Primer on Optimal Transport

Marcel Nutz
Columbia University

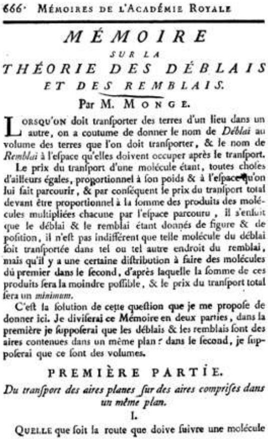
June 2025

Some References

- Santambrogio (2015), *Optimal transport for applied mathematicians*
- Villani (2003), *Topics in optimal transportation*
- Villani (2009), *Optimal transport, old and new*
- ...

Outline

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.

Le prix du tranſport d'une molécule étant, toutes choſes d'ailleurs égales, proportionnel à ſon poids & à l'eſpace qu'on lui fait parcourir, & par conſéquent le prix du tranſport total devant être proportionnel à la ſomme des produits des molécules multipliées chacune par l'eſpace parcouru, il s'enſuit que le déblai & le remblai étant donnés de figure & de poſition, il n'eſt pas indifférent que telle molécule du déblai ſoit tranſportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diſtribution à faire des molécules du premier dans le ſecond, d'après laquelle la ſomme de ces produits ſera la moindre poſſible, & le prix du tranſport total ſera un *minimum*.

C'eſt la ſolution de cette queſtion que je me propoſe de donner ici. Je diviſerai ce Mémoire en deux parties, dans la première je ſuppoſerai que les déblais & les remblais ſont des aires contenues dans un même plan ; dans le ſecond, je ſuppoſerai que ce ſont des volumes.

PREMIÈRE PARTIE.

Du tranſport des aires planes ſur des aires compriſes dans un même plan.

I.

QUELLE que ſoit la route que doive ſuivre une molécule

Gaspard Monge (1746-1818)

Marginal Spaces

- Polish space X (metrizable with a complete, separable metric)
- e.g. $\mathsf{X} = \mathbb{R}^d$
- probability measure μ on Borel σ-field $\mathcal{B}(\mathsf{X})$
- Similarly, a second space Y (often $=$ X) with probability ν
- $\mathsf{X} =$ source space, $\mathsf{Y} =$ target space

- "Transport" μ to ν by a map $T : \mathsf{X} \to \mathsf{Y}$?
- Idea: move (all) the μ-mass from location $x \in \mathsf{X}$ to location $T(x) \in \mathsf{Y}$

Pushforward

- Let $T : \mathsf{X} \to \mathsf{Y}$ be measurable
- It "pushes μ forward" to a measure $T_\#\mu$ on Y, defined by

$$T_\#\mu(B) := \mu(T^{-1}(B)), \quad B \in \mathcal{B}(\mathsf{Y})$$

  Intuitively: μ-mass at x is sent to $T(x)$
- Equivalently, $\int_{\mathsf{Y}} f \, dT_\#\mu = \int_{\mathsf{X}} f \circ T \, d\mu$ for all f bounded measurable
- In the language of probability, $\nu := T_\#\mu$ is the law of T under μ:

$$T \sim \nu \quad \text{and then} \quad \int f \, d\nu = E_\mu[f(T)]$$

- Example: for $\mu = \sum_i p_i \delta_{x_i}$ we have $T_\#\mu = \sum_i p_i \delta_{T(x_i)}$
- Atoms are always sent to atoms (of at least the same mass)

Monge Optimal Transport Problem

Given:

- Probability spaces (X,μ) and (Y,ν)
- Cost function $c:\mathsf{X}\times\mathsf{Y}\to\mathbb{R}$, continuous, often $c\geq 0$
- Some integrability, often $|c(x,y)|\leq a(x)+b(y)$ with $a\in L^1(\mu)$, $b\in L^1(\nu)$
- In Monge's original problem, $\mathsf{X}=\mathsf{Y}=\mathbb{R}^d$ and $c(x,y)=|y-x|$

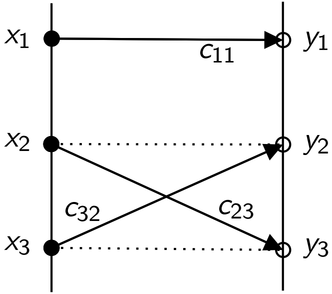
Objective:

- Find a map $T:\mathsf{X}\to\mathsf{Y}$ satisfying $\nu=T_{\#}\mu$ (transport map, Monge map) such as to minimize the total transport cost:

$$\inf_T \int c(x,T(x))\,\mu(dx)$$

Discrete Example: $n \times n$ Assignment Problem

- $\mathsf{X} = \mathsf{Y}$ have n points $\{1, \dots, n\}$ of equal mass
- A transport map T = a permutation $\sigma \in \Sigma(n)$
- $c(x_i, y_j) = c_{ij}$ cost matrix



Difficulties with Monge Formulation

Recall Monge OT problem:

$$\inf_{\{T:\, T_\# \mu = \nu\}} \int c(x, T(x))\, \mu(dx)$$

Difficulties:

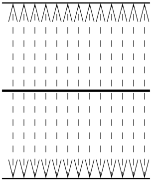
- Nonexistence of a transport map when there are atoms
- Set of transport maps is not convex and not closed in a useful way

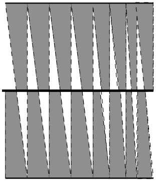
$\rightarrow$ Infimum may not be attained, even in reasonable setting

Example:

- $c(x, y) = |y - x|$ on $\mathbb{R}^2 \times \mathbb{R}^2$
- μ uniform on $[0, 1] \times \{0\}$
- ν uniform on $[0, 1] \times \{-1, 1\}$

(Figures from Villani, 2009)





Splitting Mass

- In the above example, we want to "randomize between two maps" such as to "split mass"
- E.g., mixing maps T_1 and T_2 with an independent coin flip corresponds to a stochastic kernel (cond. distribution given x): $\kappa(x, dy) = \frac{1}{2}\delta_{T_1(x)}(dy) + \frac{1}{2}\delta_{T_2(x)}(dy)$
- More generally, consider couplings or transport plans of μ, ν:

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y}) : \operatorname{proj}^{\mathsf{X}}_{\#} \pi = \mu, \operatorname{proj}^{\mathsf{Y}}_{\#} \pi = \nu\}$$

$=$ measures on $\mathsf{X} \times \mathsf{Y}$ with marginals μ and ν
$=$ joints laws of random vectors (X, Y) with $X \sim \mu$ and $Y \sim \nu$

- Each coupling can be disintegrated into marginal and a kernel $\kappa(x, dy) = \mathrm{Law}(Y|X = x)$:

$$\pi = \mu \otimes \kappa$$

Couplings vs. Transport Maps

- Monge transport T can be seen as a particular coupling π via

$$\pi = \mu \otimes \delta_T = (id, T)_{\#}\mu$$

- Initial example is the average of two such couplings
- Think of $\Pi(\mu, \nu)$ as a (closed) convex hull of the set of transport maps

- Transport plans always exist: product (independent) coupling $\mu \otimes \nu$
- Partial converse to the inclusion: if $\pi \in \Pi(\mu, \nu)$ is concentrated on the graph of a function T, then it is of Monge-type with map T

Monge–Kantorovich Optimal Transport Problem

- Find a coupling such as to minimize the total transport cost:

$$\mathrm{OT}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)\,\pi(dx,dy)$$

- Relaxation of the Monge problem
- $\Pi(\mu,\nu)$ is convex, and weakly compact by Prokhorov ($\pi_n \to \pi$ weakly $:\Longleftrightarrow \int h\,d\pi_n \to \int h\,d\pi$ for all $h \in C_b$)
- $\pi \mapsto \int c(x,y)\,\pi(dx,dy)$ is linear and lsc (if c is lsc and suitably bounded from below)
- Hence MK problem has a solution π_* as soon as $\mathrm{OT}(\mu,\nu) < \infty$
- Will not detail integrability conditions in what follows

Relation of Monge and MK Formulations

When μ is atomless:

- Monge-type couplings are weakly dense in $\Pi(\mu, \nu)$

$\rightarrow$ Value functions of formulations coincide for continuous costs

- In many important examples, minimizer π_* is unique and given by a Monge transport (later), hence relaxation is just a tool
- But not in general, as seen above

For the Assignment Problem:

- Couplings = transport polytope (doubly stochastic matrices)
- Extreme points are the permutations (Birkhoff)

$\rightarrow$ Here minimum is always attained at a Monge-type plan

(This strategy does not work in the continuous case)

Application: Optimal Transport Distances

- A natural way to lift ground cost/distance to space of distributions
- Suppose $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$ (for simplicity)
- Wasserstein (Earth's Movers) distance on (a subset of) $\mathcal{P}(\mathbb{R}^d)$

Examples:

- 1-Wasserstein distance $\mathcal{W}_1(\mu,\nu) = \mathrm{OT}(\mu,\nu)$ for $c(x,y) = |x-y|$:

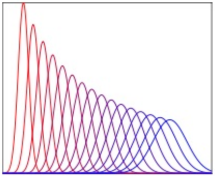
$$\mathcal{W}_1(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|\,\pi(dx,dy)$$

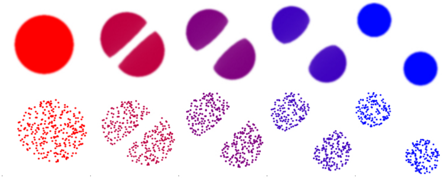
  Also called Kantorovich–Rubinstein distance
- 2-Wasserstein distance $\mathcal{W}_2(\mu,\nu) = \mathrm{OT}(\mu,\nu)^{1/2}$ for $c(x,y) = |x-y|^2$:

$$\mathcal{W}_2(\mu,\nu) = \left[\inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|^2\,\pi(dx,dy)\right]^{1/2}$$

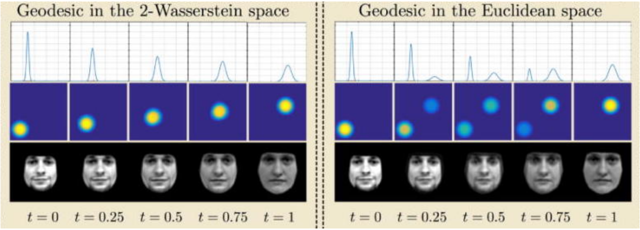
- Can compare discrete and continuous distributions, etc.
- Defines geodesics, averages (barycenters), ..., between distributions

Illustrations: 2-Wasserstein distance





$\mathcal{W}_2$ interpolation (Figures from Peyré&Cuturi 2019)



(Figure from Kolouri et al. 2017)

Outline

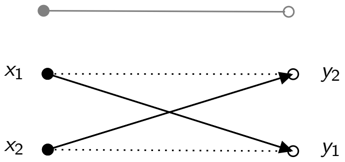
Basic Ideas

Aim: Describe optimal transports

- Suppose $\pi \in \Pi(\mu, \nu)$ is an optimal coupling for c
- Then changing π into another coupling does not improve the cost
- Try to characterize optimality in a tractable way: π is optimal $\iff$ variations of a certain type do not improve π
- Similar to saying that directional derivatives at the minimum should be ≥ 0
- Admissible variations need to preserve the marginals μ, ν

Discrete Case

- finite marginal supports $\operatorname{spt}\mu = \{x_i\}$, $\operatorname{spt}\nu = \{y_j\}$
- $\pi \in \Pi(\mu, \nu)$ is given by its weights $\pi(x_i, y_j)$
- instead of sending mass α from $x_1 \to y_1$ and $x_2 \to y_2$, permute destinations: $x_1 \to y_2$ and $x_2 \to y_1$



- This preserves the marginals: admissible variation

Discrete Case (cont.)

- Same for $k \in \mathbb{N}$ pairs (x_i, y_i) and a permutation to $(x_i, y_{\sigma(i)})$
- Can pick $\alpha > 0$ small
- OK as long as $\pi(x_i, y_i) \geq \alpha$; i.e., $(x_i, y_i) \in \operatorname{spt} \pi$
  but not feasible if $\pi(x_i, y_i) = 0$
- We get: if π is optimal and $k \in \mathbb{N}$ and $(x_i, y_i)_{1 \leq i \leq k} \in \operatorname{spt} \pi$, then

$$\sum_{i=1}^{k} c(x_i, y_i) \leq \sum_{i=1}^{k} c(x_i, y_{i+1}) \quad \text{where} \quad y_{k+1} := y_1$$

  - This only used the permutation $\sigma : i \mapsto i + 1$, but already covers all permutations, as (x_i, y_i) can be relabeled arbitrarily
  - k is finite but arbitrary

c-Cyclical Monotonicity

Definition

A set $\Gamma \subset X \times Y$ is called c-cyclically monotone if for all $k \in \mathbb{N}$ and all $(x_i, y_i)_{1 \le i \le k} \in \Gamma$,

$$\sum_{i=1}^{k} c(x_i, y_i) \le \sum_{i=1}^{k} c(x_i, y_{i+1}) \quad \text{where} \quad y_{k+1} := y_1$$

A measure is called c-cyclically monotone if it is concentrated on such a set.

- For $c(x, y) = -\langle x, y\rangle$ on $\mathbb{R}^d$, this is cyclical monotonicity in the sense of convex analysis
- A subset of a c-cyclically monotone set is again c-cyclically monotone
- If c is continuous and Γ is c-cyclically monotone, the closure of Γ is still c-cyclically monotone

$\rightarrow$ If c is continuous, a measure π is c-cyclically monotone iff $\operatorname{spt} \pi$ is

c-Cyclical Monotonicity and Optimality

- In the discrete case: this type of variation covers all possible directions, hence by convexity:

$$\pi \in \Pi(\mu, \nu) \text{ is optimal iff } \operatorname{spt} \pi \text{ is } c\text{-cyclically monotone}$$

- General case is less obvious: the variation is "infinitesimal" and only uses finite rearrangements. Nevertheless, the result turns out to be the same:

Theorem

Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be continuous and $\mathsf{OT}(\mu, \nu) < \infty$. Then $\pi \in \Pi(\mu, \nu)$ is optimal $\iff \operatorname{spt} \pi$ is c-cyclically monotone.

- The shape of the support is enough to determine optimality:

  Just need to use "good roads," no need to look into how much is transported on each road

$\Rightarrow$ If not c-cyclically monotone, blow up points to collect mass ...

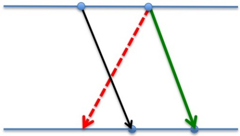
$\Leftarrow$ More difficult. Later: proof via duality

Application 1 — Scalar Case: μ, ν on $\mathbb{R}$

Theorem: Let c satisfy the Spence–Mirrlees condition $c_{xy} < 0$. Then the optimal $\pi \in \Pi(\mu, \nu)$ is unique, given by the Hoeffding–Fréchet Coupling:

- π is characterized by monotonicity:

$$\text{if } (x, y), (x', y') \in \operatorname{spt} \pi \text{ and if } x < x', \text{ then } y \leq y'.$$



- $\pi = ((F_\mu)^{-1}, (F_\nu)^{-1})_\# \, \mathrm{Leb}\,|_{[0,1]}$
- If μ is atomless, π is of Monge type with $T = (F_\nu)^{-1} \circ F_\mu$.

Examples: $c(x, y) = |y - x|^2$ (or $c(x, y) = -xy$, maximizing correlation), $c(x, y) = h(y - x)$ with h strictly convex

Application 2 — Stability of Optimal Transport Plans

- Consider a sequence of marginals (μ_n, ν_n) on the same $\mathsf{X} \times \mathsf{Y}$
- c continuous (!)
- Let $\pi_n \in \Pi(\mu_n, \nu_n)$ and $\pi_n \to \pi$ weakly
- If π_n is c-cyclically monotone for all n, then so is π

  Proof: Let $(x_i, y_i)_{1 \leq i \leq k} \subset \operatorname{spt} \pi$. Then there are $(x_i^n, y_i^n) \in \operatorname{spt} \pi_n$ with $(x_i^n, y_i^n) \to (x_i, y_i)$. For each n,

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

  As c is continuous, we can pass to the limit.

Assuming that $\mathrm{OT}(\mu_n, \nu_n)$ remains bounded as $n \to \infty$, we conclude:

Theorem: Weak limits of optimal transport plans are again optimal.

Outline

Formal Derivation

- Introduce Lagrange multipliers for the marginal constraints μ, ν
- Write $\mathcal{M} = \{\text{nonnegative measures on } \mathsf{X} \times \mathsf{Y}\}$
- $\pi \in \mathcal{M}$ has marginal μ iff $\int \varphi(x)\, d\pi = \int \varphi\, d\mu$ for all (suitably integrable) functions $\varphi : \mathsf{X} \to \mathbb{R}$, and similarly for ν
- Hence

$$\sup_{\varphi} \left[\int \varphi\, d\mu - \int \varphi(x)\, d\pi \right] = \begin{cases} 0 & \text{if } \operatorname{proj}^{\mathsf{X}}_{\#} \pi = \mu, \\ \infty & \text{if } \operatorname{proj}^{\mathsf{X}}_{\#} \pi \neq \mu \end{cases}$$

Formal Derivation

- Write $\varphi \oplus \psi$ for the function $(x, y) \mapsto \varphi(x) + \psi(y)$
- Formally derive duality by interchanging inf and sup (minimax)

$$
\begin{aligned}
\inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi &= \inf_{\pi \in \mathcal{M}} \left[\int c \, d\pi + \infty 1_{\pi \notin \Pi(\mu,\nu)} \right] \\
&= \inf_{\pi \in \mathcal{M}} \sup_{\varphi,\psi} \left[\int \varphi \, d\mu + \int \psi \, d\nu + \int (c - \varphi \oplus \psi) \, d\pi \right] \\
(?) &= \sup_{\varphi,\psi} \inf_{\pi \in \mathcal{M}} \left[\int \varphi \, d\mu + \int \psi \, d\nu + \int (c - \varphi \oplus \psi) \, d\pi \right] \\
&= \sup_{\varphi,\psi} \left[\int \varphi \, d\mu + \int \psi \, d\nu - \infty 1_{(\{\varphi \oplus \psi > c\} \neq \emptyset)} \right] \\
&= \sup_{\varphi,\psi:\, \varphi \oplus \psi \leq c} \left[\int \varphi \, d\mu + \int \psi \, d\nu \right] \qquad \leftarrow \text{dual problem}
\end{aligned}
$$

- Dual can be interpreted economically as "prices"

Weak Duality

- Dual domain $\mathcal{D} := \{(\varphi, \psi) : \varphi \in L^1(\mu), \psi \in L^1(\nu) : \varphi \oplus \psi \leq c\}$
- $(\varphi, \psi) \in \mathcal{D} \implies \int c\, d\pi \geq \int \varphi \oplus \psi\, d\pi = \int \varphi\, d\mu + \int \psi\, d\nu$
- Hence we have the "weak duality"

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi \geq \sup_{(\varphi,\psi) \in \mathcal{D}} \left[\int \varphi\, d\mu + \int \psi\, d\nu \right]$$

- Opposite inequality also holds (but not straightforward):

Theorem

Let c be lsc (and sufficiently integrable). There is no duality gap; i.e.,

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi = \sup_{(\varphi,\psi) \in \mathcal{D}} \left[\int \varphi\, d\mu + \int \psi\, d\nu \right],$$

and the sup is attained: there are dual optimizers $(\varphi_*, \psi_*) \in \mathcal{D}$.

- Let's take this for granted for the moment

Consequences of Duality

Recall $\inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi = \int \varphi_* \, d\mu + \int \psi_* \, d\nu$ and note

$$\mathsf{X} \times \mathsf{Y} = \{c \geq \varphi_* \oplus \psi_*\} = \{c = \varphi_* \oplus \psi_*\} \cup \{c > \varphi_* \oplus \psi_*\}$$

Proposition

If $\pi \in \Pi(\mu,\nu)$ is optimal for the primal problem and $(\varphi_*, \psi_*) \in \mathcal{D}$ is optimal for the dual, then π is supported on $\Gamma := \{c = \varphi_* \oplus \psi_*\}$.

Application: if one can show that Γ is a graph, once can conclude that any optimal π is given by a Monge map.

The converse is also true, and does not require a priori optimality:

Proposition ("Verification Theorem")

Let $\pi \in \Pi(\mu,\nu)$ and $(\varphi,\psi) \in \mathcal{D}$. If $\pi\{c = \varphi \oplus \psi\} = 1$, then π and (φ,ψ) are optimal for the primal/dual problem. Moreover, there is no duality gap.

c-Cyclical Monotonicity and Dual Potentials

- Given $(\varphi, \psi) \in \mathcal{D}$, the "contact set" $\Gamma := \{\varphi \oplus \psi = c\}$ is c-cyclically monotone: given $(x_i, y_i)_{1 \le i \le k} \in \Gamma$,

$$\begin{aligned}\sum_{i=1}^{k} c(x_i, y_i) &= \sum_{i=1}^{k} \varphi(x_i) + \psi(y_i) \\ &= \sum_{i=1}^{k} \varphi(x_i) + \psi(y_{i+1}) \le \sum_{i=1}^{k} c(x_i, y_{i+1})\end{aligned}$$

  where the last inequality used that $\varphi \oplus \psi \le c$ everywhere on $\mathsf{X} \times \mathsf{Y}$

- We want a converse: if $\Gamma \subset \mathsf{X} \times \mathsf{Y}$ is c-cyclically monotone, find functions $\varphi \oplus \psi \le c$ with such that $\Gamma \subset \{\varphi \oplus \psi = c\}$.
- Let's get rid of ψ first . . .

c-Conjugates

- Let $(\varphi, \psi) \in \mathcal{D}$; i.e., $\varphi \oplus \psi \leq c$
- In the dual problem, bigger is better: if $\tilde{\psi} \geq \psi$ still satisfies $\varphi \oplus \tilde{\psi} \leq c$, then $(\varphi, \tilde{\psi})$ is at least as good as (φ, ψ)
- Given φ, we can find the largest such $\tilde{\psi}$:

$$\tilde{\psi}(y) = \varphi^c(y) := \inf_{x \in X} [c(x,y) - \varphi(x)]$$

  φ^c is called the c-transform or c-conjugate of φ
- In searching for dual solutions, we can focus on φ alone: if $(\varphi, \psi) \in \mathcal{D}$, then $(\varphi, \varphi^c) \in \mathcal{D}$ is at least as good
- If $(\varphi, \psi) \in \mathcal{D}$ is a dual solution, we must have $\psi = \varphi^c$ ν-a.s.

- Hence "dual solution" sometimes refers to the single function φ
- φ often called a Kantorovich potential in this context

c-Concavity

- Symmetrically:

$$\psi^c(x) := \inf_{y \in \mathsf{Y}} [c(x,y) - \psi(y)]$$

- Improve $(\varphi, \psi) \to (\varphi, \varphi^c) \to (\varphi^{cc}, \varphi^c) \ldots$ That's it: $\varphi^{ccc} = \varphi^c$
- ψ is called c-concave if $\psi = \varphi^c$ for some function φ
- φ is c-concave iff $\varphi = \varphi^{cc}$

Connection to Convex Analysis:

- Legendre transform $f^*(y) := \sup_x [\langle x, y \rangle - f(x)]$
- f lsc is convex iff $f = f^{**}$
- For the particular cost $c(x,y) = -\langle x, y \rangle$ we have $-(\varphi^c) = (-\varphi)^*$

Summary

- We can focus on c-concave functions in $\mathcal{D}$

$\rightarrow$ With conditions on c, this gives some regularity

- We can focus on φ alone, then choose $\psi := \varphi^c$
- Dual problem often written as

$$\sup_{\varphi \in L^1(\mu)} \left[\int \varphi \, d\mu + \int \varphi^c \, d\nu \right]$$

Example: Distance Cost

Let $X = Y$ with metric d and $c(x, y) = d(x, y)$ the distance cost. Then

- φ is c-concave $\iff \mathrm{Lip}(\varphi) \leq 1$
- If φ is c-concave, then $\varphi^c = -\varphi$

Proof:

$\Rightarrow$ If $\varphi = \psi^c$ for some function ψ, then

$$\begin{aligned}\varphi(x) - \varphi(x') &= \inf_{y \in Y}[d(x, y) - \psi(y)] - \inf_{y \in Y}[d(x', y) - \psi(y)] \\ &\leq \sup_{y \in Y} |d(x, y) - d(x', y)| \leq d(x, x')\end{aligned}$$

$\Leftarrow$ Conversely, if φ is 1-Lipschitz, then $\varphi(y) - \varphi(x) \leq d(x, y)$, hence

$$-\varphi(x) = \inf_{y \in Y}[d(x, y) - \varphi(y)] = \varphi^c(x)$$

and φ is c-concave

Example cont'd: Kantorovich–Rubinstein Formula

Recall: for distance cost $c(x, y) = d(x, y)$, the optimal transport cost $\mathrm{OT}(\mu, \nu) \equiv \mathcal{W}_1(\mu, \nu)$ is the 1-Wasserstein distance. Apply duality result:

Proposition (Kantorovich–Rubinstein Formula)

Let μ, ν have finite first moments. Then

$$\mathcal{W}_1(\mu, \nu) = \sup_{\mathrm{Lip}(\varphi) \leq 1} \left[\int_{\mathsf{X}} \varphi \, d\mu - \int_{\mathsf{X}} \varphi \, d\nu \right].$$

- Useful to estimate 1-Wasserstein distances

Outline

c-Superdifferential and Rockafellar's Theorem

- Recall the contact set $\{\varphi \oplus \psi = c\}$
- Given φ, its c-superdifferential is the set $\partial^c\varphi \subset \mathsf{X} \times \mathsf{Y}$ given by

$$\partial^c\varphi = \{(x, y) : \varphi(x) + \varphi^c(y) = c(x, y)\}$$

  I.e., the contact set for $\psi := \varphi^c$

We have seen that $\partial^c\varphi$ is c-cyclically monotone. Conversely:

Theorem

Let $\Gamma \subset \mathsf{X} \times \mathsf{Y}$ be c-cyclically monotone. Then there exists a c-concave function φ such that $\Gamma \subset \partial^c\varphi$.

The proof by a direct algebraic construction.

- One can also define the section of $\partial^c\varphi$ at x: $\partial^c\varphi(x) = \{y : \varphi(x) + \varphi^c(y) = c(x, y)\}$
- For $c(x, y) = -\langle x, y\rangle$, the theorem is Rockafellar's theorem of convex analysis and $\partial^c\varphi(x)$ is the usual superdifferential at x

Construction of φ

- Let $\Gamma \subset X \times Y$ be c-cyclically monotone
- Note that the potential φ can be determined only up to a constant
- Fix some $(x_0, y_0) \in \Gamma$; we will have $\varphi(x_0) = 0$
- For $x \in \operatorname{proj}^X \Gamma$, define

$$\begin{aligned}\varphi(x) := \inf_{(x_i,y_i)\in\Gamma,\, 1\le i\le k,\, k\in\mathbb{N}} \quad & [c(x, y_k) - c(x_k, y_k)] \\ & + [c(x_k, y_{k-1}) - c(x_{k-1}, y_{k-1})] + \cdots \\ & + [c(x_1, y_0) - c(x_0, y_0)]\end{aligned}$$

- Then indeed $\varphi(x_0) = 0$ by c-cyclical monotonicity
- Shipping cost from x should be cheaper then transporting from x to x_0 and then shipping for free. In equilibrium, invariant between infimum over ways of doing that and shipping from x.

Bringing it Home

1. There exists $\pi \in \Pi(\mu, \nu)$ with $\operatorname{spt} \pi$ c-cyclically monotone
   Proof by discrete approximation and stability
   (or by control theoretic arguments from optimal π)

2. Construct c-concave function φ with $\operatorname{spt} \pi \subset \partial^c \varphi$

$\rightarrow$ π is supported in the contact set of (φ, φ^c)

$\rightarrow$ (φ, φ^c) is a dual optimizer, π is optimal transport, and no duality gap

3. More generally, any optimal $\tilde{\pi} \in \Pi(\mu, \nu)$ is concentrated on $\partial^c \varphi$, in particular is c-cyclically monotone.

(We did skip some details here. . .)

Regularity of c-Concave Functions implies Monge Solutions

- Additional knowledge on the structure of $\partial^c \varphi$ yields insights about optimizers

Suppose we know: if φ is c-concave, then $\partial^c \varphi(x)$ is a singleton for μ-a.e. x

- Let $\pi \in \Pi(\mu, \nu)$ be an optimal transport plan
- Then $\operatorname{spt} \pi$ is c-cyclically monotone, hence $\operatorname{spt} \pi \subset \partial^c \varphi$ for some c-concave φ
- By the assumption, $T(x) := \partial^c \varphi(x)$ is μ-a.e. well defined
- We must have $\pi = \mu \otimes_x \delta_{T(x)}$; i.e., π is given by Monge transport T

Convexity Argument for Uniqueness

Monge implies Uniqueness:

- Suppose we know that any optimal plan is of Monge-type
- Then we automatically have uniqueness of the optimizer

- Let $\pi_1, \pi_2 \in \Pi(\mu, \nu)$ be optimal
- Then $\pi := (\pi_1 + \pi_2)/2$ is also optimal
- If π_i is given by T_i, then $\pi = \mu \otimes (\frac{1}{2}\delta_{T_1(x)} + \frac{1}{2}\delta_{T_2(x)})$
- As we know that π is of Monge type, we must have $T_1 = T_2$ μ-a.s., and then also $\pi_1 = \pi_2$

Summary

Suppose we know (a property of c and μ):

$$\varphi \text{ any } c\text{-concave function} \Longrightarrow \partial^c\varphi(x) \text{ is a singleton for } \mu\text{-a.e. } x$$

(i.e., $\partial^c\varphi$ is the graph of a function, up to a nullset.)

Then:

- There exists a unique optimal transport plan
- It is given by a Monge map T
- $T(x) = \partial^c\varphi(x)$ for μ-a.e. x, for some c-concave function φ

Important Example: Quadratic Cost

Apply the above to quadratic cost $c(x,y) = |y-x|^2/2$ on $\mathbb{R}^d$:

- In general, costs $c(x,y)$ and $c(x,y) + a(x) + b(y)$ are "equivalent" (if a, b integrable):
  - same optimal transports
  - value functions shifted by a constant
  - c-concave functions $\varphi(x)$ shifted by $a(x)$
- $c(x,y) = (y-x)^2/2$ equivalent to $-\langle x, y\rangle$, links to regular convexity:
  - φ is c-concave iff $\bar{\varphi}(x) := |x|^2/2 - \varphi(x)$ is convex (and lsc)
  - $\partial^c \varphi(x) = \partial^- \bar{\varphi}(x)$, the subdifferential in the usual sense
  - In particular $\partial^c \varphi(x) = \{\nabla \bar{\varphi}(x)\}$ at points x of differentiability

- Recall: finite convex function is Leb-a.e. differentiable
- Assumption: $\mu \ll \text{Leb}$

$\rightarrow$ Any c-concave φ is μ-a.e. differentiable with $\partial^c \varphi(x) = \{\nabla \bar{\varphi}(x)\}$ μ-a.e.

Results Applied to Quadratic Cost

Theorem (Brenier's Theorem)

Let $c(x,y) = |y-x|^2/2$ *on* $\mathbb{R}^d$, *let* μ, ν *have finite second moments, and let* $\mu \ll \mathrm{Leb}$.

- *There exists a unique optimal transport plan* $\pi \in \Pi(\mu, \nu)$
- *It is given by a transport map* T
- $T(x) = \nabla\bar{\varphi}(x)$ μ*-a.e. for some convex function* $\bar{\varphi}$
- *The gradient of* $\bar{\varphi}$ *is* μ*-a.e. uniquely determined*

Remarks

- For suitable costs other than quadratic, "c-convex analysis" yields results similar to classical convex analysis
- Comparable results e.g. for $c(x,y) = |y-x|^p$ with $p \in (1,\infty)$
- Quadratic cost has the most detailed results and links to other fields
- Monge's $c(x,y) = |y-x|$ is qualitatively different

From Change of Variables to Monge–Ampère Equation

- Let $\mu(dx) = f(x)\,dx$ and $\nu(dy) = g(y)\,dy$ in $\mathbb{R}^d$
- Supported by open sets $F, G \subset \mathbb{R}^d$
- Given a diffeomorphism $T : F \to G$, the condition $T_{\#}\mu = \nu$ is equivalent to the following for any test function h

$$\begin{aligned}
\int_F h(T(x))f(x)\,dx &= \int_F h(T(x))\,\mu(dx) \\
&= \int_G h(y)\,\nu(dy) \\
&= \int_G h(y)g(y)\,dy \\
&= \int_F h(T(x))g(T(x))|\det(\nabla T(x))|\,dx
\end{aligned}$$

- Hence

$$f(x) = g(T(x))|\det(\nabla T(x))|, \quad x \in F$$

Monge–Ampère Equation

- Last slide: $f(x) = g(T(x))|\det(\nabla T(x))|$
- This is for any (smooth) transport map, optimal or not
- Now suppose $T = \nabla\bar{\varphi}$ is Brenier's map
- Get Monge–Ampère equation for $u := \bar{\varphi}$,

$$f(x) = g(\nabla u(x))\det(\nabla^2 u(x)), \quad \text{a.e. } x \in F$$

- In the class of convex functions u (hence $\nabla^2 u$ exists a.e.)
- Regularity theory of this PDE yields regularity for Kantorovich potential and transport map

Caffarelli's Regularity Theorem

Suppose:

- F, G bounded, connected, open
- G is convex
- f, g are bounded away from zero and infinity on F, G

Then

$$\bar{\varphi} \in C^{1,\beta}(F), \quad \text{hence} \quad T \in C^{0,\beta}(F)$$

- If in addition $f, g \in C^{k,\alpha}$, then

$$\bar{\varphi} \in C^{k+2,\alpha}(F), \quad \text{hence} \quad T \in C^{k+1,\alpha}(F)$$

"Potential gains two degrees of regularity vs. densities"

- If F, G are uniformly convex, that also holds for $\overline{F}$
- Various extension (unbounded domains, local conditions ...), convexity of G is important

Monge's Original Problem: Distance Cost on $\mathbb{R}^d$

- Cost $|y - x|$ is not strictly convex, not differentiable at $x = y$
- Leads to non-uniqueness in primal and dual problems
- E.g., on $\mathbb{R}$, if $\sup \operatorname{spt} \mu \leq \inf \operatorname{spt} \nu$, any coupling is optimal

Theorem

Let $c(x, y) = |y - x|$ on $\mathbb{R}^d$, let μ and ν have finite first moments, and let $\mu \ll \mathrm{Leb}$. Then there exists an optimal transport plan of Monge type.

- Typically the set Π_* of all optimal transport plans is not a singleton, contains non-deterministic couplings
- One way to pick a Monge map is to solve a secondary transport problem with a strictly convex cost over Π_*

Outline

Wasserstein Distances

- Let $\mathsf{X} = \mathsf{Y}$ have metric d
- Fix $p \in [1, \infty)$
- Fix any $x_0 \in \mathsf{X}$. Call $\int d(x_0, x)^p \, \mu(dx)$ the p-th moment of μ.
  In Euclidean space, $x_0 = 0$ yields the usual $\int |x|^p \, \mu(dx)$
- Results do not depend on choice of x_0
- $\mathcal{P}_p(\mathsf{X}) :=$ set of probability measures with finite p-th moment

Definition

For $\mu, \nu \in \mathcal{P}_p(\mathsf{X})$, the p-Wasserstein distance is

$$\mathcal{W}_p(\mu, \nu) = \left[\inf_{\pi \in \Pi(\mu,\nu)} \int d(x, y)^p \, \pi(dx, dy) \right]^{1/p}$$

I.e., $\mathcal{W}_p^p(\mu, \nu) = \mathrm{OT}(\mu, \nu)$ for cost $c(x, y) = d(x, y)^p$

- Note $\mathcal{W}_p(\delta_x, \delta_y) = d(x, y)$

Properties of Wasserstein Distance

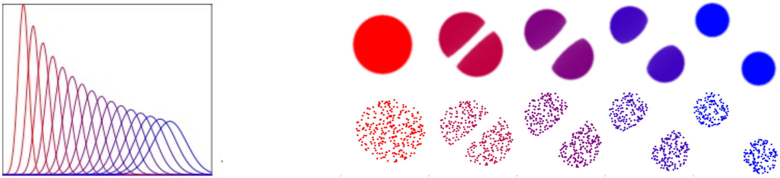
- $\mathcal{W}_p$ is a metric on $\mathcal{P}_p(\mathrm{X})$
- Under this metric, $\mathcal{P}_p(\mathrm{X})$ is complete and separable
- If X is compact, so is $\mathcal{P}_p(\mathrm{X})$
- If d is bounded on X, convergence in $\mathcal{W}_p$ is the same as weak convergence. More generally:

Proposition

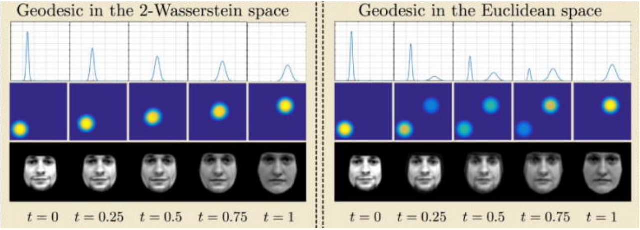
Let $\mu_n, \mu \in \mathcal{P}_p(\mathrm{X})$. The following are equivalent:

- $\mathcal{W}_p(\mu_n, \mu) \to 0$
- *$\mu_n \to \mu$ weakly (wrt. $C_b(\mathrm{X})$) and the p-th moments converge,* $\int d(x_0, x)^p\, \mu_n(dx) \to \int d(x_0, x)^p\, \mu(dx)$
- *Weak convergence wrt. cont. test functions with growth of order p:* $\int f(x)\, \mu_n(dx) \to \int f(x)\, \mu(dx)$ *for all* $f \in C(\mathrm{X})$ *with* $|f(x)| \leq const[1 + d(x_0, x)^p]$

Recall the Illustrations for $\mathcal{W}_2$ . . .



$\mathcal{W}_2$ interpolation (Figures from Peyré&Cuturi 2019)



(Figure from Kolouri et al. 2017)

Displacement Interpolation for $\mathcal{W}_2$

- Let $\mathsf{X} \subset \mathbb{R}^d$ convex
- Consider $\mu_0, \mu_1 \in \mathcal{P}_2(\mathsf{X})$ with $\mu_0, \mu_1 \ll \text{Leb}$
- Natural way to interpolate from μ_0 to μ_1 in the geometry of $\mathcal{W}_2$?

Proposition

Let $T : \mathsf{X} \to \mathsf{X}$ be the optimal map for $\mathcal{W}_2(\mu_0, \mu_1)$ (Brenier map). For $t \in [0, 1]$, let

$$T_t := (1-t)id + tT \quad \text{and} \quad \mu_t := (T_t)_\# \mu_0.$$

Then $(\mu_t)_{t\in[0,1]}$ is a constant-speed geodesic between μ_0 and μ_1 wrt. $\mathcal{W}_2$.

That is, $\mathcal{W}_2(\mu_s, \mu_t) = |t-s|\mathcal{W}_2(\mu_0, \mu_1)$ for all $s, t \in [0, 1]$.

- Trajectory $t \mapsto T_t(x)$ is a straight line in $\mathbb{R}^d$
- Similar result for $\mathcal{W}_p$