MATH 8803: Optimal Transport: Theory and Applications

Rui Gong

September 28, 2025

Acknowledgements

These notes are based on the MATH 8803 lectures given by Professor *Tobias Ried* in Fall 2025 at Georgia Institute of Technology.

Contents

1 Introduction

1.1 Monge's Original Formulation of Optimal Transport

Consider a measure μ, another measure ν, and x, y in the supports of μ and ν respectively, *what is the optimal way of moving x to y?*

1. Pile and hole
   - Pile and hole should have the same volume $\rightarrow$ normalize to 1.
   - modern way of thinking about pile and hole: probability measure on some metric space X and Y respectively, $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$. It could be point cloud: $\mu = \sum_i \alpha_i \delta_{x_i}$; or continuous densities: $\mu(dx) = f(x)dx$.
2. Transport pile to hole region: Transport described by a map $T : X \to Y$. Notice that T may be discontinuous. We need T to be measurable.
3. Transport Cost: $c : X \times Y \to [0, \infty) \cup \{\infty\}$, where $c(x, y)$ represents the cost of moving one unit of mass from x to y (how it is transported does not matter). Implicit Assumption: cost only depends on initial and final. Typical cost: $c(x, y) = |x - y|$; $c(x, y) = |x - y|^2$; $c(x, y) = \frac{1}{|x-y|}$.
4. Filling the hole completely: $\mu(T^{-1}(B)) = \nu(B)$ for every measurable $B \subseteq Y$.

Definition 1.1: Push-Forward Measure

Let $\mu \in \mathcal{P}(X)$ be a probability measure on X, $T : X \to Y$ be a measurable map between metric space X, Y. Then push-forward (or image) measure *of μ under T is the measure $T_{\#}\mu$ on Y defined by $\mu(T^{-1}(B)) = T_{\#}\mu(B)$ for every $B \subseteq Y$ measurable.*

A little bit of functional analysis

- $C_b(X)$: the Banach space of bounded continuous function on X endowed with the norm $\|f\|_\infty := \sup_{x \in X} |f(x)|$.
- $C_o(X) \subseteq C_b(X)$: closed subspace (w.r.t. $\|\cdot\|_\infty$), which is the space of continuous functions vanishing at ∞: $f \in C_o(X)$ if $f \in C_b(X)$ and for every $\epsilon > 0$, there exists a compact set $K_\epsilon \subseteq X$, such that $|f| < \epsilon$ on $X \setminus K_\epsilon$.
- $\mathcal{M}(X)$: space of finite signed measures on X. $\lambda \in \mathcal{M}(X)$ if
  (a) $\lambda(A) \in \mathbb{R}$ for any (Borel) measurable $A \subseteq X$.
  (b) for every countable disjoint union $A = \cup_{i \in \mathbb{N}} A_i$, $A_i \cap A_j = \emptyset$ for $i \neq j$, these holds
  - $\sum_{i \in \mathbb{N}} |\lambda(A_i)| < \infty$
  - $\sum_{i \in \mathbb{N}} \lambda(A_i) = \lambda(A)$.

  To every $\lambda \in \mathcal{M}(X)$, we can associate a unique non-negatable measure $|\lambda| \in \mathcal{M}_+(X)$ via $|\lambda|(A) := \sup\{\sum_{i \in \mathbb{N}} |\lambda(A_i)| : A = \cup_{i \in \mathbb{N}} A_i, A_i \cap A_j = \emptyset \text{ for } i \neq j\}$, the total variation measure of λ. $\|\lambda\| := |\lambda|(X)$ is a norm on $\mathcal{M}(X)$.

Theorem 1.2: Riesz Representation Theorem

Suppose X is separable, and locally compact. Then $\mathcal{M}(X) \cong [C_o(X)]^*$ (the dual space of $C_o(X)$). That is, every continuous linear functional $L : C_o(X) \to \mathbb{R}$ is represented in a unique way by an element of $\mathcal{M}(X)$, i.e., there exists a unique measure $\mu_L \in \mathcal{M}(X)$ s.t. $L(\varphi) = \int_X \varphi d\mu_L$.

Remark. Consider a special case of $T_{\#}\mu = \nu$; assume that T is a C^1-diffeomorphism between X, Y and $X, Y \subseteq \mathbb{R}^d$ open, and that $\mu(dx) = f(x)dx$, $\nu(dy) = g(y)dy$. Then for any $B \subseteq \mathbb{R}^d$ measurable:

$$(T_{\#}\mu)(B) = \mu(T^{-1}(B)) = \int_{\mathbb{R}^d} \mathbb{1}_{T^{-1}(B)}(x) f(x) dx = \int_{T^{-1}(B)} f(x) dx.$$

Write $y = T(x), dy = |\det DT(x)|dx$, we can write

$$(T_{\#}\mu)(B) = \nu(B) = \int_B g(y)dy = \int_{T^{-1}(B)} g(T(x))|\det DT(x)|dx,$$

which implies $f(x) = g(T(x))|\det DT(x)|$ for almost every $x \in X$ (technically remark: $f \geq \alpha$ for some $\alpha > 0$ on X).

Reminder: $T_{\#}\mu = \nu$ means:

1. $(T_{\#}\mu)(B) = \mu(T^{-1}(B)) = \nu(B)$ for any measurable subset $B \subseteq Y$.
2. $\int_y \varphi d(T_{\#}\mu) = \int_X \varphi \circ T d\mu = \int_y \varphi d\nu$, $\forall \varphi \in C_o(y)$.

A quick remark on the change of variables formula:

$$\int_y \varphi d(T_{\#}\mu) = \int_X \varphi(T(x))\mu(dx).$$

Definition 1.3: Monge's Optimal Tranposrt Problem

Given $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$,

$$\min\ I[T] = \int_X c(x, T(x))\mu(dx) \qquad (\mathcal{M})$$

over all transport maps $T : X \to Y$ (i.e., all measurable maps from X to Y such that $T_{\#}\mu = \nu$).

Remark.

- I is a highly nonlinear functional of T subject to the nonlinear constraint $T_{\#}\mu = \nu$.
- Functional relatively simple: depends only locally on T (or its pointwise values)
  - no coupling between different values of T.
  - without constraint could just minimize pointwise, i.e. find minimum $y_{\min}(x)$ of $y \mapsto c(x, y)$ for each x and get $T(x) = y_{\min}(x)$.
- constraint complicated: nonlocal, couples values of T. If we could restrict to smooth diffeomorphisms, problem requires solving highly nonlinear PDE. Also, it is not even clear whether T such that $T_{\#}\mu = \nu$ exists for given μ, ν.

1.2 The Kantorovich Optimal Transport Problem

Comparing to Monge's OT problem, Kantorovich OT problem allows measure splitting, so we are looking for probability measure on $X \times Y$.

- Pile and hole: $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$.
- Transport: probability measure $\gamma \in \mathcal{P}(X \times Y)$ (transport plan).

  $$\gamma(A \times B) = \int_{A \times B} \gamma(dxdy)$$

  is the amount of mass moved from measurable $A \subseteq X$ to measurable $B \subseteq Y$. All the mass of μ has to be transported somewhere, hence, $\gamma(A \times Y) = \mu(A)$ for all $A \subseteq X$ measurable.
- Transport cost: Let $c(x, y)$ be the cost of moving one unit of mass from x to y, then the total cost is

  $$\int_{X \times Y} c(x, y)\gamma(dxdy) = c[\gamma].$$

- Filling hole completely: $\gamma(X \times B) = \nu(B)$ for all $B \subseteq Y$ measurable. That is, the amount of mass transported to B has to be the volume of the hole in region B.

Remark. Note that, from above, $\gamma(X \times Y) = \mu(X) = \nu(Y) = 1$, so $\gamma \in \mathcal{P}(X \times Y)$, and $\gamma(A \times Y), \gamma(X \times B)$ defined marginals.

Definition 1.4: Marginals

Let $\gamma \in \mathcal{P}(X \times Y)$.

- *Marginal w.r.t. X: $M_X\gamma \in \mathcal{P}(X)$ defined via*

$$(M_X\gamma)(A) = \gamma(A \times Y) = \int_{A\times Y} \gamma(dxdy), \ \forall \text{ measurable } A \subseteq X,$$

- *Marginal w.r.t. Y: $M_Y\gamma \in \mathcal{P}(Y)$ defined via*

$$(M_Y\gamma)(B) = \gamma(X \times B) = \int_{X\times B} \gamma(dxdy), \ \forall \text{ measurable } B \subseteq Y.$$

Remark. Transport plans are probability measures on $X \times Y$ with marginals $M_X\gamma = \mu$, $M_Y\gamma = \nu$, γ is a *coupling* of the probability measure μ and ν.

Let $\Pi(\mu, \nu)$ be the set of all couplings between μ and ν.

Lemma 1.5

Let $\varphi \in L^1(X, \mu)$ and $\psi \in L^1(Y, \nu)$. Then for any coupling $\gamma \in \Pi(\mu, \nu)$. These hold

(M1) $\int_{X\times Y} \varphi(x)\gamma(dxdy) = \int_X \varphi(x)(M_X\gamma)(dx) = \int_X \varphi(x)\mu(dx)$.

(M2) $\int_{X\times Y} \psi(y)\gamma(dxdy) = \int_Y \psi(y)(M_Y\gamma)(dy) = \int_Y \psi(y)\nu(dy)$.

Proof sketch. Any function $\varphi \in L^1(X, \mu)$ can be approximated by simple functions: $\varphi = \lim_{n\to\infty} \sum_{j=1}^n \alpha_j \mathbb{1}_{A_j}$, for $A_j \subseteq X$ measurable.

$$\begin{aligned}\int_{X\times Y} \varphi(x)\gamma(dxdy) &= \lim_{n\to\infty} \sum_{j=1}^n \alpha_j \int_{A_j\times Y} \gamma(dxdy) = \lim_{n\to\infty} \sum_{j=1}^n \alpha_j \mu(A_j) \\ &= \lim_{n\to\infty} \int_X \sum_{j=1}^n \alpha_j \mathbb{1}_{A_j}(x)\mu(dx) = \int_X \varphi(x)\mu(dx).\end{aligned}$$

□

Definition 1.6: Couplings

Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. A probability measure $\gamma \in \mathcal{P}(X \times Y)$ is called coupling of μ and ν if $M_X\gamma = \mu$, $M_Y\gamma = \nu$. The set of all couplings between μ and ν is called $\Pi(\mu, \nu)$.

Definition 1.7: Kantorovich Optimal Transport Problem

Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$,

$$\min \ C[\gamma] = \int_{X\times Y} c(x, y)\gamma(dxdy) \tag{$\mathcal{K}$}$$

over all couplings $\gamma \in \Pi(\mu, \nu)$.

Structure of ($\mathcal{K}$):

(1) $\gamma \mapsto C[\gamma]$ linear function.

(2) $M_X\gamma = \mu$, $M_Y\gamma = \nu$ linear constraints.

(3) $\Pi(\mu,\nu)$ is a convex set: if $\gamma_1, \gamma_2 \in \Pi(\mu,\nu)$, $\lambda \in (0,1)$, then

(a) $\lambda\gamma_1 + (1-\lambda)\gamma_2 \in \mathcal{P}(X\times Y)$, since $\lambda\gamma_1(X\times Y)+(1-\lambda)\gamma_2(X\times Y) = 1$, and $\lambda\gamma_1(Z)+(1-\lambda)\gamma_2(Z) \geq 0$ for any $Z \subseteq X \times Y$ measurable.

(b) $(\lambda\gamma_1 + (1-\lambda)\gamma_2)(A \times Y) = \lambda\gamma_1(A\times Y) + (1-\lambda)\gamma_2(A\times Y) = \lambda\mu(A) + (1-\lambda)\mu(A) = \mu(A)$; analogously, $(\lambda\gamma_1 + (1-\lambda)\gamma_2)(X \times B) = \nu(B)$, for any $A \subseteq X, B \subseteq Y$ measurable.

$\Longrightarrow$ ($\mathcal{K}$) is a linear programming problem. *But*: it is in inifinte dimensions.

(4) Existence of couplings is trivial: independent coupling (product measure) $\gamma = \mu \otimes \nu$ defined via $\gamma(A\times B) = \mu(A)\nu(B)$ for every measurable $A \subseteq X, B \subseteq Y$, is a coupling of μ and ν.

(5) ($\mathcal{K}$) is higher-dimensional than ($\mathcal{M}$) in the following sense: consider transport plan given by a density $\tilde{\gamma} : \mathbb{R}^{2d} \to \mathbb{R}$. Discretize $\mathbb{R}^d$ by ℓ gridpoints then $\tilde{\gamma}$ corresponds to ℓ^2 real numbers. Transport $T : \mathbb{R}^d \to \mathbb{R}^d$ however only corresponds to ℓd real numbers. e.g. ℓ = number of pixels in a 2D picture, say $\ell = 500 \times 500$, then $\ell d = 500000$, but $\ell^2 = 62.5^9$.

1.3 Monge VS Kantorovich

Kantorovish problem is a relaxation of Monge Problem in the following sense: Monge = restriction of Kantorovish to sparse plan.

$$\gamma_T(dxdy) = \delta_{T(x)}(dy)\mu(dx)$$

for $T : X \to Y$ measurable such that $T_\#\mu = \nu$. For any $\varphi \in C_o(X\times Y)$,

$$\begin{aligned}\int_{X\times Y} \varphi d\gamma_T &= \int_{X\times Y} \varphi(x,y)\gamma_T(dx\times dy) = \int_{X\times Y} \varphi(x,y)\delta_{T(x)}(dy)\mu(dx) = \int_X \varphi(x,T(x))\mu(dx)\\ &= \int_X \varphi((id,T)(x))\mu(dx) = \int_X \varphi\circ(id,t)d\mu = \int_{X\times Y} \varphi(x,y)[(id,T)_\#\mu](dx\times dy).\end{aligned}$$

In other words, Monge measures that

$$\begin{aligned}\operatorname{supp}\gamma &= \{(x,y)\in X\times Y : \gamma(B_\epsilon(x,y)) > 0 \text{ for every } \epsilon > 0\}\\ &= \cap\{Z \subseteq X\times Y \text{ closed } : \gamma(Z) = 1\}\\ &\subseteq \operatorname{graph} T = \{(x,T(x)) \in X\times Y : x \in X\}\end{aligned}$$

Lemma 1.8

Let $T : X \to Y$ be measurable such that $T_\#\mu = \nu$. Then

(i) $C[\gamma_T] = I[T]$.

(ii) $\gamma_T \in \Pi(\mu,\nu)$

(iii) ($\mathcal{K}$) $\leq$ ($\mathcal{M}$).

Proof. (i) $C[\gamma_T] = \int_{X\times Y} c(x,y)\gamma_T(dx\times dy) = \int_{X\times Y} c(x,y)((id,T)_\#\mu)(dxdy) = \int_X c(x,T(x))\mu(dx) = I[T]$.

(ii) Let $A \subseteq X$, $B \subseteq Y$ measurable. Then

$$\begin{aligned}(M_X\gamma_T)(A) = \gamma_T(A\times Y) &= \int_{A\times Y} d\gamma_T = \int_X \mathbb{1}_{A\times Y}(x,y)\ \underbrace{\gamma_T(dx\times dy)}_{((id,T)_\#\mu)(dxdy)}\\ &= \int_X \underbrace{\mathbb{1}_{A\times Y}(x,T(x)}_{\mathbb{1}_A(x)}\mu(dx)\\ &= \int_A d\mu = \mu(A)\\ (M_Y\gamma_T)(B) = \int_X \underbrace{\mathbb{1}_{X\times B}(x,T(x)}_{\mathbb{1}_B(x)}\mu(dx) &= \int_Y \mathbb{1}_B(y)(T_\#\mu)(dy)\\ &= \int_A d\nu = \mu(A)\end{aligned}$$

(iii) By (i) and (ii),

$$(\mathcal{M}) = \inf_{T:X\to Y \text{ measurable}, T_\#\mu=\nu} I[T] = \inf_{T:X\to Y \text{ measurable}, T_\#\mu=\nu} C[\gamma_T] \geq \inf_{\gamma\in\Pi(\mu,\nu)} C[\gamma] = (\mathcal{K}),$$

by $\gamma_T \in \Pi(\mu,\nu)$.

□

1.4 Basic questions and examples

1. Existence Do optimal plans/maps exist? For optimal plans, yes, under reasonable assumptions on c, $\inf$ in ($\mathcal{K}$) is $\min$. For optimal maps:

   Example 1.1. Let $X = Y = \mathbb{R}^d$, $\mu = \delta_a, a \in \mathbb{R}^d$, $\nu = \frac{1}{2}(\delta_b + \delta_C)$ for $b \neq c \in \mathbb{R}^d$. Let $T : \mathbb{R}^d \to \mathbb{R}^D$ be measurable for all $A \subseteq \mathbb{R}^d$ open.

   $$(T_\#\delta_a)(A) = \delta_a(T^1(A)) = \begin{Bmatrix}1 & \text{if } a \in T^{-1}(A)\\ 0 & \text{otherwise}\end{Bmatrix} = \delta_{T(a)}(A).$$

   i.e. $T_\#\delta_a = \delta_{T(a)}$, so I cannot map the mass to two different points so $T_\#\mu = \nu$ is only possible if $b = c$.

   The class of all transport plans $\Pi(\mu,\nu)$ consist of a single measure

   $$\Pi(\mu,\nu) = \left\{\mu\otimes\nu = \frac{1}{2}\left(\delta_a\otimes\delta_b + \delta_a\otimes\delta_c\right)\right\}.$$

   Indeed, note that

   $$\begin{aligned}&\gamma\Big(\underbrace{\mathbb{R}^{2d}\setminus\{(a,b),(a,c)\}}_{=(\mathbb{R}^d\setminus\{a\}\times\mathbb{R}^d)\cup(\mathbb{R}^d\times\mathbb{R}^d\setminus\{b,c\})}\Big)\\ \leq&\gamma\left(\left(\mathbb{R}^d\setminus\{a\}\times\mathbb{R}^d\right)\right) + \gamma\left(\left(\mathbb{R}^d\times\mathbb{R}^d\setminus\{b,c\}\right)\right) = \mu(\mathbb{R}^d\setminus\{a\}) + \nu(\mathbb{R}^d\setminus\{b,c\})\\ =&\delta_a(\mathbb{R}^d\setminus\{a\}) + \frac{1}{2}(\delta_b+\delta_c)(\mathbb{R}^d\setminus\{b,c\}) = 0.\end{aligned}$$

   For any $\gamma \in \Pi(\mu,\nu)$,

   $\implies$ Any $\gamma \in \Pi(\mu,\nu)$ is supported on the points (a,b) and (a,c).

   $\implies$ $\gamma = \lambda\delta_{(a,b)} + (1-\lambda)\delta_{(a,c)} = \lambda\delta_a\otimes\delta_b + (1-\lambda)\delta_a\otimes\delta_c$, for some $\lambda \in [0,1]$.
   Since $M_Y\gamma = \lambda\delta_b + (1-\lambda)\delta_c$, $\nu = M_Y\gamma$ implies that $\lambda = \frac{1}{2}$.

2. Monge VS Kantorovich: discussed above.

3. Uniqueness Are minimizers unique? If not, can we characterize the set of minimizers?

Example 1.2. Consider $X = Y = \mathbb{R}^2$, $a = (-1, 0), b = (1, 0), a' = (0, -1), b' = (0, 1)$; $\mu = \frac{1}{2}(\delta_a + \delta_b), \nu = \frac{1}{2}(\delta_{a'} + \delta_{b'})$.

a) Consider Monge problem with quadratic distance cost

$$\inf_{T \in \mathcal{J}(\mu,\nu)} \int_{\mathbb{R}^2} |T(x) - x|^2 \mu(dx),$$

where $\mathcal{J}(\mu, \nu)$ is the set of push-forward maps, has two minimizers, defined on the support of μ:

$$T^{(1)}(a) = a', T^{(1)}(b) = b'; \quad T^{(1)}(a) = b', T^{(1)}(b) = a'.$$

Indeed, $\mathcal{J}(\mu, \nu) = \{T^{(1)}, T^{(2)}\}$, and

$$\begin{aligned}
\int_{\mathbb{R}^2} |T^{(1)}(x) - x|^2 \mu(dx) &= \frac{1}{2}\left(|T^{(1)}(a) - a|^2 + |T^{(1)}(b) - b|^2\right) \\
&= \frac{1}{2}\left(|a' - a|^2 + |b' - b|^2\right) = \frac{1}{2}(2 + 2) = 2. \\
\int_{\mathbb{R}^2} |T^{(2)}(x) - x|^2 \mu(dx) &= \frac{1}{2}\left(|T^{(2)}(a) - a|^2 + |T^{(2)}(b) - b|^2\right) \\
&= \frac{1}{2}\left(|b' - a|^2 + |a' - b|^2\right) = \frac{1}{2}(2 + 2) = 2.
\end{aligned}$$

b) Non-uniqueness in Kantorovich is even bigger:

$$\begin{aligned}
\Pi(\mu, \nu) &= \left\{ \text{ convex combinations of } \gamma_{T^{(1)}} \text{ and } \gamma_{T^{(2)}} \right\} \\
&= \left\{ \frac{1}{2}\left[(1 - \lambda)\delta_a \otimes \delta_{a'} + \lambda\delta_a \otimes \delta_{b'} + \lambda\delta_b \otimes \delta_{a'} + (1 - \lambda)\delta_b \otimes \delta_{b'}\right] : \lambda \in [0, 1] \right\}
\end{aligned}$$

4. Exact solutions

Example 1.3. Optimality of translation Let $X = Y = \mathbb{R}^d$, μ be a compactly supported probability measure on $\mathbb{R}^d$, and $\nu = (\tau_a)_{\#}\mu$, $\tau_a : \mathbb{R}^d \to \mathbb{R}^d$ translation by $a \in \mathbb{R}^d$, where $\tau_a(x) = x + a$. e.g., $\mu(dx) = f(x)dx$, then $\nu(dy) = g(y)dy$ with $g(y) = f(y - a)$ for all $y \in \mathbb{R}^d$.

$$\min\ I[T] = \int_{\mathbb{R}^d} |T(x) - x|^p \mu(dx),\ 1 < p < \infty$$

over $\mathcal{J}(\mu, \nu)$.

- Best to move each piece of mass by some distance?
- or better to move right side of pile to left side of hole (shorter distance) and make up for it by moving left side of pile to right side of hole (longer distance)?

The answer lies in *convexity* of the cost in displacement $T(x) - x$.

Definition 1.9

$\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ convex *if*

$$\Phi((1 - t)z + tz') \leq (1 - t)\Phi(z) + t\Phi(z')$$

for all $z \neq z' \in \mathbb{R}^d, t \in (0, 1)$. *It is* strictly convex *if the inequality is always strict.*

Theorem 1.10: Jensen's Inequality

$\Phi : \mathbb{R}^d \to \mathbb{R}$ convex and continuout. If μ is a probability measure on $\mathbb{R}^d$ and $v : \mathbb{R}^d \to \mathbb{R}^d$ integrable w.r.t. μ, then

$$\Phi\left(\int_{\mathbb{R}^d} u d\mu\right) \leq \int_{\mathbb{R}^d} \Phi(u) d\mu.$$

If Φ is strictly convex, inequality is strict, unless $u(x) = \bar{u} \in \mathbb{R}^d$ for μ-almost-everywhere x.

Let us prove (using convexity) that uniform transition, $T = \tau_a$ is the best:

Step 1 Introduce centers of mass of μ and ν,

$$R_\mu = \int_{\mathbb{R}^d} x\mu(dx), R_\nu = \int_{\mathbb{R}^d} y\nu(dy),$$

then $R_\nu = \int_{\mathbb{R}^d} y((\tau_a)_\# \mu)dy = \int_{\mathbb{R}^d} \tau_a(x)\mu(dx) = \int_{\mathbb{R}^d} x\mu(dx) + \int_{\mathbb{R}^d} a\mu(dx) = R_\mu + a$.

Step 2 Average displacement of any $T \in \mathcal{J}(\mu, \nu)$

$$\int_{\mathbb{R}^d} (T(x) - x)\mu(dx) = \underbrace{\int_{\mathbb{R}^d} T(x)\mu dx}_{\int_{\mathbb{R}^d} yd(T_\#\mu)(y) = R_\nu} - \underbrace{\int_{\mathbb{R}^d} x\mu(dx)}_{R_\mu} = R_\nu - R\mu = a$$

Step 3 Strict convexity of $\Phi_p : \mathbb{R}^d \to \mathbb{R}, z \to |z|^p$ for $1 < p < \infty$. By Jensen's inequality:

$$\begin{aligned} I[T] = \int_{\mathbb{R}^d} |T(x) - x|^p \mu(dx) &= \int_{\mathbb{R}^d} \Phi_p(T(x) - x)\mu(dx) \\ &\geq \Phi \underbrace{\left(\int_{\mathbb{R}^d} (T(x) - x)\mu(dx)\right)}_{a} = \Phi_p(a) = |a|^p \end{aligned}$$

for every $T \in \mathcal{J}(\mu, \nu)$.

- $T = \tau_a$ achieves equality by $\tau_a(x) - x = x + a - x = a$, so $I[\tau_a] = \int_{\mathbb{R}^d} |a|^p \mu(dx) = |a|^p$.
- Φ_p is strictly convex for $p > 1$: so equality holds if and only if $T(x) - x$ is a constant μ-a.e., which implies that $T = \tau_a$ is a unique minimizer.

Example 1.4 (Book Shifting (Gangbo-McCann 1996)). Consider $\mu(dx) = \frac{4}{3}\mathbb{1}_{[0,\frac{3}{4}]}(x)dx$ and $\nu(dx) = \frac{4}{3}\mathbb{1}_{[\frac{1}{4},1]}(y)dy$. Consider the problem

$$\inf \int_{[0,1]} |T(x) - x|\mu(dx)$$

where define $\Phi_1(x) := |x|$. Since Φ_1 is convex, a solution is given by $T_1(x) = x + \frac{1}{4}$ for $x \in [0, 3/4]$, which is minimal. Its transportation cost is $\frac{3}{4} * \frac{4}{3} * \frac{1}{4} = \frac{1}{4}$. Consider another shift:

$$T_2(x) = \begin{cases} x + \frac{3}{4}, & \text{if } x \in [0, \frac{1}{4}] \\ x, & \text{otherwise.} \end{cases}$$

The transport cost is $\frac{4}{3}\int_{[0,\frac{1}{4}]} \frac{3}{4}dx + \frac{4}{3}\int_{[\frac{1}{4},\frac{3}{4}]} 0 = \frac{1}{4}$, so T_2 is also minimal. Notce Φ_1 is convex but not strictly.

Example 1.5 (A dam and two ditches). We now consider a problem where Monge's problem has an optimal value but the inf is not attained.
Let $X = Y = \mathbb{R}^2$, and let $\mu(dx) = dx_2|_{x_1=0, x_2\in[0,1]}$, $\nu(dy) = \frac{1}{2}dy_2|_{y_1=-1, y_2\in[0,1]} + \frac{1}{2}dy_2|_{y_1=1, y_2\in[0,1]}$. The goal is

$$\min\ I[T] = \int_{\mathbb{R}^2} |T(x) - x|^2 \mu(dx) \text{ among all } T \in \mathcal{J}(\mu, \nu).$$

- $|T(x)-x| \geq d(\operatorname{supp}\mu, \operatorname{supp}\nu) = 1$ for all $x \in \operatorname{supp}\mu$ and for all $T \in \mathcal{J}(\mu,\nu)$., which implies that $I[T] \geq \int_{\mathbb{R}^2} 1\mu(dx) = 1$.
- Now we show $\inf_{T\in\mathcal{J}(\mu,\nu)} I[T] = 1$. Split up the dam into even number of segments of length $\epsilon > 0$ and consider the transport map T_2 depicted on the left. The maximal displacement of any point in the support is ϵ. Hence

$$|T_\epsilon(x)-x| \leq \sqrt{1+\epsilon^2} \implies I(T_\epsilon) = \int_{\mathbb{R}^2} |T_\epsilon(x)-x|^2 \mu(dx) \leq 1+\epsilon^2.$$

  Since ϵ can be chosen arbitrarily, $\inf I = 1$.

  Claim. No $T \in \mathcal{J}(\mu,\nu)$ achieves $I[T] = 1$.

  Proof. Assume there exists $T \in \mathcal{J}(\mu,\nu)$ such that $I(T) = 1$. This means $T(x) \in \operatorname{supp}\nu$ for μ-a.e. x; $|T(x)-x| = 1$ for μ-a.e. x. Hence, $T(0,x_2) = (\pm 1, x_2)$ for every Lesbegue-a.e. $x_2 \in [0,1]$. Set $\Omega_+ := \{x_2 \in [0,1] : [T(0,x_2)]_1 = 1\}$ and $\Omega_- := \{x_2 \in [0,1] : [T(0,x_2)]_1 = -1\}$. Then

$$T_\# \mu = dx_2|_{x_1=-1, x_2\in\Omega_-} + dx_2|_{x_1=1, x_2\in\Omega_+ \neq \nu}$$

  so $T \notin \mathcal{J}(\mu,\nu)$. Hence $\inf_{T\in\preceq,\succeq} I[T]$ is not attained. □

  Everything holds if change $|T(x)-x|^2$ to $|T(x)-x|^p, p > 0$.
  Problem: minimizing sequence T_ϵ exhibits faster and faster oscillations as $\epsilon \downarrow 0$. Hence, T_ϵ converges weakly, but not strongly in any $L^p(\mathbb{R}^2;\mu)$,

$$T_\epsilon(0,x_k) \xrightarrow[L^p]{\epsilon\downarrow 0} T_0(0,x_2) = \begin{pmatrix} 0 \\ x_2 \end{pmatrix}, \ \forall p > 1, \text{ but } T_0 \notin \mathcal{J}(\mu,\nu).$$

- Corresponding Kantororvich ProblemL

$$\inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^2\times\mathbb{R}^2} |y-x|^2 \gamma(dxdy)$$

  has simple explicit solution. Let $\gamma(dxdy) = \delta_0(x_1)\mathbb{1}_{[0,1]}(x_2)dx_2 \cdot \frac{\delta_1(y_1)+\delta_{-1}(y_1)}{2}\delta_{x_2}(y_2)$, which splits half of the mass to the left ditch and another half to the right one. It is not hard to see that the objective value of this γ is 1.

2 Multi-Marginal Optimal Transport (MMOT)

Goal: Find multivariate probability measure $\gamma \in \mathcal{P}(X_1 \times \cdots \times X_N)$. Given $N \in \mathbb{N}$, $N \geq 2$,

$$\min \int_{X_1 \times \cdots \times X_N} c(x_1, \ldots, x_N)\gamma(dx_1, \ldots, dx_N) =: c[\gamma] \tag{MMOT}$$

subject to constraint of given multivariate marginals $\gamma \in \Pi(\mu_1, \ldots, \mu_N)$, given measures $\mu_{\in}\mathcal{P}(X_i)$, where X_i are locally compact, separable, metric space and complete.

Definition 2.1: Marginal and Couplings

$\gamma \in \mathcal{P}(X_1 \times \cdots \times X_N)$ has marginals $\mu_1, \ldots, \mu_N$ with $\mu_i \in \mathcal{P}(X_i)$ if and only if $(M_{X_i}\gamma)(A_i) = \gamma(X_1 \times \cdots \times X_{i-1} \times A_i \times X_{i+1} \times \cdots X_N) = \mu_i(A_i)$ for any μ_i measurable $A_i \subseteq X_i$ for every $i \in \{1, \ldots, N\}$. Equivalently,

$$\int_{X_1 \times \cdots \times X_N} \varphi_i(x_i)\gamma(dx_1, \ldots, dx_N) = \int_{X_i} \varphi_i d\mu_i, \ \forall \varphi_i \in C_o(X_i), i \in \{1, \ldots, N\}.$$

Denote the set of all couplings as $\Pi(\mu_1, \ldots, \mu_N) := \{\gamma \in \mathcal{P}(X_1 \times \cdots \times X_N) : M_{X_i}\gamma = \mu_i, \ \forall i = 1, \ldots, N\}$.

Lemma 2.2

(MMOT) *is a linear programming problem.*

1) $\gamma \mapsto c(\gamma)$ is a linear map.

2) $\Pi(\mu_1, \ldots, \mu_N)$ is a convex subset of space of finite signed measure on $X_1 \times \cdots \times X_N$.

Remark. When $N = 2$, (MMOT) is equivalent to ($\mathcal{K}$). $\Pi(\mu_1, \ldots, \mu_N) \neq \emptyset$ since $\mu_1 \otimes \cdots \otimes \mu_N \in \Pi(\mu_1, \ldots, \mu_N)$.

Theorem 2.3

Let $c : \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_N} \to \mathbb{R} \cup \{\infty\}$ be lower semi-continuous (l.s.c.) and bounded from below. Then for any $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, $i = 1, \ldots, N$, (MMOT) *has a minimizer.*

Remark. Lower semi-continuous (sequential): Let X be a vector space with some notion of convergence. A functional $f : X \to \mathbb{R} \cup \{\infty\}$ is called (sequentially) lower semi-continuous (with respect to the given notion of convergence) if for any $x \in X$ and for any $\{x_j\}_{j\in\mathbb{N}} \subseteq X$ such that $x_j \overset{j\to\infty}{\longrightarrow} x$, we have $f(x) \leq \liminf_{j\to\infty} f(x_j)$.

Proof. Direct method of the calculus of variations.

1. Assume that $c[\gamma] = \int_{X_1 \times \cdots X_N} cd\gamma < \infty$ for at least one $\gamma \in \Pi(\mu_1, \ldots, \mu_N)$ (otherwise every coupling is a minimizer with cost ∞). Since c is bounded from below, say $c \geq \alpha$ for some $\alpha \in \mathbb{R}$, then

$$c[\gamma] = \int_{X_1 \times \cdots \times X_N} cd\gamma \geq \alpha\gamma(X_1 \times \cdots \times X_N) = \alpha > -\infty \implies \inf_{\gamma \in \Pi(X_1 \times \cdots \times X_N)} c[\gamma] \in \mathbb{R}.$$

2. Let $\{\gamma_k\}_{k\in\mathbb{N}} \subseteq \Pi(\mu_1, \ldots, \mu_k)$ be a minimizing sequence (exists by 1), i.e. $c[\gamma_k] \overset{k\to\infty}{\longrightarrow} \inf_{\gamma\in\Pi(\mu_1,\ldots,\mu_N)} c[\gamma]$.

3. Notions of convergence for measures

   - Signed Measure: $\mathcal{M}(X) = \{\mu - \nu : \mu, \nu \text{ non-negative Borel measures of finite mass}\}$. whose norm is defined as $\|\lambda\|_\pi = |\lambda|(X), |\lambda|(A) := \sup\{\sum_{i\in\mathbb{N}} |\lambda(A_i)| : A = \cup_{i\in\mathbb{N}} A_i, A_i \cap A_j = \emptyset, \forall i \neq j\}$.
   - Riesz Representation Theorem:

$$\mathcal{M}(\mathbb{R}^d) \cong [C_o(\mathbb{R}^d)]^* \cong [C_c(\mathbb{R}^d)]^*$$

where $C_o(\mathbb{R}^d) = \overline{C_c(\mathbb{R}^d)}^{\|\cdot\|_\infty}$, where $C_o(\mathbb{R}^d)$ is the set of continuous function vanishing at ∞ on $\mathbb{R}^d$, and $C_c(\mathbb{R}^d)$ is the set of compactly supported continuous function on $\mathbb{R}^d$.

- weak-$*$ convergence: $\mu_n \xrightarrow{*} \mu$ as $n \to \infty$ if and only if $\mu_n(\varphi) \overset{n\to\infty}{\longrightarrow} \mu(\varphi)$, i.e.,

$$\int \varphi d\mu_n \overset{n\to\infty}{\longrightarrow} \int \varphi d\mu, \ \forall \varphi \in C_o(\mathbb{R}^d).$$

□

Theorem 2.4: (Sequentia) Banach-Alaoglo

Let X ($C_o(\mathbb{R}^d)$) be a normed vector space, X^* ($\mathcal{M}(\mathbb{R}^d)$) be its dual. If X is separable*, then every bounded sequence in X^* possesses a weakly-$*$ convergence subsequence.*

Remark. Let $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}^d$ such that $|x_n| \to \infty$. Then $\delta_{x_n} \xrightarrow{*} 0$. For all $\varphi \in C_o(\mathbb{R}^d)$: $\int_{\mathbb{R}^d} \varphi d\delta_{x_n} = \varphi(x_n) \overset{n\to\infty}{\longrightarrow} 0$. That is, indeed, we do not capture how the δ_{x_n} behaves when x_n goes to ∞.

Definition: Narrow convergence

Narrow convergence (probabilist: weak convergence): $\mu_n \to \mu$ narrowly *as $n \to \infty$ if and only if*

$$\int_{\mathbb{R}^d} \varphi d\mu_n \to \int_{\mathbb{R}^d} \varphi d\mu, \ \forall \varphi \in C_b(\mathbb{R}^d)$$

where $C_b(\mathbb{R}^d)$ is the set of bounded continuous function on $\mathbb{R}^d$. In particular, if $\mu_n \to \mu$ narrowly as $n \to \infty$, then (pick $\varphi \equiv 1$), $\mu_n(\mathbb{R}^d) \to \mu(\mathbb{R}^d)$ (Notice that we might have this for weak-$*$ convergence as $\varphi \equiv 1$ does not vanish at ∞.

Theorem 2.5: Prokhorov

For a set $K \subseteq \mathcal{M}_+(\mathbb{R}^d)$ the following are equivalent:

1. *K is bounded and* tight*, i.e. $\sup_{\mu\in K} \mu(\mathbb{R}^d \setminus \overline{B}_R) \overset{R\to\infty}{\longrightarrow} 0$. This could be understood as the mass outside any large ball uniformly small for all measure on K.*
2. *K is relatively sequentially compact with respect to narrow convergence, i.e., every sequence in K has a narrowly convergent subsequence.*

Proof. Assume K is bounded and tight and let $\{\mu_j\}_{j\in\mathbb{N}}$ be a sequence in K.

- K bounded, and by THM 2.4 and the fact that $C_o(\mathbb{R}^d)$ is separable, there exists a subsequence $\{mu_{j_k}\}_{k\in\mathbb{N}} \subseteq K$ such that $\mu_{j_k} \xrightarrow{*} \mu \in \mathcal{M}(\mathbb{R}^d)$.
  Note that $\mu \in \mathcal{M}_+(\mathbb{R}^d)$: to see this, let $B \subseteq \mathbb{R}^d$ be compact, and approximate $\mathbb{1}(B)$ by continuous functions: Define $\varphi_B^\epsilon := (1 - \epsilon^{-1} d(x, B))_+$, where $d(x, B) = \inf_{y\in B}|x - y|$, and $B^\epsilon := \{x \in \mathbb{R}^d : d(x, B) \le \epsilon\}$. Then
  1. For $\epsilon_1 < \epsilon_2$, $B^{\epsilon_1} \subseteq B^{\epsilon_2}$, and $B = \cap_{n\in\mathbb{N}} B^{1/n}$.
  2. $\mathbb{1}_B \le \varphi_B^\epsilon \le \mathbb{1}_{B^\epsilon}$.
  3. φ_B^ϵ continuous with $|\varphi_B^\epsilon(x) - \varphi_B^\epsilon(y)| \le \epsilon^{-1}|x - y|$.

  Hence

$$\mu(B^\epsilon) = \int \mathbb{1}_{B^\epsilon} d\mu \overset{2}{\ge} \int \varphi_B^\epsilon d\mu \overset{3}{\ge} \lim_{k\to\infty} \int \varphi_B^\epsilon d\mu_{j_k} \overset{2}{\ge} \limsup_{k\to\infty} \int \mathbb{1}_B d\mu_{j_k} = \limsup_{k\to\infty} \mu_{j_k}(B) \ge 0.$$

  Note that

$$\mu(B) \overset{1}{=} \mu(\cap_{n\in\mathbb{N}} B^{1/n}) \overset{(*)}{=} \lim_{n\to\infty} \mu(B^{1/n}) \ge 0$$

so $\mu \in \mathcal{M}_+(\mathbb{R}^d)$.
$(*)$ continuity: since μ finite (signed) measure, we can write

$$\begin{aligned}
\mu(B^*) - \mu(\cap_{n\in\mathbb{N}} B^{1/n}) &= \mu(B^1 \setminus \cap_{n\in\mathbb{N}} B^{1/n} = \mu(\cup_{n\in\mathbb{N}}(B^1 \setminus B^{1/n})) \\
&= \mu\left(\dot{\cup}_{n\in\mathbb{N}}(B^1 \setminus B^{1/(n+1)}) \setminus (B^1 \setminus B^{1/n})\right) \\
&= \sum_{n\in\mathbb{N}} \mu\left((B^1 \setminus B^{1/(n+1)}) \setminus (B^1 \setminus B^{1/n})\right) \\
&= \lim_{N\to\infty} \sum_{n=1}^{N-1} \mu\left((B^1 \setminus B^{1/(n+1)}) \setminus (B^1 \setminus B^{1/n})\right) \\
&= \lim_{N\to\infty} \cup_{n=1}^{N-1} \mu\left((B^1 \setminus B^{1/(n+1)}) \setminus (B^1 \setminus B^{1/n})\right) \\
&= \lim_{N\to\infty} \mu(B^1 \setminus B^{1/N}) = \mu(B^1) - \lim_{N\to\infty} \mu(B^{1/N})
\end{aligned}$$

where the limit exists because $|\mu|(\mathbb{R}^d) < \infty$.
More generally, we have

Theorem: Portmanteau theorem

Let $\{\mu_n\} \subseteq \mathcal{P}(\mathbb{R}^d)$. Then the following are equivalent:

1. *$\mu_n \to \mu$ narrowly.*
2. *$\limsup_{n\to\infty} \mu_n(F) \le \mu(F)$ for all F closed.*
3. *$\mu(O) \le \liminf_{n\to\infty} \mu_n(O)$ for all O open.*
4. *$\mu_n(A) \to \mu(A)$ for all A measurable such that $\mu(\partial A) = 0$, where ∂A is the boundary of A.*

- *Claim.* $\int \varphi d\mu_{j_k} \to \int \varphi d\mu$ for all $\varphi \in C_b(\mathbb{R}^d)$, i.e., $\mu_{j_k} \to \mu$ narrowly as $k \to \infty$.
Truncation argument: fix $\varphi \in C_b(\mathbb{R}^d)$ and define the cutoff function ξ_R via

$$\xi_R(x) = \begin{cases} 1, & \text{if } |x| \le R, \\ 1 - (|x| - R), & \text{if } R < |x| \le R+1, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\int_{\mathbb{R}^d} \varphi d\mu_{j_k} - \int_{\mathbb{R}^d} \varphi d\mu = \underbrace{\int_{\mathbb{R}^d} \xi_R \varphi d(\mu_{j_k} - \mu)}_{\to 0 \text{ by weak-}* \text{ convergence}} + \underbrace{\int_{\mathbb{R}^d} (\varphi - \varphi\xi_R) d\mu_{j_k}}_{=: T_R^{(j_k)}} + \underbrace{\int_{\mathbb{R}^d} (\xi_R\varphi - \varphi) d\mu}_{=: T_R}.$$

Note that by tightness

$$|T_R^{(j_k)}| \le \|\varphi\|_\infty \int_{\mathbb{R}^d \setminus \bar{B}_R} d\mu_{j_k} = \|\varphi\|_\infty \mu_{j_k}(\mathbb{R}^d \setminus \bar{B}_R) =: \alpha_R \overset{R\to\infty}{\longrightarrow} 0$$

$$|T_R| \le \|\varphi\|_\infty \int_{\mathbb{R}^d \setminus \bar{B}_R} d\mu := \beta_R \overset{R\to\infty}{\longrightarrow} 0$$

Thus, $\lim_{j\to\infty} \left|\int \varphi d\mu_{j_k} - \int \varphi d\mu\right| \le \alpha_R + \beta_R$ for all $R > 0$. Letting $R \to \infty$ proves as the narrow convergence.

□

Proposition 2.6

The set of couplings $\Pi(\mu_1, \dots, \mu_N) \subseteq \mathcal{M}_+(\mathbb{R}^d)$ is

(a) Bounded and tight

(b) closed under narrow convergence.

Proof. (a) Being bounded is clear since they are probability measures.
Tightness: using marginal conditions: since $\mu_i(\mathbb{R}^d) = 1, \forall i = 1, \dots, N$, we can find, for every $\epsilon > 0$, a closed ball $\bar{B}_i$ such that $\mu(\mathbb{R}^{d_i} \backslash \bar{B}_i) < \epsilon/N$. Note that $\mathbb{R}^d \backslash (\bar{B}_1 \times \cdots \times \bar{B}_N) = \cup_{i=1}^N \left\{x = (x_1, \dots, x_N) : x_i \in \mathbb{R}^{d_i} \setminus \bar{B}_i\right\}$, then for $\gamma \in \Pi(\mu_1, \dots, \mu_N)$, we have

$$\gamma(\mathbb{R}^d \setminus \bar{B}_1 \times \cdots \times \bar{B}_N) \leq \sum_{i=1}^N \gamma(\{x : x_i \in \mathbb{R}^{d_i} \setminus \bar{B}_i\}) = \sum_{i=1}^N \mu_i(\mathbb{R}^{d_i} \setminus \bar{B}_i) < \epsilon.$$

(b) Let $\gamma_j \to \gamma$ narrowly, $\{\gamma_j\}_{j\in\mathbb{N}} \subseteq \Pi(\mu_1, \dots, \mu_N)$. Want to show that $\gamma \in \Pi(\mu_1, \dots, \mu_N)$. Take $\varphi_i \in C_o(\mathbb{R}^{d_i})$ and set $\Phi_i(x_1, \dots, x_N) = \varphi_i(x_i) = \varphi_i(x_i)$. Then $\Phi \in C_b(\mathbb{R}^d)$ (Not in $C_o(\mathbb{R}^d)$ necessarily) and by narrow convergence

$$\int \varphi_i d\mu_i = \int \Phi_i d\gamma_j \overset{\text{narrow convergence}}{\longrightarrow} \int_{\mathbb{R}^d} \Phi_i d\gamma = \int_{\mathbb{R}^{d_i}} \varphi d(M_{X_i}\gamma),\ \forall i = 1, \dots, N$$

So $M_{X_i}\gamma = \mu$ for each i, we are done.

□

Corollary 2.7

Our minimizing sequence $\{\gamma_j\}_{j\in\mathbb{N}} \subseteq \Pi(\mu_1, \dots, \mu_N)$, (s.t. $c(\gamma_j) \to \inf_{\gamma\in\Pi} c[\gamma]$) has narrowly convergent subsequence $\{\gamma_{j_k}\}_{k\in\mathbb{N}} \subseteq \Pi(\mu_1, \dots, \mu_N)$ s.t. $\gamma_{j_k} \overset{k\to\infty}{\longrightarrow} \gamma \in \mathcal{M}_+(\mathbb{R}^d)$ narrowly. Since $\Pi(\mu_1, \dots, \mu_N)$ is closed under narrow convergence, it follows that $\gamma \in \Pi(\mu_1, \dots, \mu_N)$.

Question: Is γ the minimizer we are looking for? I.e., do we have

$$c[\gamma] = c[\text{narrow } \lim_{k\to\infty} \gamma_{j_k}] \leq \liminf_{k\to\infty} c[\gamma_{j_k}] = \inf_{\gamma\in\Pi} c[\gamma]?$$

Yes if we have lower semi-continuous of c with respect to narrow convergence.

Proposition 2.8

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be lower semi-continuous and bounded from below. Then the functional $F : \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$, $F[\mu] := \int_{\mathbb{R}^d} f d\mu$ is lower semi-continuous on $\mathcal{M}_+(\mathbb{R}^d)$ with respect to narrow convergence. I.e., if $\mu_j \to \mu$ narrowly, then $F[\mu] \leq \liminf_{j\to\infty} F[\mu_j]$.

We introduce some technical tools before proving this proposition

Lemma 2.9: l.s.c. preserved under pointwise supremum

Let X be a vector space with some notion of convergence. If $\{f_\lambda\}_{\lambda\in\Lambda}$ is an arbitrary family of l.s.c. functions (with respect to notion of convergence), $f_\lambda : x \to \mathbb{R} \cup \{\infty\}$. Then the pointwise supremeum $f(x) := \sup_{\lambda\in\Lambda} f_\lambda(x)$ is l.s.c.

Proof. Let $\{x_j\}_{j\in\mathbb{N}} \subseteq X$ such that $x_j \to x \in X$. By l.s.c. of f_λ we get that

$$f_\lambda(x) \leq \liminf_{j\to\infty} f_\lambda(x_j) \leq \liminf_{j\to\infty} \sup_{\lambda\in\Lambda} f_\lambda(x_j) = \liminf_{j\to\infty} f(x_j).$$

This holds for any $\lambda \in \Lambda$. Taking the sup over all $\lambda \in \Lambda$ in this inequality gives $f(x) = \sup_{\lambda\in\Lambda} f_\lambda(x) \leq \liminf_{j\to\infty} f(x_j)$. □

Definition 2.10: Infimal Convolution

Let X be a vector space, $F, G : X \to \mathbb{R} \cup \{\infty\}$. Then their infimal convolution $F\square G : X \to \mathbb{R} \cup \{\infty\}$ is defined via

$$\begin{aligned}(F\square G)(x) &= \inf_{z\in X}(F(x-z) + G(z)) \\ &= \inf_{y\in X}(F(y) + G(x-y)) \\ &= \inf_{a+b=x}(F(a) + G(b))\end{aligned}$$

Proposition 2.11: Properties of infimal convolution

(i) If F, G bounded below, then $F\square G$ is bounded below and proper ($\not\equiv \infty$).

(ii) (Identity) $X_{\{0\}}\square F = F$, where for $A \subseteq X$,

$$X_A := \begin{cases} 0, & \text{if } x \in A. \\ \infty, & \text{else} \end{cases}$$

(iii) (Smoothing) If X is a normed vector space and F bounded below, then for any $\varepsilon \geq 0$, $\left(\frac{1}{\varepsilon}\|\cdot\|\right)\square F$ is $\frac{1}{\varepsilon}$-Lipschitz.

(iv) (Approximation) In the setting of (iii): if F is in addition l.s.c. with respect norm convergence, then

$$\left(\frac{1}{\varepsilon}\|\cdot\|\right)\square F \to F \text{ pointwise as } \varepsilon \to 0$$

Proof. (i) Let $F \geq \alpha, G \geq \beta$ for some $\alpha, \beta \in \mathbb{R}$. Then

$$F\square G(x) = \infty_{z\in X} F(x-z) + G(z) \geq \alpha + \beta,\ \forall x \in X.$$

Given F, G proper, i.e., there exists $a, b \in X$ such that $F(a) < \infty$, $G(b) < \infty$. Note that $F\square G$ is finite at $x = a + b$ since

$$(F\square G)(x) = \inf_{z\in X}(F(x-z) + G(z)) \leq F(x-b) + G(b) = F(a) + G(b) < \infty.$$

(ii) $X_{\{0\}}F(x) = \inf_{z\in X}(X_{\{0\}}(x-z) + F(z)) = F(x), \forall x \in X$, where

$$X_{\{0\}}(x-z) + F(z) = \begin{cases} F(x), & \text{if } z = x \\ \infty, & \text{else.} \end{cases}$$

(iii) Since F is proper, $F(x_0) < \infty$ for some $x_0 \in X$, which implies

$$F_\varepsilon(x) = \left(\left(\frac{1}{\varepsilon}\|\cdot\|\square F\right)\right)(x) = \inf_{z\in X}\left(\frac{1}{\varepsilon}\|x-z\| + F(z)\right) \leq \frac{1}{\varepsilon}\|x-x_0\| + F(x_0) < \infty,\ \forall x \in X.$$

Want to show: $|F_\varepsilon(x) - F_\varepsilon(x')| \le \frac{1}{\varepsilon}\|x - x'\|$ for any $x, x' \in X$.
For any $x \in X$ fixed and given $\delta > 0$, one can choose z_δ such that

$$F(z_\delta) + \frac{1}{\varepsilon}\|x - z_\delta\| \le F_\varepsilon(x) + \delta < \infty.$$

On the other hand, for any $x' \in X$:

$$F(z_\delta) + \frac{1}{\varepsilon}\|x' - z_\delta\| \ge F_\varepsilon(x').$$

Subtracting two inequalities, we obtain

$$\begin{aligned} F_\varepsilon(x') - F_\varepsilon(x) &\le F(z_\delta) + \frac{1}{\varepsilon}\|x' - z_\delta\| - \left(F(z_\delta) + \frac{1}{\varepsilon}\|x - z_\delta\| - \delta\right) \\ &\le \frac{1}{\varepsilon}\|x' - x\| + \delta. \end{aligned}$$

Then for any $\delta > 0$, $x, x' \in X$, we have the above inequality and we can switch x, x', then we have

$$\begin{aligned} F_\varepsilon(x') - F(x) &\le \frac{1}{\varepsilon}\|x - x'\| + \delta \\ F_\varepsilon - F_\varepsilon(x') &\le \frac{1}{\varepsilon}\|x - x'\| + \delta \\ \implies |F_\varepsilon - F_\varepsilon(x')| &\le \frac{1}{\varepsilon}\|x - x'\| + \delta \end{aligned}$$

let $\delta \to 0$, done. Note that $F_\varepsilon(x) > -\infty$ for every $x \in X$ simply by F being bounded below.

(iv) Heuristically, $\frac{1}{\varepsilon}\|\cdot\| \xrightarrow{\varepsilon \downarrow 0} X_{\{0\}}$, so formally, this follows with (ii).
Rigorous argument: Fix $x \in X$, WLOG, let $\varepsilon < 1$,

- $F_\varepsilon(x) := \left(\left(\frac{1}{\varepsilon}\|\cdot\| \square F\right)\right)(x)$ is increasing as $\varepsilon \downarrow 0$. Since for $\varepsilon' \le \varepsilon$.

$$F_\varepsilon(x) = \inf_{z \in X}\left(\frac{1}{\varepsilon}\|z - x\| + F(z)\right) \le \left(\frac{1}{\varepsilon'}\|z - x\| + F(z)\right) = F_{\varepsilon'}(x)$$

- $F_\varepsilon(x) = \inf_{z \in X}\left(\frac{1}{\varepsilon}\|z - x\| + F(z)\right) \le F(x)$ (by choosing $z = x$) uniformly in $\varepsilon > 0$.

Then $\lim_{\varepsilon \downarrow 0} F_\varepsilon(x) =: a \in \mathbb{R} \cup \{\infty\}$ exists.
Suppose for contradiction that $a < F(x)$. Then in particular $a < \infty$. Choose z_ε such that

$$F(z_\varepsilon) + \frac{1}{\varepsilon}\|x - z_\varepsilon\| \le F_\varepsilon(x) + \varepsilon \le a + \varepsilon,\ \forall \varepsilon > 0 \tag{$\star$}$$

which implies that

$$\frac{1}{\varepsilon}\|x - z_\varepsilon\| \le a - \underbrace{F(z_\varepsilon)}_{\ge \alpha} + \varepsilon \le a + |\alpha| + 1 < \infty,\ \forall \varepsilon < 1$$

so $\|x - z_\varepsilon\| \to 0$ as $\varepsilon \downarrow 0$. By ($\star$), $F(z_\varepsilon) \le a + \varepsilon - \frac{1}{\varepsilon}\|x - z_\varepsilon\| \le a + \varepsilon$ and by l.s.c. of F with respect to $\|\cdot\|$-convergence as $\varepsilon \to 0$, we have

$$F(x) \le \liminf_{\varepsilon \downarrow 0} F(z_\varepsilon) \le a$$

because $\varepsilon \downarrow 0$, $z_\varepsilon \to x$ w.r.t. $\|\cdot\|$ and F l.s.c., which contradicts to $a < F(x)$, which implies that $a = F(x)$, i.e., $F_\varepsilon(x) \to F(x)$ as $\varepsilon \downarrow 0$.

□

Proof of Proposition 2.8.

(1) There exists a sequence of bounded, continuous function f_n such that f_n converges pointwise monotonically from below to f. We get this by taking $f_n := \min\{n, (n|\cdot|)\square f)\}$. Then f_n is

- bounded from below (since f is).
- bounded from above (by construction)
- continuous (since f is bounded from below, by (ii) of the above proposition, we know $(n|\cdot|)\square c)$ is n-Lipschitz.
- converges pointwise to f ((iv) of the above proposition).

(2) Set $F_n := \int_X f_n d\mu$. Claim: $F[\mu] = \lim_{n\to\infty} F_n[\mu],\ \forall \mu \in \mathcal{M}_+(\mathbb{R}^d)$. Indeed, since $f_n \uparrow f$, this follows from monotone convergence theorem.

(3) Since $f_n \leq f$, we have $F_n \leq F$. By (2), $F_n \to F$ pointwise. Hence $F = \sup_{n\in\mathbb{N}} F_n$

(4) f_n is bounded and continuous, so F_n is continuous and in particular with respect narrow convergence. By Lemma 2.9, $F = \sup_n f_n$ is l.s.c..

□

3 Duality and Brenie's Theorem

(MMOT) is a convex optimization problem, so we can consider its dual formulation.

Theorem 3.1: Kantorovich Duality

Let $c : \mathbb{R}^{d_1 \times \cdots d_N} \to \mathbb{R} \cup \{\infty\}$ be l.s.c., bounded from below, and $\inf_\gamma c[\gamma] < \infty$. Then for any $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, $i = 1, \dots, N$,

$$\inf_{\gamma \in \Pi(\mu_1, \dots, \mu_N)} \underbrace{\int_{\mathbb{R}}^{d} c d\gamma}_{c[\gamma]} = \sup_{(u_1, \dots, u_N) \in \mathcal{A}(c)} \underbrace{\sum_{i=1}^{N} \int_{\mathbb{R}^{d_i}} u_i d\mu_i}_{\mathcal{J}[u]},$$

where $\mathcal{A}(c) = \left\{(u_1, \dots, u_N) : u_i \in Y(\mathbb{R}^{d_i}), \sum_{i=1}^N u_i(x_i) \le c(x_1, \dots, x_N) \text{ for a.e. } x = (x_1, \dots, x_N) \in \mathbb{R}^d\right\}$, and $C_o \subseteq Y \subseteq L^1(\mu_i)$.

Remark.

- Solution u_i of the dual are called *Kantorovich potentials.*
- Will show: u_i: Lagrangian multiplier for i-th marginal constraint; u_i: functional derivative of optimal cost with respect to i-th marginal.
- What is Kantorovish duality good for?
  - verifying explicitly (guesses of) optimal plans, so "dual certificates". Suppose we have some $\gamma \in \Pi(\mu_1, \dots, \mu_N)$ and $(u_1, \dots, u_N) \in \mathcal{A}[c]$ such that $c[\gamma] = \mathcal{J}[u]$. Then γ is optimal, and u is optimal.

$$\mathcal{J}[u] \le \sup \mathcal{J} \le \inf c \le c[\gamma] = \mathcal{J}[u].$$

  - numerics
  - modern proof of Brenier's theorem (and other sparsity theorems).

Theorem 3.2: Brenier's Theorem/Sparsity restuls, $N = 2$

If $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ are such that $\int_{\mathbb{R}^d} |x|^2 \mu_i(dx) < \infty$ and μ_1 is absolutely continuous with respect Lebesgue measure, then any optimal coupling $\pi \in \arg\min_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2}|y - x|^2 \gamma(dxdy)$ is of Monge fashion, i.e., $\operatorname{supp} \pi \subseteq \operatorname{graph} T$ for some $T : \mathbb{R}^d \to \mathbb{R}^d$ if and only if $\pi(dxdy) = \mu_1(dx)\delta_{T(x)}(dy)$ or $\pi = (Id, T)_{\#}\mu_1$.

Proof idea.

- First, $c[\gamma] = \int \frac{1}{2}|y - x|^2 \gamma(dxdy)$ is finite for $\gamma \in \Pi(\mu_1, \mu_2)$. Indeed,

$$c[\gamma] \le \int (|y|^2 + |x|^2)\gamma(dxdy) = \int_{\mathbb{R}^d} |y|^2 d\mu_2(y) + \int_{\mathbb{R}^d} |x|^2 \mu_1(dx) < \infty.$$

- $c(x, y) = \frac{1}{2}|x - y|^2$ is continuous and nonnegative.

By the existence theorem, there exists a minimizer π of $\inf c[\gamma]$. Take minimizer π of c and maximizer $u = (u_1, u_2) \in \mathcal{A}(c)$ of $\mathcal{J}$. By Kantorovish Duality,

$$\begin{aligned} 0 = c[\pi] - \mathcal{J}[u] &= \int_{\mathbb{R}^{2d}} \frac{1}{2}|y - x|^2 \pi(dxdy) - \int_{\mathbb{R}^d} u_1 d\mu_1 - \int_{\mathbb{R}^d} u_2 d\mu_2 \\ &= \int_{\mathbb{R}^{2d}} \underbrace{\left(\frac{1}{2}|y - x|^2 - u_1(x) - u_2(y)\right)}_{=: \tilde{c}(x,y) \ge 0} \pi(dxdy). \end{aligned}$$

It follows that $\tilde{c}(x,y) = 0$ for all $(x,y) \in \operatorname{supp}\pi$ by $\pi \geq 0$. Since $\tilde{c} \geq 0$, this means that $\tilde{c}$ is minimal on $\operatorname{supp}\pi$, so $\operatorname{supp}\pi \subseteq \arg\min \tilde{c}$. Thus we have $\forall (x,y) \in \operatorname{supp}\pi : 0 = \nabla_x \tilde{c}(x,y) = x - y - \nabla u_i(x)$ and $T(x) := x - \nabla u_1(x) = y$. In other words, $T(x) - x = -\nabla u_1(x)$. One has to argue that u_1 (resp. u_2) is indeed differentiable (at least a.e.). We will get back to this later. This is where we will need the absolute continuity of μ_1. □

Theorem 3.3: Fenchel-Rockefeller Duality, finite dimensional

Let $X = X^* = \mathbb{R}^n$, $F, G : X \to \mathbb{R} \cup \{+\infty\}$ convex. Suppose that there exists $z_0 \in X$ such that

- *F is finite in open neighborhood of z_0.*
- *G is finite at z_0.*

Then

$$\inf_{x \in X} (F(x) + G(x)) = \sup_{y \in X^*} (-F^*(y) - G^*(-y)).$$

Note. Amazingly, the above theorem holds also in infinite dimensional normed vector spaces!

Proof.

Step 1 Statement is true for $F, G : \mathbb{R} \to \mathbb{R}$.

- smooth ($C^1(\mathbb{R})$).
- superlinear: $F(x)/|x|, G(x)/|x| \to \infty$ as $|x| \to \infty$, so that sup in definition of Legendre transform is always attained for all y.
- strictly convex (sup in Legendre transformation is attained at unique point for all y).

Geometric meaning:

(i) left-hand side: want to minimize

$$F(x) + G(x) = F(x) - (-G(x)),$$

which gives the minimal vertical distance between graphs of F and $-G$.

(ii) r.h.s: want to maximize

$$-F^*(y) - G^*(-y)$$

which gives the maximal vertical distance between parallel tangents (maximum over slopes).

How to see this? Let $L(x) = y \cdot x + a$ (a = intersection of line with slope y with vertical axis). Then a = maximal numbers such that $L(x) \leq F(x)$ for all $x \in \mathbb{R}$, i.e., $a = \inf_{x \in \mathbb{R}}(F(x) - y \cdot x) = -F^*(y)$. Similarly for lower line $L(x) = y \cdot x + b$, where b = smallest number such that $L(x) \geq -G(x)$ for all x, i.e., $b = \sup_{x \in \mathbb{R}}(-y \cdot x - G(x)) = G^*(-y)$. We now use standard calculus to show that both sides are equal.

(I) At a point where $\inf_{x \in \mathbb{R}}(F(x) + G(x))$ is attained, we must have $F'(x) + G'(x) = 0$, i.e., $F'(x) = (-G)'(x)$, so tangents to F and $-G$ at x are parallel.

(II) At a point where $\sup_{y \in \mathbb{R}}(-F^*(y) - G^*(-y))$ is attained, we have $-F^{*\prime}(y) + G^{*\prime}(-y) = 0$. Let us compute these definitions: $F^*(y) = \sup_{x \in \mathbb{R}}(yx - F(x))$. The supremum is attained at the point where $y = F'(x)$ which implies $x_F = (F')^{-1}(y)$. Thus, $F^*(y) = yx_F - F(x_F) = y(F')^{-1}(y) - F((F')^{-1}(y))$ and therefore $F^{*\prime}(y) = (F')^{-1}(y) + y((F')^{-1})'(y) - F'((F')^{-1}(y))((F')^{-1})'(y) = (F')^{-1}(y) = x_F$. Similarly, $G^{*\prime}(-y) = x_{-G}$ which implies $x_F = x_{-G}$. Note that by construction $F'(x_F) = y = (-G)'(x_{-G})$. Hence at the common touching point $x_F = x_{-G} =: x'_*$, this holds

$$F'(x'_*) = -G'(x'_*).$$

(III) F' is strictly increasing (since F is strictly convex), $(-G)'$ is strictly decreasing (since $-G$ strictly concave), which implies that there exists only one point where $F' = -G'$, so such point is $x'_* = x_F = x_{-G}$. But at x'_*, the distance $-F^*(y) - G^*(-y)$ equals the distance $F(x'_*) - (-G(x'_*))$ by $F^*(y) = yx_F - F(x_F)$, and $G^*(-y) = -yx_G - F(x_G)$. Thus,

$$F(x'_*) + G(x'_*) = \min_{x\in\mathbb{R}} F(x) + G(x) = -F^*(y) - G^*(-y) = \max_{y\in\mathbb{R}} -F^*(y) - G^*(-y)$$

Step 2 Extension to general F and G, notice that we cannot use the previous proof: F, G may not be differentiable (not even continuous!), optimizers may not exists, etc.

(I) Set $m := \inf_{x\in X}(F(x) + G(x))$. Note that since F and G finite at $z_0 \in X$, $m < \infty$ (might have $m = -\infty$). By definition of F^* and G^*,

$$\begin{aligned}\sup_{y\in X^*}(-F^*(y) - G^*(-y)) &= \sup_{y\in X^*}(-\sup_{x\in X}(y\cdot x - F(x)) - \sup_{z\in X}(-y\cdot z - G(z)))\\ &= \sup_{y\in X^*}\inf_{x,z\in X}(F(x) + G(z) + y\cdot(z-x))\end{aligned}$$

(II) $m \geq \sup_{y\in X^*}(-F^*(y) - G^*(-y))$. This follows by fixing y and setting $z = x$:

$$\inf_{x,z\in X}(F(x) + G(z) + y\cdot(z-x)) \leq \inf_{x\in X}(F(x) + G(x) + 0) = m.$$

(III) $m \leq \sup_{y\in X^*}(-F^*(y) - G^*(-y))$. This is highly non-trivial: we show that there exists y_* in X^* such that

$$F(x) + G(z) + y_*\cdot(z-x) \geq m,\ \forall x, z \in X.$$

This is easy for $m = -\infty$: any choice of $y_* \in \mathbb{R}^d$ works. For $m > -\infty$, we rely on the following deep result:

Theorem: Separation Theorem for convex sets in $\mathbb{R}^d$

Let $X = X^* = \mathbb{R}^d$ and C, C' be any convex, nonempty, disjoint sets in X. Then there exists a separating hyperplane, i.e., there exist $v \in X^* \setminus \{0\}$ and $\alpha \in \mathbb{R}$ s.t. $v\cdot x \geq \alpha$ for all $x \in C$ and $v\cdot y \leq \alpha$ for all $y \in C'$

We apply this to

$$C := \{(x,\lambda) \in X\times\mathbb{R} : \lambda > F(x)\},\ C' := \{(z,\mu) \in X\times\mathbb{R} : \mu \leq m - G(z)\}.$$

- C, C' are convex since F, G are convex.
- C is nonempty, since $F(z_0) < \infty$, C' is nonempty, since $C(z_0) < \infty$ and $m > -\infty$.
- C and C' are disjoint: Suppose not, then there exist $(x,\lambda) \in C,\ (z,\mu) \in C'$, such that $(x,\lambda) = (z,\mu)$ which implies $\lambda > F(x), \lambda \leq m - G(x)$ so $0 > F(x) + G(x) - m$ contradicting $m = \inf_{x\in X}(F(x) + G(x))$. The separation theorem implies that there exist $(y,\alpha) \in X^*\times\mathbb{R}\setminus\{0,0\}$, s.t.

$$(y,\alpha)\cdot(x,\lambda) \geq (y,\alpha)\cdot(z,\mu)$$

for all $(x,\lambda) \in C$, $(z,\mu) \in C'$.
We now prove the required inequality:

Claim 1 $\alpha \geq 0$. Indeed, for $x = z = z_0$, we get from above that $\alpha\lambda \geq \alpha\mu \iff \alpha(\lambda - \mu) \geq 0$ whenever $\lambda > F(z_0)$ and $\mu \leq m - G(z_0)$.

Claim 2 $\alpha \neq 0$ (here we will need F finite in the neighborhood of z_0). Assume $\alpha = 0$. Since $(y,\alpha) \neq (0,0)$ we have $y \neq 0$, so we have $y\cdot x \geq y\cdot z$ for all x such that $F(x) < \infty$ and z such that $G(z) < \infty$. By assumption there exists $\varepsilon > 0$ such that $F(z_0 + \varepsilon\eta) < \infty$ for all $\eta \in X$, $|\eta| = 1$. Hence $y\cdot z_0 + \varepsilon y\cdot\eta \geq y\cdot z_0$ for all $|\eta| = 1$. Thus, $y\cdot\eta \geq 0$ for all $|\eta| = 1$. And take $-\eta$, we have

$y \cdot \eta = 0$ for all $|\eta| = 1$, so $y = 0$, contradicting $y \neq 0$. It now follows that $\alpha > 0$. Dividing by α on both sides and gives with $\tilde{y} = y/\alpha$,

$$\tilde{y} \cdot x + \lambda \geq \tilde{y} \cdot z + \mu,$$

whenever $\lambda > F(x)$, $\mu < m - G(z)$. Thus,

$$\tilde{y} \cdot x + F(x) \geq \tilde{y} \cdot z + m - G(z)$$

whenever $F(x) < \infty,\ G(z) < \infty$ as equivalently,

$$F(x) + G(z) + (-\tilde{y}) \cdot (z - x) \geq m$$

for all x, z with $F(x) < \infty$, $G(z) < \infty$. If at least one of $F(x)$ and $G(z)$ equals to ∞, the inequality holds trivially, otherwise,

$$\inf_{x,z \in X} (F(x) + G(z) + (-\tilde{y})(z - x) \geq m$$

which implies

$$\sup_{y \in X^*} (-F^*(y) - G^*(-y)) = \sup_{y \in X^*} \inf_{x,z \in X} (F(x) + G(z) + y \cdot (z - x)) \geq m$$

□

Corollary 3.4: Extension of dual solution

The supremum in Fenchel-Rockefeller is attained.

Proof. Taking infimum over x, z in

$$F(x) + G(z) + y_* \cdot (z - x) \geq m,\ \forall x, z \in X,$$

gives

$$-F^*(y_*) - G^*(y_*) \geq m,$$

so y_* realizes the supremum. □

Corollary 3.5

Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper. Then $F^{**} = F$ if and only if F is lower semi-continuous and convex.

Theorem 3.6: Fenchel-Rockefeller, inifnite dimensional

Let X be a normed vector space, and X^* be its dual. $F, G : X \to \mathbb{R} \cup \{\infty\}$ be proper and convex. Assume there exists $x_0 \in X$ such that $F(x_0) < \infty$, $G(x_0) < \infty$, and at least one of F and G is continuous at x_0. Then

$$\inf_{x \in X} (F(x) + G(x)) = \sup_{y \in X^*} (-F^*(y) - G^*(-y))$$

where $F^*(y) := \sup_{x \in X} (\langle , x\rangle - F(x))$ and $\langle \cdot, \cdot \rangle$ is the dual paring between X and X^*, e.g., $X = \mathbb{R}^d = X^* : \langle y, x \rangle = y \cdot x$; X be a Hilbert space, $X^* \cong X$, $\langle y, x \rangle$ is the inner product on X.

Remark. We will need it for $X = C_o(\mathbb{R}^d)$, $X^* = \mathcal{M}(\mathbb{R}^d)$ with $\langle \mu, \alpha \rangle = \int_{\mathbb{R}^d} \varphi d\mu$.

The proof is almost identical to the proof of the finite dimensional version. However, we will need an ∞-dim version of the "separation of convex sets" result: