Introduction to optimal transport for Bayesian statistics

Part I

Hugo Lavenant

Bocconi University

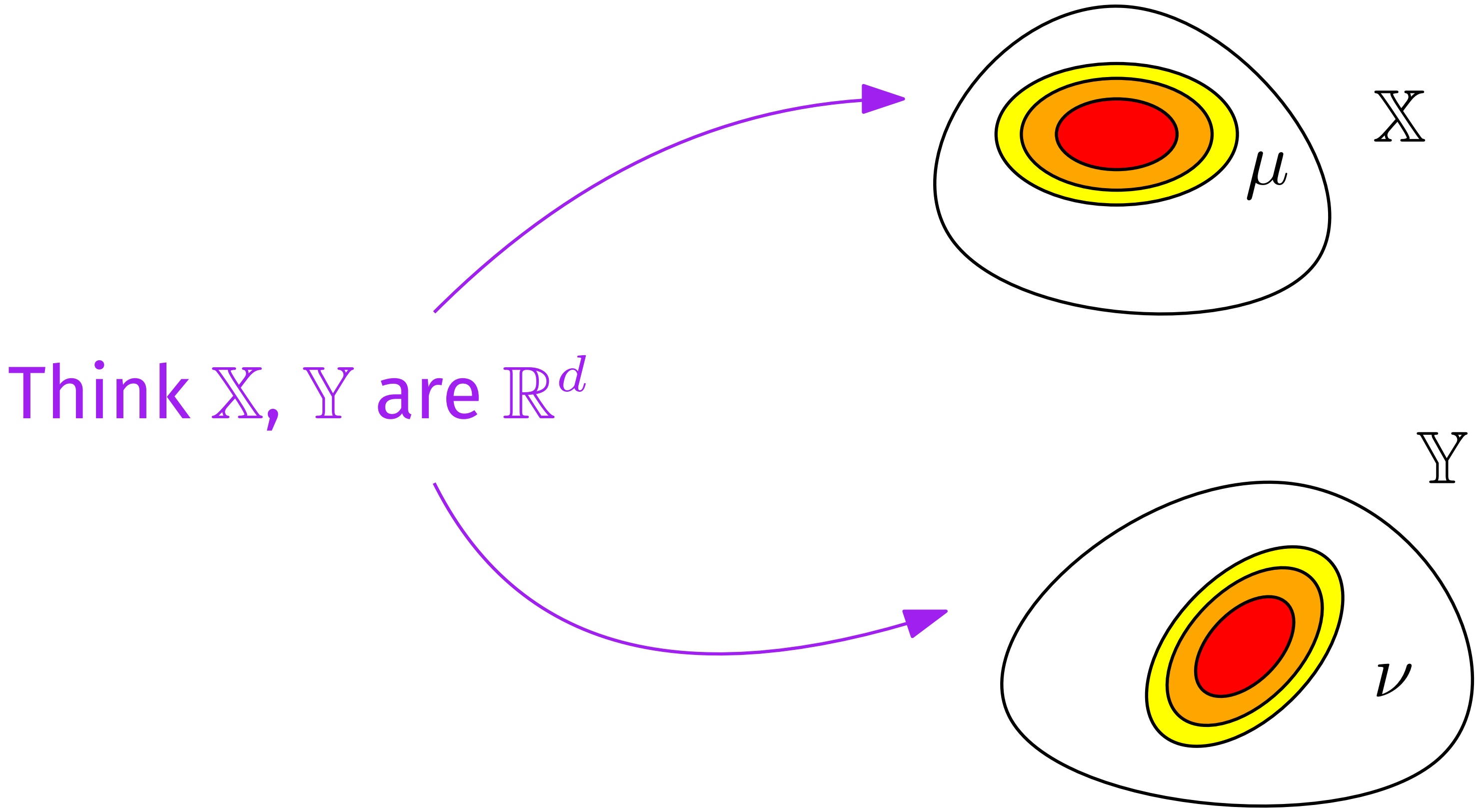
2024 ISBA World Meeting

Venice (Italy), July 1, 2024

The optimal transport problem

Inputs:

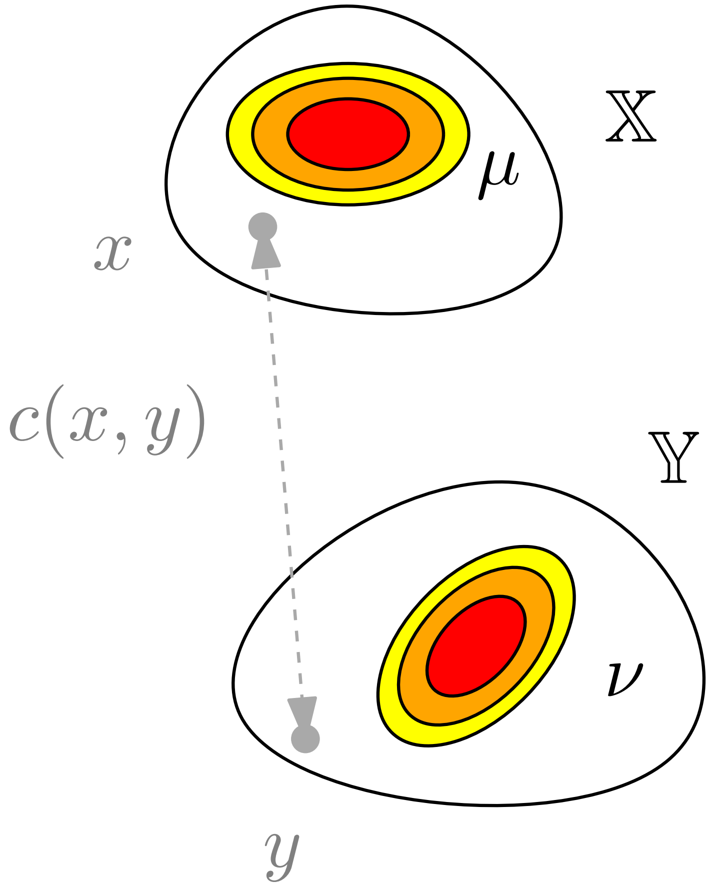
- μ, ν two probability distributions (laws of $X \sim \mu$, and $Y \sim \nu$) on $\mathbb{X}$, $\mathbb{Y}$.



The optimal transport problem

Inputs:

- μ, ν two probability distributions (laws of $X \sim \mu$, and $Y \sim \nu$) on $\mathbb{X}$, $\mathbb{Y}$.
- $c : \mathbb{X} \times \mathbb{Y} \to (-\infty, +\infty]$ "cost function."



The optimal transport problem

Inputs:

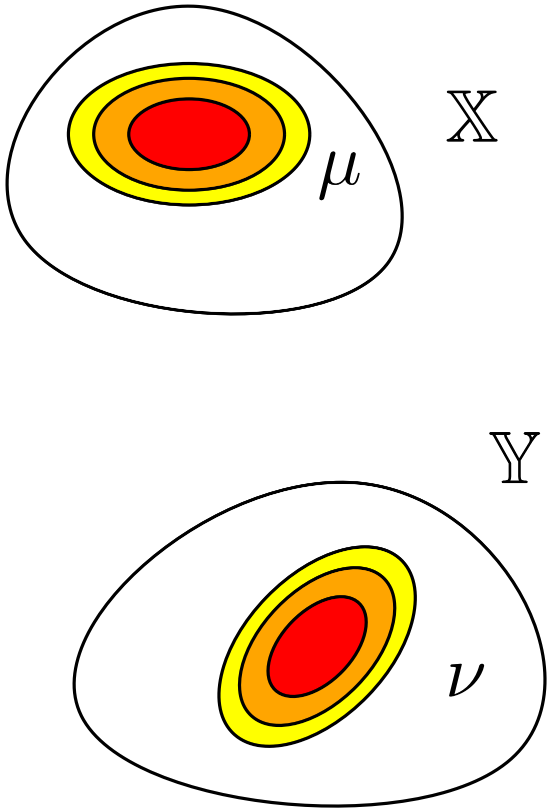
- μ, ν two probability distributions (laws of $X \sim \mu$, and $Y \sim \nu$) on $\mathbb{X}$, $\mathbb{Y}$.
- $c : \mathbb{X} \times \mathbb{Y} \to (-\infty, +\infty]$ "cost function."

Optimal transport problem

$$\inf_{\pi} \left\{ \mathbb{E}_{(X,Y)\sim\pi}(c(X,Y)) \ \text{ s.t. } \ \pi \in \Pi(\mu,\nu) \right\}$$

$\Pi(\mu,\nu)$ "couplings": laws of random variables on $\mathbb{X} \times \mathbb{Y}$ whose marginals are μ, ν.

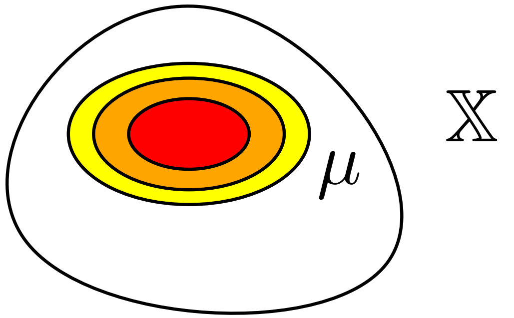
(e.g. the independent coupling is in $\Pi(\mu,\nu)$)



The optimal transport problem

Inputs:

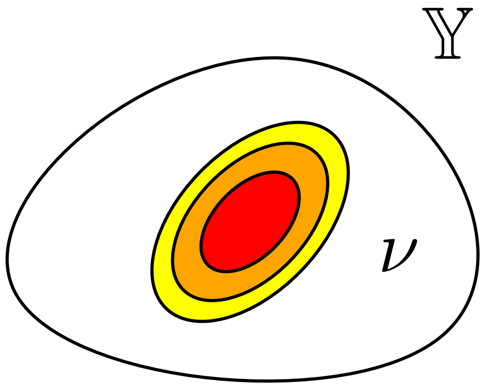
- μ, ν two probability distributions (laws of $X \sim \mu$, and $Y \sim \nu$) on $\mathbb{X}$, $\mathbb{Y}$.
- $c : \mathbb{X} \times \mathbb{Y} \to (-\infty, +\infty]$ "cost function."



Optimal transport problem

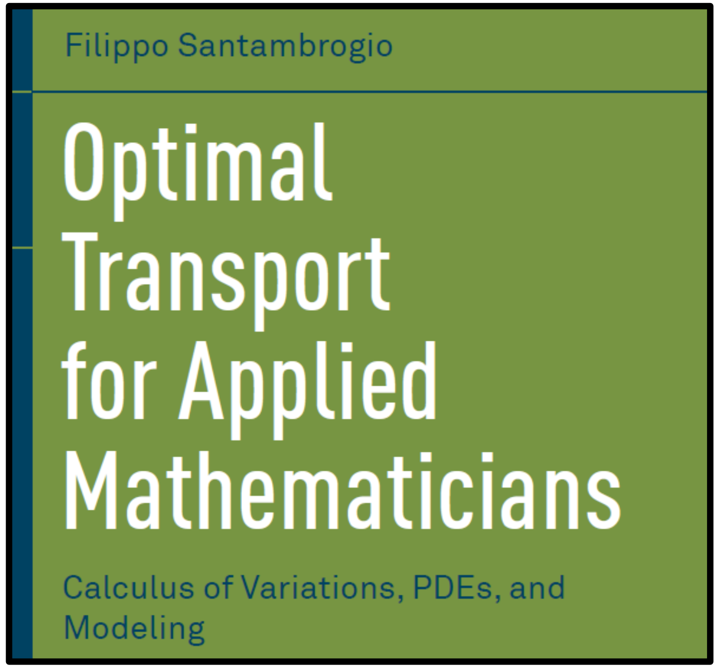
$$\inf_{\pi} \left\{ \mathbb{E}_{(X,Y)\sim\pi}(c(X,Y)) \text{ s.t. } \pi \in \Pi(\mu,\nu) \right\}$$

Outputs:



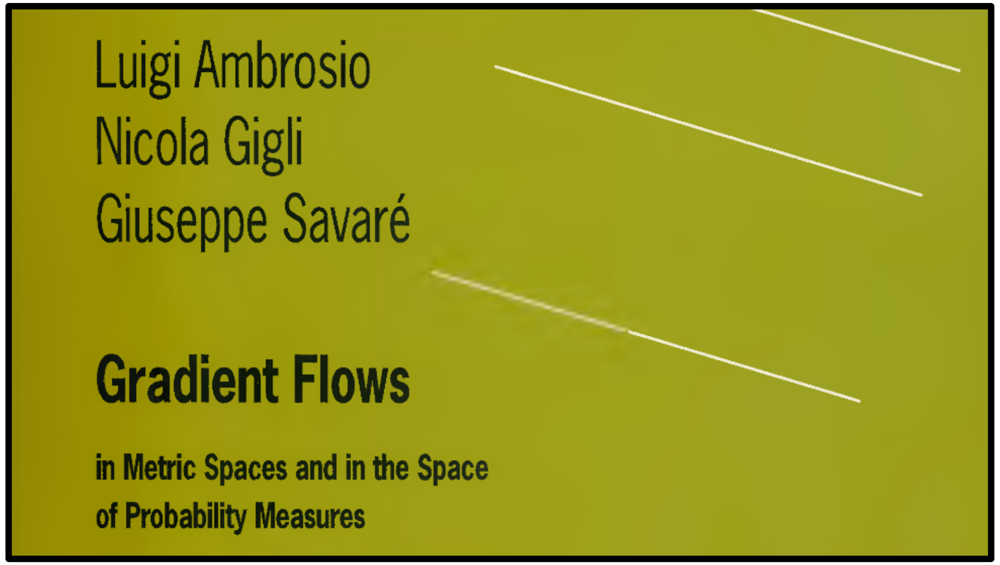
- $\mathcal{T}_c(X,Y) = \mathcal{T}_c(\mu,\nu)$ value of the transport, in $[-\infty, +\infty]$.
- π^* (if it exists) optimal coupling realizing the infimum.

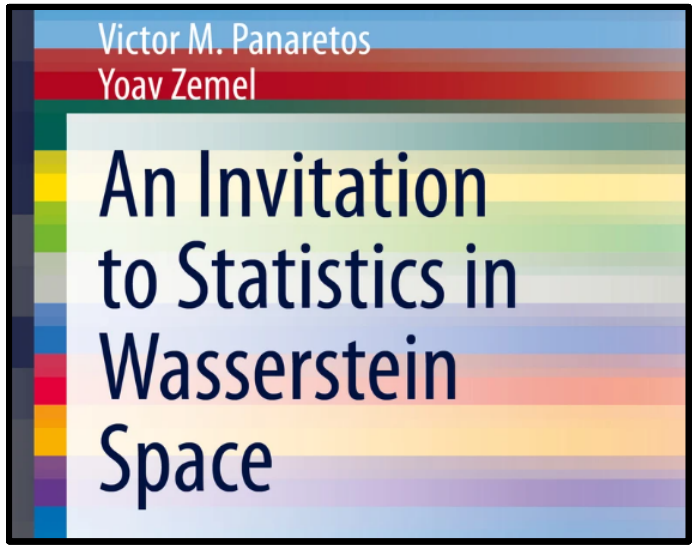
Textbooks



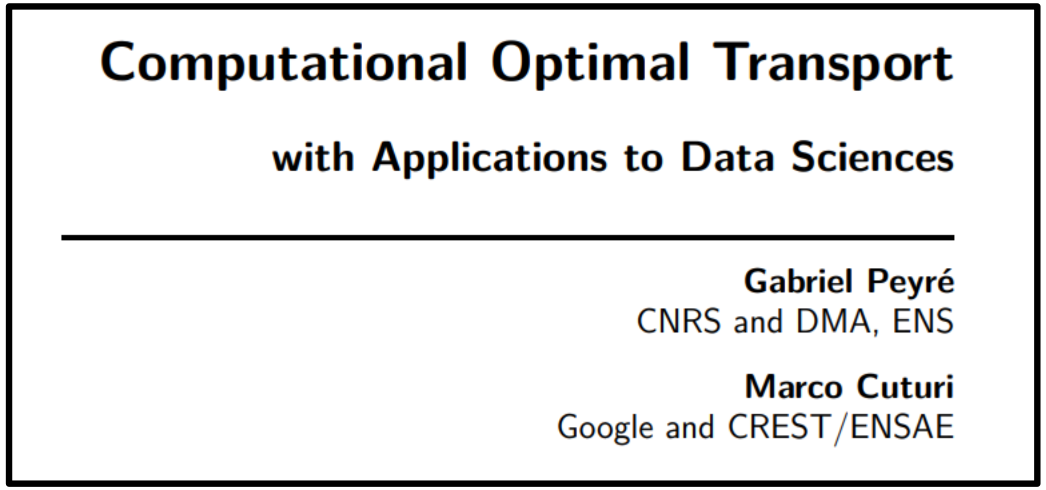
Towards calculus of variations, presents applications of OT

Theory oriented, good to find sharp results





Title self explanatory!



Computational Optimal Transport

with Applications to Data Sciences

Gabriel Peyré
CNRS and DMA, ENS

Marco Cuturi
Google and CREST/ENSAE

Title self explanatory, gentle introduction to OT in first chapters

1 - Particular case: discrete measures

2 - Particular case: one dimensional

3 - Duality

4 - Monotoncity, structure of optimal couplings

Interlude: Gaussian measures

5 - Wasserstein distances

6 - Numerical methods

1 - Particular case: discrete measures

2 - Particular case: one dimens

[Peyré & Cuturi, Chapter 2]

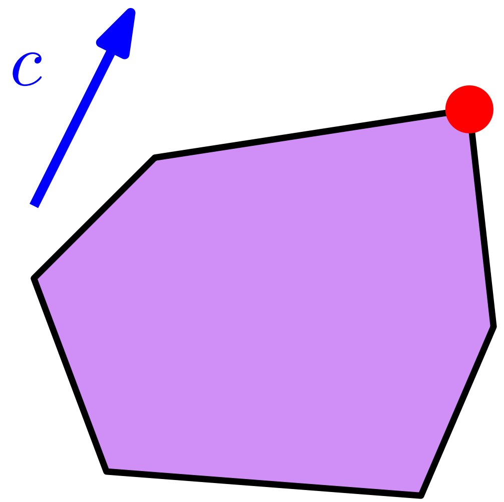
3 - Duality

4 - Monotoncity, structure of optimal couplings

Interlude: Gaussian measures

5 - Wasserstein distances

6 - Numerical methods



The “discrete” case

Assume X, Y takes a finite number of values $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

$$\mathbb{P}(X = x_i) = a_i, \qquad \mathbb{P}(Y = y_j) = b_j.$$

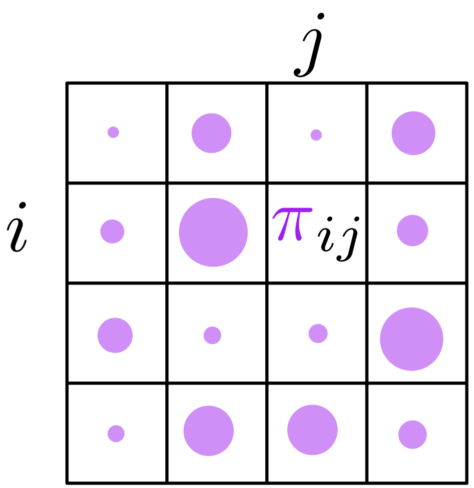
The “discrete” case

Assume X, Y takes a finite number of values $x_1, \dots, x_n$ and $y_1, \dots, y_m$.

$$\mathbb{P}(X = x_i) = a_i, \qquad \mathbb{P}(Y = y_j) = b_j.$$

$\pi \in \Pi(\mu, \nu)$ described by $\pi_{ij} = \mathbb{P}(X = x_i$ and $Y = y_j)$.

(e.g. independent coupling $\pi_{ij} = a_i b_j$.)

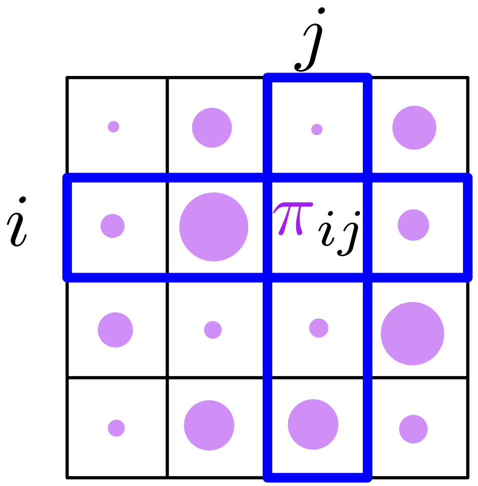


The "discrete" case

Assume X, Y takes a finite number of values $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

$$\mathbb{P}(X = x_i) = a_i, \qquad \mathbb{P}(Y = y_j) = b_j.$$

$\pi \in \Pi(\mu, \nu)$ described by $\pi_{ij} = \mathbb{P}(X = x_i$ and $Y = y_j)$.



Constraints:

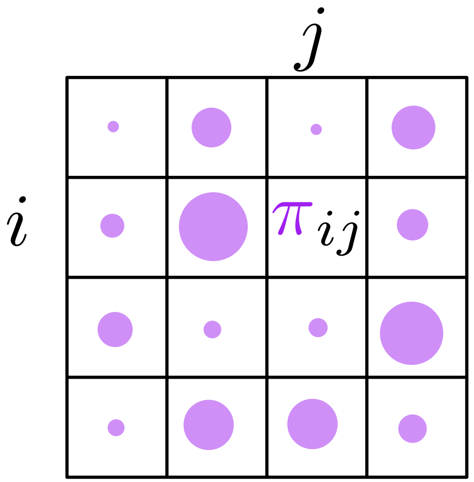
- $\pi_{ij} \geq 0$ for all i, j.
- $\begin{cases} \sum_j \pi_{ij} = a_i \\ \sum_i \pi_{ij} = b_j \end{cases}$

The “discrete” case

Assume X, Y takes a finite number of values $x_1, \dots, x_n$ and $y_1, \dots, y_m$.

$$\mathbb{P}(X = x_i) = a_i, \qquad \mathbb{P}(Y = y_j) = b_j.$$

$\pi \in \Pi(\mu, \nu)$ described by $\pi_{ij} = \mathbb{P}(X = x_i$ and $Y = y_j)$.



Constraints:

- $\pi_{ij} \geq 0$ for all i, j.
- $\begin{cases} \sum_j \pi_{ij} = a_i \\ \sum_i \pi_{ij} = b_j \end{cases}$

Objective

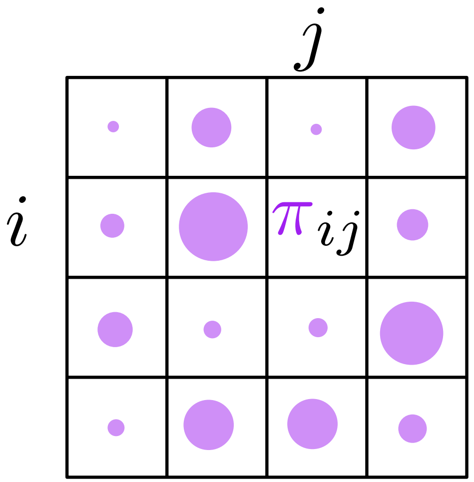
$$\mathbb{E}(c(X, Y)) = \sum_{i,j} \pi_{ij} c(x_i, y_j).$$

The “discrete” case

Assume X, Y takes a finite number of values $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

$$\mathbb{P}(X = x_i) = a_i, \qquad \mathbb{P}(Y = y_j) = b_j.$$

$\pi \in \Pi(\mu, \nu)$ described by $\pi_{ij} = \mathbb{P}(X = x_i$ and $Y = y_j)$.



Constraints:

- $\pi_{ij} \geq 0$ for all i, j.
- $\begin{cases} \sum_j \pi_{ij} = a_i \\ \sum_i \pi_{ij} = b_j \end{cases}$

Objective

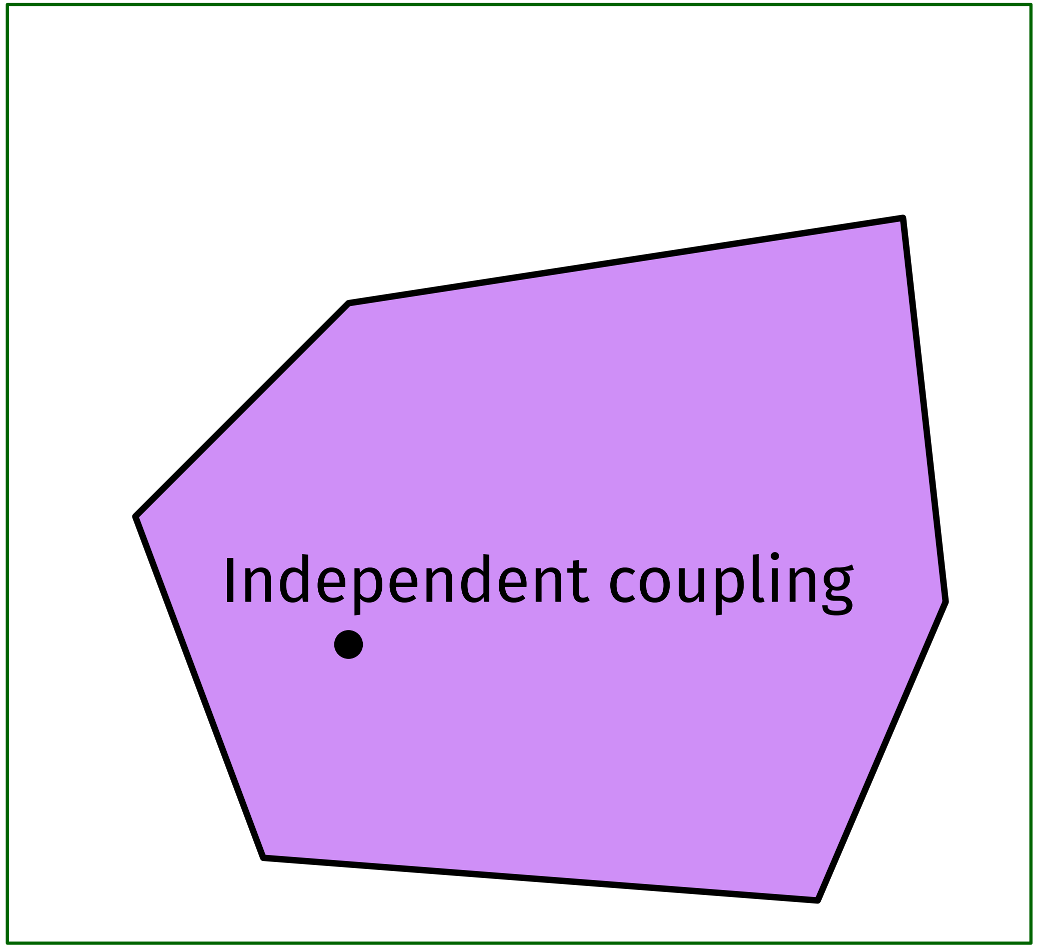
$$\mathbb{E}(c(X, Y)) = \sum_{i,j} \pi_{ij} c(x_i, y_j).$$

This is a **Linear Program** (linear objective, linear equalities and inequalities as constraints).

How to visualize a linear program?

Space $\mathbb{R}^{nm}$ of $n \times m$ matrices.

How to visualize a linear program?

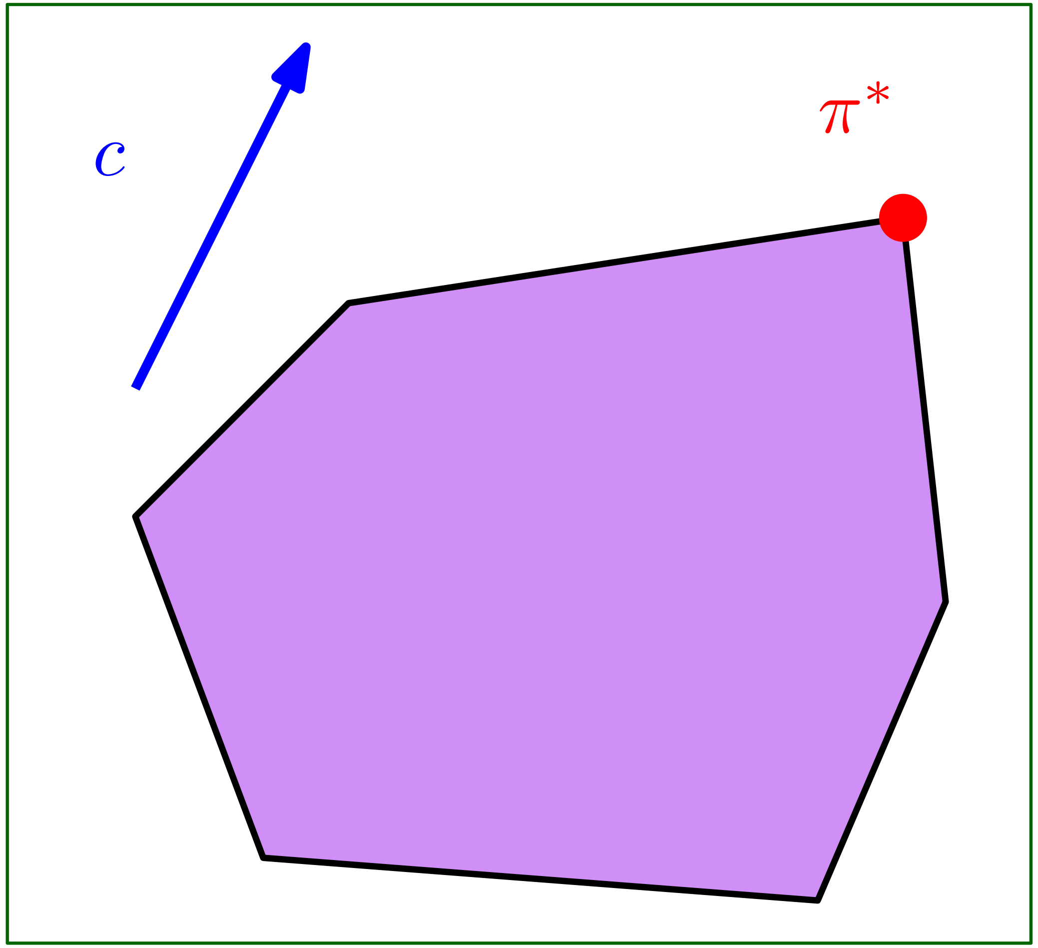


Space $\mathbb{R}^{nm}$ of $n \times m$ matrices.

- Convex polytope $\Pi(\mu, \nu)$ of admissible couplings.

$$\pi \geq 0, \quad \begin{cases} \sum_i \pi_{ij} \text{ given}, \\ \sum_j \pi_{ij} \text{ given} \end{cases}$$

How to visualize a linear program?



Space $\mathbb{R}^{nm}$ of $n \times m$ matrices.

- Convex polytope $\Pi(\mu, \nu)$ of admissible couplings.

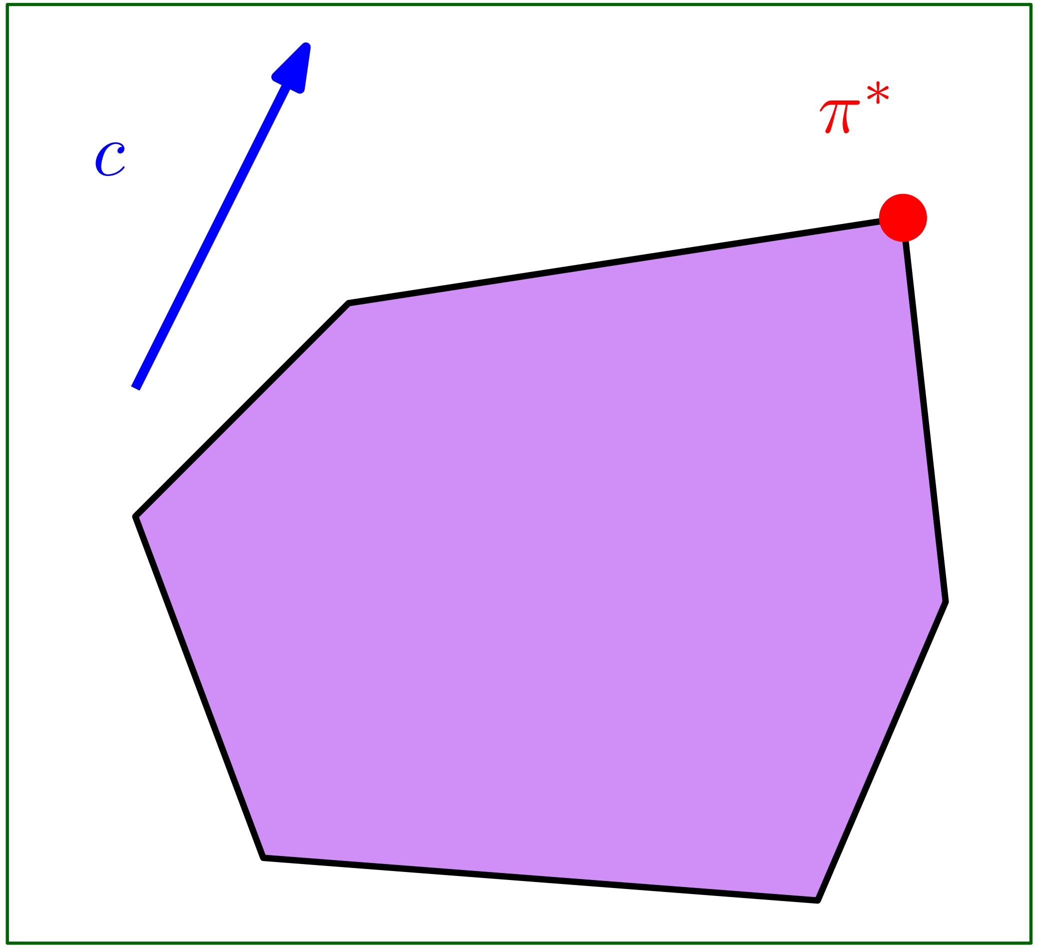
$$\pi \geq 0, \quad \begin{cases} \sum_i \pi_{ij} \text{ given,} \\ \sum_j \pi_{ij} \text{ given} \end{cases}$$

- Vector $c = c(x_i, y_j)$: direction to maximize

Maximize $\sum_{ij} \pi_{ij} c(x_i, y_j)$.

- Optimal coupling π^*!

How to visualize a linear program?



Space $\mathbb{R}^{nm}$ of $n \times m$ matrices.

- Convex polytope $\Pi(\mu, \nu)$ of admissible couplings.

$$\pi \geq 0, \quad \begin{cases} \sum_i \pi_{ij} \text{ given,} \\ \sum_j \pi_{ij} \text{ given} \end{cases}$$

- Vector $c = c(x_i, y_j)$: direction to maximize

Maximize $\sum_{ij} \pi_{ij} c(x_i, y_j)$.

- Optimal coupling π^*!

In general: optimal transport is an **infinite dimensional linear program**.

1 - Particular case: discrete measures

2 - Particular case: one dimensional

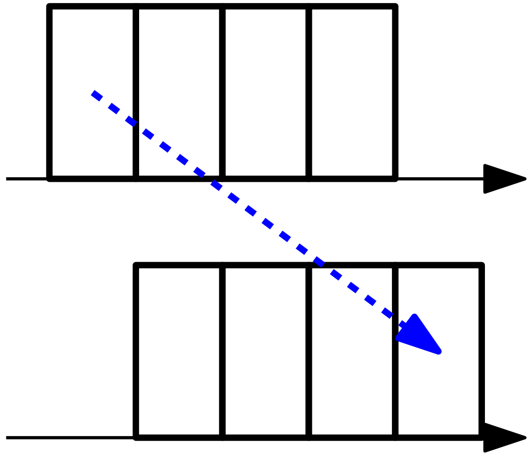
3 - Duality

[Santambrogio, Chapter 2]

4 - Monotoncity, structure of optimal couplings

Interlude: Gaussian measures

5 - Wasserstein distances



6 - Numerical methods

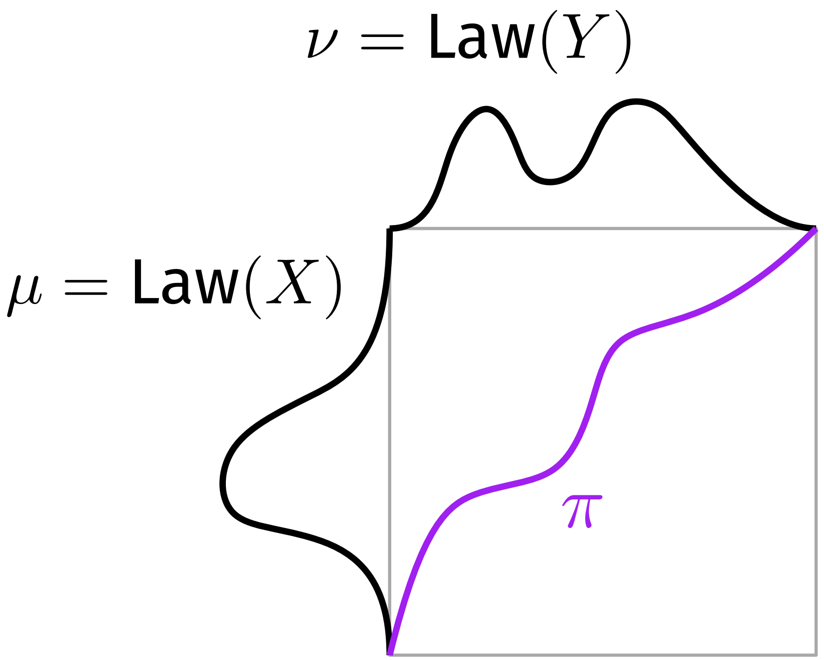
The increasing coupling

Restrict to $\mathbb{X}$, $\mathbb{Y}$ to be $\mathbb{R}$.

Lemma. If X, Y are two random variables on $\mathbb{R}$, there exists a unique coupling $(X, Y) \sim \pi$ between them which is increasing:

If $(X_1, Y_1) \sim \pi$ and $(X_2, Y_2) \sim \pi$ then

$$X_1 \leq X_2 \quad \Rightarrow \quad Y_1 \leq Y_2$$



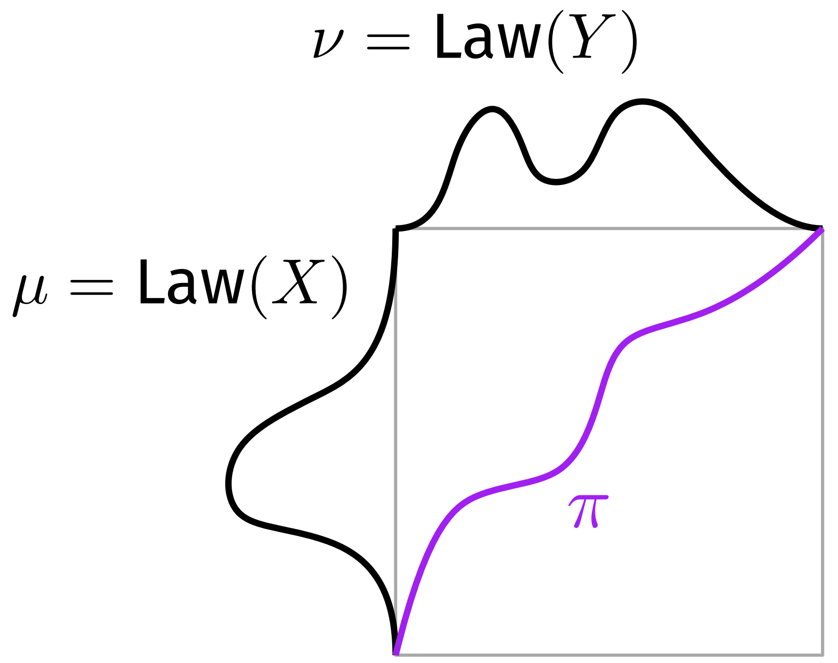
The increasing coupling

Restrict to $\mathbb{X}$, $\mathbb{Y}$ to be $\mathbb{R}$.

Lemma. If X, Y are two random variables on $\mathbb{R}$, there exists a unique coupling $(X, Y) \sim \pi$ between them which is increasing:

If $(X_1, Y_1) \sim \pi$ and $(X_2, Y_2) \sim \pi$ then

$$X_1 \leq X_2 \quad \Rightarrow \quad Y_1 \leq Y_2$$



Lemma. If X is atomless, then the increasing coupling is given by

$$Y = T(X)$$

where $T = F_Y^{-1} \circ F_X$ is a non-decreasing map.

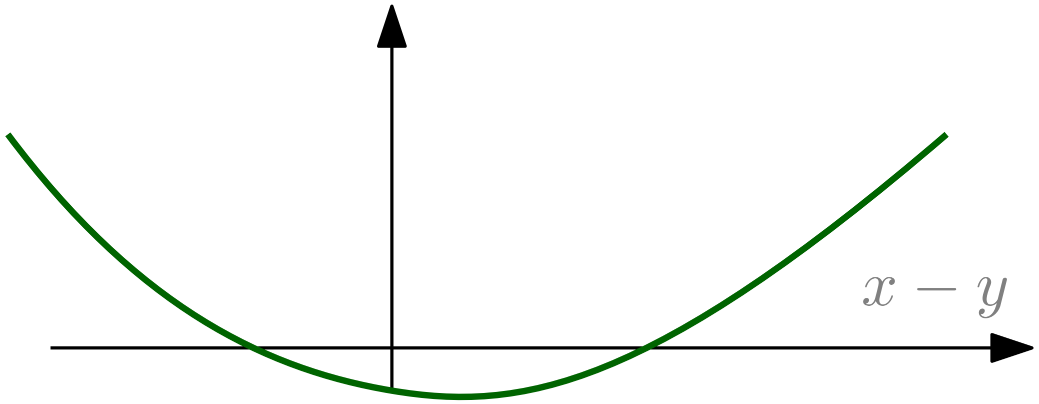
$F_X(t) = \mathbb{P}(X \leq t)$ c.d.f. of X,
F_Y^{-1} quantile function of Y.

Optimality of the increasing coupling

Proposition. Assume $c(x,y) = h(x-y)$ with $h : \mathbb{R} \to \mathbb{R}$ convex. Then **the increasing coupling** between X and Y **is optimal**, and the value is:

$$\mathcal{T}_c(X,Y) = \int_0^1 h(F_X^{-1}(u) - F_Y^{-1}(u))\,\mathrm{d}u.$$

Graph h



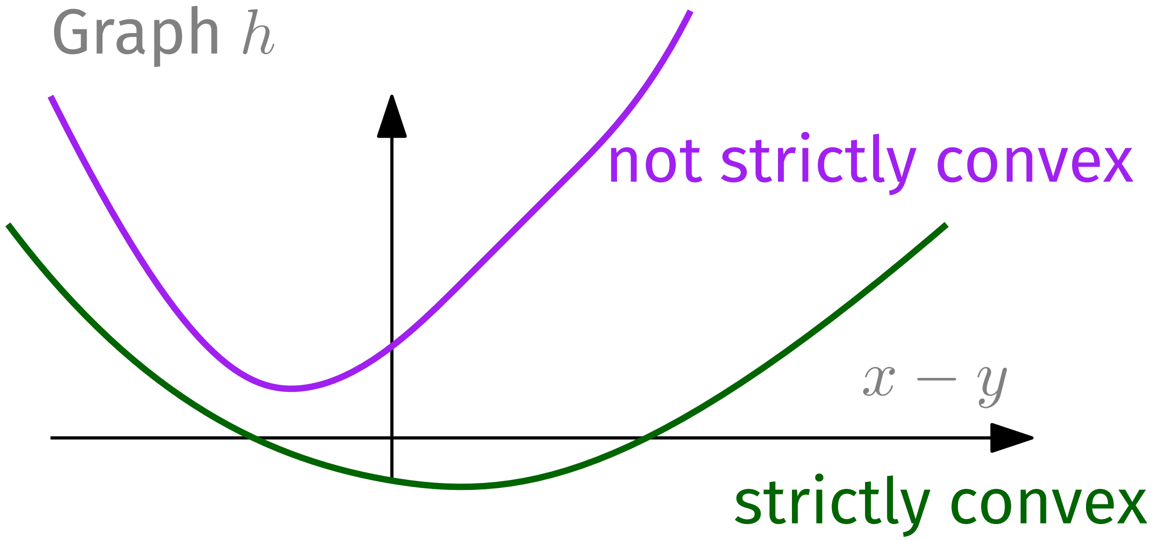
h convex: $h((1-t)a+tb) \le (1-t)h(a) + th(b)$ for all a, b and $t \in [0,1]$.

Optimality of the increasing coupling

Proposition. Assume $c(x,y) = h(x-y)$ with $h : \mathbb{R} \to \mathbb{R}$ convex. Then **the increasing coupling** between X and Y **is optimal**, and the value is:

$$\mathcal{T}_c(X,Y) = \int_0^1 h(F_X^{-1}(u) - F_Y^{-1}(u))\,\mathrm{d}u.$$

If in addition h is **strictly convex** then the increasing coupling is the **unique** optimal coupling.



h convex: $h((1-t)a+tb) \leq (1-t)h(a) + th(b)$ for all a, b and $t \in [0,1]$.

h strictly convex: there is equality above iff $a = b$ or $t \in \{0,1\}$.

The case of the Monge cost in dimension one

$\mathbb{X} = \mathbb{Y} = \mathbb{R}$ and $c(x, y) = |x - y|$.

$\rightsquigarrow$ Previous case with $h(a) = |a|$ not strictly convex.

Proposition. In this case, the value is

$$\begin{aligned} \mathcal{T}_c(X, Y) &= \int_0^1 |F_X^{-1}(u) - F_Y^{-1}(u)| \, \mathrm{d}u \\ &= \int_{-\infty}^{+\infty} |F_X(t) - F_Y(t)| \, \mathrm{d}t. \end{aligned}$$

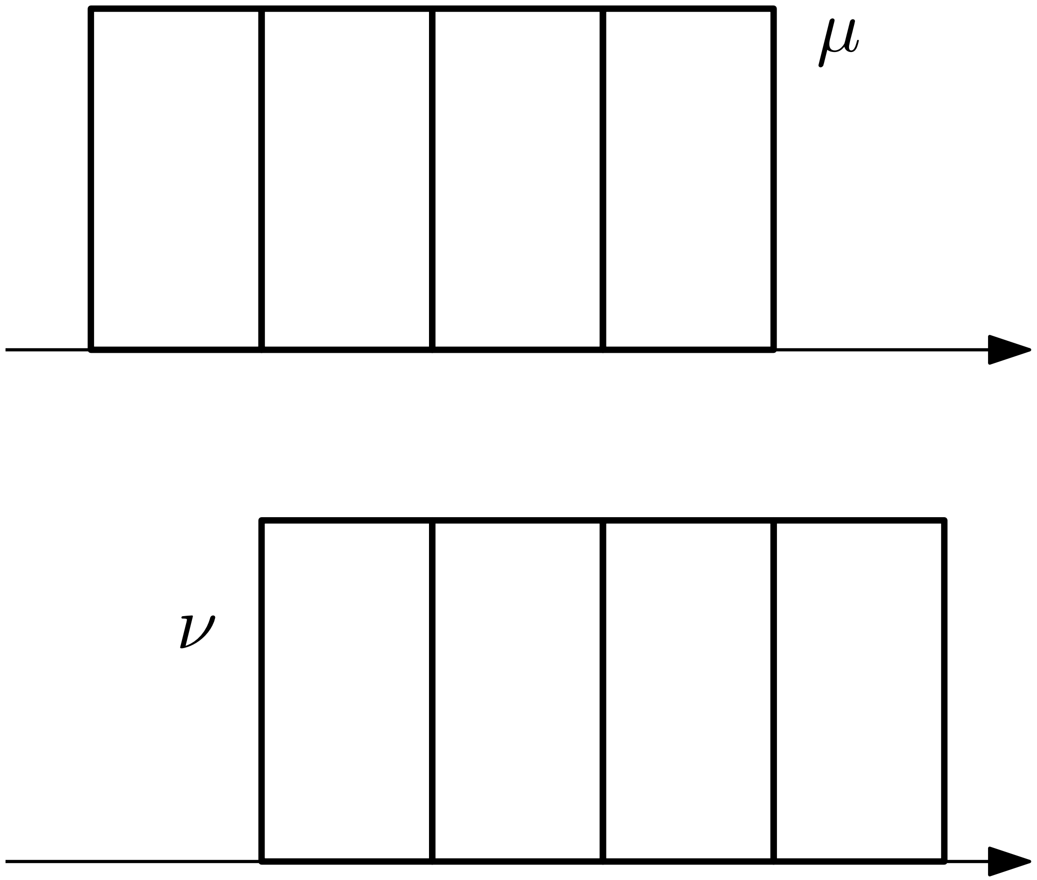
The case of the Monge cost in dimension one

$\mathbb{X} = \mathbb{Y} = \mathbb{R}$ and $c(x, y) = |x - y|$.

$\rightsquigarrow$ Previous case with $h(a) = |a|$ not strictly convex.

Proposition. In this case, the value is

$$\mathcal{T}_c(X, Y) = \int_0^1 |F_X^{-1}(u) - F_Y^{-1}(u)| \,\mathrm{d}u$$
$$= \int_{-\infty}^{+\infty} |F_X(t) - F_Y(t)| \,\mathrm{d}t.$$



But there is more than one optimal transport coupling: “book shifting example”.

The case of the Monge cost in dimension one

$\mathbb{X} = \mathbb{Y} = \mathbb{R}$ and $c(x, y) = |x - y|$.

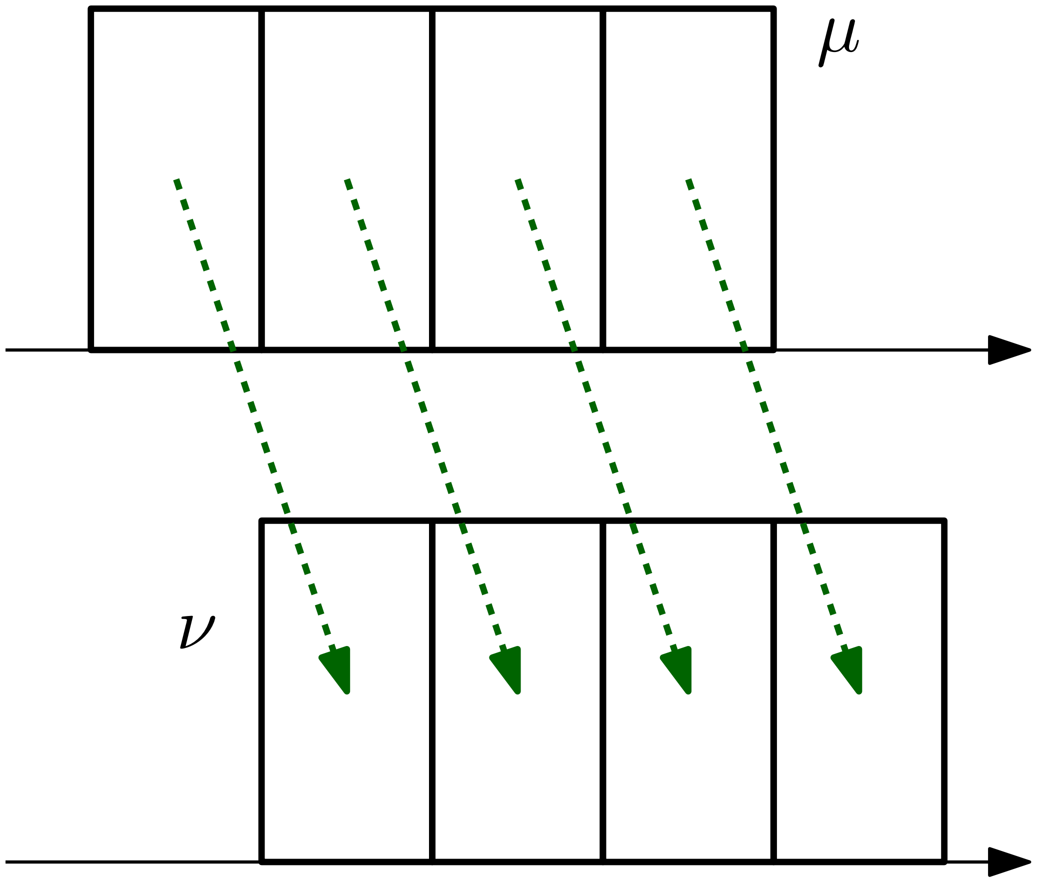
$\rightsquigarrow$ Previous case with $h(a) = |a|$ not strictly convex.

Proposition. In this case, the value is

$$\mathcal{T}_c(X, Y) = \int_0^1 |F_X^{-1}(u) - F_Y^{-1}(u)| \, \mathrm{d}u$$
$$= \int_{-\infty}^{+\infty} |F_X(t) - F_Y(t)| \, \mathrm{d}t.$$

But there is more than one optimal transport coupling: "book shifting example".

Increasing coupling



The case of the Monge cost in dimension one

$\mathbb{X} = \mathbb{Y} = \mathbb{R}$ and $c(x, y) = |x - y|$.

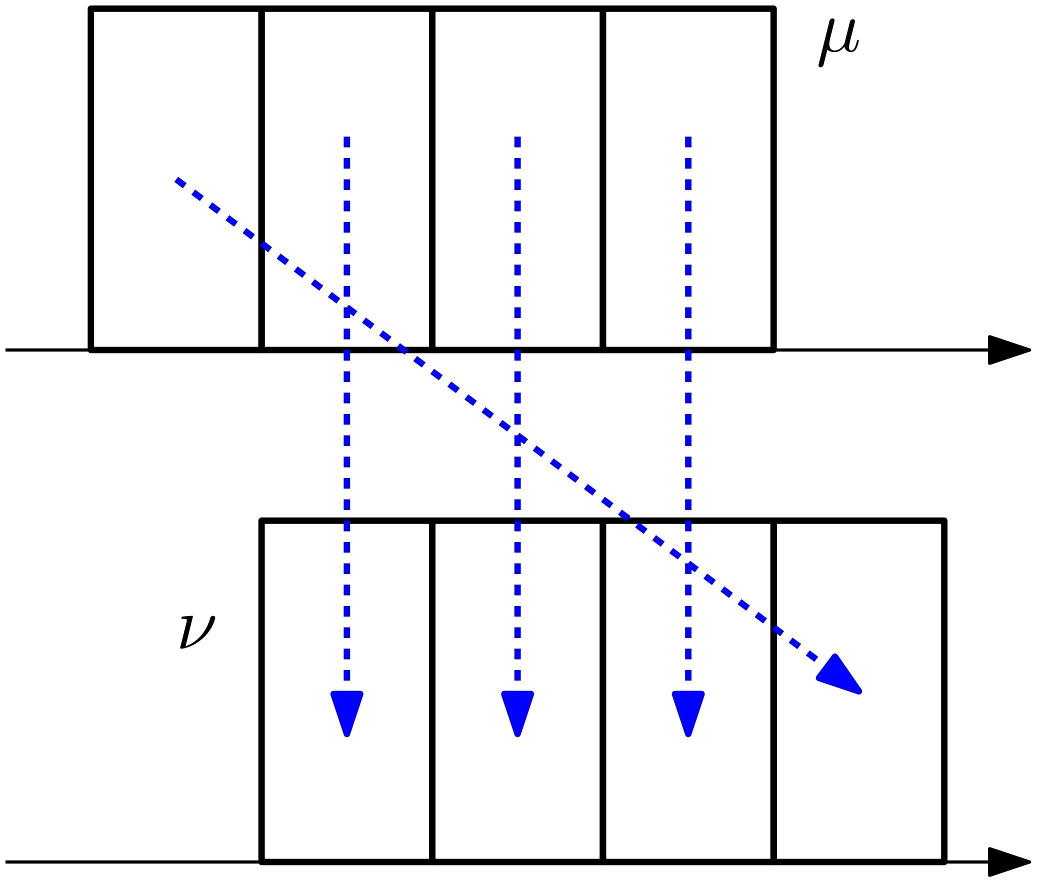
$\rightsquigarrow$ Previous case with $h(a) = |a|$ not strictly convex.

Proposition. In this case, the value is

$$\mathcal{T}_c(X, Y) = \int_0^1 |F_X^{-1}(u) - F_Y^{-1}(u)| \, \mathrm{d}u$$
$$= \int_{-\infty}^{+\infty} |F_X(t) - F_Y(t)| \, \mathrm{d}t.$$

But there is more than one optimal transport coupling: "book shifting example".

Another optimal coupling



1 - Particular case: discrete measures

2 - Particular case: one dimensional

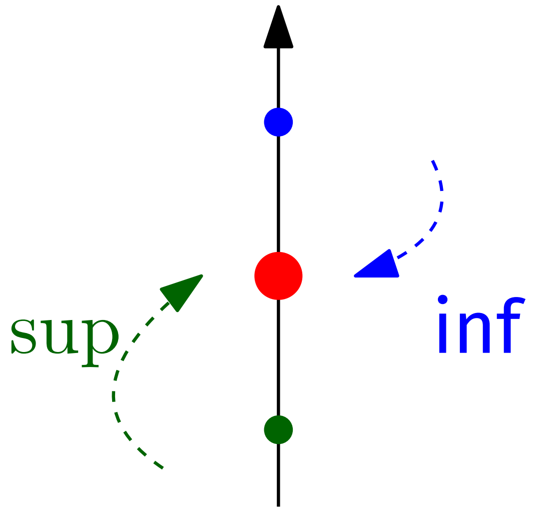
$\boxed{\textbf{3 - Duality}}$

4 - [Santambrogio, Chapter 1]
[Peyré & Cuturi, Chapter 2]

Interlude: Gaussian measures

5 - Wasserstein distances

6 - Numerical methods



Duality for the Monge cost

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

Maybe you already know the formula:

$$\boxed{\mathcal{T}_c(X, Y) = \sup_f \left\{ \mathbb{E}(f(X)) - \mathbb{E}(f(Y)) \text{ s.t. } f : \mathbb{R}^d \to \mathbb{R} \text{ is a } 1\text{-Lipschitz function} \right\}}$$

means $|f(x) - f(y)| \leq \|x - y\|$ for all x, y

Duality for the Monge cost

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

Maybe you already know the formula:

$$\boxed{\mathcal{T}_c(X, Y) = \sup_f \left\{ \mathbb{E}(f(X)) - \mathbb{E}(f(Y)) \text{ s.t. } f : \mathbb{R}^d \to \mathbb{R} \text{ is a 1-Lipschitz function} \right\}}$$

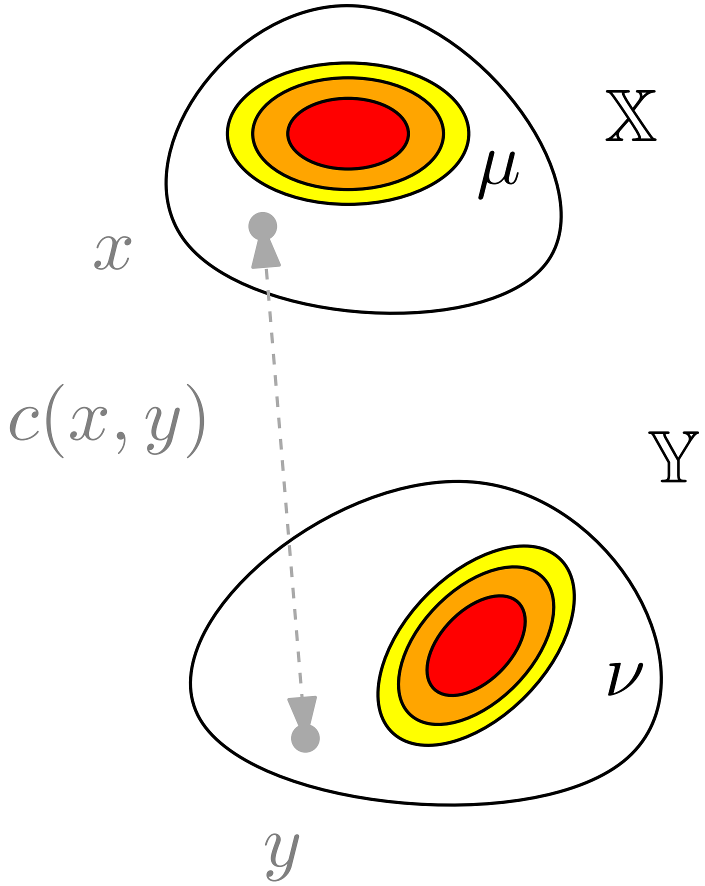
means $|f(x) - f(y)| \leq \|x - y\|$ for all x, y

How to generalize it to other cost functions? **Answer**: comes from the concept of **duality** in convex optimization.

The dual problem in the general case

Same inputs: μ, ν distributions on $\mathbb{X}, \mathbb{Y}$ and c cost function.

Metaphor: outsource the transport to a contractor.



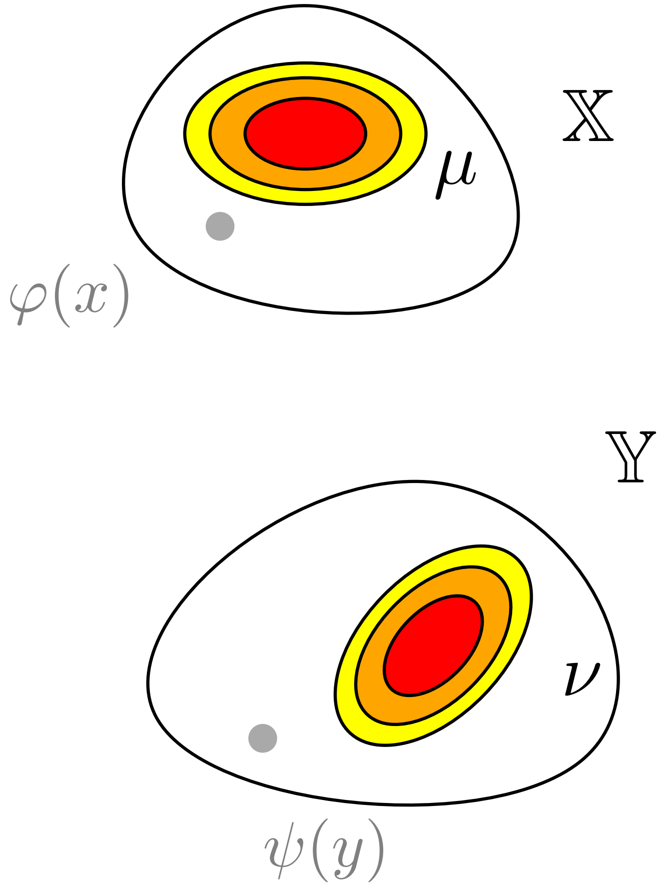
The dual problem in the general case

Same inputs: μ, ν distributions on $\mathbb{X}, \mathbb{Y}$ and c cost function.

Metaphor: outsource the transport to a contractor.

New unknowns:

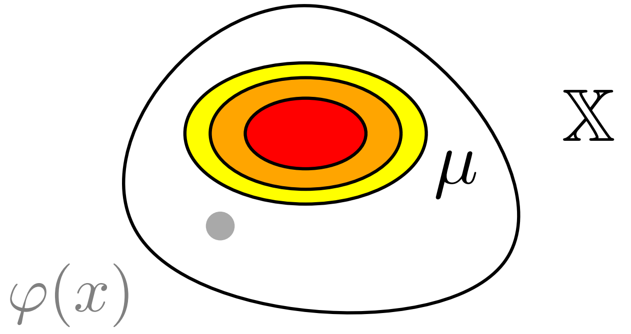
- $\varphi : \mathbb{X} \to \mathbb{R}$, with $\varphi(x)$ cost of "loading" one unit of mass in x.
- $\psi : \mathbb{Y} \to \mathbb{R}$, with $\psi(y)$ cost of "unloading" one unit of mass in y.



The dual problem in the general case

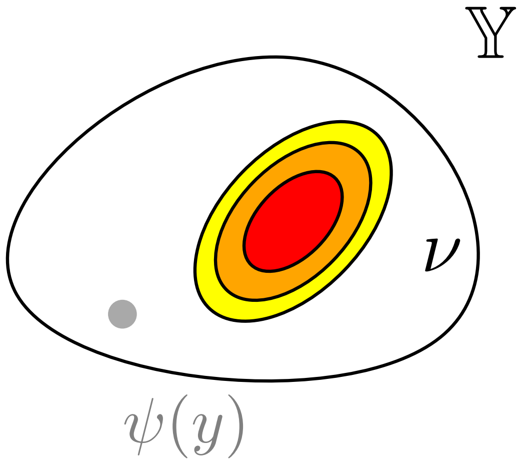
Same inputs: μ, ν distributions on $\mathbb{X}, \mathbb{Y}$ and c cost function.

Metaphor: outsource the transport to a contractor.



New unknowns:

- $\varphi : \mathbb{X} \to \mathbb{R}$, with $\varphi(x)$ cost of "loading" one unit of mass in x.
- $\psi : \mathbb{Y} \to \mathbb{R}$, with $\psi(y)$ cost of "unloading" one unit of mass in y.



Constraints of the contractor:

- For every x, y we have $\varphi(x) + \psi(y) \leq c(x, y)$.

Profit of the contractor: $\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y))$.

Contractor's problem: **maximize** profit given the constraints.

Weak and strong duality

Primal problem

$$\inf_\pi \{\mathbb{E}_\pi(c(X,Y)) \;:\; \pi \in \Pi(\mu,\nu)\}$$

π probability distribution on $\mathbb{X} \times \mathbb{Y}$

Dual problem

$$\sup\{\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \;:$$

$$\varphi(x) + \psi(y) \leq c(x,y) \text{ for all } x, y \,\}$$

φ, ψ functions on $\mathbb{X}$ and $\mathbb{Y}$

Weak and strong duality

Primal problem

$$\inf_\pi \{\mathbb{E}_\pi(c(X,Y)) \;:\; \pi \in \Pi(\mu,\nu)\}$$

π probability distribution on $\mathbb{X} \times \mathbb{Y}$

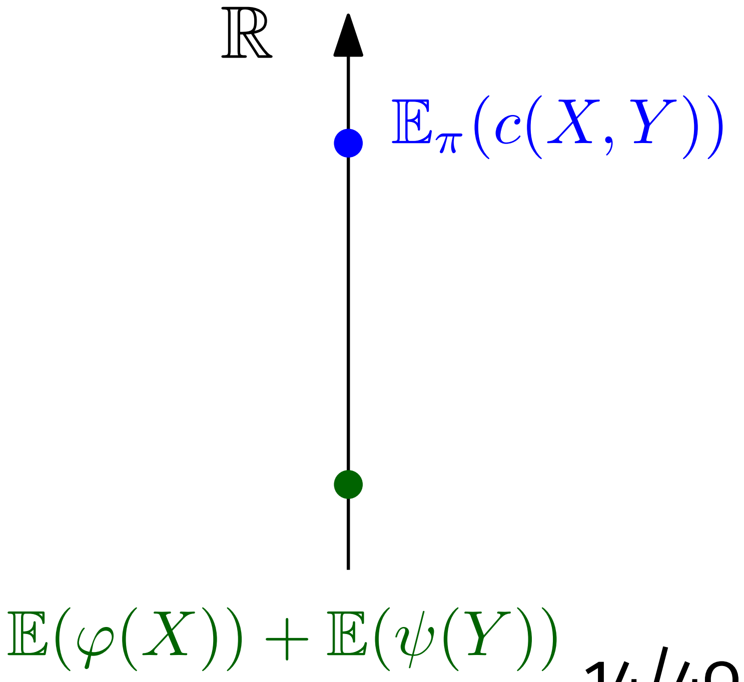
Dual problem

$$\sup\{\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \;: \\ \varphi(x) + \psi(y) \le c(x,y) \text{ for all } x, y\}$$

φ, ψ functions on $\mathbb{X}$ and $\mathbb{Y}$

Lemma (weak duality) For any π, φ, ψ satisfying the constraints,

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \le \mathbb{E}_\pi(c(X,Y)).$$



Weak and strong duality

Primal problem

$$\inf_{\pi}\{\mathbb{E}_\pi(c(X,Y)) \ : \ \pi \in \Pi(\mu,\nu)\}$$

π probability distribution on $\mathbb{X}\times\mathbb{Y}$

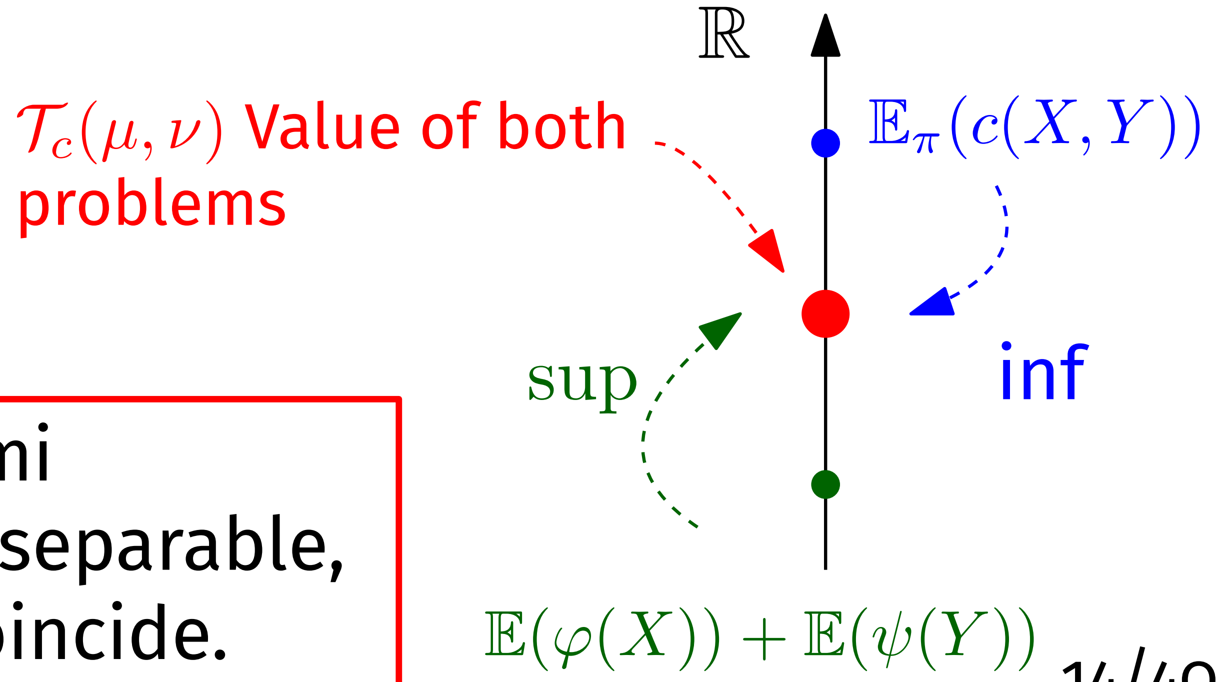
Dual problem

$$\sup\{\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \ : \ \varphi(x)+\psi(y) \leq c(x,y) \text{ for all } x,y \ \}$$

φ,ψ functions on $\mathbb{X}$ and $\mathbb{Y}$

Lemma (weak duality) For any π, φ,ψ satisfying the constraints,

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \leq \mathbb{E}_\pi(c(X,Y)).$$



Theorem (strong duality). If c lower semi continuous and $\mathbb{X}$, $\mathbb{Y}$ metric, complete, separable, then the values of the two problems coincide.

A word on attainment

Assumptions on the spaces:

- $\mathbb{X}, \mathbb{Y}$ metric, complete, separable. (e.g. $\mathbb{R}^d$)

Assumption on c

- c takes finite values and bounded from below.
- $c(x,y) \leq a(x) + b(y)$ with $a \in L^1(\mu)$ and $b \in L^1(\nu)$.
- c is continuous.

Optimal (φ^*, ψ^*): **Kantorovich potentials.**

Theorem. With these assumptions, there exists a solution π^* to the primal problem and a solution $(\varphi^*, \psi^*) \in L^1(\mu) \times L^1(\nu)$ to the dual problem.

In particular: $\mathbb{E}(\varphi^*(X)) + \mathbb{E}(\psi^*(Y)) = \mathbb{E}_{\pi^*}(c(X,Y))$.

A word on attainment

Assumptions on the spaces:

- $\mathbb{X}, \mathbb{Y}$ metric, complete, separable. (e.g. $\mathbb{R}^d$)

Assumption on c

- c takes finite values and bounded from below.
- $c(x, y) \leq a(x) + b(y)$ with $a \in L^1(\mu)$ and $b \in L^1(\nu)$.
- c is continuous.

Optimal (φ^*, ψ^*): **Kantorovich potentials.**

Theorem. With these assumptions, there exists a solution π^* to the primal problem and a solution $(\varphi^*, \psi^*) \in L^1(\mu) \times L^1(\nu)$ to the dual problem.

In particular: $\mathbb{E}(\varphi^*(X)) + \mathbb{E}(\psi^*(Y)) = \mathbb{E}_{\pi^*}(c(X, Y))$.

Moreover, $\varphi^*(x) + \psi^*(y) \leq c(x, y)$ for all x, y,

$\varphi^*(X) + \psi^*(Y) = c(X, Y)$ a.s. if $(X, Y) \sim \pi^*$.

Some remarks on duality

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \leq \mathcal{T}_c(X,Y) \leq \mathbb{E}_\pi(c(X,Y)).$$

Any admissible φ, ψ gives a lower bound

Any coupling between X and Y gives an upper bound

Some remarks on duality

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \leq \mathcal{T}_c(X, Y) \leq \mathbb{E}_\pi(c(X, Y)).$$

Any admissible φ, ψ gives a lower bound

Any coupling between X and Y gives an upper bound

Lemma (Criterion for optimality) If (φ, ψ) satisfy the constraints $\varphi(x) + \psi(y) \leq c(x, y)$ for all x, y and $\pi \in \Pi(\mu, \nu)$ such that

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) = \mathbb{E}_\pi(c(X, Y)),$$

then (φ, ψ) is **optimal for the dual problem** and π is **optimal for the primal problem**.

Some remarks on duality

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) \leq \mathcal{T}_c(X,Y) \leq \mathbb{E}_\pi(c(X,Y)).$$

Any admissible φ, ψ gives a lower bound

Any coupling between X and Y gives an upper bound

Lemma (Criterion for optimality) If (φ, ψ) satisfy the constraints $\varphi(x) + \psi(y) \leq c(x,y)$ for all x, y and $\pi \in \Pi(\mu, \nu)$ such that

$$\mathbb{E}(\varphi(X)) + \mathbb{E}(\psi(Y)) = \mathbb{E}_\pi(c(X,Y)),$$

then (φ, ψ) is **optimal for the dual problem** and π is **optimal for the primal problem**.

Remark. For the case $c(x,y) = \|x-y\|$, then we can restrict to $f = \varphi = -\psi$: we recover the formulation with Lipschitz functions.

1 - Particular case: discrete measures

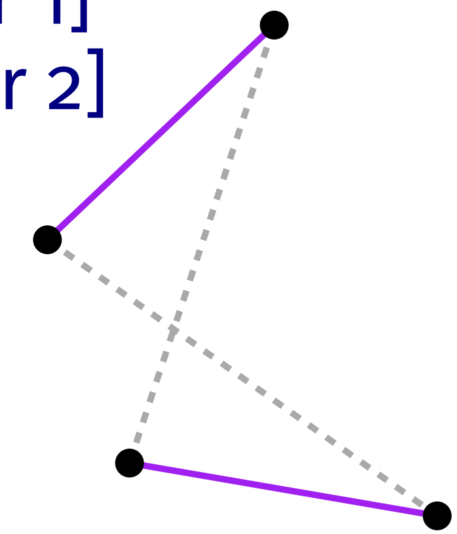
2 - Particular case: one dimensional

3 - Duality

4 - Monotoncity, structure of optimal couplings

Interlude: Gaussi

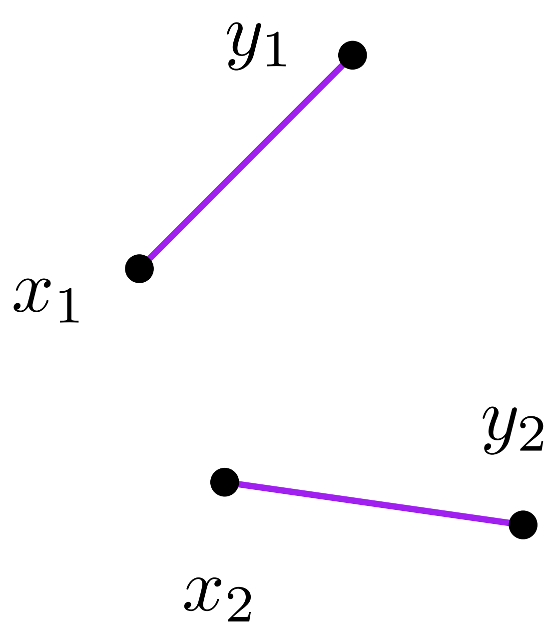
[Santambrogio, Chapter 1]
[Peyré & Cuturi, Chapter 2]



5 - Wasserstein distances

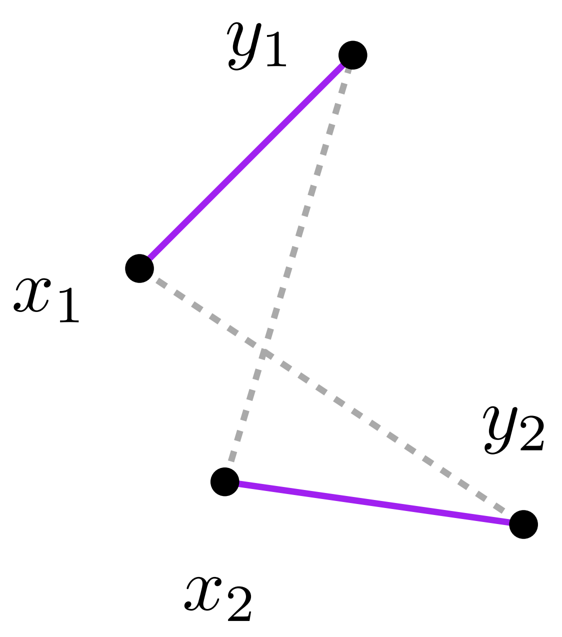
6 - Numerical methods

(c-cyclical) Monotonicity



Take π optimal and reason in the discrete case: assume $\pi((X,Y) = (x_1, y_1)) > 0$ and $\pi((X,Y) = (x_2, y_2)) > 0$

(c-cyclical) Monotonicity



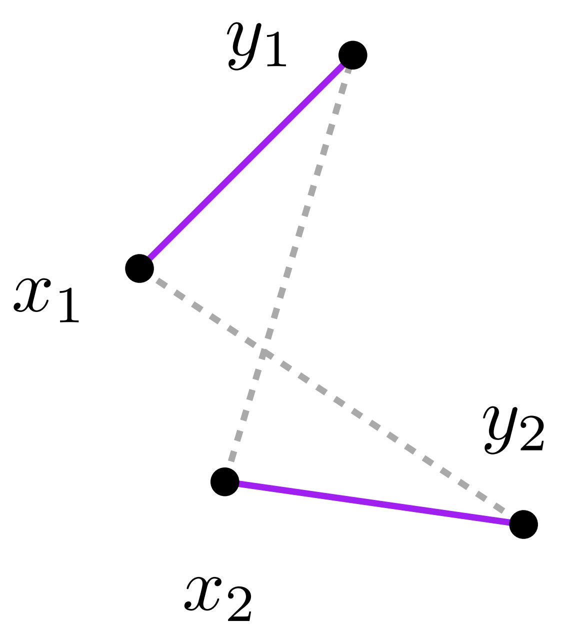
Take π optimal and reason in the discrete case: assume $\pi((X,Y)=(x_1,y_1))>0$ and $\pi((X,Y)=(x_2,y_2))>0$

Then we must have:

$$c(x_1,y_1)+c(x_2,y_2)\leq c(x_1,y_2)+c(x_2,y_1).$$

(If not just pair the other way around)

(c-cyclical) Monotonicity



Take π optimal and reason in the discrete case: assume $\pi((X,Y)=(x_1,y_1))>0$ and $\pi((X,Y)=(x_2,y_2))>0$

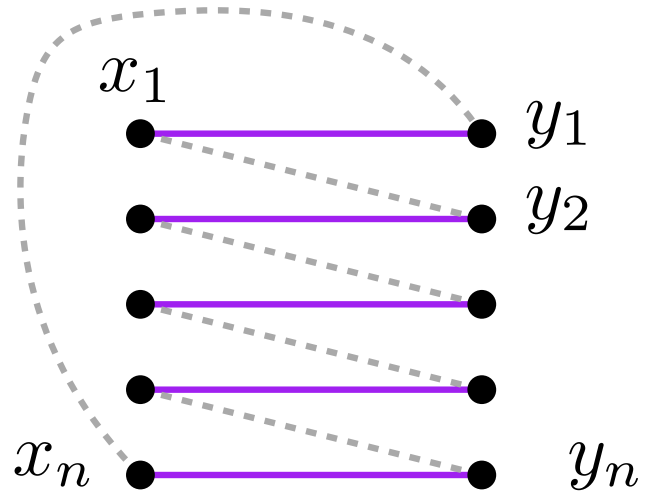
Then we must have:

$$c(x_1,y_1)+c(x_2,y_2)\le c(x_1,y_2)+c(x_2,y_1).$$

(If not just pair the other way around)

More general: if $\pi((X,Y)=(x_k,y_k))>0$ for $k=1,\dots,n$. Then we must have:

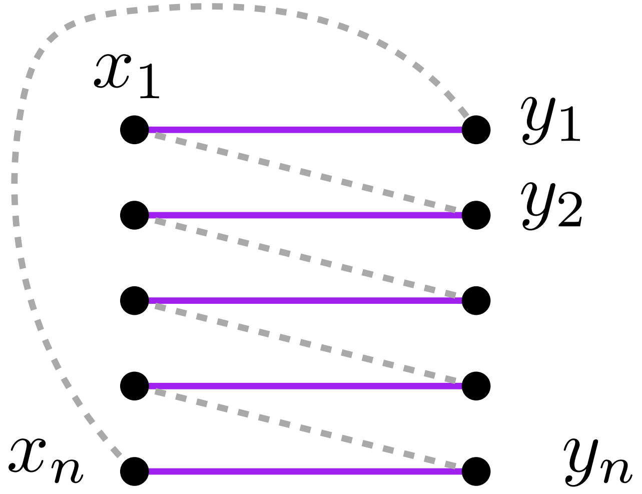
$$\sum_{k=1}^{n} c(x_k,y_k)\le \sum_{k=1}^{n} c(x_k,y_{k+1})$$



A necessary and sufficient condition for optimality

Definition. A subset Γ of $\mathbb{X} \times \mathbb{Y}$ is said c-cyclically monotone if: for every $(x_1, y_1), \dots (x_n, y_n) \in \Gamma$, we have:

$$\sum_{k=1}^{n} c(x_k, y_k) \leq \sum_{k=1}^{n} c(x_k, y_{k+1}).$$



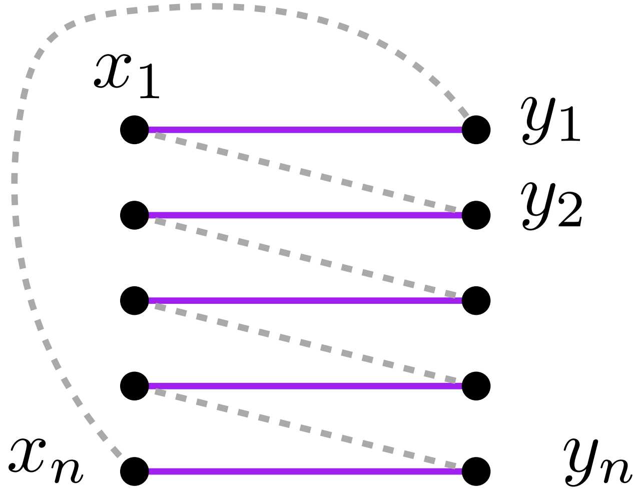
Theorem. Assume c continuous. A coupling π is optimal if and only if its topological support is c-cyclically monotone.

(x, y) in topological support if $\pi(V) > 0$ for every neighborhod V of (x, y).

A necessary and sufficient condition for optimality

Definition. A subset Γ of $\mathbb{X} \times \mathbb{Y}$ is said c-cyclically monotone if: for every $(x_1, y_1), \dots (x_n, y_n) \in \Gamma$, we have:

$$\sum_{k=1}^{n} c(x_k, y_k) \leq \sum_{k=1}^{n} c(x_k, y_{k+1}).$$



Theorem. Assume c continuous. A coupling π is optimal if and only if its topological support is c-cyclically monotone.

Remark. A lot of fine results in optimal coupling (is the coupling deterministic? Stability of optimal couplings, Brenier's theorem, etc.) start from this result.

Brenier's theorem

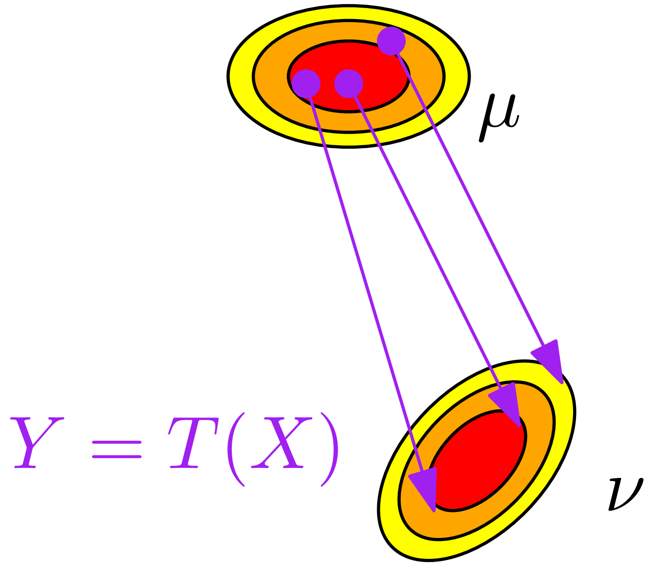
Restrict to $c(x, y) = \|x - y\|^2$ on $\mathbb{R}^d$, and $\mathbb{E}(\|X\|^2) < +\infty$, $\mathbb{E}(\|Y\|^2) < +\infty$.

Brenier's theorem

Restrict to $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d$, and $\mathbb{E}(\|X\|^2) < +\infty$, $\mathbb{E}(\|Y\|^2) < +\infty$.

Assumption: μ, the law of X, has a density with respect to the Lebesgue measure.

Theorem. In this setting, there exits a **unique** optimal coupling, and a coupling $\pi = \mathrm{Law}(X,Y)$ is the optimal one iff $Y = T(X)$ where T is the **gradient of a convex function**.



(A convex function is always differentiable almost everywhere)

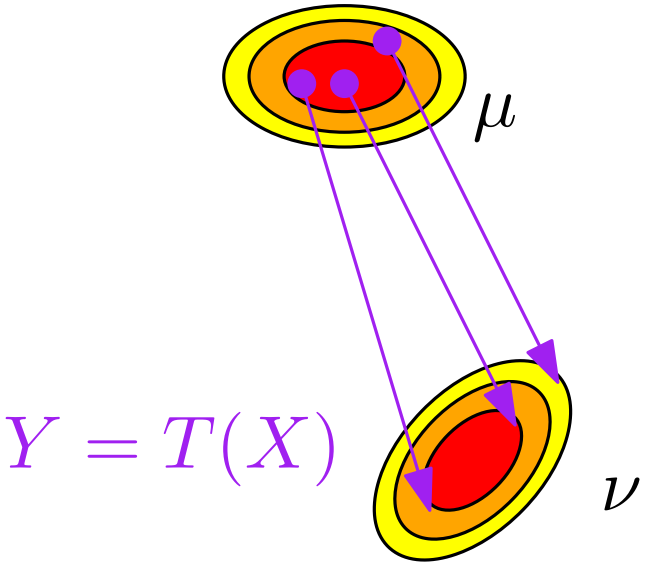
Actually, $T(x) = x - \nabla\varphi^*(x)$, where φ^* solution to the dual problem.

Brenier's theorem

Restrict to $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d$, and $\mathbb{E}(\|X\|^2) < +\infty$, $\mathbb{E}(\|Y\|^2) < +\infty$.

Assumption: μ, the law of X, has a density with respect to the Lebesgue measure.

> **Theorem**. In this setting, there exits a **unique** optimal coupling, and a coupling $\pi = \mathrm{Law}(X,Y)$ is the optimal one iff $Y = T(X)$ where T is the **gradient of a convex function**.



(A convex function is always differentiable almost everywhere)

Actually, $T(x) = x - \nabla\varphi^*(x)$, where φ^* solution to the dual problem.

Remark. If $d=1$, T is the derivative of a convex function if and only if T is non-decreasing. Consistent with previous results.

Further remarks on Brenier's theorem

Write f, g for the p.d.f. of X and Y.
Write $T = \nabla u$ where u is convex.

Then $\mathrm{Law}(T(X)) = \mathrm{Law}(Y)$ yields the **Monge-Ampère** equation for u:

$$\det(D^2 u(x)) = \frac{f(x)}{g(\nabla u(x))} \qquad \text{for all } x$$

Determinant of the Hessian matrix of u

Further remarks on Brenier's theorem

Write f, g for the p.d.f. of X and Y.
Write $T = \nabla u$ where u is convex.

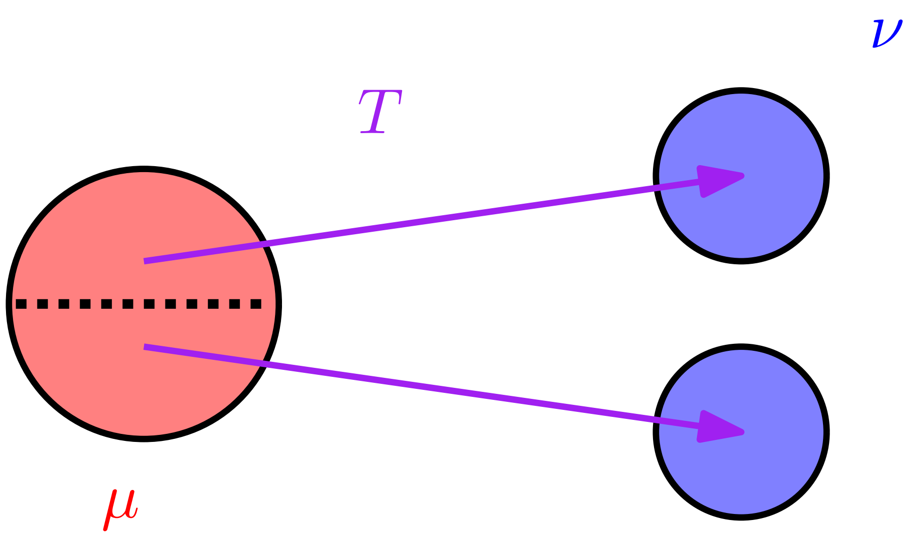
Then $\mathrm{Law}(T(X)) = \mathrm{Law}(Y)$ yields the **Monge-Ampère** equation for u:

$$\det(D^2 u(x)) = \frac{f(x)}{g(\nabla u(x))} \qquad \text{for all } x$$

Determinant of the Hessian matrix of u

Beware $T = \nabla u$ can be discontinuous:

T smooth if f, g smooth and μ, ν have **convex** support.



Further remarks on Brenier's theorem

Write f, g for the p.d.f. of X and Y.
Write $T = \nabla u$ where u is convex.

Then $\mathrm{Law}(T(X)) = \mathrm{Law}(Y)$ yields the **Monge-Ampère** equation for u:

$$\det(D^2 u(x)) = \frac{f(x)}{g(\nabla u(x))} \qquad \text{for all } x$$

Remark. The conclusion $Y = T(X)$ deterministic function of X can be obtained with much weaker assumption, e.g.

- the law of X, has a density with respect to the Lebesgue measure.
- For every x, the map $y \mapsto \nabla_x c(x, y)$ is injective (**twist condition**).
  Typically satisfied if:
  $$\det\left(\frac{\partial^2 c}{\partial x_i \partial y_j}\right)_{1 \le i,j \le d} \neq 0.$$

1 - Particular case: discrete measures

2 - Particular case: one dimensional

3 - Duality

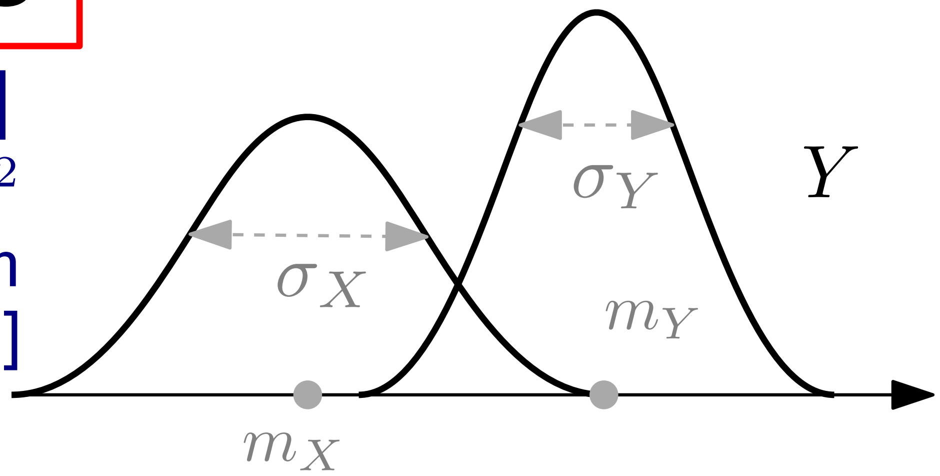
4 - Monotoncity, structure of optimal couplings

Interlude: Gaussian measures

[Peyré & Cuturi, Chapter 2]
[Gelbrich (1990) On a Formula for the L^2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces]

5 -

6 -

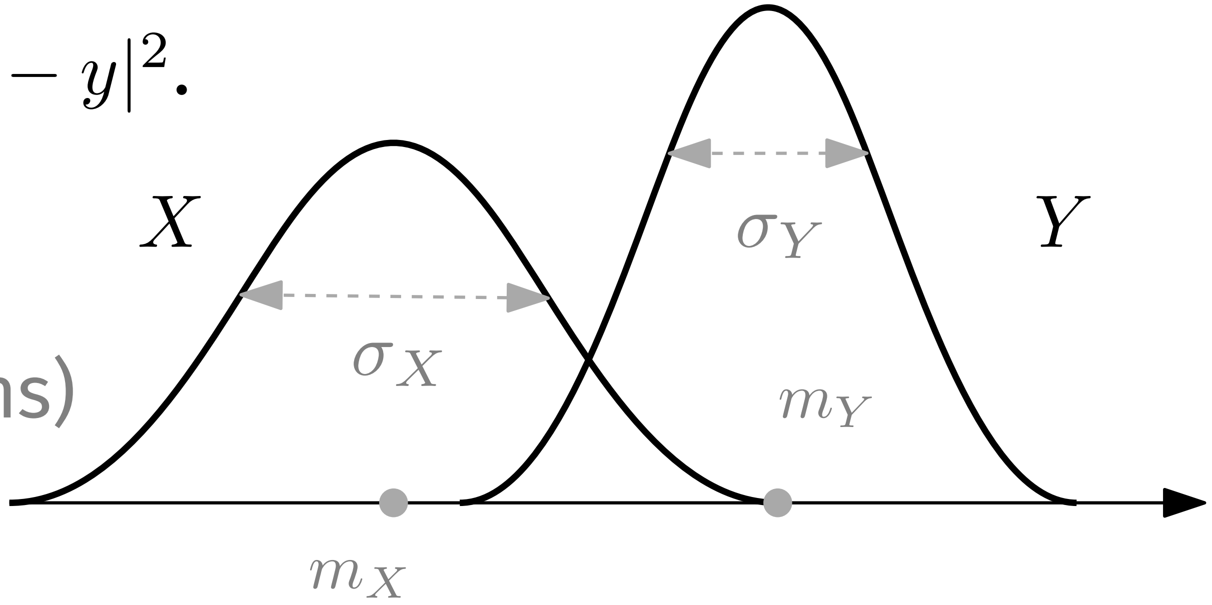


Gaussian measures in one dimension

Restriction to **quadratic** cost $c(x,y) = |x-y|^2$.

- $X \sim \mathcal{N}(m_X, \sigma_X^2)$
- $Y \sim \mathcal{N}(m_Y, \sigma_Y^2)$

  (One dimensional Gaussian distributions)

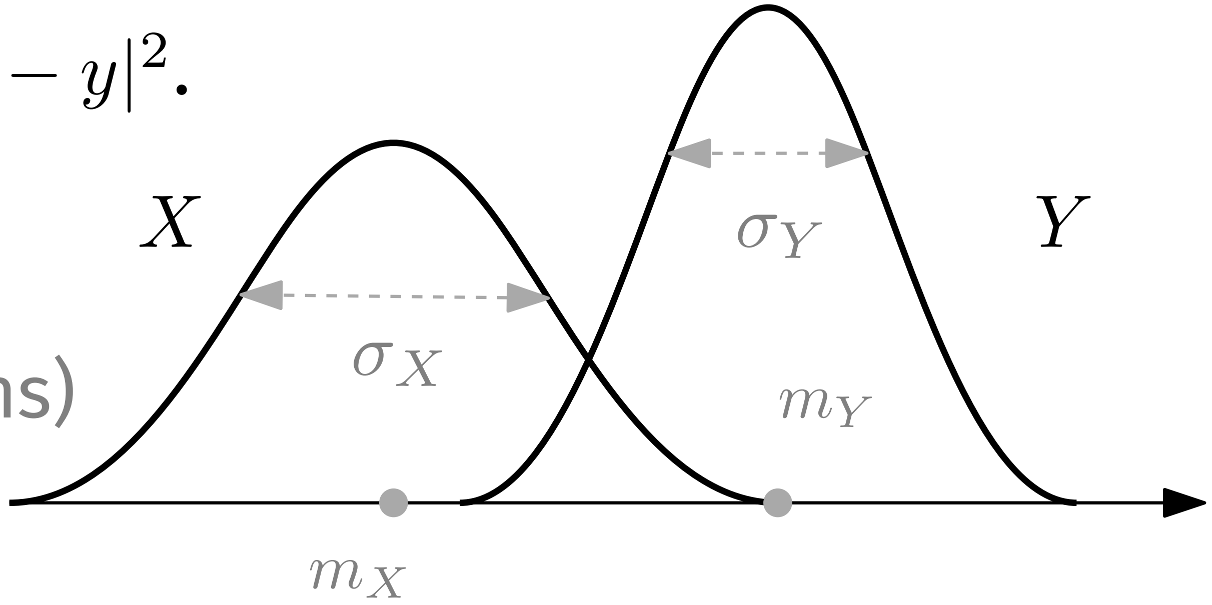


Gaussian measures in one dimension

Restriction to **quadratic** cost $c(x,y) = |x-y|^2$.

- $X \sim \mathcal{N}(m_X, \sigma_X^2)$
- $Y \sim \mathcal{N}(m_Y, \sigma_Y^2)$

(One dimensional Gaussian distributions)



Lemma. The optimal transport coupling is given by $Y = T(X)$ with

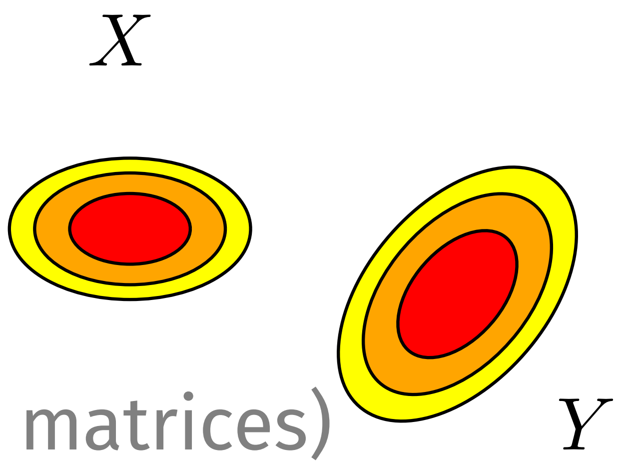
$$T(x) = m_Y - m_X + \frac{\sigma_Y}{\sigma_X}(x - m_X).$$

Moreover the value of the problem is

$$\mathcal{T}_c(X,Y) = |m_X - m_Y|^2 + |\sigma_X - \sigma_Y|^2.$$

Towards higher dimension

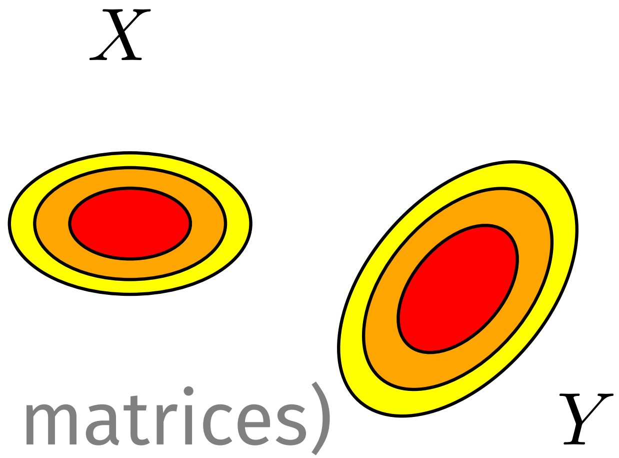
Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.



- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)

Towards higher dimension



Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)

Recall. Take $Z \sim \mathcal{N}(0, I)$. Then $m + \Sigma^{1/2} Z$ follows $\mathcal{N}(m, \Sigma)$.

Ansatz. the optimal coupling is given by:

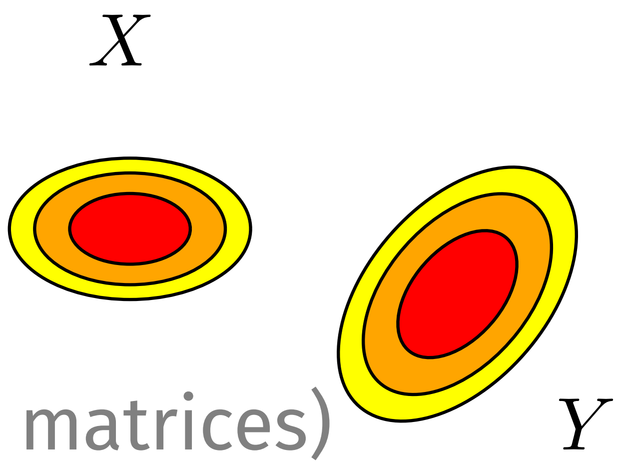
$$(m_X + \Sigma_X^{1/2} Z, m_Y + \Sigma_Y^{1/2} Z), \qquad Z \sim \mathcal{N}(0, I).$$

Towards higher dimension

Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.

- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)



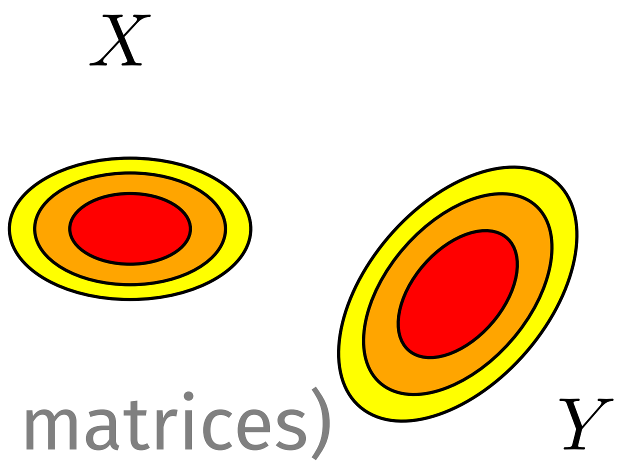
Recall. Take $Z \sim \mathcal{N}(0, I)$. Then $m + \Sigma^{1/2} Z$ follows $\mathcal{N}(m, \Sigma)$.

Ansatz. the optimal coupling is given by:

$$(m_X + \Sigma_X^{1/2} Z, m_Y + \Sigma_Y^{1/2} Z), \qquad Z \sim \mathcal{N}(0, I).$$

Is it really the optimal coupling?

Towards higher dimension



Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)

Recall. Take $Z \sim \mathcal{N}(0, I)$. Then $m + \Sigma^{1/2} Z$ follows $\mathcal{N}(m, \Sigma)$.

~~**Ansatz**. the optimal coupling is given by:~~

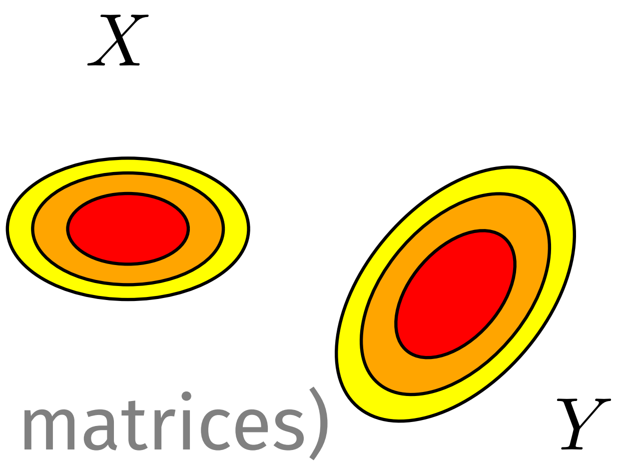
$$(m_X + \Sigma_X^{1/2} Z, m_Y + \Sigma_Y^{1/2} Z), \qquad Z \sim \mathcal{N}(0, I).$$

Is it really the optimal coupling?

No. With $m_X = m_Y = 0$, it gives $Y = \Sigma_Y^{1/2} \Sigma_X^{-1/2} X$, but $x \mapsto \Sigma_Y^{1/2} \Sigma_X^{-1/2} x$ is not the gradient of a convex function.

The correct formula for the transport between Gaussians

Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.



- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

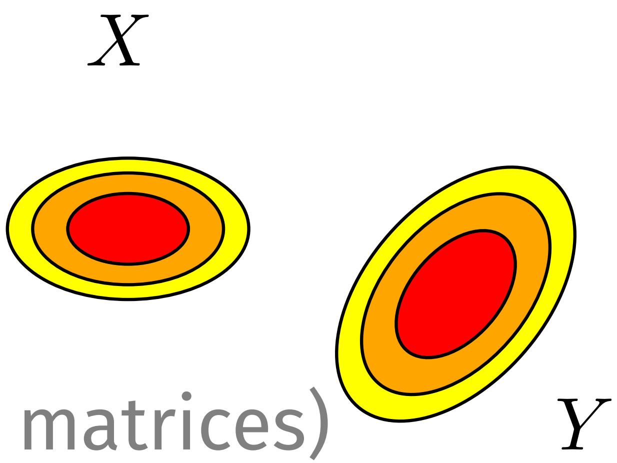
(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)

The correct formula for the transport between Gaussians

Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)



New ansatz. $Y = T(X)$ with T linear:

$$T(x) = Ax + b$$

with A symmetric semi positive definite matrix.

for T to be gradient of convex function $u(x) = \frac{1}{2}x^\top Ax + b^\top x$.

The correct formula for the transport between Gaussians

Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

- $X \sim \mathcal{N}(m_X, \Sigma_X)$
- $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$

(Multivariate Gaussian distributions: Σ_X, Σ_Y covariance matrices)

New ansatz. $Y = T(X)$ with T linear:

$$T(x) = Ax + b$$

with A symmetric semi positive definite matrix.

for T to be gradient of convex function $u(x) = \frac{1}{2}x^\top Ax + b^\top x$.

To match the covariance we must have: $\Sigma_Y = A\Sigma_X A$.

We find as **unique** solution: $A = \Sigma_X^{-1/2}(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}\Sigma_X^{-1/2}$.

Conclusion: explicit formula

Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

Theorem. Assume $X \sim \mathcal{N}(m_X, \Sigma_X)$ and $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$. Then if Σ_X invertible the optimal coupling is $Y = T(X)$ with:

$$T(x) = A(x - m_X) + m_Y - m_X, \qquad A = \Sigma_X^{-1/2}(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}\Sigma_X^{-1/2}.$$

Conclusion: explicit formula

Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

Theorem. Assume $X \sim \mathcal{N}(m_X, \Sigma_X)$ and $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$. Then if Σ_X invertible the optimal coupling is $Y = T(X)$ with:

$$T(x) = A(x - m_X) + m_Y - m_X, \qquad A = \Sigma_X^{-1/2}(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}\Sigma_X^{-1/2}.$$

The value of the problem is

$$\mathcal{T}_c(X,Y) = \|m_X - m_Y\|^2 + \mathrm{Tr}\left(\Sigma_X + \Sigma_Y - 2(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}\right).$$

Conclusion: explicit formula

Restriction to **quadratic** cost $c(x,y) = \|x-y\|^2$.

Theorem. Assume $X \sim \mathcal{N}(m_X, \Sigma_X)$ and $Y \sim \mathcal{N}(m_Y, \Sigma_Y)$. Then if Σ_X invertible the optimal coupling is $Y = T(X)$ with:

$$T(x) = A(x - m_X) + m_Y - m_X, \qquad A = \Sigma_X^{-1/2}(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}\Sigma_X^{-1/2}.$$

The value of the problem is

$$\mathcal{T}_c(X,Y) = \|m_X - m_Y\|^2 + \operatorname{Tr}\left(\Sigma_X + \Sigma_Y - 2(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2})^{1/2}\right).$$

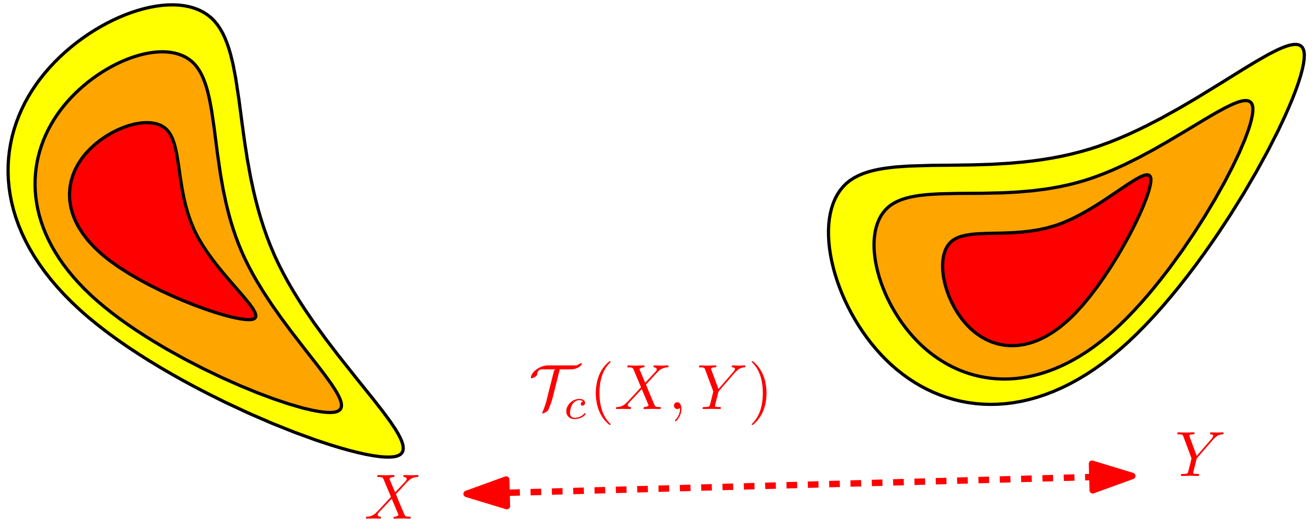
Remark. If Σ_X and Σ_Y commute we recover simpler formulas:

$$A = \Sigma_Y^{1/2}\Sigma_X^{-1/2}, \qquad \mathcal{T}_c(X,Y) = \|m_X - m_Y\|^2 + \operatorname{Tr}\left(\left(\Sigma_X^{1/2} - \Sigma_Y^{1/2}\right)^2\right)$$

Gelbrich's lower bound

Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.

Theorem. Let X, Y be random variable with finite second moments.

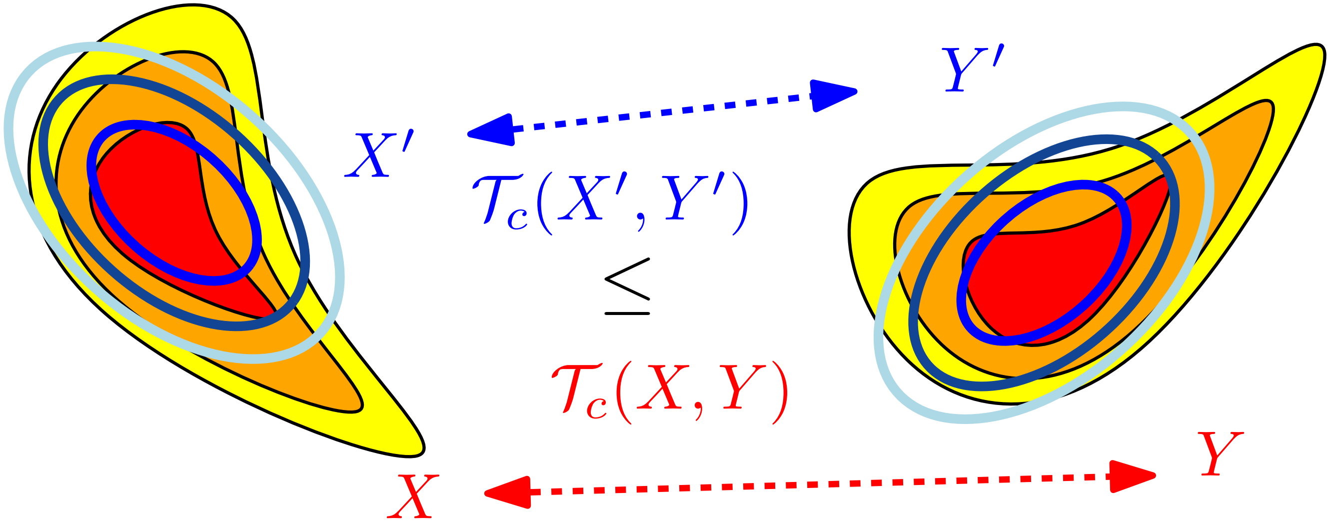


Gelbrich's lower bound

Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.

Theorem. Let X, Y be random variable with finite second moments. Define X', Y' Gaussian random variables whose mean and covariance coincide respectively with the ones of X and Y. Then:

$$\mathcal{T}_c(X, Y) \geq \mathcal{T}_c(X', Y').$$



Previous slide gives a formula for this.

Gelbrich's lower bound

Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.

Theorem. Let X, Y be random variable with finite second moments. Define X', Y' Gaussian random variables whose mean and covariance coincide respectively with the ones of X and Y. Then:

$$\mathcal{T}_c(X, Y) \geq \mathcal{T}_c(X', Y').$$

Previous slide gives a formula for this.

Remark: nothing specific to Gaussians, result generalizes to any location scale family.

Gelbrich's lower bound

Restriction to **quadratic** cost $c(x, y) = \|x - y\|^2$.

Theorem. Let X, Y be random variable with finite second moments. Define X', Y' Gaussian random variables whose mean and covariance coincide respectively with the ones of X and Y. Then:

$$\mathcal{T}_c(X, Y) \geq \mathcal{T}_c(X', Y').$$

Previous slide gives a formula for this.

Remark: nothing specific to Gaussians, result generalizes to any location scale family.

Shortest proof I know: duality! Use (φ', ψ') solution to the dual problem for (X', Y') to provide a lower bound for $\mathcal{T}_c(X, Y)$.

1 - Particular case: discrete measures

2 - Particular case: one dimensional

3 - Duality

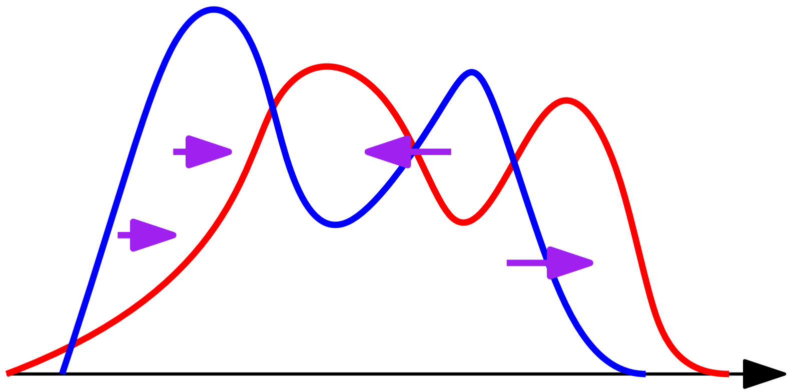
4 - Monotoncity, structure of optimal couplings

Interlude: Gaussian measures

5 - Wasserstein distances

6 -

[Santambrogio, Chapter 5]
[Ambrosio, Gigli & Savaré, Chapter 7]



Wasserstein distances

Space:

- We take $(\mathbb{X}, \mathsf{d})$ a metric, complete, separable space with distance d.
- For $p \geq 1$, $\mathcal{P}_p(\mathbb{X})$ probability distributions μ with $\mathbb{E}_{X\sim\mu}(\mathsf{d}(x_0, X)^p) < +\infty$.

Cost function:

- $c(x, y) = \mathsf{d}(x, y)^p$.

x_0 any point in $\mathbb{X}$

Definition. The **Wasserstein** distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{X})$ is

$$W_p(\mu, \nu) = (\mathcal{T}_c(\mu, \nu))^{1/p} = \min_{\pi} \left\{ \|\mathsf{d}(X, Y)\|_{L^p(\pi)} \ : \ \pi \in \Pi(\mu, \nu) \right\}.$$

Wasserstein distances

Space:

- We take $(\mathbb{X}, \mathsf{d})$ a metric, complete, separable space with distance d.
- For $p \geq 1$, $\mathcal{P}_p(\mathbb{X})$ probability distributions μ with $\mathbb{E}_{X \sim \mu}(\mathsf{d}(x_0, X)^p) < +\infty$.

x_0 any point in $\mathbb{X}$

Cost function:

- $c(x, y) = \mathsf{d}(x, y)^p$.

Definition. The **Wasserstein** distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{X})$ is

$$W_p(\mu, \nu) = (\mathcal{T}_c(\mu, \nu))^{1/p} = \min_{\pi} \left\{ \|\mathsf{d}(X, Y)\|_{L^p(\pi)} \ : \ \pi \in \Pi(\mu, \nu) \right\}.$$

Theorem. W_p defines a distance on $\mathcal{P}_p(\mathbb{X})$ which makes it a complete separable metric space.

Special cases of the Wasserstein distance

• In dimension one: F_X is the cumulative distribution of X, and $Q_X = F_X^{-1}$ is its quantile function.

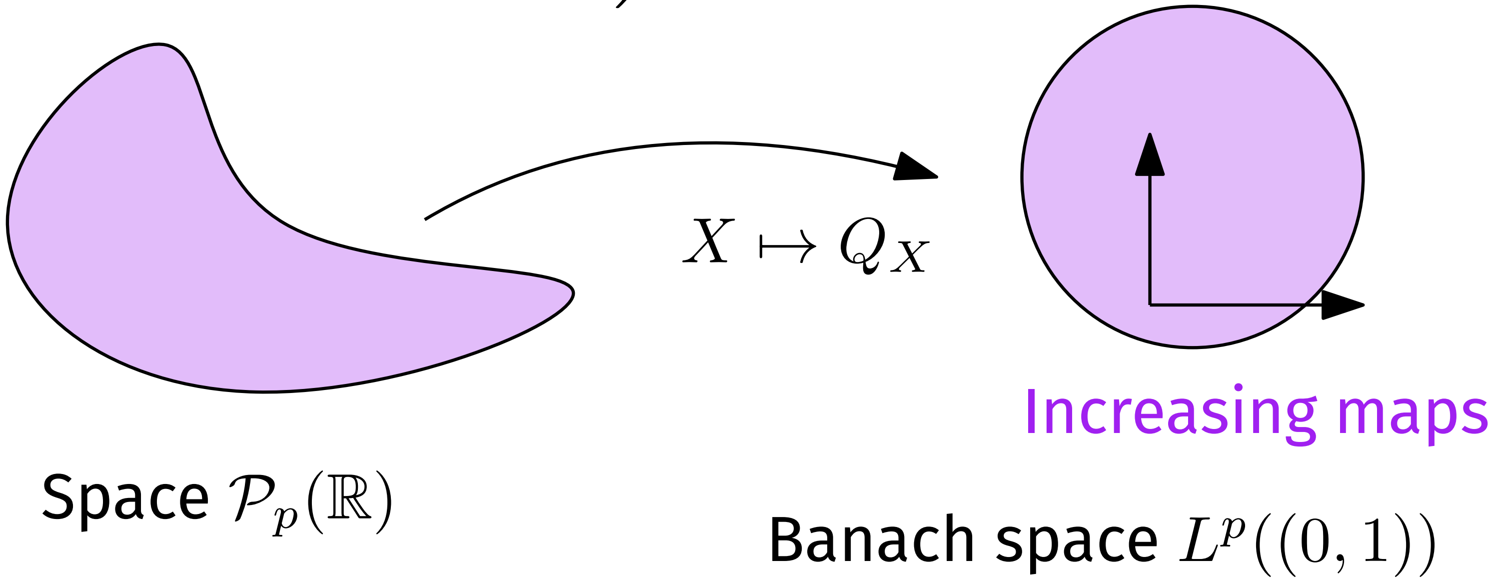
We saw: $$W_p(X,Y) = \left(\int_0^1 |Q_X(u) - Q_Y(u)|^p \, \mathrm{d}u\right)^{1/p} = \|Q_X - Q_Y\|_{L^p((0,1))}.$$

Special cases of the Wasserstein distance

- In dimension one: F_X is the cumulative distribution of X, and $Q_X = F_X^{-1}$ is its quantile function.

We saw: $$W_p(X,Y) = \left(\int_0^1 |Q_X(u) - Q_Y(u)|^p \,\mathrm{d}u\right)^{1/p} = \|Q_X - Q_Y\|_{L^p((0,1))}.$$

So $\mathcal{P}_p(\mathbb{R})$ is isometric to a convex subset of the Banach space $L^p((0,1))$.

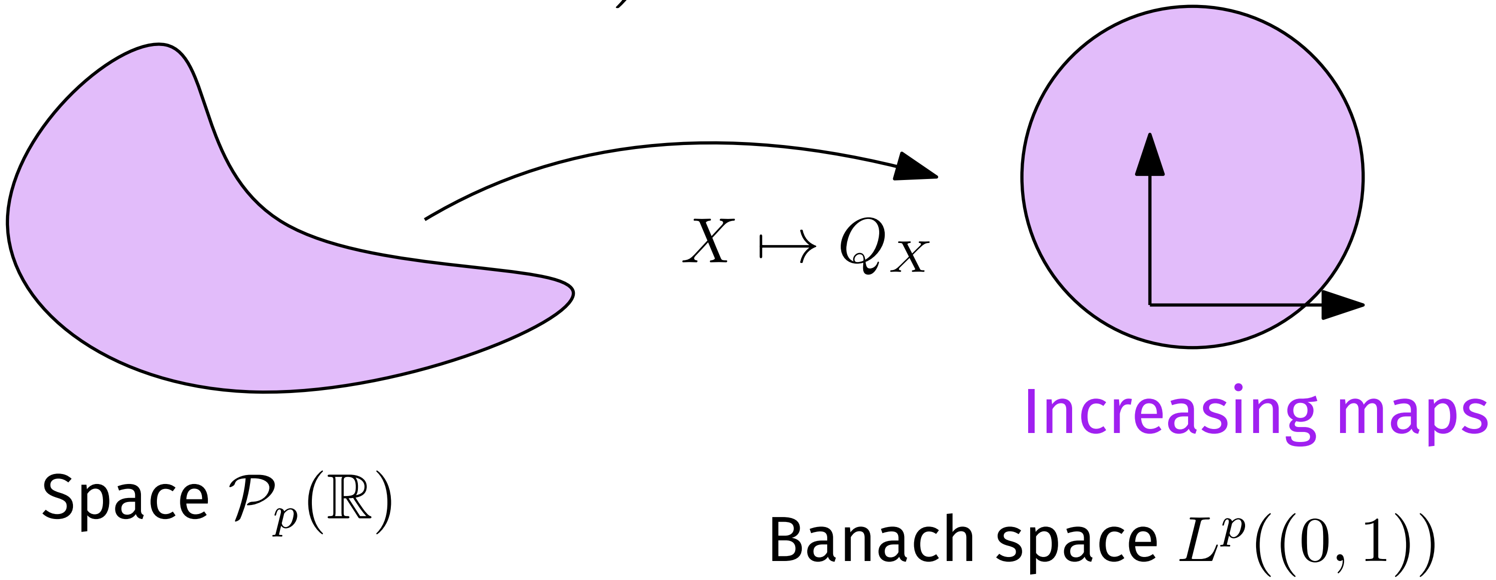


Special cases of the Wasserstein distance

- In dimension one: F_X is the cumulative distribution of X, and $Q_X = F_X^{-1}$ is its quantile function.

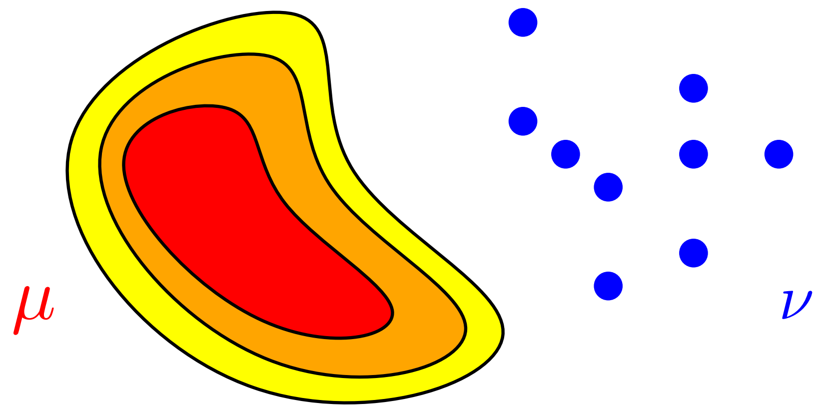
We saw: $$W_p(X,Y) = \left(\int_0^1 |Q_X(u) - Q_Y(u)|^p \, \mathrm{d}u \right)^{1/p} = \|Q_X - Q_Y\|_{L^p((0,1))}.$$

So $\mathcal{P}_p(\mathbb{R})$ is isometric to a convex subset of the Banach space $L^p((0,1))$.



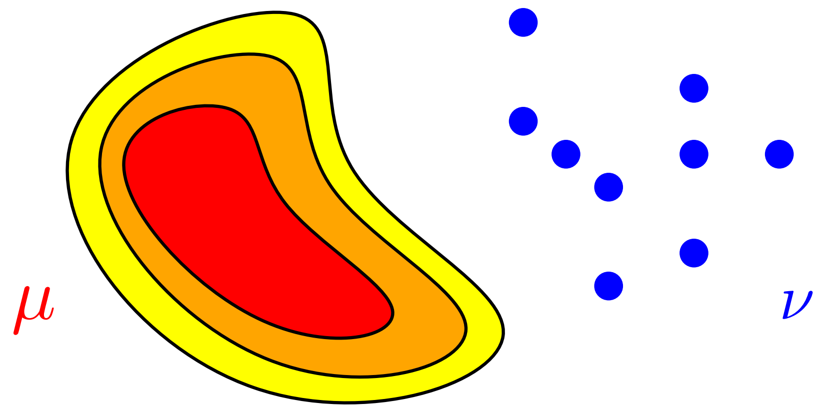
- For Gaussians measures, we have an explicit formula.

Important remark and properties



• Distance between **laws** of random variables, **no restriction on support**.

Important remark and properties



• Distance between **laws** of random variables, **no restriction on support**.

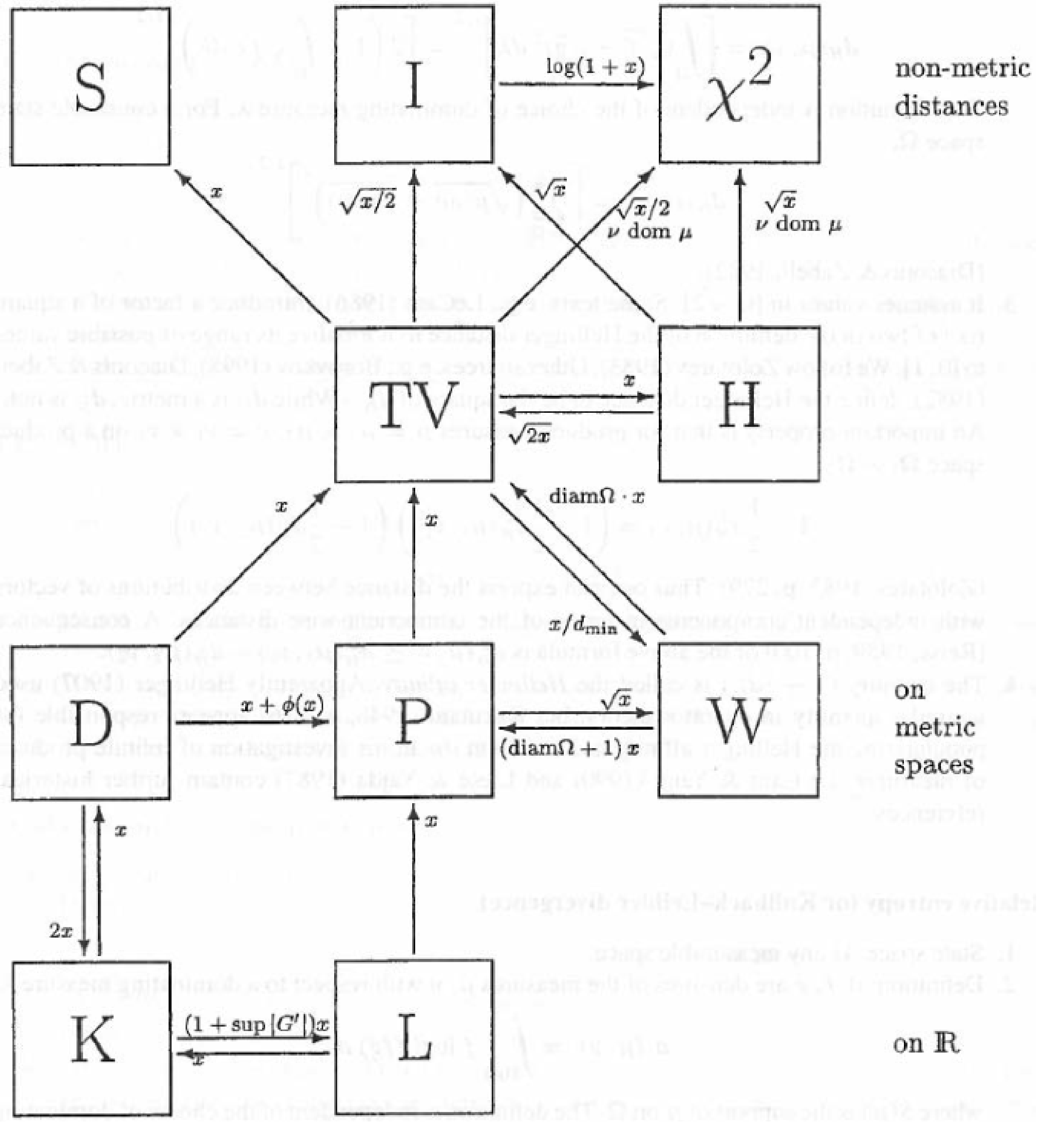
• Metrization of weak convergence with convergence p moment:

Theorem. A sequence (μ_n) is such that $W_p(\mu_n, \mu)$ converges to zero iff

$$\mathbb{E}_{X\sim\mu_n}(f(X)) \to \mathbb{E}_{X\sim\mu}(f(X))$$

for any $f : \mathbb{X} \to \mathbb{R}$ continuous and such that $|f(x)| \leq C(1 + \mathrm{d}(x_0, x)^p)$ for some C and x_0.

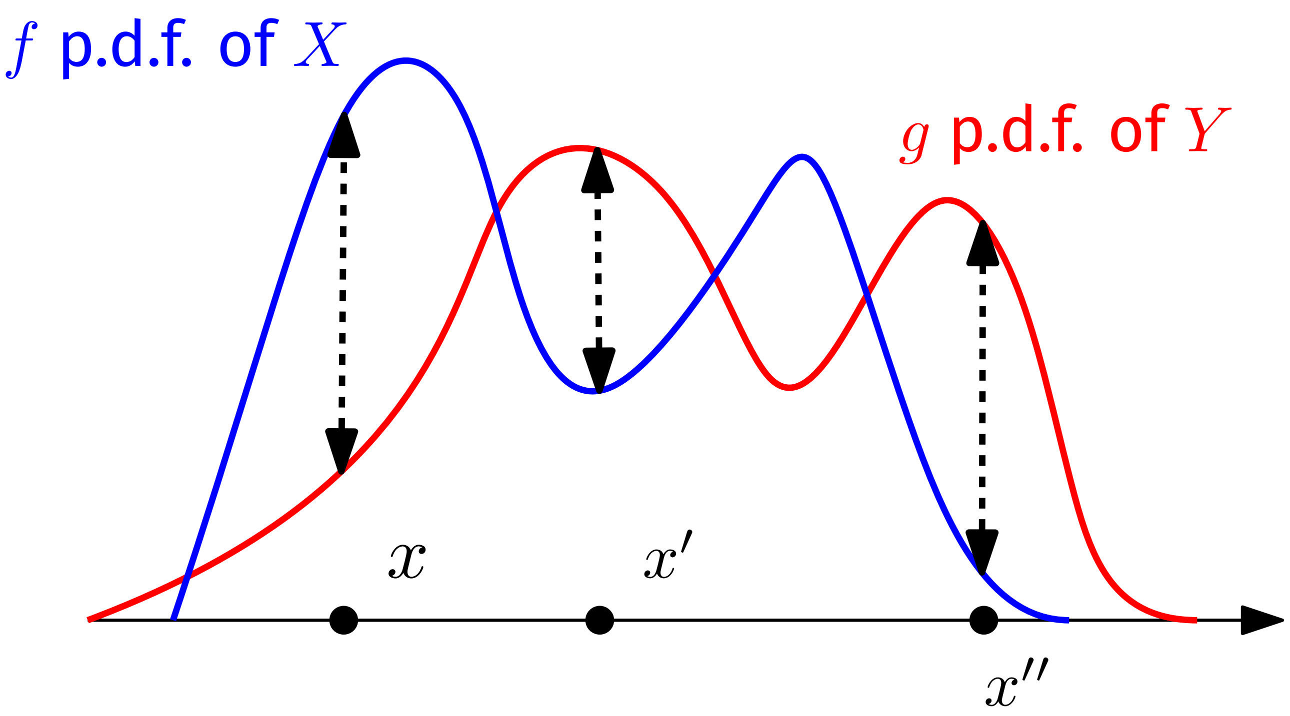
Vertical vs horizontal distance



Many distances between probabilities!

[Gibbs & Su (2002). On Choosing and Bounding Probability Metrics.]

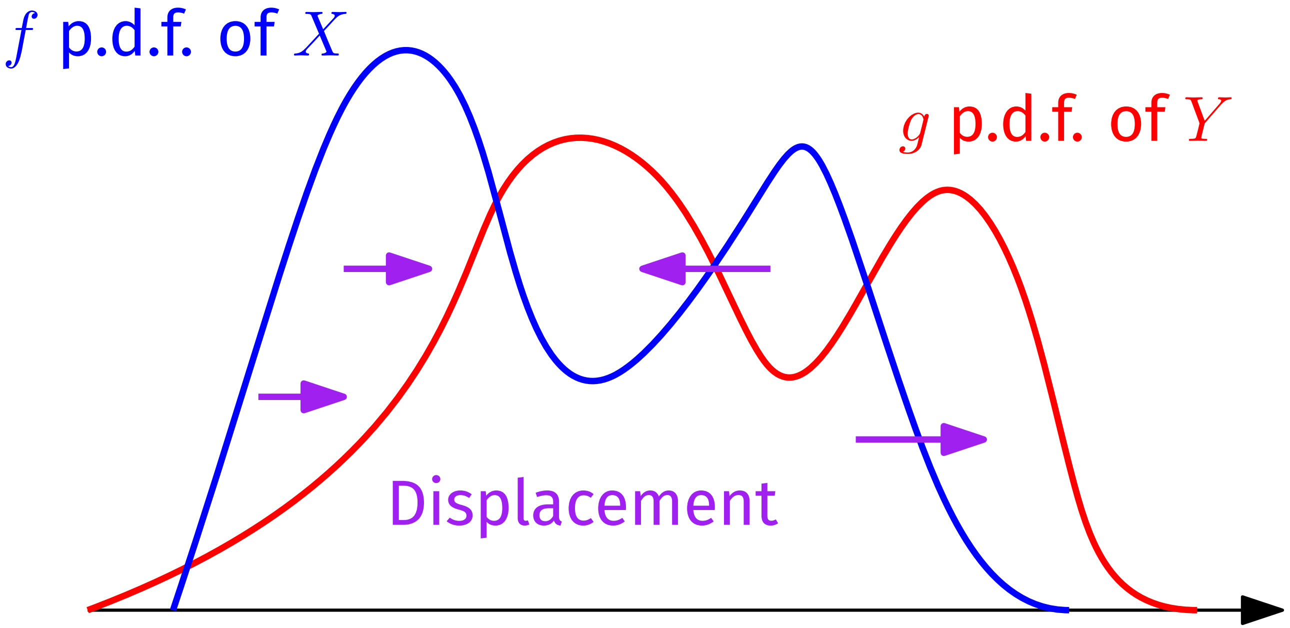
Vertical vs horizontal distance



Vertical distance: compare $f(x)$ and $g(x)$ for the same x.

(Total Variation, Hellinger, Kullback Leiber.)

Vertical vs horizontal distance

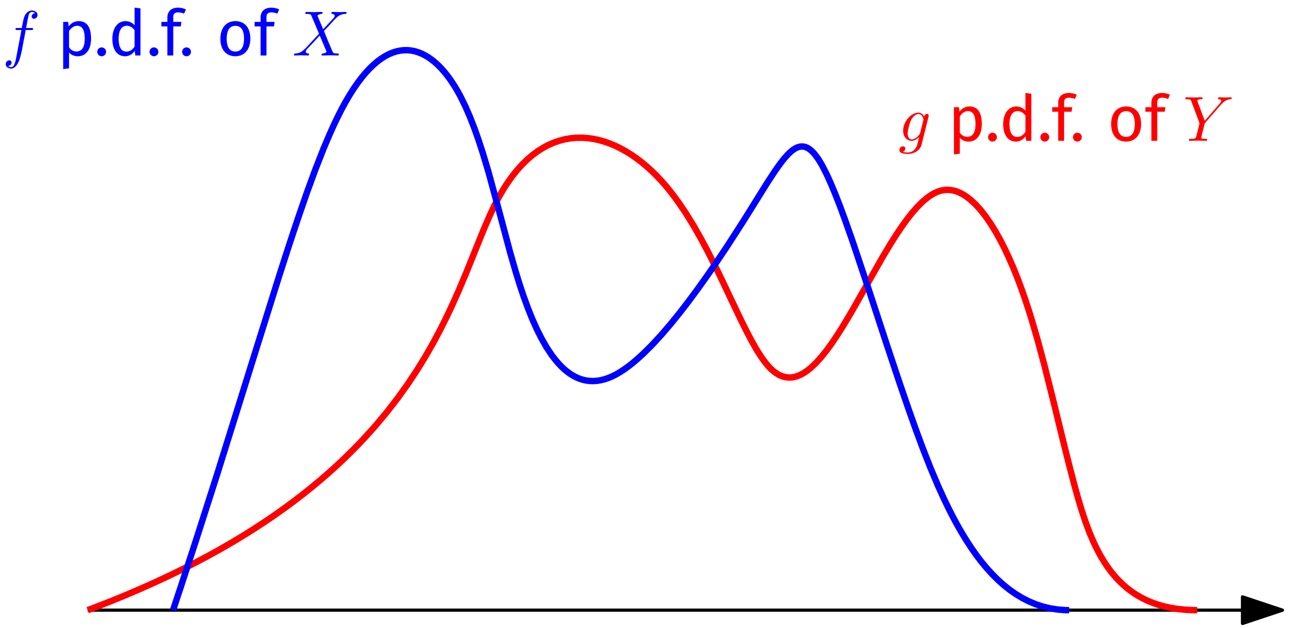


Vertical distance: compare $f(x)$ and $g(x)$ for the same x.

(Total Variation, Hellinger, Kullback Leiber.)

Transport distance. Compare $f(x)$ and $g(y)$ in different locations x and $y = T(x)$.

Vertical vs horizontal distance



Vertical distance: compare $f(x)$ and $g(x)$ for the same x.

(Total Variation, Hellinger, Kullback Leiber.)

Transport distance. Compare $f(x)$ and $g(y)$ in different locations x and $y = T(x)$.

Example. If $\mu = \delta_a$ and $\nu = \delta_b$ (that is, $X = a$ and $Y = b$ a.s.):

$$W_p(\delta_a, \delta_b) = \mathrm{d}(a, b).$$

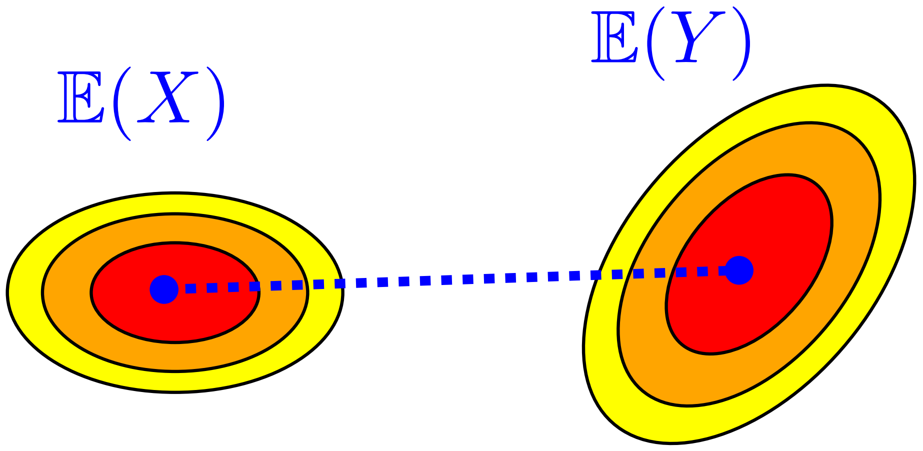
Some useful equalities and bounds in $\mathbb{R}^d$

In $\mathbb{R}^d$, for any $p \geq 1$:

$$W_p(X, Y) \geq \|\mathbb{E}(X) - \mathbb{E}(Y)\|$$

And if $p = 2$, with $\bar{X} = X - \mathbb{E}(X)$ and $\bar{Y} = Y - \mathbb{E}(Y)$ "Pythagora's identity":

$$W_2^2(X, Y) = \|\mathbb{E}(X) - \mathbb{E}(Y)\|^2 + W_2^2(\bar{X}, \bar{Y}).$$



[Gelbrich (1990) On a Formula for the L^2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces]

Some useful equalities and bounds in $\mathbb{R}^d$

In $\mathbb{R}^d$, for any $p \geq 1$:

$$W_p(X, Y) \geq \|\mathbb{E}(X) - \mathbb{E}(Y)\|$$

And if $p = 2$, with $\bar{X} = X - \mathbb{E}(X)$ and $\bar{Y} = Y - \mathbb{E}(Y)$ "Pythagora's identity":

$$W_2^2(X, Y) = \|\mathbb{E}(X) - \mathbb{E}(Y)\|^2 + W_2^2(\bar{X}, \bar{Y}).$$

Convolving decreases the distance: if Z independent from X, Y

$$W_p(X + Z, Y + Z) \leq W_p(X, Y).$$

Some useful equalities and bounds in $\mathbb{R}^d$

In $\mathbb{R}^d$, for any $p \geq 1$:

$$W_p(X,Y) \geq \|\mathbb{E}(X) - \mathbb{E}(Y)\|$$

And if $p = 2$, with $\bar{X} = X - \mathbb{E}(X)$ and $\bar{Y} = Y - \mathbb{E}(Y)$ "Pythagora's identity":

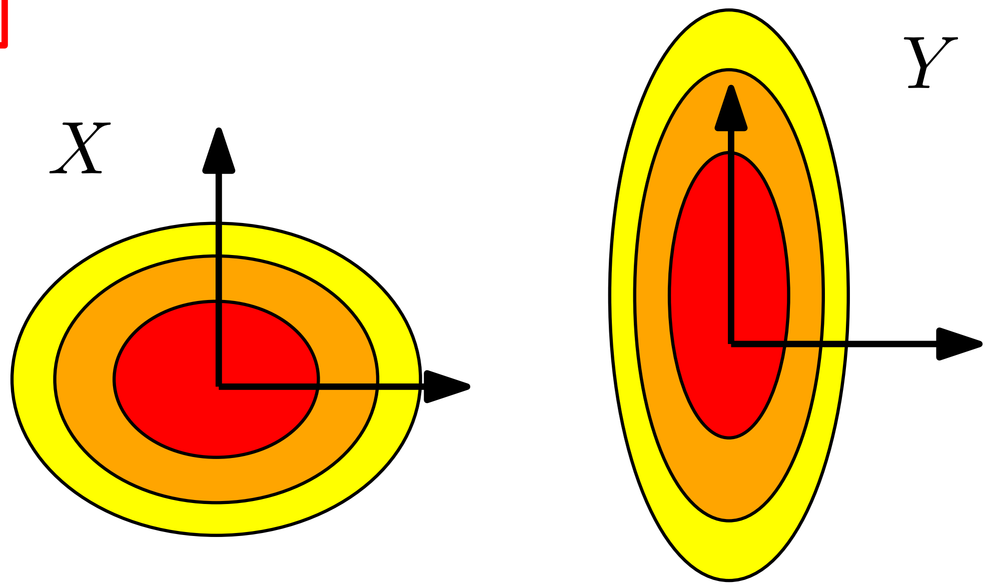
$$W_2^2(X,Y) = \|\mathbb{E}(X) - \mathbb{E}(Y)\|^2 + W_2^2(\bar{X},\bar{Y}).$$

Convolving decreases the distance: if Z independent from X, Y

$$W_p(X+Z, Y+Z) \leq W_p(X,Y).$$

Tensorization: if $X = (X_1, \dots, X_d)$ with $X_1, \dots, X_d$ independent and same for Y:

$$W_2^2(X,Y) = \sum_{i=1}^{d} W_2^2(X_i, Y_i).$$



Some useful equalities and bounds in $\mathbb{R}^d$

In $\mathbb{R}^d$, for any $p \geq 1$:

$$W_p(X,Y) \geq \|\mathbb{E}(X) - \mathbb{E}(Y)\|$$

And if $p = 2$, with $\bar{X} = X - \mathbb{E}(X)$ and $\bar{Y} = Y - \mathbb{E}(Y)$ "Pythagora's identity":

$$W_2^2(X,Y) = \|\mathbb{E}(X) - \mathbb{E}(Y)\|^2 + W_2^2(\bar{X},\bar{Y}).$$

Convolving decreases the

di … nt

W …).

$\mathrm{KL}(\nu|\mu) = \mathbb{E}_\nu[\log \mathrm{d}\nu/\mathrm{d}\mu]$ relative entropy (vertical distance!)

Tensorization: if $X = (X_1, \ldots, X_d)$ with $X_1, \ldots, X_d$ independent and same for Y:

$$W_2^2(X,Y) = \sum_{i=1}^{d} W_2^2(X_i,Y_i).$$

Entropy transport inequality. If X has law μ with density $\exp(-V)$ and V is λ-convex, for any ν

$$W_2^2(\mu,\nu) \leq \frac{2\mathrm{KL}(\nu|\mu)}{\lambda}.$$

[Villani (2008). Optimal transport, old and new. Chapter 22]

1 - Particular case: discrete measures

2 - Particular case: one dimensional

3 - Duality

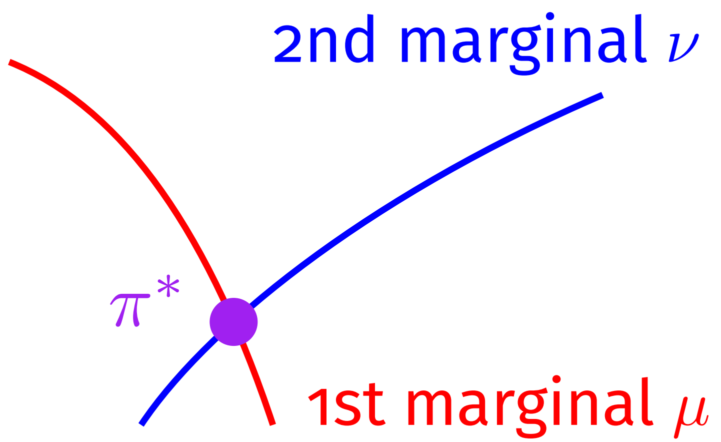
4 - Monotoncity, structure of optimal couplings

Interlude: Gaussian measures

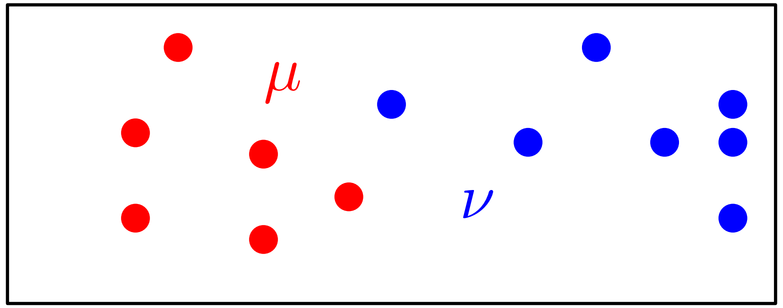
5 - Wasserstein distances

6 - Numerical methods

[Peyré & Cuturi]



A zoology of numerical methods (often convex optimization)



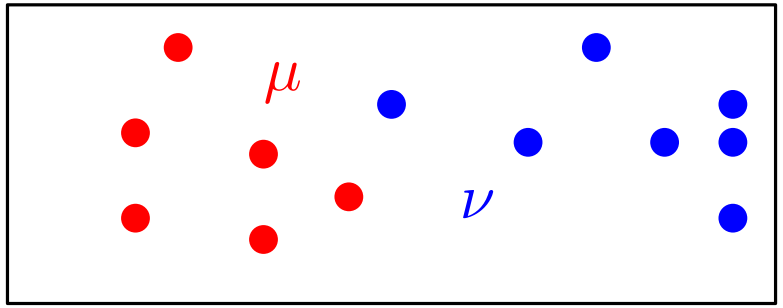
Discrete to discrete

- Simplex algorithm,
- Auction algorithm,
- Entropic regularization and Sinkhorn, etc.

[Peyré & Cuturi, Chapter 3 and 4]

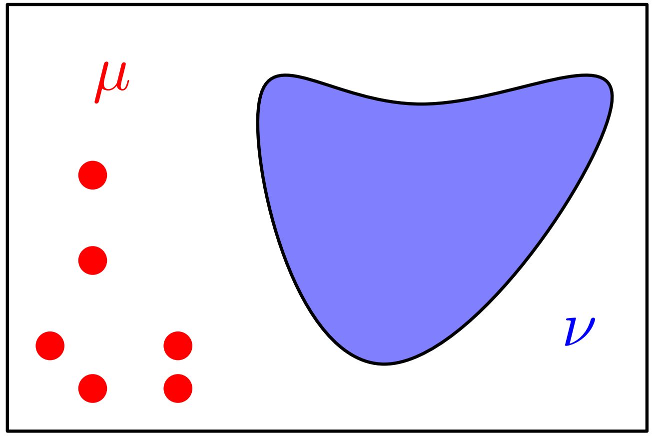
A zoology of numerical methods (often convex optimization)

Discrete to discrete



- Simplex algorithm,
- Auction algorithm,
- Entropic regularization and Sinkhorn, etc.

Semi discrete

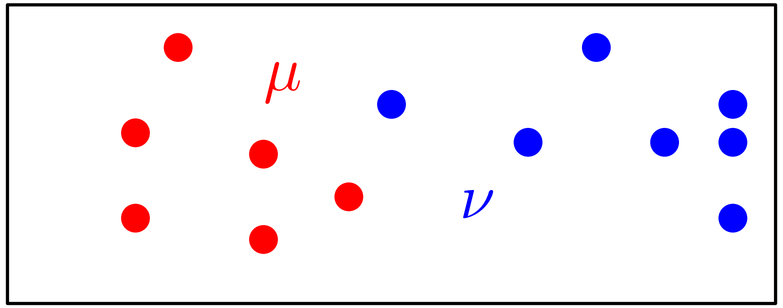


[Peyré & Cuturi, Chapter 5]

[Mérigot& Thibert (2021). Optimal transport: discretization and algorithms]

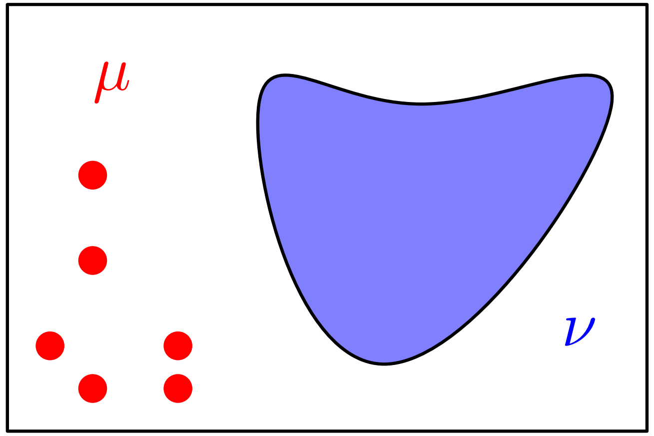
A zoology of numerical methods (often convex optimization)

Discrete to discrete



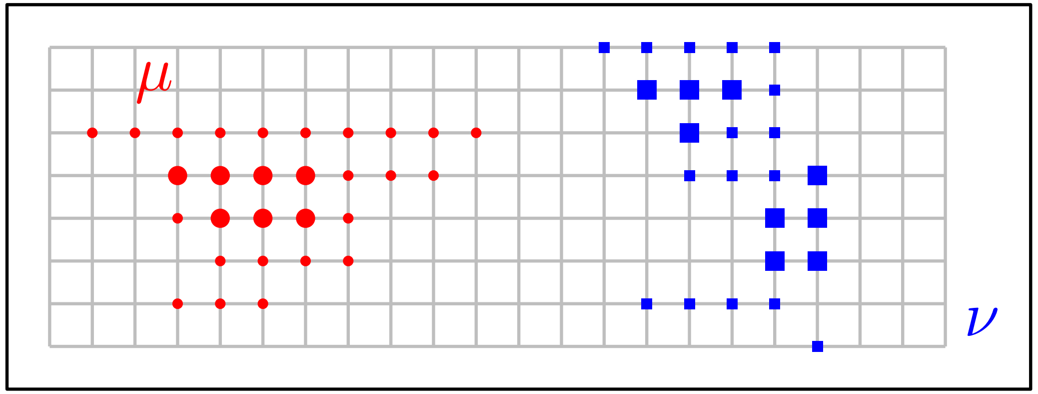
- Simplex algorithm,
- Auction algorithm,
- Entropic regularization and Sinkhorn, etc.

Semi discrete



PDEs methods

- Solving Monge-Ampère,
- dynamical formulation,
- Back and forth method, etc.



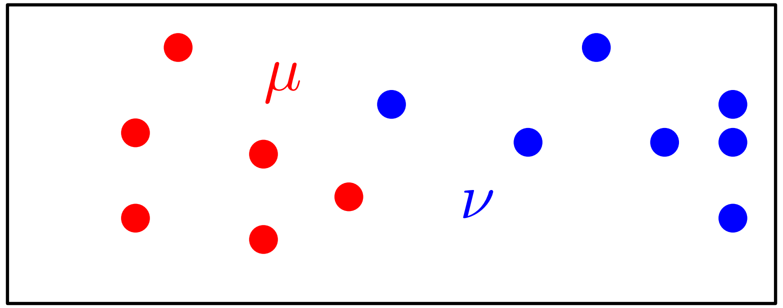
[Peyré & Cuturi, Chapter 7]

[Bonnet & Mirebeau (2022), Monotone discretization of the Monge–Ampère equation of OT]

[Jacobs & Léger (2020), A fast approach to OT: The back-and-forth method]

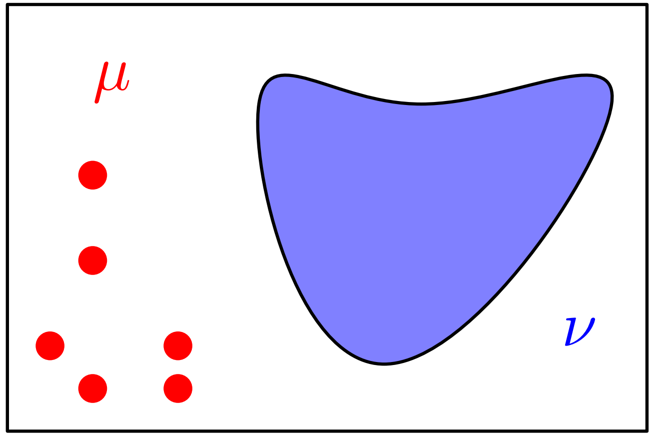
A zoology of numerical methods (often convex optimization)

Discrete to discrete



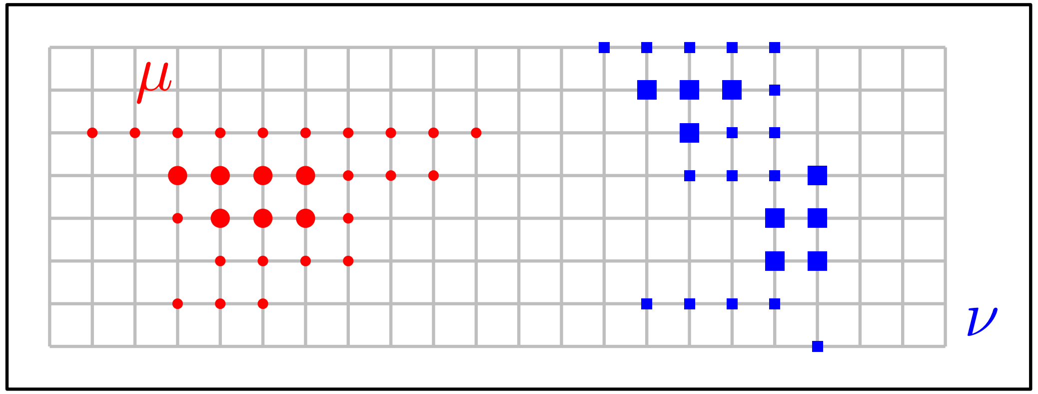
- Simplex algorithm,
- Auction algorithm,
- Entropic regularization and Sinkhorn, etc.

Semi discrete

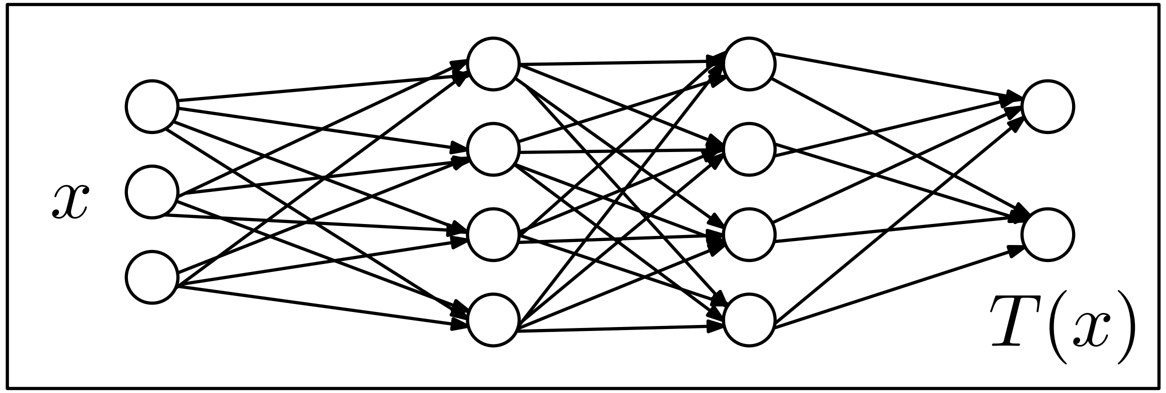


PDEs methods

- Solving Monge-Ampère,
- dynamical formulation,
- Back and forth method, etc.



And now with neural networks



[Makkuva et al (2020), Optimal transport mapping via input convex neural networks]

A reminder on the simplex

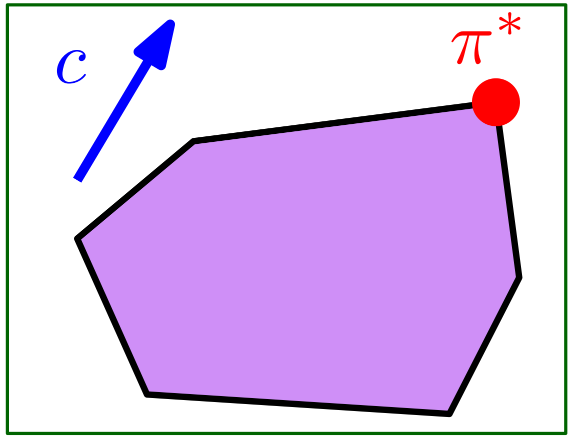
X, Y discrete: $\mathbb{P}(X = x_i) = a_i$, $\mathbb{P}(X = y_j) = b_j$ with $1 \leq i \leq n$, $1 \leq j \leq m$.

Primal

Minimize $\displaystyle\sum_{i,j} \pi_{ij} c(x_i, y_j).$

such that $\pi_{ij} \geq 0$ for all i, j.

$$\begin{cases} \sum_j \pi_{ij} = a_i \\ \sum_i \pi_{ij} = b_j \end{cases}$$



nm unknowns,
$n + m$ equality constraints.

A reminder on the simplex

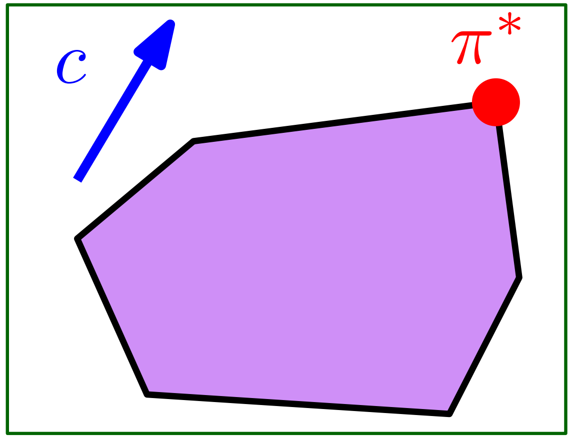
X, Y discrete: $\mathbb{P}(X = x_i) = a_i$, $\mathbb{P}(X = y_j) = b_j$ with $1 \le i \le n$, $1 \le j \le m$.

Primal

Minimize $\displaystyle\sum_{i,j} \pi_{ij} c(x_i, y_j).$

such that $\pi_{ij} \geq 0$ for all i, j.

$$\begin{cases} \sum_j \pi_{ij} = a_i \\ \sum_i \pi_{ij} = b_j \end{cases}$$



nm unknowns,
$n + m$ equality constraints.

Dual

Maximize $\displaystyle\sum_i \varphi_i a_i + \sum_j \psi_j b_j$

such that

$\varphi_i + \psi_j \le c(x_i, y_j)$ for all i, j.

$n + m$ unknowns,
nm inequality constraints.

A reminder on the simplex

X, Y discrete: $\mathbb{P}(X = x_i) = a_i$, $\mathbb{P}(X = y_j) = b_j$ with $1 \leq i \leq n$, $1 \leq j \leq m$.

Primal

Minimize $\displaystyle\sum_{i,j} \pi_{ij} c(x_i, y_j)$.

such that $\pi_{ij} \geq 0$ for all i, j.

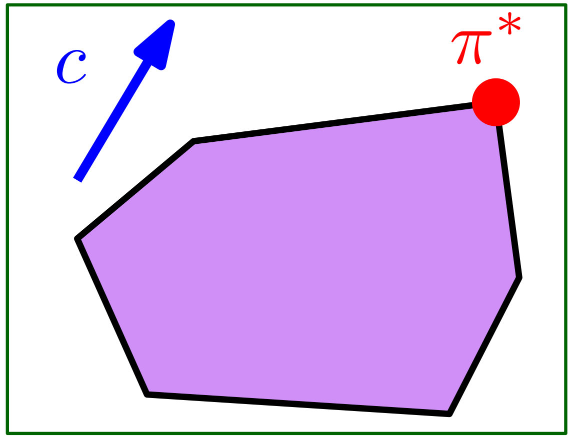
$$\begin{cases} \sum_j \pi_{ij} = a_i \\ \sum_i \pi_{ij} = b_j \end{cases}$$

Dual

Maximize $\displaystyle\sum_i \varphi_i a_i + \sum_j \psi_j b_j$

such that

$\varphi_i + \psi_j \leq c(x_i, y_j)$ for all i, j.



nm unknowns,
$n + m$ equality constraints.

$n + m$ unknowns,
nm inequality constraints.

Simplex: explore vertices of polytope. Complexity typically cubic in number points.

Entropic regularization of optimal transport

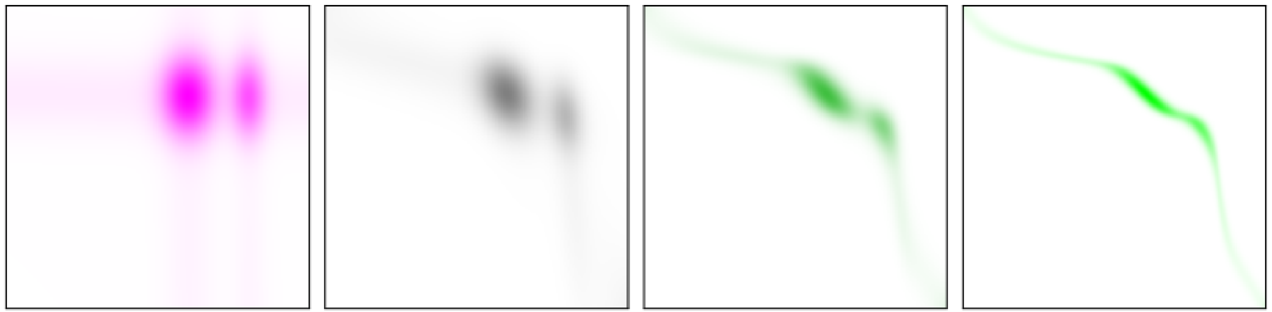
Look at a **different** problem easier to solve numerically.

Recall $\mathrm{KL}(\sigma|\theta) = \mathbb{E}_\sigma(\log \mathrm{d}\sigma/\mathrm{d}\theta)$.

$$\boxed{\mathcal{T}_{c,\varepsilon}(\mu,\nu) = \inf_{\pi} \left\{ \mathbb{E}_{(X,Y)\sim\pi}(c(X,Y)) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu) \ \text{ s.t. } \ \pi \in \Pi(\mu,\nu) \right\}}$$

Minimized if π optimal coupling

Minimized if π independent coupling



$\varepsilon = 10$ $\quad$ $\varepsilon = 1$ $\quad$ $\varepsilon = 10^{-1}$ $\quad$ $\varepsilon = 10^{-2}$

Entropic regularization of optimal transport

Look at a **different** problem easier to solve numerically.

Recall $\mathrm{KL}(\sigma|\theta) = \mathbb{E}_\sigma(\log \mathrm{d}\sigma/\mathrm{d}\theta)$.

$$\boxed{\mathcal{T}_{c,\varepsilon}(\mu,\nu) = \inf_{\pi} \{ \mathbb{E}_{(X,Y)\sim\pi}(c(X,Y)) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu) \ \text{ s.t. } \ \pi \in \Pi(\mu,\nu) \}}$$

$= \mathrm{KL}(\pi|r)$ with $r = \exp(-c/\varepsilon)\mu\otimes\nu$

Interpretation Optimal π KL projection of $r = \exp(-c/\varepsilon)\mu\otimes\nu$ on $\Pi(\mu,\nu)$.

Entropic regularization of optimal transport

Look at a **different** problem easier to solve numerically.

Recall $\mathrm{KL}(\sigma|\theta) = \mathbb{E}_\sigma(\log \mathrm{d}\sigma/\mathrm{d}\theta)$.

$$\mathcal{T}_{c,\varepsilon}(\mu,\nu) = \inf_\pi \{\mathbb{E}_{(X,Y)\sim\pi}(c(X,Y)) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu) \text{ s.t. } \pi \in \Pi(\mu,\nu)\}$$

$= \mathrm{KL}(\pi|r)$ with $r = \exp(-c/\varepsilon)\mu\otimes\nu$

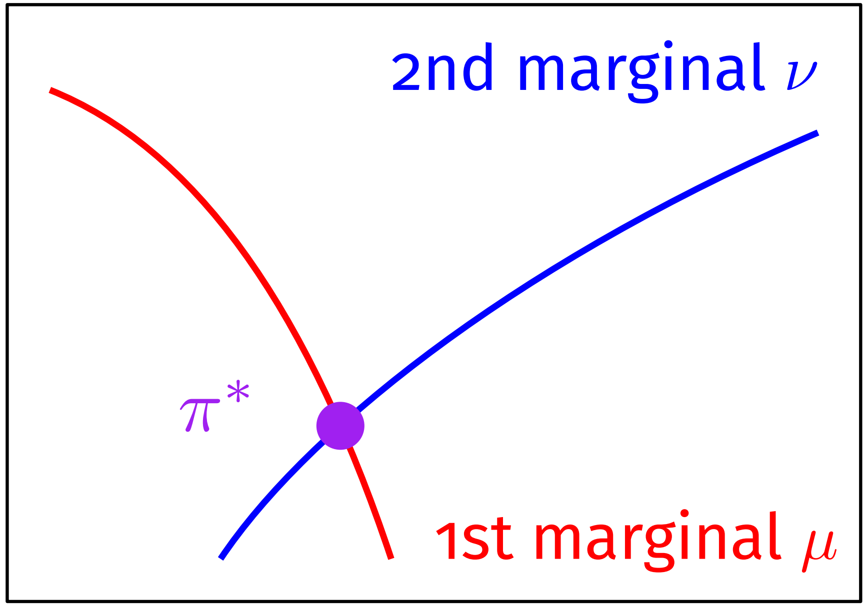
Interpretation Optimal π KL projection of $r = \exp(-c/\varepsilon)\mu\otimes\nu$ on $\Pi(\mu,\nu)$.

Proposition. A coupling $\pi \in \Pi(\mu,\nu)$ is optimal if and only if there exists $u, v : \mathbb{X} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}\pi}{\mathrm{d}\mu\otimes\nu}(x,y) = \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right).$$

Iterative Proportional Fitting Procedure (a.k.a. Sinkhorn)

Goal: find u, v such that $\pi[u,v] = \exp((u \oplus v - c)/\varepsilon)\mu \otimes \nu$ belongs to $\Pi(\mu, \nu)$.

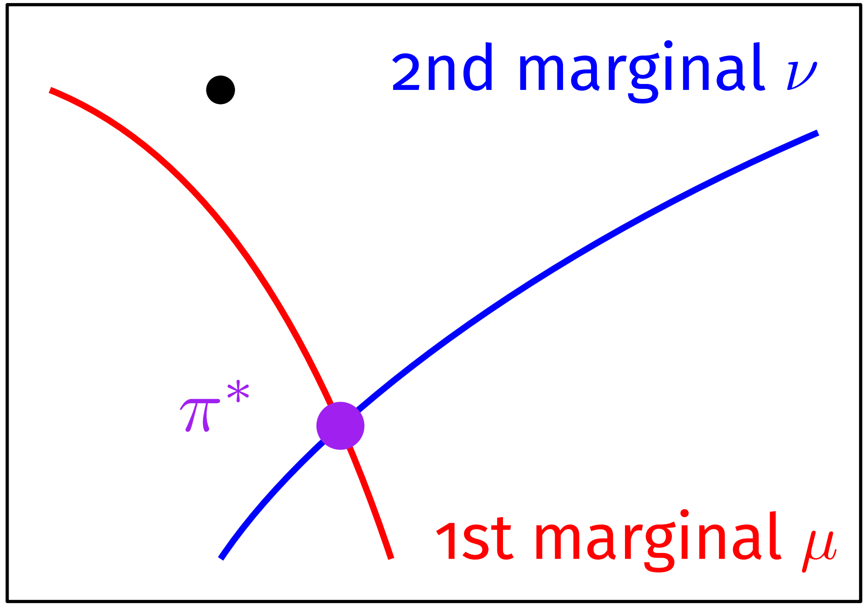


Couplings $\pi[u,v]$

Iterative Proportional Fitting Procedure (a.k.a. Sinkhorn)

Goal: find u, v such that $\pi[u,v] = \exp((u \oplus v - c)/\varepsilon)\mu \otimes \nu$ belongs to $\Pi(\mu,\nu)$.

1. Initialize $u^{(0)}$ and $v^{(0)}$.

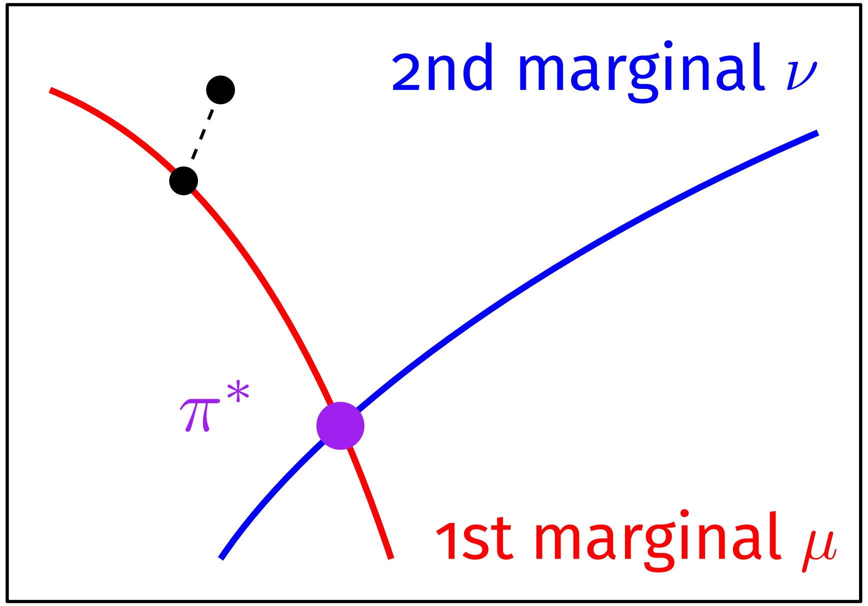


Couplings $\pi[u,v]$

Iterative Proportional Fitting Procedure (a.k.a. Sinkhorn)

Goal: find u, v such that $\pi[u,v] = \exp((u \oplus v - c)/\varepsilon)\mu \otimes \nu$ belongs to $\Pi(\mu,\nu)$.

1. Initialize $u^{(0)}$ and $v^{(0)}$.
2. Repeat: given $u^{(n)}$, $v^{(n)}$
   a. Find $u^{(n+1)}$ such that $\pi[u^{(n+1)}, v^{(n)}]$ has first marginal μ.



Couplings $\pi[u,v]$

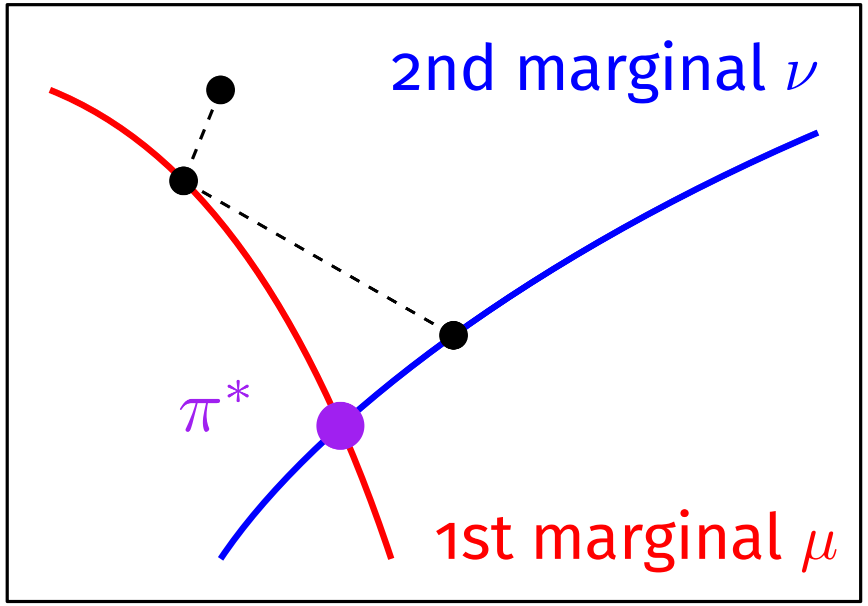
Update rule:

$$u^{(n+1)}(x) = -\varepsilon \log \int_{\mathbb{Y}} \exp\left(\frac{v^n(y) - c(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y).$$

Iterative Proportional Fitting Procedure (a.k.a. Sinkhorn)

Goal: find u, v such that $\pi[u,v] = \exp((u \oplus v - c)/\varepsilon)\mu \otimes \nu$ belongs to $\Pi(\mu,\nu)$.

1. Initialize $u^{(0)}$ and $v^{(0)}$.
2. Repeat: given $u^{(n)}$, $v^{(n)}$
   a. Find $u^{(n+1)}$ such that $\pi[u^{(n+1)}, v^{(n)}]$ has first marginal μ.
   b. Find $v^{(n+1)}$ such that $\pi[u^{(n+1)}, v^{(n+1)}]$ has 2nd marginal ν.



Couplings $\pi[u,v]$

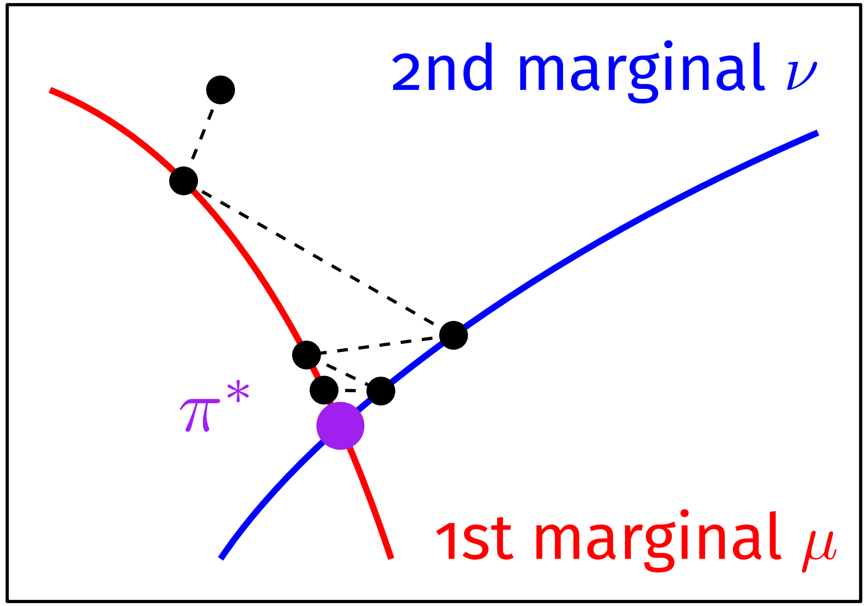
Update rule:

$$v^{(n+1)}(x) = -\varepsilon \log \int_{\mathbb{X}} \exp\left(\frac{u^{n+1}(x) - c(x,y)}{\varepsilon}\right) \, \mathrm{d}\mu(x).$$

Iterative Proportional Fitting Procedure (a.k.a. Sinkhorn)

Goal: find u, v such that $\pi[u,v] = \exp((u \oplus v - c)/\varepsilon)\mu \otimes \nu$ belongs to $\Pi(\mu,\nu)$.

1. Initialize $u^{(0)}$ and $v^{(0)}$.
2. Repeat: given $u^{(n)}$, $v^{(n)}$
   a. Find $u^{(n+1)}$ such that $\pi[u^{(n+1)}, v^{(n)}]$ has first marginal μ.
   b. Find $v^{(n+1)}$ such that $\pi[u^{(n+1)}, v^{(n+1)}]$ has 2nd marginal ν.



Couplings $\pi[u,v]$

Comments.
- **Easy** to implement (only matrix product with $\exp(-c/\varepsilon)$),
- In practice **converges quickly** (less then 20 iterations).

Additional comments on entropic optimal transport

- Lots of implementation subtelties.

POT: Python Optimal Transport

Python implementation of lots of algorithms

OptimalTransport.jl

Also exists in Julia!

downloads 202k build passing docs passing coverage 91%

Optimal Transport Tools (OTT)

JAX implementation, with automatic differentiation of the outputs

Additional comments on entropic optimal transport

- Lots of implementation subtelties.
- Why use it?

    1. Works for many problems involving OT (barycenters, gradient flows, etc.)

Additional comments on entropic optimal transport

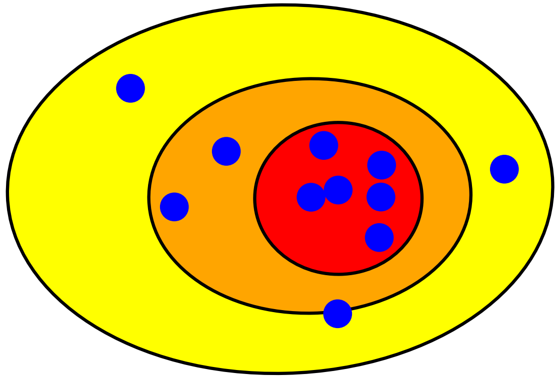
- Lots of implementation subtelties.
- Why use it?

1. Works for many problems involving OT (barycenters, gradient flows, etc.)
2. Smoother functions of the inputs μ, ν.

Additional comments on entropic optimal transport

- Lots of implementation subtelties.
- Why use it?

1. Works for many problems involving OT (barycenters, gradient flows, etc.)
2. Smoother functions of the inputs μ, ν.
3. Better sample complexity, less strong curse of dimensionality



$$\mathbb{E}\left|\mathcal{T}_{c,\varepsilon}(\mu, \mu_n) - \mathcal{T}_{c,\varepsilon}(\mu, \mu)\right| \lesssim \frac{C(\varepsilon)}{\sqrt{n}}$$

[Mena & Niles-Weed (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem]

Additional comments on entropic optimal transport

- Lots of implementation subtelties.
- Why use it?
  1. Works for many problems involving OT (barycenters, gradient flows, etc.)
  2. Smoother functions of the inputs μ, ν.
  3. Better sample complexity, less strong curse of dimensionality
  4. Debiased version for $\varepsilon \sim 1$

$$S_\varepsilon(\mu, \nu) = \mathcal{T}_{c,\varepsilon}(\mu, \nu) - \frac{1}{2}\mathcal{T}_{c,\varepsilon}(\mu, \mu) - \frac{1}{2}\mathcal{T}_{c,\varepsilon}(\nu, \nu).$$

$S_\varepsilon(\mu, \nu) \geq 0$ with equality iff $\mu = \nu$. Metrizes weak convergence.

[Feydy et al (2019) Interpolating between OT and MMD using sinkhorn divergences]

Extensions and problems using optimal transport

Extensions

- Multimarginal OT,
- Martingale OT,
- Causal OT,
- Weak OT,
- OT on graphs,
- Unbalanced OT,
- Matrix valued OT,
- Quantum OT,
- Extended OT,
- Sliced OT,
- Gromov Wasserstein distance, etc.

Extensions and problems using optimal transport

Extensions

- Multimarginal OT,
- Martingale OT,
- Causal OT,
- Weak OT,
- OT on graphs,
- Unbalanced OT,
- Matrix valued OT,
- Quantum OT,
- Extended OT,
- Sliced OT,
- Gromov Wasserstein distance, etc.

Variational problems involving optimal transport

- Barycenters,
- Gradient flows,
- Mean Field Games,
- Trajectory inference,
- Wasserstein GANs,
- Schrödinger bridges for diffusion matching, etc.

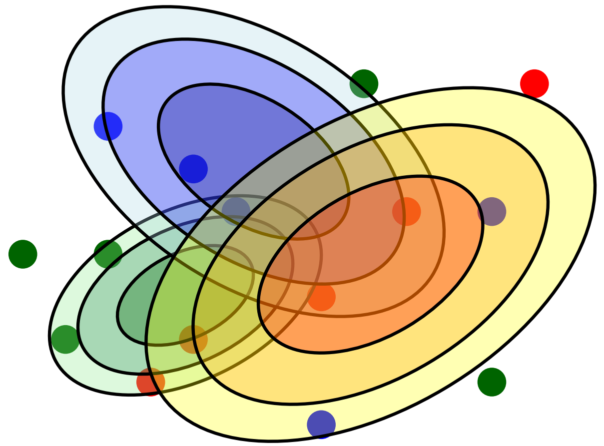
Extensions and problems using optimal transport

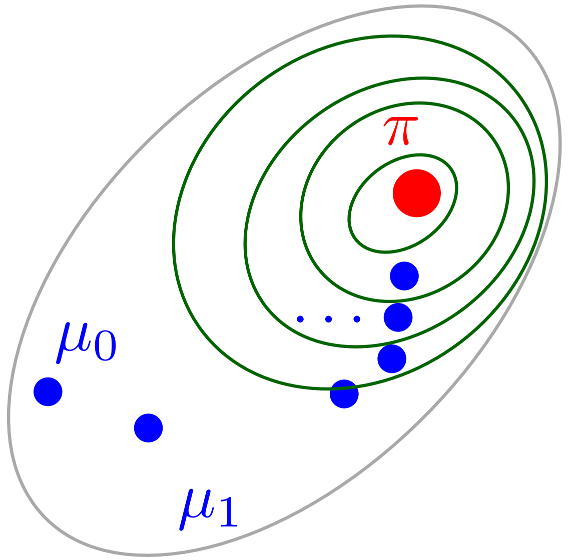
Extensions

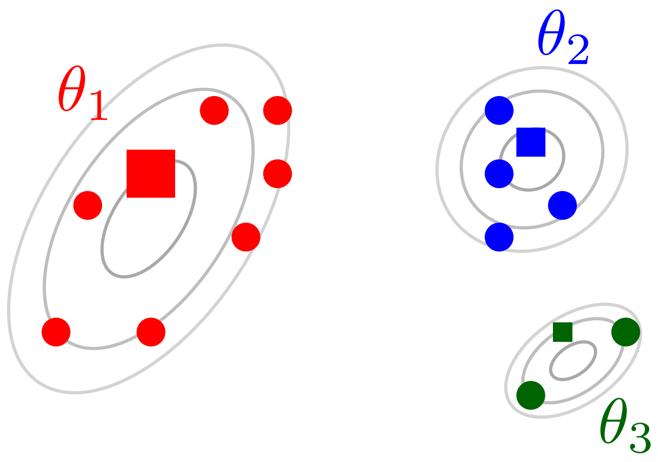
- Multimarginal OT,
- Martingale OT,

Variational problems involving optimal transport

- Barycenters,
- Gradient flows,







And optimal transport is used in Bayesian statistics!

Let's do the break first!

Introduction to optimal transport for Bayesian statistics

Part II

Hugo Lavenant

Bocconi University

2024 ISBA World Meeting

Venice (Italy), July 1, 2024

1 - Wasserstein distances in Bayesian statistics

2 - Wasserstein barycenters for model selection and scalable Bayes

Honourable mentions

(topics I researched but did not include)

3 - Looking at sampling with the geometry of optimal transport

1 - Wasserstein distances in Bayesian statistics

2 - Was

and sca

Honou

(topics I res

3 - Lool

optimal transport

[Nguyen (2013). Convergence of latent mixing measures in finite and infinite mixture models.]

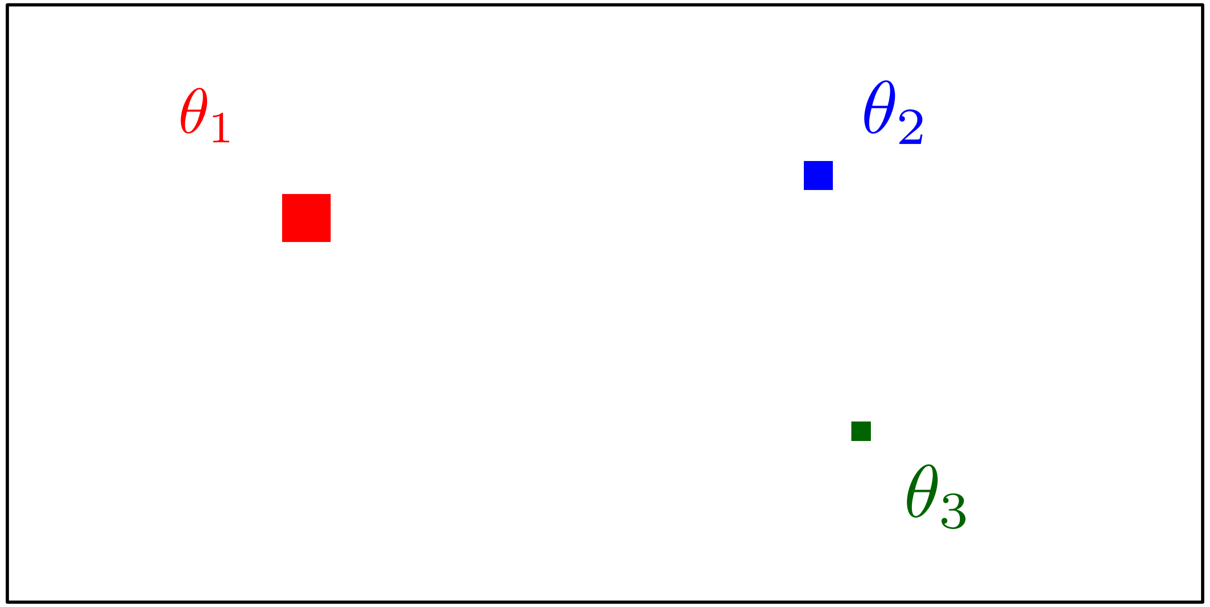
[Nguyen (2016). Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure.]

[Catalano & Lavenant (2024). Hierarchical Integral Probability Metrics]

Density estimation or parameter estimation?

Mixing measure

$$(\theta_1, \ldots, \theta_K) \sim \pi_1 \quad \text{(cluster parameter)}$$
$$(\lambda_1, \ldots, \lambda_K) \sim \pi_2 \quad \text{(weights)}$$
$$\lambda_k \geq 0, \sum \lambda_k = 1$$

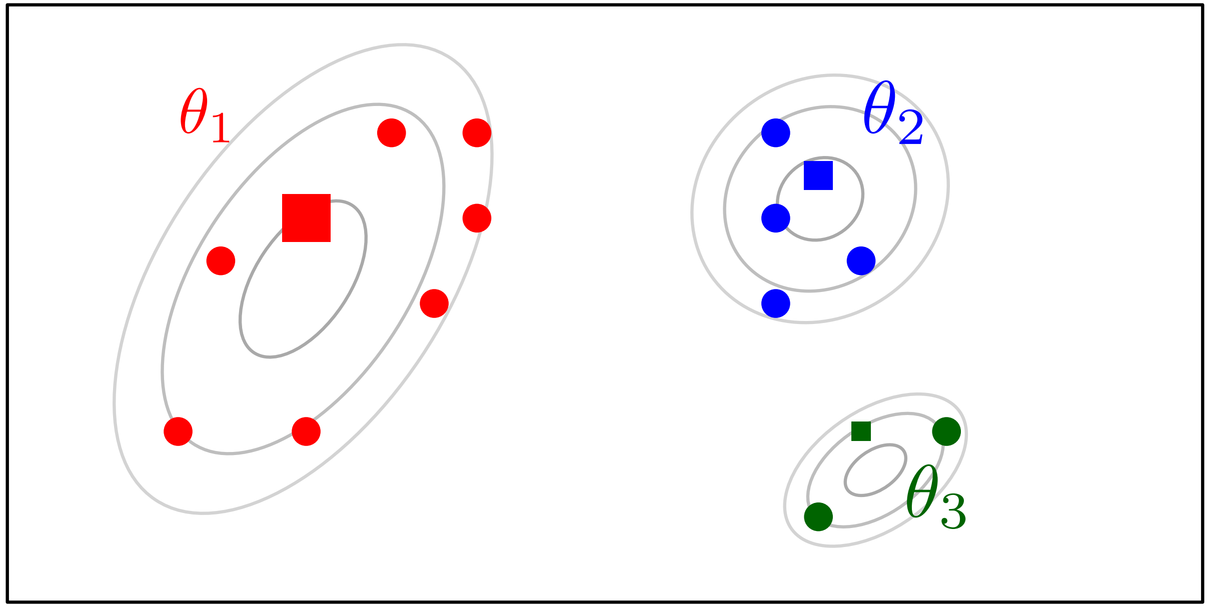


Density estimation or parameter estimation?

Mixing measure

$$(\theta_1, \ldots, \theta_K) \sim \pi_1 \qquad \text{(cluster parameter)}$$
$$(\lambda_1, \ldots, \lambda_K) \sim \pi_2 \qquad \text{(weights)}$$

$$\lambda_k \geq 0, \sum \lambda_k = 1$$



Sampling the data

iid data, for any $i = 1, \ldots, n$

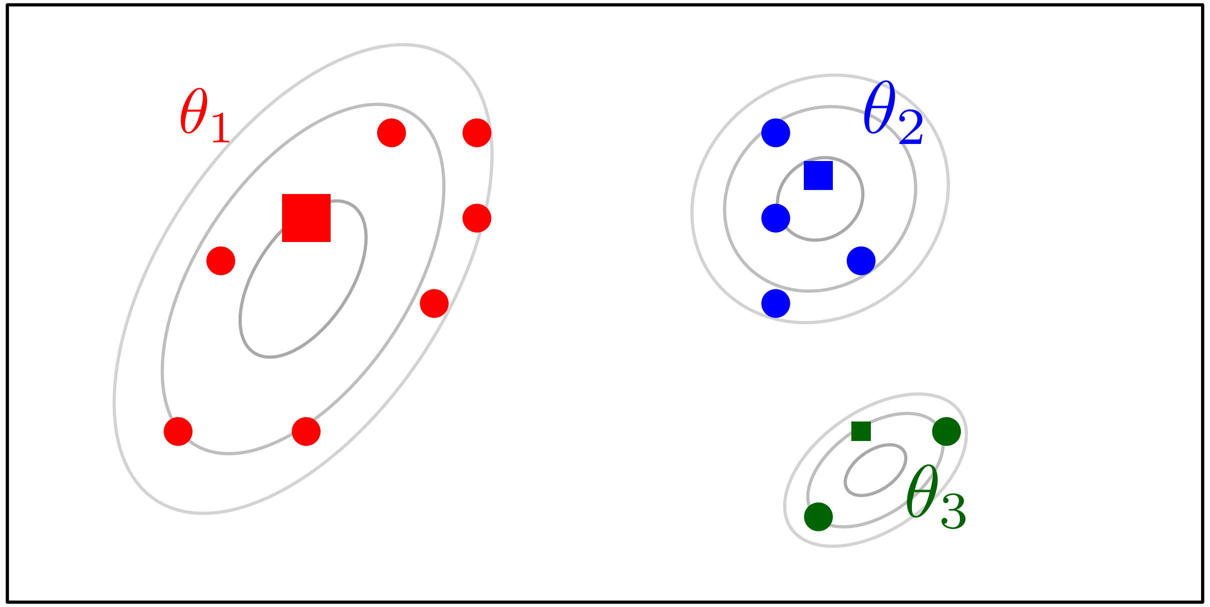
$$k_i | \theta, \lambda \in \{1, \ldots, K\} \sim (\lambda_1, \ldots, \lambda_K).$$

$$X_i | k_i, \theta, \lambda \sim f(\cdot | \theta_{k_i})$$

Density estimation or parameter estimation?

Mixing measure

$$(\theta_1, \ldots, \theta_K) \sim \pi_1 \qquad \text{(cluster parameter)}$$
$$(\lambda_1, \ldots, \lambda_K) \sim \pi_2 \qquad \text{(weights)}$$
$$\lambda_k \geq 0, \sum \lambda_k = 1$$



Sampling the data

iid data, for any $i = 1, \ldots, n$

$$k_i | \theta, \lambda \in \{1, \ldots, K\} \sim (\lambda_1, \ldots, \lambda_K).$$
$$X_i | k_i, \theta, \lambda \sim f(\cdot | \theta_{k_i})$$

Write $p_{\theta^0, \lambda^0} = \sum_k \lambda_k^0 f(\cdot | \theta_k^0)$ for the density of X_1 under (θ^0, λ^0).

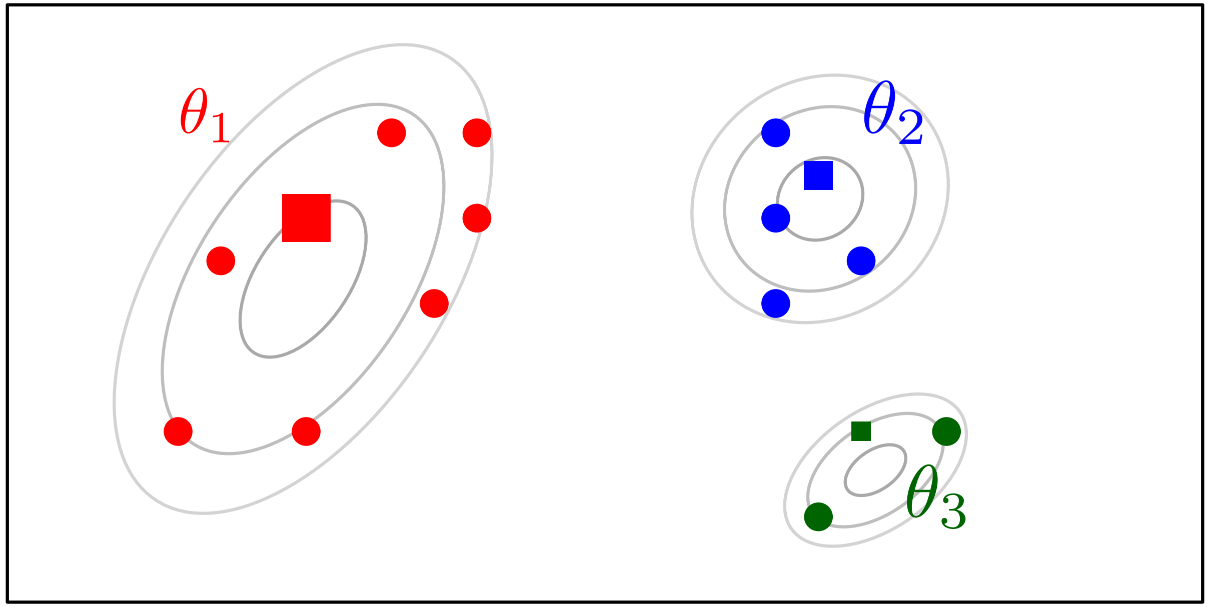
Question. If truth (θ^0, λ^0), do you want to infer:

- p_{θ^0, λ^0} (**density estimation**)?
- θ^0 and λ^0? (**parameter estimation**)

Density estimation or parameter estimation?

Mixing measure

$$(\theta_1, \dots, \theta_K) \sim \pi_1 \quad \text{(cluster parameter)}$$
$$(\lambda_1, \dots, \lambda_K) \sim \pi_2 \quad \text{(weights)}$$
$$\lambda_k \geq 0, \sum \lambda_k = 1$$



Sampling the data

iid data, for any $i = 1, \dots, n$

$$k_i | \theta, \lambda \in \{1, \dots, K\} \sim (\lambda_1, \dots, \lambda_K).$$
$$X_i | k_i, \theta, \lambda \sim f(\cdot | \theta_{k_i})$$

Write $p_{\theta^0, \lambda^0} = \sum_k \lambda_k^0 f(\cdot | \theta_k^0)$ for the density of X_1 under (θ^0, λ^0).

Question. If truth (θ^0, λ^0), do you want to infer:

- p_{θ^0, λ^0} (**density estimation**)?
- θ^0 and λ^0? (**parameter estimation**)

Problem: how to measure the quality of parameter estimation?

Mixing measure and Wasserstein distances

Mixing measure

$$(\theta_1, \ldots, \theta_K) \sim \pi_1$$ (cluster parameter)

$$(\lambda_1, \ldots, \lambda_K) \sim \pi_2$$ (weights)

$$\lambda_k \geq 0, \sum \lambda_k = 1$$

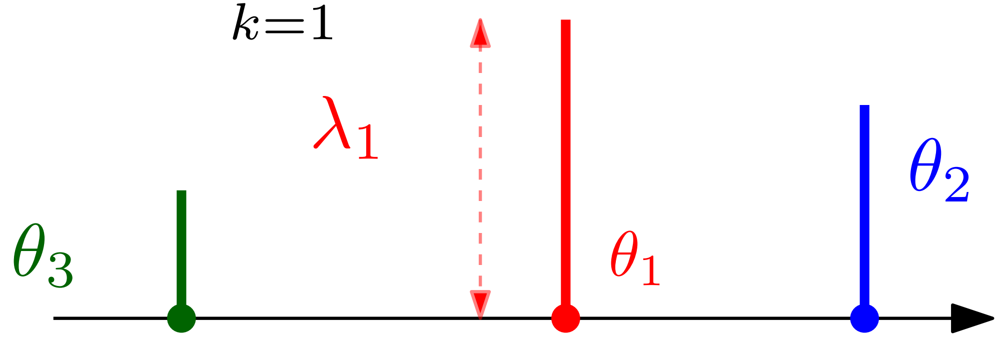
Mixing measure and Wasserstein distances

Mixing measure

$$(\theta_1, \ldots, \theta_K) \sim \pi_1 \quad \text{(cluster parameter)}$$
$$(\lambda_1, \ldots, \lambda_K) \sim \pi_2 \quad \text{(weights)}$$
$$\lambda_k \geq 0, \sum \lambda_k = 1$$

Define probability measure on Θ:

$$G = \sum_{k=1}^{K} \lambda_k \delta_{\theta_k}.$$



- No label switching issue:

  Same G for $((\lambda_1, \theta_1), (\lambda_2, \theta_2))$ and $((\lambda_2, \theta_2), (\lambda_1, \theta_1))$

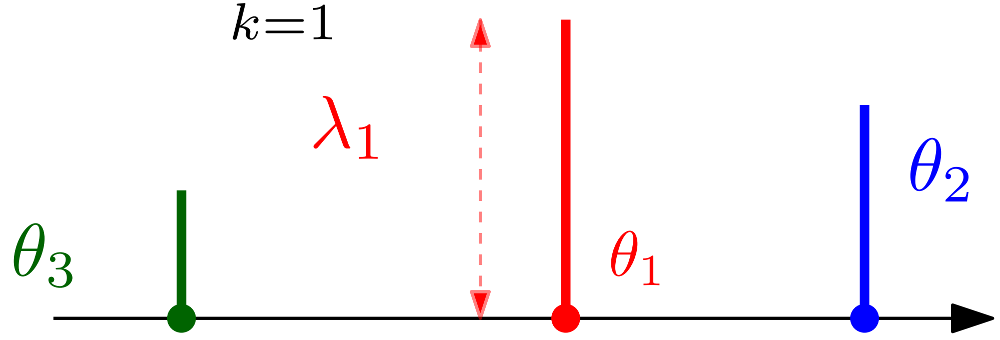
- No problem with K finite or infinite (or even G continuous).

Mixing measure

$$(\theta_1, \ldots, \theta_K) \sim \pi_1 \quad \text{(cluster parameter)}$$
$$(\lambda_1, \ldots, \lambda_K) \sim \pi_2 \quad \text{(weights)}$$
$$\lambda_k \geq 0, \; \sum \lambda_k = 1$$

Define probability measure on Θ:

$$G = \sum_{k=1}^{K} \lambda_k \delta_{\theta_k}.$$



- No label switching issue:

  Same G for $((\lambda_1, \theta_1), (\lambda_2, \theta_2))$ and $((\lambda_2, \theta_2), (\lambda_1, \theta_1))$

- No problem with K finite or infinite (or even G continuous).

Use **Wasserstein distances** to compare different G's.

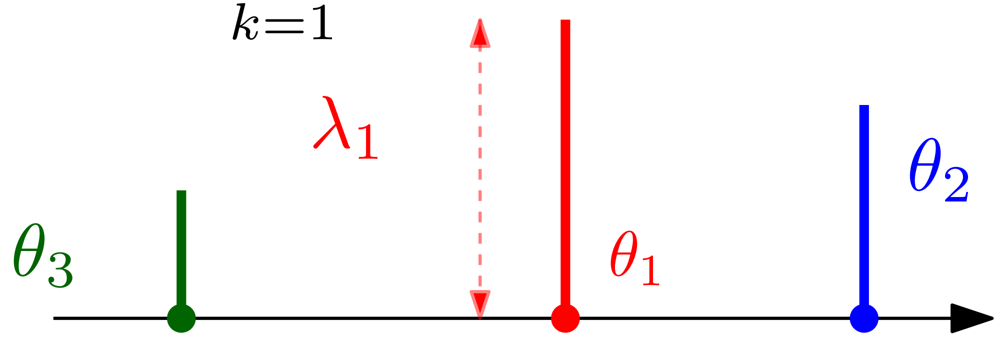
Mixing measure and Wasserstein distances

Mixing measure

$$(\theta_1, \dots, \theta_K) \sim \pi_1 \quad \text{(cluster parameter)}$$
$$(\lambda_1, \dots, \lambda_K) \sim \pi_2 \quad \text{(weights)}$$
$$\lambda_k \geq 0, \sum \lambda_k = 1$$

Define probability measure on Θ:

$$G = \sum_{k=1}^{K} \lambda_k \delta_{\theta_k}.$$



- No label switching issue:

  Same G for $((\lambda_1, \theta_1), (\lambda_2, \theta_2))$ and $((\lambda_2, \theta_2), (\lambda_1, \theta_1))$

- No problem with K finite or infinite (or even G continuous).

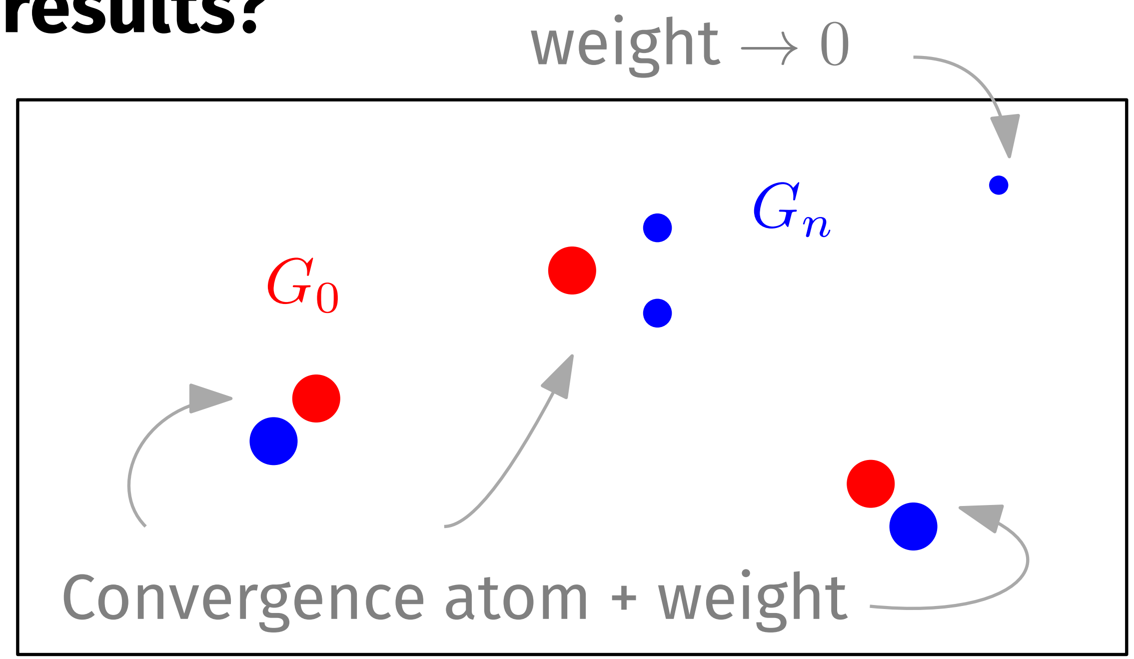
Use **Wasserstein distances** to compare different G's.

- Prior: G random, described by $\mathcal{P}(\mathcal{P}(\Theta))$ (Which distance to use?)

How to interpret results?

Lemma. If G_0 finite number of atoms and $G_n \to G_0$ in W_p distance:

- For atoms of G_n getting closer to those of G_0, convergence of weights.
- If atom of G_n does not converge to one of G_0, its weight converge to 0.

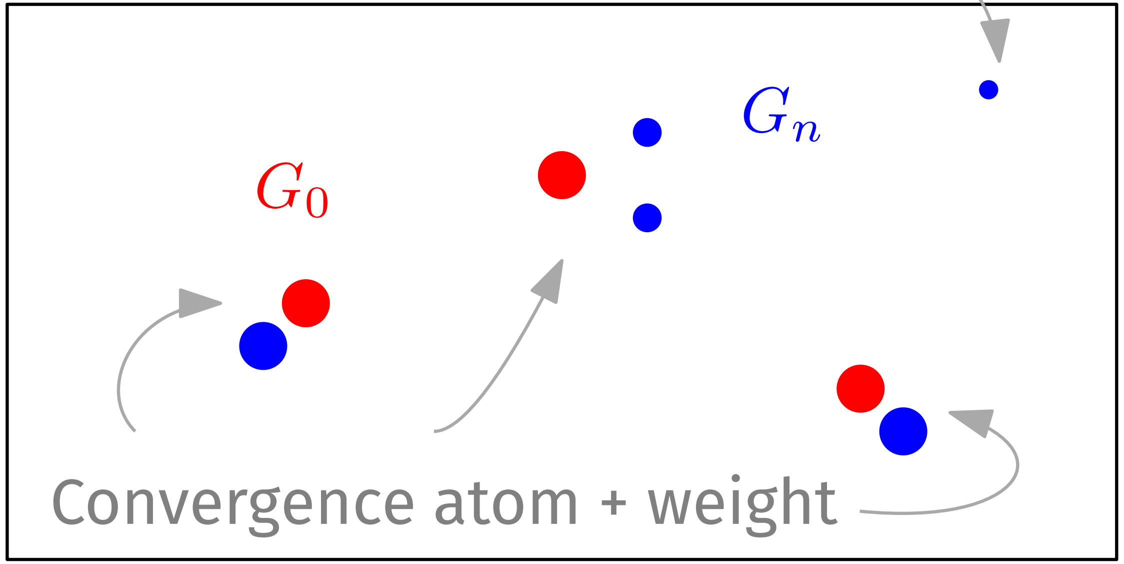


How to interpret results?

Lemma. If G_0 finite number of atoms and $G_n \to G_0$ in W_p distance:

- For atoms of G_n getting closer to those of G_0, convergence of weights.
- If atom of G_n does not converge to one of G_0, its weight converge to 0.

weight $\to 0$



How does $G \in \mathcal{P}(\Theta)$ relate to $p_G \in \mathcal{P}(\mathbb{X})$ the distribution of data?

Easy bound. Take d a distance on $\mathcal{P}(\mathbb{X})$ with d^p convex. Define W_p with base distance $d(f(\cdot|\theta), f(\cdot|\theta'))$. Then:

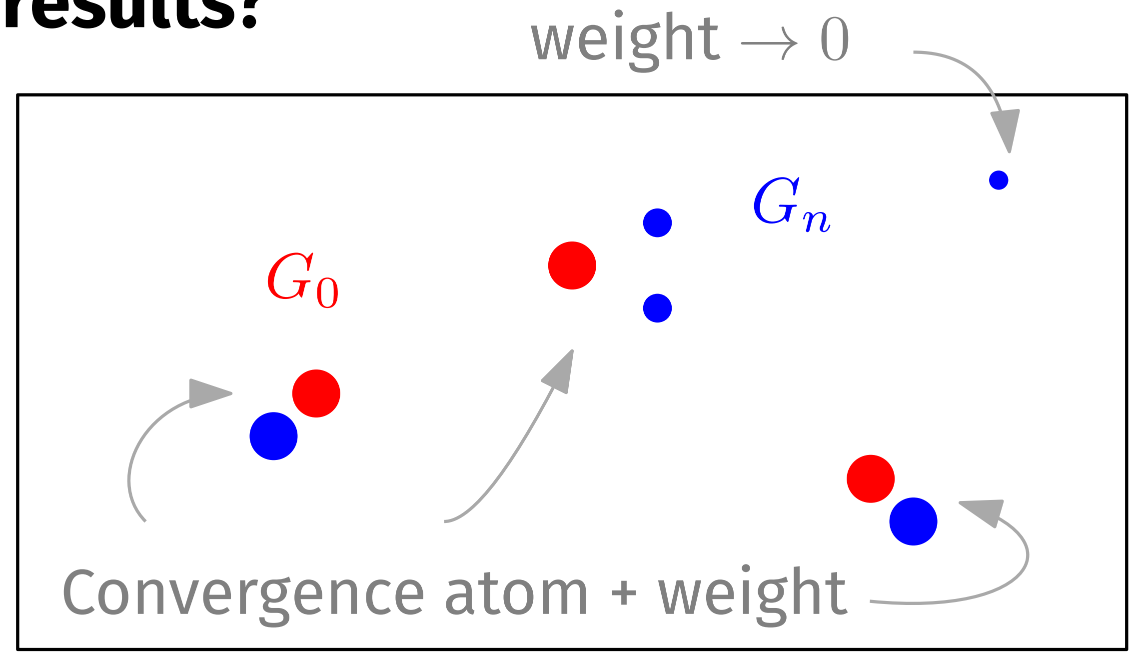
$$d(p_G, p'_G)^p \leq W_p^p(G, G').$$

e.g. Total Variation, Hellinger, KL and $p = 1$, or W_p.

How to interpret results?

Lemma. If G_0 finite number of atoms and $G_n \to G_0$ in W_p distance:

- For atoms of G_n getting closer to those of G_0, convergence of weights.
- If atom of G_n does not converge to one of G_0, its weight converge to 0.



How does $G \in \mathcal{P}(\Theta)$ relate to $p_G \in \mathcal{P}(\mathbb{X})$ the distribution of data?

Easy bound. Take d a distance on $\mathcal{P}(\mathbb{X})$ with d^p convex. Define W_p with base distance $d(f(\cdot|\theta), f(\cdot|\theta'))$. Then:

$$d(p_G, p'_G)^p \leq W_p^p(G, G').$$

e.g. Total Variation, Hellinger, KL and $p = 1$, or W_p.

Can one get, for $\alpha > 0$:

$$W_p(G, G') \lesssim d(p_G, p_{G'})^\alpha ?$$

Yes but need assumptions on f, this is a **deconvolution** problem.

(Sample of) examples: Posterior Contraction Rate

Posterior: $\pi_n^* = \mathrm{Law}(G|X_1, \ldots, X_n)$.

$$G = \sum_k \lambda_k \delta_{\theta_k} \sim \pi$$

iid data, for any $i = 1, \ldots, n$

$$theta_i | G \sim G$$

$$X_i | k_i, G \sim f(\cdot|\theta_{k_i})$$

Posterior contraction rate. Assume data from mixture with mixing measure G_0, find ε_n such that $\pi_n^* \left(W_2(G, G_0) \geq \varepsilon_n\right) \to 0.$

(Sample of) examples: Posterior Contraction Rate

Posterior: $\pi_n^* = \mathrm{Law}(G|X_1, \dots, X_n)$.

$$G = \sum_k \lambda_k \delta_{\theta_k} \sim \pi$$

iid data, for any $i = 1, \dots, n$

$$theta_i | G \sim G$$

$$X_i | k_i, G \sim f(\cdot|\theta_{k_i})$$

Posterior contraction rate. Assume data from mixture with mixing measure G_0, find ε_n such that $\pi_n^* \left(W_2(G, G_0) \geq \varepsilon_n\right) \to 0.$

$f(\cdot|\theta)$ normal distribution of mean θ

Infinite number clusters!

- $\varepsilon_n \asymp \log(n)^{1/2} n^{-1/4}$ $\quad$ G_0 **finite** number atoms.
- $\varepsilon_n \asymp \log(n)^{-1/2}$ $\quad$ G_0 Dirichlet process with support non empty interior in $\mathbb{R}^d$.

And many others: interplay between smoothness of $f(\cdot|\theta)$ and support G_0.

(Sample of) examples: Posterior Contraction Rate

Posterior: $\pi_n^* = \mathrm{Law}(G|X_1, \ldots, X_n)$.

$$G = \sum_k \lambda_k \delta_{\theta_k} \sim \pi$$

iid data, for any $i = 1, \ldots, n$

$$theta_i|G \sim G$$

$$X_i|k_i, G \sim f(\cdot|\theta_{k_i})$$

Posterior contraction rate. Assume data from mixture with mixing measure G_0, find ε_n such that $\pi_n^* \left(W_2(G, G_0) \geq \varepsilon_n\right) \to 0.$

$f(\cdot|\theta)$ normal distribution of mean θ

- $\varepsilon_n \asymp \log(n)^{1/2} n^{-1/4}$ — G_0 **finite** number atoms.
- $\varepsilon_n \asymp \log(n)^{-1/2}$ — G_0 Dirichlet process with support non empty interior in $\mathbb{R}^d$.

Infinite number clusters!

And many others: interplay between smoothness of $f(\cdot|\theta)$ and support G_0.

Extension. Take a Hierarchical Dirichlet process, study how the borrowing of information changes the rate.

Adding one more layer of randomness

G mixing measure, in $\mathcal{P}(\Theta)$.
Prior π on G: "probability over probability", in $\mathcal{P}(\mathcal{P}(\Theta))$

Adding one more layer of randomness

G mixing measure, in $\mathcal{P}(\Theta)$.
Prior π on G: "probability over probability", in $\mathcal{P}(\mathcal{P}(\Theta))$

Definition. If $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{P}(\Theta))$, the "Wasserstein over Wasserstein" distance is:

$$W_{W,p}^p(\pi_1, \pi_2) = \inf_{\gamma \in \Pi(\pi_1, \pi_2)} \mathbb{E}_{(\tilde{G}_1, \tilde{G}_2) \sim \gamma} \left[W_p^p(\tilde{G}_1, \tilde{G}_2) \right].$$

Adding one more layer of randomness

G mixing measure, in $\mathcal{P}(\Theta)$.
Prior π on G: "probability over probability", in $\mathcal{P}(\mathcal{P}(\Theta))$

Definition. If $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{P}(\Theta))$, the "Wasserstein over Wasserstein" distance is:

$$W_{W,p}^p(\pi_1, \pi_2) = \inf_{\gamma \in \Pi(\pi_1, \pi_2)} \mathbb{E}_{(\tilde{G}_1, \tilde{G}_2) \sim \gamma} \left[W_p^p(\tilde{G}_1, \tilde{G}_2) \right].$$

Example: if π_1, π_2 Dirichlet processes with same concentration parameter:

$$W_{W,p}(\pi_1, \pi_2) = W_p(P_{0,1}, P_{0,2})$$

Distance between base measures

(Extends to Species Sampling Processes)

Adding one more layer of randomness

G mixing measure, in $\mathcal{P}(\Theta)$.
Prior π on G: "probability over probability", in $\mathcal{P}(\mathcal{P}(\Theta))$

Definition. If $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{P}(\Theta))$, the "Wasserstein over Wasserstein" distance is:

$$W_{W,p}^p(\pi_1, \pi_2) = \inf_{\gamma \in \Pi(\pi_1, \pi_2)} \mathbb{E}_{(\tilde{G}_1, \tilde{G}_2) \sim \gamma} \left[W_p^p(\tilde{G}_1, \tilde{G}_2) \right].$$

Example: if π_1, π_2 Dirichlet processes with same concentration parameter:

$$W_{W,p}(\pi_1, \pi_2) = W_p(P_{0,1}, P_{0,2})$$

Distance between base measures

(Extends to Species Sampling Processes)

With Marta Catalano:

- alternative distances (e.g. between laws of Completely Random Measures),
- Use to measure dependence,
- Merging of opinions: do posterior converge if more and more data is coming but the priors are different?

[Catalano et al (2023). A Wasserstein index of dependence for random measures]
[Catalano & Lavenant (2022). Merging Rate of Opinions via Optimal Transport on Random Measures]

1 - Wasserstein distances in Bayesian statistics

2 - Wasserstein barycenters for model selection and scalable Bayes

Honourable

(topics I researched

[Agueh & Carlier (2011). Barycenters in the Wasserstein space]

[Backhoff-Veraguas, Fontbona, Rios & Tobar (2022). Bayesian learning with Wasserstein barycenters.]

3 - Looking a

optimal tran

[Srivastava, Li, & Dunson (2018). Scalable Bayes via barycenter in Wasserstein space.]

Nonparametric Model selection

$$\theta \sim \pi$$

$$X_1, \ldots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

$$\theta | X_1, \ldots, X_n$$

Posterior

Question: which value to return if asked a point estimate of p_θ?

Nonparametric Model selection

$$\theta \sim \pi$$

$$X_1, \ldots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

$$\theta | X_1, \ldots, X_n$$

Posterior

Question: which value to return if asked a point estimate of p_θ?

- p_{θ^*} for θ^* Maximum a posteriori
- p_{θ^*} for $\theta^* = \mathbb{E}(\theta | X_1, \ldots, X_n)$ posterior mean

May depend on the parametrization $\theta \to p_\theta$, not only on the random proba $p_\theta, \theta \sim \pi$.

Nonparametric Model selection

$$\theta \sim \pi$$
$$X_1, \dots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$
$$\theta | X_1, \dots, X_n$$

Posterior

Question: which value to return if asked a point estimate of p_θ?

- p_{θ^*} for θ^* Maximum a posteriori
- p_{θ^*} for $\theta^* = \mathbb{E}(\theta | X_1, \dots, X_n)$ posterior mean

May depend on the parametrization $\theta \to p_\theta$, not only on the random proba p_θ, $\theta \sim \pi$.

- $\mathbb{E}(p_\theta | X_1, \dots, X_n)$ Bayesian model average.

$\rightsquigarrow$ if $X | p_\theta = \mathcal{N}(\theta, I)$ then $\mathbb{E}(p_\theta | X_1, \dots, X_n)$ mixture of Gaussians while p_{θ_*} Gaussian.

May be hard to interpret

Nonparametric Model selection

$$\theta \sim \pi$$
$$X_1, \dots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$
$$\theta | X_1, \dots, X_n$$

Posterior

Question: which value to return if asked a point estimate of p_θ?

- p_{θ^*} for θ^* Maximum a posteriori
- p_{θ^*} for $\theta^* = \mathbb{E}(\theta | X_1, \dots, X_n)$ posterior mean

May depend on the parametrization $\theta \to p_\theta$, not only on the random proba $p_\theta, \theta \sim \pi$.

- $\mathbb{E}(p_\theta | X_1, \dots, X_n)$ Bayesian model average.

  $\rightsquigarrow$ if $X | p_\theta = \mathcal{N}(\theta, I)$ then $\mathbb{E}(p_\theta | X_1, \dots, X_n)$ mixture of Gaussians while p_{θ_*} Gaussian.

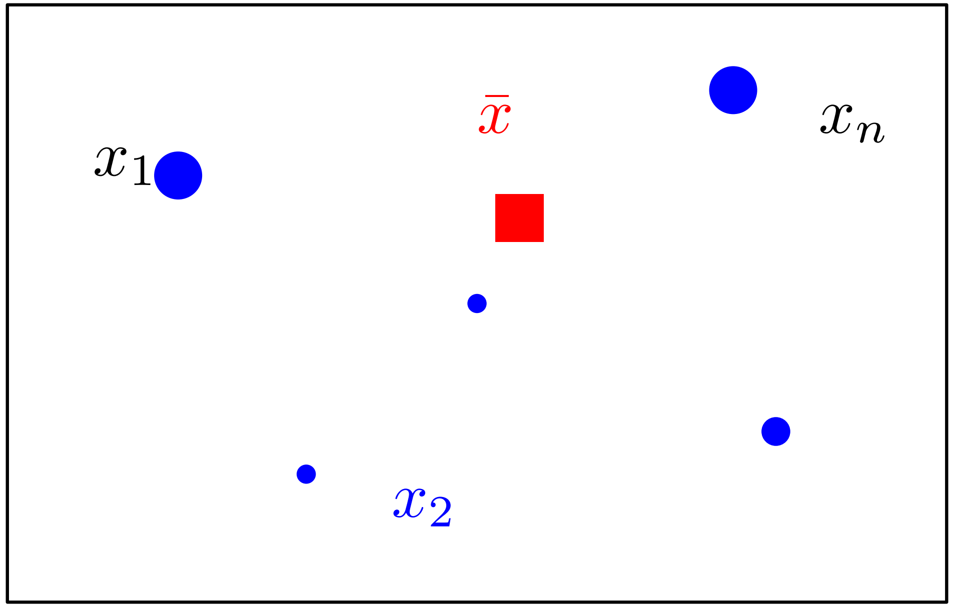
May be hard to interpret

Proposal: take the **Wasserstein barycenter** of the p_θ.

Barycenters

If $x_1, \ldots x_n$ points in $\mathbb{R}^d$ and weights $\lambda_1, \ldots, \lambda_n$ which sum up to 1,

barycenter: $$\bar{x} = \sum_{i=1}^{n} \lambda_i x_i$$



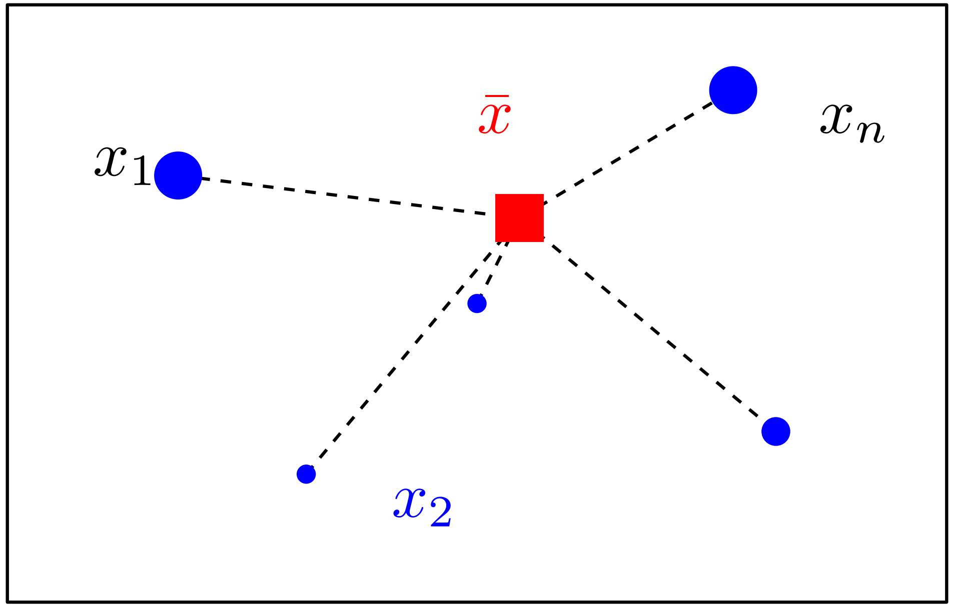
Barycenters

If $x_1, \ldots x_n$ points in $\mathbb{R}^d$ and weights $\lambda_1, \ldots, \lambda_n$ which sum up to 1,

barycenter: $$\bar{x} = \sum_{i=1}^{n} \lambda_i x_i$$

Lemma. The barycenter $\bar{x}$ minimizes

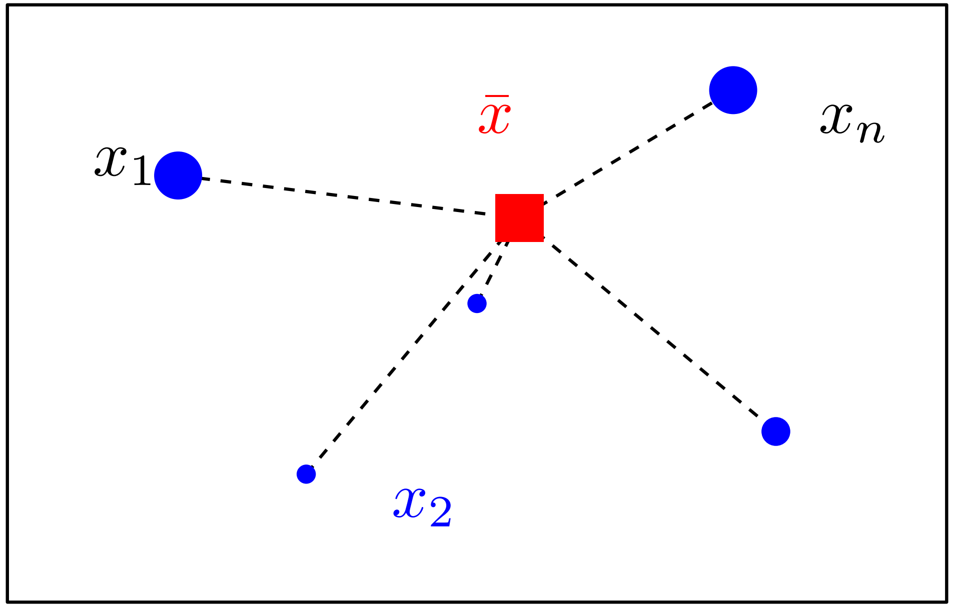
$$x \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i \|x - x_i\|^2.$$



Barycenters

If $x_1, \dots x_n$ points in $\mathbb{R}^d$ and weights $\lambda_1, \dots, \lambda_n$ which sum up to 1,

barycenter: $$\bar{x} = \sum_{i=1}^{n} \lambda_i x_i$$



Lemma. The barycenter $\bar{x}$ minimizes

$$x \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i \|x - x_i\|^2.$$

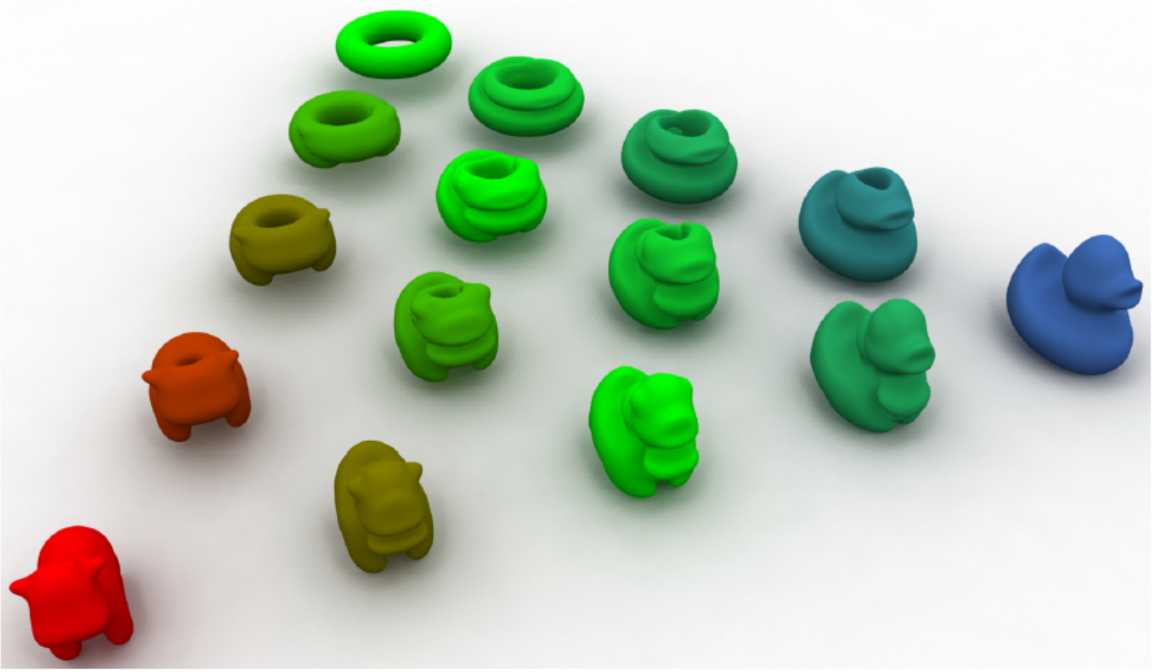
Definition. If $\mu_1, \dots, \mu_n$ in $\mathcal{P}_2(\mathbb{R}^d)$ and $\lambda_1, \dots, \lambda_n$ non-negative and sum up to 1, **a Wasserstein barycenter** is a measure $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ minimizing

$$\mu \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

Existence and uniqueness

Theorem. There always exists a barycenter. It is unique if at least one μ_i has a density with respect to Lebesgue measure.

$$\bar{\mu} \in \arg\min_{\mu} \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$



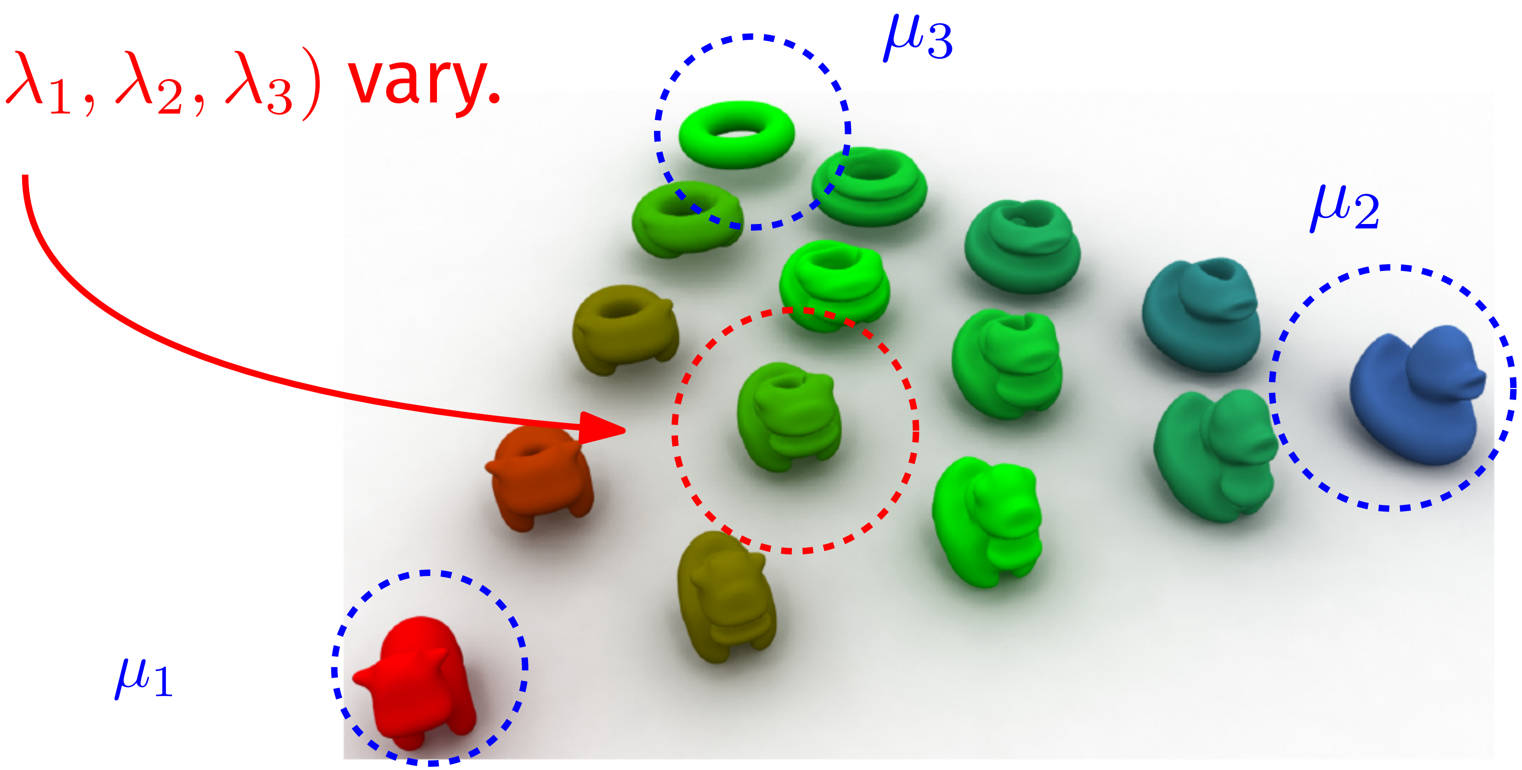
[Solomon et al (2015). Convolutional Wasserstein distances]

Existence and uniqueness

Theorem. There always exists a barycenter. It is unique if at least one μ_i has a density with respect to Lebesgue measure.

$$\bar{\mu} \in \arg\min_{\mu} \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

$\bar{\mu}$ as $(\lambda_1, \lambda_2, \lambda_3)$ vary.



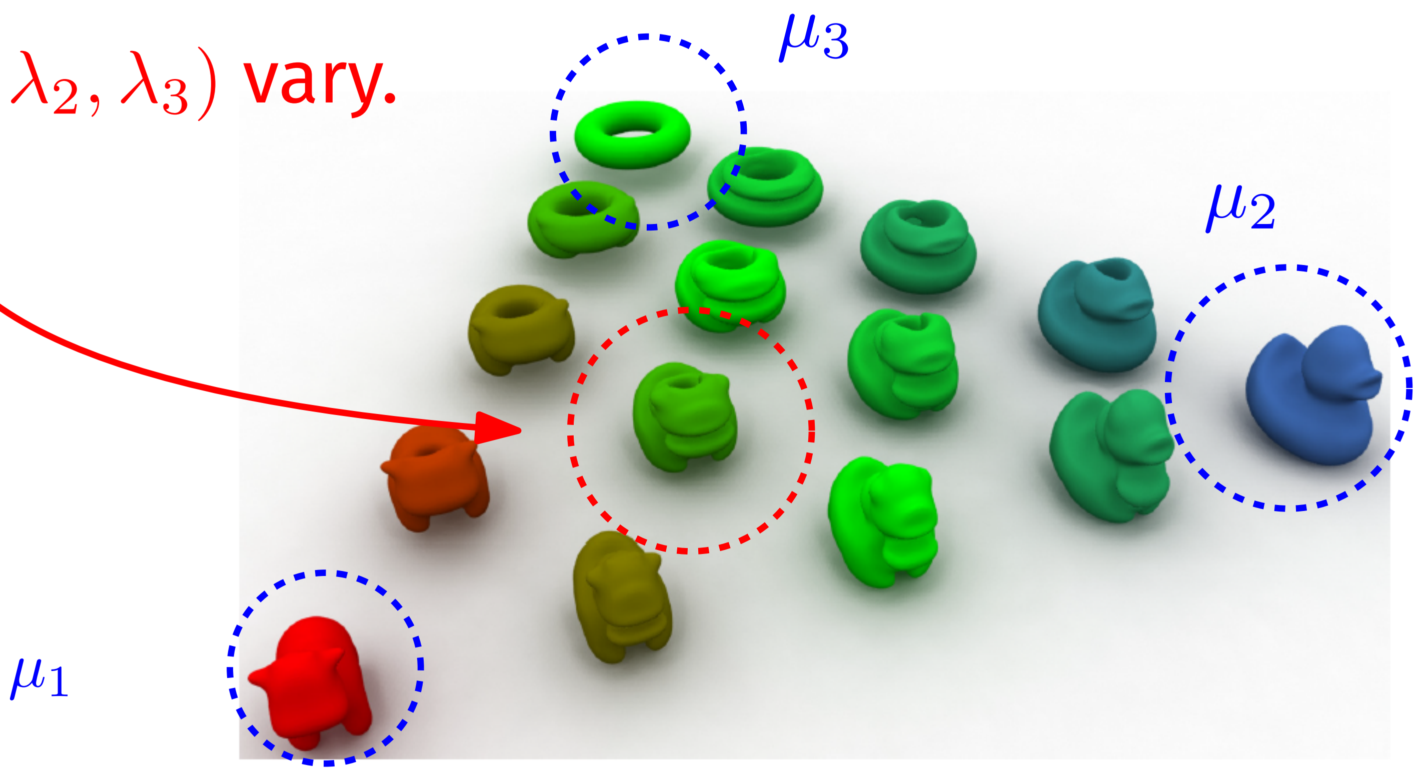
[Solomon et al (2015). Convolutional Wasserstein distances]

Existence and uniqueness

Theorem. There always exists a barycenter. It is unique if at least one μ_i has a density with respect to Lebesgue measure.

$$\bar{\mu} \in \arg\min_{\mu} \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

$\bar{\mu}$ as $(\lambda_1, \lambda_2, \lambda_3)$ vary.



Warning: lots of postprocessing to make it look good. In general **hard** to compute barycenters.

[Solomon et al (2015). Convolutional Wasserstein distances]

Numerics of Wasserstein barycenters

Inputs: $\mu_1, \ldots, \mu_n$ and $\lambda_1, \ldots, \lambda_n$.
Goal. Minimize

$$\mu \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

$\rightsquigarrow$ **Optimization** in the space of probability distributions.

Numerics of Wasserstein barycenters

Inputs: $\mu_1, \ldots, \mu_n$ and $\lambda_1, \ldots, \lambda_n$.
Goal. Minimize

$$\mu \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

If $\mu_1, \ldots, \mu_n$ all discrete supported on m atoms, Linear Programming formulation but a priori with m^n unknowns.

$\rightsquigarrow$ at least theoretically, possible to find a barycenter in polynomial time in n, m (but not in dimension) .

$\rightsquigarrow$ **Optimization** in the space of probability distributions.

[Altschuler & Boix-Adsera (2020). Wasserstein barycenters can be computed in polynomial time in fixed dimension]

Numerics of Wasserstein barycenters

Inputs: $\mu_1, \ldots, \mu_n$ and $\lambda_1, \ldots, \lambda_n$.
Goal. Minimize

$$\mu \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

If $\mu_1, \ldots, \mu_n$ all discrete supported on m atoms, Linear Programming formulation but a priori with m^n unknowns.

$\rightsquigarrow$ at least theoretically, possible to find a barycenter in polynomial time in n, m (but not in dimension) .

$\rightsquigarrow$ **Optimization** in the space of probability distributions.

Adding **Entropy**: doubly entropic barycenter. Minimize in μ

$$\frac{1}{2} \sum_{i=1}^{n} \lambda_i \mathcal{T}_{c,\varepsilon}(\mu, \mu_i) + \frac{\varepsilon}{2} \mathrm{KL}(\mu, \mu_{\text{ref}}).$$

$\rightsquigarrow$ bias $\mathcal{O}(\varepsilon^2)$, polynomial convergence of Sinkhorn like algorithm.

[Chizat (2023). Doubly regularized entropic Wasserstein barycenters]

Numerics of Wasserstein barycenters

Inputs: $\mu_1, \ldots, \mu_n$ and $\lambda_1, \ldots, \lambda_n$.
Goal. Minimize

$$\mu \mapsto \frac{1}{2} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

If $\mu_1, \ldots, \mu_n$ all discrete supported on m atoms, Linear Programming formulation but a priori with m^n unknowns.

$\rightsquigarrow$ at least theoretically, possible to find a barycenter in polynomial time in n, m (but not in dimension) .

$\rightsquigarrow$ **Optimization** in the space of probability distributions.

Adding **Entropy**: doubly entropic barycenter. Minimize in μ

$$\frac{1}{2} \sum_{i=1}^{n} \lambda_i \mathcal{T}_{c,\varepsilon}(\mu, \mu_i) + \frac{\varepsilon}{2} \mathrm{KL}(\mu, \mu_{\text{ref}}).$$

$\rightsquigarrow$ bias $\mathcal{O}(\varepsilon^2)$, polynomial convergence of Sinkhorn like algorithm.

Also: **fixed point approach**, gradient descents and variations, etc.

[Alvarez Esteban et al (2016). A fixed-point approach to barycenters in Wasserstein space]

Population barycenter

Why restrict to barycenter of a finite number of probabilities?

[Le Gouic & Loubes (2017). Existence and consistency of Wasserstein barycenters]

Population barycenter

Why restrict to barycenter of a finite number of probabilities?

$\rightsquigarrow$ Take $\tilde{p}$ a random probability distribution!

(Before: $\mathbb{P}(\tilde{p} = \mu_i) = \lambda_i$, so $\mathrm{Law}(\tilde{p}) = \sum_i \lambda_i \delta_{\mu_i}$)

Definition. Let $\tilde{p}$ a random probability distribution such that:

$$\mathbb{E}\left(\int \|x\|^2 \, \mathrm{d}\tilde{p}(x)\right) < +\infty.$$

We define a barycenter as any minimizer of

$$\mu \mapsto \frac{1}{2}\mathbb{E}\left(W_2^2(\mu, \tilde{p})\right).$$

[Le Gouic & Loubes (2017). Existence and consistency of Wasserstein barycenters]

Population barycenter

Why restrict to barycenter of a finite number of probabilities?
$\rightsquigarrow$ Take $\tilde{p}$ a random probability distribution!

(Before: $\mathbb{P}(\tilde{p} = \mu_i) = \lambda_i$, so $\mathrm{Law}(\tilde{p}) = \sum_i \lambda_i \delta_{\mu_i}$)

Definition. Let $\tilde{p}$ a random probability distribution such that:

$$\mathbb{E}\left(\int \|x\|^2 \,\mathrm{d}\tilde{p}(x)\right) < +\infty.$$

We define a barycenter as any minimizer of

$$\mu \mapsto \frac{1}{2}\mathbb{E}\left(W_2^2(\mu, \tilde{p})\right).$$

Theorem. Existence guaranteed, uniqueness if $\tilde{p}$ has a density with respect to Lebesgue with positive probability.

[Le Gouic & Loubes (2017). Existence and consistency of Wasserstein barycenters]

Population barycenter

Why restrict to barycenter of a finite number of probabilities?
$\rightsquigarrow$ Take $\tilde{p}$ a random probability distribution!

(Before: $\mathbb{P}(\tilde{p} = \mu_i) = \lambda_i$, so $\mathrm{Law}(\tilde{p}) = \sum_i \lambda_i \delta_{\mu_i}$)

Definition. Let $\tilde{p}$ a random probability distribution such that:

$$\mathbb{E}\left(\int \|x\|^2 \,\mathrm{d}\tilde{p}(x)\right) < +\infty.$$

We define a barycenter as any minimizer of

$$\mu \mapsto \frac{1}{2}\mathbb{E}\left(W_2^2(\mu, \tilde{p})\right).$$

Theorem. Existence guaranteed, uniqueness if $\tilde{p}$ has a density with respect to Lebesgue with positive probability.

Statistical question. Take $p_1, \ldots, p_n \overset{\text{i.i.d.}}{\sim} \tilde{p}$ and $\bar{\mu}_n$ their barycenter. Does it converge to $\bar{\mu}$ the population barycenter?

[Le Gouic & Loubes (2017). Existence and consistency of Wasserstein barycenters]

Population barycenter

Why restrict to barycenter of a finite number of probabilities?

$\rightsquigarrow$ Take $\tilde{p}$ a random probability distribution!

(Before: $\mathbb{P}(\tilde{p} = \mu_i) = \lambda_i$, so $\mathrm{Law}(\tilde{p}) = \sum_i \lambda_i \delta_{\mu_i}$)

Definition. Let $\tilde{p}$ a random probability distribution such that:

$$\mathbb{E}\left(\int \|x\|^2 \,\mathrm{d}\tilde{p}(x)\right) < +\infty.$$

We define a barycenter as any minimizer of

$$\mu \mapsto \frac{1}{2}\mathbb{E}\left(W_2^2(\mu, \tilde{p})\right).$$

Theorem. Existence guaranteed, uniqueness if $\tilde{p}$ has a density with respect to Lebesgue with positive probability.

Statistical question. Take $p_1, \ldots, p_n \overset{\text{i.i.d.}}{\sim} \tilde{p}$ and $\bar{\mu}_n$ their barycenter. Does it converge to $\bar{\mu}$ the population barycenter?

✓ Law of large numbers

∼ Central limit theorem: under strong assumptions.

[Le Gouic & Loubes (2017). Existence and consistency of Wasserstein barycenters]

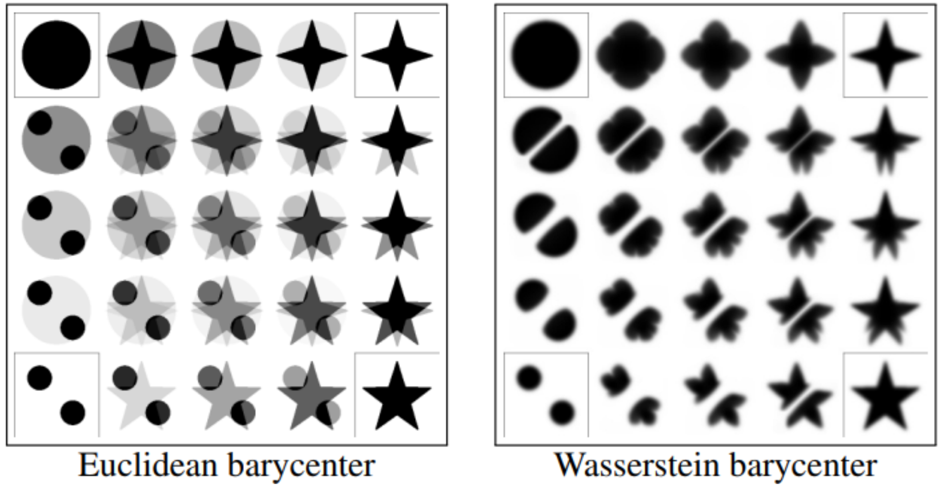
Comparision with linear barycenter

$\tilde{p}$ random probability distribution, two choices:

$$\bar{\mu}^W, \qquad \bar{\mu}^L = \mathbb{E}(\tilde{p})$$

Wasserstein barycenter

$\bar{\mu}^L(A) = \mathbb{E}(\tilde{p}(A))$ “linear” barycenter



[Solomon et al (2015). Convolutional Wasserstein distances]

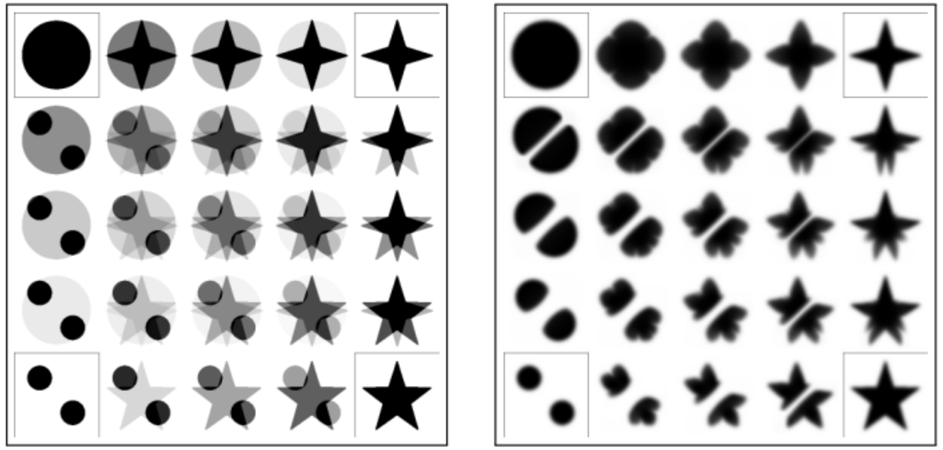
Comparision with linear barycenter

$\tilde{p}$ random probability distribution, two choices:

$$\bar{\mu}^W, \qquad \bar{\mu}^L = \mathbb{E}(\tilde{p})$$

Wasserstein barycenter

$\bar{\mu}^L(A) = \mathbb{E}(\tilde{p}(A))$ "linear" barycenter



Euclidean barycenter   Wasserstein barycenter

[Solomon et al (2015). Convolutional Wasserstein distances]

Proposition. We have: $\mathbb{E}_{\bar{\mu}^W}(X) = \mathbb{E}_{\bar{\mu}^L}(X),$

And with Var the variance

$$\mathrm{Var}(\bar{\mu}^W) \leq \mathbb{E}[\mathrm{Var}(\tilde{p})] \leq \mathrm{Var}(\bar{\mu}^L).$$

(Actually $\bar{\mu}^W \leq \bar{\mu}^L$ for the convex order)

Thus same mean but the Wasserstein barycenter is more concentrated.

Back to the model selection problem

$$\theta \sim \pi$$

$$X_1, \ldots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

Posterior

$$\theta | X_1, \ldots, X_n$$

Definition. The Bayesian Wasserstein Barycenter is:

$$\bar{\mu} \in \arg\min_{\mu} \mathbb{E}(W_2^2(\mu, p_\theta) | X_1, \ldots, X_n).$$

Wasserstein barycenter of $\tilde{p}$ with $\tilde{p} = p_\theta$ and θ follows posterior distribution.

Back to the model selection problem

$$\theta \sim \pi$$
$$X_1, \ldots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$
$$\theta | X_1, \ldots, X_n$$

Posterior

Definition. The Bayesian Wasserstein Barycenter is:

$$\bar{\mu} \in \arg\min_{\mu} \mathbb{E}(W_2^2(\mu, p_\theta) | X_1, \ldots, X_n).$$

Wasserstein barycenter of $\tilde{p}$ with $\tilde{p} = p_\theta$ and θ follows posterior distribution.

✓ • Does not depend on parametrization $\theta \mapsto p_\theta$, can be defined for nonparametric models.

✓ • For data in 1d, corresponds to linear average of quantile functions.

✓ • Smaller variance than the Bayesian model average.

Back to the model selection problem

$$\theta \sim \pi$$

$$X_1, \dots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

Posterior

$$\theta | X_1, \dots, X_n$$

Definition. The Bayesian Wasserstein Barycenter is:

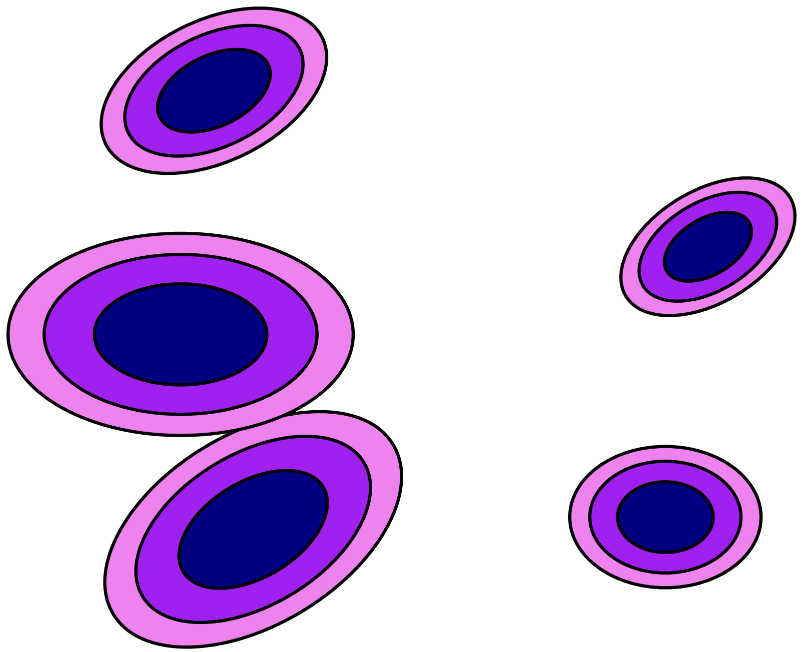
$$\bar{\mu} \in \arg\min_{\mu} \mathbb{E}(W_2^2(\mu, p_\theta) | X_1, \dots, X_n).$$

Wasserstein barycenter of $\tilde{p}$ with $\tilde{p} = p_\theta$ and θ follows posterior distribution.

- Does not depend on parametrization $\theta \mapsto p_\theta$, can be defined for nonparametric models.
- For data in 1d, corresponds to linear average of quantile functions.
- Smaller variance than the Bayesian model average.
- If $X_1, \dots X_n \dots \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$, possible to study consistency $\bar{\mu} \to p_{\theta_0}$ (but more complicated because of second moments!).
- Numerics: more complicated!

What about Gaussians?

Assume $\tilde{p} = \mathcal{N}(\tilde{m}, \tilde{\Sigma})$ with random mean $\tilde{m}$ and random covariance $\tilde{\Sigma}$ invertible with positive probability.



What about Gaussians?

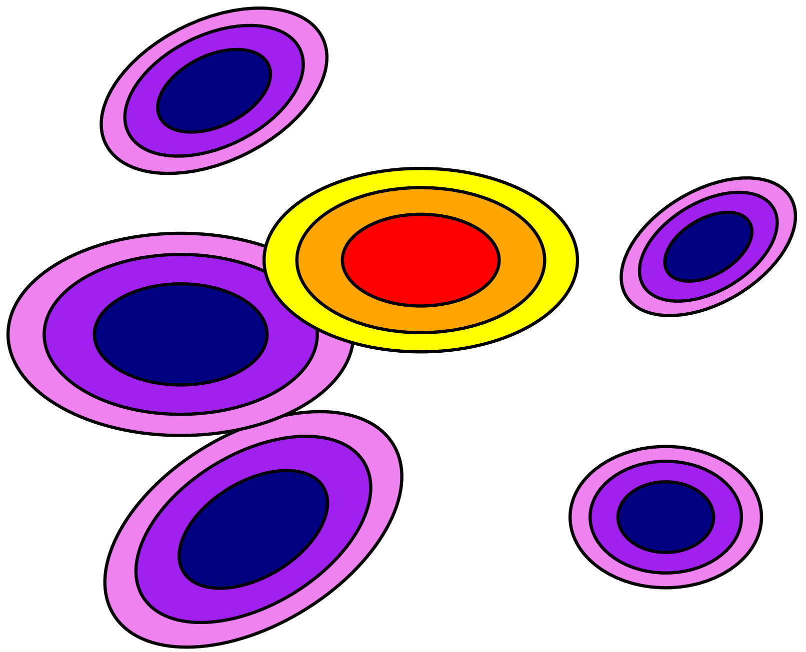
Assume $\tilde{p} = \mathcal{N}(\tilde{m}, \tilde{\Sigma})$ with random mean $\tilde{m}$ and random covariance $\tilde{\Sigma}$ invertible with positive probability.

Theorem. The barycenter is Gaussian, with

$$\bar{m} = \mathbb{E}(\tilde{m})$$

and $\bar{\Sigma}$ unique solution to:

$$\bar{\Sigma} = \mathbb{E}\left(\left(\bar{\Sigma}^{1/2}\tilde{\Sigma}\bar{\Sigma}^{1/2}\right)^{1/2}\right).$$



Can be computed by a fixed point iteration in space $d \times d$ S.D.P. matrices.

What about Gaussians?

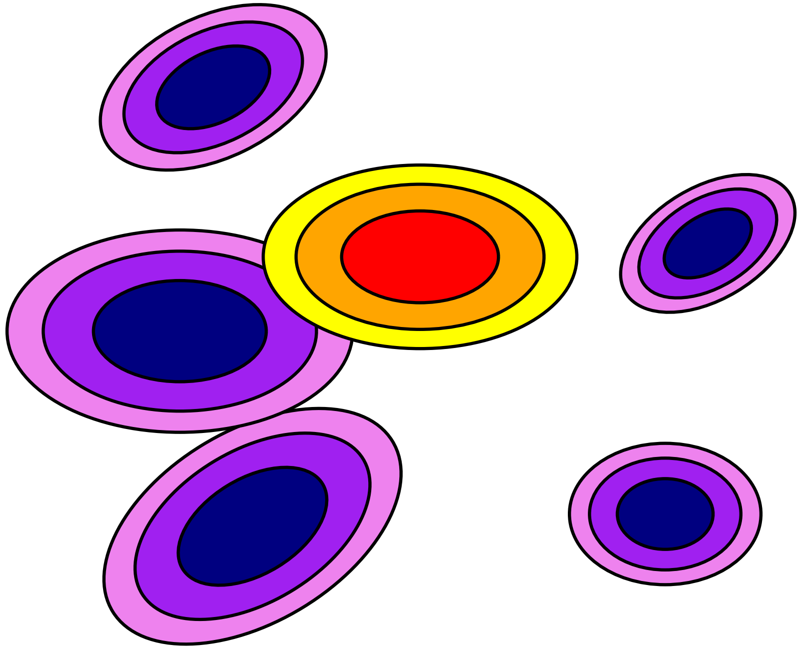
Assume $\tilde{p} = \mathcal{N}(\tilde{m}, \tilde{\Sigma})$ with random mean $\tilde{m}$ and random covariance $\tilde{\Sigma}$ invertible with positive probability.

Theorem. The barycenter is Gaussian, with

$$\bar{m} = \mathbb{E}(\tilde{m})$$

and $\bar{\Sigma}$ unique solution to:

$$\bar{\Sigma} = \mathbb{E}\left(\left(\bar{\Sigma}^{1/2}\tilde{\Sigma}\bar{\Sigma}^{1/2}\right)^{1/2}\right).$$



(Clear that Wasserstein barycenter more concentrated than linear barycenter)

Remark. If $\tilde{\Sigma}$ deterministic, then the barycenter has the same covariance.

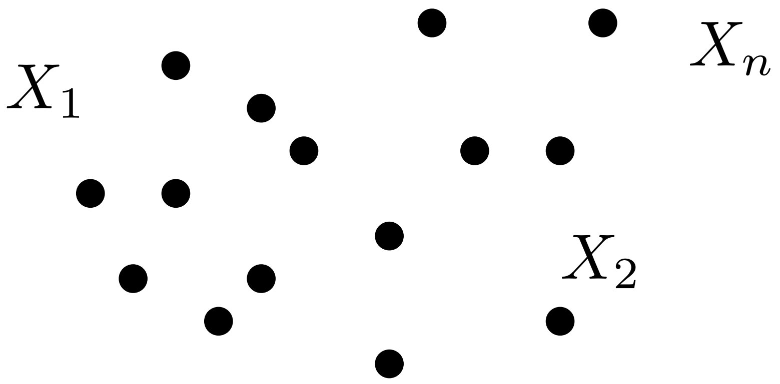
Lemma. If $AI \leq \tilde{\Sigma} \leq BI$ a.s. for $A, B \in \mathbb{R}$ then $AI \leq \bar{\Sigma} \leq BI$

Other use of Wasserstein barycenters: for scalable Bayes

$$\theta \sim \pi$$

$$X_1, \dots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

$$\theta | X_1, \dots, X_n$$ Posterior π_n^*



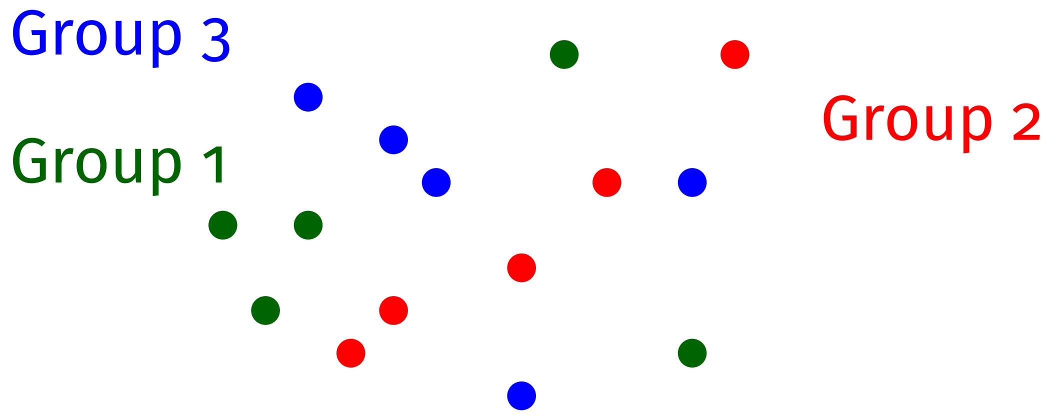
Other use of Wasserstein barycenters: for scalable Bayes

$$\theta \sim \pi$$

$$X_1, \ldots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

$$\theta | X_1, \ldots, X_n$$

Posterior π_n^*



1. Partition the data in k groups of m elements, $n = km$.

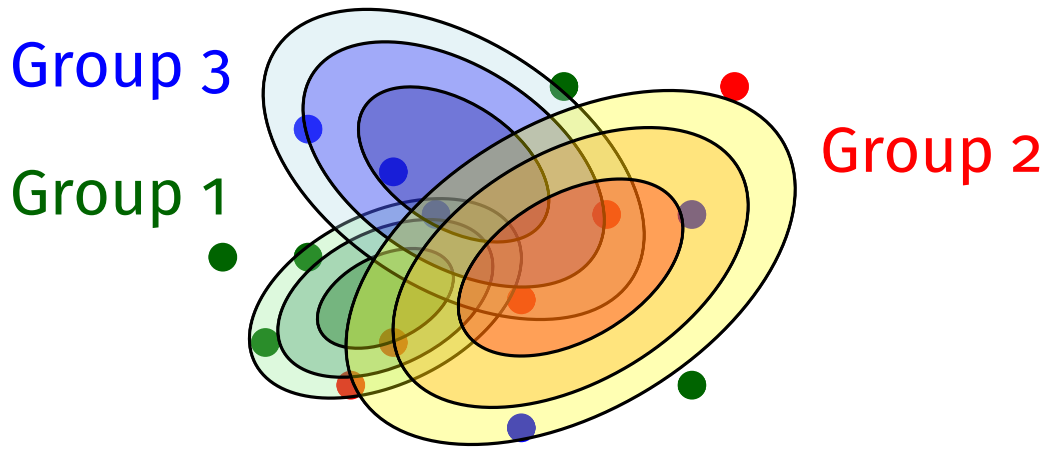
Other use of Wasserstein barycenters: for scalable Bayes

$$\theta \sim \pi$$

$$X_1, \dots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

$$\theta | X_1, \dots, X_n$$

Posterior π_n^*



1. Partition the data in k groups of m elements, $n = km$.
2. For each group j, compute $\pi_{n,j}^*$ posterior with only data of group j

(likelihood rescaled to do as if there were n data and not m)

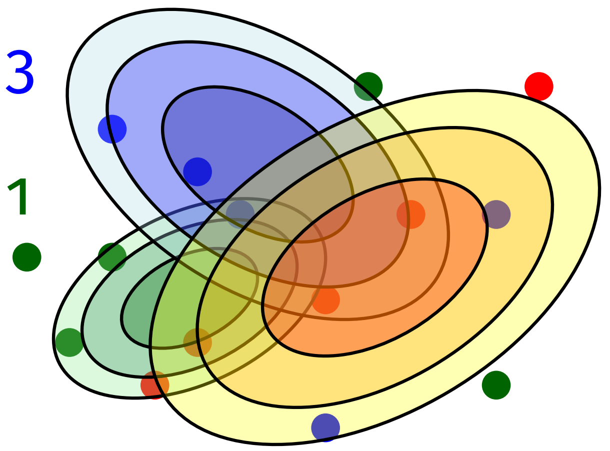
Other use of Wasserstein barycenters: for scalable Bayes

$$\theta \sim \pi$$

$$X_1, \ldots, X_n | \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$

$$\theta | X_1, \ldots, X_n$$

Posterior π_n^*



1. Partition the data in k groups of m elements, $n = km$.
2. For each group j, compute $\pi_{n,j}^*$ posterior with only data of group j

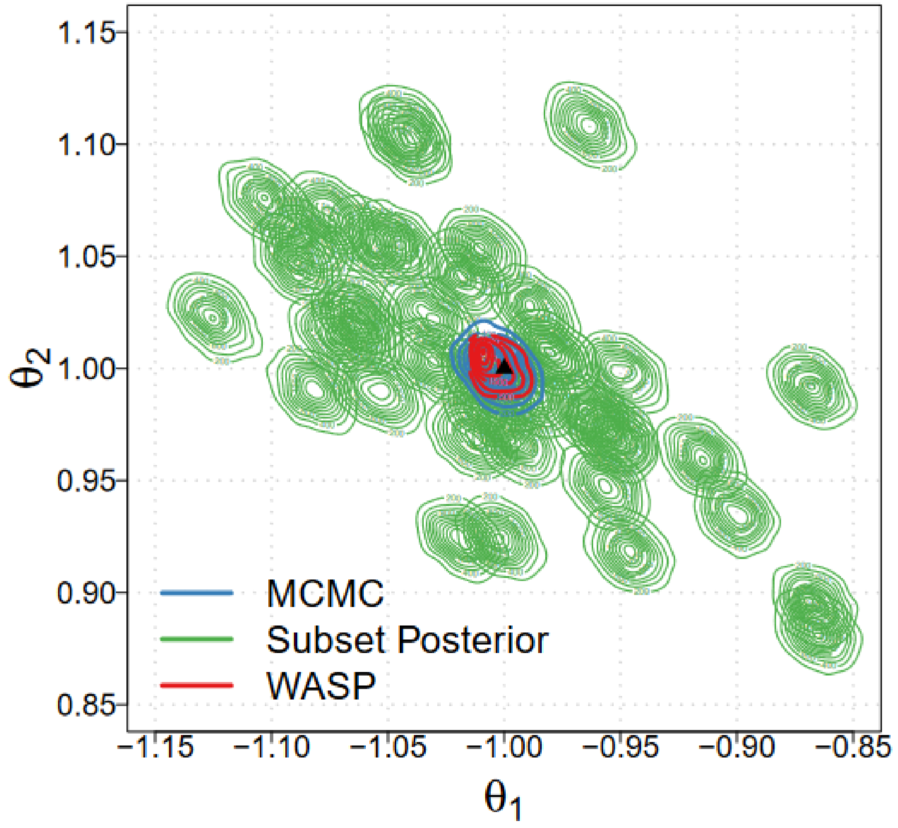
   (likelihood rescaled to do as if there were n data and not m)

3. Compute the **Wasserstein posterior** as the Wasserstein barycenter of $\pi_{n,j}^*$ for $j = 1, \ldots, k$.

Why? More tractable, enable distributed computations.

Findings

Logistic regression

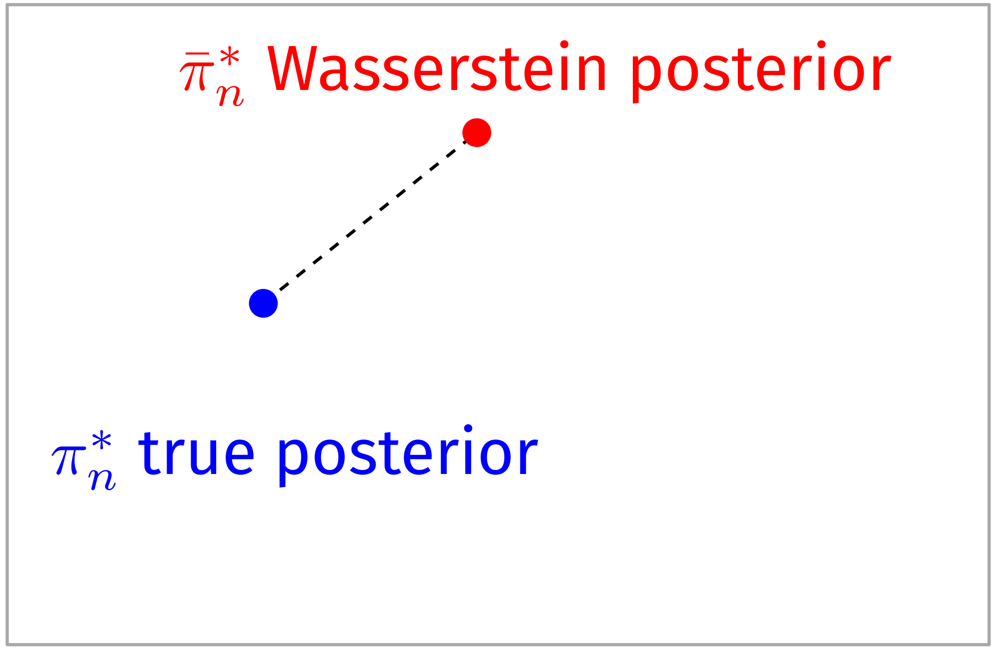


Recall: $\mathrm{Var}(\bar{\mu}^W) \le \mathbb{E}[\mathrm{Var}(\tilde{p})]$

Good: we want variance $\to 0$ as $n \to +\infty$.

Findings

Space of posteriors



Recall: $\mathrm{Var}(\bar{\mu}^W) \leq \mathbb{E}[\mathrm{Var}(\tilde{p})]$

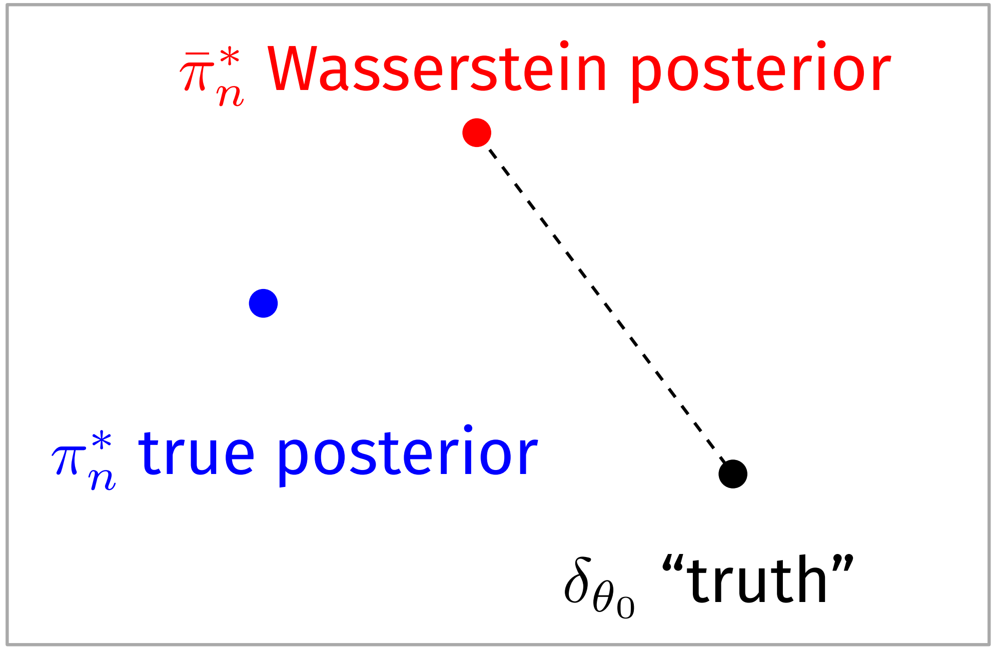
Good: we want variance $\to 0$ as $n \to +\infty$.

Ideal result: $n^{1/2} W_2(\bar{\pi}_n^*, \pi_n^*) \to 0$ as $m \to +\infty$.

Proved in 1d, iid data and regular parametric model.

Findings

Space of posteriors



Recall: $\mathrm{Var}(\bar{\mu}^W) \leq \mathbb{E}[\mathrm{Var}(\tilde{p})]$

Good: we want variance $\to 0$ as $n \to +\infty$.

Ideal result: $n^{1/2} W_2(\bar{\pi}_n^*, \pi_n^*) \to 0$ as $m \to +\infty$.

Proved in 1d, iid data and regular parametric model.

Alternative: $W_2(\bar{\pi}_n^*, \delta_{\theta_0}) \to 0$ at parametric rate if data come from p_{θ_0}.

Proved in multi dimension up to log factor, inid setup, $k \sim \log n$.

1 - Wasserstein distances in Bayesian statistics

2 - Wasserstein barycenters for model selection and scalable Bayes

Honourable mentions
(topics I researched but did not include)

3 - Looking at sampling with the geometry of optimal transport

Some other topics about optimal transport and Bayesian statistics

Stability of posterior with respect to the data

How to estimate $W_2(\pi_*^1, \pi_*^2)$, where π_*^1, π_*^2 correspond to posterior distribution for the same prior but different data.

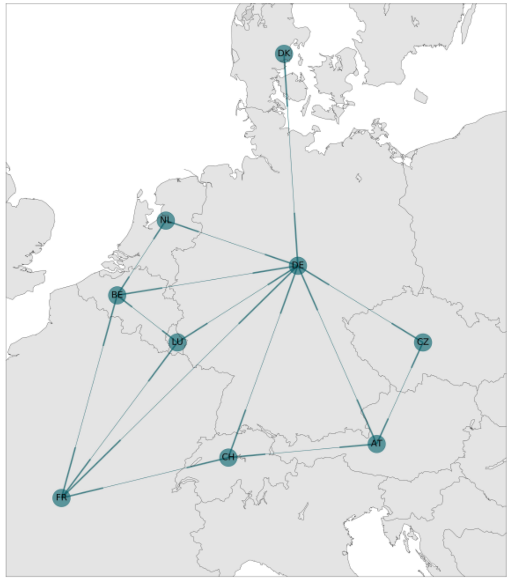
Also: Wasserstein ABC, construction couplings between Markov chains with OT, etc.

[Dolera & Mainini (2023). Lipschitz continuity of probability kernels in the OT framework]
[Dolera, Favaro & Mainini (2023). Strong posterior contraction rates via W dynamics]
[Camerlenghi et al (2022). Wasserstein posterior contraction rates in non-dominated Bayesian nonparametric models]

Some other topics about optimal transport and Bayesian statistics

Stability of posterior with respect to the data

How to estimate $W_2(\pi_*^1, \pi_*^2)$, where π_*^1, π_*^2 correspond to posterior distribution for the same prior but different data.



(**Example**: infer people's preference from the observation of migration data)

Also: Wasserstein ABC, construction couplings between Markov chains with OT, etc.

Bayesian statistics for inverse optimal transport

Given (an approximation of) an optimal coupling, how to find the inputs of the problem (marginals and **cost function** c)?

[Stuart & Wolfram (2020). Inverse optimal transport]

[Chi, Wang & Shafto (2022). Discrete Probabilistic Inverse Optimal Transport]

1 - Wasserstein distances in Bayesian statistics

2 - Wasserstein barycenters for model selection and scalable Bayes

Honourable mentions

(topics I researched but did not include)

3 - Looking at sampling with the geometry of optimal transport

[Ambrosio, Gigli & Savaré. Second Part]

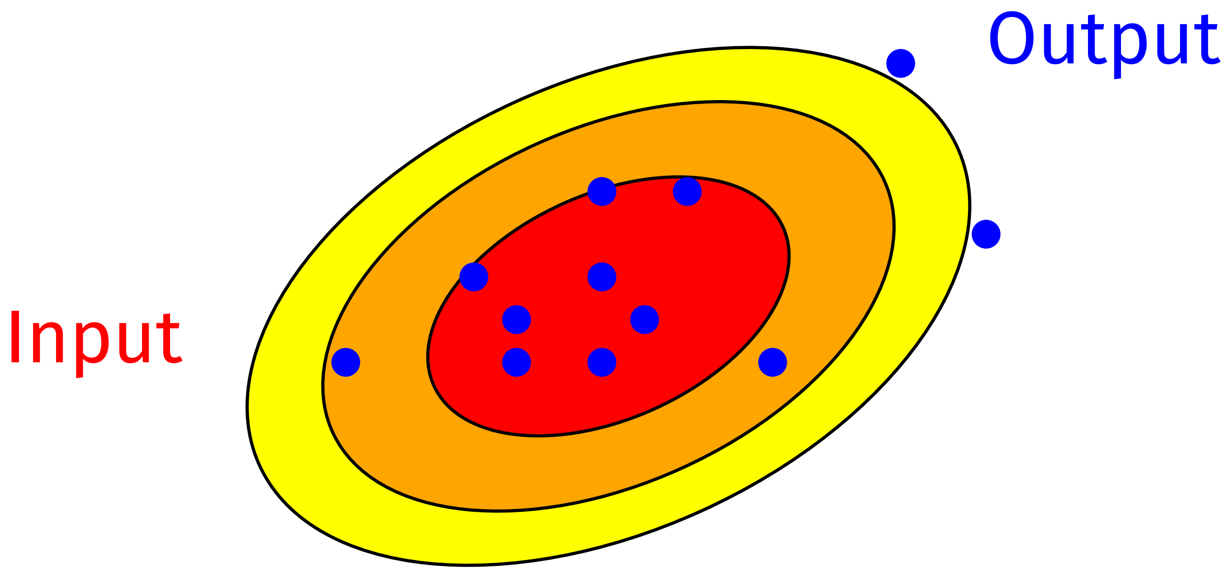
[Chewi (2024+). Log-concave sampling]

The sampling problem

Input. Target density π with probability density function proportional to $\exp(-V)$, with $V : \mathbb{X} \to \mathbb{R}$.

e.g. a posterior distribution!

Goal. Produce samples from π.



The sampling problem

Input. Target density π with probability density function proportional to $\exp(-V)$, with $V : \mathbb{X} \to \mathbb{R}$.

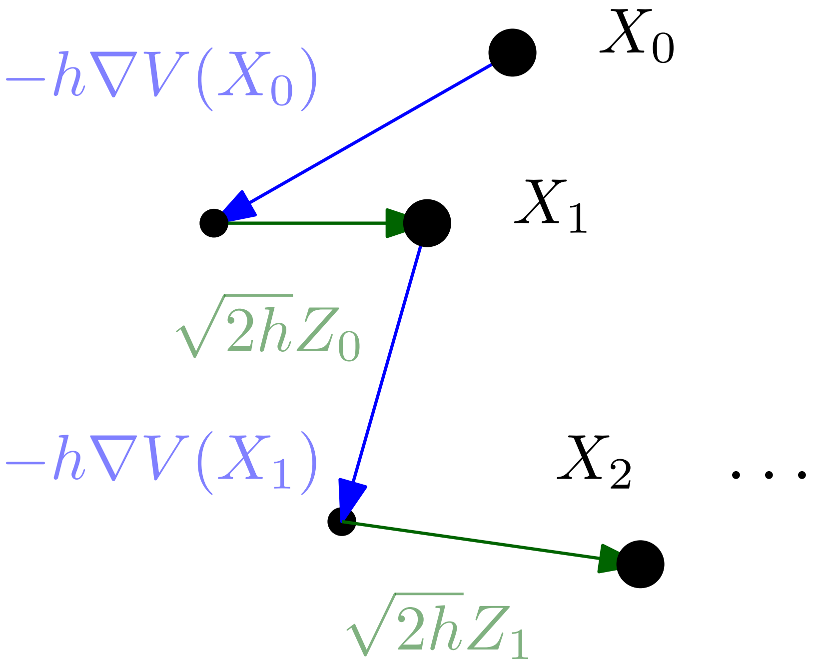
e.g. a posterior distribution!

Goal. Produce samples from π.

Langevin Monte Carlo.

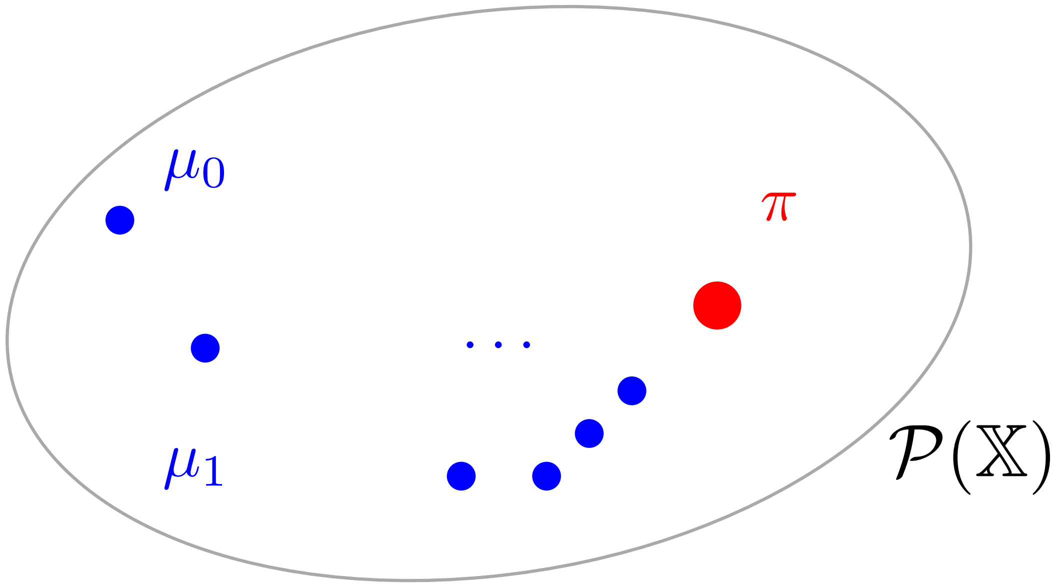
1. Initialize X_0, fix time step $h > 0$.
2. Iterate $X_{n+1} = X_n - h\nabla V(X_n) + \sqrt{2h} Z_n$

$$Z_1, \ldots, Z_n, \ldots, \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$$



Sampling as optimization

Only keep track of $\mu_n = \mathrm{Law}(X_n)$.

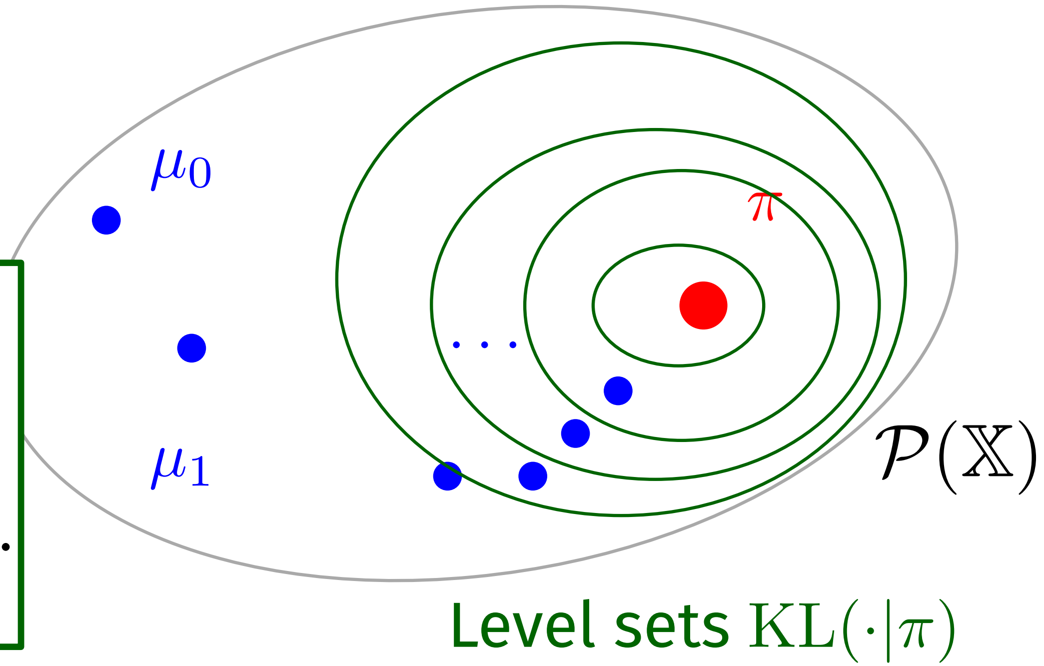


Sampling as optimization

Only keep track of $\mu_n = \mathrm{Law}(X_n)$.

Look at $\mathrm{KL}(\mu|\pi)$. It is ≥ 0 and $= 0$ iff $\mu = \pi$:

$$\mathrm{KL}(\mu|\pi) = \mathbb{E}_\mu \left[\log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right] = \int V \,\mathrm{d}\mu + \int \mu \log \mu.$$

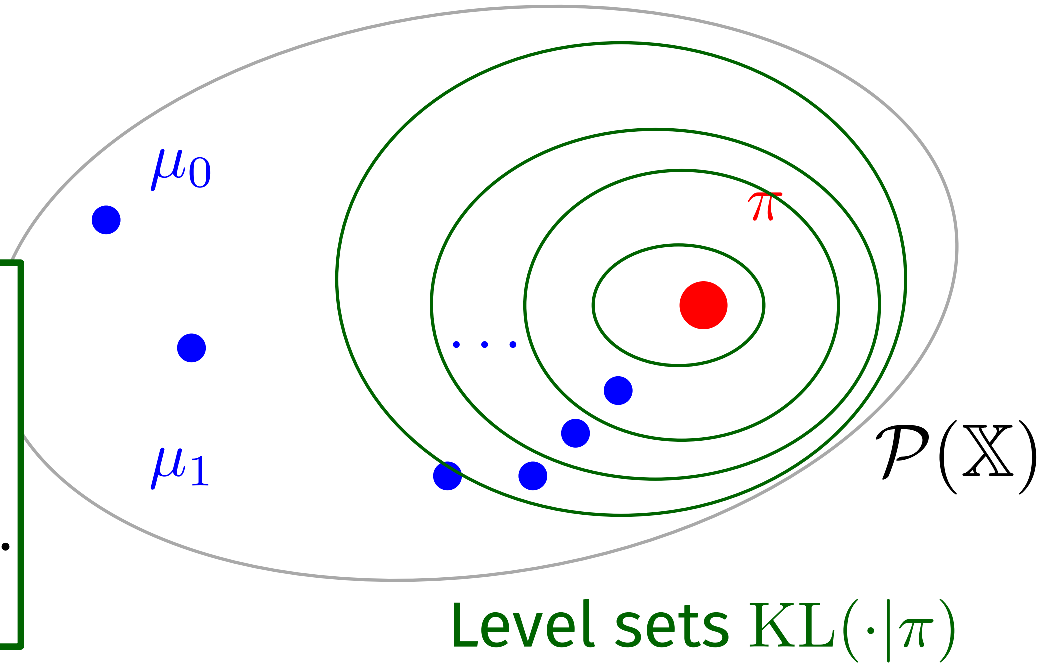


Sampling as optimization

Only keep track of $\mu_n = \mathrm{Law}(X_n)$.

Look at $\mathrm{KL}(\mu|\pi)$. It is ≥ 0 and $= 0$ iff $\mu = \pi$:

$$\mathrm{KL}(\mu|\pi) = \mathbb{E}_\mu \left[\log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right] = \int V \, \mathrm{d}\mu + \int \mu \log \mu.$$

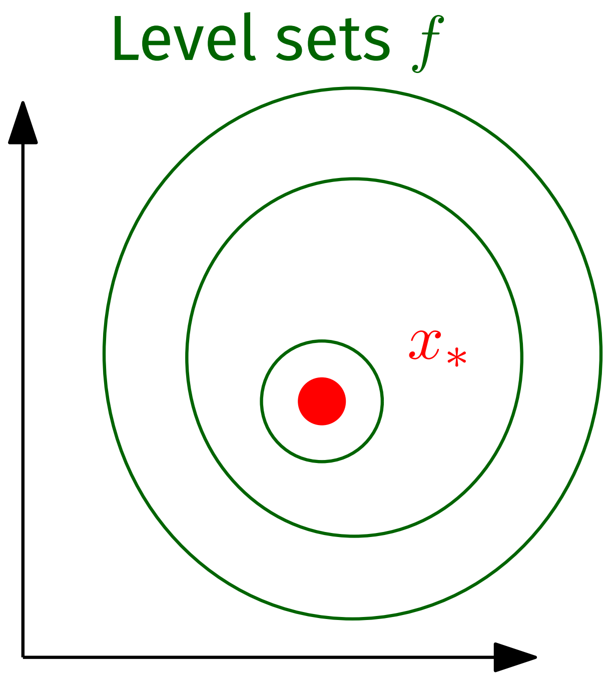


Sampling is **optimization** on $\mathcal{P}(\mathbb{X})$: we want to find the minimum (π) of a function ($\mathrm{KL}(\cdot|\pi)$) defined $\mathcal{P}(\mathbb{X})$.

To analyze this optimization task we use the **geometry** of optimal transport.

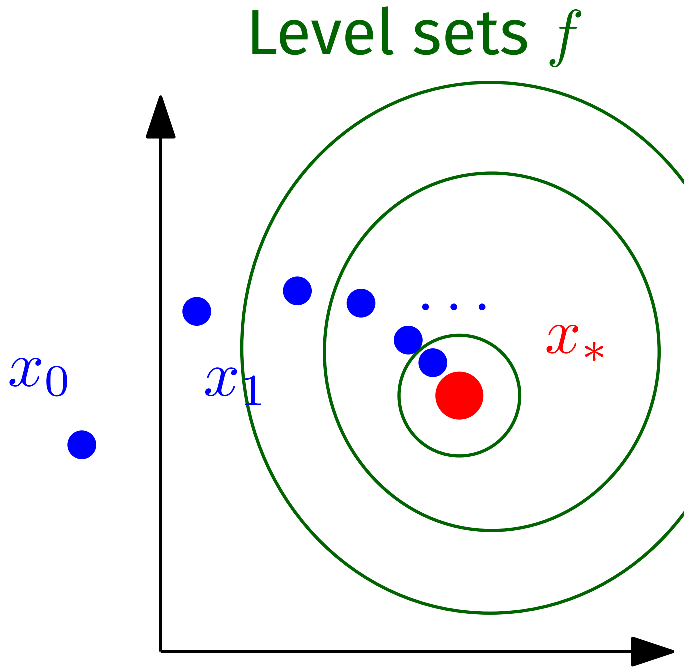
Optimization: gradient descent and variants

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$.
Goal: Find x_* the point of minimum of f



Optimization: gradient descent and variants

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$.
Goal: Find x_* the point of minimum of f



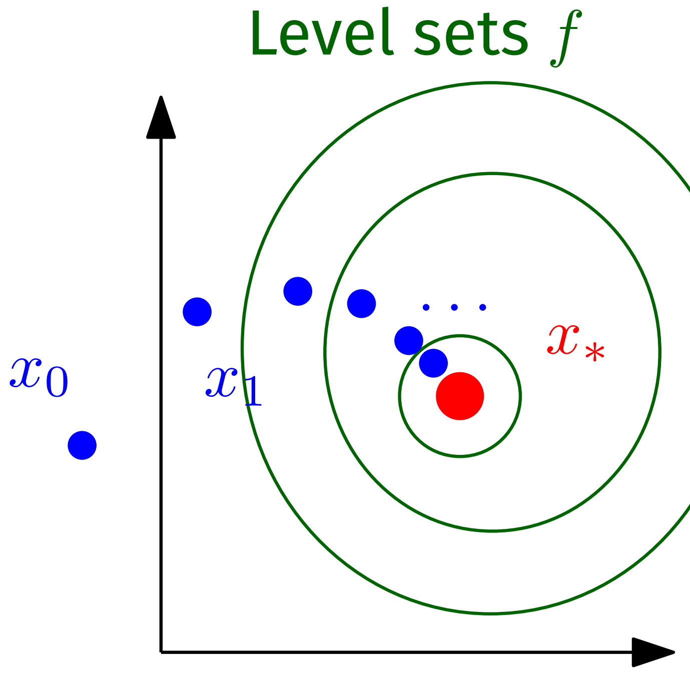
Given initial guess x_0 and stepsize h:

Gradient descent

$$x_{n+1} = x_n - h\nabla f(x_n)$$

Optimization: gradient descent and variants

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$.
Goal: Find x_* the point of minimum of f



Given initial guess x_0 and stepsize h:

Gradient descent

$$x_{n+1} = x_n - h\nabla f(x_n)$$

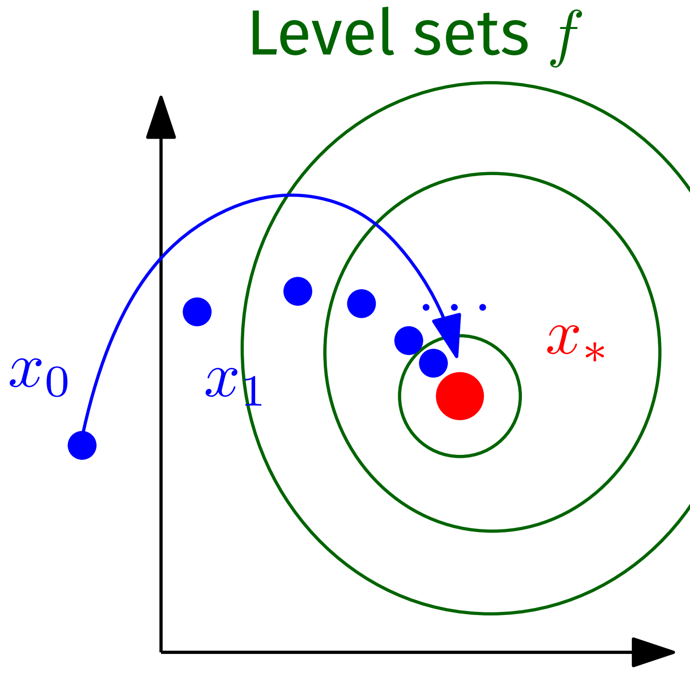
Proximal step (Implicit)

$$x_{n+1} = x_n - h\nabla f(x_{n+1})$$

$$\Leftrightarrow x_{n+1} \in \arg\min_x \left(f(x) + \frac{\|x - x_n\|^2}{2h} \right)$$

Optimization: gradient descent and variants

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$.
Goal: Find x_* the point of minimum of f



Given initial guess x_0 and stepsize h:

Gradient descent

$$x_{n+1} = x_n - h\nabla f(x_n)$$

Proximal step (Implicit)

$$x_{n+1} = x_n - h\nabla f(x_{n+1})$$

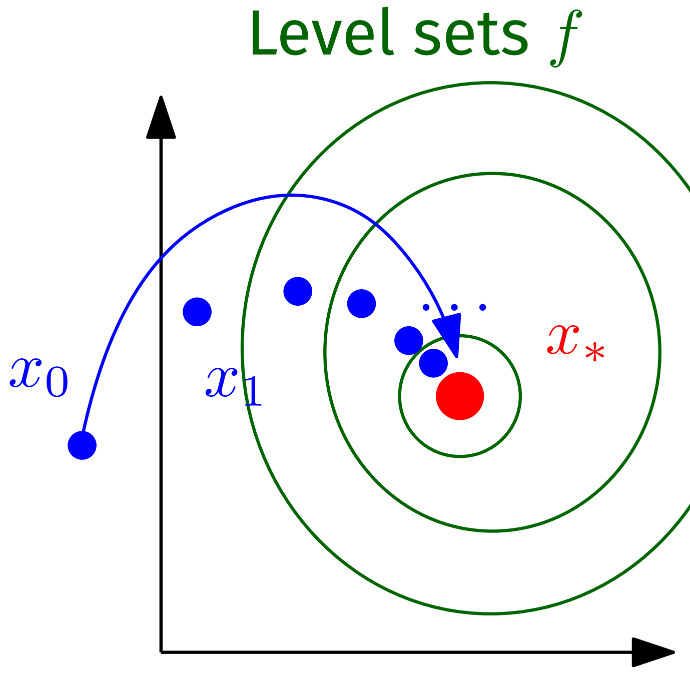
$$\Leftrightarrow x_{n+1} \in \arg\min_x \left(f(x) + \frac{\|x - x_n\|^2}{2h} \right)$$

Gradient flow $(h \to 0)$

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t).$$

Optimization: gradient descent and variants

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$.
Goal: Find x_* the point of minimum of f



Given initial guess x_0 and stepsize h:

Gradient descent

$$x_{n+1} = x_n - h\nabla f(x_n)$$

Proximal step (Implicit)

$$x_{n+1} = x_n - h\nabla f(x_{n+1})$$

$$\Leftrightarrow x_{n+1} \in \arg\min_x \left(f(x) + \frac{\|x - x_n\|^2}{2h} \right)$$

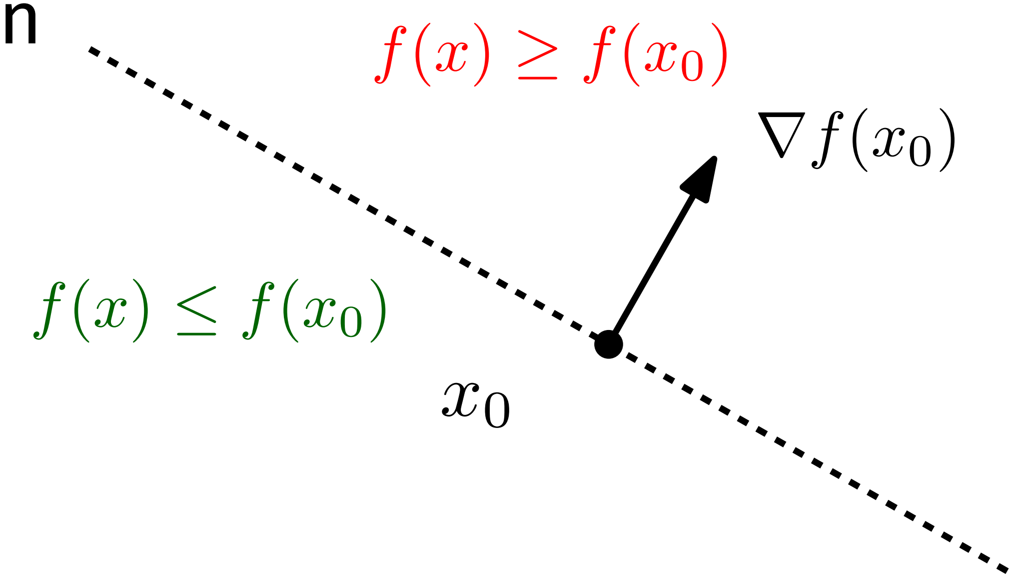
Gradient flow ($h \to 0$)

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t).$$

But what is the link with **geometry**?

Why choose the gradient?

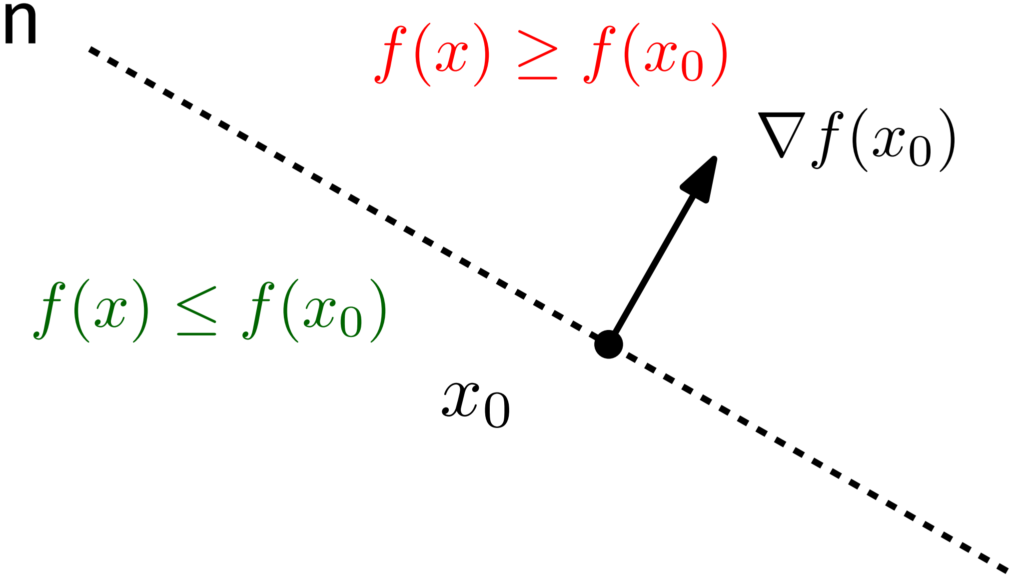
Why move in direction $-\nabla f(x_0)$? The function decreases in many other directions!



Why choose the gradient?

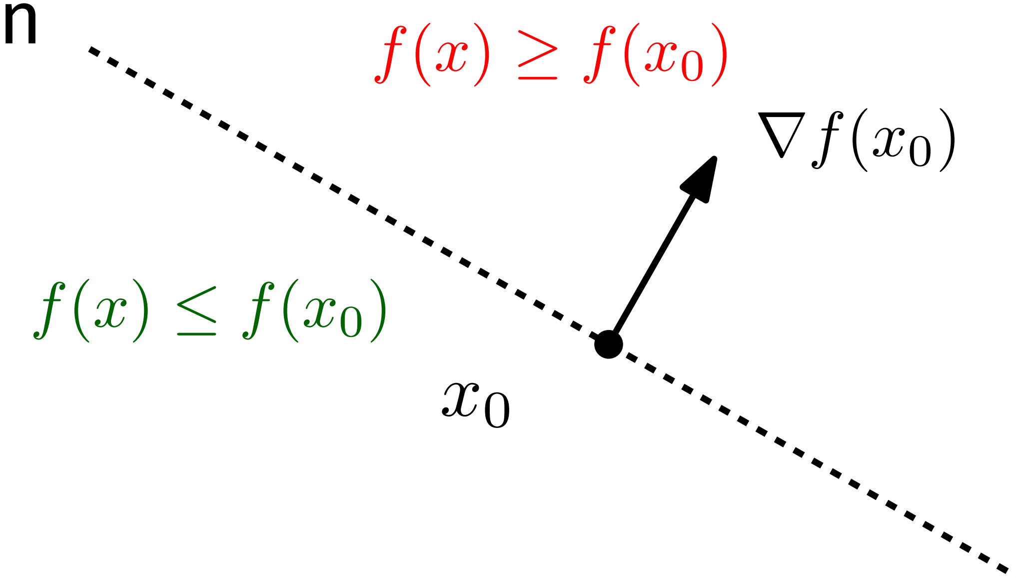
Why move in direction $-\nabla f(x_0)$? The function decreases in many other directions!

$-\nabla f(x_0)$ direction where f **decreases the most** (at first order) if distances are measured with Euclidean distance $\|\cdot\|$.



Why choose the gradient?

Why move in direction $-\nabla f(x_0)$? The function decreases in many other directions!



$-\nabla f(x_0)$ direction where f **decreases the most** (at first order) if distances are measured with Euclidean distance $\|\cdot\|$.

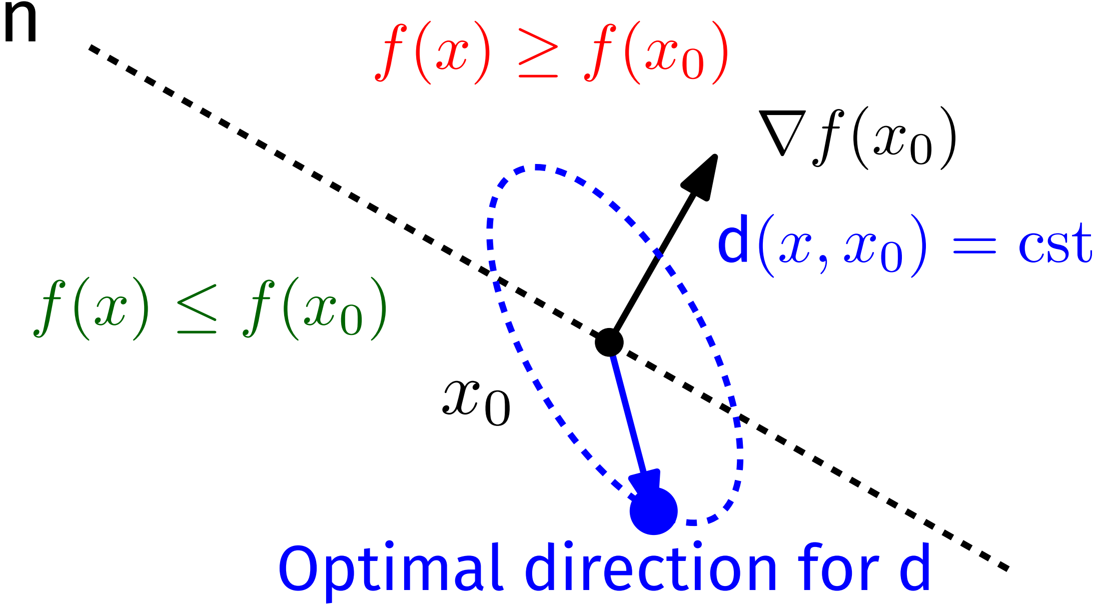
In general take d a distance on $\mathbb{R}^d$

(e.g. $\mathrm{d}^2(x, y) = (x-y)^\top Q(x-y)$, with Q p.d. matrix)

Choosing distance $\Leftrightarrow$ choosing geometry

Why choose the gradient?

Why move in direction $-\nabla f(x_0)$? The function decreases in many other directions!



$-\nabla f(x_0)$ direction where f **decreases the most** (at first order) if distances are measured with Euclidean distance $\|\cdot\|$.

In general take d a distance on $\mathbb{R}^d$

Choosing distance $\Leftrightarrow$ choosing geometry

(e.g. $\mathrm{d}^2(x, y) = (x-y)^\top Q(x-y)$, with Q p.d. matrix)

At first order $f(x) \simeq f(x_0) + \nabla f(x_0)^\top (x - x_0)$.

$\rightsquigarrow$ Choose x with $\nabla f(x_0)^\top (x - x_0)$ minimal under constraint $\mathrm{d}(x, x_0) = \mathrm{cst}$.

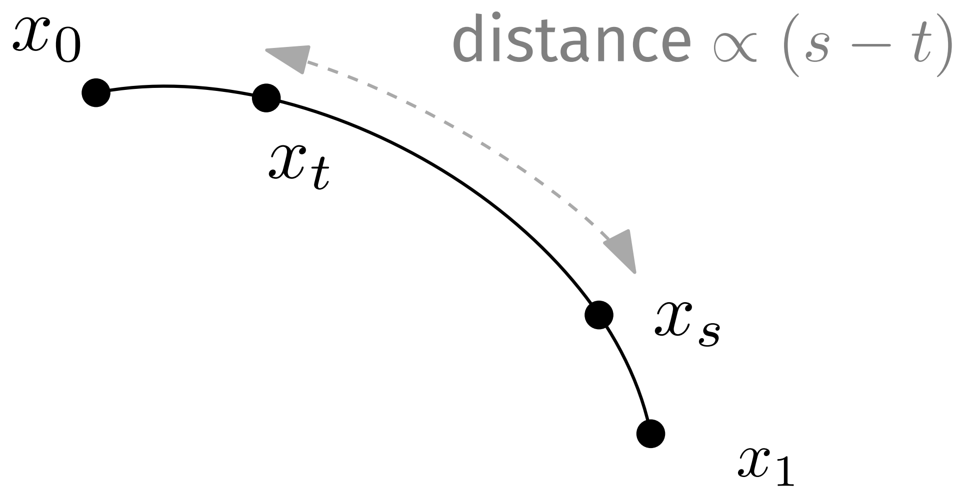
(e.g. $x - x_0 \propto -Q^{-1}\nabla f(x_0)$)

From straight lines to geodesics

Fix d distance on $\mathbb{R}^d$.

Definiton. We call $(x_t)_{t\in[0,1]}$ a geodesic if for any $0 \le t \le s \le 1$:

$$\mathsf{d}(x_t, x_s) = (s-t)\mathsf{d}(x_0, x_1).$$

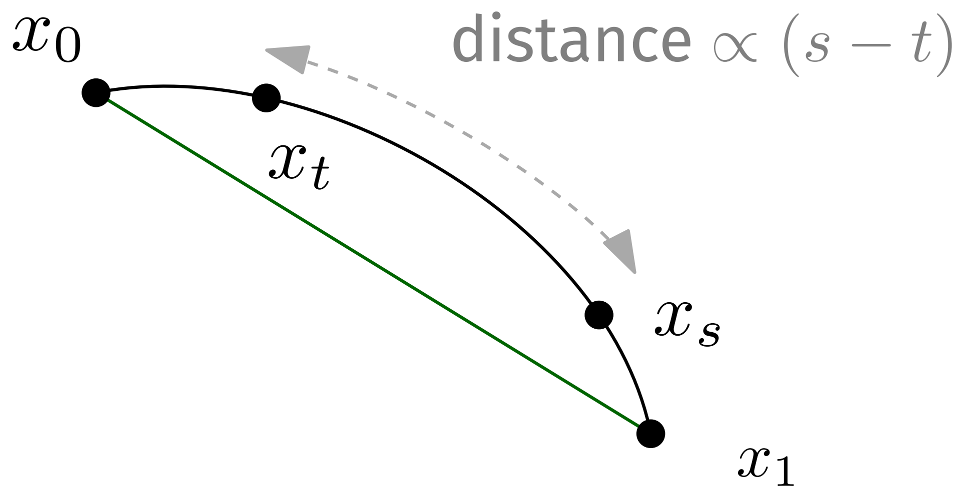


From straight lines to geodesics

Fix d distance on $\mathbb{R}^d$.

Definiton. We call $(x_t)_{t\in[0,1]}$ a geodesic if for any $0 \le t \le s \le 1$:

$$\mathsf{d}(x_t, x_s) = (s-t)\mathsf{d}(x_0, x_1).$$



Example. If $\mathsf{d}^2(x,y) = (y-x)^\top Q(y-x)$ (or is a norm) then $x_t = (1-t)x + ty$ is the unique geodesic joining x to y.

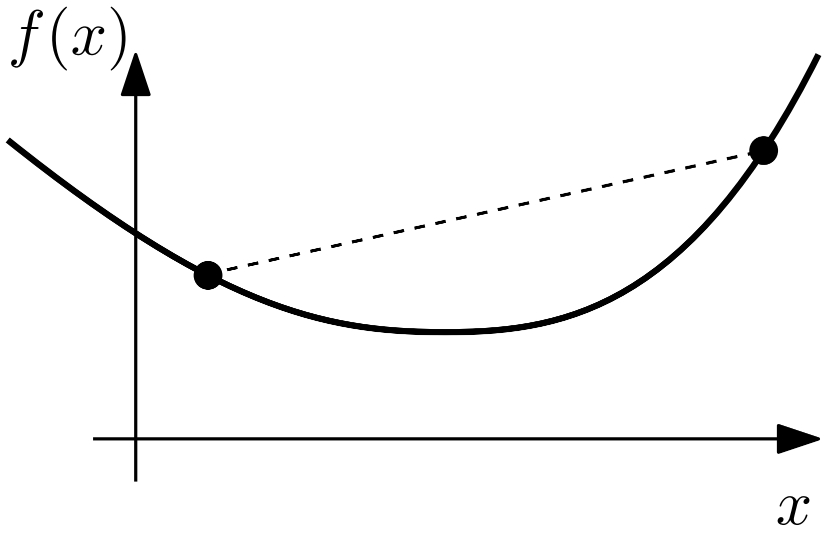
Convexity and smoothness

Recall f is convex if for all x, y, $t \in [0,1]$

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

Equivalent if f smooth:

$$D^2 f(x) \geq 0 \ \text{ for all } x.$$



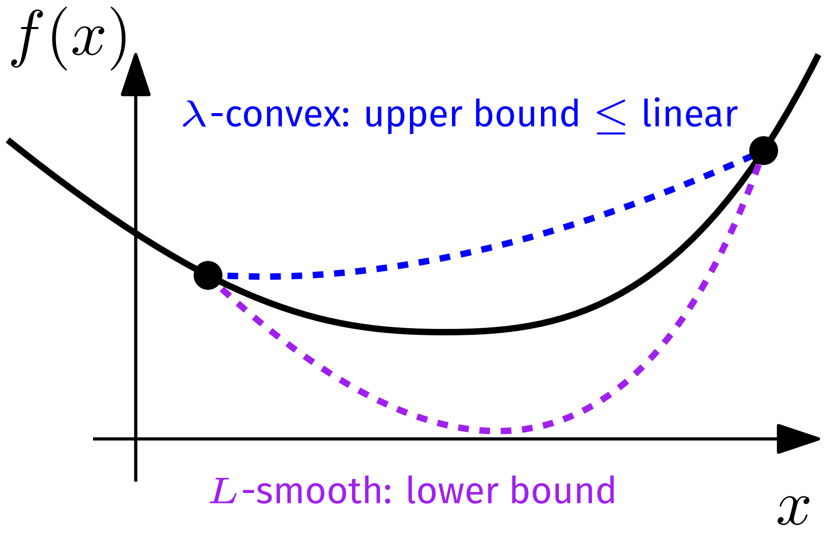
Convexity and smoothness

Recall f is convex if for all x, y, $t \in [0,1]$

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

Equivalent if f smooth:

$$D^2 f(x) \geq 0 \text{ for all } x.$$



For f C^2, it is λ convex and L smooth if $\lambda I \leq D^2 f \leq LI$ everywhere.

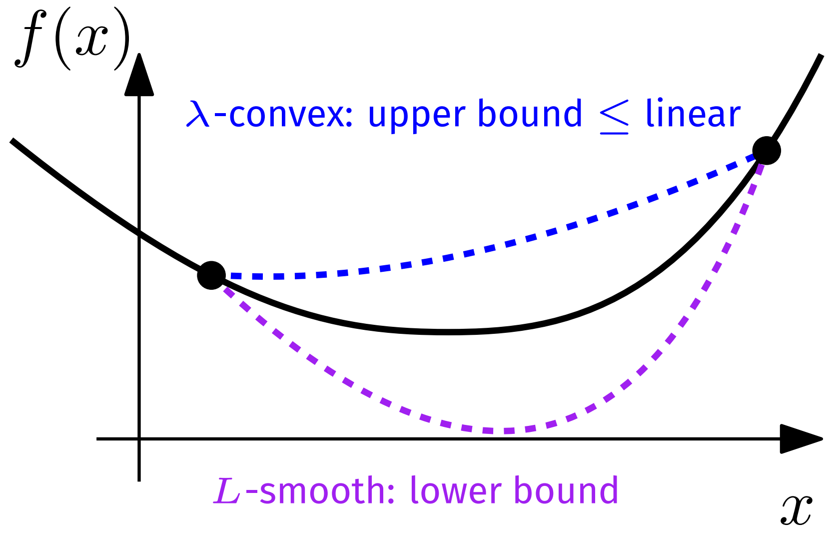
Convexity and smoothness

Recall f is convex if for all x, y, $t \in [0,1]$

$$f((1-t)x + ty) \le (1-t)f(x) + tf(y)$$

Equivalent if f smooth:

$$D^2 f(x) \ge 0 \text{ for all } x.$$



For f C^2, it is λ convex and L smooth if $\lambda I \le D^2 f \le LI$ everywhere.

For distance d, we say f is λ convex and L smooth if there exists a **geodesic** (x_t) joining x to y such that for any $t \in [0,1]$:

$$\frac{\lambda t(1-t)}{2}\mathsf{d}^2(x,y) \le (1-t)f(x) + tf(y) - f(x_t) \le \frac{Lt(1-t)}{2}\mathsf{d}^2(x,y)$$

(e.g. if $\mathsf{d}^2(x,y) = (x-y)^\top Q(x-y)$ means $\lambda Q \le D^2 f \le LQ$ everywhere.)

Gradient descent and variants in another geometry

On $\mathbb{R}^d$,

Riemannian distance: d distance and $\mathsf{d}^2(x,y) \simeq (y-x)^\top Q(x)(y-x)$ if $y \to x$.

Denote $\nabla_{\mathsf{d}} f(x) = Q(x)^{-1}\nabla f(x)$ gradient in geometry of d.
Assume f is $\lambda > 0$ convex and L smooth in geometry of d.

Gradient descent and variants in another geometry

On $\mathbb{R}^d$,

Riemannian distance: d distance and $\mathsf{d}^2(x,y) \simeq (y-x)^\top Q(x)(y-x)$ if $y \to x$.

Denote $\nabla_{\mathsf{d}} f(x) = Q(x)^{-1}\nabla f(x)$ gradient in geometry of d.
Assume f is $\lambda > 0$ convex and L smooth in geometry of d.

Gradient descent ($h \leq 1/L$)

$x_{n+1} = x_n - h\nabla_{\mathsf{d}} f(x_n)$ then $\boxed{f(x_n) - f(x_*) \leq (1-h\lambda)^n \left(f(x_0) - f(x_*)\right)}$

Gradient descent and variants in another geometry

On $\mathbb{R}^d$,

Riemannian distance: d distance and $\mathsf{d}^2(x,y) \simeq (y-x)^\top Q(x)(y-x)$ if $y \to x$.

Denote $\nabla_{\mathsf{d}} f(x) = Q(x)^{-1}\nabla f(x)$ gradient in geometry of d.
Assume f is $\lambda > 0$ convex and L smooth in geometry of d.

Gradient descent ($h \leq 1/L$)

$$x_{n+1} = x_n - h\nabla_{\mathsf{d}} f(x_n)$$

then

$$\boxed{f(x_n) - f(x_*) \leq (1-h\lambda)^n \left(f(x_0) - f(x_*)\right)}$$

Proximal step (Implicit)

$$x_{n+1} = x_n - h\nabla_{\mathsf{d}} f(x_{n+1})$$

$$\Leftrightarrow x_{n+1} \in \arg\min_x \left(f(x) + \frac{\mathsf{d}^2(x,x_n)}{2h}\right)$$

then

$$\boxed{f(x_n) - f(x_*) \leq \left(\frac{1}{1+2h\lambda}\right)^n \left(f(x_0) - f(x_*)\right)}$$

Gradient descent and variants in another geometry

On $\mathbb{R}^d$,

Riemannian distance: d distance and $\mathrm{d}^2(x,y) \simeq (y-x)^\top Q(x)(y-x)$ if $y \to x$.

Denote $\nabla_{\mathrm{d}} f(x) = Q(x)^{-1}\nabla f(x)$ gradient in geometry of d.
Assume f is $\lambda > 0$ convex and L smooth in geometry of d.

Gradient descent ($h \leq 1/L$)

$x_{n+1} = x_n - h\nabla_{\mathrm{d}} f(x_n)$ then $\boxed{f(x_n) - f(x_*) \leq (1-h\lambda)^n \left(f(x_0) - f(x_*)\right)}$

Proximal step (Implicit)

$x_{n+1} = x_n - h\nabla_{\mathrm{d}} f(x_{n+1})$

$\Leftrightarrow x_{n+1} \in \arg\min_x \left(f(x) + \frac{\mathrm{d}^2(x,x_n)}{2h}\right)$ then $\boxed{f(x_n) - f(x_*) \leq \left(\frac{1}{1+2h\lambda}\right)^n \left(f(x_0) - f(x_*)\right)}$

Gradient flow ($h \to 0$)

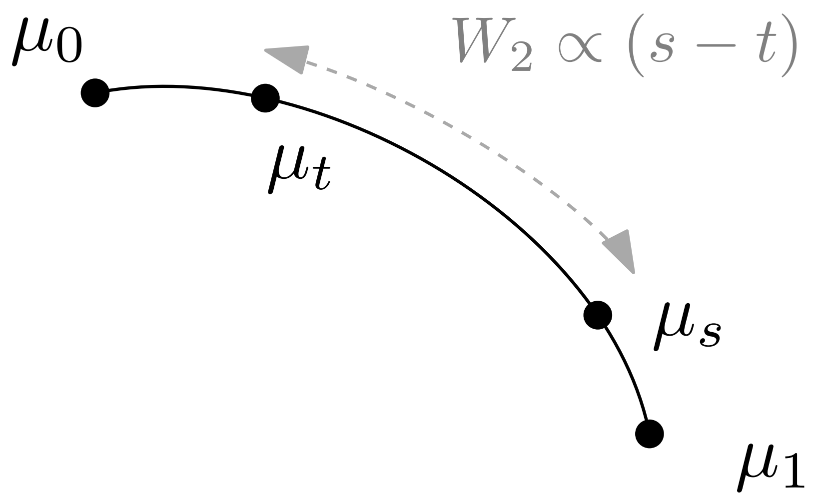
$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla_{\mathrm{d}} f(x_t).$ then $\boxed{f(x_t) - f(x_*) \leq \exp(-2\lambda t)\left(f(x_0) - f(x_*)\right)}$

Let's move to Wasserstein geometry: geodesics

On $\mathcal{P}_2(\mathbb{R}^d)$, we take d the Wasserstein 2 distance W_2.

Definiton. We call $(\mu_t)_{t\in[0,1]}$ a geodesic if for any $0 \leq t \leq s \leq 1$:

$$W_2(\mu_t, \mu_s) = (s-t)W_2(\mu_0, \mu_1).$$

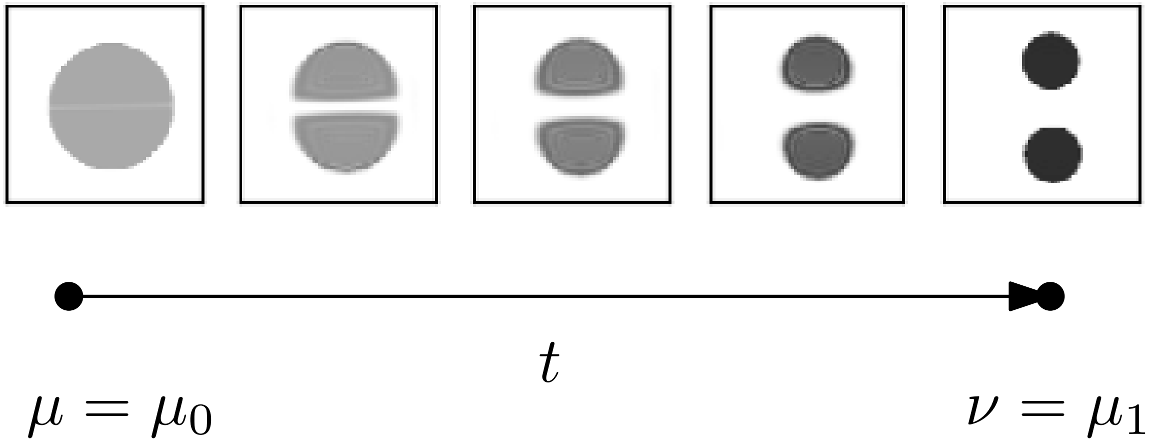


Let's move to Wasserstein geometry: geodesics

On $\mathcal{P}_2(\mathbb{R}^d)$, we take d the Wasserstein 2 distance W_2.

Definiton. We call $(\mu_t)_{t\in[0,1]}$ a geodesic if for any $0 \leq t \leq s \leq 1$:

$$W_2(\mu_t, \mu_s) = (s-t)W_2(\mu_0, \mu_1).$$



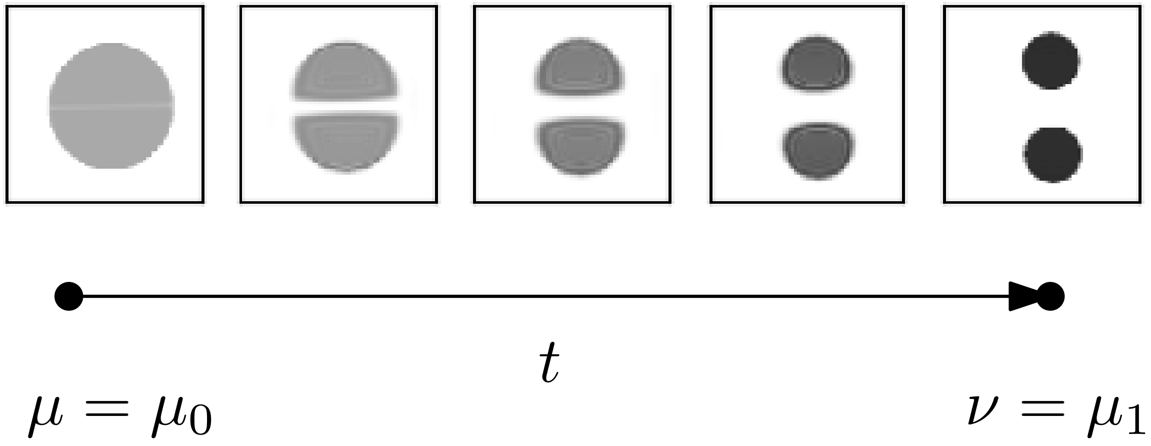
Theorem. Take (X, Y) optimal coupling between μ and ν. Then a geodesic between μ and ν is given by $\mu_t = \mathrm{Law}((1-t)X + tY)$.

Let's move to Wasserstein geometry: geodesics

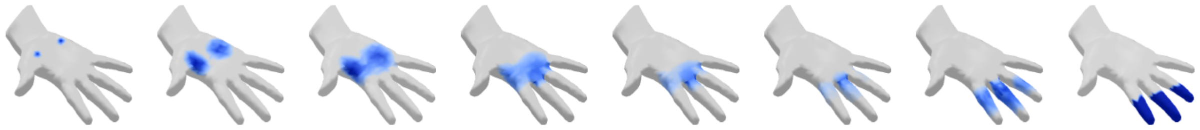
On $\mathcal{P}_2(\mathbb{R}^d)$, we take d the Wasserstein 2 distance W_2.

Definiton. We call $(\mu_t)_{t\in[0,1]}$ a geodesic if for any $0 \le t \le s \le 1$:

$$W_2(\mu_t, \mu_s) = (s-t)W_2(\mu_0, \mu_1).$$



Theorem. Take (X, Y) optimal coupling between μ and ν. Then a geodesic between μ and ν is given by $\mu_t = \mathrm{Law}((1-t)X + tY)$.



(Also works when base space is a manifold!)

[Lavenant et al (2018). Dynamical OT on discrete surfaces]

Wasserstein geometry: gradient flow of entropy

Theorem. Assume $V : \mathbb{R}^d \to \mathbb{R}$ is λ-convex, so that $\pi \propto \exp(-V)$ is ($\lambda-$)log concave. Then $\mathrm{KL}(\cdot|\pi)$ is λ-convex in the W_2 geometry.

Wasserstein geometry: gradient flow of entropy

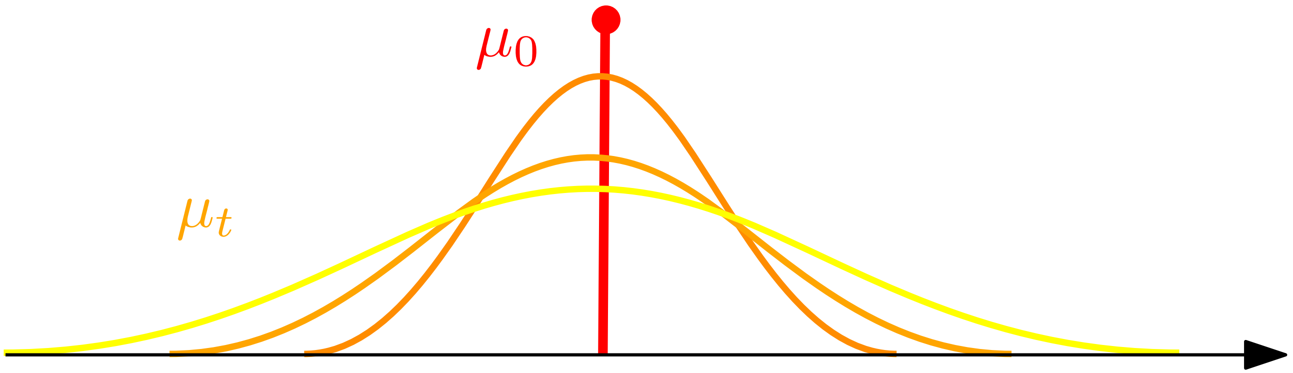
Theorem. Assume $V : \mathbb{R}^d \to \mathbb{R}$ is λ-convex, so that $\pi \propto \exp(-V)$ is $(\lambda-)$log concave. Then $\mathrm{KL}(\cdot|\pi)$ is λ-convex in the W_2 geometry.

Theorem. The gradient flow

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = -\nabla_{W_2}\mathrm{KL}(\mu_t|\pi)$$

is the Fokker Planck equation:

$$\frac{\partial\mu}{\partial t} = \Delta\mu_t + \mathrm{div}(\mu_t \nabla V).$$



$\mu_t = \mathrm{Law}(X_t)$ law of diffusion $(X_t)_t$ with:

$$\mathrm{d}X_t = -\nabla V(X_t) + \sqrt{2}\mathrm{d}B_t.$$

Remark. Possible to write:

$$\nabla_{W_2}\mathrm{KL}(\mu|\pi)(x) = -\nabla V(x) - \nabla \log\mu(x).$$

[Jordan, Kinderlehrer & Otto (1998). The Variational Formulation of the FP Equation]

Wasserstein geometry: gradient flow of entropy

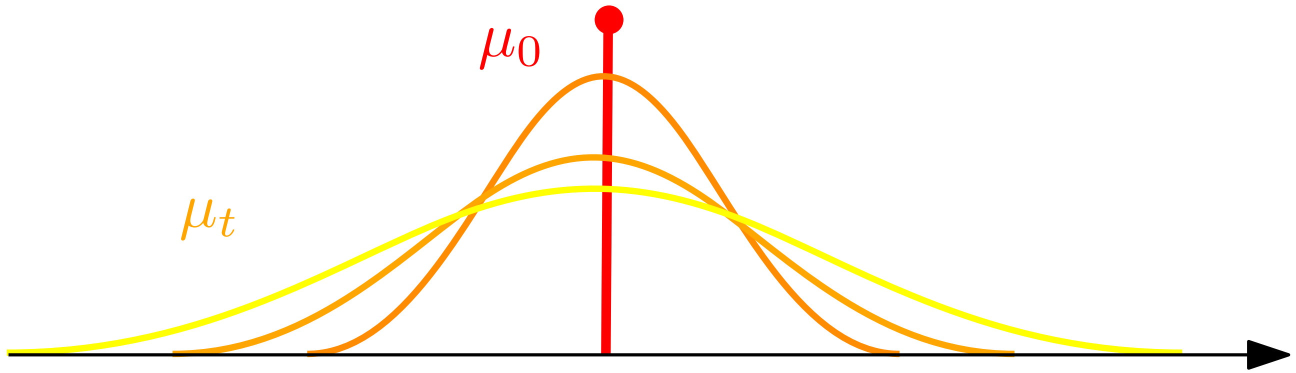
Theorem. Assume $V : \mathbb{R}^d \to \mathbb{R}$ is λ-convex, so that $\pi \propto \exp(-V)$ is $(\lambda-)$log concave. Then $\mathrm{KL}(\cdot|\pi)$ is λ-convex in the W_2 geometry.

Theorem. The gradient flow

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = -\nabla_{W_2}\mathrm{KL}(\mu_t|\pi)$$

is the Fokker Planck equation:

$$\frac{\partial\mu}{\partial t} = \Delta\mu_t + \mathrm{div}(\mu_t\nabla V).$$



Consequence. We have:

$$\mathrm{KL}(\mu_t|\pi) \leq \exp(-2\lambda t)\mathrm{KL}(\mu_0|\pi).$$

Remark. Possible to write:

$$\nabla_{W_2}\mathrm{KL}(\mu|\pi)(x) = -\nabla V(x) - \nabla\log\mu(x).$$

[Jordan, Kinderlehrer & Otto (1998). The Variational Formulation of the FP Equation]

What about gradient descent?

Proximal step

$$\mu_{n+1} \in \arg\min_{\mu} \left(\mathrm{KL}(\mu|\pi) + \frac{W_2^2(\mu, \mu_n)}{2h} \right)$$

(Converge to Wasserstein gradient flow as $h \to 0$)

Theorem. If V is λ-convex

$$\mathrm{KL}(\mu_n|\pi) \leq \left(\frac{1}{1 + 2h\lambda} \right)^n \mathrm{KL}(\mu_0|\pi)$$

(But no direct way to implement these iterations)

What about gradient descent?

Proximal step

$$\mu_{n+1} \in \arg\min_{\mu} \left(\mathrm{KL}(\mu|\pi) + \frac{W_2^2(\mu,\mu_n)}{2h} \right)$$

(Converge to Wasserstein gradient flow as $h \to 0$)

Theorem. If V is λ-convex

$$\mathrm{KL}(\mu_n|\pi) \leq \left(\frac{1}{1+2h\lambda} \right)^n \mathrm{KL}(\mu_0|\pi)$$

(But no direct way to implement these iterations)

Langevin Monte Carlo:

$$X_{n+1} = X_n - h\nabla V(X_n) + \sqrt{2h} Z_n$$

$$Z_n \sim \mathcal{N}(0, I)$$

✓ Is a Wasserstein gradient step of $\mu \mapsto \int V \mathrm{d}\mu$

✗ Not directly a gradient step in Wasserstein geometry

No big surprise: (unadjusted) Langevin Monte Carlo is not so easy to analyze.

Example of results for Langevin Monte Carlo

Langevin Monte Carlo:

$$X_{n+1} = X_n - h\nabla V(X_n) + \sqrt{2h} Z_n \qquad Z_n \sim \mathcal{N}(0, I)$$

Theorem. In $\mathbb{R}^d$ assuming V is λ-convex and L smooth and with $h \leq 1/L$:

$$W_2^2(\mu_n, \pi) \leq \exp(-n\lambda h) W_2^2(\mu_0, \pi) + \mathcal{O}\left(\frac{\lambda}{L} dh\right).$$

Rate of convergence of gradient descent.

bias term because $h > 0$.

[Chewi (2024+). Log-concave sampling, Section 4.3]

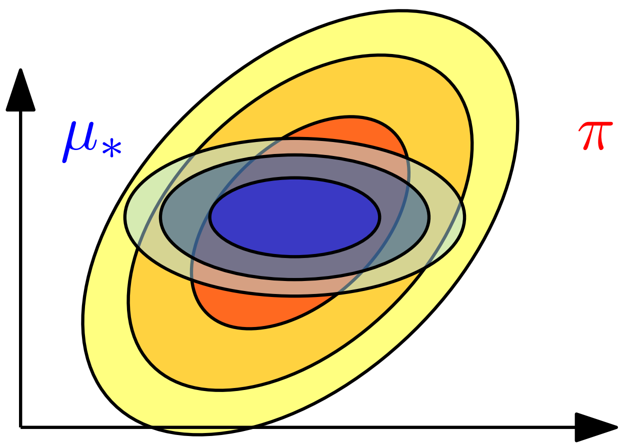
Different context (variational inference), same geometry

Definition. The mean field approximation of π is the law μ_* of $(X_1, \dots, X_K)$ with $X_1, \dots, X_K$ independent which minimizes

$$\mu \mapsto \mathrm{KL}(\mu|\pi)$$

Algorithm. Random Scan Coordinate Variational inference.
Initialize $X_1, \dots, X_K$, then iterate:

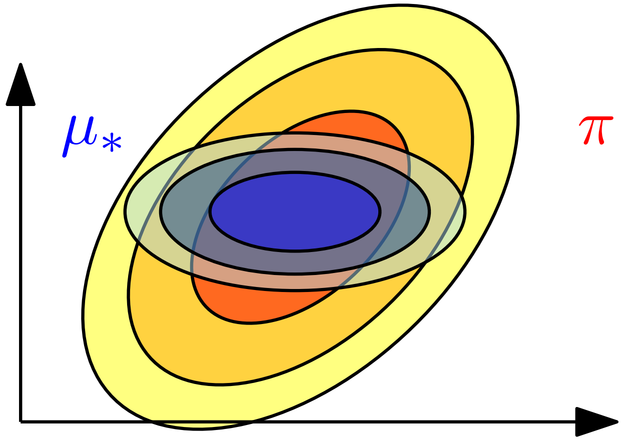
1. Select index k_n at random,
2. Change law of X_{k_n} only such that it minimizes $\mathrm{KL}(\mu|\pi)$.



Different context (variational inference), same geometry

Definition. The mean field approximation of π is the law μ_* of $(X_1, \dots, X_K)$ with $X_1, \dots, X_K$ independent which minimizes

$$\mu \mapsto \mathrm{KL}(\mu|\pi)$$



Important: space of factorized distributions **convex** for W_2 geometry.

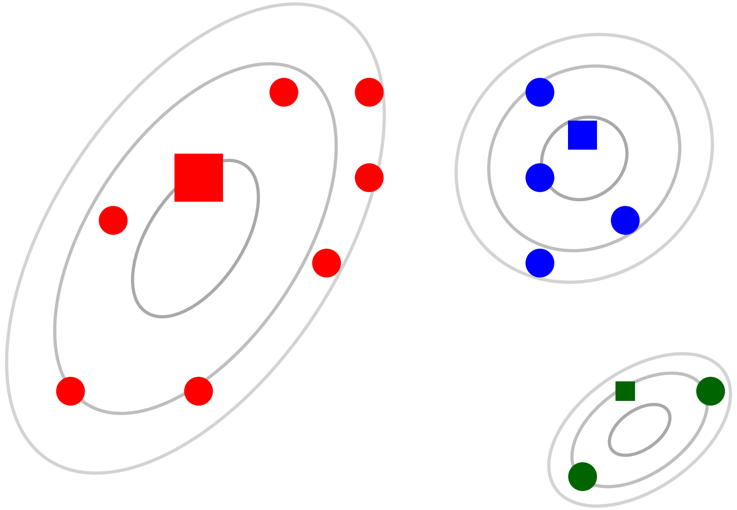
Algorithm. Random Scan Coordinate Variational inference.
Initialize $X_1, \dots, X_K$, then iterate:

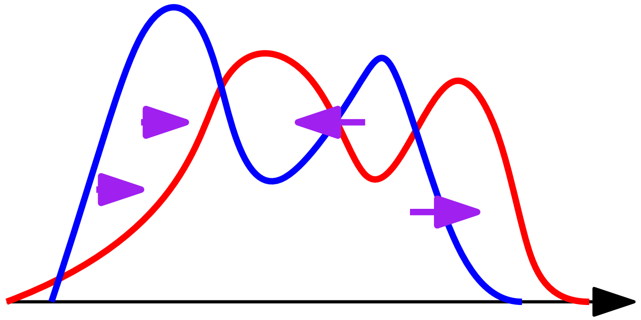
1. Select index k_n at random,
2. Change law of X_{k_n} only such that it minimizes $\mathrm{KL}(\mu|\pi)$.

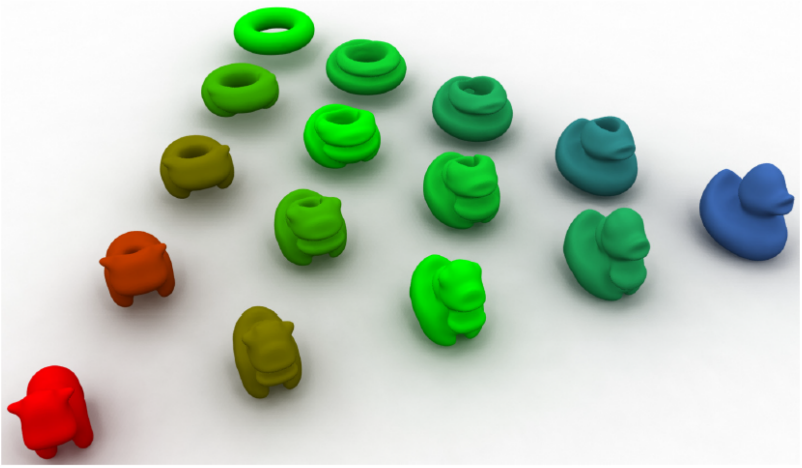
Theorem. Assume V is λ-convex and L smooth, then with $\mu_n = \mathrm{Law}(X_n)$:

$$\mathbb{E}(\mathrm{KL}(\mu_n|\pi) - \mathrm{KL}(\mu^*|\pi)) \leq \left(1 - \frac{\lambda}{KL}\right)^n (\mathrm{KL}(\mu_0|\pi) - \mathrm{KL}(\mu^*|\pi)).$$

[Arnese & Lacker (2024). Convergence of CAVI for log-concave measures via OT]
[Lavenant & Zanella (2024). Convergence rate of RS-CAVI under log-concavity]







The End