Outline

Part 1: Theoretical foundations

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

Mém. de l'Académie Royale
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. Monge.

Loasqu'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, et le nom de Remblai à l'espace qu'elles doivent occuper après le transport.

Figure from Villani (2008).
Origins of OT: Monge (1781)

Modern formulation:

Given

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

Solve

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

and find a minimizer $T^*$, where $T_{#} \mu(A) = \mu(T^{-1}(A))$ 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)

Problem studied by Kantorovich:

Given

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

Solve

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

and find a minimizer $\pi^*$, where $\Pi$ is the set of couplings of $\mu$ and $\nu$.

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 $\mathcal{X}$.

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

Monge problem (MP):

\[ \min_{T \# \mu = \nu} \int_X c(x, T(x)) \, d\mu(x). \]

Kantorovich problem (KP):

\[ \min_{\pi \in \Pi(\mu, \nu)} \int_{X^2} c(x, y) \, d\pi(x, y). \]

- If \( T \) is feasible for (MP), it also gives rise to a feasible solution for (KP): \( \pi = (\text{Id}, T) \# \mu. \)

- The set of maps \( \{ T : T \# \mu = \nu \} \) 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

Primal problem:

\[
\min_{\pi \in \Pi(\mu, \nu)} \int_{X^2} c(x, y) d\pi(x, y).
\]

▶ 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

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

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 $\pi \mapsto \int c_n d\pi$ is continuous.
- Show $K(\pi) = \sup_n \int c_n d\pi$ by monotone convergence, giving lower semicontinuity of $K(\pi)$.

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

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

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

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

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

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

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

If we can interchange* the “min” and the “sup”, we get

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

*Needs formal justification.
Duality

Note that

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

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

In turn, we get the dual problem (DP)

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

Duality: Intuition

We hire a logistics company to transport “déblais” distributed as $\mu$ to “remblais” distributed as $\nu$. 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 $\int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{X}} \psi \, d\nu$. 
Duality

Dual problem (DP):

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

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

Definition

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

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

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

Many important properties follow from cyclical monotonicity.

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

- $\pi^*$ is unique,
- $\pi^* = (\text{Id}, T) \# \mu$, i.e. (MP) and (KP) coincide,
- $T(x) = \nabla \varphi^*(x)$, where $\varphi^*$ 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 \(\mu, \nu \in P_p(\mathcal{X})\),

\[
W_p(\mu, \nu) = \left( \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X}^2} d^p(x, y) \, d\pi \right)^{1/p}
\]

defines the \(p\)-Wasserstein distance between \(\mu\) and \(\nu\).

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

Defines a proper distance on \(P_p(\mathcal{X})\):

- \(W_p(\mu, \nu) \geq 0\), with equality if and only if \(\mu = \nu\).
- \(W_p(\mu, \nu) = W_p(\nu, \mu)\).
- \(W_p(\mu, \nu) \leq W_p(\mu, \rho) + W_p(\rho, \nu)\).

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.


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 $\mu$ and $\nu$.

Figure from Cuturi and Solomon (2017).
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 \( \rho_k \) converges weakly to \( \rho \) in \( \mathcal{P}_p(\mathcal{X}) \).


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 \( \mu \) 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

\[
\begin{align*}
\triangleright (F_{\mu}^{-1})_{\#} \text{Unif}[0, 1] &= \mu, \\
\triangleright (F_{\mu})_{\#} \mu &= \text{Unif}[0, 1], \text{ provided } \mu \ll \text{Leb}.
\end{align*}
\]

Hence, if \( \mu \ll \text{Leb} \), then \( T = F_{\nu}^{-1} \circ F_{\mu} \) is such that \( T_{\#} \mu = \nu \).

**Exercise:** Show that \( T \) solves (MP) and (KP), and that

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

Questions?