Overview on optimal transport

Bernhard Schmitzer

Oberwolfach, November 2023

Disclaimer

- material selected here is based on personal experience with the subject so far
- need to skip many interesting aspects, mostly skip technical aspects here
- goal is not that all details here will be understood at first pass (many of you will already know them); important: now get the big picture, can follow up on details on second pass through notes
- first parts contain no references. any textbook will provide rigorous background.

1 Introduction: Monge and Kantorovich

1.1 Monge formulation of OT

Problem statement. Given a pile of sand and a hole to fill; what is the most efficient way to fill the hole with the sand?

Mathematical modelling of problem.

- space X in which problem lives
- cost function $c : X \times X \to \mathbb{R}$, $c(x, y)$ is amount of work to move one unit of sand from x to y
- $T : X \to X$ indicates for each position x that sand is to be sent to $T(x)$
- how to describe pile and hole? as probability measures $\mu, \nu \in \mathcal{P}(X)$. mass in region $A \subset X$? $\mu(A) = \int_A \mathrm{d}\mu$

Examples for measures.

- density. let $f(x)$ be height of sand pile at x. then volume in region A given by

$$\mu(A) = \int_A f(x)\,\mathrm{d}x$$

  where $\mathrm{d}x$ denotes integration against the 'usual' Lebesgue measure

- Dirac measure δ_x, unit mass at x,

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

- combination of Diracs, $\mu := \sum_{i=1}^{N} m_i \delta_{x_i}$

Push-forward.

- if we take mass from x to $T(x)$, how does this transform pile μ?
- denote transformed measure by $T_{\#}\mu$: 'push-forward of μ under T'
- which mass particles get mapped into A?

$$T^{-1}(A) := \{x \in X \,|\, T(x) \in A\}$$

therefore, $T_{\#}\mu(A) = \mu(T^{-1}(A))$

Monge problem.

- look for map T that fills hole and causes least amount of work
- to fill the hole, need $\nu = T_{\#}\mu$
- total work associated with map T:

$$\int_X c(x, T(x))\,\mathrm{d}\mu(x)$$

- so arrive at:

$$\inf \left\{ \int_X c(x, T(x))\,\mathrm{d}\mu(x) \middle| T : X \to X, T_{\#}\mu = \nu \right\}$$

- main issue: feasible set is highly non-trivial, may even be empty

1.2 Kantorovich formulation

Transport plans.

- new concept: transport plan $\pi \in \mathcal{P}(X \times X)$
  - intuition: $\pi(x, y)$ gives (infinitesimal) amount of mass that goes from x to y
  - or: $\pi(A \times B)$ gives amount of mass that goes from $A \subset X$ to $B \subset X$
- need that π transports μ onto ν, so need

$$\pi(A \times X) = \mu(A)\,\forall\, A \subset X, \qquad \pi(X \times A) = \nu(A)\,\forall\, A \subset X.$$

- can write this with push-forward: let

$$p_i : X \times X \to X, \qquad (x_1, x_2) \mapsto x_i$$

then find

$$p_1^{-1}(A) = \left\{(x, y) \in X \times X \middle| p_1(x, y) \in A\right\} = A \times X$$

so need for all $A \subset X$ that

$$\mu(A) = \pi(A \times X) = \pi(p_1^{-1}(A)) = p_{1\#}\pi(A)$$

and therefore $p_{1\#}\pi = \mu$.

- so the set of admissible transport plans can be written as

$$\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(X \times X) | p_{1\#}\pi = \mu,\, p_{2\#}\pi = \nu\}$$

- $\Pi(\mu, \nu) \neq \emptyset$ since $\mu \otimes \nu \in \Pi(\mu, \nu)$ where

$$(\mu \otimes \nu)(A \times B) := \mu(A) \cdot \nu(B).$$

Kantorovich problem.

- cost associated with plan:
$$\int_{X\times X} c(x,y)\,\mathrm{d}\pi(x,y)$$
- Kantorovich problem:
$$\inf\left\{\int_{X\times X} c(x,y)\,\mathrm{d}\pi(x,y)\middle|\pi\in\Pi(\mu,\nu)\right\}$$
objective is linear, feasible set is non-empty, convex, 'polyhedral'

Example: discrete setting.

- let $X=\{x_1,\ldots,x_N\}$, discrete set of points
- identify $\mathcal{P}(X)$ with probability simplex
$$\sigma_N := \left\{\mu\in\mathbb{R}^N_+\middle|\sum_{i=1}^N \mu_i = 1\right\}.$$
measure represented by vector $\mu\in\sigma_N$: $\sum_i \mu_i\delta_{x_i}$
- identify $\Pi(\mu,\nu)$ with
$$\left\{\pi\in\mathbb{R}^{N\times N}_+\middle|\sum_j \pi_{i,j}=\mu_i\,\forall\, i,\ \sum_i \pi_{i,j}=\nu_j\,\forall\, j\right\}$$
for convenience introduce row- and column sum operators:
$$(P_1\pi)_i=\sum_j \pi_{i,j}, \qquad (P_2\pi)_j=\sum_i \pi_{i,j}$$
$P_i:\mathbb{R}^{N\times N}\to\mathbb{R}^N$, linear, can be represented as discrete matrices.
- $c: X\times X\to\mathbb{R}$ reduced to matrix in $\mathbb{R}^{N\times N}$, by abuse of notation: $c_{i,j}:=c(x_i,x_j)$
$$\int_{X\times X} c\,\mathrm{d}\pi = \sum_{i,j} c_{i,j}\pi_{i,j} =: \langle c,\pi\rangle$$
- arrive at finite-dimensional linear program; can in principle be solved with standard solvers; in fact: even a min cost flow problem, for which there are special, more efficient variants of the simplex algorithm

1.3 Kantorovich duality

Dual problem. We now give a formal derivation of the Kantorovich dual problem. For simplicity consider the discrete case.

- primal problem:
$$\inf_{\pi\in\mathbb{R}^{N\times N}_+} \langle c,\pi\rangle \quad \text{s.t. } P_1\pi=\mu,\ P_2\pi=\nu$$
- add Lagrange multipliers $\phi,\psi\in\mathbb{R}^N$ for constraints:
$$=\inf_{\pi\in\mathbb{R}^{N\times N}_+}\ \sup_{\phi,\psi\in\mathbb{R}^N} \langle c,\pi\rangle+\langle\phi,\mu-P_1\pi\rangle+\langle\psi,\nu-P_2\pi\rangle$$
- duality theorem for (finite-dimensional) linear programs: can swap order of inf and sup; also: re-arrange terms
$$=\sup_{\phi,\psi\in\mathbb{R}^N}\langle\phi,\mu\rangle+\langle\psi,\nu\rangle+\inf_{\pi\in\mathbb{R}^{N\times N}_+}\langle c-P_1^\top\phi-P_2^\top\psi,\pi\rangle$$

- here use transpose (or adjoint) of P_i. find for P_1:

$$\langle P_1^\top \phi, \pi \rangle_{\mathbb{R}^{N\times N}} := \langle \phi, P_1 \pi \rangle_{\mathbb{R}^N} = \sum_i \phi_i \cdot (P_1 \pi)_i =$$

$$\sum_i \phi_i \sum_j \pi_{i,j} = \sum_{i,j} \phi_i \pi_{i,j} = \sum_{i,j} (P_1^\top \phi)_{i,j} \pi_{i,j}$$

- now explicitly evaluate infimum over π, can do this entry-wise for each i, j:

$$\inf_{\pi \in \mathbb{R}_+^{N\times N}} \sum_{i,j} \hat{c}_{i,j} \pi_{i,j} = \sum_{i,j} \inf_{\pi \in \mathbb{R}_+} \hat{c}_{i,j} \cdot \pi = \begin{cases} 0 & \text{if } \hat{c}_{i,j} \geq 0, \\ -\infty & \text{else} \end{cases} = \begin{cases} 0 & \text{if } \hat{c} \geq 0, \\ -\infty & \text{else} \end{cases}$$

- so arrive at dual problem:

$$\sup_{\phi,\psi \in \mathbb{R}^N} \langle \phi, \mu \rangle + \langle \psi, \nu \rangle \quad \text{s.t.} \quad \underbrace{(\phi_i + \psi_j)}_{=:(\phi \oplus \psi)_{i,j}} \leq c_{i,j} \, \forall \, i, j$$

- continuous version:

$$\sup_{\phi,\psi : X \to \mathbb{R}} \int_X \phi \, \mathrm{d}\mu + \int_X \psi \, \mathrm{d}\nu \quad \text{s.t.} \quad \phi(x) + \psi(y) \leq c(x,y) \, \forall \, x, y$$

- A particular property of the dual problem is that it is invariant under constant shifts of the dual variables, $(\phi, \psi) \to (\phi + \lambda, \psi - \lambda)$ for $\lambda \in \mathbb{R}$. The shifted dual variables are still dual feasible and have the same objective value.

Primal-dual optimality condition.

- from previous derivation know for $\pi \in \Pi(\mu, \nu)$ and $\phi, \psi : X \to \mathbb{R}$ with $\phi \oplus \psi \leq c$ that

$$\int c \, \mathrm{d}\pi \geq \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$$

  with equality if and only if π is optimal (primal) plan and (ϕ, ψ) are dual optimal. so we have:

$$0 \leq \int c \, \mathrm{d}\pi - \int \phi \underbrace{\mathrm{d}\mu}_{=\mathrm{d}P_1\pi} - \int \psi \underbrace{\mathrm{d}\nu}_{=\mathrm{d}P_2\pi} = \int_{X\times X} \underbrace{[c(x,y) - \phi(x) - \psi(y)]}_{\geq 0} \underbrace{\mathrm{d}\pi(x,y)}_{\geq 0}$$

- so equality if and only if $c(x,y) = \phi(x) + \psi(y)$ $\pi(x,y)$-almost everywhere; in discrete case:

$$[\pi_{i,j} > 0] \quad \Leftrightarrow \quad [c_{i,j} = \phi_i + \psi_j]$$

***c*-concave functions.**

- In dual problem, for fixed ψ, find best ϕ. Since $\mu \geq 0$, try to make ϕ at each point as large as possible without violating the constraint $c \geq \phi \oplus \psi$. So set

$$\phi(x) = \inf_{y \in X} c(x,y) - \psi(y) =: \psi^c(x).$$

  This is called c-transform (assume here for simplicity that c is symmetric, otherwise need to define 'forward and backwards transform' separately).

- A function ϕ that can be written as $\phi = \psi^c$ is called c-concave. c-concavity is a strong structural property and various 'big' theorems in OT are derived from special properties of c-concave functions. For $c(x,y) = \|x - y\|^2$, ϕ is c-concave, iff $x \mapsto \|x\|^2 - \phi(x)$ is convex.

- Alternating maximization of the dual potentials ϕ and ψ is easy, but not a good optimization algorithm. In fact, $\phi^{ccc} = \phi^c$, so the iterations become stationary after three iterations, without guarantee of optimality. The auction and Sinkhorn algorithms can be interpreted as fixes of this issue, such that alternating dual maximization becomes an efficient (approximate) algorithm.

Differentiability of Kantorovich cost.

- Consider function $C : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$, $(\mu, \nu) \mapsto \inf_{\pi \in \Pi(\mu,\nu)} \int c, \mathrm{d}\pi$. Can also be written via dual:

$$C(\mu, \nu) = \sup_{\substack{\phi,\psi:X\to\mathbb{R},\\ \phi\oplus\psi\leq c}} \int \phi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$$

- This is a supremum over a collection of linear functions, parametrized by slopes (ϕ, ψ). Convex analysis: C is lower semi-continuous and sub-gradients are given by optimal (ϕ, ψ).
- For (μ, ν) let (ϕ, ψ) be optimal duals. Then

$$C(\mu + \delta\mu, \nu) \geq C(\mu, \nu) + \int \phi \mathrm{d}\delta\mu.$$

  This can in principle be used to predict small changes in transport cost; but careful: this holds for any dual optimal ϕ.
- So if $\delta\mu$ is not mean zero, can just add big λ to ϕ and make this as large as we want. Consistent: $C(\mu + \delta\mu, \nu) = \infty$ if $\mu + \delta\mu \notin \mathcal{P}(X)$: there is no feasible transport plan.
- If $\delta\mu$ has zero mean and ϕ is unique up to constant shifts (e.g. in setting of Benier theorem) then this yields a first order estimate. But no simple regularity theory as for finite-dimensional differentiable functions.

1.4 Relation between Monge and Kantorovich problem; Brenier's theorem

Kantorovich as relaxation of Monge

- for $\mu, \nu \in \mathcal{P}(X)$, assume $T : X \to X$ is such that $T_{\#}\mu = \nu$, then define measure by

$$\pi := (\mathrm{id}, T)_{\#}\mu \quad \text{where} \quad (\mathrm{id}, T) : X \to X \times X, \quad x \mapsto (x, T(x))$$

- find: $p_{1\#}\pi = p_{1\#}(\mathrm{id}, T)_{\#}\mu = [p_1 \circ (\mathrm{id}, T)]_{\#}\pi = \mathrm{id}_{\#}\, \mu = \mu$
  likewise: $p_{2\#}\pi = p_{2\#}(\mathrm{id}, T)_{\#}\mu = [p_2 \circ (\mathrm{id}, T)]_{\#}\pi = T_{\#}\mu = \nu$
  so $\pi \in \Pi(\mu, \nu)$.
- compare transport costs:

$$\int_X c(x, T(x))\, \mathrm{d}\mu(x) = \int_X [c \circ (\mathrm{id}, T)](x)\, \mathrm{d}\mu(x) = \int_{X\times X} c\, \mathrm{d}[(\mathrm{id}, T)_{\#}\mu](x, y) = \int_{X\times X} c\, \mathrm{d}\pi$$

  here we used the change of variables formula for push-forward measures
- so each Monge transport map (not necessarily optimal) induces a Kantorovich transport plan with the same cost. we find:

$$\inf_{\substack{T:X\to\mathbb{R},\\ T_{\#}\mu=\nu}} \int_X c(x, T(x))\, \mathrm{d}\mu(x) \geq \inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times X} c(x, y)\, \mathrm{d}\pi(x, y)$$

  this inequality can be strict, in particular if there are no feasible transport maps.

Brenier's theorem.

- In special cases it can be shown that the above inequality is in fact an equality and that the optimal Kantorovich plan is indeed induced by an optimal Monge map. For the special case of $X \subset \mathbb{R}^d$ with $c(x, y) = \|x - y\|^2$ this is called Brenier's theorem.
- Assume that μ has a Lebesgue density (a slightly weaker condition would suffice) and that μ an ν 'decay fast enough to zero as $\|x\| \to +\infty$'. Then the minimizing π in the Kantorovich problem is unique. It has the form $\pi = (\mathrm{id}, T)_{\#}\mu$ for a map $T : X \to X$ with $T_{\#}\mu = \nu$.
- In addition, T is the gradient of a convex potential $\phi : X \to \mathbb{R}$, i.e. $T = \nabla\phi$. (At least Lebesgue-almost everywhere, and ϕ may be non-differentiable on a μ-negligible set.)
- This can be proved via the above primal-dual relation and indeed the convex potential ϕ is very closely related to the dual Kantorovich potential $\hat{\phi}$ in the previous section via $\phi(x) = \frac{1}{2}(\|x\|^2 - \hat{\phi}(x))$.

Semi-discrete transport.

- As above, let $X \subset \mathbb{R}^d$, $c(x,y) = \|x-y\|^2$, let $\mu = f \cdot \mathcal{L}$ (where $\mathcal{L}$ denotes the Lebesgue measure) and $\nu = \sum_i m_i \delta_{y_i}$.
- By Brenier's theorem, the optimal π is unique and induced by a map T, $\pi = (\mathrm{id}, T)_{\#}\mu$. Clear: T must be piecewise constant, mapping regions $C_i \subset X$ to the mass locations y_i.
- This might model assignments of students (a lot of them, essentially 'continuously' distributed over the country or city) to schools (discrete centers), or customers to supermarkets and similar situations with continuous demand and discrete supply.
- Locally, at first, customers simply want to go to the nearest market, where $c(x, y_i) = \min_j c(x, y_j)$. So customers in

  $$V_i := \{x \in X | c(x, y_i) = \min_j c(x, y_j)\}$$

  would all go to market i. The sets $\{V_i\}_i$ are called Voronoi cells. But then the demand $\mu(V_i)$ might not match the supply m_i. Economically, we need to introduce prices to act as additional incentive in addition to travel distance, to influence customers decisions. If $\mu(V_i) > m_i$, then the market i should raise its prices to 'repulse' some customers, if $\mu(V_i) < m_i$, then lower the prices to attract near customers from other nearby markets. OT provides a natural description of this phenomenon.
- By the primal-dual optimality relation, one has $c = \phi \oplus \psi$ π-almost everywhere. The optimal ϕ can be written as c-transform of ψ, so get

  $$\phi(x) = \psi^c(x) = \min_i c(x, y_i) - \psi_i.$$

  For a given x, mass can only be transported to some y_i that is minimizing in this expression. I.e. location i can only receive mass from within the region

  $$C_i(\psi) := \{x \in X | c(x, y_i) - \psi_i = \min_j c(x, y_j) - \psi_j\}.$$

  Now we see that $-\psi$ acts as the aforementioned price.
- For the squared distance, the above condition can be written as

  $$x \in C_i(\psi) \quad \text{iff} \quad \langle x, y_i \rangle - \tfrac{1}{2}(\|y_i\|^2 - \psi_i) = \max_j \langle x, y_j \rangle - \tfrac{1}{2}(\|y_j\|^2 - \psi_j).$$

  From this we deduce that the boundaries between adjacent cells are straight lines and that in fact each cell is a convex polytope.
- Eliminating ϕ via the c-transform, one can obtain a finite-dimensional unconstrained dual problem in terms of ψ:

  $$\sup_{\psi \in \mathbb{R}^N} J(\psi) \quad \text{where} \quad J(\psi) := \int \psi^c \mathrm{d}\mu + \sum_i \psi_i m_i$$

  The integral can be written as

  $$\sum_i \int_{C_i(\psi)} \left[\|x - y_i\|^2 - \psi_i\right] \mathrm{d}\mu(x).$$

  One can show that J is differentiable and

  $$\partial_{\psi_i} J(\psi) = -\int_{C_i(\psi)} \mathrm{d}\mu + m_i.$$

  This corresponds nicely with the economic interpretation: if the demand $\mu(C_i)$ is lower than the supply m_i, the partial derivative for ψ_i is positive.
- Note that if we add the same constant λ to all prices, then all C_i remain the same. Economically, this invariance is not so realistic. Eventually, customers will stop going to the market entirely and try to find alternatives. This setting might be modelled more appropriately with unbalanced transport.

2 Wasserstein distance

2.1 Definition and metric axioms

Definition.

- Let (X,d) be a metric space, e.g. $X \subset \mathbb{R}^d$, $d(x,y) = \|x-y\|^2$ or a curved surface.
- Let $p \in [1,\infty)$, $\mu,\nu \in \mathcal{P}(X)$. Set:

$$W_p(\mu,\nu) := \inf \left\{ \int_{X\times X} d^p(x,y)\,\mathrm{d}\pi(x,y) \middle| \pi \in \Pi(\mu,\nu) \right\}^{1/p}$$

- claim: W_p is a metric on $\mathcal{P}(X)$, called Wasserstein distance (technically: a metric on probability measures with finite p-th moment)

Symmetry, definiteness.

- Since $d(x,y) \geq 0$, and $\pi \geq 0$ for all admissible plans, one has $W_p(\mu,\nu) \geq 0$.
- Symmetry, $W_p(\mu,\nu) = W_p(\nu,\mu)$: let $\pi \in \Pi(\mu,\nu)$, set $\hat{\pi} := (p_2,p_1)_\#\pi$ for which ones finds $\hat{\pi} \in \Pi(\nu,\mu)$. Further:

$$\int d^p\,\mathrm{d}\hat{\pi} = \int d^p \circ (p_2,p_1)\mathrm{d}\pi = \int d^p \mathrm{d}\pi$$

  where we used the symmetry of d. Consequently, for any plan in $\Pi(\mu,\nu)$, the 'reversed plan' is in $\Pi(\nu,\mu)$, has the same transport cost, and vice versa. Hence, both problem have the same infimal value.

- Definiteness, part 1: $W_p(\mu,\mu) = 0$. In this case, $\pi = (\mathrm{id},\mathrm{id})_\#\mu$ is a feasible transport plan. We get:

$$0 \leq W_p(\mu,\mu)^p \leq \int_{X\times X} d^p\,\mathrm{d}\pi = \int_X d^p \circ (\mathrm{id},\mathrm{id})\,\mathrm{d}\mu = \int_X d(x,x)^p\,\mathrm{d}\mu(x) = 0$$

- Definiteness, part 2: $[W_p(\mu,\nu) = 0] \Rightarrow [\mu = \nu]$. Let π be an optimal plan for $W_p(\mu,\nu)$ (do not address existence here, argument can also be made with 'almost optimal' plans). Then:

$$0 = \int_{X\times X} d(x,y)^p\,\mathrm{d}\pi(x,y)$$

  Since $d \geq 0$ and $\pi \geq 0$, this is only possible if $d(x,y) = 0$ $\pi(x,y)$-almost everywhere. So π must live exclusively on the 'diagonal' and thus have the form $\pi = (\mathrm{id},\mathrm{id})_\#\rho$ for some $\rho \in \mathcal{P}(X)$. Now use $\pi \in \Pi(\mu,\nu)$: $\mu = p_{1\#}\pi = \rho = p_{2\#}\pi = \nu$.

2.2 Triangle inequality

The metric d must satisfy

$$d(x,y) + d(y,z) \geq d(x,z) \quad \text{for all} \quad x,y,z \in X.$$

The same must hold for W_p on triplets of probability measures $\mu,\nu,\rho \in \mathcal{P}(X)$.

Simple proof for transport maps.

- We sketch the proof for the simple case where the optimal plans for $W_p(\mu,\nu)$ and $W_p(\nu,\rho)$ are induced by maps $S,T : X \to X$, $S_\#\mu = \nu$, $T_\#\nu = \rho$.

- In this case, $(T \circ S)_{\#}\mu = \rho$ and so $T \circ S$ is a feasible (not necessarily optimal) transport map from μ to ρ for $W_p(\mu, \rho)$. Therefore we get

$$\begin{aligned}
W_p(\mu,\rho) &\leq \left[\int d(x, T(S(x)))^p \,\mathrm{d}\mu(x)\right]^{1/p} \\
&\leq \left[\int [d(x,S(x)) + d(S(x), T(S(x)))]^p \,\mathrm{d}\mu(x)\right]^{1/p} \\
&\leq \left[\int d(x,S(x))^p \,\mathrm{d}\mu(x)\right]^{1/p} + \left[\int d(S(x), T(S(x)))^p \,\mathrm{d}\mu(x)\right]^{1/p} \\
&= W_p(\mu,\nu) + \left[\int d(x,T(x))^p \,\mathrm{d}(S_{\#}\mu)(x)\right]^{1/p} = W_p(\mu,\nu) + W_p(\nu,\rho)
\end{aligned}$$

Gluing lemma.

- The proof strategy also works when the optimal plans are not map-like. Will not go through all steps, but merely show how two plans $\pi \in \Pi(\mu,\nu)$, $\lambda \in \Pi(\nu,\rho)$ can be 'composed'.
- We will construct a measure $\eta \in \mathcal{P}(X \times X \times X)$ with

$$(p_1, p_2)_{\#}\eta = \pi, \qquad\qquad (p_2, p_3)_{\#}\eta = \lambda.$$

  $\eta(x, y, z)$ can intuitively be interpreted as (infinitesimal) mass going from x to z via y.
- For simplicity, do this in discrete setting, but same idea works in continuum with slightly more complex notation, using the disintegration theorem. Denote mass at individual points by $\mu(x)$, $\pi(x, y)$, et cetera. Then $\pi \in \Pi(\mu,\nu)$ implies $\sum_y \pi(x, y) = \mu(x)$ and so forth.
- Consider mass arriving in y after being transported according to π. How should we send the mass along? Use 'conditional probability'

$$\lambda_y(z) := \frac{\lambda(y,z)}{\nu(y)}$$

  to distribute mass. (In the continuous setting, this conditional probability is provided by the disintegration.)
- Then we set

$$\eta(x,y,z) := \pi(x,y)\frac{\lambda(y,z)}{\nu(y)} = \frac{\pi(x,y)\lambda(y,z)}{\nu(y)}$$

  (Be a little careful when $\eta(y) = 0$. In this case no mass will arrive at y and no mass will leave. So simply set $\eta(x, y, z) = 0$.)
- Then find

$$\sum_z \eta(x,y,z) = \sum_z \frac{\pi(x,y)\lambda(y,z)}{\nu(y)} = \frac{\pi(x,y)}{\nu(y)} \sum_z \lambda(y,z) = \pi(x,y).$$

  Likewise, find $\sum_x \eta(x, y, z) = \lambda(y, z)$ and thus $\sum_y \eta(\cdot, y, \cdot) \in \Pi(\mu, \rho)$.
- Some things to take away:
  - this construction also works in continuum, via disintegration
  - evaluating Wasserstein cost of $\sum_y \eta(\cdot, y, \cdot)$ yields triangle inequality, similar to above
  - perfectly reasonable to have 'transport plans' between more than two marginals

2.3 Shortest paths

For $x, y \in X$, a curve $\gamma_{(x,y)} : [0, 1] \to X$ is called (constant speed) geodesic from x to y if

$$\gamma_{(x,y)}(0) = x, \qquad \gamma_{(x,y)}(1) = y, \qquad d(\gamma_{(x,y)}(s), \gamma_{(x,y)}(t)) = |s-t| \cdot d(x,y) \,\forall\, s,t \in [0,1].$$

We will call (X, d) geodesic, if such curves exist for all pairs (x, y). We will now show: If (X, d) is geodesic, then so is $(\mathcal{P}(X), W_p)$.

Constructing geodesics.

- For simplicity, assume $X \subset \mathbb{R}^d$, $d(x,y) = \|x-y\|$, and let the optimal plan π be induced by a map T, $\pi = (\mathrm{id}, T)_{\#}\mu$. Easy to generalize, mostly just more complex notation.
- Intuition for geodesic in $\mathcal{P}(X)$ from μ to ν: each mass particle moves along geodesic in X from initial to final position. Particle at x moves to $T(x)$, so it will move on curve

$$\gamma_{(x,T(x))} : [0,1] \to X, \qquad \gamma_{(x,T(x))}(t) := (1-t)\cdot x + t \cdot T(x)$$

  Taking each particle from its initial position x to $\gamma_{(x,T(x))}(t)$ means we apply the push-forward of

$$f_t : X \to X, \qquad f_t(x) := \gamma_{(x,T(x))}(t)$$

  to μ. So our conjectured geodesic takes the form

$$\rho : [0,1] \to \mathcal{P}(X), \qquad \rho(t) := f_{t\#}\mu.$$

  It is easy to verify: $\rho(0) = \mu$, $\rho(1) = \nu$.

Estimating distances along geodesic.

- How do we estimate the distance between $\rho(0) = \mu$ and $\rho(s)$? Need to guess a transport map. Clear: particle from x moves to $f_t(x) = \gamma_{(x,T(x))}(t)$, so use this as transport map candidate: $T_{0,s} := f_s$. We find:

$$W_p(\mu, \rho(s))^p \le \int \| \underbrace{T_{0,s}(x) - x}_{=s(T(x)-x)} \|^p \, \mathrm{d}\mu(x) = s^p \int \|T(x) - x\|^p \mathrm{d}\mu(x) = s^p W_p(\mu,\nu)^p$$

- How about optimal map from $\rho(s)$ to $\rho(t)$? Mass from $f_s(x)$ moves to $f_t(x)$. For $s < 1$ can show: f_s is invertible. So use as candidate: $T_{s,t} := f_t \circ f_s^{-1}$. Find:

$$\begin{aligned} W_p(\rho(s), \rho(t))^p &\le \int \|T_{s,t}(x) - x\|^p \mathrm{d}\rho(s) = \int \|f_t \circ f_s^{-1} - \mathrm{id}\,\|^p \mathrm{d}(f_{s\#}\mu) \\ &= \int \|f_t - f_s\|^p \mathrm{d}\mu = |t-s|^p \int \|T - \mathrm{id}\,\|^p \mathrm{d}\mu = |t-s|^p W_p(\mu,\nu)^p. \end{aligned}$$

- With same arguments can bound

$$W_p(\rho(t), nu)^p \le (1-t)^p W_p(\mu,\nu)^p.$$

- If any of these inequalities were strict, we would violate the triangle inequality. The above inequalities provide

$$W_p(\mu, \rho(s)) + W_p(\rho(s), \rho(t)) + W_p(\rho(t), \nu) \le W_p(\mu, \nu)$$

  whereas the triangle inequality gives the opposite inequality. Thus, all inequalities above must be equalities.
- The proof strategy generalizes to non-map transport plans and more general metric spaces.

2.4 Continuity equation and Benamou–Brenier formula

Continuity equation.

- Let $X \subset \mathbb{R}^d$, $d(x,y) = \|x-y\|$, set $p = 2$ for simplicity. Consider a curve of measures $\rho(t) := \gamma(t,\cdot)_{\#}\mu$, assume γ differentiable in time, $\gamma(t,\cdot)$ invertible.

- A particle starting at $t = 0$ at x will move with along the path $t \mapsto \gamma(t, x)$. Therefore its (Lagrangian) velocity is $\dot{\gamma}(t, x)$. But at time t it is at $\gamma(t, x)$, so an external observer (who cannot distinguish the mass particles) will see the (Eulerian) velocity field

$$v(t, \cdot) := \dot{\gamma}(t, \cdot) \circ \gamma(t, \cdot)^{-1}.$$

- If all mass particles of ρ follow a (Eulerian) velocity field v, (ρ, v) will satisfy the continuity equation

$$\partial_t \rho + \operatorname{div}(v \cdot \rho) = 0,$$

in our case with the boundary conditions $\rho(0) = \mu$ and $\rho(1) =: \nu := \gamma(1, \cdot)_{\#}\mu$.

- This equation is to be understood in a distributional sense. This means that for every differentiable test function $\phi \in C^1([0, 1] \times X)$ one imposes that

$$\int_0^1 \int_X \partial_t \phi(t, \cdot) \, \mathrm{d}\rho(t) \, \mathrm{d}t + \int_0^1 \int_X \nabla \phi(t, \cdot) \cdot v(t, \cdot) \, \mathrm{d}\rho(t) \, \mathrm{d}t = \int_X \phi(1, \cdot) \, \mathrm{d}\nu - \int_X \phi(0, \cdot) \, \mathrm{d}\mu$$

- Now we show that $\rho(t) := \gamma(t, \cdot)_{\#}\mu$ and $v(t, \cdot) := \dot{\gamma}(t, \cdot) \circ \gamma(t, \cdot)^{-1}$ solve the continuity equation. Set $F(t) := \int_X \phi(t, \cdot) \, \mathrm{d}\rho(t)$. Then

$$\int_0^1 \left[\frac{\mathrm{d}}{\mathrm{d}t} F(t) \right] \mathrm{d}t = F(1) - F(0)$$

and

$$F(1) = \int \phi(1, \cdot) \mathrm{d}\nu, \qquad F(0) = \int \phi(0, \cdot) \mathrm{d}\mu.$$

For $F(t)$ one has

$$F(t) = \int_X \phi(t, \cdot) \, \mathrm{d}(\gamma(t, \cdot)_{\#}\mu) = \int_X \phi(t, \gamma(t, \cdot)) \, \mathrm{d}\mu$$

and so

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F(t) &= \int_X [\partial_t \phi(t, \gamma(t, \cdot)) + \nabla \phi(t, \gamma(t, \cdot)) \cdot \dot{\gamma}(t, \cdot)] \, \mathrm{d}\mu \\ &= \int_X \Big[\partial_t \phi(t, \cdot) + \nabla \phi(t, \cdot) \cdot \underbrace{\dot{\gamma}(t, \gamma(t, \cdot)^{-1})}_{=v(t, \cdot)} \Big] \mathrm{d}\rho(t) \end{aligned}$$

Benamou–Brenier formula.

- Geodesics in $(\mathcal{P}(\mathbb{R}^d), W_2)$ are of the above form and therefore solve the continuity equation. Question: which of the solutions of the continuity equation yields the shortest path? Answer: the one with the lowest time-average kinetic energy (in this case also called: action).

- For a pair (ρ, v) that solves the continuity equation, set

$$BB(\rho, v) := \int_0^1 \int_X \|v(t, \cdot)\|^2 \, \mathrm{d}\rho(t) \, \mathrm{d}t.$$

Then one has

$$W_2(\mu, \nu)^2 = \inf \{ BB(\rho, v) | (\rho, v) \text{ solve CE between } \mu \text{ and } \nu \}$$

- Sketch of inequality $BB \leq W$: take (ρ, v) generated by shortest path. Recall: $\gamma(t, \cdot) = (1-t) \operatorname{id} + t \cdot T$. Lagrangian velocity: $\dot{\gamma}(t, \cdot) = T - \operatorname{id}$. Have already shown that (ρ, v) solve the continuity equation. Now plug into BB to get:

$$\begin{aligned} \inf BB \leq BB(\rho, v) &= \int_0^1 \int_X \|v(t, \cdot)\|^2 \mathrm{d}\rho(t) \, \mathrm{d}t = \int_0^1 \int_X \|\dot{\gamma}(t, \cdot) \circ \gamma(t, \cdot)^{-1}\|^2 \mathrm{d}\gamma(t, \cdot)_{\#}\mu \mathrm{d}t \\ &= \int_0^1 \int_X \|\dot{\gamma}(t, \cdot)\|^2 \, \mathrm{d}\mu \mathrm{d}t = \int_0^1 \int_X \|T - \operatorname{id}\|^2 \, \mathrm{d}\mu \mathrm{d}t = W_2(\mu, \nu)^2 \end{aligned}$$

- For converse inequality, for given (ρ, v) follow the individual mass particles. Let $\gamma(t, x)$ be the solution of the following ODE:

$$\gamma(0, x) = x, \qquad \dot{\gamma}(t, x) = v(t, \gamma(t, x))$$

This means, we recover the Lagrangian coordinate from the Eulerian velocity field. (Mathematically this only works if v if is sufficiently regular. Need some smoothing arguments in rigorous proof.) Can then show with similar calculations as above: $\rho(t) = \gamma(t, \cdot)_{\#}\rho(0) = \gamma(t, \cdot)_{\#}\mu$. Now look at BB functional:

$$\begin{aligned}
BB(\rho, v) &= \int_0^1 \int_X \|v(t, \cdot)\|^2 \mathrm{d}\rho(t)\, \mathrm{d}t = \int_0^1 \int_X \|v(t, \gamma(t, \cdot))\|^2 \mathrm{d}\mu\, \mathrm{d}t \\
&= \int_X \int_0^1 \|v(t, \gamma(t, \cdot))\|^2 \mathrm{d}t\, \mathrm{d}\mu = \int_X \int_0^1 \|\dot{\gamma}(t, \cdot))\|^2 \mathrm{d}t\, \mathrm{d}\mu \geq \int_X \left\| \int_0^1 \dot{\gamma}(t, \cdot)) \mathrm{d}t \right\|^2 \mathrm{d}\mu \\
&\geq \int_X \|\gamma(1, \cdot) - \gamma(0, \cdot)\|^2 \, \mathrm{d}\mu \geq W_2(\mu, \nu)^2
\end{aligned}$$

where we used Jensen's inequality when pulling out $\| \cdot \|^2$ from the integral, and the fact that $\gamma(0, \cdot) = \mathrm{id}$ and that $\gamma(1, \cdot)$ is a feasible transport map from μ to ν.

3 Entropic optimal transport

3.1 Introduction

Definition 3.1 (Setting)**.**

- X will be a compact metric space, $C(X)$ denotes continuous (real-valued) functions
- $\mathcal{M}(X)$, $\mathcal{M}_+(X)$, $\mathcal{P}(X)$ will be (signed) Radon measures, non-negative measures, and probability measures, respectively
- We will use a lot the duality between continuous functions and measures on X.
- $c \in C(X \times X)$ will be a continuous cost function
- p_1, p_2 denote the marginal projection operators:

$$\int_X \phi(x) \mathrm{d}(p_1\gamma)(x) := \int_{X\times X} \phi(x) \mathrm{d}\gamma(x,y)$$

  correspond to push-forward by map $(x_1, x_2) \mapsto x_i$.

- For finite spaces $X = \{x_1, \ldots, x_n\}$, we usually identify $\mathcal{M}(X) \simeq C(X) \simeq \mathbb{R}^n$ (and likewise with product spaces). Then for $\mu \in \mathcal{M}(X)$ we denote by μ_i the mass at x_i, et cetera.

Definition 3.2 (KL divergence)**.** Let

$$\varphi : \mathbb{R} \to [0, \infty], \qquad s \mapsto \begin{cases} s\log(s) - s + 1 & \text{for } s > 0, \\ 1 & \text{for } s = 0, \\ +\infty & \text{for } s < 0. \end{cases}$$

Note that φ is convex, proper and lower-semicontinuous. Then for $\mu, \nu \in \mathcal{M}(X)$, the Kullback–Leibler (KL) divergence of μ w.r.t. ν is given by

$$\mathrm{KL}(\mu|\nu) := \begin{cases} \int_X \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu & \text{if } \mu \ll \nu,\ \nu \geq 0, \\ +\infty & \text{else.} \end{cases}$$

Definition 3.3 (Entropic transport problem)**.** For $\mu, \nu \in \mathcal{P}(X)$, a continuous cost function $c \in C(X \times X)$, regularization strength $\varepsilon > 0$ and a reference measure $\rho \in \mathcal{M}_+(X \times X)$, the entropic transport problem is given by

$$C_\varepsilon(\mu, \nu) := \inf \left\{ \int_{X\times X} c \,\mathrm{d}\gamma + \varepsilon \,\mathrm{KL}(\gamma|\rho) \middle| \gamma \in \Gamma(\mu, \nu) \right\}$$

Remark 3.4 (Motivation)**.**

- Will show: entropic transport problem has unique solution; gives stability of solutions w.r.t. fluctuations in marginals or cost function; even differentiability. Useful for downstream applications.
- Can be solved efficiently with simple numerical method: Sinkhorn algorithm
- Will allow for more reliable statistical estimation of optimal transport between sampled empirical measures. (Beyond the scope of this course.)

Remark 3.5 (Choice of reference measure)**.** There are many different potential choices for the reference measure $\rho \in \mathcal{M}_+(X \times X)$. Common choices are:

(i) For a finite space $X = \{x_1, \ldots, x_n\}$ one often uses the Shannon entropy, i.e. one sets ρ to be the counting measure, $\rho_{i,j} = 1$.

(ii) In some applications, if $X \subset \mathbb{R}^d$, it may be natural to choose $\rho = \mathcal{L}^{2d} \llcorner (X \times X)$. This will only work well, if all measures of interest μ and ν are dominated by the Lebesgue measure.

(iii) The most agnostic choice is probably $\rho = \mu \otimes \nu$. Then, the optimal objective is always finite (for bounded cost functions) and we will show that regular dual solutions exist.

Lemma 3.6. KL is jointly weak* lower-semicontinuous and convex in both arguments. For fixed $\nu \in \mathcal{M}_+(X)$, $\mu \mapsto \mathrm{KL}(\mu|\nu)$ is strictly convex.

Proof. Joint lsc follows from [Ambrosio *et al.*, 2000, Theorem 2.34] since the function φ is non-negative, convex and lower-semicontinuous. Join convexity follows from the fact that the function $(r, s) \mapsto s \log(s/r) - s + r$. Strict convexity in the first argument follows from strict convexity of φ. □

Lemma 3.7. If the infimal objective is finite, the entropic optimal transport problem has a unique solution.

Proof. Since $\Gamma(\mu, \nu)$ is bounded (and weak* closed), any minimizing sequence has weak* cluster points that also lie in $\Gamma(\mu, \nu)$. Clearly, the map $\gamma \mapsto \int_{X \times X} c \mathrm{d}\gamma$ is continuous. By the above Lemma the KL term is lower-semicontinuous. Hence, any cluster point must be a minimizer.
Assume the objective is finite. Let $\gamma_1, \gamma_2 \in \Gamma(\mu, \nu)$ be two minimizers, and therefore have $\mathrm{KL}(\gamma_i|\rho) < \infty$. By linearity of the term $\gamma \mapsto \int_{X \times X} c \mathrm{d}\gamma$ and strict convexity of the KL-term, if $\gamma_1 \neq \gamma_2$, then $(\gamma_1 + \gamma_2)/2$ would be an even better candidate. Hence $\gamma_1 = \gamma_2$ and the minimizer must be unique. □

Remark 3.8. Choose $c(x, y) = d(x, y)^p$ for a metric d on X and $\varepsilon > 0$. Then the map

$$W_\varepsilon : \mathcal{P}(X)^2 \ni (\mu, \nu) \mapsto C_\varepsilon(\mu, \nu)^{1/p}$$

is no longer a metric on $\mathcal{P}(X)$. (Unless something really boring happens, like X being only a single point and choosing ρ to be the probability measure on that single point.)
In general one finds that $W_\varepsilon(\mu, \mu) > 0$ and that it violates the triangle inequality. The former issue (along with something called 'entropic bias') can be fixed by the Sinkhorn divergence [Feydy *et al.*, 2018]. While they do not satisfy the triangle inequality, they are a useful (and statistically robust) notion of similarity in many applications. The question of a metric induced by entropic optimal transport is (to my knowledge) still open.

3.2 Convergence as $\varepsilon \to 0$

Definition 3.9 (Setting)**.** Let $(\varepsilon_n)_{n \in \mathbb{N}}$ be a strictly positive, decreasing sequence with limit $\lim_{n \to \infty} \varepsilon_n = 0$. Let $\mu, \nu \in \mathcal{P}(X)$ and choose as reference measure $\rho = \mu \otimes \nu$. Set

$$\begin{aligned} E_n(\gamma) &:= \int_{X \times X} c \,\mathrm{d}\gamma + \varepsilon_n \,\mathrm{KL}(\gamma|\mu \otimes \nu) + \iota_{\Gamma(\mu,\nu)}(\gamma), \\ E(\gamma) &:= \int_{X \times X} c \,\mathrm{d}\gamma + \iota_{\Gamma(\mu,\nu)}(\gamma). \end{aligned}$$

Let $(\gamma_n)_{n \in \mathbb{N}}$ be a sequence of minimizers of E_n (existence shown above, with finite optimal objective). We will now show that, up to selection of subsequences, $(\gamma_n)_n$ converges to some γ that minimizes E, and that the optimal objectives converge.

Proposition 3.10 (Finite space)**.** Let $X = \{x_1, \dots, x_n\}$ be a finite space. Then the sequence $(\gamma_n)_{n \in \mathbb{N}}$ is precompact and any cluster point γ minimizes E. The optimal objectives converge.

Proof. For simplicity, w.l.o.g. assume that μ and ν are strictly positive (otherwise, simply remove points x_i where $\mu_i = 0$ from the first marginal space, and similarly with the second marginal). Then ρ is strictly positive, and by finiteness of X, ρ is bounded away from zero by some finite constant. Therefore, by continuity of φ on its domain $[0, \infty]$, the domain of

$$\mathbb{R}^{n \times n} \ni \gamma \mapsto \mathrm{KL}(\gamma|\rho) = \sum_{i,j} \varphi(\gamma_{i,j}/\rho_{i,j}) \cdot \rho_{i,j}$$

is $\mathbb{R}_+^{n \times n}$ and it is continuous on this domain. The set $\Gamma(\mu, \nu) \subset \mathbb{R}_+^{n \times n}$ is compact and therefore $\mathrm{KL}(\cdot|\rho)$ is bounded on $\Gamma(\mu, \nu)$ (say, by some constant $C < \infty$). Therefore, $E_n(\gamma_n) - E(\gamma_n) \in [0, \varepsilon_n \cdot C]$.

Since $\Gamma(\mu,\nu)$ is compact and non-empty, $(\gamma_n)_n$ will have cluster points (that also lie in $\Gamma(\mu,\nu)$). Let now γ be any such cluster point, for simplicity denote by $(\gamma_n)_n$ a convergent subsequence. Clearly, E is continuous on $\Gamma(\mu,\nu)$. Therefore we find:

$$\lim_n E_n(\gamma_n) = \lim_n E(\gamma_n) = E(\gamma)$$

Finally, if there were some $\tilde{\gamma} \in \Gamma(\mu,\nu)$ with $E(\tilde{\gamma}) < E(\gamma)$, then for sufficiently big n (and some suitable $\delta > 0$) we would have

$$E_n(\tilde{\gamma}) = E(\tilde{\gamma}) + \varepsilon_n \cdot \mathrm{KL}(\tilde{\gamma}|\mu \otimes \nu) \le E(\tilde{\gamma}) + \varepsilon_n \cdot C < E(\gamma) - \delta \le E(\gamma_n) \le E_n(\gamma_n)$$

which contradicts the optimality of γ_n for E_n. Therefore, γ must minimize E. □

For continuous X the situation is more involved, since $\mathrm{KL}(\cdot|\mu\otimes\nu)$ will in general not be bounded (or even finite) on $\Gamma(\mu,\nu)$ and might be infinite for minimizers of E. We will show convergence of minimizers by means of Γ-convergence. Much more general versions of the following result are possible (on non-compact spaces, with less regular cost functions, and with approximate marginals). We focus on a few key properties of the problem here.

Lemma 3.11. Let $(\gamma_n)_n$ be a sequence in $\mathcal{M}(X \times X)$ that converges weak* to $\gamma \in \mathcal{M}(X \times X)$. Then

$$\liminf_n E_n(\gamma_n) \ge E(\gamma).$$

Proof. Since $\Gamma(\mu,\nu)$ is weak* closed, if $\gamma \notin \Gamma(\mu,\nu)$, we will eventually have that $\gamma_n \notin \Gamma(\mu,\nu)$ and thus $E(\gamma) = E_n(\gamma_n) = \infty$ for sufficiently large n. So assume $\gamma \in \Gamma(\mu,\nu)$ from now on.
Then, using weak* continuity of the linear cost term and non-negativity of the entropy term, we obtain

$$\begin{aligned}\liminf_n E_n(\gamma_n) &= \liminf_n \int_{X\times X} c\,\mathrm{d}\gamma_n + \varepsilon_n \cdot \mathrm{KL}(\gamma_n|\mu\otimes\nu)\\ &\ge \lim_n \int_{X\times X} c\,\mathrm{d}\gamma_n = E(\gamma).\end{aligned}$$

□

The lim-sup inequality is considerably more involved. For $\gamma \in \Gamma(\mu,\nu)$ with $\mathrm{KL}(\gamma|\mu\otimes\nu) = \infty$ we need to construct an approximating sequence $(\gamma_n)_n$ with finite entropy (diverging in a controlled way) while preserving the marginals. We do this here via the *block approximation trick* [Carlier *et al.*, 2017].

Definition 3.12 (Block approximation)**.** Let $\gamma \in \Gamma(\mu,\nu)$. For a length scale $L > 0$, denote by $\{X_{L,i}\}_{i=1}^{n_L}$ a (measurable) partition of X into n_L sets, each of which with diameter at most L. Such a partition exists by compactness of X. Denote in the following

$$\mu_{L,i} := \mu(X_{L,i}), \qquad \nu_{L,i} := \nu(X_{L,i}), \qquad \gamma_{L,i,j} := \gamma(X_{L,i} \times X_{L,j}),$$

and finally

$$\lambda_{L,i,j} := \begin{cases} \frac{\mu \llcorner X_{L,i} \otimes \nu \llcorner X_{L,j}}{\mu_{L,i}\cdot\nu_{L,j}} & \text{if } \mu_{L,i}\cdot\nu_{L,j} > 0,\\ 0 & \text{else.}\end{cases}$$

Then the *block approximation* of γ at scale L is given by

$$\gamma_L := \sum_{i,j=1}^{n_L} \gamma_{L,i,j} \cdot \lambda_{L,i,j}.$$

Lemma 3.13. $\gamma_L \in \Gamma(\mu,\nu)$.

Proof. First observe: $\gamma_L \geq 0$. Next, observe that

$$\sum_{j=1}^{n_L} \gamma_{L,i,j} = \sum_{j=1}^{n_L} \gamma(X_{L,i} \times X_{L,j}) = \gamma(X_{L,i} \times X) = \mu(X_{L,i}) = \mu_{L,i}.$$

In particular, this implies $\gamma_{L,i,j} > 0 \Rightarrow \mu_{L,i} > 0$. And of course likewise for the other marginal. Now, for any measurable $A \subset X$ one has

$$\begin{aligned}
\gamma_L(A \times X) &= \sum_{i,j=1}^{n_L} \gamma_{L,i,j} \cdot \lambda_{L,i,j}(A \times X) \\
&= \sum_{\substack{i,j=1,\dots,n_L: \\ \gamma_{L,i,j}>0}} \frac{\gamma_{L,i,j}}{\mu_{L,i} \cdot \nu_{L,j}} \cdot \mu(X_{L,i} \cap A) \cdot \nu(X_{L,j} \cap X) \\
&= \sum_{\substack{i,j=1,\dots,n_L: \\ \gamma_{L,i,j}>0}} \frac{\gamma_{L,i,j}}{\mu_{L,i}} \cdot \mu(X_{L,i} \cap A) \\
&= \sum_{\substack{i=1,\dots,n_L: \\ \mu_{L,i}>0}} \mu(X_{L,i} \cap A) = \mu(A).
\end{aligned}$$

The same computation applies for the second marginal, which completes the proof. ☐

Lemma 3.14. Equip $X \times X$ with the metric $D((x_1, x_2), (y_1, y_2)) := d(x_1, y_1) + d(x_2, y_2)$ (which yields a compact metric space). Then $W_p(\gamma, \gamma_L) \leq 2L$ and in particular $\gamma_L \xrightharpoonup{*} \gamma$ as $L \to 0$.

Proof. A potential transport plan from γ to γ_L involves moving mass only within products of partition cells $X_{L,i} \times X_{L,j}$, which have diameter bounded by $2L$ in D. This yields the Wasserstein bound, which implies the weak* convergence. ☐

Lemma 3.15. $\mathrm{KL}(\gamma_L | \mu \otimes \nu) \leq 2 \log(n_L)$.

Proof.

$$\begin{aligned}
\mathrm{KL}(\gamma_L | \mu \otimes \nu) &= \int_{X \times X} \varphi\left(\tfrac{\mathrm{d}\gamma_L}{\mathrm{d}\mu \otimes \nu}\right) \mathrm{d}\mu \otimes \nu \\
&= \sum_{\substack{i=1,\dots,n_L: \\ \mu_{L,i}>0}} \sum_{\substack{j=1,\dots,n_L: \\ \nu_{L,j}>0}} \varphi\left(\frac{\gamma_{L,i,j}}{\mu_{L,i} \cdot \nu_{L,j}}\right) \cdot \mu_{L,i} \cdot \nu_{L,j} \\
&= \sum_{\substack{i=1,\dots,n_L: \\ \mu_{L,i}>0}} \sum_{\substack{j=1,\dots,n_L: \\ \nu_{L,j}>0}} \left[\gamma_{L,i,j} \cdot \log\left(\frac{\gamma_{L,i,j}}{\mu_{L,i} \cdot \nu_{L,j}}\right) - \gamma_{L,i,j} + \mu_{L,i} \cdot \nu_{L,j}\right] \\
&\leq - \sum_{\substack{i=1,\dots,n_L: \\ \mu_{L,i}>0}} \sum_{\substack{j=1,\dots,n_L: \\ \nu_{L,j}>0}} \gamma_{L,i,j} \cdot [\log(\mu_{L,i}) + \log(\nu_{L,j})] \\
&\leq - \sum_{\substack{i=1,\dots,n_L: \\ \mu_{L,i}>0}} \mu_{L,i} \cdot \log(\mu_{L,i}) - \sum_{\substack{j=1,\dots,n_L: \\ \nu_{L,j}>0}} \nu_{L,j} \cdot \log(\nu_{L,j}) \\
&\leq -2 \log(1/n_L) = 2 \log(n_L)
\end{aligned}$$

where we used that $\mathbb{R}_+^{n_L} \ni p \mapsto \sum_{i=1}^{n_L} p_i \log(p_i)$ is convex and minimized among 'probability vectors' by the 'uniform' one $p_i = 1/n_L$. ☐

Using now the block approximation and its basic properties that we established, we conclude:

Lemma 3.16 (Lim sup)**.** For any $\gamma \in \Gamma(\mu, \nu)$ there is a sequence $(\gamma_n)_n$, converging weak* to γ, such that

$$\lim_n E_n(\gamma_n) = E(\gamma).$$

Proof. Let $(L_n)_n$ be a positive sequence. Then

$$E_n(\gamma_n) \leq \int_{X\times X} c \,\mathrm{d}\gamma_n + 2\varepsilon_n \log(n_{L_n}).$$

Choosing now $(L_n)_n$ decreasing, such that $\varepsilon_n \log(n_{L_n}) \to 0$, and with the weak* convergence of $(\gamma_{L_n})_n$ to γ one obtains the result. □

Together, lim-inf and lim-sup inequality provide:

Proposition 3.17. Let $(\gamma_n)_n$ be a sequence of minimizers for E_n (their existence was established earlier). Then the sequence is weak* precompact and any cluster point minimizes E.

Proof. Weak* precompactness is obtained from compactness of $\Gamma(\mu,\nu)$. Any cluster point γ then satisfies

$$\liminf_n E_n(\gamma_n) \geq E(\gamma).$$

(Restriction to subsequences is no issue here, since $E_n(\gamma_n)$ can be seen to be non-increasing, and thus all subsequences have the same limit inferior.) If there were some other $\tilde{\gamma}$, with $E(\tilde{\gamma}) < E(\gamma)$, then a recovery sequence $(\tilde{\gamma}_n)_n$ for $\tilde{\gamma}$, constructed via the block approximation as above, would satisfy

$$E(\tilde{\gamma}) = \lim_n E_n(\tilde{\gamma}_n)$$

and thus for sufficiently big n one would obtain the contradiction

$$E_n(\tilde{\gamma}_n) < E_n(\gamma_n).$$

□

3.3 Duality

Theorem 3.18 (Fenchel–Rockafellar)**.** Let (X,X^*), (Y,Y^*) be two couples of topologically paired spaces. Let $A : X \to Y$ be a bounded linear operator. Let G and F be proper convex functions, defined on X and Y respectively, with values in $(-\infty,\infty]$. If there exists $x \in X$ such that G is finite at x and F is continuous at Ax, then

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

The supremum is attained.

Proposition 3.19 (Duality for entropic OT)**.** A dual problem for the entropic OT problem is given by

$$C_\varepsilon(\mu,\nu) = \sup\left\{\int_X \phi \,\mathrm{d}\mu + \int_X \psi \,\mathrm{d}\nu - \varepsilon \int_{X\times X} \left[\exp\left(\frac{\phi\oplus\psi - c}{\varepsilon}\right) - 1\right] \mathrm{d}\rho \middle| \phi,\psi \in C(X)\right\}.$$

Here $\phi \oplus \psi$ denotes the function $(x,y) \mapsto \phi(x) + \psi(y)$.

Proof. Setting

$$\begin{aligned}
F &: \mathcal{M}(X)^2 \to (-\infty,\infty], & F &= \iota_{\{(\mu,\nu)\}},\\
G &: \mathcal{M}(X\times X) \to (-\infty,\infty], & \gamma &\mapsto \int_{X\times X} c\,\mathrm{d}\gamma + \varepsilon\,\mathrm{KL}(\gamma|\rho),\\
A &: \mathcal{M}(X\times X) \to \mathcal{M}(X)^2, & \gamma &\mapsto (p_1\gamma, p_2\gamma)
\end{aligned}$$

we can write the entropic transport problem as

$$\inf_{\gamma\in\mathcal{M}(X\times X)} F(A\gamma) + G(\gamma).$$

We find that we cannot directly apply the FR duality theorem, since F is nowhere continuous. We will still proceed for now and observe in the end, that we can indeed apply the theorem in the ‘reverse’ direction by flipping the roles of primal and dual problem (and keeping careful track of minus signs).

We find:

$$F^*((\phi,\psi)) = \sup_{(\rho,\sigma)\in\mathcal{M}(X)^2} \int_X \phi\,\mathrm{d}\rho + \int_X \psi\,\mathrm{d}\sigma - \iota_{\{(\mu,\nu)\}}(\rho,\sigma) = \int_X \phi\,\mathrm{d}\mu + \int_X \psi\,\mathrm{d}\nu.$$

For the adjoint of A:

$$\langle(\phi,\psi), A\gamma\rangle = \int_X \phi \mathrm{d}p_1\gamma + \int_X \psi \mathrm{d}p_2\gamma = \int_{X\times X} [\phi(x)+\psi(y)]\mathrm{d}\gamma(x,y)$$

and so $A^*(\phi,\psi) = \phi\oplus\psi$. For G we first observe that $G(\gamma) = \int_{X\times X} c\,\mathrm{d}\gamma + \varepsilon\,\mathrm{KL}(\gamma|\mu\otimes\nu)$. That is, it is obtained from KL first by a positive re-scaling, and then by adding a linear term. Using the simple relations

$$[f(x) = \varepsilon\cdot g(x)] \quad\Rightarrow\quad [f^*(z) = \varepsilon\cdot g^*(z/\varepsilon)]$$

$$[f(x) = \langle a,x\rangle + g(x)] \quad\Rightarrow\quad [f^*(z) = g^*(z-a)]$$

we obtain that $G^*(\xi) = \varepsilon\,\mathrm{KL}^*((\xi-c)/\varepsilon|\rho)$ where KL^* denotes the conjugation with respect to the first argument. One obtains that

$$\begin{aligned}\mathrm{KL}^*(\xi|\rho) &= \sup_{\gamma\in\mathcal{M}(X\times X)} \int \xi\,\mathrm{d}\gamma - \mathrm{KL}(\gamma|\rho)\\ &= \sup_{u\in L^1(\rho)} \int [\xi\cdot u - \varphi(u)]\,\mathrm{d}\rho = \int \varphi^*(\xi)\,\mathrm{d}\rho\end{aligned}$$

and a brief explicit computation yields that $\varphi^*(s) = \exp(s)-1$. Then, formally writing down

$$\sup_{(\phi,\psi)\in C(X)^2} -F^*(-(\phi,\psi)) - G^*(A^*(\phi,\psi))$$

yields the above expression for the dual problem.
It remains to show that we can actually apply the FR theorem. For this we now observe that the functions F^* and G^* are globally finite and continuous, hence the 'reverse constraint qualifications' are satisfied. This implies that FR also provides the existence of optimal entropic OT plans, which we had already established earlier by direct methods. □

Proposition 3.20 (Primal-dual optimality conditions for Fenchel–Rockafellar duality)**.** x and $y*$ are primal and dual optimal in the FR-primal-dual problem pair above if and only if

$$[Ax\in\partial F^*(-y^*) \Leftrightarrow -y^*\in\partial F(Ax)] \wedge [x\in\partial G^*(A^*y^*) \Leftrightarrow A^*y^*\in\partial G(x)].$$

Proof. The Fenchel–Young inequality states that

$$F(x) + F^*(y*) \geq \langle x, y*\rangle$$

with equality if and only if $x\in\partial F^*(y^*)$ or equivalently $y^*\in\partial F(x)$.
Now consider the primal dual gap of the above problem pair:

$$\begin{aligned}0 &\leq [F(Ax)+G(x)] - [-F^*(-y^*) - G^*(A^*y)]\\ &= [F(Ax)+F^*(-y^*)+\langle Ax,-y^*\rangle] + [G(x)+G^*(A^*y)+\langle x, A^*y^*\rangle]\end{aligned}$$

By the Fenchel–Young inequality, this can be zero if and only if both parentheses are zero, which happens if and only if the subdifferential conditions for both apply, which are the stated PD optimality conditions. □

Proposition 3.21 (Application to entropic OT)**.** A pair $\gamma\in\mathcal{M}(X\times X)$, $(\phi,\psi)\in C(X)^2$ are primal-dual optimal if and only if

$$p_1\gamma = \mu, \qquad\qquad p_2\gamma = \nu, \qquad\qquad \gamma = \exp([\phi\oplus\psi - c]/\varepsilon)\cdot\rho.$$

Proof. Consider the condition $A\gamma\in\partial F^*(-(\phi,\psi))$. Since F^* is the linear pairing $(\alpha,\beta)\mapsto\langle\alpha,\mu\rangle+\langle\beta,\nu\rangle$, it is subdifferential is the singleton (μ,ν) at all points. With A being the marginal projection operator, this translates to the two marginal constraints for γ.
The function $G^*(\xi) = \varepsilon\int[\exp([\xi-c]/\varepsilon)-1]\,\mathrm{d}\rho$ is differentiable with

$$\frac{\mathrm{d}}{\mathrm{d}t}G^*(\xi+t\cdot\eta)|_{t=0} = \int \exp([\xi-c]/\varepsilon)\eta\mathrm{d}\rho.$$

The subdifferential is therefore given by $\partial G^*(\xi) = \exp([\xi-c]/\varepsilon)\rho$. Inserting now the argument $\xi = \phi\oplus\psi$ yields the expression for γ. □

3.4 Sinkhorn algorithm

Remark 3.22 (Choice of reference measure and existence of optimal dual solutions)**.** For this section we fix the choice $\rho = \mu \otimes \nu$. This will have some useful consequences, in particular existence of continuous (and even more regular of c is 'nice') optimal dual solutions.

The Sinkhorn algorithm can be interpreted as alternating block optimization on the dual problem, alternatingly fixing one of the functions ϕ or ψ and optimizing over the other.

Lemma 3.23. For fixed $\psi \in C(X)$, an optimal ϕ in the dual entropic OT problem is given by

$$\phi(x) = -\varepsilon \cdot \log \left(\int_X \exp([\psi(y) - c(x, y)]/\varepsilon)\, \mathrm{d}\nu(y) \right).$$

Of course, the corresponding Lemma for the second marginal also holds.

Proof. The dual entropic OT objective is concave and differentiable. Hence, its maximizers can be determined by studying the first order optimality conditions. Let

$$J(\phi, \psi) := \int \phi \, \mathrm{d}\mu + \int \psi \mathrm{d}\nu - \varepsilon \int_{X \times X} [\exp([\phi \oplus \psi - c]/\varepsilon) - 1] \, \mathrm{d}\mu \otimes \nu.$$

One finds that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} J(\phi + t \cdot \eta, \psi)|_{t=0} &= \int_X \eta \, \mathrm{d}\mu - \int_{X \times X} \exp([\phi(x) + \psi(y) - c(x, y)]/\varepsilon) \eta(x) \mathrm{d}\mu(x) \mathrm{d}\nu(y) \\ &= \int_X \left[1 - \exp(\phi(x)/\varepsilon) \int_X \exp([\psi(y) - c(x, y)]/\varepsilon) \mathrm{d}\nu(y) \right] \eta(x) \, \mathrm{d}\mu(x) \end{aligned}$$

For this to be zero for all $\eta \in C(X)$, we need that the bracket is zero μ-almost everywhere. Resolving this expression for ϕ yields the given expression. □

Definition 3.24 (Sinkhorn algorithm)**.** For some initial $\psi^{(0)} \in C(X)$, set recursively for $\ell \in \{0, 1, \ldots\}$,

$$\begin{aligned} \phi^{(\ell+1)}(x) &= -\varepsilon \cdot \log \left(\int_X \exp([\psi^{(\ell)}(y) - c(x, y)]/\varepsilon) \, \mathrm{d}\nu(y) \right), \\ \psi^{(\ell+1)}(y) &= -\varepsilon \cdot \log \left(\int_X \exp([\phi^{(\ell+1)}(x) - c(x, y)]/\varepsilon) \, \mathrm{d}\mu(x) \right). \end{aligned}$$

We refer to this procedure as the Sinkhorn algorithm.

Remark 3.25 (Primal interpretation of dual optimality condition)**.** Recall the primal-dual optimality condition $\gamma = \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \mu \otimes \nu$ and the marginal constraint $p_1 \gamma = \mu$. Together these imply for all $\eta \in C(X)$,

$$\begin{aligned} \int_X \eta(x) \, \mathrm{d}\mu(x) &= \int_{X \times X} \eta(x) \, \mathrm{d}\gamma(x, y) \\ &= \int_X \left[\exp(\phi(x)/\varepsilon) \int_X \exp([\psi(y) - c(x, y)]/\varepsilon) \mathrm{d}\nu(y) \right] \eta(x) \, \mathrm{d}\mu(x). \end{aligned}$$

This is precisely the dual optimality condition for ϕ obtained in the above lemma. Hence, this dual optimality condition has the primal interpretation that ϕ is chosen just right, such that the implied primal γ has the prescribed first marginal μ. Alternating maximization therefore also has the interpretation of alternating re-scaling. Consequently, the Sinkhorn algorithm is also known as iterative proportional fitting procedure.
Finally, these re-scalings can also be interpreted as KL projections of the 'old' γ onto one of the two marginal constraints, and thus the algorithm can also be interpreted as alternating projection method.

Remark 3.26 (Entropic c-transform)**.** The map $\psi \mapsto \psi^{c,\nu,\varepsilon} := \phi$ where ϕ is given by the locally optimal ϕ,

$$\phi(x) = -\varepsilon \cdot \log\left(\int_X \exp([\psi(y) - c(x,y)]/\varepsilon)\,\mathrm{d}\nu(y)\right)$$

is sometimes referred to as entropic c-transform.
The 'classical' c-transform is given by

$$\psi^c(x) = \inf_{y\in X} c(x,y) - \psi(y)$$

and it plays an important role in the analysis of unregularized optimal transport and some numerical methods, such as the auction algorithm.
Applying the classical c-transform in unregularized optimal transport will however always become stationary after at most three steps ($\phi^{ccc} = \phi^c$) and need not be converge to dual maximizers. This is related to the fact that the unregularized dual problem is constrained and therefore non-smooth. Alternating maximization may therefore get stuck in the 'ridge' corresponding to the constraint.

Similar to the classical c-transform, $\psi^{c,\nu,\varepsilon}$ inherits some regularity from c.

Lemma 3.27. Assume that c has a modulus of continuity ω in its first argument, i.e. $c(x+\delta,y) \leq c(x,y) + \omega(|\delta|)$ (here $\omega : [0,\infty) \to [0,\infty)$ is continuous and $\omega(0) = 0$), then $\psi^{c,\nu,\varepsilon}$ has the same modulus of continuity.

Proof.

$$\begin{aligned}
\psi^{c,\nu,\varepsilon}(x+\delta) &= -\varepsilon \cdot \log\left(\int_X \exp([\psi(y) - c(x+\delta,y)]/\varepsilon)\,\mathrm{d}\nu(y)\right) \\
&\leq -\varepsilon \cdot \log\left(\int_X \exp([\psi(y) - c(x,y) - \omega(|\delta|)]/\varepsilon)\,\mathrm{d}\nu(y)\right) \\
&\leq \psi^{c,\nu,\varepsilon}(x) + \omega(|\delta|).
\end{aligned}$$

□

Remark 3.28. Indeed, one can show even higher-order Sobolev type regularity of entropic c-transforms for the squared distance cost, see for instance [Genevay *et al.*, 2019]. This is a crucial step for statistical stability of empirical (entropic) optimal transport. This higher-order regularity deteriorates as $\varepsilon \to 0$ and in the limit one obtains the 'cursed' convergence rates of classical optimal transport.

This can be used for a simple convergence proof of the Sinkhorn algorithm, based on compactness.

Proposition 3.29. The Sinkhorn algorithm converges (up to subsequences and optimal constant shifts) to a solution of the dual entropic OT problem. In particular, optimal dual solutions exist.

Proof. Since $c \in C(X \times X)$, there exists a modulus of continuity for both arguments. Hence, the Sinkhorn iterates all have the same modulus of continuity and they are therefore equi-continuous. Adding a constant shift to each $\phi^{(\ell+1)}$ such that $\phi^{(\ell+1)}(x_0) = 0$ for some arbitrary fixed $x_0 \in X$ will merely result in a corresponding shift in the opposite direction in $\psi^{(\ell+1)}$. In particular the sequence $(\phi^{(\ell)})_\ell$ will also be equibounded, as will be the sequence $(\psi^{(\ell)})_\ell$.
So by the Arzela–Ascoli theorem, there exists a pair of cluster points (ϕ, ψ) such that a suitable subsequence of iterates (with some constant shifts) converges uniformly to these two functions. Since the entropic c-transform is continuous, this means that $\phi = \psi^{c,\nu,\varepsilon}$ and $\psi = \phi^{c^\top,\mu,\varepsilon}$ (here $c^\top$ denotes the 'flipped' cost function with first and second argument flipped).
This implies that $\gamma := \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \mu \otimes \nu$ satisfies both marginal constraints, and thus the tripled $(\gamma, (\phi, \psi))$ is primal and dual optimal for the entropic OT problem. □

Remark 3.30.

- The convergence (and optimality proof) above does not extend to the case $\varepsilon = 0$, since being a fixed point of the c-transforms is not sufficient for optimality in the unregularized problem.

- The choice $\rho = \mu \otimes \nu$ was important in this section, as only by this choice does the expression for $\psi^{c,\nu,\varepsilon}$ obtain spatial regularity independent of μ. This yields the compactness of the iterates and existence of a continuous dual solutions.

Remark 3.31 (Matrix-scaling formulation)**.** Introduce the functions

$$u := \exp(\phi/\varepsilon), \qquad v := \exp(\psi/\varepsilon),$$

(and likewise applied to all the iterates of the Sinkhorn algorithm). Then the iterations can be written as

$$u^{(\ell+1)}(x) = 1/\int \exp(-c(x,y)/\varepsilon) v^{(\ell)}(y)\,\mathrm{d}\nu(y),$$
$$v^{(\ell+1)}(y) = 1/\int \exp(-c(x,y)/\varepsilon) u^{(\ell+1)}(x)\,\mathrm{d}\mu(x).$$

This can easily be expressed as matrix-vector multiplications in the discrete setting. This might be a tad faster than the logsumexp-version, but is also numerically more prone to issues, especially for small ε. There are many tricks for running the Sinkhorn algorithm stable and efficiently at small ε.

Remark 3.32 (Speed of convergence)**.**

- [Franklin and Lorenz, 1989]: linear convergence of dual iterates to maximizer in Hilbert's projective metric. But: contraction ratio approaches 1 like $1 - \exp(-\|c\|_\infty/\varepsilon)$ as $\varepsilon \to 0$.
- [Schmitzer, 2019]: convergence of an asymmetric ('auction-like') Sinkhorn algorithm in $O(1/\varepsilon)$ iterations (measured in L^1-error of primal iterate marginal constraints)
- [Berman, 2020]: convergence of the Sinkhorn algorithm for the W_2 distance on the Torus in $O(1/\varepsilon)$ iterations, by showing that the iterates asymptotically follow a non-linear PDE
- ε-scaling very efficient in practice (at least on 'normal problems') but no proof for its efficiency yet (as far as I am aware).
- There are several variants of Sinkhorn, intended to be faster, such as the 'Greenkhorn' algorithm.

Remark 3.33 (Flexibility of the Sinkhorn algorithm)**.** One of the biggest strengths of the Sinkhorn algorithm is that it can easily be adapted to related problems, such as optimal transport barycenters, multi-marginal transport problems (only efficient, if there is some trick to handle the high problem dimensionality), and unbalanced transport problems. See for instance: [Benamou *et al.*, 2015; Peyré, 2015; Chizat *et al.*, 2018; Benamou *et al.*, 2019b]

4 Multi-marginal transport

4.1 Introduction

Preface.

- Here use regularity setting of previous section.
- We have seen above (in the gluing lemma) that it is perfectly natural to consider 'transport plans' or 'couplings' between more than two marginals.

Definition 4.1 (Multi-marginal optimal transport problem)**.** For some $N \in \mathbb{N}$, let $\vec{\mu} = (\mu_1, \ldots, \mu_N) \in \mathcal{P}(X)^N$ be a collection of N probability measures. The set of corresponding multi-marginal couplings is given by

$$\Pi(\vec{\mu}) := \left\{\pi \in \mathcal{P}(X^N) \,\middle|\, p_{i\#}\pi = \mu_i \text{ for } i = 1, \ldots, N\right\}.$$

For a multi-marginal cost function $c \in C(X^N)$ the multi-marginal transport problem is given by

$$C(\vec{\mu}) := \int \left\{ \int_{X^N} c(\vec{x}) \, \mathrm{d}\pi(\vec{x}) \middle| \pi \in \Pi(\vec{\mu}) \right\}$$

Proposition 4.2. For the above regularity setting (X compact, c continuous), minimizers for the multi-marginal transport problem exist.

Remark 4.3 (Motivation)**.**
- 'matching for teams' Carlier and Ekeland [2010]: each μ_i could represent agents of a different type (e.g. electricians, carpenters, painters, . . . ; but also: a potential construction site), and one of each is required to finish a project (e.g. a house). $c(\vec{x})$ denotes the cost of the agents x_1 to x_N being paired to build one house (e.g. factoring in all travel distances and potentially pairwise animosities between the agents that may not want to work together). Then the above problem finds the most efficient grouping (generalization of pairing) of agents to finish all projects.
- we will study below: Wasserstein barycenter problem
- later this week: interaction between particles that move along paths.

Proposition 4.4 (Duality)**.** One has

$$C(\vec{\mu}) = \sup \left\{ \sum_i \int_X \phi_i \, \mathrm{d}\mu_i \middle| \vec{\phi} = (\phi_1, \ldots, \phi_N) \in C(X)^N, \bigoplus_i \phi_i \leq c \right\}$$

where $(\bigoplus_i \phi_i)(\vec{x}) = \sum_i \phi_i(x_i)$.

Proof.
- We re-use a lot from standard Kantorovich duality, but here use Fenchel–Rockafellar duality as in entropic transport, to obtain the dual problem.
- Write the dual problem as

$$\sup_{\vec{\phi}} -F(A\vec{\phi}) - G(-\vec{\phi})$$

for

$$\begin{aligned}
&G : C(X)^N \to \mathbb{R} \cup \{\infty\}, && G(\vec{\phi}) = \sum_i \int_X \phi_i \, \mathrm{d}\mu_i, \\
&A : C(X)^N \to C(X^N), && A\vec{\phi} = \bigoplus_i \phi_i, \\
&F : C(X^N) \to \mathbb{R} \cup \{\infty\}, && F(\xi) = \begin{cases} 0 & \text{if } \xi \leq c, \\ +\infty & \text{else.} \end{cases}
\end{aligned}$$

- The formal dual problem is then given by

$$\inf_{\pi \in \mathcal{M}(X^N)} F^*(\pi) + G^*(A^*\pi)$$

where we find

$$\begin{aligned}
&G^* : \mathcal{M}(X)^N \to \mathbb{R} \cup \{\infty\}, && G^*(\vec{\nu}) = \begin{cases} 0 & \text{if } \nu_i = \mu_i \text{ for all } i, \\ +\infty & \text{else,} \end{cases} \\
&A^* : \mathcal{M}(X^N)^N \to \mathcal{M}(X)^N, && A^*\pi = (p_{1\#}\pi, \dots, p_{N\#}\pi), \\
&F^* : \mathcal{M}(X^N) \to \mathbb{R} \cup \{\infty\}, && F^*(\pi) = \begin{cases} \int_{X^N} c \, \mathrm{d}\pi & \text{if } \pi \geq 0, \\ +\infty & \text{else.} \end{cases}
\end{aligned}$$

- Since c is bounded from below and G is linear, easy to find $\vec{\phi} \in C(X)^N$ such that G finite at $\vec{\phi}$ and F continuous at $A\vec{\phi}$. So strong duality holds. □

4.2 Wasserstein barycenter

Motivation.

- Similar to Euclidean space or Riemannian manifolds one can wonder what the weighted center of mass of a tuple of probability measures $\vec{\mu} = (\mu_1, \dots, \mu_N)$ in $\mathcal{P}(X)^N$ with weights $\vec{\lambda} = (\lambda_1, \dots, \lambda_N)$ ($\lambda_i > 0$, $\sum_{i=1}^N \lambda_i = 1$) with respect to the squared Wasserstein distance is.
- For simplicity, in this subsection we assume X is a compact, convex subset of $\mathbb{R}^d$.
- The seminal paper on this is Agueh and Carlier [2011]. It is discussed in the textbooks Santambrogio [2015] and Peyré and Cuturi [2019]. I collected a few basic computations for convenience in Friesecke *et al.* [2021], from which mos of the following is taken. There is a ton of exciting literature on algorithms and applications.

Definition 4.5 (Coupled-two-marginal formulation for Wasserstein barycenter)**.** For $\vec{\mu} \in \mathcal{P}(X)^N$ set

$$W_{\mathrm{C2M}}(\vec{\mu})^2 := \inf \left\{ \sum_{i=1}^N \lambda_i \cdot W(\mu_i, \nu)^2 \middle| \nu \in \mathcal{M}_+(X) \right\}. \tag{1}$$

This is a nested optimization problem where one needs to minimize over $\nu \in \mathcal{M}_+(X)$ and over each $\pi \in \mathcal{M}_+(X^2)$ within the $W(\mu_i, \nu)^2$ terms. Hence, we refer to this as the coupled-two-marginal formulation, as opposed to the multi-marginal formulation introduced below. Since $W(\mu_i, \nu)^2 = +\infty$ when $\|\mu_i\| \neq \|\nu\|$ (as the feasible set in the Kantorovich formulation for $W(\mu_i, \nu)^2$ is empty), we need not add the constraint $\nu \in \mathcal{P}(X)$, as it is enforced automatically.

Proposition 4.6. Minimizers ν of (1) exist. A minimizer is called Wasserstein barycenter of $\vec{\mu}$ with weights $\vec{\lambda}$.

A proof can be found in Agueh and Carlier [2011] or follows from standard arguments about weak∗ compactness of bounded measures and weak∗ continuity of the Wasserstein distance on compact metric spaces.
Complementarily, the Wasserstein barycenter problem can also be formulated as a multi-marginal transport problem on X^N with a suitable cost function.

Definition 4.7 (Multi-marginal formulation for Wasserstein barycenter)**.**

$$c_{\mathrm{W,MM}}(\vec{x}) := \inf_{y \in X} \sum_{i=1}^N \lambda_i \, |x_i - y|^2 = \sum_{i=1}^N \lambda_i \, |x_i - T(\vec{x})|^2 = \sum_{i,j=1}^N \frac{\lambda_i \, \lambda_j}{2} |x_i - x_j|^2 \tag{2}$$

$$\text{where} \qquad T(\vec{x}) := \sum_{i=1}^N \lambda_i \, x_i \tag{3}$$

takes the points $\vec{x}$ to the unique minimizer $y = T(\vec{x})$ in the first line.

$$W_{\text{MM}}(\vec{\mu})^2 := \inf\left\{ \int_{X^N} c_{\text{W,MM}}\,\text{d}\pi \,\middle|\, \pi \in \mathcal{M}_+(X^N),\, p_{i\sharp}\pi = \mu_i \right\} \tag{4}$$

Proposition 4.8 (Agueh and Carlier [2011])**.** $W_{\text{C2M}}(\vec{\mu})^2 = W_{\text{MM}}(\vec{\mu})^2$. ν is a minimizer of (1) if and only if there exists a minimizer π of (4) such that $T_\sharp \pi = \nu$ (with T given by (3)). Consequently, for minimizers π of (4) we will also call $T_\sharp\pi$ a barycenter.

Proof.

- Let π be a minimizer of (4), set $\nu := T_\sharp \pi$ and $\pi_i := (p_i, T)_\sharp \pi$ for $i = 1, \ldots, N$, where $(p_i, T) : X^N \to X^2$, $\vec{x} \mapsto (x_i, T(\vec{x}))$.

- One finds that $p_{1\sharp}\pi_i = (p_1 \circ (p_i, T))_\sharp \pi = p_{i\sharp}\pi = \mu_i$ and similarly $p_{2\sharp}\pi_i = \nu$, so that $\pi_i \in \Pi(\mu_i, \nu)$. Therefore, one finds that

$$\begin{aligned}
W_{\text{MM}}(\vec{\mu})^2 &= \int_{X^N} c_{\text{W,MM}}\,\text{d}\pi = \int_{X^N} \left[\sum_{i=1}^N \lambda_i |x_i - T(\vec{x})|^2\right] \text{d}\pi(\vec{x}) \\
&= \sum_{i=1}^N \lambda_i \int_{X^N} |p_i(\vec{x}) - T(\vec{x})|^2\,\text{d}\pi(\vec{x}) \\
&= \sum_{i=1}^N \lambda_i \int_{X^2} |x - y|^2\,\text{d}[(p_i, T)_\sharp\pi](x, y) \\
&= \sum_{i=1}^N \lambda_i \int_{X^2} |x - y|^2\,\text{d}\pi_i(x, y) \geq W_{\text{C2M}}(\vec{\mu})^2\,.
\end{aligned} \tag{5}$$

- Conversely, let now $\hat{\nu}$ be a minimizer of (1) and let $\hat{\pi}_i \in \Pi(\mu_i, \hat{\nu})$ be an optimal plan for $W(\mu_i, \hat{\nu})^2$ for $i = 1, \ldots, N$. Further, let $(\hat{\pi}_i^y)_{y\in X}$ be the disintegration of $\hat{\pi}_i$ w.r.t. its second marginal. Introduce now the measure $\hat{\pi} \in \mathcal{M}_+(X^N)$ via

$$\int_{X^N} \phi\,\text{d}\hat{\pi} := \int_{X^{N+1}} \phi(\vec{x})\,\text{d}\hat{\pi}_1^y(x_1)\ldots\text{d}\hat{\pi}_N^y(x_N)\,\text{d}\hat{\nu}(y) \tag{6}$$

for test functions $\phi \in C(X^N)$.

- One then finds for $\phi \in C(X)$ that

$$\int_X \phi \circ p_i\,\text{d}\hat{\pi} = \int_{X^2} \phi(x_i)\,\text{d}\hat{\pi}_i^y(x_i)\,\text{d}\hat{\nu}(y) = \int_{X^2} \phi \circ p_1\,\text{d}\hat{\pi}_i = \int_X \phi\,\text{d}\mu_i$$

and therefore that $p_{i\sharp}\hat{\pi} = \mu_i$.

- Consequently,

$$\begin{aligned}
W_{\text{MM}}(\vec{\mu})^2 &\leq \int_{X^N} c_{\text{W,MM}}\,\text{d}\hat{\pi} = \int_{X^{N+1}} \left(\inf_{z\in X} \sum_{i=1}^N \lambda_i |x_i - z|^2\right) \text{d}\hat{\pi}_1^y(x_1)\ldots\text{d}\hat{\pi}_N^y(x_N)\,\text{d}\hat{\nu}(y) \\
&\leq \int_{X^{N+1}} \left(\sum_{i=1}^N \lambda_i |x_i - y|^2\right) \text{d}\hat{\pi}_1^y(x_1)\ldots\text{d}\hat{\pi}_N^y(x_N)\,\text{d}\hat{\nu}(y) \\
&= \sum_{i=1}^N \lambda_i \int_{X^2} |x_i - y|^2 \text{d}\hat{\pi}_i^y(x_i)\,\text{d}\hat{\nu}(y) = \sum_{i=1}^N \lambda_i \int_{X^2} |x_i - y|^2 \text{d}\hat{\pi}_i(x_i, y) \\
&= \sum_{i=1}^N \lambda_i\, W(\mu_i, \hat{\nu})^2 = W_{\text{C2M}}(\vec{\mu})^2.
\end{aligned} \tag{7}$$

- Combining (5) and (7) one finds that $W_{\text{MM}}(\vec{\mu})^2 = W_{\text{C2M}}(\vec{\mu})^2$ and that ν constructed from π is optimal for $W_{\text{C2M}}(\vec{\mu})^2$ and $\hat{\pi}$ constructed from $\hat{\nu}$ is optimal for $W_{\text{MM}}(\vec{\mu})^2$.

- In addition, by equality of $W_{\mathrm{MM}}(\vec{\mu})^2$ and $W_{\mathrm{C2M}}(\vec{\mu})^2$ the second inequality in (7) must be an equality and thus one must have that y is a minimizer of $z \mapsto \sum_{i=1}^N \lambda_i |x_i - z|^2$ $\mathrm{d}\hat{\pi}_1^y(x_1) \dots \mathrm{d}\hat{\pi}_N^y(x_N)\,\mathrm{d}\hat{\nu}(y)$-almost everywhere, i.e. $y = T(\vec{x})$ almost surely. Therefore, one finds

$$\int_{X^N} \phi \circ T \,\mathrm{d}\hat{\pi} = \int_X \phi \,\mathrm{d}\hat{\nu}$$

for $\phi \in C(X)$ and thus $T_\sharp \hat{\pi} = \hat{\nu}$. □

Remark 4.9.
- From a numerical perspective, the naive dimensionality of multi-marginal problems increases exponentially with N, as one must handle measures (or functions) on the space X^N.

- Above we have seen that the multi-marginal problem for the Wasserstein barycenter has an equivalent formulation as a 'pairwise coupled' transport problem. This is possible, since in the multi-marginal transport cost $c_{\mathrm{W,MM}}$ the x_i do not 'interact directly', but merely via the 'proxy' y. More generally, multi-marginal problems with 'tree-structured' cost functions can be reduced to coupled pairwise problems, i.e. they are 'simpler' than general multi-marginal problems.

- It is easy to derive Sinkhorn-type algorithms for multi-marginal transport problems and/or their pairwise couplings via entropic regularization. See for instance Benamou *et al.* [2015, 2019b]; Haasler *et al.* [2021]; Beier *et al.* [2021] (and probably many more).

4.3 Outlook: measures on paths, superposition principle

Definition of path measures.

- Let $\mathcal{X} := H^1([0,1], X)$ be the Sobolev space of paths $[0,1] \to X$ with square-integrable weak derivative. By the Sobolev embedding theorem, curves in $\mathcal{X}$ are continuous.

- A measure $\lambda \in \mathcal{P}(\mathcal{X})$ can be interpreted as a collection of traveling particles. For instance, $\lambda := \delta_\gamma$ for some $\gamma \in \mathcal{X}$ represents a single particle, moving on the path γ.

- For $t \in [0,1]$, the map

$$\mathrm{ev}_t : \mathcal{X} \to X, \qquad\qquad \gamma \mapsto \gamma(t)$$

is called *evaluation map*. For a distribution of paths, $\lambda \in \mathcal{P}(\mathcal{X})$, $\mathrm{ev}_{t\#}\lambda \in \mathcal{P}(X)$ gives the distribution of positions at time t.

Benamou–Brenier energy of a measure path and of measure of paths.

- Let $\rho : [0,1] \to \mathcal{P}(X)$ be a path of measures. Recall the Benamou–Brenier energy:

$$BB(\rho, v) := \int_0^1 \int_X \|v(t,\cdot)\|^2 \,\mathrm{d}\rho(t)\,\mathrm{d}t.$$

Introduce now the path energy:

$$E(\rho) := \inf \{BB(\rho, v) | v \text{ such that } (\rho, v) \text{ solve CE between } \rho(0) \text{ and } \rho(1)\}$$

- Likewise, introduce energy for measure of paths $\lambda \in \mathcal{P}(\mathcal{X})$:

$$\mathcal{E}(\lambda) := \int_{\mathcal{X}} c(\gamma)\,\mathrm{d}\lambda(\gamma)$$

where $c(\gamma) = \|\gamma\|_{H^1}^2$.

Theorem 4.10 (Superposition principle Lisini [2007])**.**

(1) For $\lambda \in \mathcal{P}(\mathcal{X})$ with $\mathcal{E}(\lambda) < \infty$, set $\rho : [0,1] \ni t \mapsto \mathrm{ev}_{t\#}\lambda$. Then $E(\rho) \leq \mathcal{E}(\lambda)$.

(2) For all $\rho : [0,1] \mapsto \mathcal{P}(X)$ with $E(\rho) < \infty$, there exists some $\lambda \in \mathcal{P}(\mathcal{X})$ such that $\rho(t) = \mathrm{ev}_{t\#}\lambda$ and $E(\rho) = \mathcal{E}(\lambda)$.

Remark 4.11. So curves of measures with sufficient regularity (= finite BB energy, which corresponds to being absolutely continuous in W_2 in a metric sense), can be disintegrated into curves of individual particles (measure on curves) that yield the same energy. In more detail, the theorem also gives the expected statements on the Lagrangian and Eulerian velocity fields.

Remark 4.12 (BB energy of a path as inf-dim multi-marginal problem)**.**

- By the above, for given $\rho : [0,1] \to \mathcal{P}(X)$ we can compute $E(\rho)$ as

$$E(\rho) = \inf \{BB(\rho, v) | v \text{ such that } (\rho, v) \text{ solve CE between } \rho(0) \text{ and } \rho(1)\}$$

  or

$$E(\rho) = \inf \left\{ \int_{\mathcal{X}} c(\gamma) \, \mathrm{d}\lambda(\gamma) \middle| \lambda \in \mathcal{P}(\mathcal{X}) \text{ such that } \mathrm{ev}_{t\#}\lambda = \rho(t) \text{ for all } t \in [0,1] \right\}$$

- Formally the latter can be interpreted as infinite-dimensional multi-marginal problem, with a tree-structured cost (in fact: chain-structured), and the former is the corresponding decomposition into pairwise problems.

- For some large $N \in \mathbb{N}$, purely formally, we can approximate the former as

$$E_{N,1}(\rho) := \sum_{i=0}^{N-1} W_2^2 \left(\rho(\tfrac{i}{N}), \rho(\tfrac{i+1}{N})\right),$$

  and the latter as

$$E_{N,2}(\rho) := \inf \left\{ \int_{X^N} c(\vec{x}) \, \mathrm{d}\lambda(\vec{x}) \middle| \lambda \in \mathcal{P}(X^N) \text{ such that } p_{i\#}\lambda = \rho(i/N) \text{ for all } i \in \{0, \ldots, N\} \right\}.$$

- This is a good starting point for rich modelling possibilities, e.g. for congestion constraints, and a neat bridge to connect optimal transport with optimal control and mean field games. See, for instance, Carlier *et al.* [2008]; Benamou *et al.* [2019a]; Sarrazin [2020] (some arbitrary references I could list from the top of my head; not exhaustive!).

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems.* Oxford mathematical monographs. Oxford University Press, 2000.

F. Beier, J. von Lindheim, S. Neumayer, and G. Steidl. Unbalanced multi-marginal optimal transport. arXiv:2103.10854, 2021.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Imaging Sci.*, 37(2):A1111–A1138, 2015.

Jean-David Benamou, Guillaume Carlier, Simone Di Marino, and Luca Nenna. An entropy minimization approach to second-order variational mean-field games. *Mathematical Models and Methods in Applied Sciences*, pages 1–31, 2019.

Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numerische Mathematik*, 142(1):33–54, 2019.

Robert J. Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations. *Numerische Mathematik*, 145:771–836, 2020.

G. Carlier and I. Ekeland. Matching for teams. *Econ Theory*, 42(2):397–418, 2010.

G. Carlier, C. Jimenez, and F. Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM J. Control Optim*, 47(3):1330–1350, 2008.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, 87:2563–2609, 2018.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. arXiv:1810.08278, 2018.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

Gero Friesecke, Daniel Matthes, and Bernhard Schmitzer. Barycenters for the Hellinger–Kantorovich distance over Rd. *SIAM J. Math. Anal.*, 53(1):62–110, 2021.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583, 2019.

Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Multimarginal optimal transport with a tree-structured cost and the schrödinger bridge problem. *SIAM Journal on Control and Optimization*, 59(4):2428–2453, 2021.

Stefano Lisini. Characterization of absolutely continuous curves in Wasserstein spaces. *Calc. Var. Partial Differential Equations*, 28(1):85–120, 2007.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM J. Imaging Sci.*, 8(4):2323–2351, 2015.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser Boston, 2015.

Clément Sarrazin. Lagrangian discretization of variational mean field games. arXiv:2010.11519, 2020.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.*, 41(3):A1443–A1481, 2019.