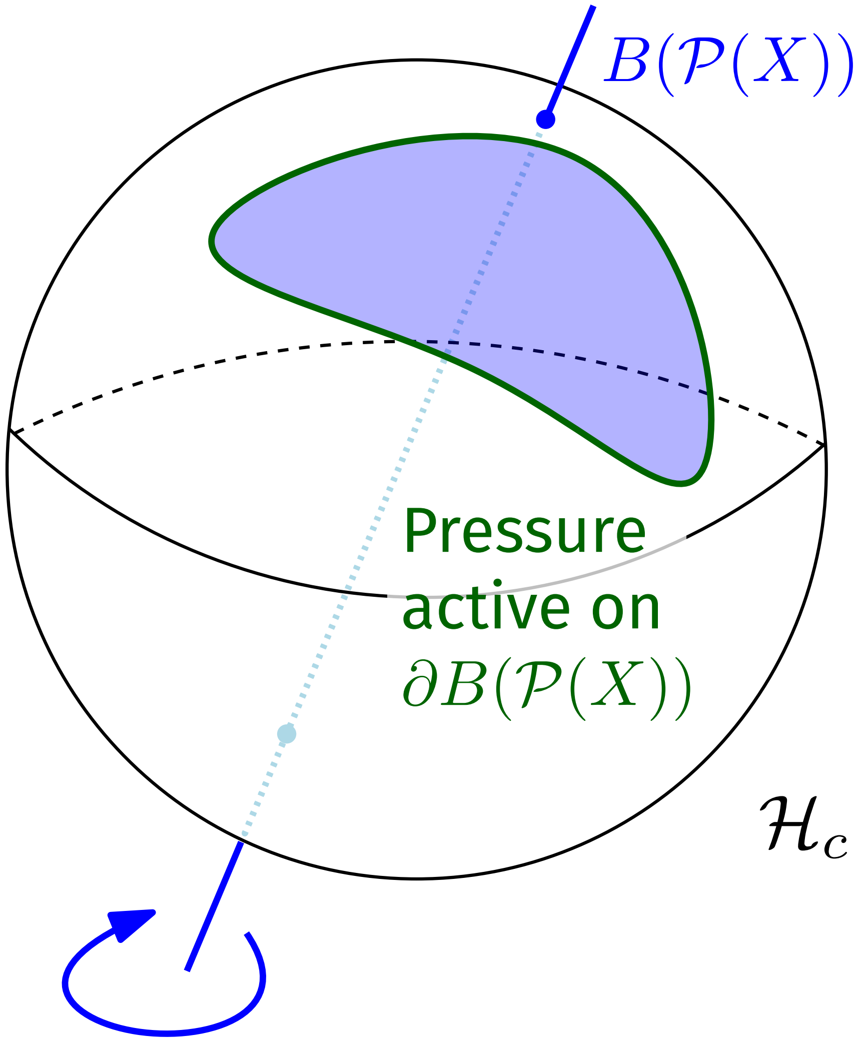
$p_t \leq 0$, and $p_t = 0$ on $\mathrm{supp}(\mu_t)$

multiplication by $V$ in $\mathcal{H}_c$

$V^*$ Adjoint of multiplication by $V$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$

$\frac{2}{\varepsilon}(V - V^*)$ skew-symmetric: generates group of unitary operators, but unbounded.

Pressure: in the polar cone of $B(\mathcal{M}_+(X))$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$, maintains $\mu_t \geq 0$.




"Rotation" generated by $\frac{2}{\varepsilon}(V - V^*)$

# Sinkhorn flow in the Hilbert space $\mathcal{H}_c$

Recall $b_t = \exp(-f_{\mu_t}/\varepsilon) \in \mathcal{H}_c$ and $V$

**Sinkhorn flow** in the $b$-variable.

$$\dot{b}_t + \frac{2}{\varepsilon}(V - V^*)\, b_t + p_t = 0.$$

multiplication by $V$ in $\mathcal{H}_c$

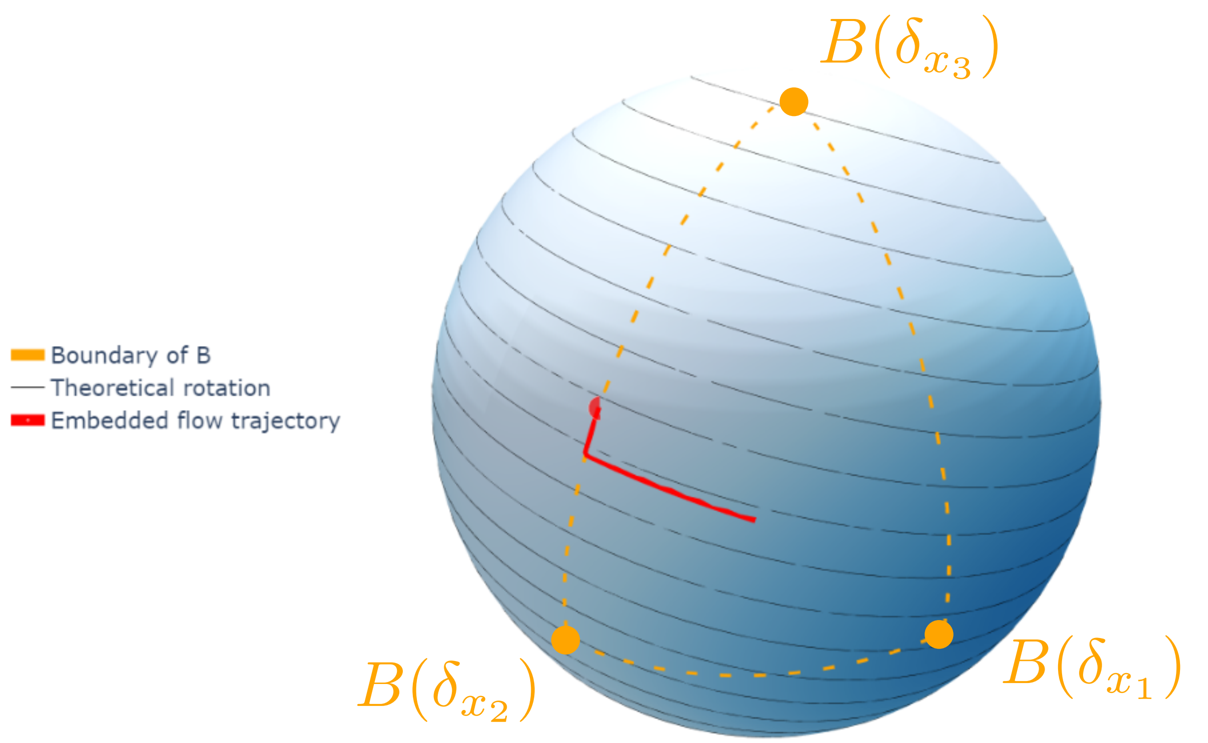
$V^*$ Adjoint of multi by $V$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$

$\frac{2}{\varepsilon}(V - V^*)$ skew-symmetric: genera group of unitary operators, but unb

Pressure: in the polar cone of $B(\mathcal{M}_+(X))$ for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$, maintains $\mu_t$

$X = \{x_1, x_2, x_3\}$

$V(x_3) \leq V(x_2) \leq V(x_1)$



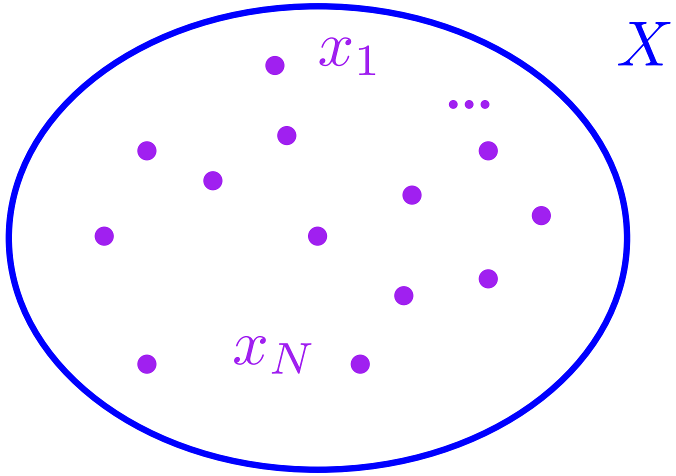

Obtained by solving the **SJKO** scheme.

# Main theoretical results

**Theorem** (on the Sinkhorn flow).

1. **Existence**: for any $b_0 = B(\mu_0)$, there exists a solution, with $(b_t) \in H^1([0,+\infty), \mathcal{H}_c)$.

**Proof idea**: approximate $X$ by a finite space $X_N = \{x_1, \dots, x_N\}$. For measures supported on $X_N$, the Sinkhorn flow is a maximal monotone evolution.



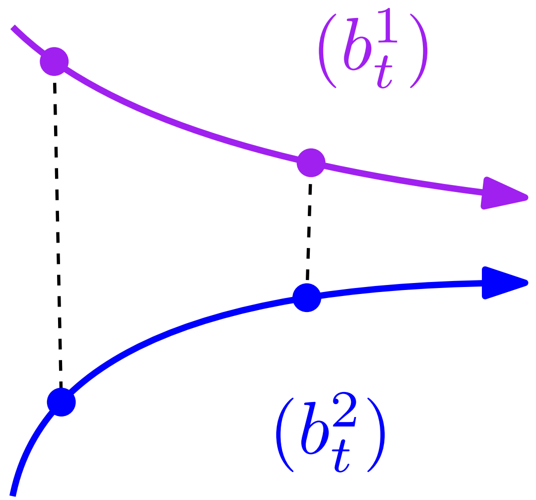
# Main theoretical results

**Theorem** (on the Sinkhorn flow).

1. **Existence**: for any $b_0 = B(\mu_0)$, there exists a solution, with $(b_t) \in H^1([0, +\infty), \mathcal{H}_c)$.

2. The flow is **non-expansive** in $\mathcal{H}_c$: for two flows $(b_t^1)$ and $(b_t^2)$, we have $\|b_t^2 - b_t^1\|_{\mathcal{H}_c} \leq \|b_0^2 - b_0^1\|_{\mathcal{H}_c}$ for all $t \geq 0$. It implies **uniqueness**.

**Proof idea**: Maximal monotone operators are non-expansive.

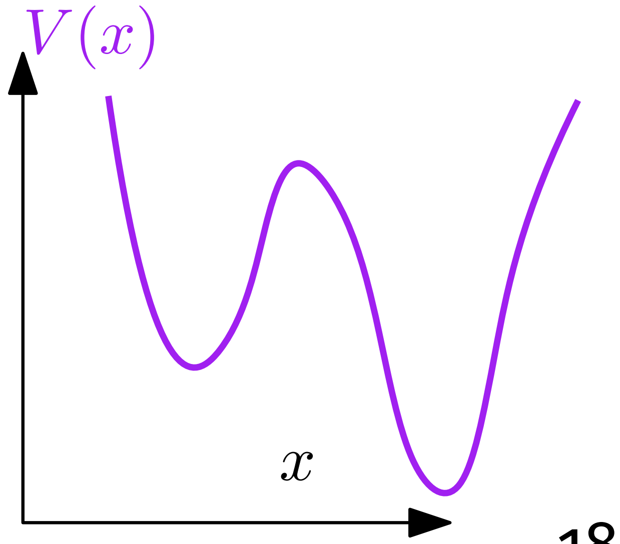


# Main theoretical results

**Theorem** (on the Sinkhorn flow).

1. **Existence**: for any $b_0 = B(\mu_0)$, there exists a solution, with $(b_t) \in H^1([0, +\infty), \mathcal{H}_c)$.
2. The flow is **non-expansive** in $\mathcal{H}_c$: for two flows $(b_t^1)$ and $(b_t^2)$, we have $\|b_t^2 - b_t^1\|_{\mathcal{H}_c} \leq \|b_0^2 - b_0^1\|_{\mathcal{H}_c}$ for all $t \geq 0$. It implies **uniqueness**.
3. Convergence to **global minimum**: $E(\mu_t) \to \min E$ as $t \to +\infty$.

**Recall**. The flow of the Wasserstein GF $\partial \mu_t = \mathrm{div}(\mu_t \nabla V)$ gets trapped in local minima.



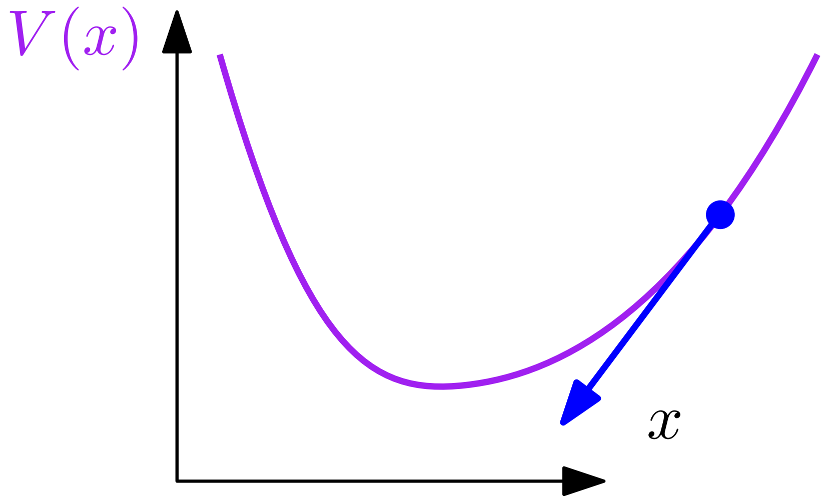

**Proof idea**. The only critical points of $E$ are global minima because vertical perturbations (teleportation) is allowed in the Sinkhorn geometry (no convexity of $V$ needed).

## Examples

**Proposition.** If $X = \mathbb{R}^d$, $V$ convex, $c$ quadratic cost and $\mu_0 = \delta_{x_0}$ then $\mu_t = \delta_{x_t}$ with

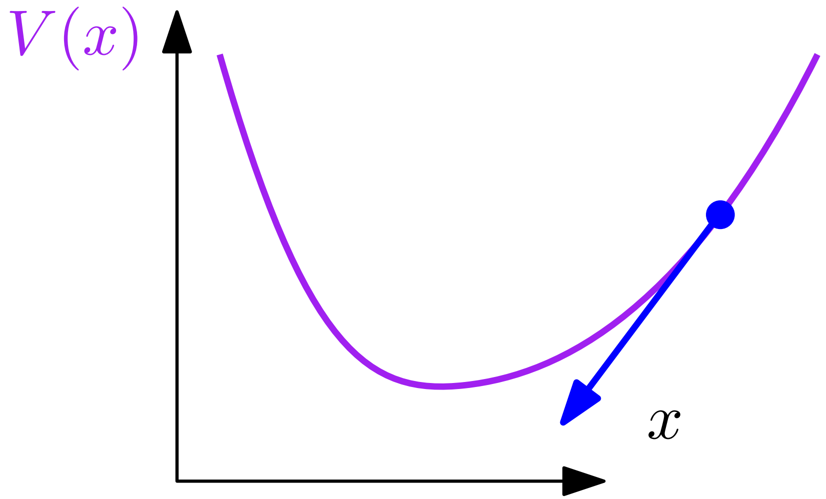
$$\dot{x}_t \in -\partial V(x_t).$$

(Same as Wasserstein GF).



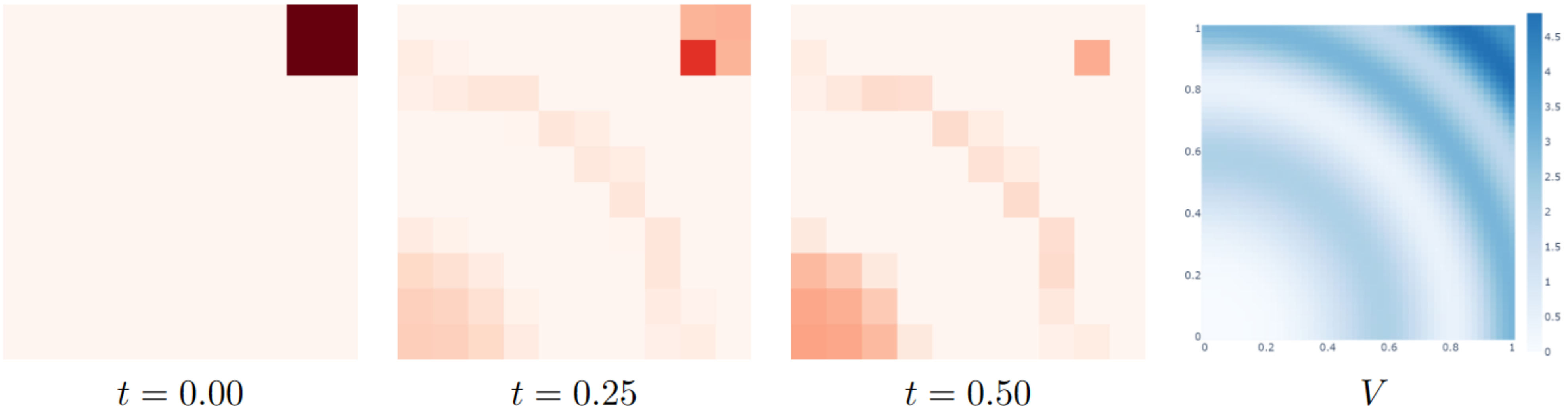
# Examples

**Proposition.** If $X = \mathbb{R}^d$, $V$ convex, $c$ quadratic cost and $\mu_0 = \delta_{x_0}$ then $\mu_t = \delta_{x_t}$ with
$$\dot{x}_t \in -\partial V(x_t).$$
(Same as Wasserstein GF).




But if $V$ not convex there can be **teleportation**!

Here a **Lagrangian** discretization won't work.



# Link with the Sinkhorn JKO scheme

**Sinkhorn JKO:**

$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu, \mu_k^\tau)}{2\tau}.$$

with $S_\varepsilon$ Sinkhorn divergence and $E(\mu) = \int V \, \mathrm{d}\mu$ **potential energy**.

**Sinkhorn flow** in the $b$-variable:

$$\dot{b}_t + \frac{2}{\varepsilon}(V - V^*)\, b_t + p_t = 0.$$

with $b_t = \exp(-f_{\mu_t}/\varepsilon)$.

# Link with the Sinkhorn JKO scheme

**Sinkhorn JKO:**

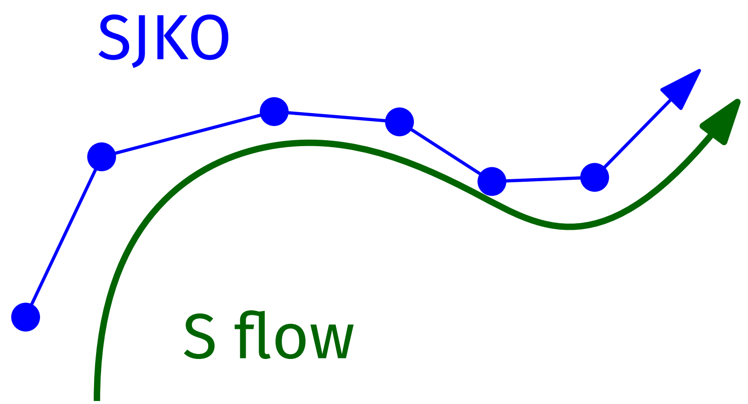
$$\mu_{k+1}^\tau \in \arg\min_\mu \; E(\mu) + \frac{S_\varepsilon(\mu, \mu_k^\tau)}{2\tau}.$$

with $S_\varepsilon$ Sinkhorn divergence and $E(\mu) = \int V \,\mathrm{d}\mu$ **potential energy**.

**Sinkhorn flow** in the $b$-variable:

$$\dot{b}_t + \frac{2}{\varepsilon}(V - V^*)\, b_t + p_t = 0.$$

with $b_t = \exp(-f_{\mu_t}/\varepsilon)$.

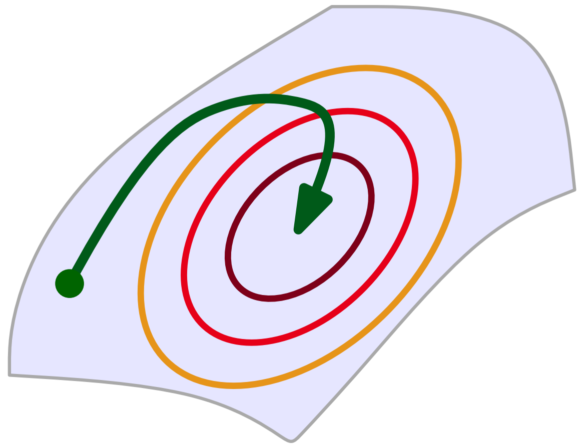


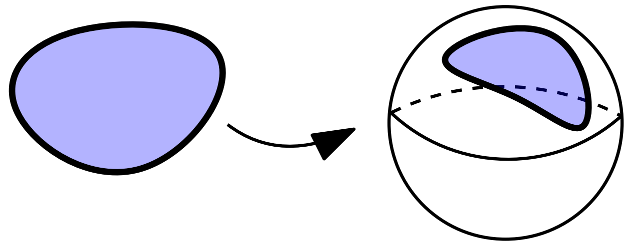

**Proposition.** If $X$ is a **finite set**, the solutions of the Sinkhorn JKO scheme, properly interpolated in time, converge to the Sinkhorn flow as $\tau \to 0$ in $C([0,T], \mathcal{P}(X))$.

# Future works

**What I have not presented:**

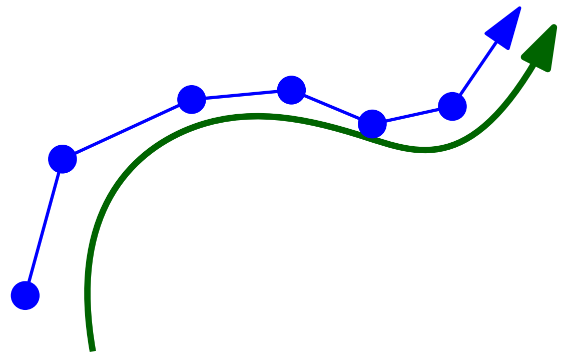
- Explicit computations for Gaussians, two points space,
- Proof that Sinkhorn divergence is not jointly convex,
- Proof that Sinkhorn divergence is not a metric.





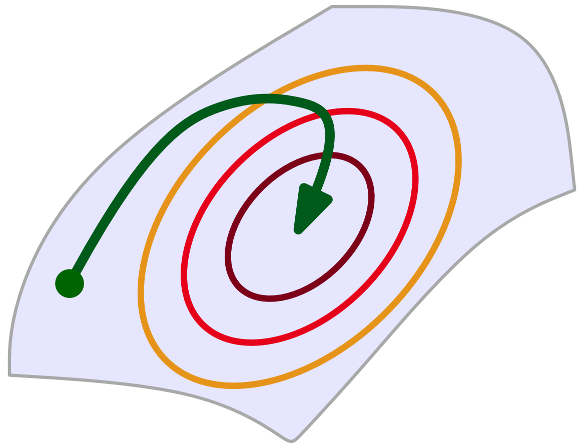
**Some topics we are working on:**

- Extend the convergence SJKO $\to$ Sinkhorn flow,
- Numerical approximation of geodesics,
- Limit $\varepsilon \to 0$ towards optimal transport,
- Homogeneization when space is refined.

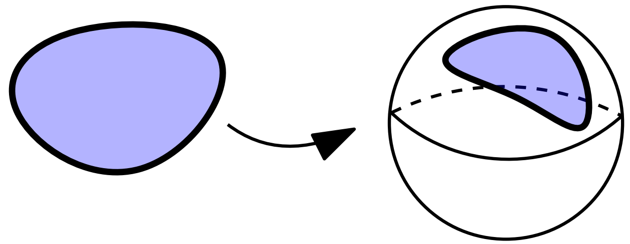


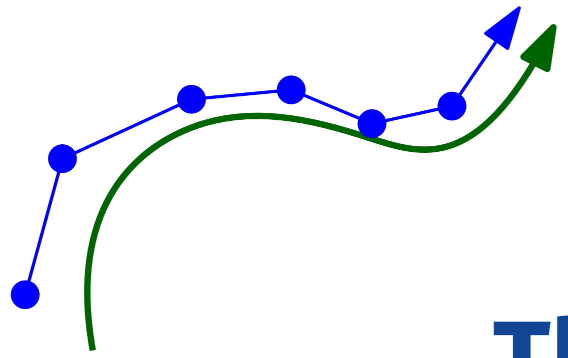
# Future works

## What I have not presented:



- Explicit computations for Gaussians, two points space,
- Proof that Sinkhorn divergence is not jointly convex,
- Proof that Sinkhorn divergence is not a metric.





## Some topics we are working on:

- Extend the convergence SJKO $\to$ Sinkhorn flow,
- Numerical approximation of geodesics,
- Limit $\varepsilon \to 0$ towards optimal transport,
- Homogeneization when space is refined.

# Thank you for your attention