Optimal transport bootcamp

Espen Bernton

Department of Statistics, Columbia University

January 15, 2019

Outline

Part 1: Theoretical foundations

- Origins of OT: Monge & Kantorovich problems
- Primal and dual formulations
- Wasserstein distance
- Some important properties

Outline

Part 2: Computation and applications

- Exact and approximate computation
- Some statistical properties
- OT as a loss function
- Application: Generative models

Part 1: Theoretical foundations

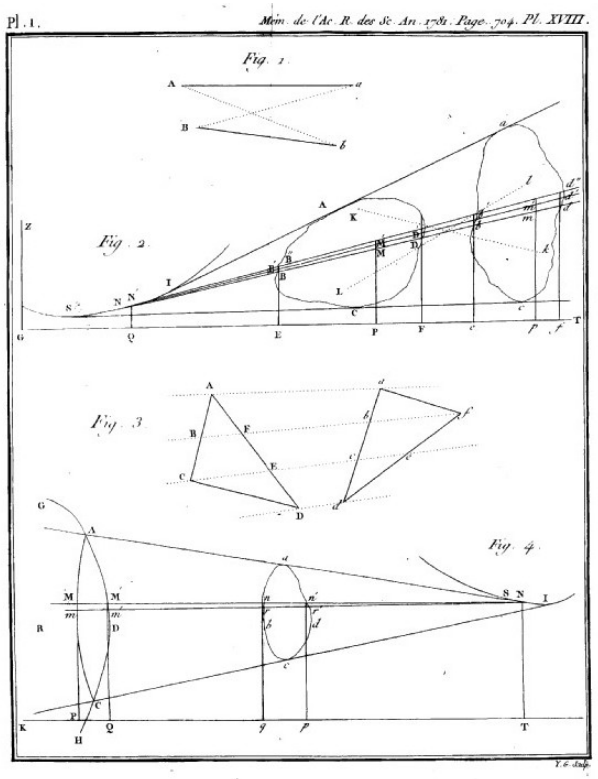
Origins of OT: Monge (1781)

666 · MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.



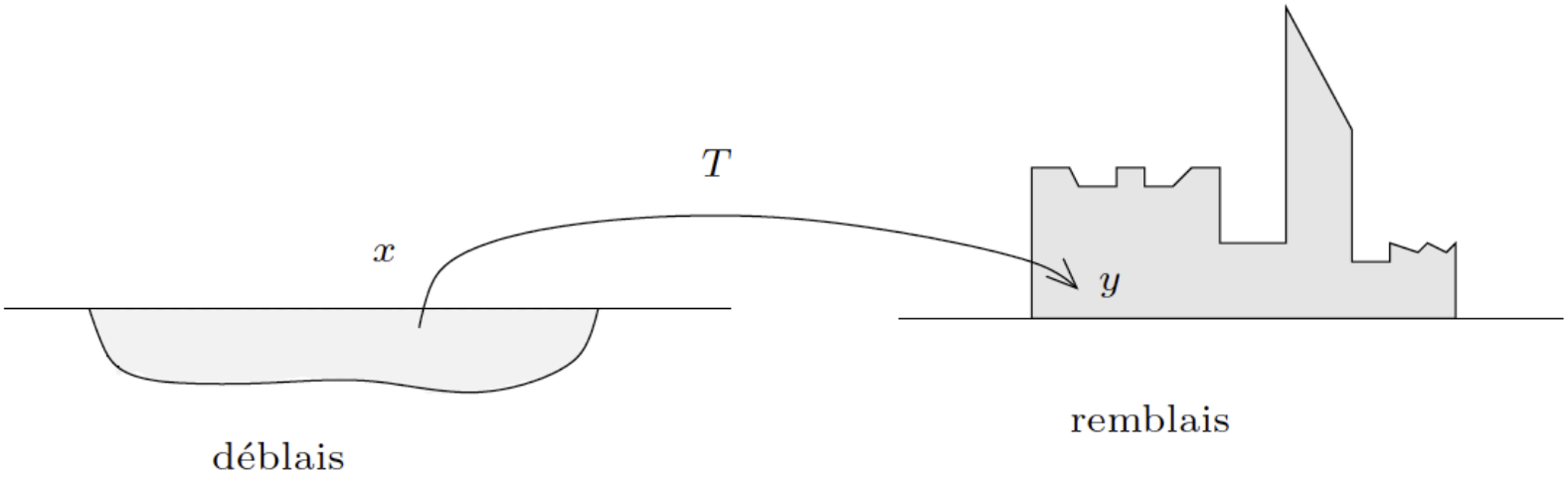


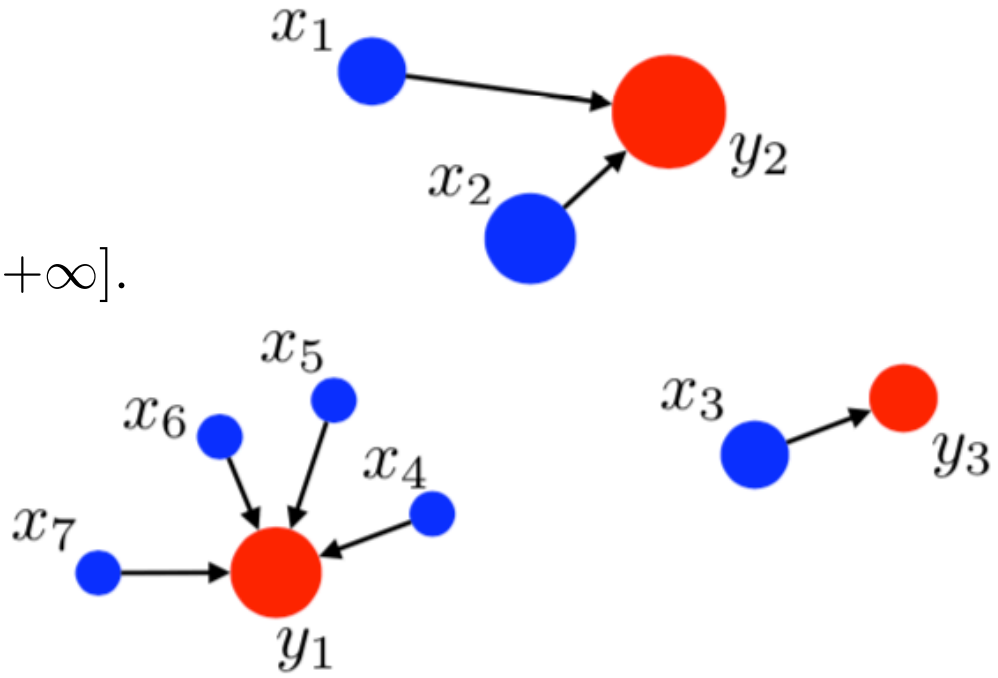
Figure from Villani (2008).

Origins of OT: Monge (1781)

<u>Modern formulation:</u>

Given

- Distributions μ, ν on $\mathcal{X}$,
- Cost function $c : \mathcal{X}^2 \to [0, +\infty]$.



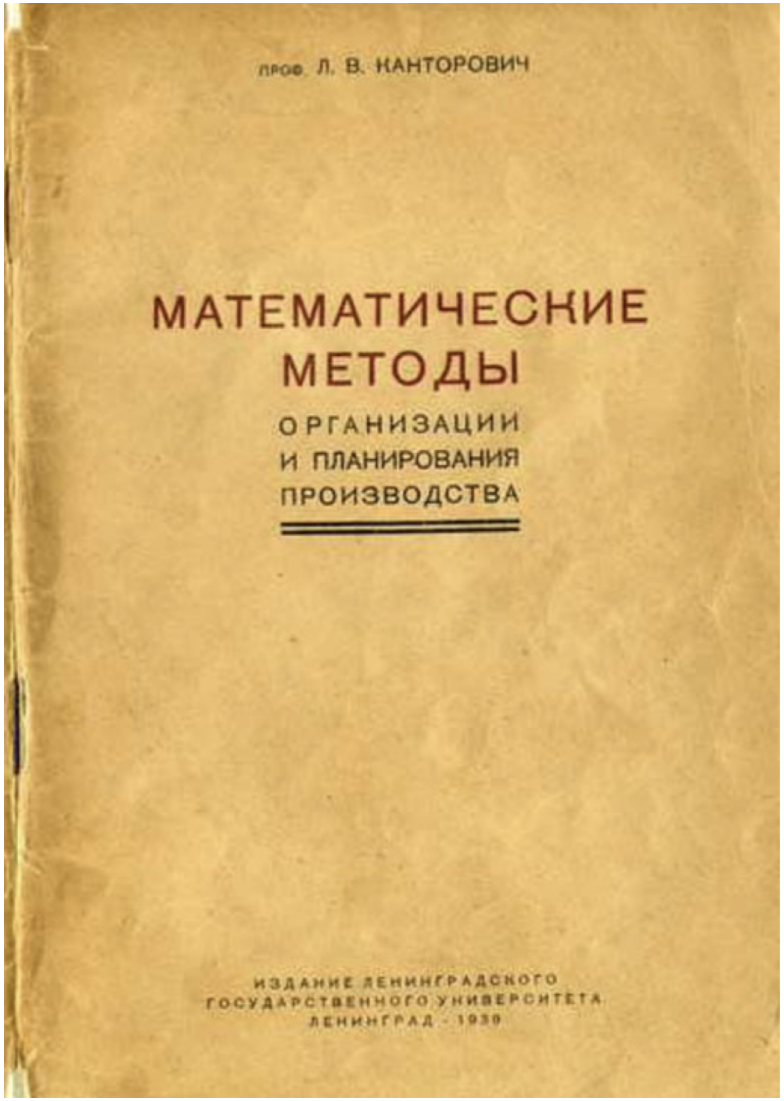
Solve

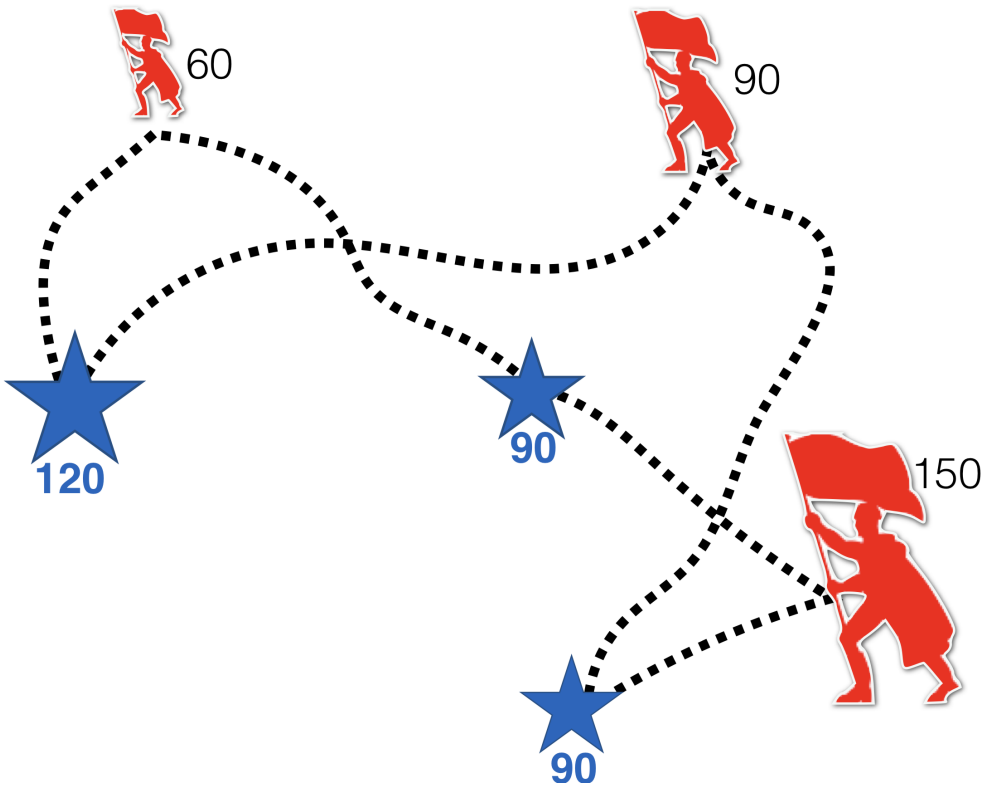
$$\min_{\boldsymbol{T}_{\#}\mu=\nu} \int_{\mathcal{X}} c(x, \boldsymbol{T}(x)) \mathrm{d}\mu(x),$$

and find a minimizer $\boldsymbol{T}^\star$, where $\boldsymbol{T}_{\#}\mu(A) = \mu(\boldsymbol{T}^{-1}(A))$ for $A \subset \mathcal{X}$.

Figure from Peyré and Cuturi (2019).

Origins of OT: Kantorovich (1942)





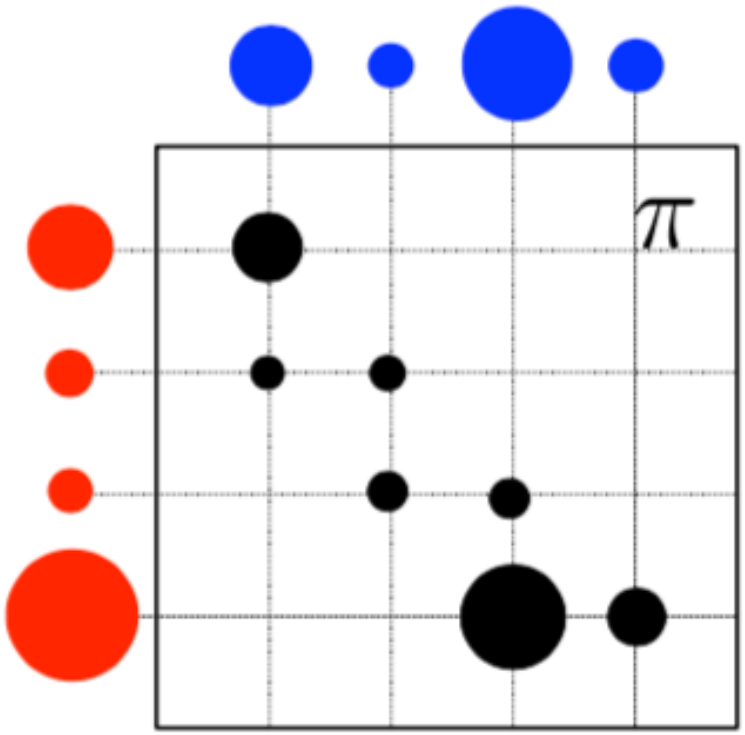
Figures from Vershik (2013); Cuturi and Solomon (2017).

Origins of OT: Kantorovich (1942)

<u>Problem studied by Kantorovich:</u>

Given

- Distributions μ, ν on $\mathcal{X}$,
- Cost function $c : \mathcal{X}^2 \to [0, +\infty]$.



Solve

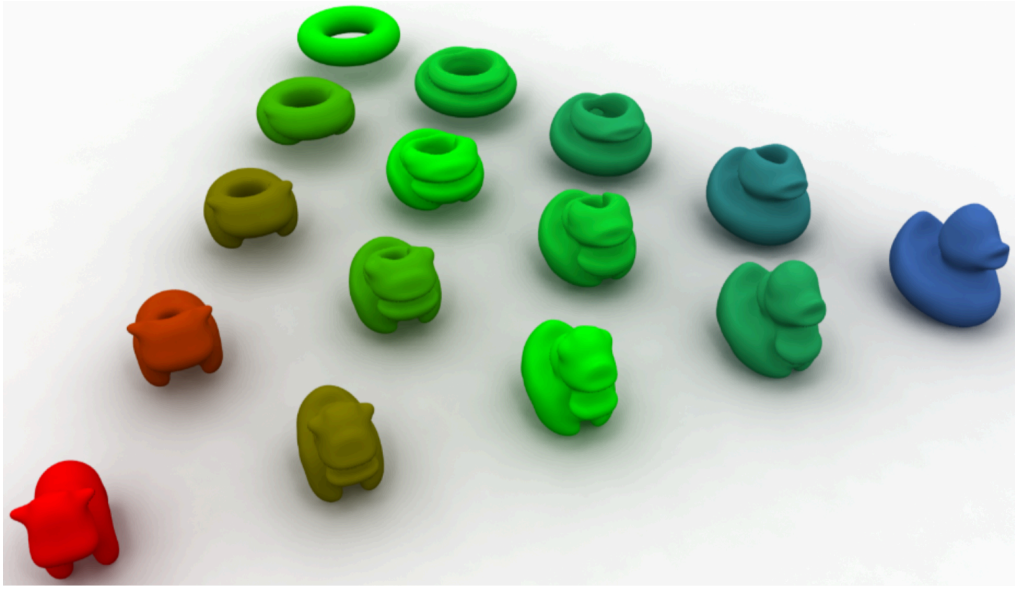
$$\min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x,y) \mathrm{d}\boldsymbol{\pi}(x,y),$$

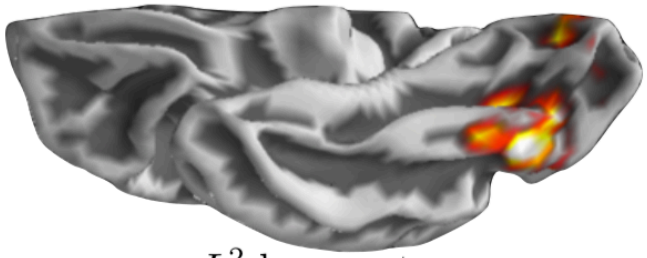
and find a minimizer $\boldsymbol{\pi}^\star$, where Π is the set of couplings of μ and ν.

Figure from Peyré and Cuturi (2019).

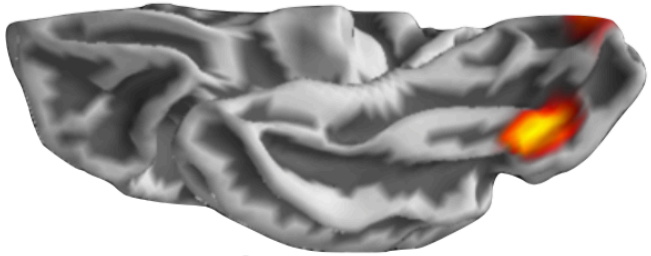
Origins of OT

Lifts the **geometric structure** captured by c on $\mathcal{X}$ up to the space of probability measures over $\mathcal{X}$.

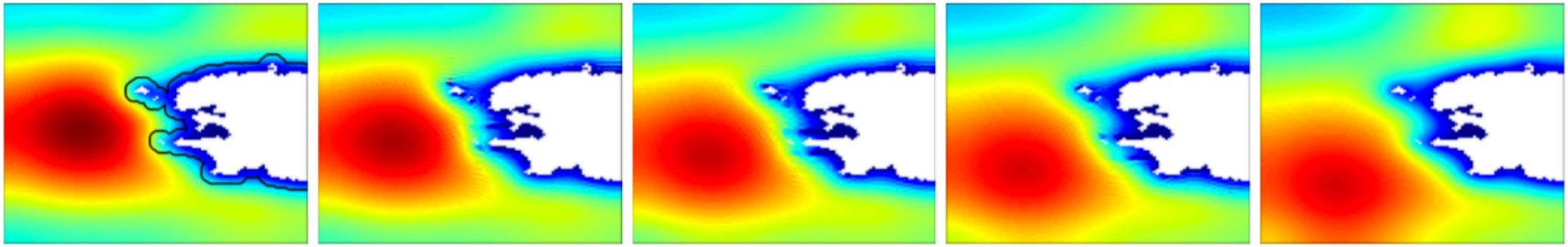




L^2 barycenter



W_2^2 barycenter



$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

Figures from Solomon et al. (2015); Gramfort et al. (2015); Papadakis et al. (2014).

Two OT problems

Monge problem (MP):

$$\min_{\boldsymbol{T}_{\#}\mu=\nu} \int_{\mathcal{X}} c(x, \boldsymbol{T}(x))\mathrm{d}\mu(x).$$

Kantorovich problem (KP):

$$\min_{\boldsymbol{\pi}\in\Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x, y)\mathrm{d}\boldsymbol{\pi}(x, y).$$

- If $\boldsymbol{T}$ is feasible for (MP), it also gives rise to a feasible solution for (KP): $\boldsymbol{\pi} = (\mathrm{Id}, \boldsymbol{T})_{\#}\mu$.
- The set of maps $\{\boldsymbol{T} : \boldsymbol{T}_{\#}\mu = \nu\}$ might be empty, but $\Pi(\mu, \nu)$ is always non-empty: $\mu \otimes \nu \in \Pi(\mu, \nu)$.

Two OT problems

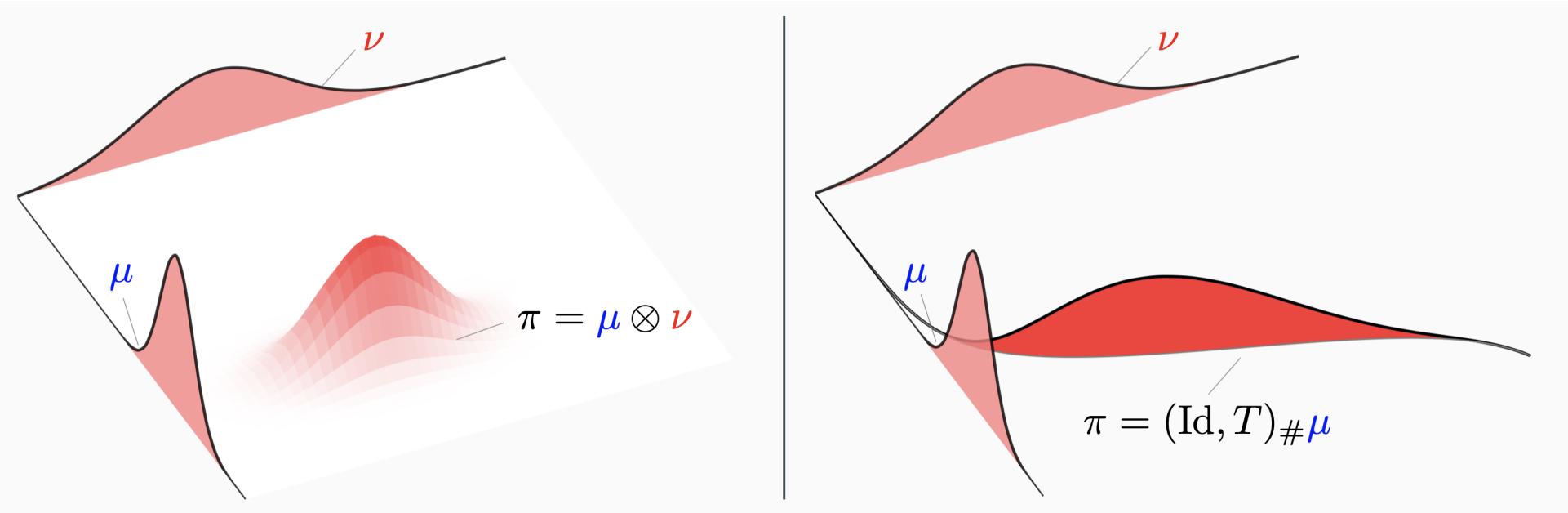


Figure from Chizat (2019).

The primal problem

<u>Primal problem:</u>

$$\min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x,y) \mathrm{d}\boldsymbol{\pi}(x,y).$$

- Does a minimizer exist?
- If so, does only one exist, or could there be many?
- What can be said about the form of the minimizer(s)?

Existence of minimizers

Theorem

Suppose that $\mathcal{X}$ is a Polish space. Suppose that $c : \mathcal{X}^2 \to [0, +\infty]$ is lower semicontinuous. Then (KP) admits a solution.

Proof idea: Show that $\Pi(\mu, \nu)$ is compact and that

$$K(\boldsymbol{\pi}) := \int_{\mathcal{X}^2} c(x, y) \mathrm{d}\boldsymbol{\pi}(x, y)$$

is lower semicontinuous for an appropriate notion of topology.

- Topology of weak convergence.
- Show compactness with Prokhorov's theorem.
- Take sequence $c_n \uparrow c$, where c_n continuous and bounded, so that each map $\boldsymbol{\pi} \mapsto \int c_n \mathrm{d}\boldsymbol{\pi}$ is continuous.
- Show $K(\boldsymbol{\pi}) = \sup_n \int c_n \mathrm{d}\boldsymbol{\pi}$ by monotone convergence, giving lower semicontinuity of $K(\boldsymbol{\pi})$.

See Santambrogio (2015, Theorem 1.8) for a complete proof.

Duality

To derive the dual of (KP), first note that the constraint $\boldsymbol{\pi} \in \Pi(\mu, \nu)$ can be encoded as follows:

$$\sup_{\varphi,\psi} \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{X}} \psi \, \mathrm{d}\nu - \int_{\mathcal{X}^2} \{\varphi(x) + \psi(y)\} \, \mathrm{d}\boldsymbol{\pi}$$

$$= \begin{cases} 0 & \text{if } \boldsymbol{\pi} \in \Pi(\mu, \nu), \\ +\infty & \text{otherwise,} \end{cases}$$

where $\varphi \in L_1(\mu)$ and $\psi \in L_1(\nu)$.

Duality

Using this formulation of the constraint, (KP) can be expressed

$$\min_{\boldsymbol{\pi}\geq 0}\left\{\int_{\mathcal{X}^2} c\,\mathrm{d}\boldsymbol{\pi} + \sup_{\varphi,\psi}\int_{\mathcal{X}}\varphi\,\mathrm{d}\mu + \int_{\mathcal{X}}\psi\,\mathrm{d}\nu - \int_{\mathcal{X}^2}\{\varphi(x)+\psi(y)\}\,\mathrm{d}\boldsymbol{\pi}\right\}.$$

If we can interchange* the "min" and the "sup", we get

$$\sup_{\varphi,\psi}\left\{\int_{\mathcal{X}}\varphi\,\mathrm{d}\mu + \int_{\mathcal{X}}\psi\,\mathrm{d}\nu + \inf_{\boldsymbol{\pi}\geq 0}\int_{\mathcal{X}^2}\{c(x,y)-\varphi(x)-\psi(y)\}\,\mathrm{d}\boldsymbol{\pi}\right\}.$$

*Needs formal justification.

Duality

Note that

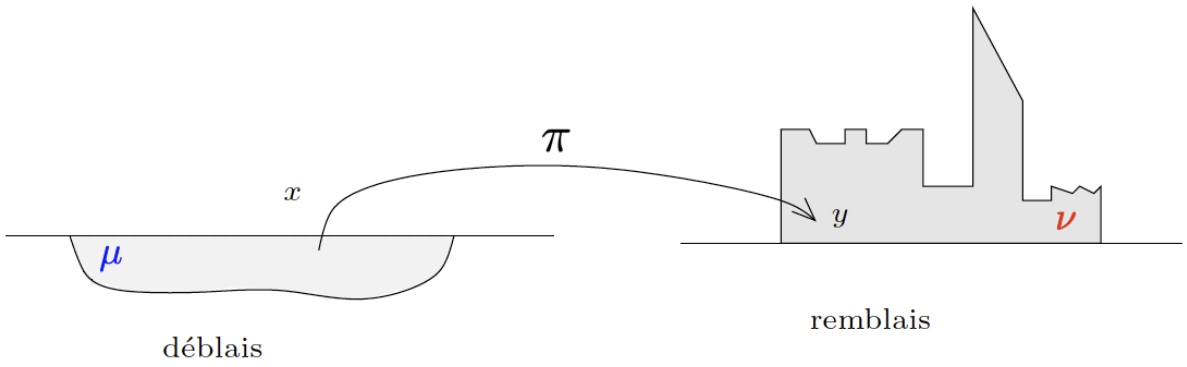
$$\inf_{\boldsymbol{\pi}\geq 0} \int_{\mathcal{X}^2} \{c(x,y) - \varphi(x) - \psi(y)\} \, \mathrm{d}\boldsymbol{\pi} = \begin{cases} 0 & \text{if } \varphi \oplus \psi \leq c \text{ on } \mathcal{X}^2, \\ -\infty & \text{otherwise,} \end{cases}$$

where $\varphi \oplus \psi(x,y) := \varphi(x) + \psi(y)$.

In turn, we get the dual problem (DP)

$$\max \left\{ \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{X}} \psi \, \mathrm{d}\nu : \varphi \oplus \psi \leq c, \text{ and } \varphi \in L_1(\mu), \psi \in L_1(\nu) \right\}.$$

Duality: Intuition



We hire a logistics company to transport "déblais" distributed as μ to "remblais" distributed as ν. They charge

- $\varphi(x)$ to pick up unit mass at x, regardless of destination.
- $\psi(y)$ to drop off unit mass at y, regardless of origin.

The fair price for moving a unit mass from x to y is $c(x, y)$.

- We only accept offers such that $\varphi(x) + \psi(y) \leq c(x, y)$.

The company wants to maximize the total price $\int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{X}} \psi \, \mathrm{d}\nu$.

Duality

<u>Dual problem (DP):</u>

$$\max\left\{\int_{\mathcal{X}} \varphi \,\mathrm{d}\mu + \int_{\mathcal{X}} \psi \,\mathrm{d}\nu : \varphi \oplus \psi \leq c, \text{ and } \varphi \in L_1(\mu), \psi \in L_1(\nu)\right\}.$$

Duality

Theorem

Suppose that $\mathcal{X}$ is a Polish space. Suppose that $c : \mathcal{X}^2 \to [0, +\infty]$ is lower semicontinuous. Then $\min$ *(KP)* $= \max$ *(DP).*

Proof: See Santambrogio (2015, Theorem 1.43 and Remark 1.45).

Furthermore, at optimality we have that

- $\varphi^\star(x) + \psi^\star(y) = c(x, y)$ for $\boldsymbol{\pi}^\star$-almost every (x, y).
- $\varphi^\star$ and $\psi^\star$ are *c-conjugate*: $\psi^\star(y) = \inf_{x \in \mathcal{X}} c(x, y) - \varphi^\star(x)$.
- $\boldsymbol{\pi}^\star$ is supported on a *c-cyclically monotone* set.

Cyclical monotonicity

Definition

A set $\mathcal{S} \subset \mathcal{X}^2$ is c-cyclically monotone if for any $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{S}$,

$$\sum_{i=1}^n c(x_i, y_i) \leq \sum_{i=1}^n c(x_i, y_{\sigma(i)}),$$

for any permutation $\sigma \in \mathrm{Perm}(n)$.

Many important properties follow from cyclical monotonicity.

For instance, suppose that $\mathcal{X} = \mathbb{R}^d$, $c(x, y) = \|x - y\|^2$, and $\mu \ll \mathrm{Leb}$. From cyclical monotonicity, McCann (1995) then derived that

- $\boldsymbol{\pi}^\star$ is unique,
- $\boldsymbol{\pi}^\star = (\mathrm{Id}, \boldsymbol{T})_{\#}\mu$, i.e. (MP) and (KP) coincide,
- $\boldsymbol{T}(x) = \nabla\varphi^\star(x)$, where $\varphi^\star$ is a convex function.

Summary so far

- Motivated and defined the Monge and Kantorovich problems.
- Showed the existence of minimizers in (KP).
- Derived the dual problem (DP) to the primal (KP).
- Illustrated that (DP) and cyclical monotonicity can be used to derive important properties of the minimizers.

Wasserstein distance

Definition

Let $(\mathcal{X}, d)$ be a Polish space. Then for any $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$,

$$\mathcal{W}_p(\mu, \nu) = \left(\min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} d^p(x, y) \mathrm{d}\boldsymbol{\pi} \right)^{1/p}$$

defines the p-Wasserstein distance between μ and ν.

- $\mathcal{P}_p(\mathcal{X})$ is the set of distributions on $\mathcal{X}$ with finite p-th moment.

Defines a proper distance on $\mathcal{P}_p(\mathcal{X})$:

- $\mathcal{W}_p(\mu, \nu) \geq 0$, with equality if and only if $\mu = \nu$.
- $\mathcal{W}_p(\mu, \nu) = \mathcal{W}_p(\nu, \mu)$.
- $\mathcal{W}_p(\mu, \nu) \leq \mathcal{W}_p(\mu, \rho) + \mathcal{W}_p(\rho, \nu)$.

Proof: See Villani (2008, Definition 6.1).

Topological properties of $(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p)$

Theorem

If $(\mathcal{X}, d)$ is a Polish space and $p \geq 1$, then $(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p)$ is also Polish.

Proof: See Villani (2008, Theorem 6.18).

Important geometric concepts on $(\mathcal{P}_p(\mathcal{X}), \mathcal{W}_p)$:

- geodesics, or "shortest paths", between distributions,
- barycenters, or "geometric centers", of collections of distributions,
- Riemannian-like structure $\implies$ Otto calculus,
- gradients flows of functionals on $\mathcal{P}_2(\mathcal{X})$ etc...

Geodesic between μ and ν

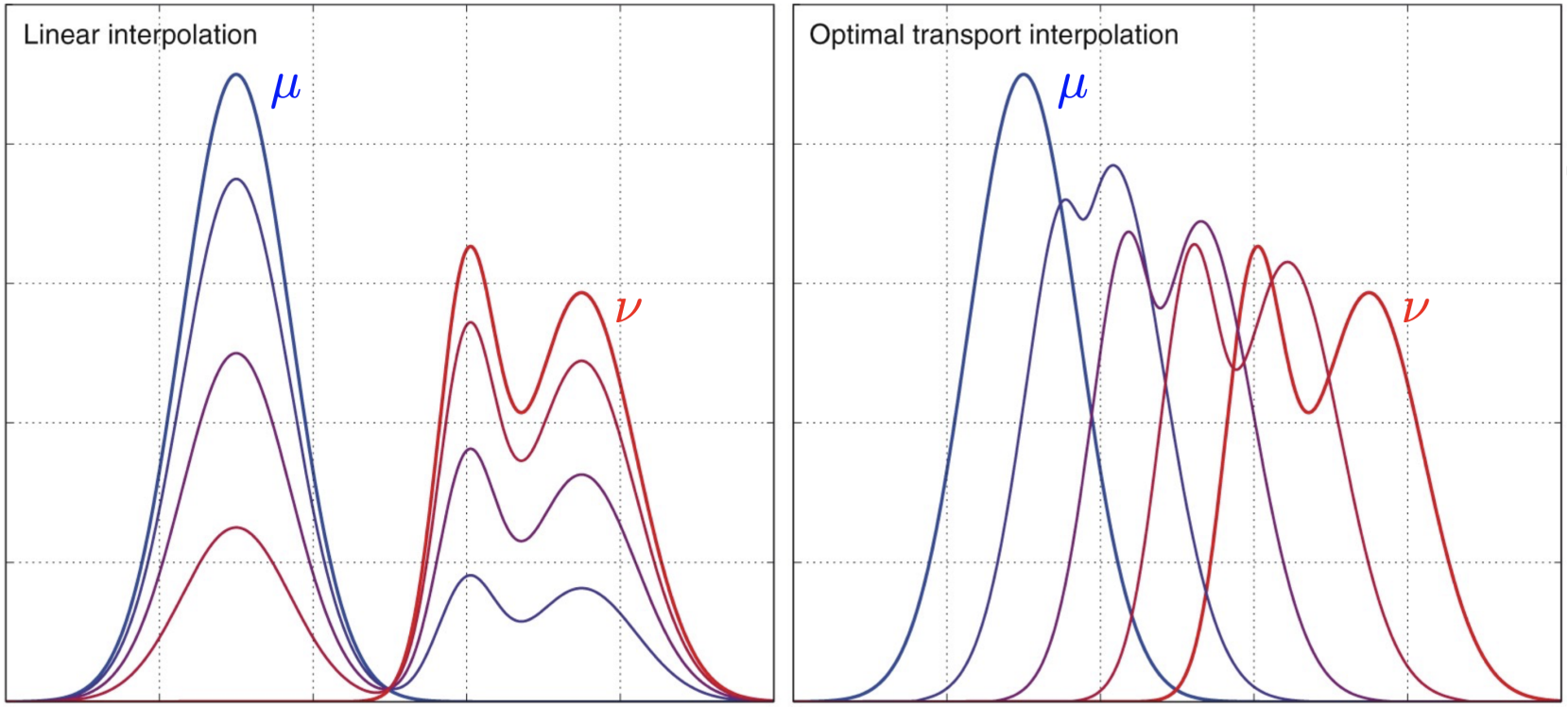


Figure from Cuturi and Solomon (2017).

$\mathcal{W}_p$ metrizes weak convergence

Definition

A sequence of distributions ρ_k converges weakly to ρ in $\mathcal{P}_p(\mathcal{X})$ if

- ρ_k converges weakly to ρ in the usual sense,
- there exists $x_0 \in \mathcal{X}$ such that

$$\lim_{k\to\infty} \int_{\mathcal{X}} d(x_0, x)^p \mathrm{d}\rho_k = \int_{\mathcal{X}} d(x_0, x)^p \mathrm{d}\rho.$$

Theorem

$\mathcal{W}_p$ metrizes weak convergence in $\mathcal{P}_p(\mathcal{X})$, that is $\mathcal{W}_p(\rho_k, \rho) \to 0$ if and only if ρ_k converges weakly to ρ in $\mathcal{P}_p(\mathcal{X})$.

Proof: See Villani (2008, Theorem 6.9).

Important in statistics and ML:

- If $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ and $\{x_i\}_{i=1}^k$, $\{y_i\}_{i=1}^k$ are s.t. $\hat{\mu}_k = \frac{1}{k}\sum \delta_{x_i} \to \mu$ and $\hat{\nu}_k = \frac{1}{k}\sum \delta_{y_i} \to \nu$, then $\mathcal{W}_p(\hat{\mu}_k, \hat{\nu}_k) \to \mathcal{W}_p(\mu, \nu)$.

Special case: $\mathcal{X} = \mathbb{R}$ and $d(x, y) = |x - y|$

The CDF and quantile function of μ are given by

$$F_\mu(x) = \mu((-\infty, x]), \quad \text{and} \quad F_\mu^{-1}(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq t\}.$$

An elementary result says that

- $\left(F_\mu^{-1}\right)_\# \mathrm{Unif}[0,1] = \mu$,
- $(F_\mu)_\# \mu = \mathrm{Unif}[0,1]$, provided $\mu \ll \mathrm{Leb}$.

Hence, if $\mu \ll \mathrm{Leb}$, then $\boldsymbol{T} = F_\nu^{-1} \circ F_\mu$ is such that $\boldsymbol{T}_\# \mu = \nu$.

<u>*Exercise:*</u> Show that $\boldsymbol{T}$ solves (MP) and (KP), and that

$$\mathcal{W}_p(\mu, \nu) = \left(\int_0^1 \left| F_\mu^{-1}(t) - F_\nu^{-1}(t) \right|^p \mathrm{d}t \right)^{1/p}.$$

Hint: See Santambrogio (2015, Chapter 2).

Questions?