3.36pt

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. **Today!**
Lecture 3	Entropic regularization of optimal transport.
Seminar	Optimal Brownian stopping with free target and the supercooled Stefan problem.
Lecture 4	Application of optimal transport to developmental processes.
Lecture 5	Multimarginal optimal transport. Wasserstein barycentre.
Lecture 6	Optimal marginale 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

Recall

$$MK(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)d\pi(x,y).$$

Theorem (Brenier 80's)

Suppose

- $X = Y = \mathbb{R}^n$, μ, ν *are probability measures, compactly supported.*
- $\mu \ll$ *Lebesgue,*
  - *that is,* $d\mu = fdm$, *for the Lebesgue measure m, f measurable;*
- $c(x,y) = |x-y|^2$.

Then,

- *there exists unique optimal solution* π^* *to* $MK(\mu,\nu)$;
- π^* *is given by a measurable mapping* $T : \mathbb{R}^n \to \mathbb{R}^n$ *defined* μ*-a.e., that is,* $\pi^* = (id \times T)_{\#}\mu$;
- T *is given by a convex function* $\phi : \mathbb{R}^n \to \mathbb{R}$ *in the sense that* $T(x) = \nabla\phi(x)$ *for* μ*-a.e.* x.

In this case, $d_W^2(\mu,\nu) = \int |x - \nabla\phi(x)|^2 d\mu(x)$.

Wasserstein distance between probability measures on $\mathbb{R}^d$

Let $X = \mathbb{R}^d$, $dist(x, y) = |x - y|$.
(More generally, $(X, dist)$ can be a separable, complete metric space;e.g. path space $C([0, 1]; \mathbb{R}^d)$ with uniform metric on curves.)

- A **distance** between probability measures μ, ν on X.

$$d_W(\mu, \nu) = \sqrt{\min_{\pi \in \Pi(\mu,\nu)} \int_X \int_X \mathrm{dist}^2(x, y) d\pi(x, y)}$$

  called **the Wasserstein distance**.
  For $p > 1$, p-Wasserstein distance: replace 2 with p.

- **Triangle inequality:** $d_W(\mu_1, \mu_3) \leq d_W(\mu_1, \mu_2) + d_W(\mu_2, \mu_3)$.

- $P(X)$ =“the space of probability measures on X", becomes a natural **metric space** with d_W:

  - Isometric imbedding $X \ni x \mapsto \delta_x \in P(X)$. $d_W(\delta_x, \delta_y) = dist(x, y)$.

Wasserstein distance and weak* topology

- **weak* topology:** $\mu_k \to \mu$ in weak* **iff** $\forall f \in C_0(\mathbb{R}^d), \quad \int f d\mu_k \to \int f d\mu$.

- **Theorem:** For $\mu_k, \mu \in P(\mathbb{R}^d)$,

$$\lim_{k\to\infty} d_W(\mu_k, \mu) = 0$$

iff (1) $\mu_k \to \mu$ in weak* and (2) $d_W(\delta_0, \mu_k) \to d_W(\delta_0, \mu)$.
(Proof is long but straightforward.)

Wasserstein geodesics

A curve $\sigma : [0,1] \to P(\mathbb{R}^d)$ is said to be a (d_W-length minimizing) **geodesic** if $\forall s, t \in [0,1]$,

$$d_W(\sigma(s), \sigma(t)) = |s - t| d_W(\sigma(0), \sigma(1)).$$

Notation: $P_2(\mathbb{R}^d) = \{\mu \in P(\mathbb{R}^d) \mid d_W(\delta_0, \mu) < \infty\}$.

Theorem: A geodesic exists between any $\mu_0, \mu_1 \in P_2(\mathbb{R}^d)$. More precisely,

Theorem (McCann's displacement interpolation)

- *Let* $c(x,y) = |x-y|^2$ *and* $\mu_0, \mu_1 \in P_2(\mathbb{R}^d)$.
- *Let* $\pi_0 \in \Pi_{op}(\mu_0, \mu_1)$ *(*$\leftarrow$ *the set of optimal transport plans).*
- *For each* $s \in [0,1]$, *define* $I_s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ *by* $I_s(x,y) = (1-s)x + sy$.
- *Let* $\mu_s := (I_s)_{\#}\pi_0$. *(*$\leftarrow$ *"Displacement interpolation between* μ_0 *and* μ_1*.")*

Then

- $s \mapsto \mu_s$ *is a* d_W*-length minimizing geodesic between* μ_0 *and* μ_1.
- *Moreover,* $\pi_s := (I_0 \times I_s)_{\#}\pi_0 \in \Pi_{op}(\mu_0, \mu_s)$.

Example

If $\pi_0 = (id \times \nabla\psi)_{\#}\mu_0$ ($\leftarrow$ Here $\nabla\psi$ is the Monge solution of Brenier.)
then, $\mu_s = ((1-s)id + s\nabla\psi)_{\#}\mu_0$.

Differential Geometry on the space of probability measures

- **Notation:** $P_{2,ac}(\mathbb{R}^d) = P_2(\mathbb{R}^d) \cap \{\mu \mid \mu \ll Leb\}$.
- We can consider "smooth" (in weak sense) curves $\rho : [-\delta, \delta] \to P_{2,ac}(\mathbb{R}^d)$ as a "smooth" (in weak sense) family of probability measures.

- For $\rho \in P_{2,ac}(\mathbb{R}^d)$, *roughly speaking,* the tangent space $T_\rho P_{2,ac}(\mathbb{R}^d)$, is given as

$$T_\rho P_{2,ac}(\mathbb{R}^d) \text{ " } = \text{ " } \left\{ \frac{\partial \rho}{\partial t}\Big|_{t=0} \;\Big|\; \text{for a smooth curve } \rho(t),\ -\delta \le t \le \delta, \text{ in } P_{2,ac}(\mathbb{R}^d) \right\}$$

- An infinitesimal version of Wasserstein metric?
  Question: How to define metric (norm) $\| \frac{\partial \rho}{\partial t} \|_\rho$ at $T_\rho P_{2,ac}(\mathbb{R}^d)$ such that

$$d_W^2(\mu, \nu) = \inf_{\text{curve } \rho_t \text{ in } P_{2,ac} \text{ with } \rho_0 = \mu, \rho_1 = \nu} \left\{ \int_0^1 \left\| \frac{\partial \rho_t}{\partial t} \right\|_{\rho_t}^2 dt \right\} ?$$

  It is natural to set

$$\left\| \frac{\partial \rho_t}{\partial t} \right\|_{\rho_t}^2 = \left| \frac{d}{d\epsilon}\Big|_{\epsilon=0} d_W(\rho_t, \rho_{t+\epsilon}) \right|^2$$

Infinitesimal mass transport and continuity equation

- Underling idea: **Mass changes due to motion by vector fields.**

  change of mass distribution $\leftrightarrow$ vector fields.

- **Infinitesimal mass transport:**

$$(T_\epsilon)_\# \rho_t = \rho_{t+\epsilon} \Longleftrightarrow \rho_{t+\epsilon}(T_\epsilon(x)) \det(\nabla T_\epsilon(x)) = \rho_t(x).$$

When $T_\epsilon = id + \epsilon \vec{V} + o(\epsilon)$, differential the righthand side in ϵ at $\epsilon = 0$, and get

$$\partial_t \rho_t + \nabla \rho_t \cdot \vec{V} + \rho_t div \vec{V} = 0$$

That is,

$$\partial_t \rho_t + div(\rho_t \vec{V}) = 0 \quad \textbf{"continuity equation"}$$

- - A pair $(\rho, V) = (\rho_t, V_t)_{0 \le t \le 1}$ (time dependent distribution ρ_t and vector field V_t): is said to be *admissible* if it satisfies the continuity equation in a weak sense.
  - "Energy" of (ρ, V):

$$\int_0^1 \int |V_t|^2 \rho_t dx dt.$$

($\leftarrow$ convex in ρ and ρV.)

  - Length distance between $\mu, \nu \in P_{2,ac}(\mathbb{R}^d)$.

$$\tilde{d}_W(\mu, \nu) = \sqrt{\inf_{(\rho, V) \text{ admissible}, \rho_0 = \mu, \rho_1 = \nu} \int_0^1 \int |V_t|^2 \rho_t dx dt}.$$

  - Both the functional and constraint are linear in the mass ρ and the momentum ρV.

Infinitesimal optimal transport

- Take the optimal transport T_ϵ with $(T_\epsilon)_\# \rho_t = \rho_{t+\epsilon}$.
- Brenier $\Rightarrow$ $T_\epsilon = \nabla \psi_\epsilon$ for some convex ψ_ϵ.
- So,

$$T_\epsilon = \nabla \psi_\epsilon = id + \epsilon \nabla u + o(\epsilon) \quad \text{for some function } u : \mathbb{R}^d \to \mathbb{R} \text{ with } \frac{d}{d\epsilon}\Big|_{\epsilon=0} T_\epsilon = \nabla u.$$

- It follows

$$\begin{aligned} d_W^2(\rho_t, \rho_{t+\epsilon}) &= \int |x - T_\epsilon(x)|^2 \rho_t dx \\ &= \int |\epsilon \nabla u(x) + o(\epsilon)|^2 \rho_t dx = \epsilon^2 \int |\nabla u|^2 \rho_t dx + o(\epsilon). \end{aligned}$$

- Therefore,

$$\left| \frac{d}{d\epsilon}\Big|_{\epsilon=0} d_W(\rho_t, \rho_{t+\epsilon}) \right|^2 = \sqrt{\int |\nabla u|^2 \rho_t dx}$$

- We can define

$$\left\| \frac{\partial \rho_t}{\partial t} \right\|_{\rho_t}^2 := \int |\nabla u|^2 \rho_t dx \qquad \text{where } \nabla u = \lim_{\epsilon \to 0} T_\epsilon \text{ for } (T_\epsilon)_\# \rho_t = \rho_{t+\epsilon}.$$

Remark:

- Given a curve ρ_t there can be infinitely many $\vec{V}$ satisfying the continuity equation $\partial_t \rho + div \rho \vec{V} = 0$.
- The gradient vector field ∇u with the continuity equation $\partial_t \rho + div \rho \nabla u = 0$, is the one that has the smallest L^2 norm with respect to ρ:

$$\int |\nabla u|^2 \rho dx = \inf_{\text{admissible } \vec{V}} \int |\vec{V}|^2 \rho dx$$

- ∇u is the optimal infinitesimal transport!

Summary

- Continuity equation is an infinitesimal mass transport.
- Optimal infinitesimal mass transport is the continuity equation with the vector field $\vec{V}$ given by the gradient ∇u of a function!
- We have the correspondence

$$\partial_t \rho \quad \longleftrightarrow \quad \nabla u$$

with

$$\partial_t \rho + div(\rho \nabla u) = 0$$

- $$\left\| \frac{\partial \rho_t}{\partial t} \right\|_{\rho_t}^2 = \int |\nabla u|^2 \rho_t dx \qquad \text{where } \nabla u = \lim_{\epsilon \to 0} T_\epsilon \text{ for } (T_\epsilon)_\# \rho_t = \rho_{t+\epsilon}.$$

Benamou-Brenier

Theorem (Benamou-Brenier '97)

For $\mu, \nu \in P_{2,ac}(\mathbb{R}^d)$,

$$
\begin{aligned}
d_W^2(\mu,\nu) &= \inf_{\text{curve } \rho_t \text{ in } P_{2,ac} \text{ from } \rho_0=\mu \text{ to } \rho_1=\nu} \left\{ \int_0^1 \left\| \frac{\partial \rho_t}{\partial t} \right\|_{\rho_t}^2 dt \right\} \\
&= \inf_{\partial_t \rho + div(\rho \nabla u)=0,\ \rho_0=\mu, \rho_1=\nu} \left\{ \int_0^1 |\nabla u|^2 \rho dx dt \right\} \\
&= \tilde{d}_W^2(\mu,\nu) \\
&= \inf_{(\rho,V) \text{ admissible}, \rho_0=\mu, \rho_1=\nu} \int_0^1 \int |V_t|^2 \rho_t dx dt.
\end{aligned}
$$

Proof.

See e.g. [Villani, Topics in Optimal Transport]. □

Otto's metric on $P_{2,ac}(\mathbb{R}^d)$: an infinitesimal Wasserstein metric

- Recall the infinitesimal optimal transport equation, that is, the continuity equation + gradient vector field:

$$\partial_t \rho + div(\rho \nabla u) = 0.$$

- This gives the correspondence

$$\partial_t \rho \longleftrightarrow \nabla u.$$

while

$$\|\partial_t \rho\|_\rho^2 = \int |\nabla u|^2 \rho dx.$$

- Then, we can define the W_2 **Riemannian metric** for $\partial_t \rho^1, \partial_t \rho^2 \in T_\rho(P_{2,ac}(\mathbb{R}^d))$:

$$\langle \partial_t \rho^1, \partial_t \rho^2 \rangle_\rho = \int \langle \nabla u_1, \nabla u_2 \rangle \rho dx$$

with $\partial_t \rho_1 + div(\rho_1 \nabla u_1) = 0, \quad \partial_t \rho_2 + div(\rho_2 \nabla u_2) = 0.$

- Each metric $\langle , \rangle_\rho$ at $T_\rho(P_{2,ac}(\mathbb{R}^d))$ depends on ρ!

Otto's calculus: the gradient $grad_W$ with respect to the Wasserstein metric.

- Given a functional $\mathcal{F} : P_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$,
  and a curve ρ_t, $-\delta \leq t \leq \delta$, with $\partial_t \rho + div(\rho \nabla u) = 0$,

$$\frac{d}{dt}\Big|_{t=0} \mathcal{F}(\rho(t)) = \left\langle grad_W \mathcal{F}(\rho), \frac{\partial \rho}{\partial t}\Big|_{t=0} \right\rangle_\rho$$
$$= \int \langle ?, \nabla u \rangle \rho dx.$$

- For the correspondence

$$\partial_t \rho \longleftrightarrow \nabla u.$$

  what is the counterpart for $grad_W \mathcal{F}(\rho)$?

$$grad_W \mathcal{F}(\rho) \longleftrightarrow \quad ?$$

Otto's calculus: a key calculation

- Consider $\mathcal{F}(\rho) = \int U(\rho) dx$.
- Then

$$\begin{aligned}
\frac{d}{dt}\mathcal{F}(\rho) &= \int \frac{\delta U}{\delta \rho}(\rho)\partial_t \rho dx \\
&= -\int \frac{\delta U}{\delta \rho}(\rho) div(\rho \nabla u) dx \quad \text{(from } \partial_t \rho + div(\rho \nabla u) = 0.\text{)} \\
&= \int \langle \nabla \left(\frac{\delta U}{\delta \rho}(\rho) \right), \nabla u \rangle \rho dx.
\end{aligned}$$

- Therefore we have the correspondence:

$$grad_W \mathcal{F}(\rho) \longleftrightarrow \nabla \left(\frac{\delta U}{\delta \rho}(\rho) \right).$$

Gradient flows

- **Gradient flow**: **It is the steepest descent!**

- Given $F : X \to \mathbb{R}$, the gradient flow of F, is the curve $x(t)$ in X that satisfies

$$\frac{d}{dt}x(t) = -\nabla F(x(t))$$

  where the gradient ∇ is determined by the choice of Riemannian metric.
- Many physical systems can be understood as gradient flows of certain physical quantities, e.g. energy, entropy, etc.

Otto's calculus: Gradient flows, entropy and the heat equation

Example

$$\textbf{(Mathematical) Entropy:} \quad Ent(\rho) = \int \rho \log \rho dx.$$

Physical entropy is the negative of the mathematical one, by convention.

- Let $U(\rho) = \rho \log \rho$.
- Then

$$\partial_\rho U = \log \rho + 1. \text{ Thus } \nabla \partial_\rho U = \nabla[\log \rho(x)].$$

- So,

$$-grad_{W_2} Ent \quad \longleftrightarrow \quad -\nabla[\log \rho(x)].$$

- Therefore, **the gradient flow of the entropy functional with respect to the Wasserstein metric** is:

$$\partial_t \rho_t + div(\rho_t(-\nabla \log \rho_t)) = 0 \quad \text{that is,} \quad \partial_t \rho - div \nabla \rho = 0 \quad \textbf{the heat equation!}.$$

Remark: Many (nonlinear) diffusion equations, e.g. the porous medium equation, can be written as the $grad_W$ flow of a certain functional on $P_{2,ac}(\mathbb{R}^d)$.

Next: Lecture 3: Entropic regularization of optimal transport.

See you next week!