Lectures on Optimal Transport. May, 2022. KAIST

Young-Heon Kim (Department of Mathematics, University of British Columbia)

Lecture 1	The Monge-Kantorovich problem and duality.
Lecture 2	Wasserstein geometry of the space of probability measures.
Lecture 3	Entropic regularization of optimal transport. **Today!**
Lecture 4	Application of optimal transport to developmental processes.
Lecture 5	Multimarginal optimal transport. Wasserstein barycentre.
Seminar	Optimal Brownian stopping with free target and the supercooled Stefan problem.
Lecture 6	Optimal martingale transport
Lecture 7	Optimal Brownian martingale transport

Some references for the lectures

- Lecture 1, 2, and 3:
  - Villani: Topics in Optimal Transport. Book
  - Villani: Optimal Transport. Old and New. Book
  - Cuturi & Payré: Computational Optimal Transport. Book
- Lecture 4:
  - Schiebinger: https://broadinstitute.github.io/wot/tutorial/
  - Kim, Lavenant, Schiebinger, Zhang: Towards a mathematical theory of trajectory inference. https://arxiv.org/abs/2102.09204
- Lecture 5
  - Cuturi & Payré: Computational Optimal Transport. Book
  - Kim & Pass: Wasserstein Barycenters over Riemannian manifolds. Adv. in Math. 2017.
- Lecture 6
  - Ghoussoub, Kim, & Lim: Structure of optimal martingale transport in general dimensions. Ann. Prob. 2019.
- Lecture 7
  - Ghoussoub, Kim, & Palmer: PDE Methods For Optimal Skorokhod Embeddings. Calc. Var. 2019.
  - Ghoussoub, Kim, & Palmer: A solution to the Monge transport problem for Brownian martingales. Ann. Prob. 2021.
  - I. Kim & Y. Kim.The Stefan problem and free targets of optimal Brownian martingale transport. Preprint. 2021

- **Computation of Optimal Transport: Entropic regularization**
- **Schrödinger problem**
    - Entropy regularized OT has theoretical implications.

Part 1. Entropic regularization of optimal transport

Computation of Optimal Transport

Find a transport plan π that solves

$$
\begin{array}{rl}
\text{Minimize} & \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\pi_{ij} \\
\text{subject to} & \sum_{j=1}^{n} \pi_{ij} = \mu_i \\
 & \sum_{i=1}^{m} \pi_{ij} = \nu_j \\
 & \pi_{ij} \geq 0 \quad \text{for all } i \text{ and } j.
\end{array}
$$

- Simpler notation:

$$
\min_{\pi \in \Pi(\mu,\nu)} \sum_{i,j} c_{ij}\pi_{ij}
$$

This is a linear programming problem, but, it is still costly to solve if m, n are large.

- the dimension for the decision variable π is mn.
- the number of constraints is $mn + m + n$.

Entropic regularization

[Marco Cuturi 2013]

Solve for small $\epsilon > 0$, $\min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{i,j} c_{ij}\pi_{ij} + \epsilon S(\pi) \right]$

- Here

$$S(\pi) = \sum_{i,j} [\pi_{ij} \log \pi_{ij} - \pi_{ij}]$$

  with the convention that $0 \log 0 = 0$.

- $x \in [0, \infty] \mapsto x \ln x - x$ is strictly convex.
- the function $S(\pi)$] is strictly convex, and so is $\pi \mapsto \sum_{ij} c_{ij}\pi_{ij} + \epsilon S(\pi)$.
- $\min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{ij} c_{ij}\pi_{ij} + \epsilon S(\pi) \right]$ has unique optimal solution π^{ϵ}.
- Finding the optimal solution π^{ϵ} is (relatively) easy. In fact, much faster!

$$\min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{ij} c_{ij}\pi_{ij} + \epsilon S(\pi) \right]$$

- The smaller ϵ, the closer to the original optimal solution.

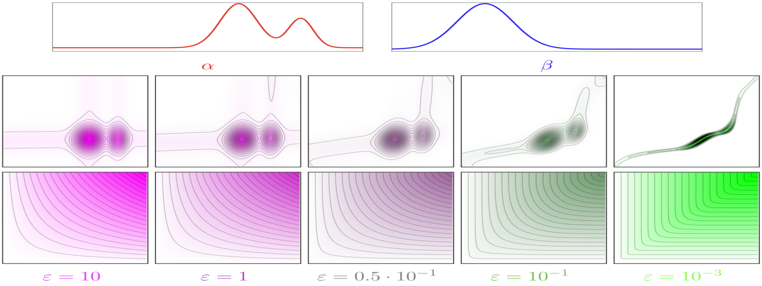

Image from the book “Computational Optimal Transport” by Cuturi and Peyré

Convergence as $\epsilon \to 0^+$

Theorem

As $\epsilon \to 0^+$,

$$\min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{i,j} c_{ij}\pi_{ij} + \epsilon S(\pi) \right] \longrightarrow \min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{i,j} c_{ij}\pi_{ij} \right]$$

$$\pi^{\epsilon} \longrightarrow \pi^{*} \text{ (in weak*)}$$

where π^* is the optimal solution of the original optimal transport problem.

Proof.

See e.g. Theorem 2.7., Carlier et al., SIAM J. Math. Anal. 49 (2017), no. 2, 1385–1418. MR 3635459. □

Duality for entropy regularized OT

- $$\begin{array}{ll} \text{Minimize} & \sum_{ij} c_{ij}\pi_{ij} + \epsilon S(\pi) \\ \text{subject to} & \sum_{j=1}^{n} \pi_{ij} = \mu_i \\ & \sum_{i=1}^{m} \pi_{ij} = \nu_j \\ & \pi_{ij} \geq 0 \end{array}$$

  is equivalent to

- $$\min_{\pi \geq \mathbf{0}} \max_{\phi,\psi} \left\{ \sum_{ij}^{m} c_{ij}\pi_{ij} + \epsilon S(\pi) \right.$$
  $$\left. + \sum_i \phi_i \left[\mu_i - \sum_{j=1}^{n} \pi_{ij} \right] + \sum_j \psi_i \left[\nu_j - \sum_{i=1}^{m} \pi_{ij} \right] \right\}$$

Duality for entropy regularized OT

$$\min_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{ij} c_{ij}\pi_{ij} + \epsilon S(\pi) \right\}$$

$$= \min_{\pi\geq\mathbf{0}} \max_{\phi,\psi} \left\{ \sum_{ij}^{m} c_{ij}\pi_{ij} + \epsilon S(\pi) \right.$$

$$\left. + \sum_i \phi_i \left[\mu_i - \sum_{j=1}^{n} \pi_{ij} \right] + \sum_j \psi_i \left[\nu_j - \sum_{i=1}^{m} \pi_{ij} \right] \right\}$$

$$= \max_{\phi,\psi} \min_{\pi\geq\mathbf{0}} [\cdots\cdots] \qquad \text{(justifiable)}$$

$$= \max_{\phi,\psi} \left\{ \sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j \right.$$

$$\left. + \min_{\pi\geq\mathbf{0}} \left[\sum_{ij}^{m} [c_{ij} - \phi_i - \psi_j]\pi_{ij} + \epsilon S(\pi) \right] \right\}$$

Lemma

$$\min_{\pi \geq \mathbf{0}} \Big[\sum_{ij}^{m} [c_{ij} - \phi_i - \psi_j] \pi_{ij} + \epsilon S(\pi) \Big]$$
$$= -\epsilon \sum_{ij} e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]}$$

and minimum at $\pi = e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]}$.

Proof.

- $F(\pi) := \sum_{ij}^{m} [c_{ij} - \phi_i - \psi_j] \pi_{ij} + \epsilon S(\pi)$ is a strictly convex function.
- Using $S(\pi) = \sum_{ij} [\pi_{ij} \log \pi_{ij} - \pi_{ij}]$, we get $(\partial_\pi F)_{ij} = c_{ij} - \phi_i - \psi_j + \epsilon \log \pi_{ij}$, and $\partial_\pi F(\pi) \to -\infty$ as $\pi_{ij} \to 0$ for any (i, j).
- So, it has a minimum when $\partial_\pi F(\pi) = 0$.
- And

$$\partial_\pi F = 0 \text{ iff } \pi = e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]}.$$

- Plug-in $\pi = e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]}$ to $F(\pi)$ and get

$$-\epsilon \sum_{ij} e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]}.$$

$\square$

Non essential assumption for simplicity

From now on, for technical simplicity, let us assume that $\mu_i, \nu_j > 0$ for all i, j.

Duality

Therefore we get

$$\min_{\pi\in\Pi(\mu,\nu)}\left\{\sum_{ij} c_{ij}\pi_{ij} + \epsilon S(\pi)\right\} \quad \textbf{Primal}$$

$$= \max_{\phi,\psi}\left[\sum_i \phi_i\mu_i + \sum_j \psi_j\nu_j - \epsilon\sum_{ij} e^{-\frac{1}{\epsilon}[c_{ij}-\phi_i-\psi_j]}\right] \quad \textbf{Dual}.$$

- **Weak duality:** The primal objective values $\geq$ the dual objective values.
- For optimal primal solution π^ϵ and optimal dual solution $(\phi^\epsilon, \psi^\epsilon)$, we have

$$\sum_{ij} c_{ij}\pi^\epsilon_{ij} + \epsilon S(\pi^\epsilon)$$

$$= \sum_i \phi^\epsilon_i\mu_i + \sum_j \psi^\epsilon_j\nu_j - \epsilon\sum_{ij} e^{-\frac{1}{\epsilon}[c_{ij}-\phi^\epsilon_i-\psi^\epsilon_j]}$$

- **Exercise:** The previous lemma implies that

$$\pi^\epsilon_{ij} = e^{-\frac{1}{\epsilon}[c_{ij}-\phi^\epsilon_i-\psi^\epsilon_j]}.$$

- So, the dual optimal solution gives the primal optimal solution explicitly!
- Notice that the optimal solution π^ϵ_{ij} is positive everywhere. This is in contrast to the ordinary optimal transport where the optimal transport plan is concentrated along a special set in the product space.

Optimality condition for the dual

$$\textbf{Dual} \quad \max_{\phi,\psi} \left[\sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \epsilon \sum_{ij} e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]} \right].$$

- The objective function for dual is strictly concave due to the exponential terms.
- $(\phi^\epsilon, \psi^\epsilon)$ is a dual optimal solution if and only if

$$0 = \frac{\partial}{\partial_{\phi,\psi}} \Big|_{(\phi,\psi)=(\phi^\epsilon,\psi^\epsilon)} \left[\sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \epsilon \sum_{ij} e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i - \psi_j]} \right]$$

if and only if

$$\sum_j e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i^\epsilon - \psi_j^\epsilon]} = \mu_i$$

$$\sum_i e^{-\frac{1}{\epsilon}[c_{ij} - \phi_i^\epsilon - \psi_j^\epsilon]} = \nu_j.$$

How to find optimal solution

We need to find some $u_i^\epsilon := e^{-\frac{1}{\epsilon}\phi_i^\epsilon}$, $v_j^\epsilon := e^{-\frac{1}{\epsilon}\psi_j^\epsilon}$, so that

$$e^{-\frac{1}{\epsilon}c_{ij}} u_i^\epsilon v_j^\epsilon$$

satisfies

$$\sum_j e^{-\frac{1}{\epsilon}c_{ij}} u_i^\epsilon v_j^\epsilon = \mu_i$$
$$\sum_i e^{-\frac{1}{\epsilon}c_{ij}} u_i^\epsilon v_j^\epsilon = \nu_j.$$

which is

$$u_i^\epsilon = \frac{\mu_i}{\sum_j e^{-\frac{1}{\epsilon}c_{ij}} v_j^\epsilon}$$
$$v_j^\epsilon = \frac{\nu_j}{\sum_i e^{-\frac{1}{\epsilon}c_{ij}} u_i^\epsilon}$$

Optimality condition

- Define the matrix $K = [K_{ij}] = [e^{-\frac{1}{\epsilon} c_{ij}}]$. Rewrite

$$u_i^\epsilon = \frac{\mu_i}{\sum_j e^{-\frac{1}{\epsilon} c_{ij}} v_j^\epsilon}$$
$$v_j^\epsilon = \frac{\nu_j}{\sum_i e^{-\frac{1}{\epsilon} c_{ij}} u_i^\epsilon}$$

as

$$u_i = \frac{\mu_i}{[K\vec{v}]_i}$$
$$v_j = \frac{\nu_j}{[K^T\vec{u}]_j}$$

Simply,

$$\vec{u} = \frac{\vec{\mu}}{K\vec{v}}$$
$$\vec{v} = \frac{\vec{\nu}}{K^T\vec{u}}$$

[Sinkhorn algorithm, 1964]

- $$\vec{u}^{(l+1)} \stackrel{def}{=} \frac{\vec{\mu}}{K\vec{v}^{(l)}}$$
  $$\vec{v}^{(l+1)} \stackrel{def}{=} \frac{\vec{\nu}}{K^T\vec{u}^{(l+1)}}$$
- Send $(u^{(l)}, v^{(l)})$ to $(u^{(l+1)}, v^{(l)})$ then to $(u^{(l+1)}, v^{(l+1)})$.
- At the limit $l \to \infty$, we get

$$\vec{u}^{(\infty)} = \frac{\vec{\mu}}{K\vec{v}^{(\infty)}}$$
$$\vec{v}^{(\infty)} = \frac{\vec{\nu}}{K^T\vec{u}^{(\infty)}}$$ (satisfying optimality condition).

- The convergence to the solution $(u^{(\infty)}, v^{(\infty)})$ is **exponentially fast.**
- Note $K = [K_{ij}] = [e^{-\frac{1}{\epsilon}c_{ij}}]$ depends on ϵ in our case.
- Smaller the ϵ, slower the algorithm (still exponentially fast).

Contraction

Hibert projective metric on $\mathbb{R}^n_{+,*}$.

$$\forall (\mathbf{u}, \mathbf{u}') \in (\mathbb{R}^n_{+,*})^2, \quad d_{\mathcal{H}}(\mathbf{u}, \mathbf{u}') \stackrel{\text{def.}}{=} \log \max_{i,j} \frac{\mathbf{u}_i \mathbf{u}'_j}{\mathbf{u}_j \mathbf{u}'_i}.$$

Theorem 4.1. Let $\mathbf{K} \in \mathbb{R}^{n \times m}_{+,*}$; then for $(\mathbf{v}, \mathbf{v}') \in (\mathbb{R}^m_{+,*})^2$

$$d_{\mathcal{H}}(\mathbf{K}\mathbf{v}, \mathbf{K}\mathbf{v}') \leq \lambda(\mathbf{K}) d_{\mathcal{H}}(\mathbf{v}, \mathbf{v}'), \text{ where } \begin{cases} \lambda(\mathbf{K}) \stackrel{\text{def.}}{=} \frac{\sqrt{\eta(\mathbf{K})}-1}{\sqrt{\eta(\mathbf{K})}+1} < 1, \\ \eta(\mathbf{K}) \stackrel{\text{def.}}{=} \max_{i,j,k,\ell} \frac{\mathbf{K}_{i,k}\mathbf{K}_{j,\ell}}{\mathbf{K}_{j,k}\mathbf{K}_{i,\ell}}. \end{cases}$$

Figure 4.7 illustrates this theorem.

Images from the book "Computational Optimal Transport" by Cuturi and Peyré

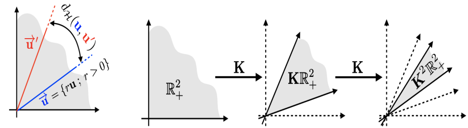

Figure 4.7: Left: the Hilbert metric $d_{\mathcal{H}}$ is a distance over rays in cones (here positive vectors). Right: visualization of the contraction induced by the iteration of a positive matrix $\mathbf{K}$.

So, we see:

- $$\begin{aligned} d_{\mathcal{H}}(\vec{u}^{(l+1)}, \vec{u}^{(l)}) &= d_{\mathcal{H}}\left(\frac{\vec{\mu}}{K\vec{v}^{(l)}}, \frac{\vec{\mu}}{K\vec{v}^{(l-1)}}\right) \\ &= d_{\mathcal{H}}(K\vec{v}^{(l)}, K\vec{v}^{(l-1)}) \\ &\leq \lambda(K) d_{\mathcal{H}}(\vec{v}^{(l)}, \vec{v}^{(l-1)}) \end{aligned}$$
- $$d_{\mathcal{H}}(\vec{v}^{(l+1)}, \vec{v}^{(l)}) \leq \lambda(K) d_{\mathcal{H}}(\vec{u}^{(l+1)}, \vec{u}^{(l)}).$$

$$\min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{i,j} c_{ij}\pi_{ij} + \epsilon S(\pi) \right]$$

- The smaller ϵ, the closer to the original optimal solution.

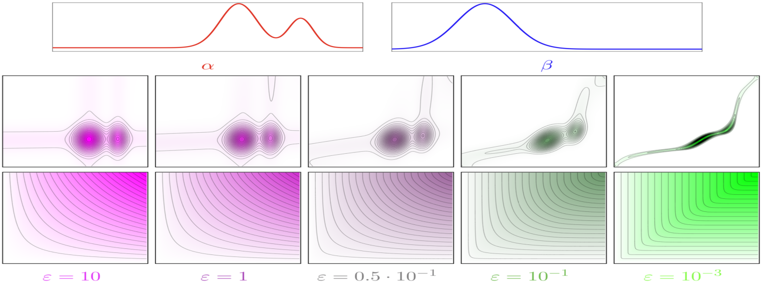

Image from the book "Computational Optimal Transport" by Cuturi and Peyré

$$\min_{\pi \in \Pi(\mu,\nu)} \left[\sum_{i,j} c_{ij}\pi_{ij} + \epsilon S(\pi) \right]$$

Entropic regularization of OT makes OT problems effectively computable.

- Entropic regularization gives an approximate optimal solution to the OT problem.
- Applying the Sinkhorn algorithm, we get those approximate solutions in a very fast way.
- In many practical applications, the approximate optimal solutions are good enough.

There are many (practical) applications of OT.

Part 2: Schrödinger's problem

A reference:

- Christian Léonard: *A survey of the Schrödinger problem and some of its connections with optimal transport*. https://arxiv.org/abs/1308.0215.

The Schrödinger problem: motivation

- Suppose:
  - Particles follow a given stochastic process, e.g. the Brownian motion $dW_t^\epsilon = \sqrt{\epsilon} dW_t$ with diffusivity ϵ, in $\mathbb{R}^d$ with the law R_ϵ.
  - We observed
    - $\mu \approx$ the given initial configuration of given many (N, $N \gg 1$) independent particles at $t = 0$.
    - $\nu \approx$ the given final configuration of those particles at $t = 1$.
- Note that the law of large numbers tells that with high probability the final distribution should look like the heat flow of μ at time 1.
- Since $N < \infty$ still there is a chance for ν to be different from it, but, with small probability.
- What if ν is very different from the heat flow?
- **Schrödinger's question:** "Conditionally on this very rare event, what is the most likely path (i.e. transport plan) of the whole system between the times $t = 0$ and $t = 1$?"
- How does it look like as $N \to \infty$"
- It turned out that this is reduced to a minimization problem of **the relative entropy**.

The Schrödinger problem

- Let

$$R_\epsilon(x,y) := C_\epsilon e^{-\frac{1}{\epsilon}|x-y|^2}\mu(x) \quad \text{where } C_\epsilon = C\epsilon^{-d/2} \text{ and } \int_{\mathbb{R}^d} R_\epsilon dy = 1.$$

  - It gives the law of the distribution of Brownian motion W_t^ϵ starting from the distribution μ, at time 1 with the diffusion coefficient ϵ, that is, $dW_t^\epsilon = \sqrt{\epsilon} dW_t$, where W_t is the standard Brownian motion.
- Relative entropy with respect to R_ϵ:

$$Ent(\pi \big| \, R_\epsilon) := \int_{\mathbb{R}^d \times \mathbb{R}^d} \pi \log \frac{\pi}{R^\epsilon} dxdy.$$

  - The relative entropy $Ent(\pi \big| \, R_\epsilon)$ compares a transport plan to the transport by the Brownian motion.
- **Schrödinger's problem:**

$$\min_{\pi \in \Pi(\mu,\nu)} Ent(\pi \big| \, R_\epsilon).$$

Relation to the entropy regularized OT

- Recall

$$R_\epsilon(x,y) := C_\epsilon e^{-\frac{1}{\epsilon}|x-y|^2}\mu(x) \quad \text{where } C_\epsilon = C\epsilon^{-d/2} \text{ and } \int_{\mathbb{R}^d} R_\epsilon dy = 1.$$

$$Ent(\pi| \, R_\epsilon) := \int_{\mathbb{R}^d\times\mathbb{R}^d} \pi \log \frac{\pi}{R^\epsilon} dxdy.$$

- Notice that for $\pi \in \Pi(\mu,\nu)$,

$$\begin{aligned} \epsilon Ent(\pi| \, R_\epsilon) = & \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2 d\pi + \epsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} \pi \log \pi \, dxdy \\ & - \epsilon \int_{R^d} \mu \log \mu dx + \epsilon \left[\frac{d}{2} \log \epsilon - \log C \right]. \end{aligned}$$

- Therefore,

$$\min_{\pi\in\Pi(\mu,\nu)} Ent(\pi| \, R_\epsilon) \text{ is equivalent to } \min_{\pi\in\Pi(\mu,\nu)} \left[\int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2 d\pi(x,y) + \epsilon S(\pi) \right]$$

  with **the same optimal solution** π^ϵ**.**

- **For fixed** ϵ**,** the results between these two are interchangeable.

As $\epsilon \to 0^+$:

Theorem

As $\epsilon \to 0^+$*,*

$$\epsilon \min_{\pi \in \Pi(\mu,\nu)} [Ent(\pi \mid R_\epsilon)] \longrightarrow \min_{\pi \in \Pi(\mu,\nu)} \left[\int |x-y|^2 d\pi \right]$$

$$\pi^\epsilon \longrightarrow \pi^* \text{ (in weak*)}$$

Proof.

See e.g. Theorem 3.3. in Christian Léonard: *From the Schrödinger problem to the Monge-Kantorovich problem.* https://doi.org/10.1016/j.jfa.2011.11.026. □

There are many more results and recent developments.

- A quantitative convergence of $\pi^\epsilon \to \pi^*$ is available in certain cases.