§2 Optimality Conditions*

Jiayao Zhang†

November 1, 2019

In the last meeting, we went through Monge's and Kantorovich's formulation of the optimal transport problem, the dual formulation of Kantorovich's relaxation, the Wasserstein W_p distance, and showed that Kantorovich's problem is *always* solvable. It remains unclear at this moment, how can we find them, if they always exist?

Today, we will take the first step by focusing on the structure of the optimal transport plans, namely, what are the necessary and sufficient conditions for the optimal transport plan? We will first introduce tools from convex analysis, using which we are able to prove **the fundamental theorem of optimal transport** and understand its nice geometric implications. We start by recapping what we did last time.

1 Recap

Recall that given two Polish spaces X, Y, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and some cost $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ that is measurable, **Monge's optimal transport problem**, aims at

$$\begin{aligned} \min_{T} \quad & \mathbb{E}_{Z\sim\mu}[c(Z, T(Z)] = \int_X c(x, T(x))\,\mathrm{d}\mu, \\ \text{subject to} \quad & T_{\#}\mu = \nu. \end{aligned} \tag{1}$$

Any maps T that preserves mass is called a **transport map**. However, Monge's formulation may be ill-posed, **Kantorovich's relaxation** circumvents this issue, where we instead

$$\begin{aligned} \min_{\gamma} \quad & \mathbb{E}_{(x,y)\sim\gamma}[c(x, y)] = \int_{X\times Y} c(x, y)\,\mathrm{d}\gamma(x, y), \\ \text{subject to} \quad & \gamma \in \Pi(\mu, \nu), \end{aligned} \tag{2}$$

where

$$\Pi(\mu, \nu) \coloneqq \left\{\gamma \in \mathcal{P}(X \times Y) : \pi^X_{\#}\gamma = \mu, \pi^Y_{\#}\gamma = \nu\right\} \tag{3}$$

*This note is based on [PC+19] and [AG13], and was presented at Penn optimal transport reading group.

†University of Pennsylvania, jiayaozhang@acm.org.

is the set of all **couplings** between μ and ν, and $\pi^{(\cdot)}$ is the natural projection. Kantorovich's problem also assumes a dual formulation:

$$\begin{aligned} \max \quad & \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu \\ \text{subject to} \quad & \varphi(x) + \psi(y) \leq c(x,y), \quad \forall (x,y) \in X \times Y, \\ & \varphi \in L^1(\mu), \quad \psi \in L^1(\nu). \end{aligned} \tag{4}$$

Furthermore, we have

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int c(x,y) \, \mathrm{d}\gamma(x,y) = \sup_{\varphi,\psi} \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu, \tag{5}$$

under mild assumptions.

2 Some notions from convex analysis

We assume throughout the spaces X and Y are Polish, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and the cost function $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ is measurable. Given the cost function c, we can define several transforms.

Definition 1. *c-Transforms. Let $\varphi : X \to \mathbb{R} \cup \{\pm\infty\}$, $\psi : Y \to \mathbb{R} \cup \{\pm\infty\}$,*

- *their c_+-transforms $\varphi^{c_+} : Y \cup \{-\infty\}$ and $\psi^{c_+} : X \cup \{-\infty\}$ are defined as*

$$\varphi^{c_+}(y) \coloneqq \inf_{x \in X} \{c(x,y) - \varphi(x)\}, \quad \psi^{c_+}(x) \coloneqq \inf_{y \in Y} \{c(x,y) - \psi(y)\}, \tag{6}$$

- *their c_--transforms[1] $\varphi^{c_-} : Y \cup \{+\infty\}$ and $\psi^{c_-} : X \cup \{+\infty\}$ are defined as*

$$\varphi^{c_-}(y) \coloneqq \sup_{x \in X} \{c(x,y) - \varphi(x)\}, \quad \psi^{c_-}(x) \coloneqq \sup_{y \in Y} \{c(x,y) - \psi(y)\}. \tag{7}$$

With this definition, we observe the following proposition.

Proposition 2. *Let $\varphi, \phi : X \to \mathbb{R} \cup \{\pm\infty\}$, c given, then*

- *$\varphi \leq \phi$ implies $\varphi^{c_+} \geq \phi^{c_+}$.*
- *$(\varphi^{c_+})^{c_+} \geq \varphi$.*
- *$((\varphi^{c_+})^{c_+})^{c_+} = \varphi^{c_+}$.*

Proof. The first property is by definition:

$$\varphi^{c_+} = \inf_x c(x,y) - \varphi(x) \geq \inf_x c(x,y) - \phi(x) = \phi^{c_+}. \tag{8}$$

The second property is by noting for any $x \in X$, $\inf_{z \in X} (c(z,y) - \varphi(z)) \leq c(x,y) - \varphi(x)$:

$$\begin{aligned} (\varphi^{c_+})^{c_+}(x) &= \inf_y c(x,y) - \varphi^{c_+}(y) = \inf_{y \in Y} \left\{ c(x,y) - \inf_{z \in X} (c(z,y) - \varphi(z)) \right\} \\ &\geq \inf_{y \in Y} \{c(x,y) - c(x,y) + \varphi(x)\} = \varphi(x). \end{aligned} \tag{9}$$

[1] Also known as the Legendre transform.

The last property is by combining the first two:

$$(\varphi^{c+})^{c+} \geq \varphi \quad \Rightarrow \quad ((\varphi^{c+})^{c+})^{c+} \leq \varphi^{c+}, \qquad (\varphi^{c+})^{c+} \geq \varphi \quad \Rightarrow \quad ((\varphi^{c+})^{c+})^{c+} \geq \varphi^{c+}. \tag{10}$$

□

If we let φ, ψ, and c be the same as the dual formulation of Kantorovich's relaxation, and write $\mathcal{S}$ for the feasible set for φ and ψ, i.e.,

$$\mathcal{S} \coloneqq \{(\varphi, \psi) : \varphi \in L_1(\mu), \psi \in L_1(\nu), \varphi(x) + \psi(y) \leq c(x, y) \quad \forall (x, y) \in X \times Y\}. \tag{11}$$

For any fixed φ, in order for $(\varphi, \psi) \in \mathcal{S}$, it is necessary that $\psi(y) \leq c(x, y) - \varphi(x)$ for *all* x, y, and hence $\varphi^{c+} = \inf_{x \in X} c(x, y) - \varphi(x)$ gives the largest possible ψ such that (φ, ψ) is still feasible. More precisely, we have the following proposition.

Proposition 3. *Let φ, ψ, c be the same as in the dual formulation of Kantorovich's problem, $\mathcal{S}$ the feasible set, we have*

- $(\varphi, \varphi^{c+}), (\psi^{c+}, \psi) \in \mathcal{S}$.
- *Furthermore,*

$$\langle \varphi, \mu \rangle + \langle \psi, \nu \rangle \leq \langle \varphi, \mu \rangle + \langle \varphi^{c+}, \nu \rangle, \quad \langle \varphi, \mu \rangle + \langle \psi, \nu \rangle \leq \langle \psi^{c+}, \mu \rangle + \langle \varphi, \nu \rangle. \tag{12}$$

The proof is left to the reader; a tricky part is to actually show $\varphi^{c+} \in L_1(\nu)$ (resp. $\psi^{c+} \in L_1(\mu)$), which is easier when we assume $c \in L_1(\mu \times \nu)$. Although it may seems promising at the first glance that to solve the dual problem by iterating

$$\langle \varphi, \mu \rangle + \langle \psi, \nu \rangle \leq \langle \varphi, \mu \rangle + \langle \varphi^{c+}, \nu \rangle \leq \langle (\varphi^{c+})^{c+}, \mu \rangle + \langle \varphi^{c+}, \nu \rangle \leq \cdots, \tag{13}$$

the fact that $((\varphi^{c+})^{c+})^{c+} = \varphi^{c+}$ eliminates this possibility (unless there are additional structures of the problem available). Before we dive into the main theorems, we introduce a few more notions.

Definition 4. ***Cyclical and c-cyclical monotonicity.*** *A set $\Gamma \subset X \times Y$ is cyclically monotone if for any $N \in \mathbb{N}$, $\{(x_i, y_i) \in X \times Y\}_{i=1}^N \subset \Gamma$ we have*

$$\sum_{i \in [N]} \langle x_i, y_i \rangle \geq \sum_{i \in [N]} \langle x_i, y_{\sigma(i)} \rangle, \quad \forall \sigma \in S_N, \tag{14}$$

where S_N is the set of permutations on N elements.

Analogously, $\Gamma \subset X \times Y$ is c-cyclical monotone if for any $N \in \mathbb{N}$, $\{(x_i, y_i) \in X \times Y\}_{i=1}^N \subset \Gamma$

$$\sum_{i \in [N]} c(x_i, y_i) \leq \sum_{i \in [N]} c(x_i, y_{\sigma(i)}), \quad \forall \sigma \in S_N. \tag{15}$$

Note the reverse in the direction of the inequality. By definition, cyclical monotone sets are c-cyclical monotone under $c = -\langle \cdot, \cdot \rangle$ up to a constant.

Definition 5. ***c-concavity/convexity.*** *A function $\varphi : X \to \mathbb{R} \cup \{-\infty\}$, is c-concave (resp. c-convex) if there exists $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ such that $\varphi = \psi^{c+}$ (resp. $\varphi = \psi^{c-}$). Similar definitions for any function $\psi : Y \to \mathbb{R} \cup \{-\infty\}$. By symmetry, the $-\infty$ can be replaced by $+\infty$ in above definitions.*

In practise, it suffices to work with one of them, e.g., c-concavity, since φ is c-convex iff $-\varphi$ is c-concave. The name may due to the following claim.

Claim 6. *If c is convex, φ is c-convex iff φ is convex and l.s.c.*

The need for being l.s.c. in Claims 6 is from the fact that we allow functions to assume values from the extended real line. The forward implication is proved by definition; and the other direction can be argued from contrapositive.

Definition 7. ***c-superdifferential/subdifferential.*** *Let $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ be c-concave. The c-superdifferential $\partial^{c+}\varphi \subset X \times Y$ and c-subdifferential $\partial^{c-}\varphi \subset X \times Y$ are defined as*

$$\begin{aligned} \partial^{c+}\varphi &\coloneqq \{(x,y) \in X \times Y : \varphi(x) + \varphi^{c+}(y) = c(x,y)\}, \\ \partial^{c-}\varphi &\coloneqq \{(x,y) \in X \times Y : \varphi(x) + \varphi^{c-}(y) = c(x,y)\}. \end{aligned} \tag{16}$$

We will mainly work with ∂^{c+}. Given y, write $\partial^{c+}\varphi(y)$ be the set of $x \in X$ such that $(x,y) \in \partial^{c+}\varphi$. Just like many other things with a name prefixed by "super", c-superdifferentials may be equivalently defined by an inequality,

$$(x,y) \in \partial^{c+}\varphi \quad \Leftrightarrow \quad \varphi(x) - c(x,y) \geq c(x',y) - \varphi(x'), \quad \forall x' \in X, \tag{17}$$

and you can now guess the equivalent alternative definition for subdifferentials.

Indeed, $\partial^{c+}\varphi$ and $\partial^{c-}\varphi$ are in some sense the **complementary slackness** condition for the dual problem. In the next section we will see how to characterize optimally using these convex anlaysis notions, and the conditions for strong duality are also sufficient and necessary for this formulation, as we may reasonably anticipate.

3 Fundamental theorem of optimal transport

We are now ready for the fundamental theorems that characterize the sufficient and necessary condition for optimal transport plans. We start from an example, consider the case where $X = Y = \mathbb{R}$, μ, ν supported on finite subsets, and with $c(x,y) = |x-y|^2$, then the optimal plan γ^* must be such that

$$\pi^X_\# \gamma^* = \mu, \quad \pi^Y_\# \gamma^* = \nu, \tag{18}$$

which translates to the following equivalent condition points x_i, y_i:

$$\sum_{i \in [N]} |x_i - y_i|^2 \leq \sum_{i \in [N]} |x_i - y_{\sigma(i)}|^2, \tag{19}$$

for any point (x_i, y_i) and any valid N that γ^* put non-zero mass on. Noting in this example, $\mathrm{supp}(\gamma^*)$ is c-cyclical monotone, and by openning up squares, $\mathrm{supp}(\gamma)$ is indeed cyclical monotone in the ordinary sense, which is indeed a general phenomena.

Theorem 8. ***Necessary condition for optimal transport plans.*** *If c is continuous and bounded from below, then the support of optimal transport plan γ is necessarily c-cyclical monotone.*

In fact, this is almost sufficient; and we can relax continuity to l.s.c. But for the time being, let's see why this is the case. The key here lies in the smoothness of c and the definition of c-cyclical monotone.

Proof. [2] First note that from last meeting, we know c being l.s.c. implies the Kantorovich's problem is solvable. Let γ be an optimal transport plan whose domain is not c-cyclical monotone, then for some $N \in \mathbb{N}$, $\{(x_i, y_i)\}^N$, there is some permutation $\sigma \in \mathcal{S}_N$ such that

$$\sum_{i\in[N]} c(x_i, y_i) - \sum_{i\in[N]} c(x_i, y_{\sigma(i)}) = \Delta > 0. \tag{20}$$

Now consider a small δ-ball $B_i \coloneqq B_\delta(x_i, y_i)$ around (x_i, y_i). The continuity of c implies that if we make δ small enough, for any $(x, y) \in B_i$ we have

$$c(x, y) \geq c(x_i, y_i) - \epsilon, \tag{21}$$

and any $(x, y) \in B_i' \coloneqq B_\delta(x_i, y_{\sigma(i)})$,

$$c(x, y) \leq c(x_i, y_{\sigma(i)}) + \epsilon. \tag{22}$$

Now we will construct another plan $\tilde{\gamma}$ that is a coupling but with a strictly lower cost. First since $(x_i, y_i) \in \operatorname{supp}(\gamma)$, we have $\gamma(B_i), \gamma(B_i') > 0$ for all i. We can thus define conditional measures $\gamma_i = \gamma | B_i$ where

$$\gamma_i(E) = \frac{\gamma(E \cap B_i)}{\gamma(B_i)}, \quad \forall E \in \mathcal{B}(X \times Y). \tag{23}$$

Write $\mu_i = \pi^X_\# \gamma_i$, $\nu_i = \pi^Y_\# \gamma_i$, and let $\tilde{\gamma}_i = \mu_i \otimes {\nu_{\sigma(i)}}$[3]. Consider

$$\tilde{\gamma} \coloneqq \gamma - \alpha \sum_{i\in[N]} \gamma_i + \alpha \sum_{i\in[N]} \tilde{\gamma}_i, \tag{24}$$

where $\alpha > 0$ is chosen such that $\tilde{\gamma}$ is positive. We can easily check that $\pi^X_\# \tilde{\gamma} = \mu$, $\pi^Y_\# \tilde{\gamma} = \nu$, and

$$\int c \,\mathrm{d}\gamma - \int c \,\mathrm{d}\tilde{\gamma} = \alpha \sum_{i\in[N]} \left(\int_{X\times Y} c \,\mathrm{d}\gamma_i - \int_{X\times Y} c \,\mathrm{d}\tilde{\gamma}_i \right). \tag{25}$$

By construction, noting γ_i (resp. γ_i') concentrates on B_i (resp. B_i'), we have

$$\int_{X\times Y} c \,\mathrm{d}\gamma_i \geq \frac{1}{\gamma(B_i)} \int_{B_i} \left(c(x_i, y_i) - \epsilon\right) \,\mathrm{d}\gamma = c(x_i, y_i) - \epsilon, \tag{26}$$

$$\int_{X\times Y} c \,\mathrm{d}\tilde{\gamma}_i = \int_{X\times Y} c \,\mathrm{d}\mu_i \,\mathrm{d}\nu_{\sigma(i)} \leq \frac{1}{\gamma(B_i')} \int_{B_i'} \left(c(x_i, y_{\sigma(i)}) + \epsilon\right) \,\mathrm{d}\gamma = c(x_i, y_{\sigma(i)}) + \epsilon. \tag{27}$$

Hence

$$\int c \,\mathrm{d}\gamma - \int c \,\mathrm{d}\tilde{\gamma} \geq \alpha \left(\sum_{i\in[N]} c(x_i, y_i) - \sum_{i\in[N]} c(x_i, y_{\sigma(i)} - 2N\epsilon \right), \tag{28}$$

and if we set $\epsilon \leq \Delta/(2N)$, $\tilde{\gamma}$ is strictly better than γ, a contradiction. □

[2] This proof is due to [San19].

[3] In fact any element in $\Pi(\mu_i, \nu_{\sigma(i)})$ works.

Theorem 9. ***Fundamental theorem of optimal transport.***

If c is continuous and bounded from below and for some $f \in L_1(\mu)$, $g \in L_1(\nu)$ we have for all $(x, y) \in X \times Y$,

$$c(x, y) \leq f(x) + g(y), \tag{29}$$

then TFAE:

(a) $\gamma \in \Gamma(\mu, \nu)$ is optimal.

(b) $\operatorname{supp}(\gamma)$ is c-cyclically monotone.

(c) There exists a c-concave function φ such that $\varphi^+ \in L_1(\mu)$ and $\operatorname{supp}(\gamma) \subset \partial^{c+}\varphi$.

Note that under the condition of this theorem, the strong duality holds, a result stated by not proved last time, as we expected earlier.

Proof.

- **(a) $\Rightarrow$ (b).** *Proven.*
- **(b) $\Rightarrow$ (c).** *Omitted.* It suffices to show that for any c-cyclically monotone $\Gamma \subset X \times Y$, there is some c-concave function φ such that $\Gamma \subset \partial^{c+}\varphi$, and $\varphi^+ \in L_1(\mu)$. This is a result from convex analysis.
- **(c) $\Rightarrow$ (a).** Let φ be given, since $\operatorname{supp}(\gamma) \subset \partial^{c+}\varphi$, $\varphi(x) + \varphi^{c+}(y) = c(x, y)$ for all $(x, y) \in \operatorname{supp}(\gamma)$. But $\varphi(x) + \varphi^{c+}(y) \leq c(x, y)$ for all $(x, y) \in X \times Y$, hence for any $\gamma' \in \Pi(\mu, \nu)$,

$$\begin{aligned} \int c \,\mathrm{d}\gamma &= \int_{X\times Y} \varphi(x) + \varphi^{c+}(y) \,\mathrm{d}\gamma(x, y) = \int \varphi \,\mathrm{d}\mu + \int \varphi^{c+} \,\mathrm{d}\nu \\ &= \int_{X\times Y} \varphi(x) + \varphi^{c+}(y) \,\mathrm{d}\gamma'(x, y) \leq \int c \,\mathrm{d}\gamma'. \end{aligned} \tag{30}$$

□

4 Consequences

4.1 Direct consequences of the fundamental theorem

1. The condition on the existence of $f \in L_1(\mu)$, $g \in L_1(\nu)$ is only needed for proving (b) $\Rightarrow$ (c). In other words, even c does not satisfy this condition, (b) will still be necessary and (c) sufficient for γ being optimal.
2. The sufficient condition (c) implies that being optimal is a property of the support being cyclically monotone wrt the given cost. This is by itself free of μ, ν. Hence if γ is optimal, for any measure γ' whose support is contained in $\operatorname{supp}(\gamma)$, γ' is optimal for transporting $\pi^X_\# \gamma'$ to $\pi^Y_\# \gamma'$ under the same cost.
3. If Monge map T exists, the support of $(\mathrm{id} \times T)_\# \mu$ is c-cyclically monotone. Hence if this support is the whole space X, then T is *the* optimal for transporting any $\mu \in \mathcal{P}(X)$ to $\nu = T_\# \mu$. Indeed, this gives us a **charaterization of Monge map** - its image must be in the c-superdifferential of some c-concave function.

4. A similar but stronger result holds for optimal plans: if $\text{supp}(\gamma) \subset \partial^{c+}\varphi$ for some optimal transport plan γ, then all optimal plan γ' must also be such that $\text{supp}(\gamma') \subset \partial^{c+}\varphi$.

4.2 A concrete example

Let $X = Y = \mathbb{R}^d$, and c the L_2 distance, we know if μ is absolutely continuous wrt the Lebesgue measure λ, the optimal plan γ is supported on the graph of a Monge map T, i.e.,

$$\gamma = (\text{id}, T)_{\#}\mu. \tag{31}$$

The necessary condition, formulated in terms of c-convex functions, implies there is some c-convex function φ such that

$$T(x) \subset \partial^{c-}\varphi, \tag{32}$$

but φ is *convex* since c is convex, and $\partial^{c-}\varphi$ is indeed the ordinary subgradient. At the point x of continuity of φ, we have

$$T(x) = \nabla\varphi(x), \tag{33}$$

and this gives us a characterization of Monge maps in this setting - it is the gradient of some convex function (λ-a.e.). This is a special case of the Brenier's theorem (as in the COT textbook) we discussed last time.

References

[AG13] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[San19] Filippo Santambrogio. c-cyclical monotonicity of the support of optimal transport plans. `https://www.math.u-psud.fr/~filippo/SupportcCM.pdf`, 2019. [Online; accessed 01-Nov-2019].