Overview on optimal transport

Bernhard Schmitzer

Chamonix / Les Houches, March 2023

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 $X = \{x_1, \dots, x_N\}$, 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 belonging to vector: $\sum_i \mu_i \delta_{x_i}$
- 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.
- $c : X \times X \to \mathbb{R}$ becomes matrix in $\mathbb{R}^{N \times N}$,

$$\int_{X \times X} c \, \mathrm{d}\pi = \sum_{i,j} c_{i,j} \pi_{i,j} =: \langle c, \pi \rangle$$

- 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:

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

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

- 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 be sufficient) 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 small 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. These 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 $\pi \geq 0$, one has $W_p(\mu, \nu) \geq 0$.
- Symmetry, $W_p(\mu, \nu) = W_p(\nu, \mu)$: let $\pi \in \Pi(\mu, \nu)$, set $\hat{\pi} := (p_2, p_1)_{\#}\pi$ and find $\hat{\pi} \in \Pi(\nu, \mu)$. Further:

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

  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:

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

- 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). Then:

$$0 = \int_{X \times X} d(x, y)^2 \, \mathrm{d}\pi(x, y)$$

  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 show how two plans $\pi \in \Pi(\mu,\nu)$, $\lambda \in \Pi(\nu,\rho)$ can be concatenated.
- 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. 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
  - evaluating Wasserstein cost of η 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 $\gamma : [0,1] \to X$ is called (constant speed) geodesic from x to y if

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

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, $\pi = (\mathrm{id}, T)_{\#}\mu$.
- 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

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

  Taking each particle from its initial position to $\gamma(t,\cdot)$ means we apply the push-forward of $\gamma(t,\cdot)$ to μ. So our conjectured geodesic takes the form

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

  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 $\gamma(t,x)$, so use this as transport map candidate: $T_{0,s} := \gamma(t,\cdot)$. 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 $\gamma(s,x)$ moves to $\gamma(t,x)$. For $s < 1$ can show: $\gamma(s,\cdot)$ is invertible. So use as candidate: $T_{s,t} := \gamma(t,\cdot) \circ \gamma(s,\cdot)^{-1}$. Find:

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

- 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 (Lagrangian) velocity $\dot\gamma(t,x)$. But at time t 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

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

  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

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

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

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

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

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

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

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

This means, we recover the Lagrangian coordinate from the Eulerian velocity field. (Mathematically this only works if v if sufficiently regular.) 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 regularization and Sinkhorn algorithm

3.1 Entropic regularization

KL-divergence / relative entropy.

- For $\rho, \sigma \in \mathcal{P}(X)$ set

$$\mathrm{KL}(\rho|\sigma) := \begin{cases} \int \varphi(\frac{\mathrm{d}\rho}{\mathrm{d}\sigma})\mathrm{d}\sigma & \text{if } \rho \ll \sigma, \\ +\infty & \text{else.} \end{cases},$$

  where

$$\varphi(s) := \begin{cases} s\log(s) - s + 1 & \text{if } s > 0, \\ 1 & \text{if } s = 0, \\ +\infty & \text{else.} \end{cases}$$

  and similarly on other spaces.

- One has KL is jointly convex in both of its arguments, $\mathrm{KL}(\cdot|\sigma)$ is strictly convex, $\mathrm{KL}(\rho|\sigma) \geq 0$ with equality if and only if $\rho = \sigma$.

Regularized primal problem.

- Now, for some (small) parameter $\varepsilon \geq 0$, add $\varepsilon \cdot \mathrm{KL}(\pi|\mu\otimes\nu)$ to primal Kantorovich objective. The regularized problem is given by:

$$\inf\left\{\int c\,\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi|\mu\otimes\nu)\middle|\pi \in \Pi(\mu,\nu)\right\}$$

  In principle other choices instead of $\mu\otimes\nu$ are possible for the reference measure, but they can all be reduced to this case.

- Minimizers exist under reasonable assumptions. The minimizer is unique for $\varepsilon > 0$. In reasonable settings, one has $\pi_\varepsilon \to \pi$ as $\varepsilon \to 0$ where π is (some) minimizer of the unregularized problem.

Motivation. Considering the regularized problem can be useful for various reasons:

- Relation to modelling stochastic processes: optimal π becomes 'blurry' and non-deterministic, even in cases where optimal unregularized solution would be given by a Monge map.

- Statistical reasons: if true μ an ν are unknown, but only finite samples are available, then can show that optimal value of entropic problem converges faster to true value, compared to the unregularized problem; in particular if X is high-dimensional.

- Uniqueness of optimal π and differentiability of optimal value: when OT is used in a bigger data processing pipeline, it is reassuring to know that optimal transport plan is unique and that gradient-based optimization can be used.

- Simple numerical method: for $\varepsilon > 0$ the problem can be solved with Sinkhorn's algorithm, which is particularly simple and also apt for parallelization on GPUs.

Derivation of dual problem.

- for simplicity, we will do this again to the discrete setting

- as before, argue via Lagrangian: primal problem is equivalent to

$$\inf_{\pi\in\mathbb{R}_+^{N\times N}} \sup_{\phi,\psi\in\mathbb{R}^N} \langle c,\pi\rangle + \varepsilon\ \mathrm{KL}(\pi|\mu\otimes\nu) + \langle\phi, \mu - P_1\pi\rangle + \langle\psi, \nu - P_2\pi\rangle$$

- swap order of minimization, reorder terms

$$\sup_{\phi,\psi\in\mathbb{R}^N} \langle\phi,\mu\rangle + \langle\psi,\nu\rangle + \inf_{\pi\in\mathbb{R}_+^{N\times N}} \left[\left\langle c - P_1^\top\phi - P_2^\top\psi, \pi\right\rangle + \varepsilon\ \mathrm{KL}(\pi|\mu\otimes\nu)\right]$$

- since KL is acting 'entry-wise' on π, the min can be performed for each entry of π separately. let us solve the following sub-problem:

$$\inf_{s\geq 0} \hat{c}\cdot s + \varepsilon\,\varphi(s/t)\cdot t$$

  where t takes the role of $\mu_i\cdot\nu_j$.

- try first order optimality condition:

$$0 = \partial_s[\hat{c}\cdot s + \varepsilon\,\varphi(s/t)\cdot t] = \hat{c} + \varepsilon\log(s/t) \quad \Rightarrow \quad s = \exp(-\hat{c}/\varepsilon)\cdot t > 0$$

  this value lies in the region where φ' is defined. by strict convexity of φ this must be the unique minimizer. we get:

$$\inf_{s\geq 0} \hat{c}\cdot s + \varepsilon \underbrace{\varphi(s/t)\cdot t}_{=s\log(s/t)-s+t} = \hat{c}\,\exp(-\hat{c}/\varepsilon)\cdot t + \varepsilon\,[\exp(-\hat{c}/\varepsilon)\cdot(-\hat{c}/\varepsilon) - \exp(-\hat{c}/\varepsilon) + 1]\cdot t$$
$$= -\varepsilon\cdot[\exp(-\hat{c}/\varepsilon) - 1]\cdot t$$

- back to full problem, arrive at regularized dual:

$$(\ldots) = \sup_{\phi,\psi\in\mathbb{R}^N} \langle\phi,\mu\rangle + \langle\psi,\nu\rangle - \varepsilon\sum_{i,j}\left[\exp\left(-\tfrac{c_{i,j}-\phi_i-\psi_j}{\varepsilon}\right) - 1\right]\mu_i\nu_j$$

- discussion: the term $-\varepsilon\exp\left(-\frac{c_{i,j}-\phi_i-\psi_j}{\varepsilon}\right)$ acts like a smooth approximation of the constraint $c_{i,j}-\phi_i-\psi_j\geq 0$. if the constraint is violated, the term tends to $+\infty$ as $\varepsilon\to 0$. if the constraint is satisfied, the penalty tends to 0.

- the last term, $-\varepsilon\cdot(-1)$ is constant and tends to 0 as $\varepsilon\to 0$.

- so we have a smooth, unconstrained approximation of the original dual problem

- observe: still have the invariance under constant shifts of ϕ and ψ, but by convexity of exp the objective is strictly concave up to these constant shifts, i.e. dual maximizers are unique up to these shifts

Primal-dual optimality condition.

- For the unregularized problem we obtained the primal-dual optimality condition

$$c - \phi\oplus\psi = 0 \quad \pi\text{-almost everywhere.}$$

  Now study the equivalent relation for the regularized case. We will show $\pi\in\Pi(\mu,\nu)$ and $\phi,\psi: X\to\mathbb{R}$ are primal-dual optimal if and only if

$$\pi = \exp\left(\tfrac{-c+\phi\oplus\psi}{\varepsilon}\right)\cdot\mu\otimes\nu.$$

- For simplicity, again we consider only the discrete setting. Recall the derivation of the regularized dual problem, when we explicitly minimized over the entries of π. We obtained:

$$s\cdot\hat{c} + \varepsilon\,\varphi(s/t)\cdot t \geq -\varepsilon[\exp(-\hat{c}/\varepsilon) - 1]\cdot t$$

  where $s=\pi_{i,j}$, $t=\mu_i\nu_j$, with equality if and only if $s=\exp(-\hat{c}/\varepsilon)\cdot t$.

- Now apply this to the primal dual gap:

$$[\langle c,\pi\rangle + \varepsilon\ \mathrm{KL}(\pi|\mu\otimes\nu)] - \left[\langle\phi,\mu\rangle + \langle\psi,\nu\rangle - \varepsilon\sum_{i,j}\left[\exp\left(\tfrac{-c_{i,j}+\phi_i+\psi_j}{\varepsilon}\right)-1\right]\mu_i\nu_j\right]$$

(now use $\pi\in\Pi(\mu,\nu)$, i.e. $P_1\pi=\mu,\dots$)

$$= \sum_{i,j}\left[(c_{i,j}-\phi_i-\psi_j)\cdot\pi_{i,j} + \varepsilon\,\varphi(\pi_{i,j}/\mu_i\nu_j)\cdot\mu_i\nu_j\right] + \varepsilon\left[\exp\left(\tfrac{-c_{i,j}+\phi_i+\psi_j}{\varepsilon}\right)-1\right]\mu_i\nu_j\Big]$$

(now for each i,j the corresponding term is of the above form, so we get:)

≥ 0

with equality if and only if $\pi_{i,j}=\exp\left(\frac{-c_{i,j}+\phi_i+\psi_j}{\varepsilon}\right)$ for all i,j.

3.2 Sinkhorn algorithm

Derivation as alternating dual maximization.

- Consider now the dual objective for fixed ψ and optimize over ϕ. The objective can be written as:

$$\sum_i\left[\mu_i\cdot\phi_i - \varepsilon\mu_i\exp(\phi_i/\varepsilon)\cdot\sum_j\exp\left(-\tfrac{c_{i,j}-\psi_j}{\varepsilon}\right)\nu_j\right] + \langle\psi,\nu\rangle + \varepsilon\sum_{i,j}\mu_i\nu_j$$

- So we can optimize over each ϕ_i individually. Take partial derivative and set to zero:

$$0 = \mu_i\left[1-\exp(\phi_i/\varepsilon)\cdot\sum_j\exp\left(-\tfrac{c_{i,j}-\psi_j}{\varepsilon}\right)\nu_j\right]$$

- Resolve for ϕ, analogous formula for optimization over ψ:

$$\phi_i = -\varepsilon\log\left[\sum_j\exp\left(-\tfrac{c_{i,j}-\psi_j}{\varepsilon}\right)\nu_j\right],\qquad \psi_j = -\varepsilon\log\left[\sum_i\exp\left(-\tfrac{c_{i,j}-\phi_i}{\varepsilon}\right)\mu_i\right]$$

- Now, if we start with some initial $\phi^{(0)}$, $\psi^{(0)}$, then generate $\phi^{(1)}$ by optimizing over ϕ, then $\psi^{(1)}$ by optimizing over ψ and keep on iterating, the update rule is given by:

$$\phi_i^{(\ell)} := -\varepsilon\log\left[\sum_j\exp\left(-\tfrac{c_{i,j}-\psi_j^{(\ell-1)}}{\varepsilon}\right)\nu_j\right],\quad \psi_j^{(\ell)} := -\varepsilon\log\left[\sum_i\exp\left(-\tfrac{c_{i,j}-\phi_i^{(\ell)}}{\varepsilon}\right)\mu_i\right]$$

for $\ell\geq 1$.

Reformulation with scaling factors.

- Define the matrix $k\in\mathbb{R}_{++}^{N\times N}$ via $k_{i,j} := \exp(-c_{i,j}/\varepsilon)$. Introduce the scaling factors $u^{(\ell)}, v^{(\ell)}\in\mathbb{R}_+^N$ via

$$u_i^{(\ell)} := \exp(\phi_i^{(\ell)}/\varepsilon),\qquad\qquad v_j^{(\ell)} := \exp(\psi_j^{(\ell)}/\varepsilon).$$

- Then the above iterations for $\phi^{(\ell)}$ and $\psi^{(\ell)}$ can be equivalently rewritten as

$$u_i^{(\ell)} := \frac{1}{\sum_j k_{i,j}\,v_j^{(\ell-1)}\,\nu_j},\qquad\qquad v_j^{(\ell)} := \frac{1}{\sum_i k_{i,j}\,u_i^{(\ell)}\,\mu_i}.$$

- This can be compactly written as

$$u^{(\ell)} := \frac{1}{k.\,\mathrm{diag}(\nu).v^{(\ell-1)}}, \qquad\qquad v^{(\ell)} := \frac{1}{k^\top.\,\mathrm{diag}(\mu).u^{(\ell)}}.$$

  where the . denotes matrix-vector multiplication and the fraction of to vectors is two be understood entry-wise. This is the famous Sinkhorn algorithm and its main loop can be written in two lines in most scientific computing environments.

- Note that since $\phi, \psi \in \mathbb{R}^N$, one has $u, v = \exp(\phi/\varepsilon), \exp(\psi/\varepsilon) \in \mathbb{R}^N_{++}$ and also $k = \exp(-c/\varepsilon) \in \mathbb{R}^{M\times N}_{++}$, the division is always well-defined. However, numerically this may become an issue.

Corresponding primal sequence.

- Recall the primal-dual optimality condition:

$$\pi_{i,j} = \exp\left(-\frac{c_{i,j} - \phi_i - \psi_j}{\varepsilon}\right)\mu_i\nu_j = u_i \cdot k_{i,j} \cdot v_j \cdot \mu_i\nu_j$$

  So we can associate a sequence of primal iterates with the dual iterates. Note the following:

$$\sum_j u_i^{(\ell)}\, k_{i,j}\, v_j^{(\ell-1)}\mu_i\nu_j = \sum_j \frac{1}{\sum_{j'} k_{i,j'}\, v_{j'}^{(\ell-1)}\nu_{j'}} k_{i,j}\, v_j^{(\ell-1)}\mu_i\nu_j = \mu_i$$
$$\sum_i u_i^{(\ell)}\, k_{i,j}\, v_j^{(\ell)}\mu_i\nu_j = \sum_i \frac{1}{\sum_{i'} k_{i',j}\, u_{i'}^{(\ell)}\mu_{i'}} k_{i,j}\, u_i^{(\ell)}\,\mu_i\nu_j = \nu_j$$

- So after a u-update, the primal iterate satisfies the row-constraints, after a v-update it satisfies the column-constraints.

- The updates can be interpreted as re-scaling each row or column such that those constraints are satisfied.

- Since the map from dual to primal iterates is continuous, convergence of dual iterates implies convergence of primal iterates.

- By stationarity of the optimal duals (under further iterations), the limit of the primal iterates satisfies row and column constraints and therefore, by the primal-dual optimality condition, must be the unique primal minimizer.

Some comments on the Sinkhorn algorithm.

- The algorithm becomes numerically unstable and slow as $\varepsilon \to 0$. Good tricks for numerical stabilization exist.

- Some results on speed of convergence:

  - Franklin, Lorenz, Linear Algebra and its Applications, (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, SIAM J. Sci. Comput. (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, Numerische Mathematik (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.

- One of big strengths of Sinkhorn 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 reduce the problem dimensionality), and unbalanced transport problems.