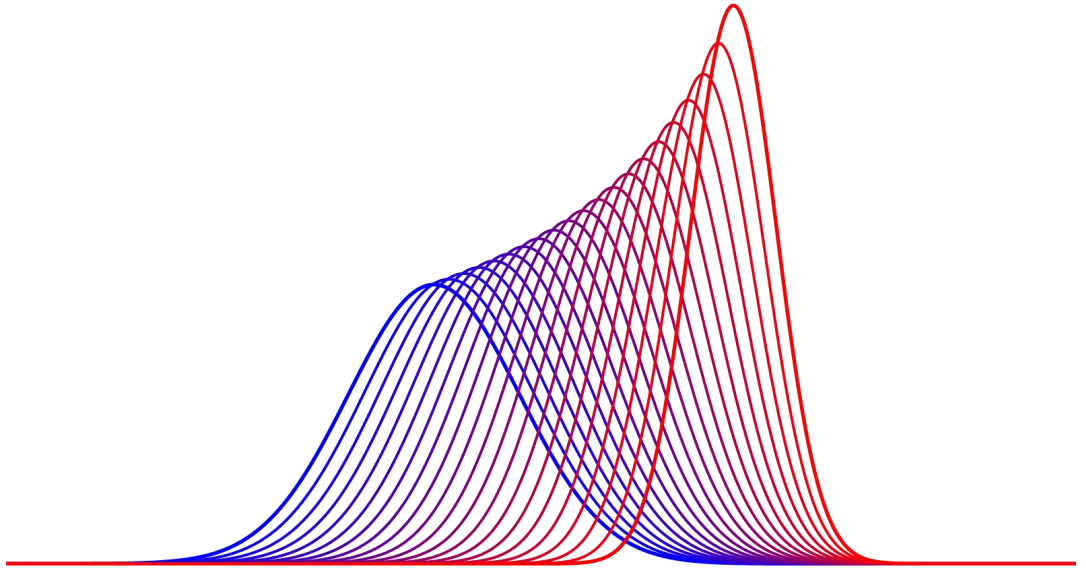
An introduction to Optimal Transport

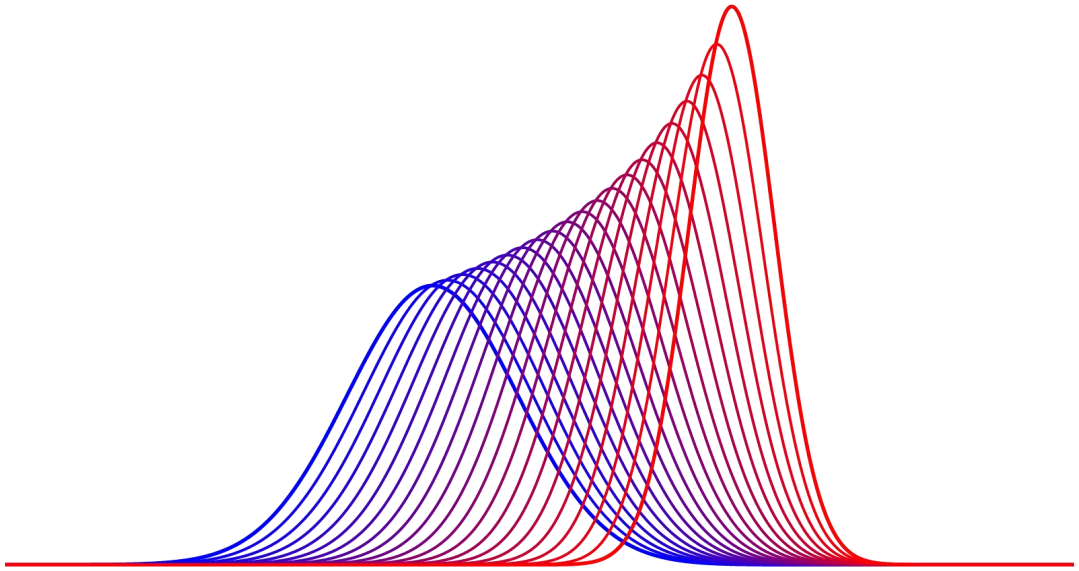
- Foundamental and Advanced Topics

Bangyan Liao

liaobangyan@westlake.edu.cn



WESTLAKE UNIVERSITY



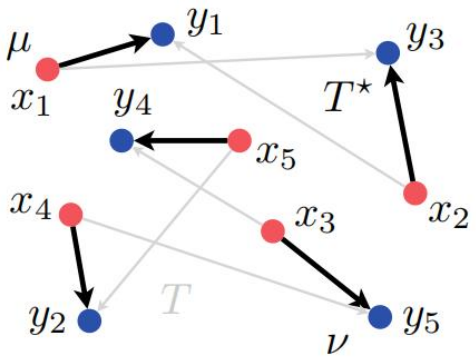
Motivation

What is the optimal transport?

Why we care about it?

Optimal Transport
Motivation

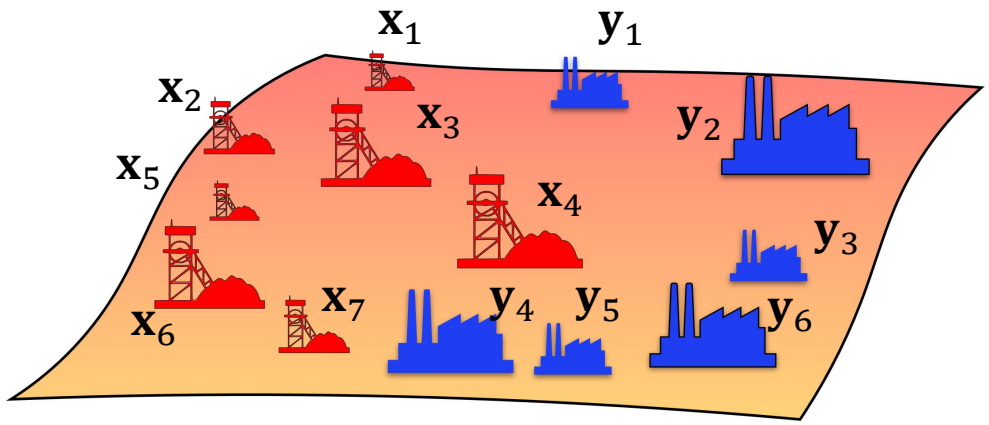
The **Optimal Transport theory** focuses on **finding** the **optimal mapping** between different **probability measures**.



Discrete Measure

$$\min_{T}\left\{\sum_{i} c(x_i, T(x_i)) \;\middle|\; \forall j \in [\![m]\!], \mathbf{y}_j = \sum_{i:T(x_i)=y_j} \mathbf{x}_i\right\}$$

Formulation

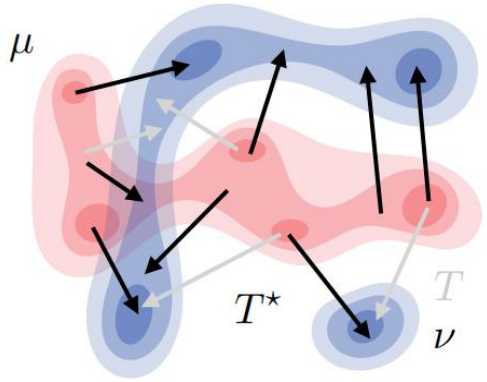


Example

Considering many mines and factories with different scales, how to find the transport map with the lowest transportation cost?

Optimal Transport
Motivation

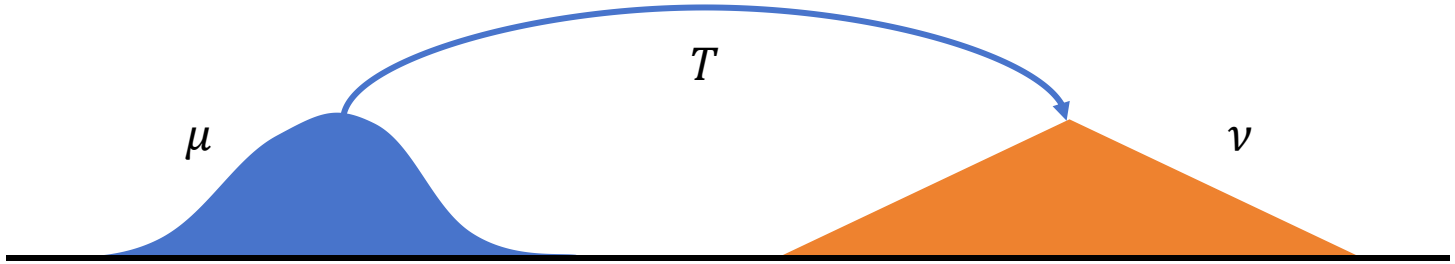
The **Optimal Transport theory** focuses on **finding** the **optimal mapping** between different **probability measures**.



Continuous Measure

$$\min_{T} \left\{ \int_x c(x, T(x))\, da(x) \,\middle|\, T_{\#}\mu = \nu \right\}$$

Formulation



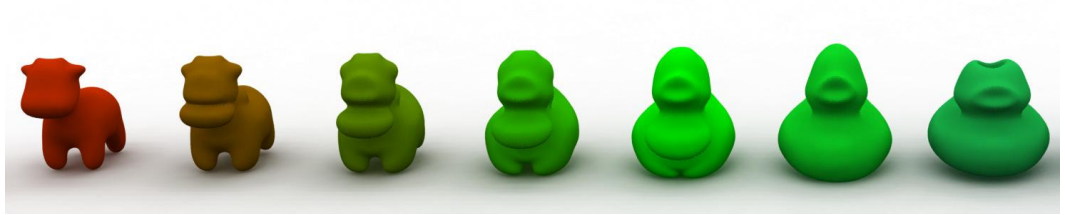
Example

Given two piles of sand,
how to find a minimum cost transportation
(optimal transport map) ?

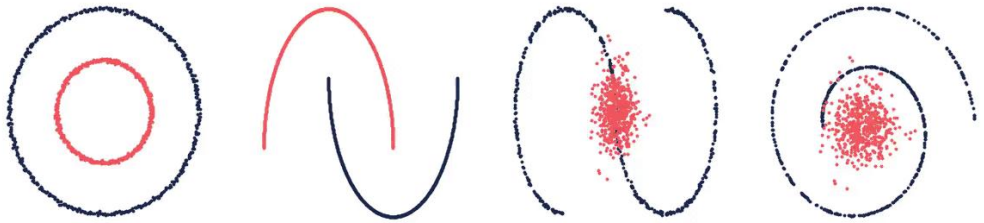
Various Measures in Optimal Transport

Motivation

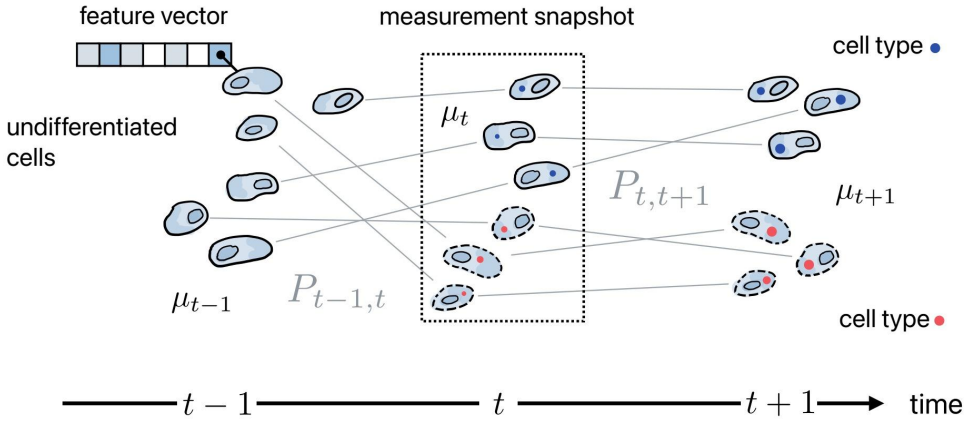
Shapes as measures



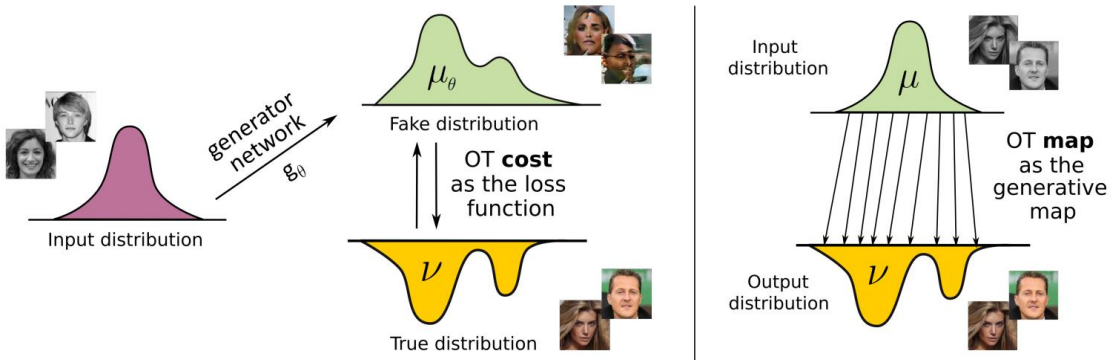
Particles as measures



Cells as measures



Images as measures



(a) OT cost as the loss for the generative model.

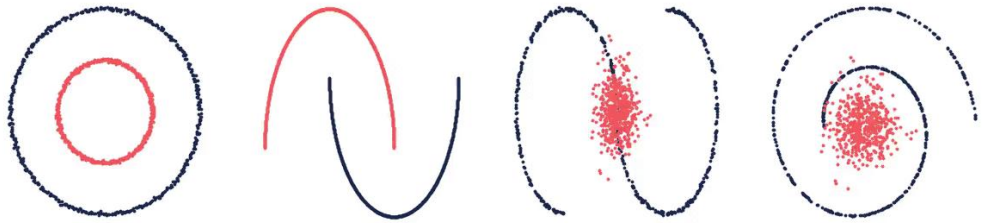
(b) OT map as the generative model.

Various Applications in Optimal Transport

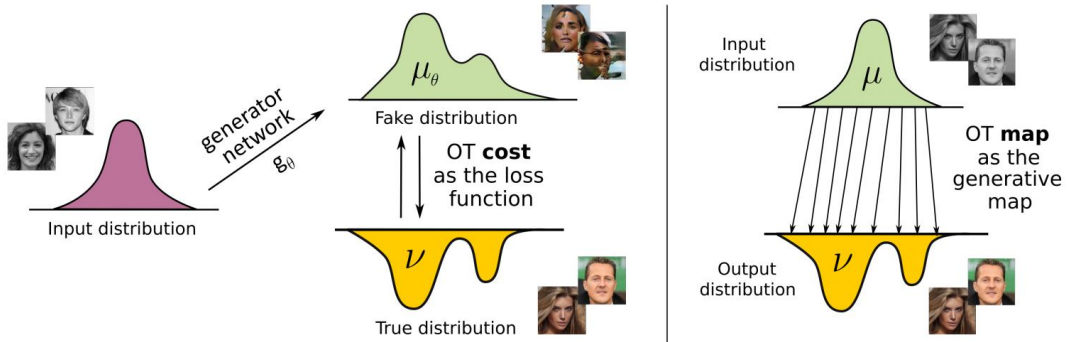
Motivation

Image Domain Transformation

Particle Interpolation



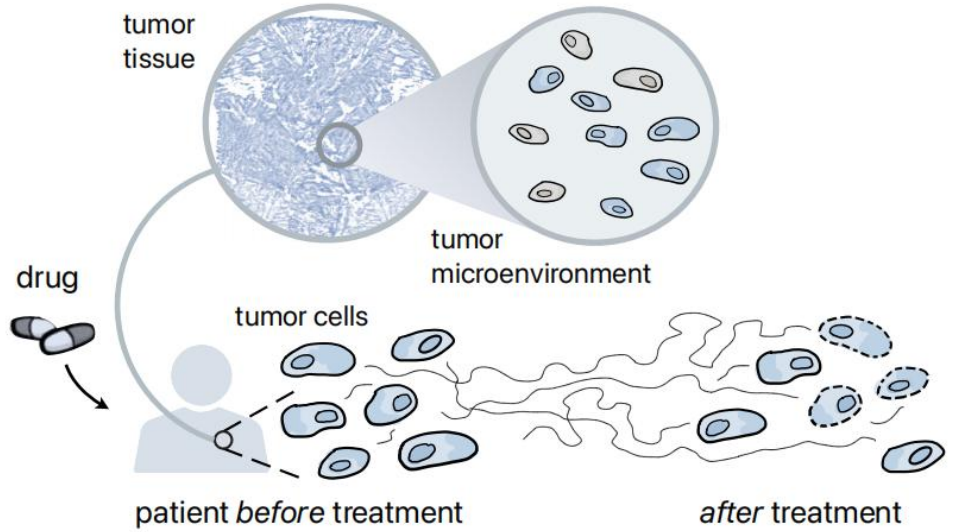
Deep Generative Model



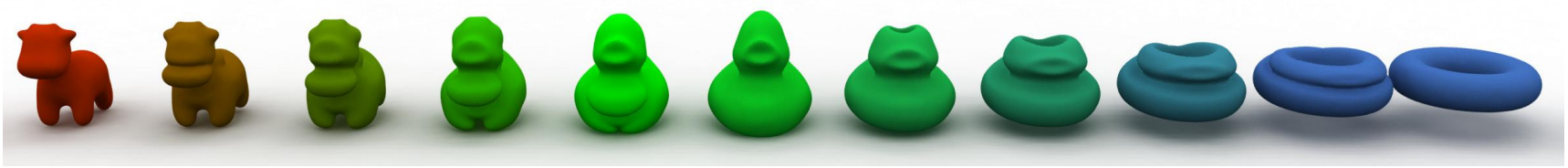
(a) OT cost as the loss for the generative model.

(b) OT map as the generative model.

Medicine and Biology



Shape Interpolation



Motivation

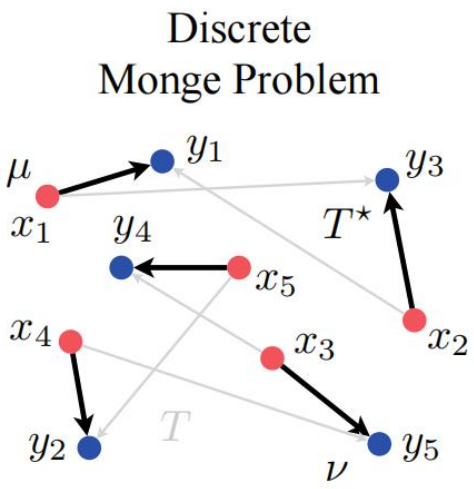
Foundamental topics in Optimal Transport

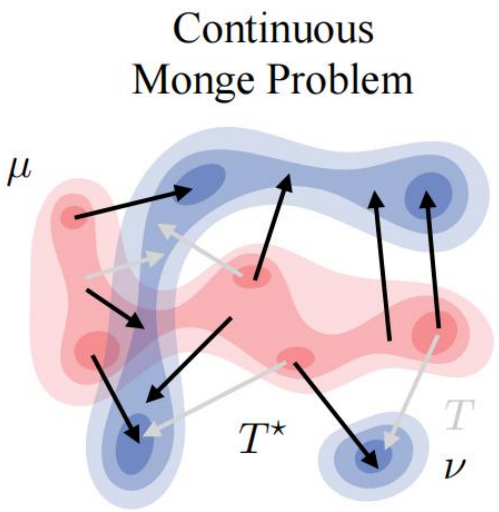
- Monge Problem and Kantorovich Relaxation
- Wasserstein Distance and Brenier Theorem
- Kantorovich Duality Theory and C-transform
- Semi-dual Formulation and Gangbo-McCann Theorem
- Entropy Regularization and Sinkhorn Algorithm
- Extensions of OT

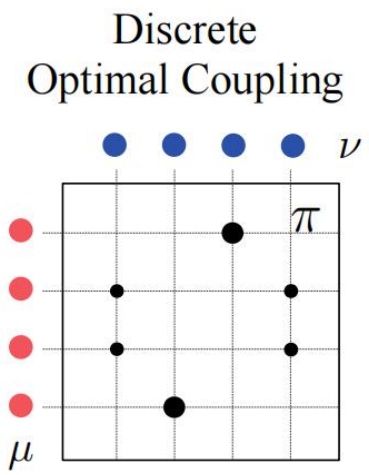
Advanced topics in Optimal Transport

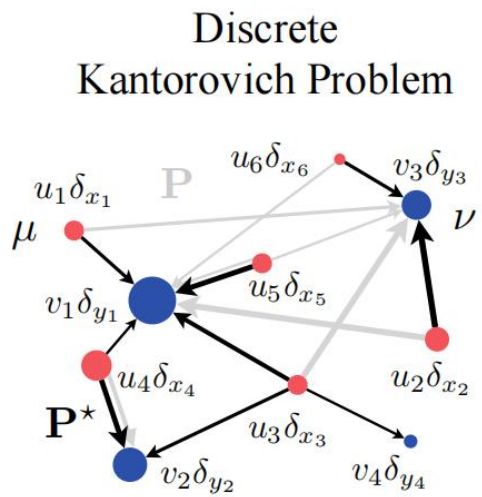
- The Riemannian Structure of Optimal Transport
- Wasserstein Gradient Flow and JKO Scheme
- Introduction to the Schrödinger Bridge Problem

Outline









Foundamental topics in Optimal Transport

- Monge Problem and Kantorovich Relaxation
- Wasserstein Distance and Brenier Theorem
- Kantorovich Duality Theory and C-transform
- Semi-dual Formulation and Gangbo-McCann Theorem
- Entropy Regularization and Sinkhorn Algorithm
- Extensions of OT

Outline Story

Foundamental topics in Optimal Transport

Monge Problem and Kantorovich Relaxation

We will introduce the **basic** and **relaxed** formulation of OT.

Wasserstein Distance and Brenier Theorem

Based on which, the **wasserstein distance** will be introduced and the **Brenier Theorem** gives us the tight relaxation.

Kantorovich Duality Theory and C-transform

We will introduce the **duality** formulation of OT.

Semi-dual Formulation and Gangbo-McCann Theorem

We will introduce the **semi-dual** formulation of OT.

Entropy Regularization and Sinkhorn Algorithm

We will introduce the most important **regularization** and **algorithm**.

Extensions of OT

Prerequisite - Probability Measure $\mathcal{M}_+^1$

Foundamental topics in Optimal Transport

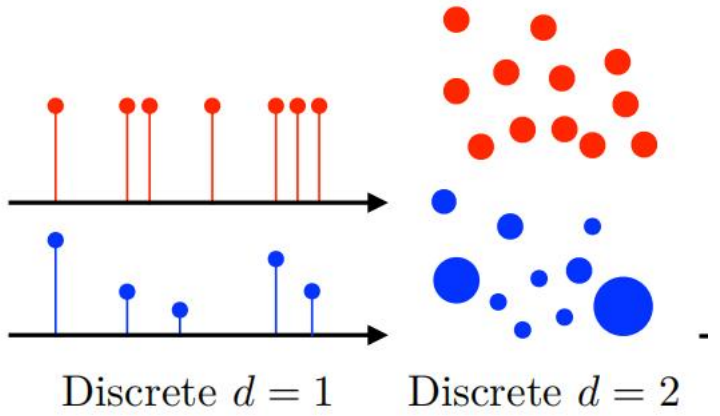
Discrete Probabilistic Measure

Definition (informal): A discrete probabilistic measure $a \in \mathcal{M}_+^1(\mathcal{X})$ is a collection of weighted discrete points:

$$a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \qquad s.t.\ \mathbf{a}_i > 0, \sum_{i=1}^{n} \mathbf{a}_i = 1$$

where δ_x is a Dirac function at position x. The integration of continuous function $f \in \mathcal{C}(\mathcal{X})$ against a discrete probabilistic measure computes a sum

$$\int_{\mathcal{X}} f(x)\, da(x) = \sum_{i=1}^{n} \mathbf{a}_i f(x_i)$$



Example

Remark.

1. in the rest of talk, we will denote the probabilistic measure as $\mathcal{M}_+^1(\mathcal{X})$ and continuous function family as $\mathcal{C}(\mathcal{X})$.
2. In practice, $a \in \mathcal{M}_+^1(\mathcal{X})$ is just an abstract symbol, we can use vector $\mathbf{a}$ and tuple $\{x_i\}_{i=1}^n$ to represent this probabilistic measure.

Prerequisite - Probability Measure $\mathcal{M}_+^1$

Foundamental topics in Optimal Transport

General Probabilistic Measure

Definition (formal): In mathematics, a probability measure is a real-valued function defined on a set of events in a σ-algebra that satisfies measure properties such as countable additivity. -- Wikipedia

Definition (informal): Probability measure can be viewed as representing the distribution of random variables. A random variable X on $\mathcal{X}$ is actually a map X: $\Omega \to \mathcal{X}$ from some abstract probability space $(\Omega, \mathbb{P})$ to the space $\mathcal{X}$ (e.g. $\mathbb{R}^d$). Its distribution a is the probability measure $a \in \mathcal{M}_+^1(\mathcal{X})$ such that:

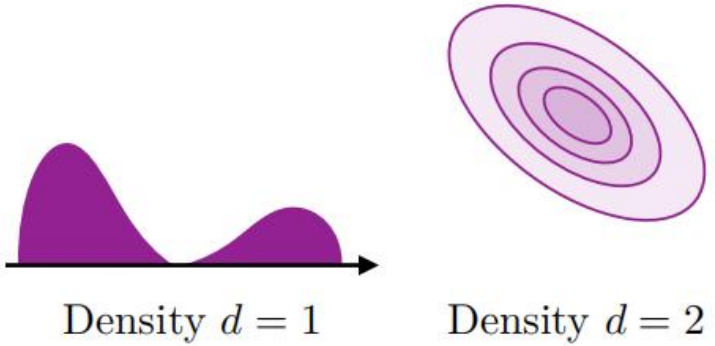
$$\mathbb{P}(X \in A) = a(A) = \int_A da(x)$$

and

$$\mathbb{P}(X \in \mathcal{X}) = a(\mathcal{X}) = \int_{\mathcal{X}} da(x) = 1.$$

If the probability measure a **have a density** $da(x) = \rho_a dx$ with respect to the Lebesgue measure on $\mathcal{X} = \mathbb{R}^d$, then for any continuous function $h \in \mathcal{C}(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} h(x)\, da(x) = \int_{\mathbb{R}^d} h(x)\, \rho_a(x) dx\,.$$



Example

Remark.

1. in the rest of talk, we will use $\int_{\mathbb{R}^d} h(x)\, da(x)$ to denote the integration over some probability measure (distribution).
2. this formulation can also represent the discrete probabilistic measure.

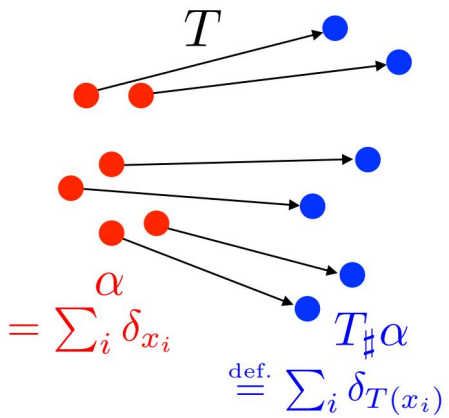
Prerequisite - Pushforward (Discrete Measure Version)

Foundamental topics in Optimal Transport

Push-forward operator and measure

Definition: For a continuous map $T: \mathcal{X} \to \mathcal{Y}$, we can define the corresponding **push-forward** operator $T_{\#}: \mathcal{M}_+^1(\mathcal{X}) \to \mathcal{M}_+^1(\mathcal{Y})$ and the **push-forward** measure $\beta = T_{\#}a \in \mathcal{M}_+^1(\mathcal{Y})$ of some $a \in \mathcal{M}_+^1(\mathcal{X})$.

For **discrete measure** $\beta = T_{\#}a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{T(x_i)}$



"$T_{\#}$ moving the position of all the points and weights in the support of the measure"

Prerequisite - Pushforward (Arbitrary Measure Version)

Foundamental topics in Optimal Transport

Push-forward operator and measure

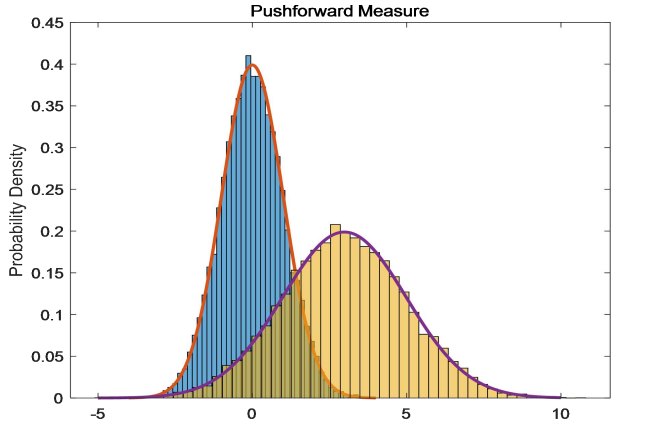
Definition: For a continuous map $T: \mathcal{X} \to \mathcal{Y}$, we can define the corresponding **push-forward** operator $T_{\#}: \mathcal{M}_+^1(\mathcal{X}) \to \mathcal{M}_+^1(\mathcal{Y})$ and the **push-forward** measure $\beta = T_{\#}a \in \mathcal{M}_+^1(\mathcal{Y})$ of some $a \in \mathcal{M}_+^1(\mathcal{X})$.

For **arbitrary measure**, the map $T: \mathcal{X} \to \mathcal{Y}$ must satisfies

$$\forall h \in \mathcal{C}(\mathcal{Y}), \int_{\mathcal{Y}} h(y)\, d\beta(y) = \int_{\mathcal{X}} h(T(x))\, da(x)$$

Equivalently, for any measurable set $B \subset \mathcal{Y}$, one has

$$\beta(B) = a(\{x \in \mathcal{X}: T(x) \in B\}) = a(T^{-1}(B))$$



$\boldsymbol{a = \mathcal{N}(0,1)} \quad \boldsymbol{\beta = T_{\#}a = \mathcal{N}(3,2)}$

$T(X) = 2X + 3$

Remark.

1. $\boldsymbol{T_{\#}}$ preserves positivity and total mass
2. push-forward can be seen as a kind of mapping (like function)
3. push-forward operator is "**linear**"

Prerequisite - Pushforward (Arbitrary Measure Version)

Foundamental topics in Optimal Transport

Push-forward for multivariate densities

Explicitly doing the push-forward T for measures with densities (ρ_a, ρ_β) on $\mathbb{R}^d$ (assuming is smooth and bijective) shows that a push-forward T acts on densities linearly as a change of variables in the integration formula.

$$\rho_a(x) = |det(T'(x))|\, \rho_\beta(T(x))$$

where $T'(x) \in \mathbb{R}^{d\times d}$ is the Jacobian matrix of T. This is similar to the change of variable theorem in integration.

Identity Map as push-forward

Similar to the identity matrix, we can also define the identity map as Id: $Id_{\#}a = a$.

Monge Problem and Kantorovich Relaxation (Discrete Measure Version)

Foundamental topics in Optimal Transport

Monge Problem between measures

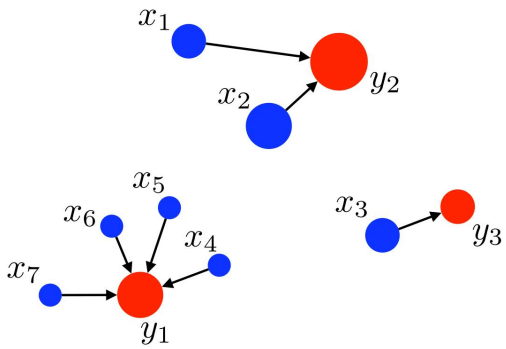
Definition: Given the discrete measures $a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$

the **Monge problem** seeks a **map** $T: \{x_1, \dots, x_n\} \to \{y_1, \dots, y_n\}$ which minimizes

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \,\middle|\, T_{\#}a = \beta \right\},$$

where c is the **ground cost** (e.g. $c = (x_i - T(x_i))^2$) and $T_{\#}a = \beta$ can be understanded as

$$\forall j \in [\![m]\!], \mathbf{b}_j = \sum_{i:T(x_i)=y_j} \mathbf{a}_i$$

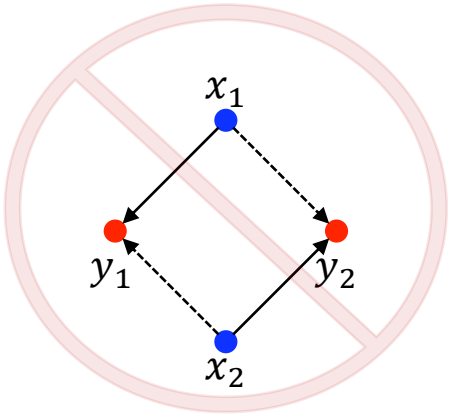


Example

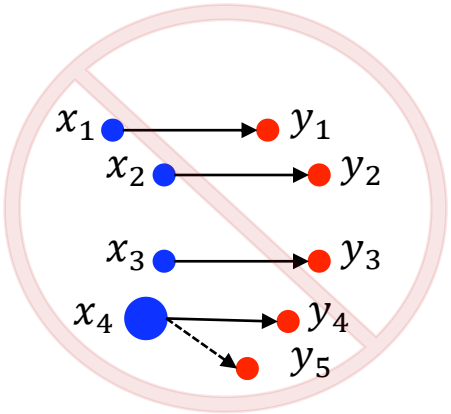
Remark.

1. Monge Problem's uniqueness and existence is not guaranteed (push forward operator dosen't allow mass splitting)
2. Monge Problem is a combinatorical problem and is non-convex

Can we make this problem easier to be solved?



The Monge map may not be unique.



The Monge map may not exist.

Monge Problem and Kantorovich Relaxation (Discrete Measure Version)

Foundamental topics in Optimal Transport

Kantorovich Relaxation Problem between measures

Definition: Given the discrete measures $a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$

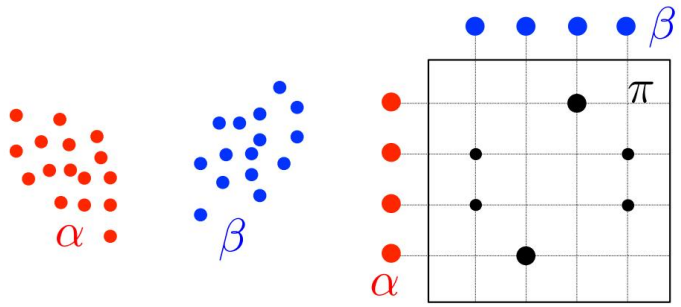
the **Kantorovich Relaxation problem** seeks a **plan(coupling)** $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_i)}$

which minimizes

$$\mathcal{L}_c(a, \beta) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def}}{=} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}$$

where

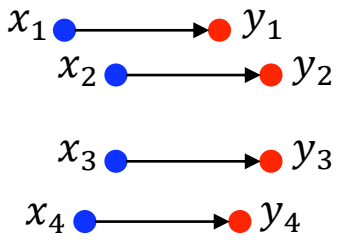
$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{ \mathbf{P} \in \mathbb{R}_{+}^{n \times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \}$$



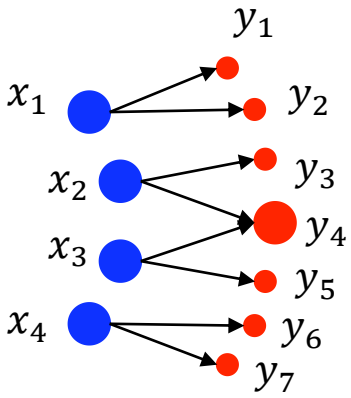
Example

Remarks.

1. Kantorovich Relaxation allows mass spliting
2. The coupling exists under mild condition and is unique if it is strongly convex
3. Monge map is a special type of coupling $\pi = (Id \times T(\cdot))$
4. Kantorovich Relaxation is a convex constrainted problem



Monge Problem



Kantorovich Relaxation

Monge Problem and Kantorovich Relaxation (Arbitrary Measure Version)

Foundamental topics in Optimal Transport

Monge Problem between arbitary measures

Definition: Given the measures $a \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, and the ground cost c: $(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, the **Monge problem** seeks a **map** T: $\mathcal{X} \to \mathcal{Y}$ which minimizes

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x))\, da(x) \,\middle|\, T_{\#}a = \beta \right\}$$

Kantorovich Relaxation between arbitary measures

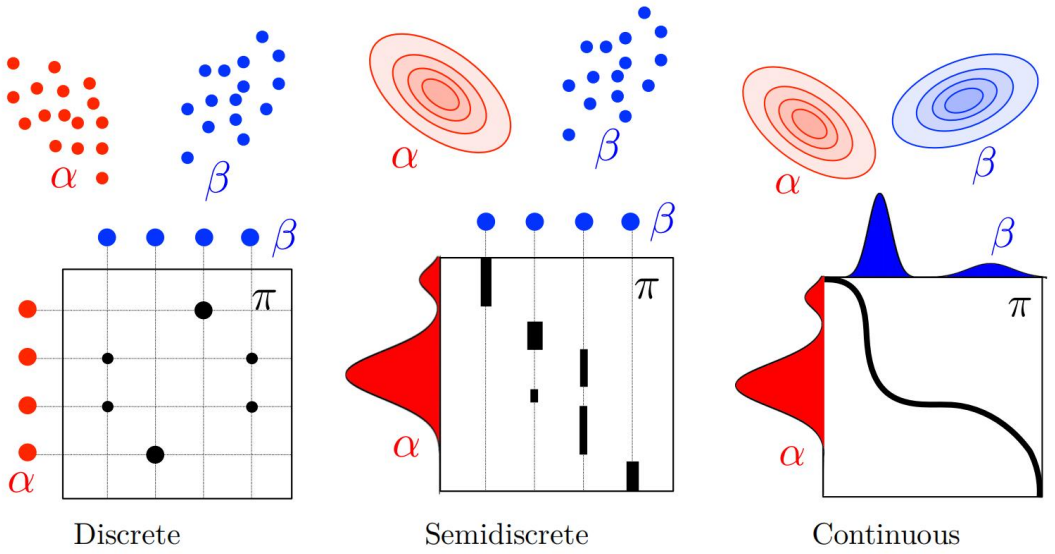
Definition: Given the measures $a \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the **Kantorovich Relaxation problem** seeks a **plan(coupling)** $\pi \in \mathcal{U}(a, \beta)$ which minimizes

$$\mathcal{L}_c(a, \beta) \stackrel{\text{def}}{=} \min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)\, d\pi(x, y) \right\},$$

where

$$\mathcal{U}(a, \beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$

Here, $P_{\mathcal{X}\#}$ and $P_{\mathcal{Y}\#}$ are the push-forwards of the projections $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$ respectively.



Example

Can we define $\mathcal{L}_c(a, \beta)$ as a distance between distributions?
-> **Wasserstein Distance**

Outline Overview

Foundamental topics in Optimal Transport

Monge Problem

$$\min_{T} \left\{ \int_{\mathcal{X}} c(x, T(x))\, da(x) \,\middle|\, T_{\#}a = \beta \right\}$$

Kantorovich Relaxation

$$\min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) \right\}$$

$$\mathcal{U}(a,\beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_{+}^{1}(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$

Kantorovich Relaxation

Wasserstein Distance and Brenier Theorem

Foundamental topics in Optimal Transport

Wasserstein Distance

We assume $\mathcal{X} = \mathcal{Y}$ and $c(\cdot\,,\cdot)$ is a distance on $\mathcal{X}$, i.e.
$1) c(x,y) = c(y,x) \geq 0$;
$2) c(x,y) = 0$ if and only if $x = y$;
$3)\ \forall (x,y,z) \in \mathcal{X}^3, c(x,y) \leq c(x,z) + c(y,z)$

Then, the wasserstein distance between measures can be defined as

$$\mathcal{W}_p(a,\beta) \stackrel{\text{def}}{=} \left(\min_{\pi \in \mathcal{U}(a,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)^p \, d\pi(x,y) \right)^{1/p}$$

If $p = 2$, we can get some interesting properties
-> **Brenier Theorem**

Remarks.
1. $\mathcal{W}_p(a,\beta)$ can be proved to satisfy the distance property.
2. $\mathcal{W}_p(a,\beta)$ can be defined on arbitrary measures (discrete-discrete, discrete-continuous, continuous-continuous)

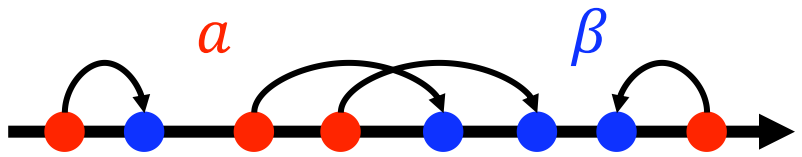
Brenier Theorem

Definition: In the case $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$, and at least one of the two input measures has a **density** with respect to the Lebesgue measure, then the optimal **plan(coupling)** $\pi \in \mathcal{U}(a, \beta)$ is **unique** and we can express the plan as $\pi = (Id, T)_{\#}a$, i.e.

$$\forall h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x, y) d\pi(\mathrm{x}, \mathrm{y}) = \int_{\mathcal{X}} h(x, T(x))\, da(\mathrm{x})$$

Furthermore, this map T is uniquely defined as the **gradient of a convex function** $T(x) = \nabla\varphi(x)$, where φ is the unique convex function such that $(\nabla\varphi)_{\#}a = \beta$.

Remark. according to the monotonicity of the gradient of a convex function f i.e.
$\langle x - y, f'(x) - f'(y)\rangle \geq 0, \quad \forall x, y \in \mathbb{R}^d$
if $d = 1$



Benefit.
1. Brenier theorem guarantees to produce unique optimal plan (relaxation is tight).
2. Brenier theorem shows that $T(x) = \nabla\varphi(x)$, we can parameterize this plan easily throuth deep network (e.g. ICNN).

It seems that we have not guarantee the push forward property (mass preserving).
-> **Monge-Ampere equation**

Monge-Ampere equation

Definition: For measures with densities, one obtains that φ is the **unique convex function** which solves the following **Monge-Ampere** equation:

$$det(\partial^2\varphi(x))\ \rho_\beta(\nabla\varphi(x)) = \rho_a(x)$$

where $\partial^2\varphi(x)$ is the Hessian of φ.

Proof.

Recall the push forward opterator $\beta = T_\# a$ for measures with density

$$\rho_a(x) = |det(T'(x))|\ \rho_\beta(T(x))$$

Combining it with the **Brenier Theorem** $(\nabla\varphi)_\# a = \beta$.

■

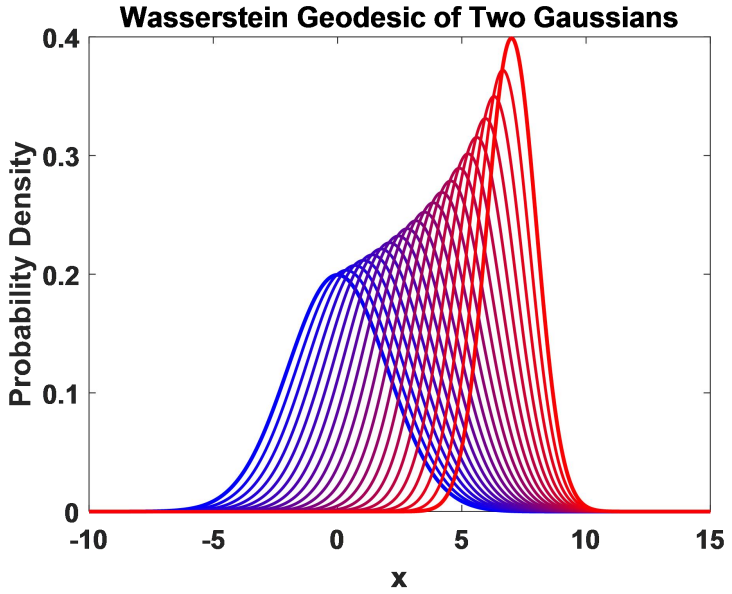
Benefit.

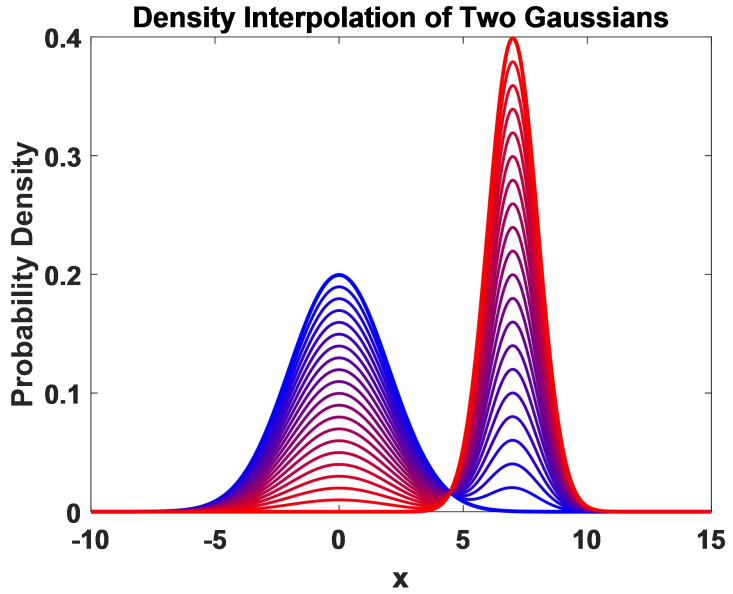
1. The Brenier Theorem guarantees the uniqueness and exsistence.
2. The $\rho_a(x) = |det(T'(x))|\ \rho_\beta(T(x))$ guarantees the push forward property.
3. The final equation is a PDE.

Wasserstein Geodesic (McCann's Interpolation) (Displacement Interpolation)

Definition: In the case where there exists an optimal transport map $T: \mathcal{X} \to \mathcal{Y}$ with $T_{\#}a_0 = a_1$, then we can interpolate the intermediate probabilistic measure a_t through the McCann's interpolation:

$$a_t = ((1-t)Id + tT)_{\#}a_0, \quad t \in [0,1]$$





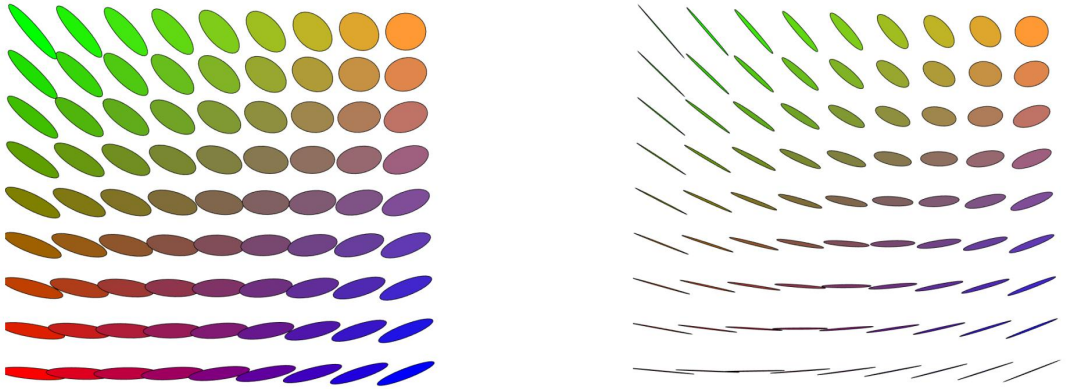
Example

Remarks.
We will see in the next section that, this interpolation path is actually a geodesic path of the Riemannian structure of the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$.

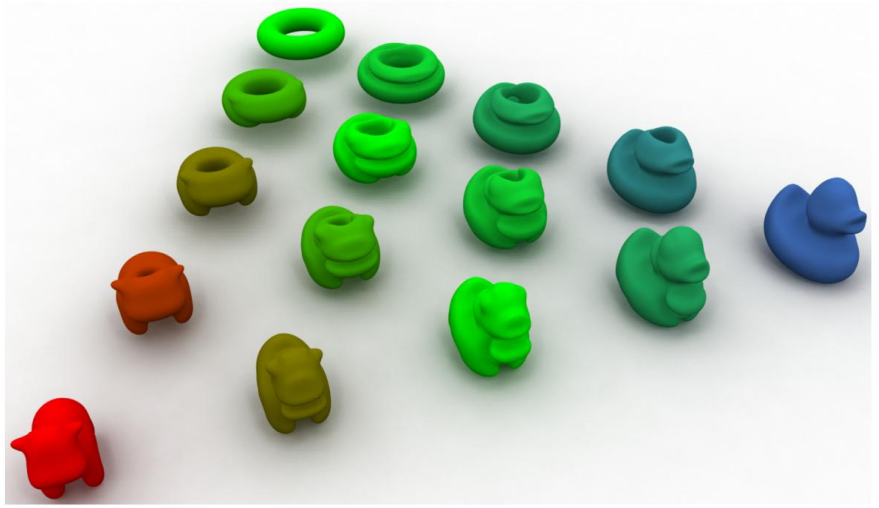
Wasserstein Barycenters

Definition: Given a set of input measure $(\beta_s)_{[1,\,S]}$ defined on some space $\mathcal{X}$, the barycenter problem can be defined as:

$$\min_{a \in \mathcal{M}_+^1(\mathcal{X})} \sum_{s=1}^{S} \lambda_s \mathcal{W}_p(a, \beta_s)$$



Example of 2D Gaussian

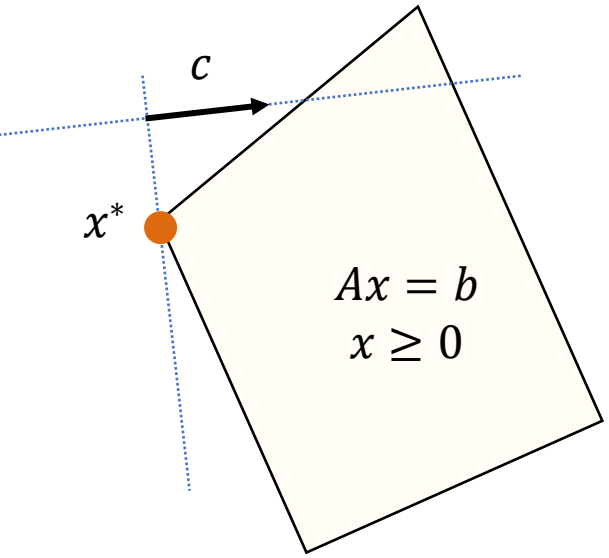


Example of shape

Remarks.
The barycenter problem can be seen as a generazed version of displacement interpolation.

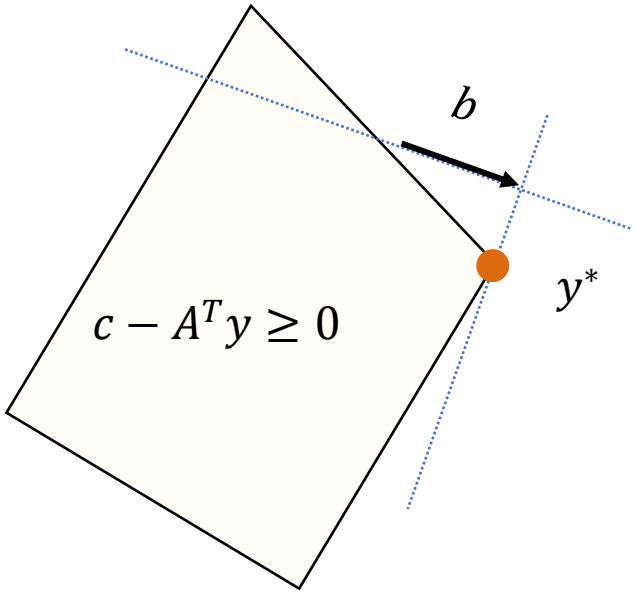
Primal Linear Programming

$$\min_{x\in\mathbb{R}^n} \langle c, x\rangle$$
$$s.t. \quad Ax = b$$
$$x \geq 0$$



Dual Linear Programming

$$\max_{y\in\mathbb{R}^m} \langle b, y\rangle$$
$$s.t. \quad c - A^T y \geq 0$$



Proof.

The primal problem can be reimplemented as

$$\min_{x\in\mathbb{R}^n} \max_{\mu\in\mathbb{R}^m,\lambda\in\mathbb{R}^n_+} \langle c, x\rangle + \langle Ax - b, \mu\rangle + \langle \lambda, -x\rangle$$
$$\geq \max_{\mu\in\mathbb{R}^m,\lambda\in\mathbb{R}^n_+} \min_{x\in\mathbb{R}^n} \langle c, x\rangle + \langle Ax - b, \mu\rangle + \langle \lambda, -x\rangle$$
$$= \max_{\mu\in\mathbb{R}^m,\lambda\in\mathbb{R}^n_+} \min_{x\in\mathbb{R}^n} \langle c + A^T\mu - \lambda, x\rangle - \langle b, \mu\rangle$$

notice that

$$\min_{x\in\mathbb{R}^n} \langle c + A^T\mu - \lambda, x\rangle - \langle b, \mu\rangle$$
$$= \begin{cases} -\langle b, \mu\rangle & if\ c + A^T\mu - \lambda = 0 \\ -\infty & otherwise \end{cases}$$

then $\max_{\mu\in\mathbb{R}^m,\lambda\in\mathbb{R}^n_+} \min_{x\in\mathbb{R}^n} \langle c + A^T\mu - \lambda, x\rangle - \langle b, \mu\rangle \implies$

$$\max_{y\in\mathbb{R}^m} \langle b, y\rangle$$
$$s.t. \quad c - A^T y = \lambda \geq 0$$

where $y = -\mu$.

■

Prerequisite - Linear Programming

Foundamental topics in Optimal Transport

Primal Linear Programming

$$\min_{x\in\mathbb{R}^n} \langle c, x\rangle$$
$$s.t. \quad Ax = b$$
$$x \geq 0$$

Dual Linear Programming

$$\max_{y\in\mathbb{R}^m} \langle b, y\rangle$$
$$s.t. \quad c - A^T y \geq 0$$

Strong Duality in Linear Programming i.e. $\langle c, x^*\rangle = \langle b, y^*\rangle$.

1. If one of problems is feasible, then strong duality holds.
2. If one of problems achieve finite optimal solution, then strong duality holds.
3. A primal feasible point x^* is an optimal solution if and only if there exists a dual feasible point y^* such that $\langle c, x^*\rangle = \langle b, y^*\rangle$

Complementary Slackness

If strong duality holds, we can get the complementary slackness.

$$x_i^*(c - A^T y^*)_i = 0, \forall i$$

Proof.

Just substitute the optimal variable into the last equation in the weak duality proof $\langle c - A^T y, x\rangle = 0$.

■

Weak Duality

For any pair of feasible primal variable x and dual variable y, we have

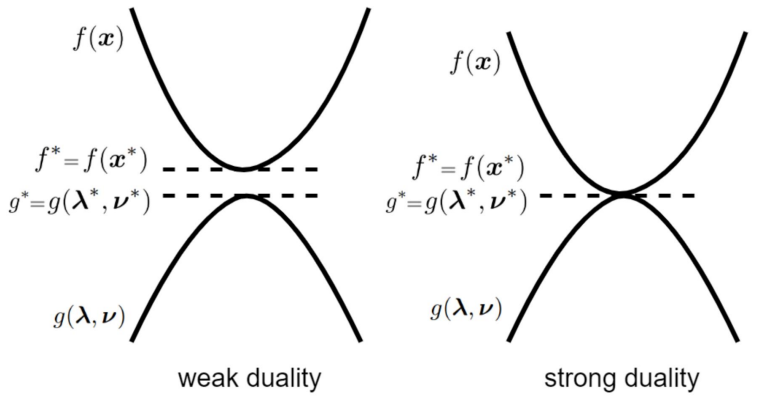
$$\langle c, x\rangle \geq \langle b, y\rangle$$

Proof.

$$\begin{aligned}&\langle c, x\rangle - \langle b, y\rangle\\ &= \langle c, x\rangle - \langle Ax, y\rangle\\ &= \langle c, x\rangle - \langle A^T y, x\rangle\\ &= \langle c - A^T y, x\rangle \geq 0\end{aligned}$$

■

Remark.



Kantorovich Duality Theory and C-transform (Discrete Measure Version)

Kantorovich Duality Theory

Definition: Given the discrete measures $a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$

the **Kantorovich duality problem** seeks two **dual variables** $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ which maximizes

$$\mathcal{L}_c(a, \beta) \stackrel{\text{def}}{=} \max_{(\mathbf{f},\mathbf{g}) \in \mathbf{R}(\mathbf{C})} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle$$

where the set of admissible dual variables is

$$\mathbf{R}(\mathbf{C}) \stackrel{\text{def}}{=} \left\{ (\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m \mid \forall (i,j) \in [\![n]\!] \times [\![m]\!], \mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{ij} \right\}$$

Proof.

Recall the Kantorovich relaxation problem

$$\mathcal{L}_c(a, \beta) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle$$

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \}$$

The lagangian function of Kantorovich relaxation problem is

$$\mathcal{L}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{C}, \mathbf{P} \rangle + \langle \mathbf{a} - \mathbf{P}\mathbf{1}_m, \mathbf{f} \rangle + \langle \mathbf{b} - \mathbf{P}^T \mathbf{1}_n, \mathbf{g} \rangle$$

Then the lagangian dual problem defines as

$$\max_{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m} \min_{\mathbf{P} \in \mathbb{R}_+^{n \times m}} \mathcal{L}(\mathbf{P}, \mathbf{f}, \mathbf{g})$$

$$\max_{(\mathbf{f},\mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle + \min_{\mathbf{P} \in \mathbb{R}_+^{n \times m}} \langle \mathbf{C} - \mathbf{f}\mathbf{1}_m{}^T - \mathbf{1}_n \mathbf{g}^T, \mathbf{P} \rangle$$

only when $(\mathbf{f}, \mathbf{g}) \in \mathbf{R}(\mathbf{C})$, $\min_{\mathbf{P} \in \mathbb{R}_+^{n \times m}} \langle \mathbf{C} - \mathbf{f}\mathbf{1}_m{}^T - \mathbf{1}_n \mathbf{g}^T, \mathbf{P} \rangle$ has finite value 0.

■

Remark.

1. If this problem has feasible solution and bounded below, then the strong duality holds.
2. The complementary slackness describes the relationship between the optimal primal and dual variables.

$$\mathbf{P}_{ij}^* (\mathbf{C}_{ij} - f_i^* + g_j^*) = 0$$

Kantorovich Duality Theory and C-transform (Arbitrary Measure Version)

Foundamental topics in Optimal Transport

Kantorovich Duality Theory

Definition: Given the measures $a \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the **Kantorovich duality problem** seeks two **dual potentials** $f \in \mathcal{C}(\mathcal{X})$ and $g \in \mathcal{C}(\mathcal{Y})$

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup_{(f,g)\in\mathcal{R}(\mathrm{c})} \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} g(y)\, d\beta(y)$$

where the set of admissible dual potentials is

$$\mathcal{R}(\mathrm{c}) \stackrel{\text{def}}{=} \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mid \forall (x,y), f(x) + g(y) \leq c(x,y)\}$$

Benefit.

1. Compared to the primal Kantorovich relaxation problem, the optimized variables in the dual problem have much smaller size.

Kantorovich Duality Theory and C-transform (Arbitrary Measure Version)

Foundamental topics in Optimal Transport

c-transform

Definition: Given the a function $f \in \mathcal{C}(\mathcal{X})$ and ground cost function $c: (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, the c-transforms are defined as:

$$\forall y \in \mathcal{Y}, \quad f^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - f(x)$$

for $g \in \mathcal{C}(\mathcal{Y})$, we can also define the corresponding $\bar{c}$-transform

$$\forall x \in \mathcal{X}, \quad g^{\bar{c}}(x) := \inf_{y \in \mathcal{Y}} c(x, y) - g(y)$$

Remark.
1. c-transform can be seen as a generalization of the famous Legendre transform (convex conjugate).

This c-transform gives us the best function when fixing another function, which shows that

$$\int_{\mathcal{X}} f \, da + \int_{\mathcal{Y}} g \, d\beta \leq \int_{\mathcal{X}} f \, da + \int_{\mathcal{Y}} f^c \, d\beta \leq \int_{\mathcal{X}} f^{c\bar{c}} \, da + \int_{\mathcal{Y}} f^c \, d\beta = \int_{\mathcal{X}} f^{c\bar{c}} \, da + \int_{\mathcal{Y}} f^{c\bar{c}c} \, d\beta$$

Remark.
1. similar to the Legendre transform, we have $f^c = f^{c\bar{c}c}$ which shows that this "alternative minimization" can not converge.

There are constraints between theses two **dual potentials** which may be hard to optimize. Can we use this c-transform to reduce constraints?
-> **Semi-dual formulation**

Overview

Foundamental topics in Optimal Transport

Monge Problem

$$\min_{T} \left\{ \int_{\mathcal{X}} c(x, T(x))\, da(x) \,\middle|\, T_{\#}a = \beta \right\}$$

Kantorovich Relaxation

$$\min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) \right\}$$

$$\mathcal{U}(a,\beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_{+}^{1}(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$

Kantorovich Duality

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} g(y)\, d\beta(y)$$

$$\mathcal{R}(c) \stackrel{\text{def}}{=} \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mid \forall (x,y), f(x) + g(y) \leq c(x,y)\}$$

Kantorovich Relaxation

Langrange Duality

Semi-dual Formulation and Gangbo-McCann Theorem

Foundamental topics in Optimal Transport

Semi-dual Formulation

Definition: Given the measures $a \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the **Semi-dual formulation** seeks a **dual potential** which maximise

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} f^c(y)\, d\beta(y)$$

where $f^c(y)$ is the c-transform defined as

$$\forall y \in \mathcal{Y},\quad f^c(y) := \inf_{x \in \mathcal{X}} c(x,y) - f(x)$$

Benefit.
This formulation transforms the constraints of two functions f and g into an implicit optimization problem. This formulation greatly simplifies the complexity of the problem, and we will see many theories based on this formulation.

Remark.
1. f^* is **necessarily** c-concave, where f is a c-concave function if exists $g \in \mathcal{C}(\mathcal{Y})$ s.t. $f = g^{\bar{c}}$.

Gangbo-McCann Theorem

Definition: Given the ground cost c: $(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, the relationship between the optimal transport **map** T^*: $\mathcal{X} \to \mathcal{Y}$ and the optimal **dual potential** f^* is given by the expression

$$T^*(x) = \nabla_x c(x, \cdot)^{-1} \circ \nabla f^*(x)$$

Proof.

According to the complementary slackness property, a point (x_0, y_0) in the support of the coupling, necessarily has such property:

$$\pi^*(x_0, y_0) > 0 \iff f^*(x_0) + g^*(y_0) = c(x_0, y_0)$$

Replacing g by the c-transform of f, we have

$$f^{c*}(y_0) = c(x_0, y_0) - f^*(x_0)$$

By definition of the c-transfrom, f^{c*} is given by

$$f^{c*}(y_0) = \inf_{x \in \mathcal{X}} c(x, y_0) - f^*(x)$$

this shows that x_0 is the minimizer of $c(x_0, y_0) - f^*(x_0)$, then according the first order optimality condition

$$\nabla_x (c(x_0, y_0) - f^*(x_0)) = 0 \implies \nabla_x c(x_0, y_0) = \nabla_x f^*(x_0)$$

Then after inversion, we have $y_0 = T^*(x_0) = \nabla_x c(x_0, \cdot)^{-1} \circ \nabla f^*(x_0)$

■

Semi-dual Formulation and Gangbo-McCann Theorem

Foundamental topics in Optimal Transport

Gangbo-McCann Theorem given $c(x,y) = \frac{1}{2}\|x-y\|_2^2$

Definition: Given the ground cost $c(x,y) = \|x-y\|_2^2$, the relationship between the optimal transport **map** $T^*: \mathcal{X} \to \mathcal{Y}$ and the optimal **dual potential** f^* is given by the expression

$$T^*(x) = x - \nabla f^*(x)$$

Proof.

Recall the optimality condition and $\nabla_x(\frac{1}{2}\|x-y\|_2^2) = x - y$

$$\nabla_x c(x_0, y_0) = \nabla_x f^*(x_0)$$
$$x_0 - y_0 = \nabla_x f^*(x_0)$$
$$y_0 = x_0 - \nabla_x f^*(x_0)$$

■

Remark.

This theorem implicitly prove the **Brenier theorem**.

Proof.

$$T^*(x) = x - \nabla f^*(x) = \nabla\left(\frac{1}{2}\|x\|_2^2 - f^*(x)\right)(x) = \nabla\varphi(x)$$

■

where $\varphi(x)$ is a convex function ($-f^*(x)$ is convex).

Outline Overview

Foundamental topics in Optimal Transport

Monge Problem

$$\min_{T} \left\{ \int_{\mathcal{X}} c(x, T(x))\, da(x) \,\middle|\, T_{\#}a = \beta \right\}$$

Kantorovich Relaxation

$$\min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) \right\}$$

$$\mathcal{U}(a,\beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_{+}^{1}(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$

Kantorovich Duality

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} g(y)\, d\beta(y)$$

$$\mathcal{R}(c) \stackrel{\text{def}}{=} \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mid \forall (x,y), f(x) + g(y) \leq c(x,y)\}$$

Semi-dual Formulation

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} f^c(y)\, d\beta(y)$$

$$\forall y \in \mathcal{Y},\ \ f^c(y) := \inf_{x\in\mathcal{X}} c(x,y) - f(x)$$

Kantorovich Relaxation

Langrange Duality

C-transform

Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport

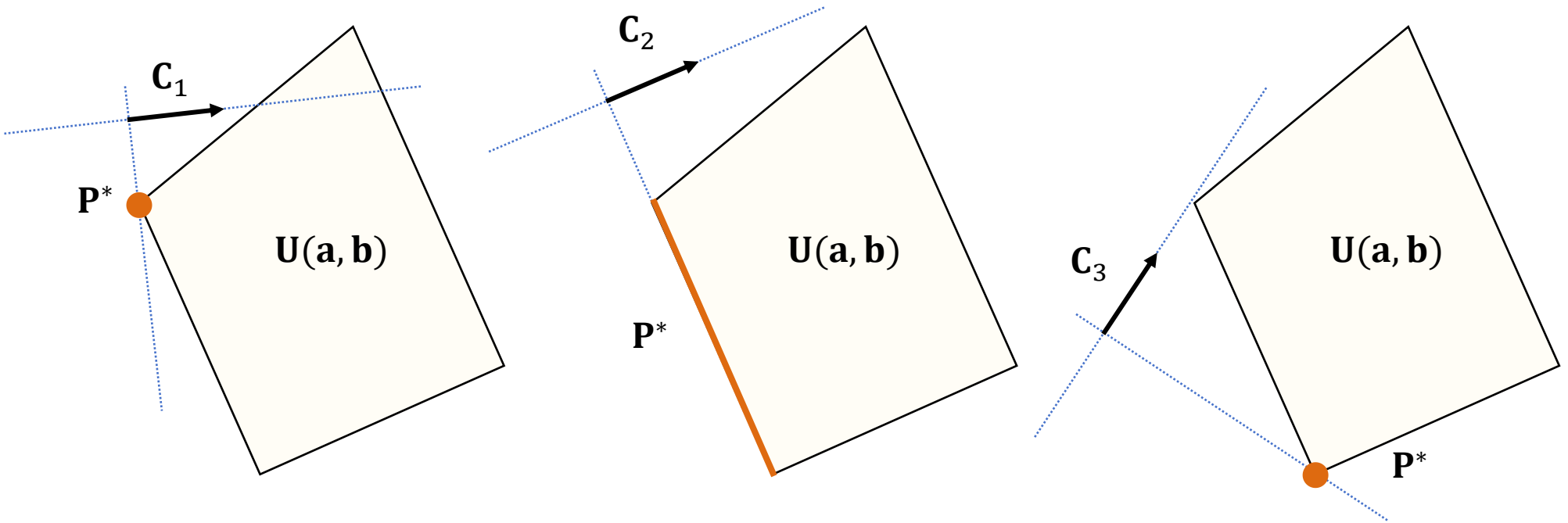
Why Entropy Regularization

Kantorovich Relaxation

$$\mathcal{L}_c(a,\beta) \overset{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\mathbf{P}\rangle \overset{\text{def}}{=} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\,\mathbf{P}\in\mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

Kantorovich Relaxation is a linear programming problem which is non-differentiable and has high time complexity O((n+m)nmlog(n+m)).



Entropy Regularization and Sinkhorn Algorithm (Discrete Measure Version)

Foundamental topics in Optimal Transport

Entropy Regularization

Definition: Given the discrete measures $a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$ the discrete entropy of the coupling matrix $\mathbf{P}$ is defined as

$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j} \, (log \, \mathbf{P}_{i,j} \, - \, 1)$$

Then the **entropy regularized Kantorovich Relaxation** can be defined as:

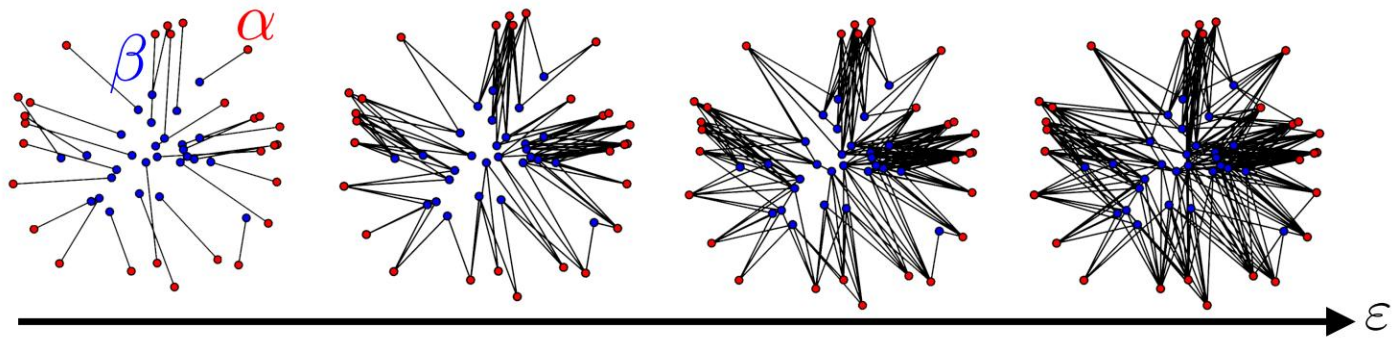
$$\mathcal{L}_c^{\varepsilon}(a, \beta) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon \mathbf{H}(\mathbf{P})$$

where

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \}$$

Remarks.

1. this problem is ε strongly convex
2. the solution is unique
3. instead of the “deterministic” of Kantorovich Relaxation, the entroy regularization is “probabilistic”



Example

Entropy Regularization and Sinkhorn Algorithm (Arbitrary Measure Version)

Foundamental topics in Optimal Transport

Entropy Regularization

Definition: Given the measures $a \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the continuous entropy can be defined as:

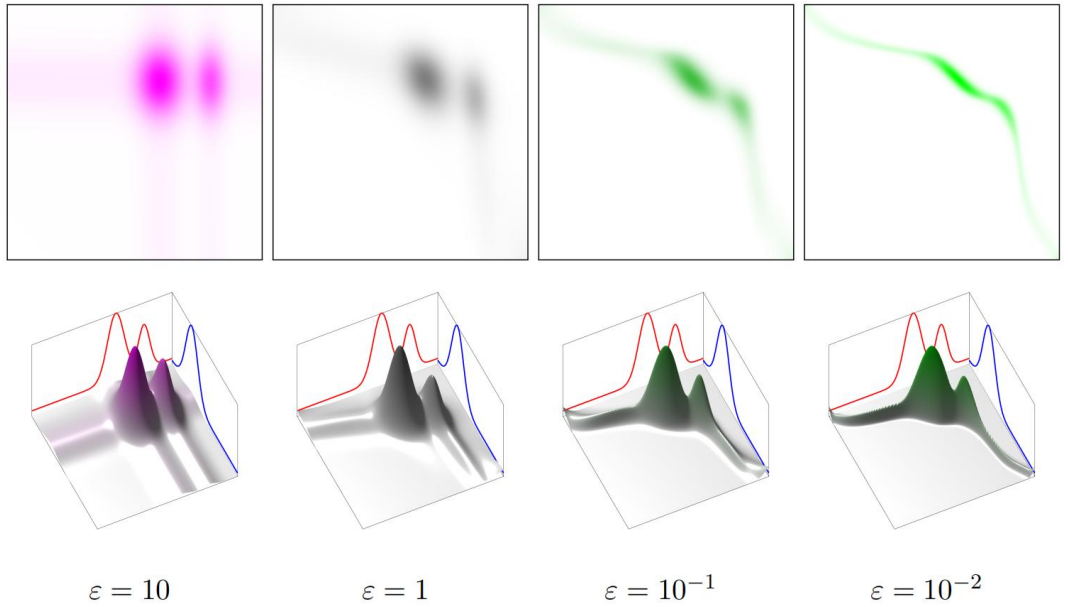
$$\mathbf{H}(\pi) \stackrel{\text{def}}{=} -\iint \pi(x,y)\, log\pi(x,y)\, dxdy$$

Then the **entropy regularized Kantorovich Relaxation** can be defined as:

$$\mathcal{L}_c^{\varepsilon}(a,\beta) \stackrel{\text{def}}{=} \min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) - \varepsilon \mathbf{H}(\pi) \right\},$$

where

$$\mathcal{U}(a,\beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$



Example

Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport

Why Entropy Regularization

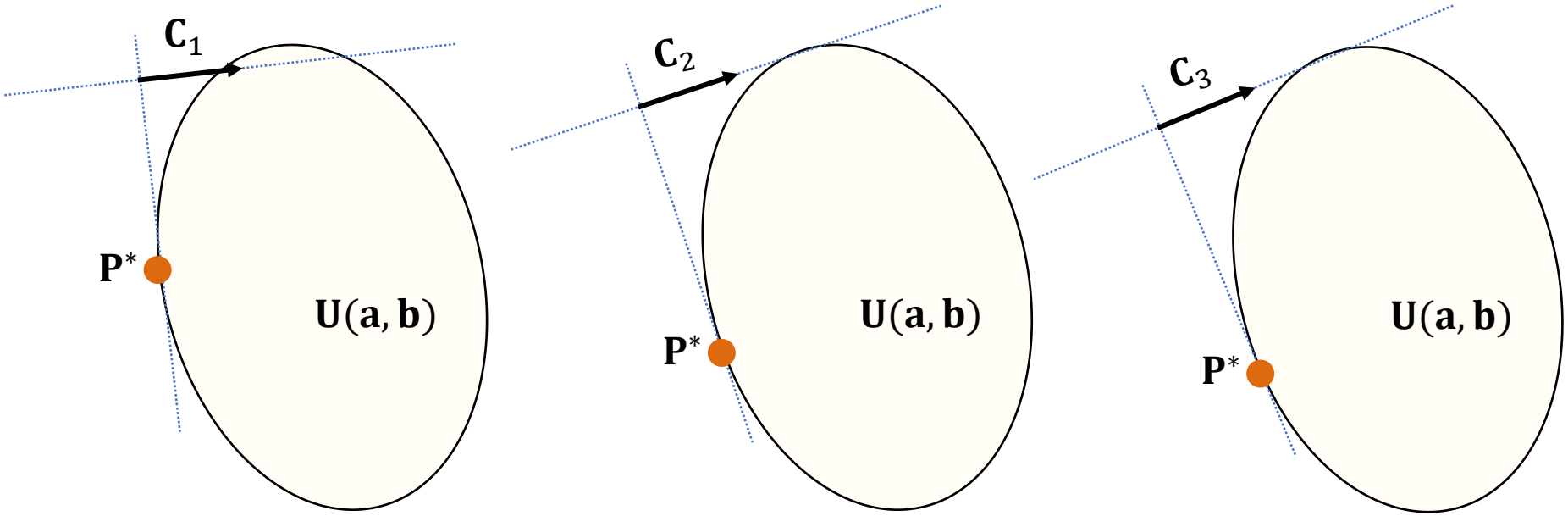
Entropy Regularization

$$\mathcal{L}_c^{\varepsilon}(a,\beta) \overset{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\mathbf{P}\rangle - \varepsilon\mathbf{H}(\mathbf{P})$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\,\mathbf{P}\in\mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

$$\mathbf{H}(\mathbf{P}) \overset{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}\,(log\ \mathbf{P}_{i,j} - 1)$$

Kantorovich Relaxation is ε strongly convex, differentiable and can be solved by Sinkhorn algorithm (can be implemented in GPU).



Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport

Why Entropy Regularization

Kantorovich Relaxation

$$\mathcal{L}_c(a,\beta) \overset{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\mathbf{P}\rangle \overset{\text{def}}{=} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\mathbf{P}\in\mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

Kantorovich Relaxation is a linear programming problem which is non-differentiable and has high time complexity O((n+m)nmlog(n+m)).

Entropy Regularization

$$\mathcal{L}_c^{\varepsilon}(a,\beta) \overset{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C},\mathbf{P}\rangle - \varepsilon\mathbf{H}(\mathbf{P})$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\mathbf{P}\in\mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

$$\mathbf{H}(\mathbf{P}) \overset{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}\,(log\,\mathbf{P}_{i,j} - 1)$$

Kantorovich Relaxation is ε strongly convex, differentiable and can be solved by Sinkhorn algorithm (can be implemented in GPU).

Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport

Entropy Regularization

$$\{\mathbf{P}_\varepsilon, \mathcal{L}_c^\varepsilon(a,\beta)\} \overset{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon \mathbf{H}(\mathbf{P})$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\, \mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

$$\mathbf{H}(\mathbf{P}) \overset{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}\,(log\ \mathbf{P}_{i,j}\ -\ 1)$$

The convergence of the Entropy Regularization

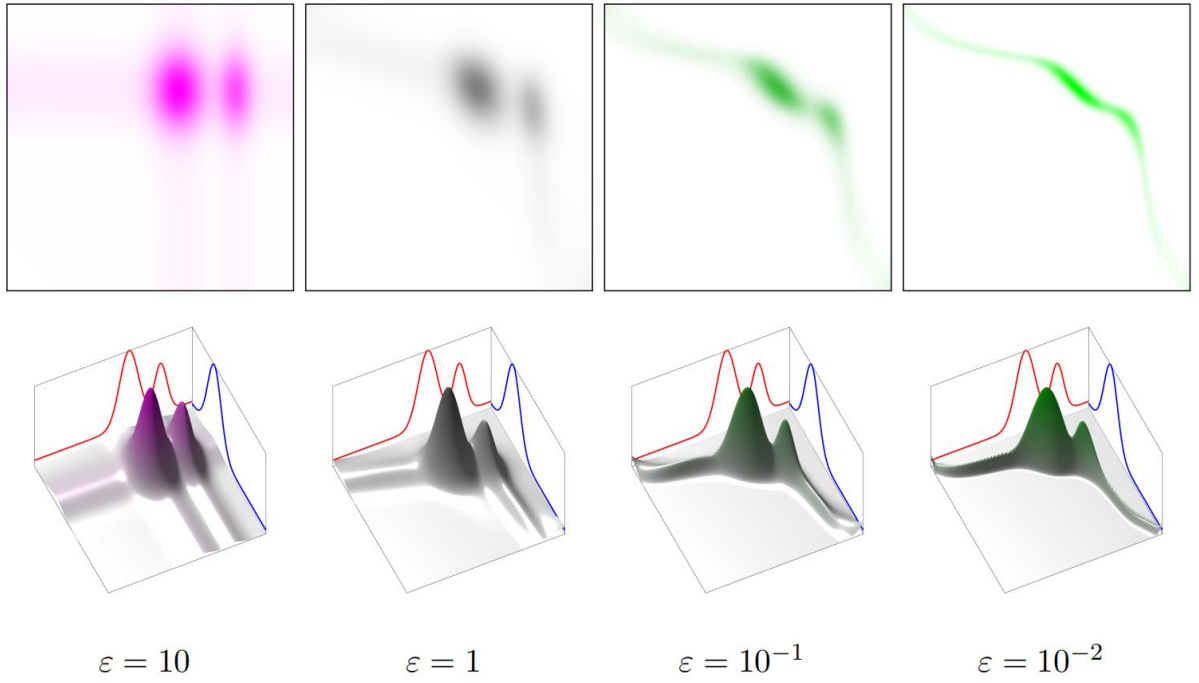
The unique solution $\mathbf{P}_\varepsilon$ converges to the optimal solution with maximal entropy within the set of all optimal solution of the Kantorovich problem, namely

$$\mathbf{P}_\varepsilon \overset{\varepsilon\to 0}{\longrightarrow} \operatorname*{argmin}_{\mathbf{P}} \left\{ -\mathbf{H}(\mathbf{P}) \,\middle|\, \mathbf{P} = \operatorname*{argmin}_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P}\rangle \right\}$$

$$\mathcal{L}_c^\varepsilon(a,\beta) \overset{\varepsilon\to 0}{\longrightarrow} \mathcal{L}_c(a,\beta) = \langle \mathbf{C}, \mathbf{P}\rangle$$

$$\mathbf{P}_\varepsilon \overset{\varepsilon\to\infty}{\longrightarrow} \mathbf{a}\mathbf{b}^T$$

Can we solver this problem?
-> **celebrated Sinkorn Algorithm !**



Example

Remarks.
1. This tells us if $\varepsilon \to \infty$, the optimal coupling converges to degenerate solution (rank 1) and if $\varepsilon \to 0$, the optimal coupling converges to the original Kantorovich problem. We can use it with confidence that as we decrease the value of ε, the approxiamation will become tighter and tighter.

Outline Overview

Foundamental topics in Optimal Transport

Monge Problem

$$\min_T \left\{ \int_{\mathcal{X}} c(x, T(x))\, da(x) \,\middle|\, T_{\#}a = \beta \right\}$$

Kantorovich Relaxation

$$\min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) \right\}$$

$$\mathcal{U}(a,\beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$

Kantorovich Duality

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup_{(f,g)\in\mathcal{R}(c)} \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} g(y)\, d\beta(y)$$

$$\mathcal{R}(c) \stackrel{\text{def}}{=} \{(f,g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mid \forall (x,y), f(x) + g(y) \le c(x,y)\}$$

Semi-dual Formulation

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \sup \int_{\mathcal{X}} f(x)\, da(x) + \int_{\mathcal{Y}} f^c(y)\, d\beta(y)$$

$$\forall y \in \mathcal{Y},\quad f^c(y) := \inf_{x\in\mathcal{X}} c(x,y) - f(x)$$

Entropy Regularization

$$\mathcal{L}_c^{\varepsilon}(a,\beta) \stackrel{\text{def}}{=} \min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) - \varepsilon \mathbf{H}(\pi) \right\}$$

$$\mathcal{U}(a,\beta) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = a \text{ and } P_{\mathcal{Y}\#}\pi = \beta \right\}$$

$$\mathbf{H}(\pi) \stackrel{\text{def}}{=} - \iint \pi(x,y)\, log\pi(x,y)\, dxdy$$

Kantorovich Relaxation

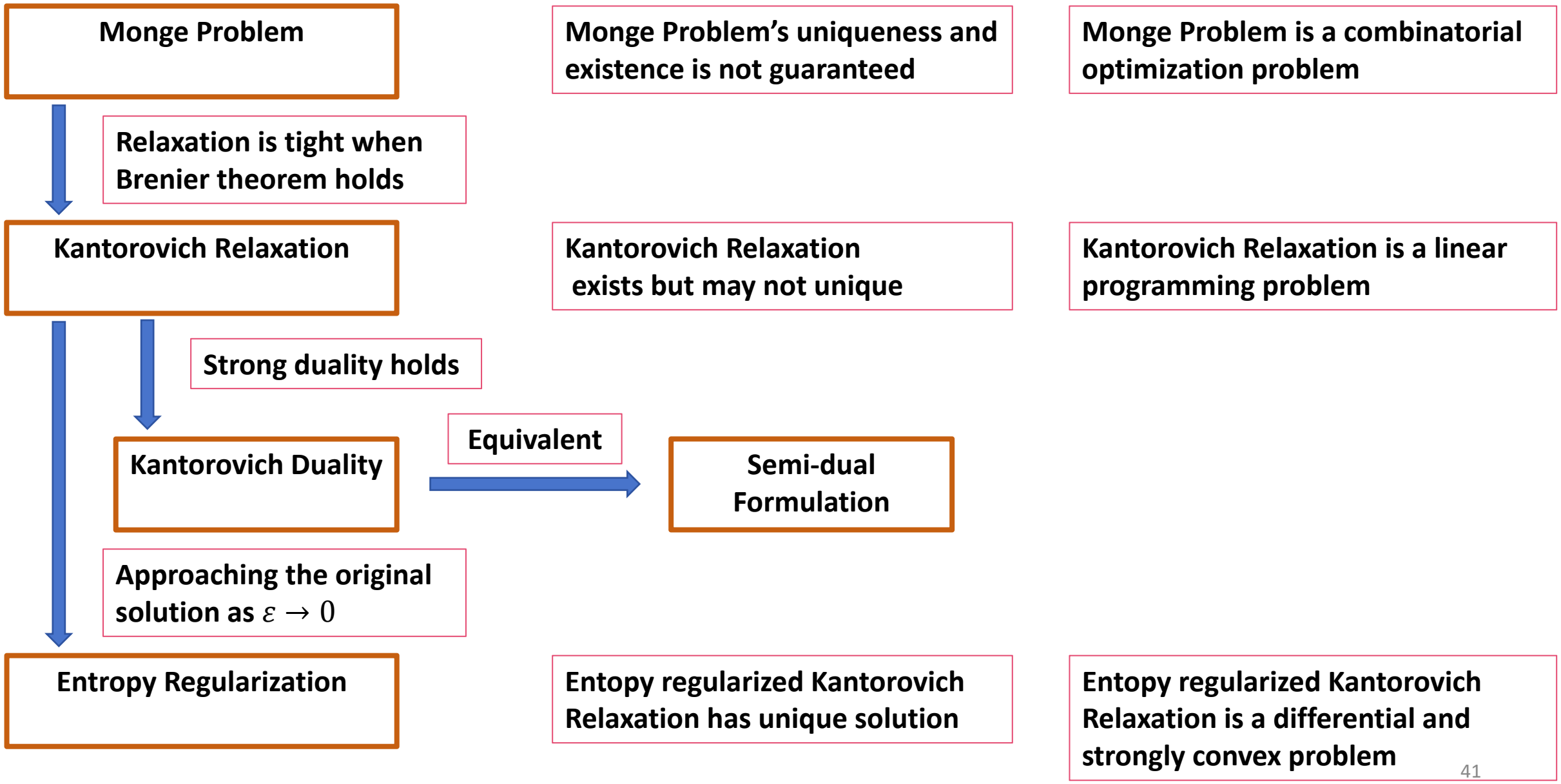
Langrange Duality

C-transform

Entropy Regularization

Property Overview

Foundamental topics in Optimal Transport



Closed-form solution

Surprisingly, for the discrete version of our entropy regularized Kantorovich problem, there exists the closed-form solution as

$$\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon}\, e^{-\mathbf{C}_{i,j}/\varepsilon}\, e^{\mathbf{g}_j/\varepsilon} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j$$
$$(\mathbf{P} = \mathrm{diag}(\mathbf{u})\mathbf{K}\mathrm{diag}(\mathbf{v}))$$

Proof.

Recall the lagangian function of Kantorovich relaxation problem with entropy regularization as

$$\mathcal{L}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon \mathbf{H}(\mathbf{P}) + \langle \mathbf{a} - \mathbf{P}\mathbf{1}_m, \mathbf{f} \rangle + \langle \mathbf{b} - \mathbf{P}^T \mathbf{1}_n, \mathbf{g} \rangle$$

Since this problem is strongly convex, we can derive its first order optimality condition:

$$\frac{\partial \mathcal{L}(\mathbf{P}, \mathbf{f}, \mathbf{g})}{\partial \mathbf{P}_{i,j}} = \mathbf{C}_{i,j} + \varepsilon log(\mathbf{P}_{i,j}) - \mathbf{f}_i - \mathbf{g}_j = 0$$

which confirms that

$$\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon}\, e^{-\mathbf{C}_{i,j}/\varepsilon}\, e^{\mathbf{g}_j/\varepsilon}$$

■

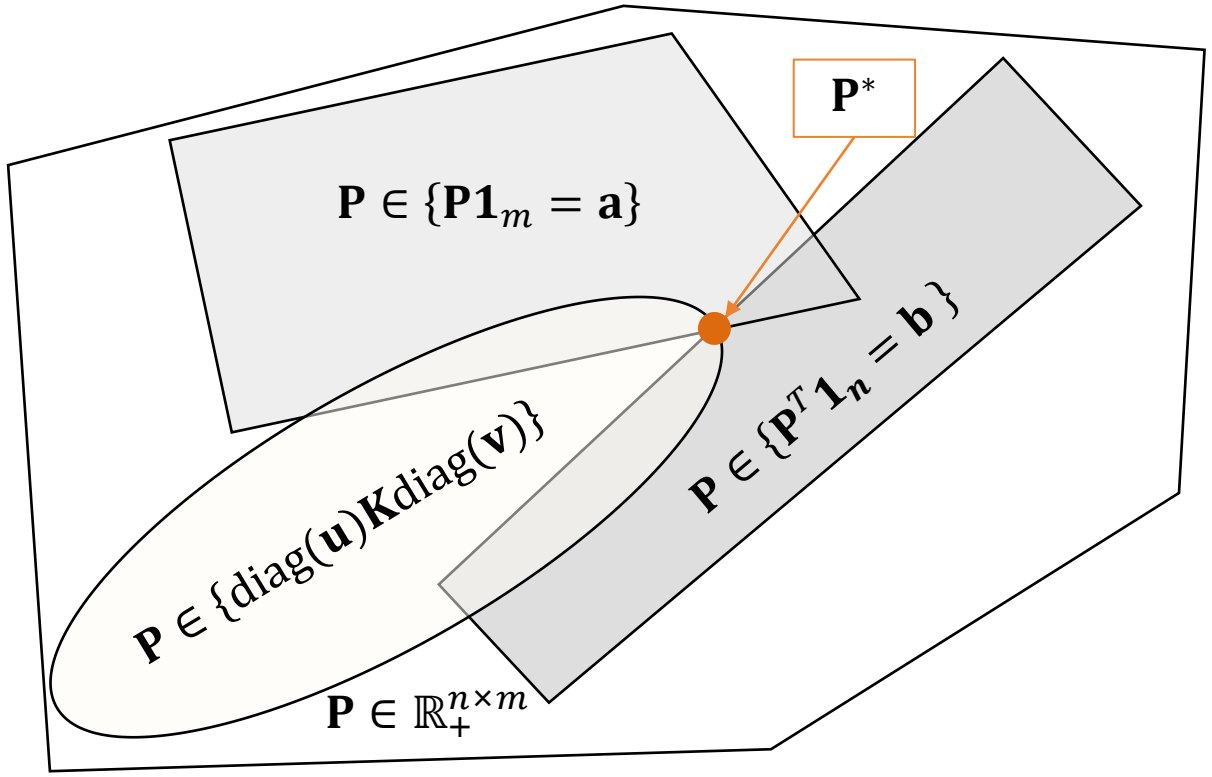
Entropy Regularization

$$\mathcal{L}_c^{\varepsilon}(a, \beta) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon \mathbf{H}(\mathbf{P})$$
$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \}$$
$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j} \, (log\, \mathbf{P}_{i,j} - 1)$$

Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport

Geometric Interpolation



Entropy Regularization

$$\mathcal{L}_c^{\varepsilon}(a,\beta) \stackrel{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon \mathbf{H}(\mathbf{P})$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \stackrel{\text{def}}{=} \{\, \mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

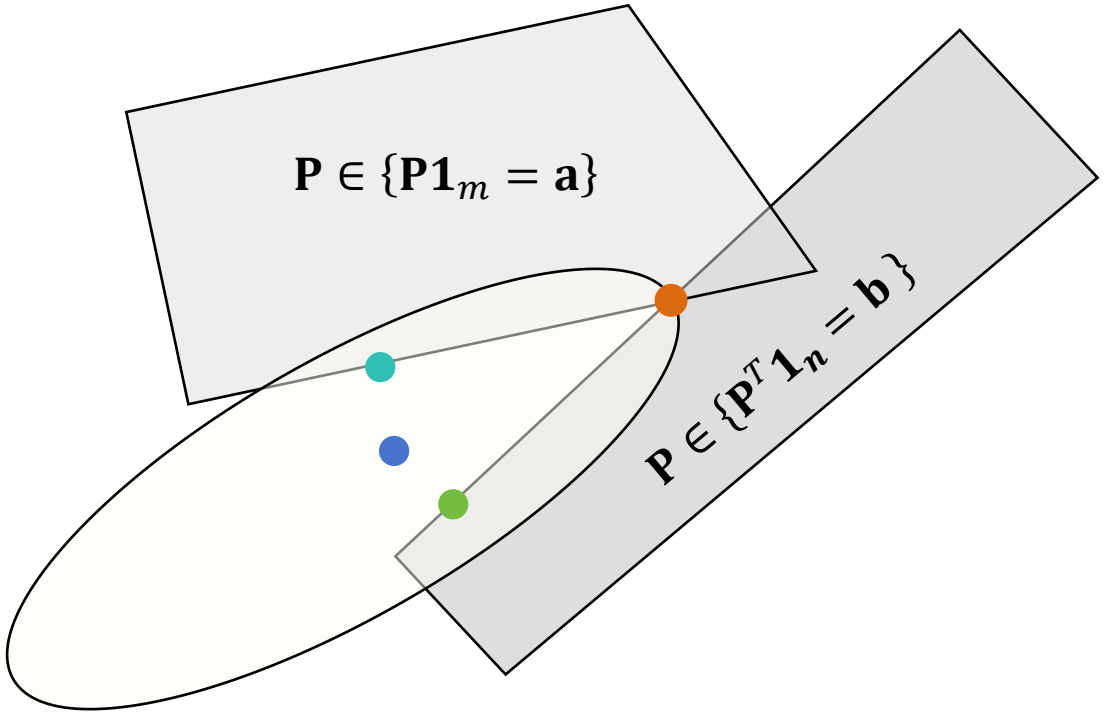
$$\mathbf{H}(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}\,(log\ \mathbf{P}_{i,j}\ -\ 1)$$

Remarks.

1. This is the “geometric” relationship of different sets (there is only one coupling matrix fullfilling all three constraints).

Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport



Benefit.
For this descrete measure input, the Sinkhorn algorithm is very efficient since it only involves the matrix element-wise division.

Sinkhorn Algorithm

Given an initialized coupling $\mathbf{P}$, we can alternately project this coupling on these two sets until convergence. Sinkhorn algorithm guarantees the convergence to the unique solution with global linear convergence rate.

Projection on set $\{\mathbf{P}\mathbf{1}_m = \mathbf{a}\}$

Given current $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{K}$, we can get the projection $\mathbf{P}$ on this set by

$$\mathbf{u} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$$

Projection on set $\{\mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$

Given current $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{K}$, we can get the projection $\mathbf{P}$ on this set by

$$\mathbf{v} = \frac{\mathbf{b}}{\mathbf{K}^T\mathbf{u}}$$

Entropy Regularization and Sinkhorn Algorithm

Foundamental topics in Optimal Transport

The KL projection interpretation of entropy regularization

The equivalent KL projection interpretation is defined as

$$\mathbf{P}_\varepsilon = \mathrm{Proj}^{\mathrm{KL}}_{\mathbf{U}(\mathbf{a},\mathbf{b})}(\mathbf{K}) \overset{\text{def}}{=} \arg\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathrm{KL}(\mathbf{P}|\mathbf{K})$$

where

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\, \mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

$$\mathrm{KL}(\mathbf{P}|\mathbf{K}) \overset{\text{def}}{=} \sum_{i,j} \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}}\right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}$$

$$\mathbf{K}_{i,j} \overset{\text{def}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$$

Entropy Regularization

$$\mathcal{L}_c^\varepsilon(a,\beta) \overset{\text{def}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon\mathbf{H}(\mathbf{P})$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \overset{\text{def}}{=} \{\, \mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

$$\mathbf{H}(\mathbf{P}) \overset{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}\,(log\ \mathbf{P}_{i,j}\ -\ 1)$$

The KL projection interpretation of Sinkhorn Algorithm

The equivalent KL projection interpretation of iterative projection

$$\mathbf{P}^{\ell+1} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^1_{\mathbf{a}}}(\mathbf{P}^\ell) \text{ and } \mathbf{P}^{\ell+2} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{C}^2_{\mathbf{b}}}(\mathbf{P}^{\ell+1})$$

where

$$\mathcal{C}^1_{\mathbf{a}} \overset{\text{def}}{=} \{\mathbf{P} : \mathbf{P}\mathbf{1}_m = \mathbf{a}\} \text{ and } \mathcal{C}^2_{\mathbf{b}} \overset{\text{def}}{=} \{\mathbf{P} : \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\}$$

Remarks.

We can generalise the discrete version to the general measure.

$$\mathbf{P}_\varepsilon \overset{\text{def}}{=} \arg\min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \mathrm{KL}(\mathbf{P}|\mathbf{K})$$

$$d\mathbf{K} \overset{\text{def}}{=} e^{-\frac{c(x,y)}{\varepsilon}} d\alpha(x) d\beta(y)$$

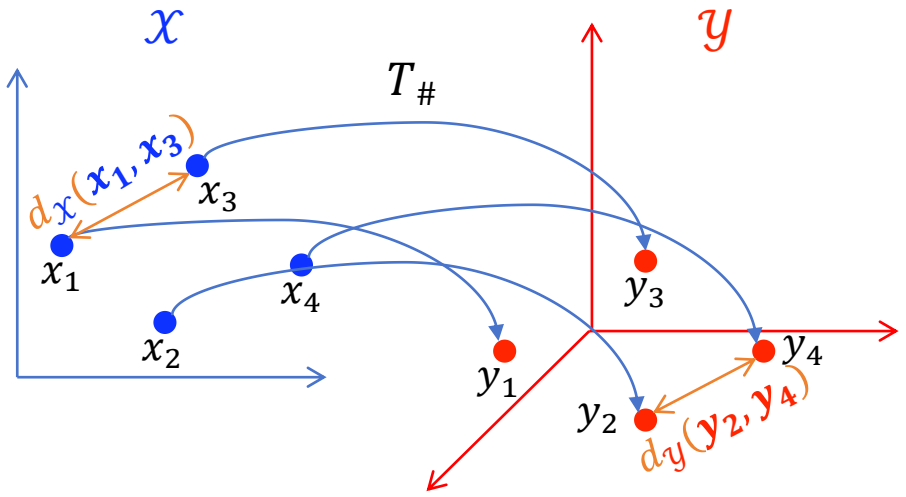
which is known as the “static Schrödinger problem”.

Gromov-Wasserstein Distance

Definition: The general setting corresponds to computing couplings between metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, a_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}}, a_{\mathcal{Y}})$, where $(d_{\mathcal{X}}, d_{\mathcal{Y}})$ are distances, while $a_{\mathcal{X}}$ and $a_{\mathcal{Y}}$ are measures on their respective spaces.

$$\mathcal{GW}\big((a_{\mathcal{X}}, d_{\mathcal{X}}), (a_{\mathcal{Y}}, d_{\mathcal{Y}})\big)^2 =$$

$$\min_{\pi \in \mathcal{U}(a_{\mathcal{X}}, a_{\mathcal{Y}})} \int_{\mathcal{X}^2 \times \mathcal{Y}^2} \left| d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y') \right|^2 d\pi(x, y) d\pi(x', y')$$



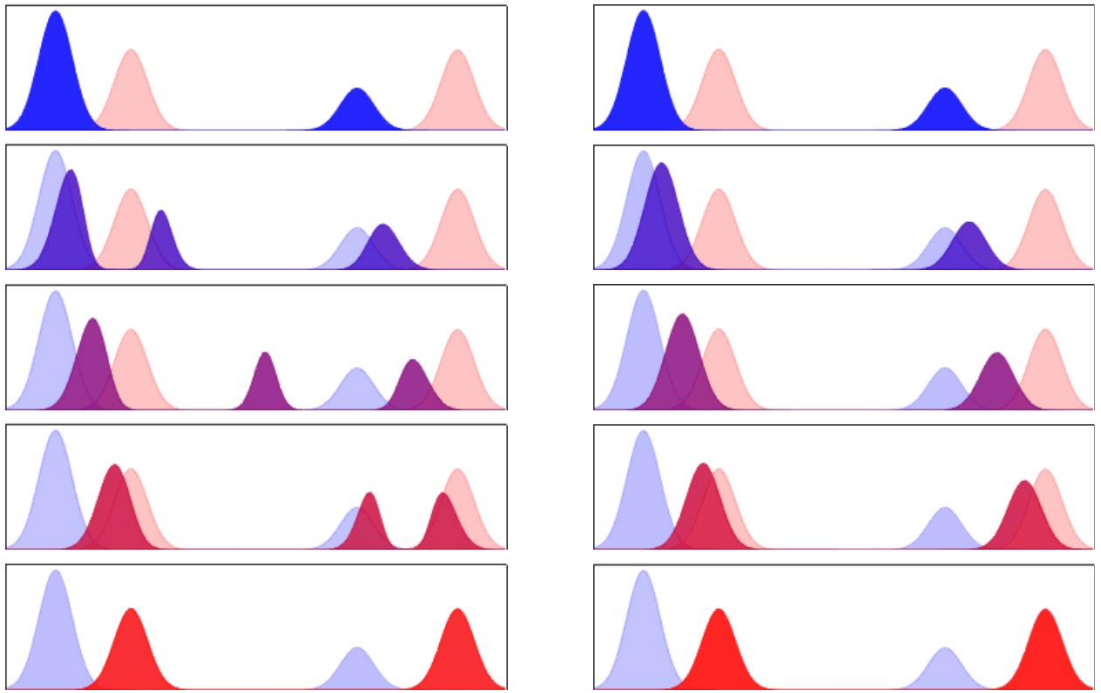
Example

Benefit.
1. If the space $\mathcal{X}$ and space $\mathcal{Y}$ are different to each other, it may be hard to define a ground cost c: $(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ among them. Then, this **Gromov-Wasserstein Distance** can be utilized.
2. **Gromov-Wasserstein Distance** defines a distance metric measure spaces up to isometrics: if there exists a bijection φ: $\mathcal{X} \to \mathcal{Y}$ such that $d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(\varphi(x), \varphi(x'))$.

Unbalanced Optimal Transport

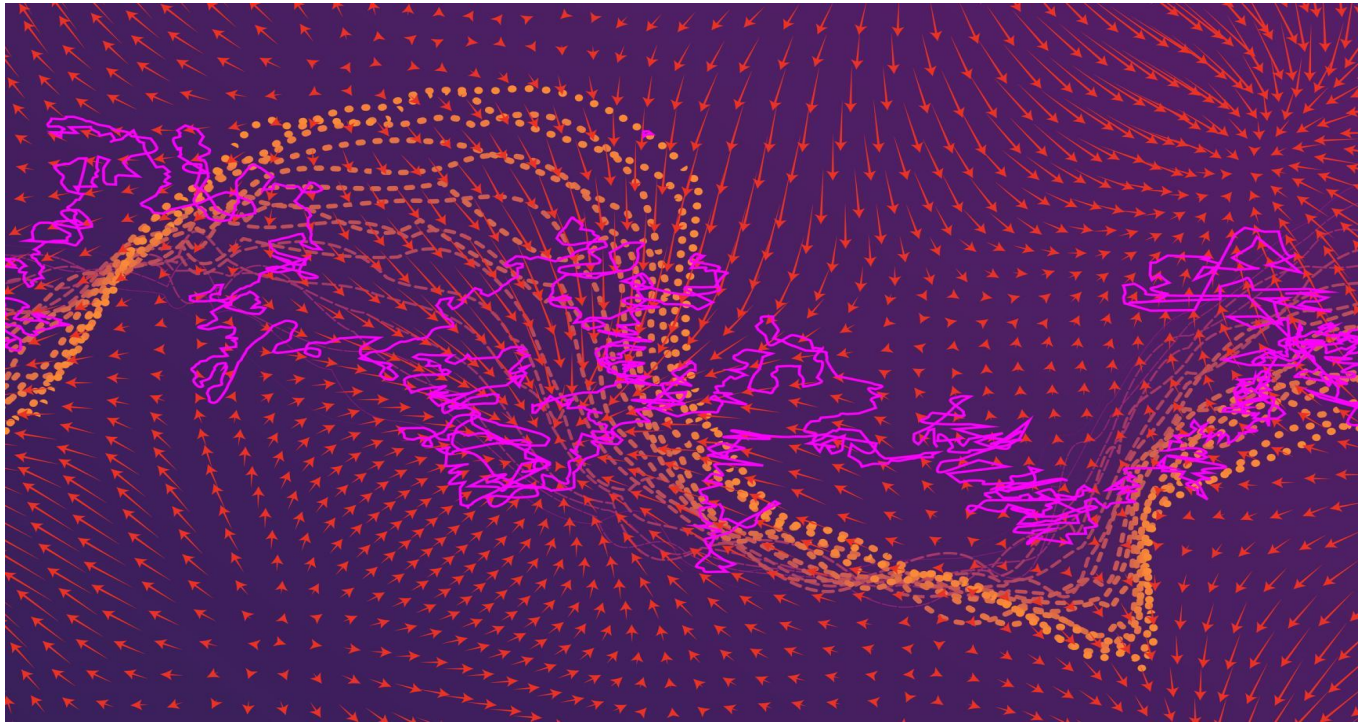
Definition: Given the measures $a \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$, the unbalanced version of **Kantorovich Relaxation problem** is defined as

$$\mathcal{L}_c(a,\beta) \stackrel{\text{def}}{=} \min_{\pi \in \mathcal{U}(a,\beta)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) + \tau KL(P_{\mathcal{X}\#}\pi|a) + \tau KL(P_{\mathcal{Y}\#}\pi|\beta) \right\},$$



Classical OT ($\tau = +\infty$) Ubalanced OT ($\tau = 1$)

Example



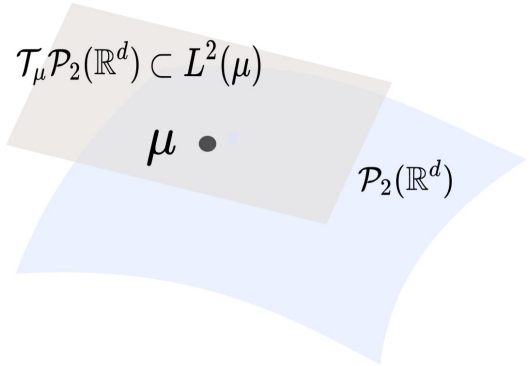
Advanced topics in Optimal Transport

- The Riemannian Structure of Optimal Transport
- Wasserstein Gradient Flow and JKO Scheme
- Introduction to the Schrödinger Bridge Problem

Outline Story

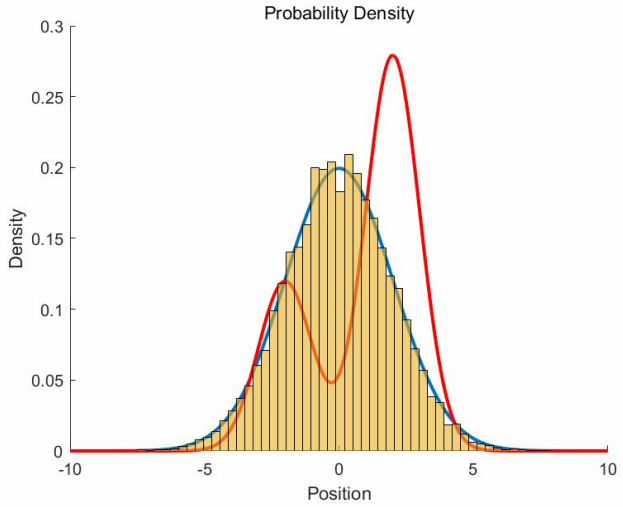
Foundamental topics in Optimal Transport

The Riemannian Structure of Optimal Transport



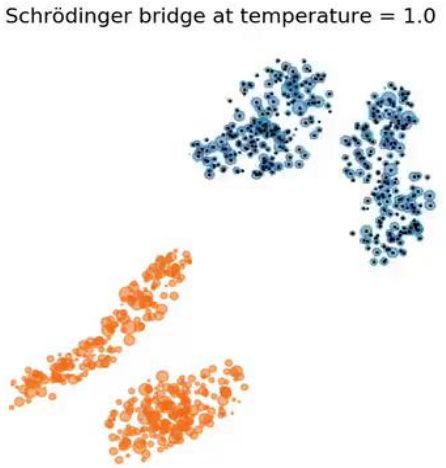
Wasserstein Gradient Flow and JKO Scheme

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)$$



Introduction to the Schrödinger Bridge Problem

$$\min_{\overline{\pi} \in \overline{\mathcal{U}}(\mu_0, \mu_1)} \mathrm{KL}(\overline{\pi} | \overline{\mathcal{K}})$$



Prerequisite - Smooth Manifold

Advanced topics in Optimal Transport

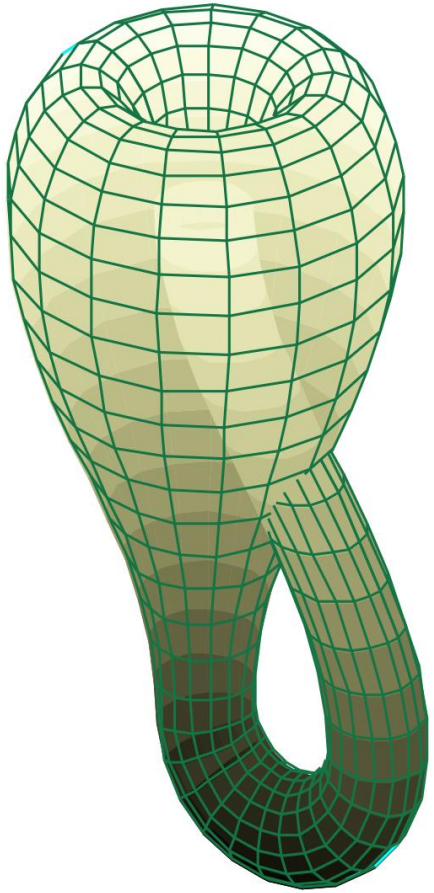
Manifold

Definition: A manifold is a topological space that locally resembles Euclidean space near each point. More precisely, an n-dimensional manifold is a topological space with the property that each point has a neighborhood that is **homeomorphic** to an open subset of n-dimensional Euclidean space.

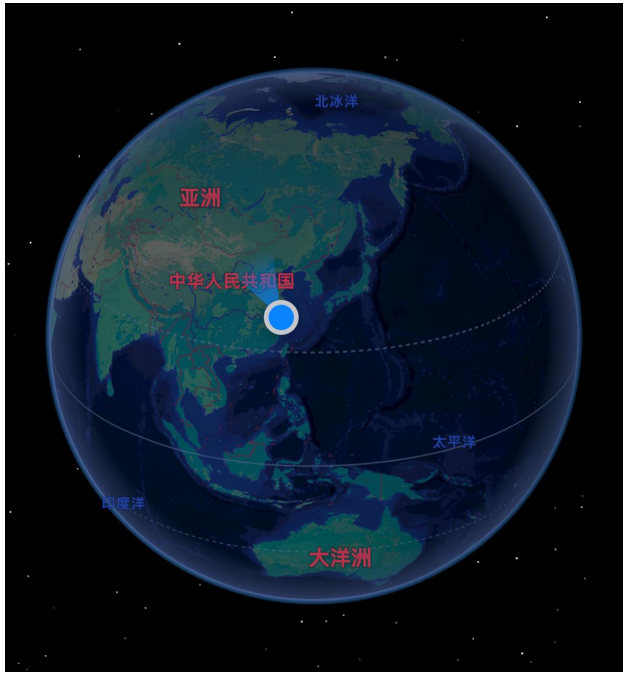
Remark.
A manifold is a space that looks like a plane locally. Just like the surface of the Earth appears flat but is actually spherical.

Smooth Manifold

Definition: A smooth manifold is a type of manifold that is locally similar enough to a vector space to allow one to apply calculus.



Example of Non Smooth Manifold



Example of Smooth Manifold

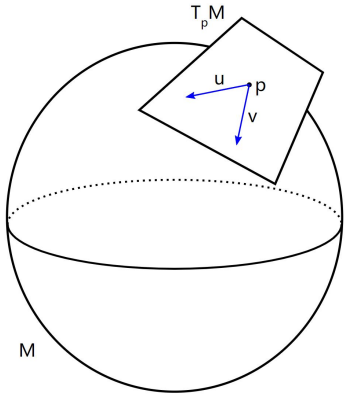
Prerequisite - Riemannian Manifold

Advanced topics in Optimal Transport

Riemannian metric and Riemannian Manifold

Definition: A **Riemannian metric** g on a smooth manifold M assigns to each p a positive-definite inner product $g_p: T_pM \times T_pM \to \mathbb{R}$ for each **tangent space** T_pM in a **smooth way**. This induces a norm $\|\cdot\|_p: T_pM \to \mathbb{R}$ defined by $\|v\|_p = \sqrt{g_p(v,v)}$. A smooth manifold M endowed with a Riemannian metric g is a **Riemannian manifold**, denoted (M, g).

Remark.
A Riemannian manifold is a geometric space on which many geometric notions such as distance, angles, length, volume, and curvature are defined.

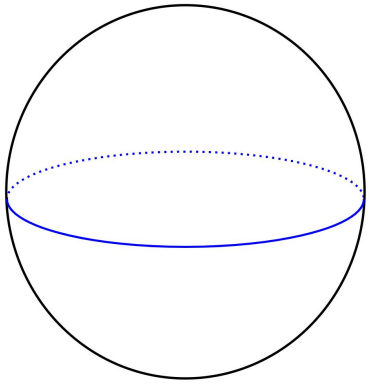


Example A tangent plane of the sphere with two vectors in it. A Riemannian metric allows one to take the inner product of these vectors.

Geodesics

Geodesics are curves with no intrinsic acceleration. They are the generalization of straight lines in Euclidean space to arbitrary Riemannian manifolds. An ant living in a Riemannian manifold walking straight ahead without making any effort to accelerate or turn would trace out a geodesic.

Remark.
Geodesics can seen as a locally minimal-distance curve on the manifold connecting the startpoint and endpoint.



Example In the round sphere the maximal geodesics are great circles.

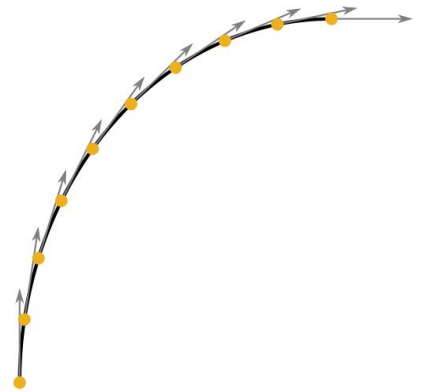
Prerequisite - Metric space of Riemannian Manifold

Advanced topics in Optimal Transport

Length of curve on Riemannian Manifold

Definition: An admissible curve is a piecewise smooth curve γ: $[0,1] \to M$ whose velocity $\gamma'(t) \in T_{\gamma(t)}M$ is nonzero everywhere it is defined. The nonnegative function $t \longrightarrow \|\gamma'(t)\|_{\gamma(t)} = \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))}$ is defined on the interval $[0,1]$ except for at finitely many points.
The length $\mathcal{L}(\gamma)$ of an admissible curve γ is defined as

$$\mathcal{L}(\gamma) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)}\, dt$$



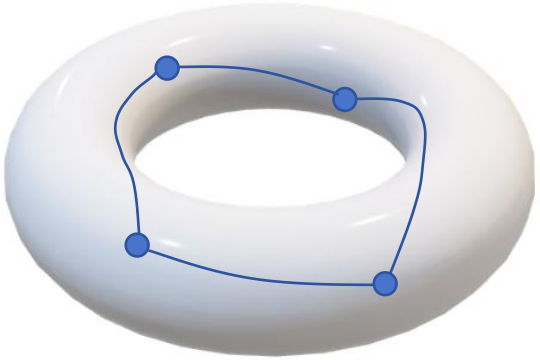
Example The length of a curve can also be expressed by integrating the curve derivative.

Metric space induced by Riemannian metric

Definition: For (M, g) a Riemannian manifold, define d_g: $M \times M \to [0, \infty)$ by

$$d_g(p, q) = \inf\{\, \mathcal{L}(\gamma) \colon\ \gamma \text{ an admissible curve with } \gamma(0) = p, \gamma(1) = q\}$$

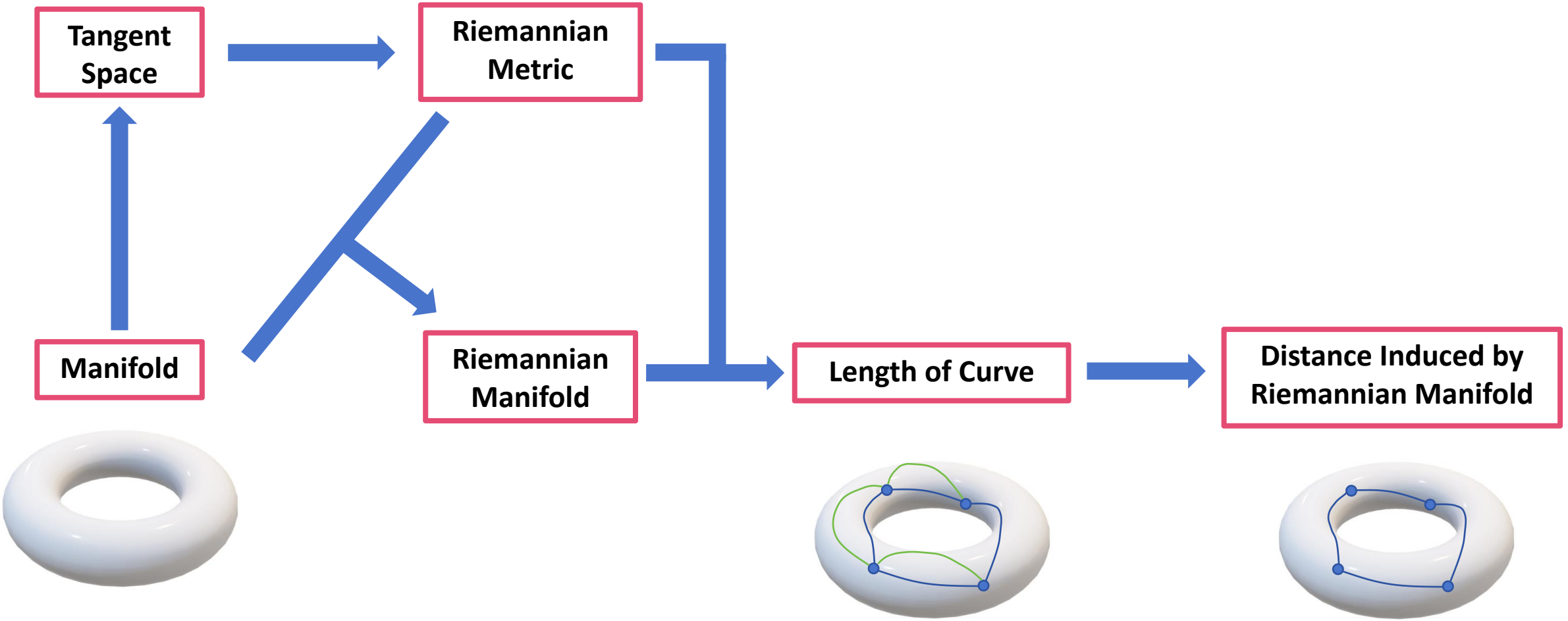
Then (M, d_g) is a metric space. Besides, this minimum distance curve is the geodesics connecting p and q.



Example The minimum length of a curve denotes the distance (metric) between points on manifold.

The Riemannian Structure of Optimal Transport

Advanced topics in Optimal Transport



$$\mathcal{L}(\gamma) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)}\, dt$$

$$d_g(p,q) = \inf_{\gamma(0)=p,\,\gamma(1)=q} \mathcal{L}(\gamma)$$

Wasserstein Space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$

Let $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second moments

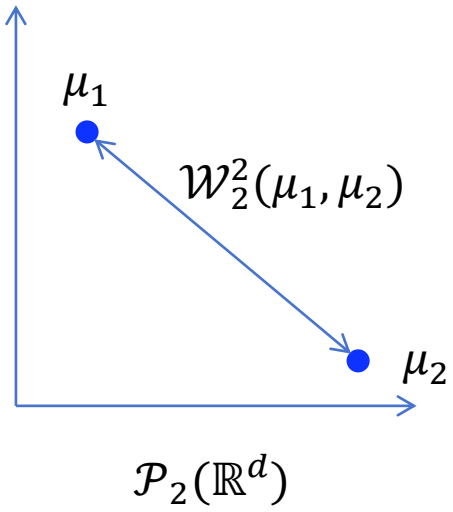
$$\mathcal{P}_2(\mathbb{R}^d) = \left\{\mu \in \mathcal{M}_+^1(\mathbb{R}^d)\,,\ \int \|x\|^2 d\mu(x) < \infty\right\}$$

$\mathcal{P}_2(\mathbb{R}^d)$ is endowed with the Wasserstein-2 distance from optimal transport

$$\mathcal{W}_2^2(a,\beta) = \inf_{\pi \in \mathcal{U}(a,\beta)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \, d\pi(x,y)$$

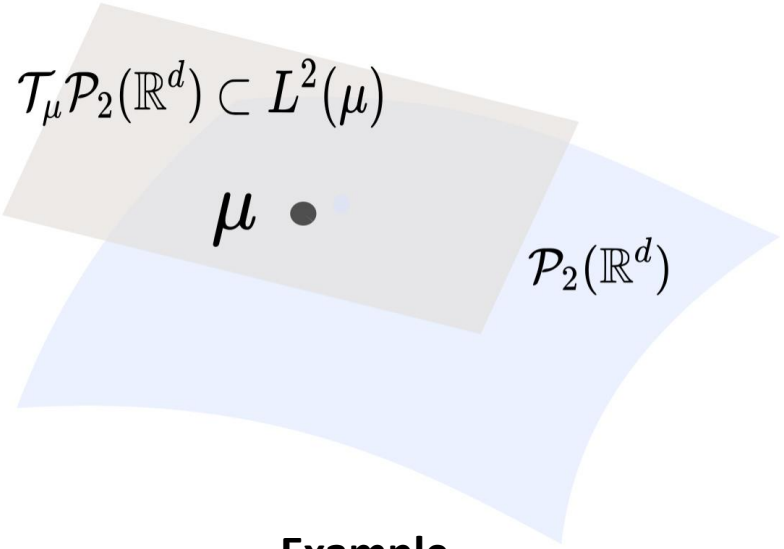
Then, the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is called the Wasserstein space.

Example



Riemannian interpretation of $\mathcal{P}_2(\mathbb{R}^d)$

We can view the whole probability measure $\mathcal{P}_2(\mathbb{R}^d)$ as a infinite dimension Riemannian manifold (a probability measure is a **positive infinite dimension vector with sum equals 1**).



Example

Riemannian metric of $\mathcal{P}_2(\mathbb{R}^d)$

Denote $L^2(\mu)$ the space of vector-valued, square-integrable function

$$L^2(\mu) = \left\{ f: \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \|f(x)\|^2 d\mu(x) < \infty \right\}$$

Then the Riemannian metric $g_\mu: T_\mu \mathcal{P}_2(\mathbb{R}^d) \times T_\mu \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ of $\mathcal{P}_2(\mathbb{R}^d)$ is the inner product between functions in $L^2(\mu)$

$$\int_{\mathbb{R}^d} \langle f(x), g(x) \rangle_{\mathbb{R}^d} \, d\mu(x)$$

Tangent space $T_\mu \mathcal{P}_2(\mathbb{R}^d)$

Remark.
The tangent space is really complicated, we only leave out an informal definition for understanding.

$$T_\mu \mathcal{P}_2(\mathbb{R}^d) = \begin{Bmatrix} f: \mathbb{R}^d \to \mathbb{R}^d \; f \in L^2(\mu) \\ f = \lambda\,(T - Id)_{\#}\mu \\ \lambda > 0, T \text{ is an optimal transport map} \end{Bmatrix}$$

Among this tangent space, there is a very interesting velocity field which defines the instantaneous probabilistic evolution.

Velocity Field $v_t(x)$ at μ_t

Let $\mu_t = (\boldsymbol{X}(t))_{\#}\mu_0$ denotes the evolving probabilistic measure under the process $\boldsymbol{X}(t)$.

Definition. Given a smooth test function ψ, the velocity field $v_t(x)$ at μ_t is defined as

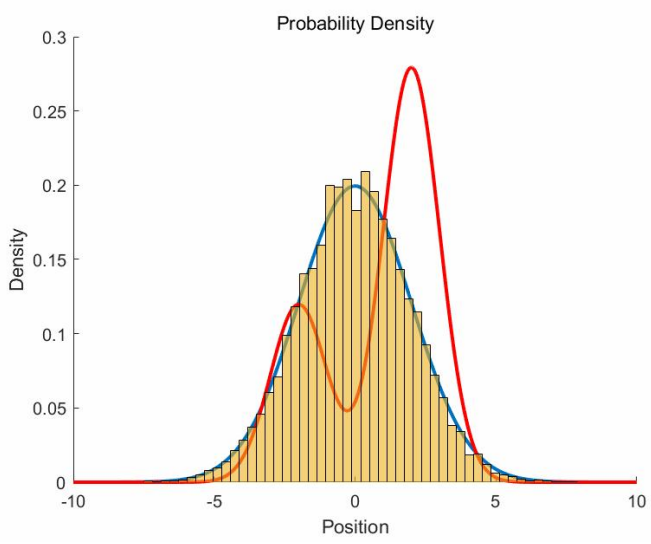
$$\frac{d}{dt}\int \psi(x)\,\mu_t(x)dx = \int \nabla\psi(x) \cdot v_t(x)\,\mu_t(x)dx$$

Remark.

1. There exists another definition of velocity field as

$$v_t(x) = \lim_{\delta \to 0} \frac{(T_{\mu_t \to \mu_{t+\delta}} - Id)(x)}{\delta}$$

2. We can also view $v_t(x)$ as the velocity field of particles.



Continuity Equation

Definition. Let $\mu_t = (\boldsymbol{X}(t))_{\#}\mu_0$ denotes the evolving probabilistic measure under the process $\boldsymbol{X}(t)$ and v_t represents the velocity field of μ_t. The continuity equation describes the relationship between flow field and probabilistic measure:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0$$

Proof. Let ψ be a smooth test function, and consider the function $t \to \int \psi(x)\, \mu_t(x) dx$, we have

$$\int \psi(x) \frac{\partial \mu_t(x)}{\partial t} dx = \frac{d}{dt} \int \psi(x)\, \mu_t(x) dx$$

$$= \frac{d}{dt} \int \psi(\boldsymbol{X}(x,t))\, \mu_0(x) dx$$

$$= \int \nabla \psi(\boldsymbol{X}(x,t)) \cdot v_t(\boldsymbol{X}(x,t))\, \mu_0(x) dx$$

$$= \int \nabla \psi(x) \cdot v_t(x)\, \mu_t(x) dx = - \int \psi(x)\, \nabla \cdot (v_t(x)\, \mu_t(x))\, dx$$

■

Remark.
This continuity equation links the velocity of local particles $\boldsymbol{X}(t)$ and the global law μ_t.

Given this equation, we can now define the length of curve on the riemannian manifold of probabilistic measure.

The Riemannian Structure of Optimal Transport

Advanced topics in Optimal Transport

Length of "curve" on Riemannian Manifold

Definition: Given the measures $a \in \mathcal{M}_+^1(\mathbb{R}^d)$ and $\beta \in \mathcal{M}_+^1(\mathbb{R}^d)$, the Riemannian metric induces a distance (length of curve) as:

$$d_{g_\mu}(a,\beta) = \inf_{\mu_t \in \mathcal{M}_+^1(\mathbb{R}^d),\, v_t \in L^2(\mu_t)} \left\{ \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, d\mu_t(x)dt \;\middle|\; \frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0, \mu_0 = a, \mu_1 = \beta \right\}$$

Benamou - Brenier Formulation

Definition: Recall the Wasserstein distance

$$\mathcal{W}_2^2(a,\beta) \overset{\text{def}}{=} \left(\min_{\pi \in \mathcal{U}(a,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)^2 \, d\pi(x,y) \right)$$

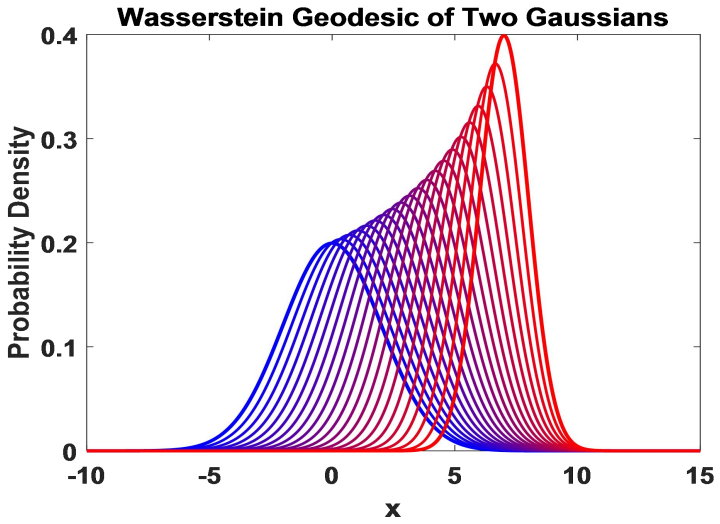
Benamou - Brenier Formulation denotes that $\mathcal{W}_2^2(a,\beta) = d_{g_\mu}(a,\beta)$

Remark.
This **Riemannian Structure of Optimal Transport** is known as Otto Calculas, It explains the geometric meaning of Wasserstein distance.

Recall the Wasserstein Geodesic

Definition. Given the optimal transport map T_a^β, the wasserstein geodesic between a and β is defined as

$$\mu_t = ((1-t)Id + t\, T_a^\beta)_\# a, \quad t \in [0,1]$$



Remark.
Another benifits of this displacement interpolation is that, this interpolation is linear (constant velocity geodesic).

The Riemannian Structure of Optimal Transport

Advanced topics in Optimal Transport

Tangent Space

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0$$

Riemannian Metric

$$\int_{\mathbb{R}^d} \langle f(x), g(x) \rangle_{\mathbb{R}^d} \, d\mu(x)$$

$$\mathcal{W}_2^2(a,\beta) \overset{\text{def}}{=} \left(\min_{\pi \in \mathcal{U}(a,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)^2 \, d\pi(x,y) \right)$$

Wasserstein2 Distance

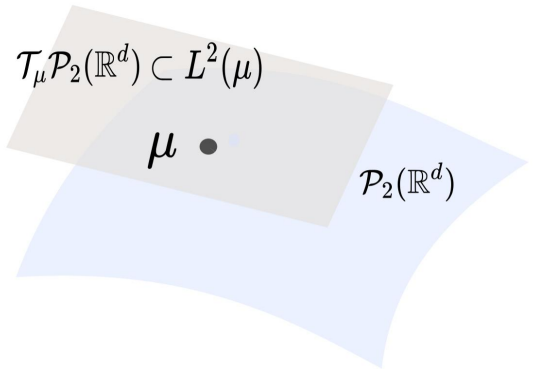
Manifold $\mathcal{P}_2(\mathbb{R}^d)$

Riemannian Manifold

Length of Curve

$$\int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, d\mu_t(x) dt$$

Distance Induced by Riemannian Manifold

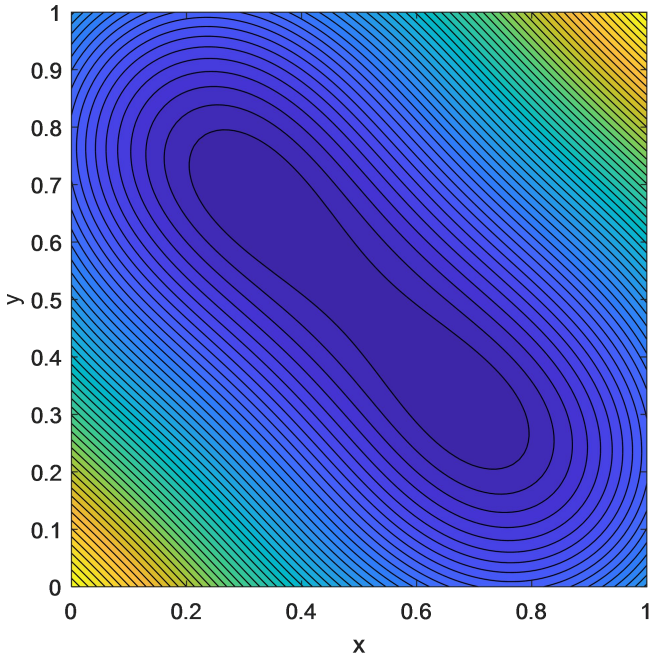


$$d_{g_\mu}(a,\beta) = \inf_{\mu_t \in \mathcal{M}_+^1(\mathbb{R}^d),\, v_t \in L^2(\mu_t)} \left\{ \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, d\mu_t(x) dt \;\middle|\; \frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0, \mu_0 = a, \mu_1 = \beta \right\}$$

Prerequisite - Gradient Flow in Euclidean space

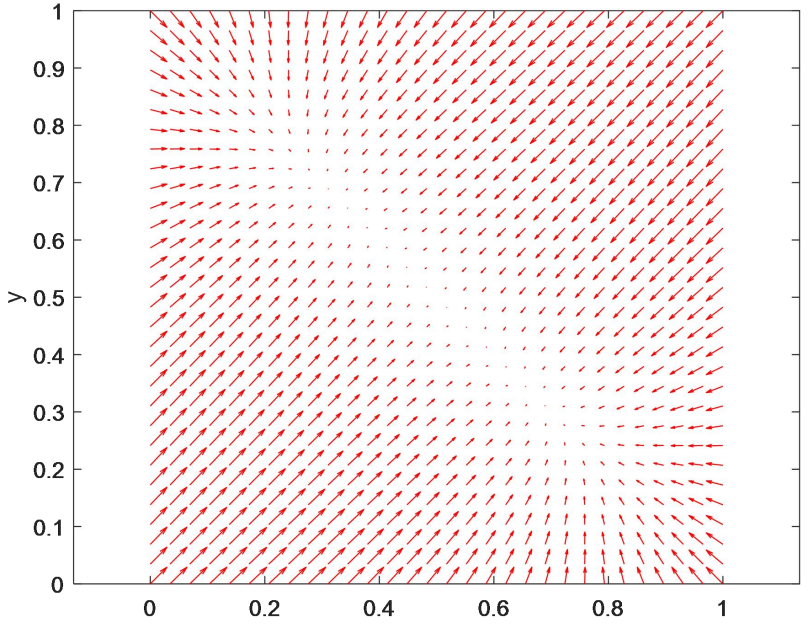
Advanced topics in Optimal Transport

Gradient flow in Euclidean space $\mathbb{R}^d$



$$V(x,y) = cos(\pi x) * cos(\pi y) + 4(x-0.5)^2 + 4(y-0.5)^2$$

Example of energy function



$$\nabla V(x,y) = \begin{bmatrix} -\pi * sin(\pi x) * cos(\pi y) + 8(x-0.5) \\ -\pi * cos(\pi x) * sin(\pi y) + 8(y-0.5) \end{bmatrix}$$

Example of field of negetive gradient flow

Prerequisite - Gradient Flow in Euclidean space

Advanced topics in Optimal Transport

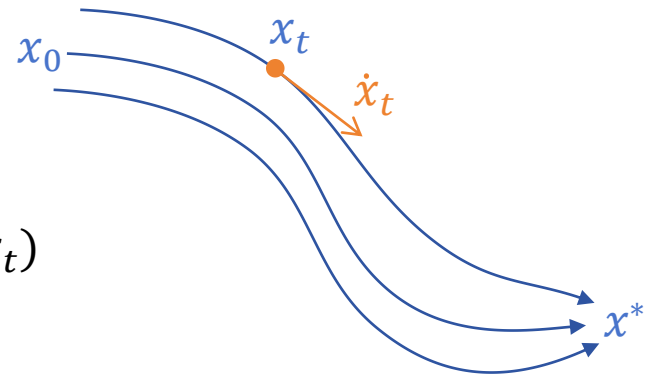
Optimization in Euclidean space $\mathbb{R}^d$

$$\min_{x \in \mathbb{R}^d} V(x)$$

where V: $\mathbb{R}^d \to \mathbb{R}$ describes the loss function

Euclidean Gradient Flow

The gradient flow describes the evolution of a point in the gradient vector field. We can represent it by a curve $x_t = x(t)$ that changes over time, and this curve satisfies the following ordinary differential equation (ODE)

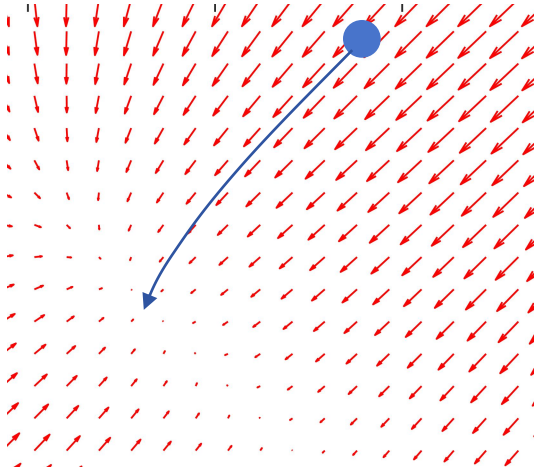


Definition.

$$\begin{cases} \dot{x}_t = \dfrac{\partial x_t}{\partial t} = -\nabla V(x_t) \\ x(0) = x_0 \end{cases}$$

Example

Gradient flow implicitly defines many odes



$$V(x,y) = cos(\pi x) * cos(\pi y) + 4(x-0.5)^2 + 4(y-0.5)^2$$

How to solve it in practice?
-> **Time discretization**

Time discretization of gradient flow in Euclidean space $\mathbb{R}^d$

There are two time discretization methods of this gradient flow

- **Gradient descent**

$$x_{m+1} = x_m - \gamma \nabla V(x_m)$$

i.e. **Forward Euler (explicit)**

$$\frac{x_{m+1} - x_m}{\gamma} = -\nabla V(x_m)$$

- **Proximal point algorithm**

$$x_{m+1} = \arg\min_x V(x) + \frac{1}{2\gamma}\|x - x_m\|^2$$

i.e. **Backward Euler (implicit)**

$$\frac{x_{m+1} - x_m}{\gamma} = -\nabla V(x_{m+1})$$

Remark.
1. Proximal point algorithm are originaly used to solve the non-smooth optimization problem.
2. If the optimization is set up in another space, we must replace the original metric $\frac{1}{2\gamma}\|x - x_m\|^2$.

How to define the gradient flow in Wasserstein space?
What is the explicit and implicit discretization in Wasserstein space?

Wasserstein Gradient Flow and JKO Scheme

Advanced topics in Optimal Transport

Variational Optimization in the Wasserstein Space

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)$$

Wasserstein Gradient Flow $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$

Definition: The **Wasserstein Gradient Flow** $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$ at μ_t can be defined as

$$\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) = \nabla\left(\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}\right)$$

where the $\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}$ represents the first variation for some linear perturbation ξ (s.t. $\varepsilon\xi + \mu \in \mathcal{P}_2(\mathbb{R}^d), \mu \in \mathcal{P}_2(\mathbb{R}^d)$)

$$\mathcal{F}(\mu + \varepsilon\xi) \approx \mathcal{F}(\mu) + \varepsilon \int \frac{\delta\mathcal{F}(\mu)}{\delta\mu}(x)\, d\xi(x)$$

Example

Variational Optimization Problem	$\mathcal{F}(\mu_t)$
Potential Energy	$\int V(x)\, d\mu_t(x)$
Negative Entropy	$\frac{1}{2}\int log(\mu_t(x))\, d\mu_t(x)$
KL Divergence	$\mathrm{KL}(\mu_t \mid \exp(-V(x))) = \int V(x) + log(\mu_t(x))\, d\mu_t(x)$

Two perspectives of Wasserstein Gradient Flow $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$

$$\dot{x}_t = v_t = -\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t), \quad x_0 \sim \mu_0$$

Langrangian (Particles)

$$\frac{\partial\mu_t}{\partial t} = \nabla \cdot \left(\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\, \mu_t\right), \quad \mu_{t=0} = \mu_0$$

Eulerian (Law/Density)

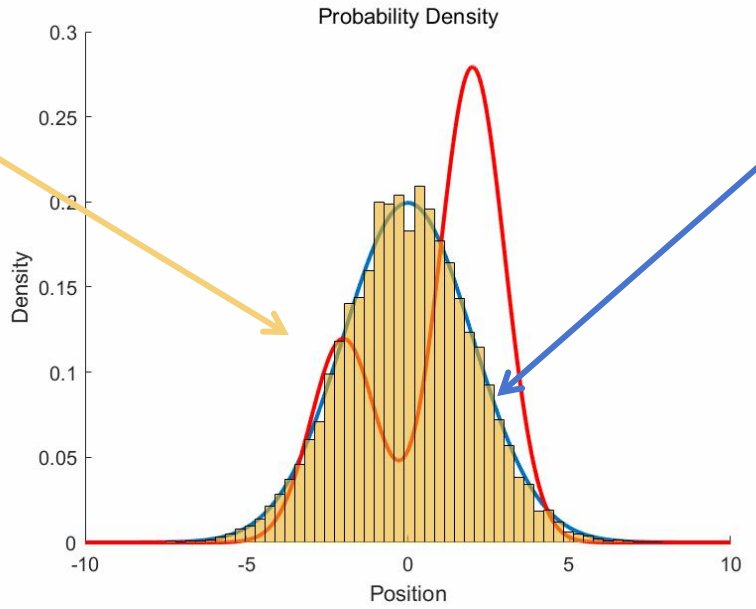
Example

$$\dot{x}_t = v_t = -\nabla_{W_2}\mathcal{F}(\mu_t), \quad x_0 \sim \mu_0$$

Langrangian (Particles)

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot \left(\nabla_{W_2}\mathcal{F}(\mu_t)\, \mu_t\right), \quad \mu_{t=0} = \mu_0$$

Eulerian (Law/Density)



Remark.
Wasserstein gradient flow is extremely **elegant**. It tells us that sampling is just a variational optimization over some functional in Wasserstein space, and its optimization trajectory follows the geodesic of the probability space manifold. In the machine learning community, many scholars use this tool to analyze the convergence properties of sampling algorithm.

Wasserstein Gradient Flow $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) = \nabla(\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t})$

Proof.

Recall the definition of the gradient definition.

$$d\,\mathcal{F}(\mu)[v] = \lim_{t\to 0}\frac{\mathcal{F}((Id+tv)_{\#}\mu) - \mathcal{F}(\mu)}{t}$$

we expand $\mu_t = (Id+tv)_{\#}\mu$

$$\mu_t = \mu + t - \nabla\cdot(\mu v)$$

then we expand $\mathcal{F}((Id+tv)_{\#}\mu)$

$$\mathcal{F}((Id+tv)_{\#}\mu) = \mathcal{F}(\mu + t - \nabla\cdot(\mu v)) = \mathcal{F}(\mu) - t\int\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}(\nabla\cdot(\mu v))dx$$

then by integration-by-parts

$$\mathcal{F}((Id+tv)_{\#}\mu) = \mathcal{F}(\mu) + t\int\nabla\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}\cdot v dx$$

finally

$$\langle\nabla_{\mathcal{W}_2}\mathcal{F}(\mu), v\rangle_{\mu} = d\,\mathcal{F}(\mu)[v] = \int\nabla\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}\cdot v\,\mu_t dx$$

$$\nabla_{\mathcal{W}_2}\mathcal{F}(\mu) = \nabla\frac{\delta\mathcal{F}(\mu_t)}{\delta\mu_t}$$

■

Wasserstein Gradient Flow Example

Variational Optimization Problem	$\mathcal{F}(\mu_t)$	$\frac{\delta \mathcal{F}(\mu_t)}{\delta \mu_t}$	$\nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t) = \nabla \frac{\delta \mathcal{F}(\mu_t)}{\delta \mu_t}$	$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0$
Potential Energy	$\int V(x)\, d\mu_t(x)$	V	∇V	$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\nabla V\, \mu_t)$
Negative Entropy	$\frac{1}{2}\int log(\mu_t(x))\, d\mu_t(x)$	$\frac{1}{2} log(\mu_t) + \frac{1}{2}$	$\frac{1}{2}\nabla log(\mu_t)$	$\frac{\partial \mu_t}{\partial t} = \frac{1}{2}\Delta \mu_t$ **heat equation**
KL Divergence	$\mathrm{KL}(\mu_t \mid \exp(-V(x))) = \int V(x) + log(\mu_t(x)) d\mu_t(x)$	$V + log(\mu_t) + 1$	$\nabla log\left(\frac{\mu_t}{\exp(-V(x))}\right)$	$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t + \nabla \cdot (\nabla V\, \mu_t)$ **Fokker-Planck Equation**

Time discretization of gradient flow in Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$

There are two time discretization methods of the wasserstein gradient flow

- **Wasserstein gradient descent**

$$\mu_{m+1} = (Id - \gamma \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_m))_{\#}\mu_m$$

i.e. Forward Euler (explicit)

- **JKO flow of JKO scheme**

$$\mu_{m+1} \in \mathrm{JKO}_\gamma \mathcal{F}(\mu_m) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \left\{ \mathcal{F}(\mu) + \frac{1}{2\gamma} \mathcal{W}_2^2(\mu, \mu_m) \right\}$$

i.e. Backward Euler (implicit)

Remark.
There exists another proof which shows that as the step size $\gamma \to 0$, the JKO flow becomes the forward euler wasserstein gradient descent.

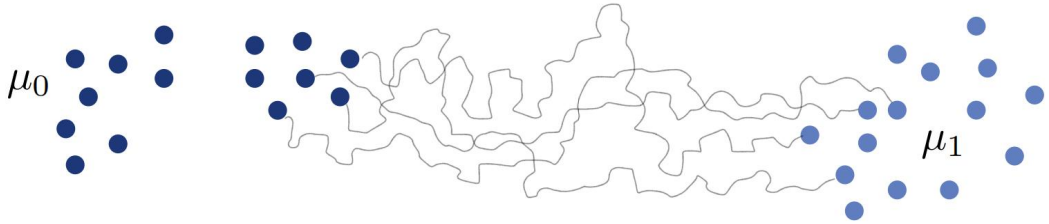
First Approache - Disintegration of measures
We will see the connection between the static and dynamic formulation of **Schrödinger Bridge**.

Second Approache - Girsanov Theorem
We will reveal the stochastic control perspective of the **Schrödinger Bridge**.

IPF (iterative proportional fitting) Algorithm
IPF (iterative proportional fitting) is the classical solver for **Schrödinger Bridge**.

Path space and measure

Definition: Path space $\overline{\mathcal{X}}$ is an abstract space which contains all possible paths (i.e. curves) γ: $[0,1] \to \mathcal{X}$ on the space $\mathcal{X}$ (i.e. $\mathbb{R}^d$). Informally, the dynamics of "particles" between input measures μ_0 and μ_1 at times $t = 0,1$ is described by the path measure $\overline{\pi} \in \mathcal{M}_+^1(\overline{\mathcal{X}})$.



Example
path measure for discrete measures

OT over the path space

Definition: Given the measures $\mu_0 \in \mathcal{M}_+^1(\mathcal{X})$ and $\mu_1 \in \mathcal{M}_+^1(\mathcal{Y})$, the dynamical version of classical **OT** seeks a path measure $\overline{\pi}$ which minimizes

$$\mathcal{W}_2^2(\mu_0, \beta) = \min_{\overline{\pi} \in \overline{\mathcal{U}}(\mu_0, \mu_1)} \left\{ \int_{\overline{\mathcal{X}}} \mathcal{L}(\gamma)^2 \, d\overline{\pi}(\gamma) \right\},$$

where

$$\overline{\mathcal{U}}(\mu_0, \mu_1) \stackrel{\text{def}}{=} \{ \overline{\pi} \in \mathcal{M}_+^1(\overline{\mathcal{X}}) \mid \overline{P}_{0\#}\overline{\pi} = \mu_0 \text{ and } \overline{P}_{1\#}\overline{\pi} = \mu_1 \}$$

Here, $\overline{P}_{0\#}$ and $\overline{P}_{1\#}$ are the push-forwards of the projections.

Remark.
The connection between the optimal coupling π^* and the optimal path measure $\overline{\pi}^*$ is that $\overline{\pi}^*$ only gives mass to geodesics joining pairs of points in proportion prescribed by π^*. For the discrete measure, we have

$$\pi^* = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)} \qquad \overline{\pi}^* = \sum_{i,j} \mathbf{P}_{i,j} \delta_{\gamma_{x_i, y_j}}$$

Introduction to the Schrödinger Bridge Problem (Discrete Measure Version)

Advanced topics in Optimal Transport

Recall the Schrödinger Bridge Problem (static)

Definition: Given the discrete measures $a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$, the **static Schrödinger Bridge** seeks a coupling $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_i)}$ which minimizes

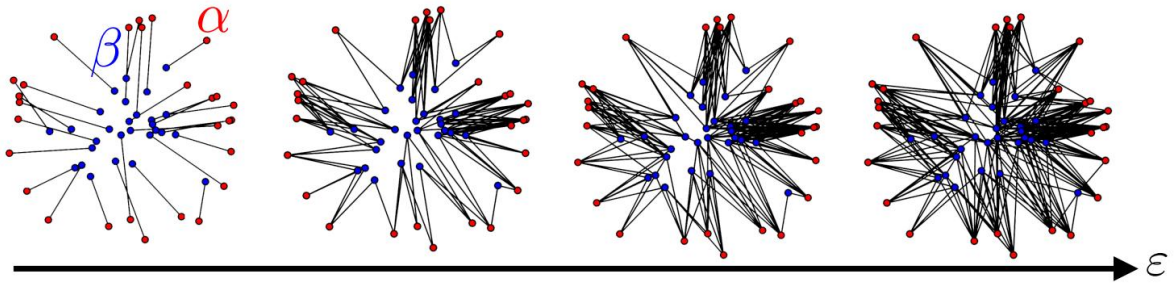
$$\mathbf{P}_\varepsilon = \mathrm{Proj}^{\mathrm{KL}}_{\mathbf{U}(\mathbf{a},\mathbf{b})}(\mathbf{K}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \mathrm{KL}(\mathbf{P}|\mathbf{K})$$

where

$$\mathbf{U}(\mathbf{a},\mathbf{b}) \stackrel{\text{def}}{=} \{ \mathbf{P} \in \mathbb{R}_+^{n \times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \}$$

$$\mathrm{KL}(\mathbf{P}|\mathbf{K}) \stackrel{\text{def}}{=} \sum_{i,j} \mathbf{P}_{i,j} \log\left(\frac{\mathbf{P}_{i,j}}{\mathbf{K}_{i,j}}\right) - \mathbf{P}_{i,j} + \mathbf{K}_{i,j}$$

$$\mathbf{K}_{i,j} \stackrel{\text{def}}{=} e^{-\frac{\mathbf{C}_{i,j}}{\varepsilon}}$$



Example

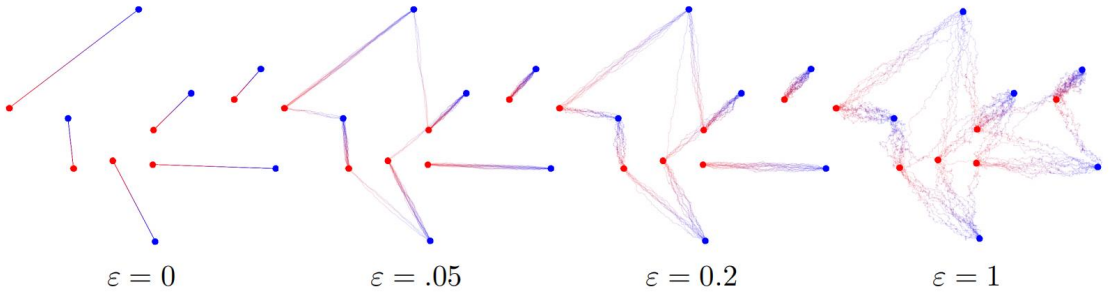
Schrödinger Bridge Problem (dynamic)

Definition: Given the discrete measures $a = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$, the dynamic **Schrödinger Bridge** seeks a path measure $\overline{\pi}$ which minimizes

$$\min_{\overline{\pi} \in \overline{\mathcal{U}}(\mu_0, \mu_1)} \mathrm{KL}(\overline{\pi}|\overline{\mathcal{K}})$$

where $\overline{\mathcal{K}}$ defines the path measure of a Brownian motion with ε volatity ($dX = \varepsilon d\mathbb{W}$) and

$$\overline{\mathcal{U}}(a,\beta) \stackrel{\text{def}}{=} \{ \overline{\pi} \in \mathcal{M}_+^1(\overline{\mathcal{X}}) \mid \overline{P}_{0\#}\overline{\pi} = a \text{ and } \overline{P}_{1\#}\overline{\pi} = \beta \}$$



Example

Introduction to the Schrödinger Bridge Problem (Arbitrary Measure Version)

Advanced topics in Optimal Transport

Recall the Schrödinger Bridge Problem (static)

Definition: Given the measures $\mu_0 \in \mathcal{M}_+^1(\mathcal{X})$ and $\mu_1 \in \mathcal{M}_+^1(\mathcal{Y})$, the **static Schrödinger Bridge** seeks a coupling π which minimizes

$$\pi^* \stackrel{\text{def}}{=} \arg\min_{\pi \in \mathcal{U}(\mu_0,\mu_1)} \mathrm{KL}(\mathbf{P}|\mathbf{K})$$

where

$$\mathcal{U}(\mu_0,\mu_1) \stackrel{\text{def}}{=} \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) \mid P_{\mathcal{X}\#}\pi = \mu_0, P_{\mathcal{Y}\#}\pi = \mu_1 \right\}$$

$$\mathrm{KL}(\mathbf{P}|\mathbf{K}) \stackrel{\text{def}}{=} \int \log\left(\frac{d\mathbf{P}(x,y)}{d\mathbf{K}(x,y)}\right) d\mathbf{P}(x,y)$$

$$d\mathbf{K}(x,y) \stackrel{\text{def}}{=} e^{-\frac{c(x,y)}{\varepsilon}} d\alpha(x) d\beta(y)$$

Schrödinger Bridge Problem (dynamic)

Definition: Given the measures $\mu_0 \in \mathcal{M}_+^1(\mathcal{X})$ and $\mu_1 \in \mathcal{M}_+^1(\mathcal{Y})$, the **dynamic Schrödinger Bridge** seeks a path measure $\overline{\pi}$ which minimizes

$$\min_{\overline{\pi} \in \overline{\mathcal{U}}(\mu_0,\mu_1)} \mathrm{KL}(\overline{\pi}|\overline{\mathcal{K}})$$

where $\overline{\mathcal{K}}$ defines the path measure of a Brownian motion with ε volatity ($dX = \varepsilon d\mathbb{W}$) and

$$\overline{\mathcal{U}}(\mu_0,\mu_1) \stackrel{\text{def}}{=} \{ \overline{\pi} \in \mathcal{M}_+^1(\overline{\mathcal{X}}) \mid \overline{P}_{0\#}\overline{\pi} = \mu_0 \text{ and } \overline{P}_{1\#}\overline{\pi} = \mu_1 \}$$

Is there any connection between static and dynamic formulation?

First Approache - Disintegration of measures

$$\mathbb{P}(path) = \mathbb{P}(path \mid x(0) = x_0, x(1) = x_1) \cdot \mathbb{P}_{0,t}(x_0, x_1)$$

$$\min_{\overline{\pi} \in \overline{\mathcal{U}}(\mu_0, \mu_1)} \mathrm{KL}(\overline{\pi} | \overline{\mathcal{K}}) = \int \log\left(\frac{d\mathbb{P}}{d\mathbb{W}}\right) d\mathbb{P} = \int \log\left(\frac{d\mathbb{P}_{0,t}(x_0, x_1)}{d\mathbb{W}^{\varepsilon}_{0,t}(x_0, x_1)}\right) d\mathbb{P}_{0,t}(x_0, x_1) + \int \log\left(\frac{d\mathbb{P}(path|x_0, x_1)}{d\mathbb{W}^{\varepsilon}(path|x_0, x_1)}\right) d\mathbb{P}(path|x_0, x_1)$$

Schrödinger Bridge Problem (static)

$$= 0 \text{ for } d\mathbb{P}(path|x_0, x_1) = d\mathbb{W}^{\varepsilon}(path|x_0, x_1)$$

Brownian Bridge

Schrödinger Bridge Problem (dynamic)

Equivalent to the Entropy Regularized OT

Proof.

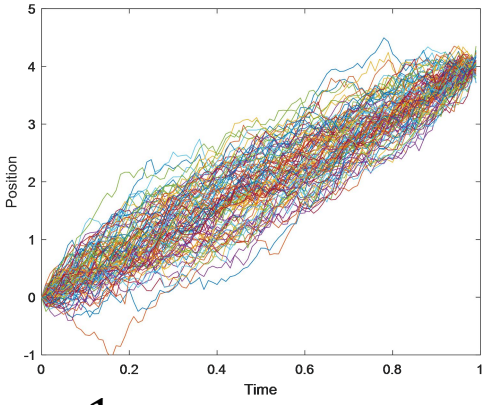
$$\min_{\pi \in \mathcal{U}(a,\beta)} \int \log\left(\frac{d\mathbb{P}_{0,t}(x_0, x_1)}{d\mathbb{W}_{0,t}(x_0, x_1)}\right) d\mathbb{P}_{0,t}(x_0, x_1)$$

$$= -\mathbf{H}(\pi) - \int \log d\mathbb{W}^{\varepsilon}_{0,t}(x_0, x_1)\, d\mathbb{P}_{0,t}(x_0, x_1)$$

since $d\mathbb{W}^{\varepsilon}_{0,t}(x_0, x_1) = d\mathbb{Q}(x_0)\, \mathcal{N}(x_1|x_0, \varepsilon)$

$$= -\mathbf{H}(\pi) - \int -\frac{\|x_0 - x_1\|^2}{2\varepsilon}\, d\mathbb{P}_{0,t}(x_0, x_1) + const$$

$$= \int \|x_0 - x_1\|^2\, d\mathbb{P}_{0,t}(x_0, x_1) - 2\varepsilon\, \mathbf{H}(\pi) \quad \blacksquare$$



$$dX_t = \frac{1}{1-t}(X_1 - X_t)dt + \varepsilon dW$$

$$\mathbb{P}(X_t|X_0) = \mathcal{N}((1-t)X_0 + tX_1, t(1-t))$$

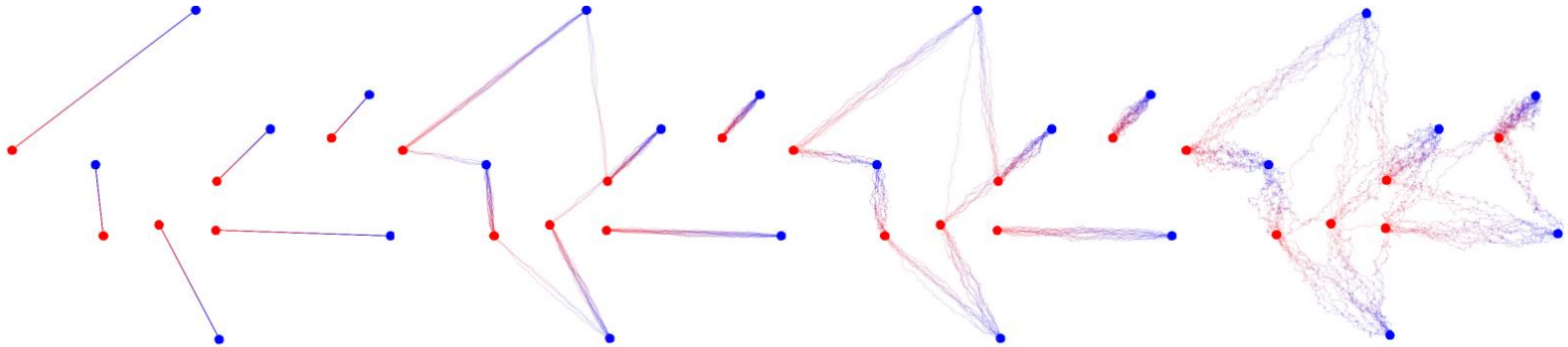
First Approache - Disintegration of measures

$$\mathbb{P}(path) = \mathbb{P}(path \mid x(0) = x_0, x(1) = x_1) \cdot \mathbb{P}_{0,t}(x_0, x_1)$$

$$\min_{\overline{\pi} \in \overline{\mathcal{U}}(\mu_0, \mu_1)} \mathrm{KL}(\overline{\pi} | \overline{\mathcal{K}}) = \int \log\left(\frac{d\mathbb{P}}{d\mathbb{W}}\right) d\mathbb{P} = \int \log\left(\frac{d\mathbb{P}_{0,t}(x_0, x_1)}{d\mathbb{W}^{\varepsilon}_{0,t}(x_0, x_1)}\right) d\mathbb{P}_{0,t}(x_0, x_1) + \int \log\left(\frac{d\mathbb{P}(path|x_0, x_1)}{d\mathbb{W}^{\varepsilon}(path|x_0, x_1)}\right) d\mathbb{P}(path|x_0, x_1)$$

Remark.

As $\varepsilon \to 0$, the whole problem becomes more deterministics (less stochastic and less regularized)



$\varepsilon = 0$ $\varepsilon = .05$ $\varepsilon = 0.2$ $\varepsilon = 1$

First Approache - Disintegration of measures

$$\mathbb{P}(path) = \mathbb{P}(path \mid x(0) = x_0, x(1) = x_1) \cdot \mathbb{P}_{0,t}(x_0, x_1)$$

Schrödinger Bridge (dynamic) = **Schrödinger Bridge (static)** × **Brownian Bridge**

Remark.

1. If we have the optimal path measure $\overline{\pi}^*$, then $\pi^* = \overline{\pi}^*_{0,1}$.
2. If we have the optimal coupling π^*, then $\overline{\pi}^* = \pi^* \cdot \mathbb{W}^{\varepsilon}(path|x_0, x_1)$.
3. In the discrete setting, we have

$$\pi^* = \sum_{i,j} \mathbf{P}^*_{i,j} \delta_{(x_i, y_j)} \quad \text{and} \quad \overline{\pi}^* = \sum_{i,j} \mathbf{P}^*_{i,j} \mathbb{W}_{x_i, y_j}$$

Second Approache - Girsanov Theorem

Given the SDE

$$dX_t = v(t, X_t)dt + \varepsilon dW$$

The law $\mathbb{P}$ of X_t and law $\mathbb{W}$ of dW has the following property

$$\int_{paths} \log\left(\frac{d\mathbb{P}}{d\mathbb{W}}\right) d\mathbb{P} = \frac{1}{2\varepsilon^2}\int \|v(t, X_t)\|^2 d\mathbb{P}$$

$$= \frac{1}{2\varepsilon^2}\int_{\mathbb{R}^d}\int_0^1 \|v(t, X_t)\|^2 \, dt \, d\mu_t(x) = \frac{1}{2\varepsilon^2}\mathbb{E}\left\{\int_0^1 \|v(t, X_t)\|^2 \, dt\right\}$$

Entropy Regularization OT (Stochastic Control Perspective)

$$\min_v \mathbb{E}\left\{\int_0^1 \|v(t, X_t)\|^2 \, dt\right\}$$

$$dX_t = v(t, X_t)dt + \varepsilon dW$$

$$X_0 \sim \mu_0, \qquad X_1 \sim \mu_1$$

Fokker-Planck Equation

Entropy Regularization OT (Fluid Dynamic Perspective)

$$\min_{v\,\mu_t} \int_{\mathbb{R}^d}\int_0^1 \|v(t, X_t)\|^2 \, dt \, d\mu_t(x)$$

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v\,\mu_t) - \frac{\varepsilon^2}{2}\Delta\mu_t = 0$$

$$\mu_{t=0} \sim \mu_0, \qquad \mu_{t=1} \sim \mu_1$$

Second Approache - Girsanov Theorem

Entropy Regularization OT
Dynamic Schrödinger Bridge
(Stochastic Control Perspective)

$$\min_{v} \mathbb{E}\left\{\int_0^1 \|v(t,X_t)\|^2\, dt\right\}$$
$$dX_t = v(t,X_t)dt + \varepsilon dW$$
$$X_0 \sim \mu_0, \qquad X_1 \sim \mu_1$$

Fokker-Planck Equation

Entropy Regularization OT
(Fluid Dynamic Perspective)

$$\min_{v\,\mu_t} \int_{\mathbb{R}^d} \int_0^1 \|v(t,X_t)\|^2\, dt\, d\mu_t(x)$$
$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v\,\mu_t) - \frac{\varepsilon^2}{2}\Delta \mu_t = 0$$
$$\mu_{t=0} \sim \mu_0, \qquad \mu_{t=1} \sim \mu_1$$

Entropy Regularization

Entropy Regularization

Kantorovich OT
(Stochastic Control Perspective)

$$\min_{v} \mathbb{E}\left\{\int_0^1 \|v(t,X_t)\|^2\, dt\right\}$$
$$dX_t = v(t,X_t)dt$$
$$X_0 \sim \mu_0, \qquad X_1 \sim \mu_1$$

Kantorovich OT
(Benamou - Brenier Formulation)

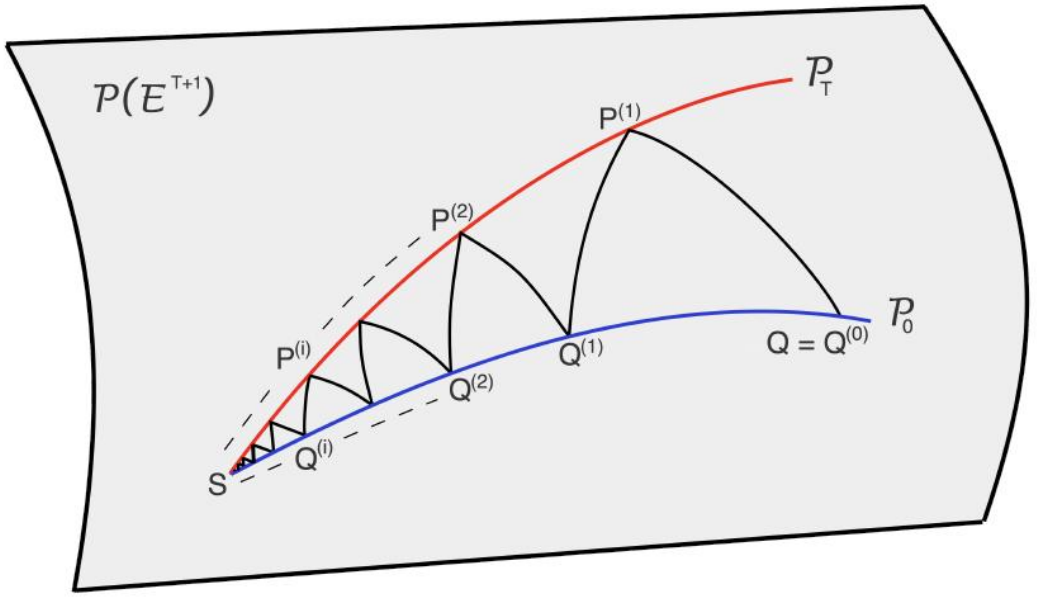
$$\min_{v\,\mu_t} \int_{\mathbb{R}^d} \int_0^1 \|v(t,X_t)\|^2\, dt\, d\mu_t(x)$$
$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v\,\mu_t) = 0$$
$$\mu_{t=0} \sim \mu_0, \qquad \mu_{t=1} \sim \mu_1$$

Remark.
The **Stochastic Control Perspective** has closed form solutions, but this is far more beyond this tutorial.

IPF (iterative proportional fitting) Algorithm

set μ^0 follow the brownian motion

$$\mu^{2n+1} = \arg\min\{\mathrm{KL}(\mu \mid \mu^{2n}), \quad \mu_{t=1} = \mu_1\}$$
$$\mu^{2n+2} = \arg\min\{\mathrm{KL}(\mu \mid \mu^{2n+1}), \quad \mu_{t=0} = \mu_0\}$$



Remark.

1. Since this problem is strongly convex, this IPF can always converge to the optimal solution.
2. IPF can be seen as the continuous version of Sinkorn algorithm.

- Computational Optimal Transport (Chapter 1)

  https://arxiv.org/abs/1803.00567

- ICML 2023 Tutorial Optimal Transport in Learning, Control, and Dynamical Systems

  https://icml.cc/virtual/2023/tutorial/21559

- Statistical Applications of Wasserstein Gradient Flows

  https://www.youtube.com/watch?v=EBA0NyY4Myc

- A SURVEY ON OPTIMAL TRANSPORT FOR MACHINE LEARNING: THEORY AND APPLICATIONS

  https://arxiv.org/pdf/2106.01963

- Recent Advances in Optimal Transport for Machine Learning

  https://arxiv.org/pdf/2306.16156

- POT doc

  https://pythonot.github.io/

Recommended reading materials