Kantorovich Duality and the Wasserstein Space

Zachary McNulty

April 2023

Throughout the following assume the standard conditions: the cost $c(x, y)$ is lower semicontinuous and bounded below by the sum $a(x) + b(y)$ of two upper semicontinuous functions $a \in L^1(\mu), b \in L^1(\nu)$, and $\mathcal{X}, \mathcal{Y}$ are Polish spaces.

1 Kantorovich Duality

Instead of trying to view optimal transport as "finding the minimal cost transport plan" we can rephrase the problem as "trying to determine the value/price of each piece of mass at a given location $x \in \mathcal{X}$ or $y \in \mathcal{Y}$". Namely, if a given x is "close" in terms of $c(x, \cdot)$ to many different y we might suppose any mass at x is more valuable because it is easy to transport. Similarly, if a given y is "far" from most values of x then we might consider mass there more valuable because it is hard to get to. Hence introduce two *price functions* $\psi : \mathcal{X} \to \mathbb{R}, \phi : \mathcal{Y} \to \mathbb{R}$ which keep track of these prices/values. Call a pair of prices *competitive* if:

$$\phi(y) - \psi(x) \leq c(x, y)$$

Namely, our cost function already gives us some notion of the relationship between these prices because we always transport mass at x to y at cost $c(x, y)$. The goal then is to maximize the profit $\phi(y) - \psi(x)$ according to this constraint, giving us the Kantorovich dual problem:

$$\sup_{\phi,\psi} \left\{ \int \phi(y)\, d\nu - \int \psi(x)\, d\mu \quad : \quad \phi - \psi \leq c \right\} \tag{1}$$

Integrating the above inequality with respect to any transport plan $\pi \in \Pi(\mu, \nu)$ gives weak duality.

Theorem 1 (Weak Duality)

$$\sup_{\phi,\psi} \left\{ \int \phi(y)\, d\nu - \int \psi(x)\, d\mu \quad : \quad \phi - \psi \leq c \right\} \leq \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi c(X, Y)$$

Some of the main benefits of the dual formulation of the problem are:

1. **Lowerbounds for optimal cost**: any pair of competitive prices provides a lower bound.

2. **Functional Analysis:** Gives access to functional analysis tools (we get to work over a space of functions rather than a space of measures)

Once we have such duality it is natural to ask (i) what the optimizing pairs (ϕ, ψ) and optimizing plans $\pi \in \Pi(\mu, \nu)$ tend to look like, and (ii) when we have equality, aka strong duality. First we focus on the former.

We start with a simple observation that it is always in our best interests to make ϕ as large as possible and ψ as small as possible. In particular, as we are forced to have $\phi \leq \psi + c$ and $\psi \geq \phi - c$ we might as well choose these inequalities to be as tight as possible. Namely given a competitive pair (ϕ, ψ) we can always improve it by replacing these functions with their c-transforms (note the different defintions between $\mathcal{X}$ and $\mathcal{Y}$):

$$\phi^c(x) := \sup_{y \in \mathcal{Y}} [\phi(y) - c(x, y)] \tag{2}$$

$$\psi^c(y) := \inf_{x \in \mathcal{X}} [\psi(x) + c(x, y)] \tag{3}$$

Any function of the form ϕ^c for some ϕ is called c-convex. Likewise such a function ψ^c is called c-concave.

Example: In the case $c(x, y) = -xy$ we obtain the standard relationship between convexity and the (negative of) the Legendre transform: a function is convex iff it is the Legendre transform of some function.

Example: In the case $c(x, y) = d(x, y)$ is the underlying metric, then the c-convex functions are just the 1-Lipschitz functions. Moreover, 1-Lipschitz functions are self-conjugate

Now to study an optimal transport plan. The intuition is that if a plan π is optimal, there ought to be no way to "reroute" mass while lowering the overall cost. Namely, there should not be a sequence of pairs $\{(x_i, y_i)\}_{i=1}^N$ in the support of π so that:

$$\sum_{i=1}^N c(x_i, y_i) > \sum_{i=1}^N c(x_i, y_{i+1})$$

This would mean we could take some mass at x_i that was originally going to y_i and instead send it to y_{i+1}. By doing this in a cyclic manner, we preserve the total amount of mass each y_i receives just by modifying the sources. This motivates the definition of a c-cyclically monotone set: a set $\Gamma \subseteq \mathcal{X} \times \mathcal{Y}$ so:

$$\sum_{i=1}^N c(x_i, y_i) \leq \sum_{i=1}^N c(x_i, y_{i+1}) \tag{4}$$

and we call a transport plan c-cyclically monotone if it is concentrated on such a set. Note since cyclces generate all permutations, this is equivalent to the saying now permutation of the sources can lower the overall cost. Now we are ready to state the main result:

Theorem 2 (Kantorovich Duality) *Under our standard assumptions:*

1. *We have strong duality*

2. *We can restrict the dual problem to bounded continuous c-convex/concave functions and their c-transforms (e.g. (ϕ, ϕ^c) or (ψ^c, ψ)).*

3. *π is optimal iff it is c-cylically monotone iff $\phi(y) - \psi(x) \leq c(x, y)$ with equality π a.s.*

Proof Outline: The idea is to construct a c-cyclically monotone π (which will turn out to be optimal) and use this to construct $\psi(x)$. Then show that:

$$\int \psi^c(y)\, d\nu - \int \psi(x) = \int c(x, y)\, d\pi(x, y)$$

By weak duality, this implies both $\pi, (\psi^c, \psi)$ are optimal once we show this choice of prices are competitive: $\psi^c(y) + \psi(x) \leq c(x, y)$. We procede in a few steps by a series of approximations:

- **Step 1**: For c continuous μ, ν discrete, construct c-cyclically monotone π . This is simple as then transport plans are essentially just a finite stochastic matrix.

- **Step 2:** For c-continuous, construct a c-cyclically monotone π for general μ, ν by approximating these distributions by discrete (random) measures and applying step 1. Then take limits.

- **Step 3**: Given such a π from step 2, let $\Gamma \subseteq \mathcal{X} \times \mathcal{Y}$ be its support. Define a candidate "price function" for fixed $(x_0, y_0) \in \Gamma$:

  $$\phi(x) := \sup_m \sup_{(x_i, y_i) \in \Gamma} \left\{ \sum_{i=0}^{m} c(x_i, y_i) - c(x_{i+1}, y_i) \right\}$$

  where $x_{m+1} = x$. Namely, the "value" $\phi(x)$ of having an additional unit of mass at x is the largest improvement in cost we could obtain by rerouting some of our transports to utilize this new mass. Show that $\psi^c(y) - \psi(x) = c(x, y)$ on Γ.

  Step 4: approximate a general cost function by an increasing sequence of continous (bounded) ones using lower-semicontinuity.

□

1.1 Applications of Duality

The next result shows that under mild assumptions optimal transport plans are stable under limits

Theorem 3 (Stability) *Suppose c is continuous and bounded from below. If $\mu_k \Rightarrow \mu, \nu_k \Rightarrow \nu$ and $C(\mu_k, \nu_k) < \infty$. Then there is some subsequence along which the optimal transport plans $\pi_k \Rightarrow \pi$ the optimal transport plan for μ, ν.*

Proof Outline: The weak convergence of μ_k, ν_k implies tightness of the corresponding sequences and hence tightness of the optimal plans π_k. By Prokorohov's theorem we can hence extra a convergent subsequence $\pi_k \Rightarrow \pi$. It is then just a matter of showing the c-cyclical monotonicity of the π_k is preserved, which follows from the weak convergence and continuity of c. □

Our next aim is to use duality to study transport inequalities, inequalities of the form:

$$C(\mu,\nu) \leq F(\mu) \quad \forall \mu \in \mathcal{P}(x)$$

The following says for nice F we can represent these inequalities in terms of a Legendre transform Λ over $C_b(\mathcal{X})$. Specifically:

$$\Lambda(\phi) := \sup_{\mu \in P(\mathcal{X})} \left(\int_{\mathcal{X}} \phi \, d\mu - F(\mu) \right)$$

Theorem 4 (Dual Transport Inequalities) *Suppose the cost function is bounded from below. Let $F : P(\mathcal{X}) \to \mathbb{R}$ be convex and lower semicontinuous. Then:*

$$C(\mu,\nu) \leq F(\mu) \quad \forall \mu \in P(\mathcal{X}) \iff \Lambda\left(\int_{\mathcal{Y}} \phi \, d\nu - \phi^c \right) \leq 0 \quad \forall \phi \in C_b(\mathcal{X})$$

Proof outline: Unrolling definitions note:

$$\Lambda\left(\int_{\mathcal{Y}} \phi \, d\nu - \phi^c \right) = \sup_{\mu \in P(\mathcal{X})} \left(\int_{\mathcal{Y}} \phi \, d\nu - \int_{\mathcal{X}} \phi^c \, d\mu - F(\mu) \right)$$

The first two terms are straight out of the Kantorovich dual problem, so the result will come immediately just by applying duality basically. □

Example (KL Divergence): Let $\mathcal{X} = \mathcal{Y}$ and $F(\mu) := H(\mu|\nu) = \int \rho \log \rho$ for $\rho = \frac{d\mu}{d\nu}$. Then an explicit calculation yields $\Lambda(\phi) = \log\left(\int e^{\phi} \, d\nu\right)$. Hence:

$$C(\mu,\nu) \leq H(\mu|\nu) \quad \forall \mu \in P(\mathcal{X}) \iff e^{\int_{\mathcal{Y}} \phi \, d\nu} \leq \int e^{\phi^c} \, d\nu$$

Lastly, we develop a criterion for an existence of a deterministic coupling:

Theorem 5 (Existence of Deterministic Coupling) *If $C(\mu,\nu) < \infty$ and any c-convex $\psi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ has:*

$$\mu\left(\{x \in \mathcal{X} \ : \ \psi^c(y) - \psi(x) = c(x,y) \textit{ for more than one } y\}\right) = 0$$

Then there is a unique solution to the optimal transport problem and it is deterministic.

Proof Outline: This follows from the fact that at optimality π a.s. we have $\psi^c(y) - \psi(x) = c(x,y)$. Hence for μ a.e. x we can set $T(x)$ to be the unique y mentioned in the assumption. □

Example: Again in the case $c(x,y) = -xy$ where the c-convex functions are just standard convex functions we have:

$$\psi(x) = \sup_y (\psi^c(y^*) + xy) \to \nabla \psi^c(y) = -x$$

Since ψ^c is concave, $\nabla \psi^c(y)$ is nonincreasing. Hence it only has countably many "flat" points in its range where $\nabla \psi^c(y) = -x$ can occur. Thus the only way for the set in the statement of Theorem 5 to have nonzero measure is if μ has an atom. Hence as long as μ is nonatomic, a deterministic coupling exists (e.g. we have seen an example of this before in the increasing rearrangement, example 3 on pg 19).

2 Wasserstein Distances

When $\mathcal{X} = \mathcal{Y}$ can think of the optimal transport cost $C(\mu,\nu)$ as a sort of "distance" between the probability distributions μ, ν. This is not a true metric in general (it does not necessarily satisfy any of the properties in fact) but if we choose an underlying metric to use as our cost function $c(x,y)$ then we do indeed get a metric, and our hope is that in some sense this metric metrizes the topology of weak convergence. This is indeed the case. For $p \geq 1$ define the Wasserstein distance of order p:

$$W_p(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left[\mathbb{E} d(X,Y)^p\right]^{1/p} \tag{5}$$

Restricted to the Wasserstein space, denoted $P_p(\mathcal{X})$, where $W_p(\delta_0,\mu) < \infty$ then W_p is indeed a metric. Namely, this is just the space of measures have "finite moment of order p" within the metric space. If (X,d) is bounded then this is just all of $P(\mathcal{X})$.

Since W_1 is just an optimal transport cost under a specific choice of cost $c(x,y) = d(x,y)$ we can apply Kantorovich duality. We already saw the c-convex functions are just 1-Lipschitz and such functions are self-conjugate. Hence:

$$W_1(\mu,\nu) = \sup_{1-Lip} \left\{ \int_{\mathcal{X}} \psi \, d\mu - \int_{\mathcal{X}} \psi \, d\nu \right\}$$

2.1 Metrizes Weak Convergence in $P_p(\mathcal{X})$

We introduce a slightly stronger notion than weak convergence. We say μ_k converges weakly in $P_p(\mathcal{X})$ to μ if $\mu_k \to \mu$ AND $\int d(x_0,x)^p \, d\mu_k \to \int d(x_0,x)^p \, d\mu$. Namely, on top of the standard weak convergence we also assume convergence of the p^{th} moments. This is stronger than standard weak convergence because in general d^p is not a bounded function. If $\mathcal{X}$ is of bounded diameter then these two notions are equivalent.

Theorem 6 *W_p metrizes weak convergence in $P_p(\mathcal{X})$*

Proof Outline: Assume $\mu_k \to \mu$ in W_p. By a technical lemma, this implies $\{\mu_k\}$ is tight and hence by Prokohorov's theorem implies it has a weakly convergent subsequence. Use lower semicontinuity of W_p to show whole sequence converges. Use the arithmetic lemma:

$$d(x_0,x)^p \leq (1+\epsilon) d(x_0,y)^p + C_\epsilon d(x,y)^p$$

to separate x,y. Apply an optimal transfer plan $\pi_k \in \Pi(\mu_k,\mu)$ and take its marginals.

Conversely assume $\mu_k \to \mu$ weakly in $P_p(\mathcal{X})$. Again take an optimal $\pi_k \in \Pi(\mu_k,\mu)$ and use tightness / Prokorhov's theorem to extract convergent subsequence. Stability of optimal transport plans implies the limit π is optimal as well, and the optimal coupling of (μ,μ) is trivially just the identity. Apply a truncation:

$$W_p(\mu_k,\mu)^p = \int_{\mathcal{X}} d(x,y)^p \, d\pi_k = \int_{\mathcal{X}} \left[d(x,y)^p \wedge R + (d(x,y)^p - R^p)_+\right] d\pi_k$$

The first term is bounded/cts so just apply standard weak convergence. Use convergence in $P_p(\mathcal{X})$ (the "finite p^{th} moments condition) to handle the second term. □

2.2 Controlled by Weighted TV distance

Note this text defines:

$$||\mu - \nu||_{TV} = 2 \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{P}_\pi(X \neq Y) \tag{6}$$

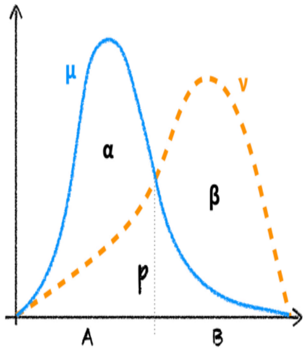
which is twice the definition I am used to.

Next we aim to show if we weight the TV distance we get an upperbound of W_p.

Theorem 7 **(W_p controlled by weighted TV)** *Suppose $p \geq 1$ and $p^{-1} + q^{-1} = 1$. Fix $x_0 \in \mathcal{X}$. Then:*

$$W_p(\mu,\nu) \leq 2^{1/q} \left(\int d(x_0,x)^p \, d|\mu - \nu|(x) \right)^{1/p}$$

Proof outline: The idea is to use the same coupling π of μ, ν that minimizes 6. The idea is to keep all the mass that μ, ν share fixed and distribute the rest uniformly. Then a simple calculation yields the result. □.



2.3 Topological Properties of Wasserstein Space

The next theorem shows the Wasserstein space shares many of the properties of the underlying metric space $\mathcal{X}$.

Theorem 8 *If $(\mathcal{X}, d)$ is Polish then so is $(P_p(\mathcal{X}), W_p)$. Moreover, any measure can be approximated by ones with finite support (discrete measures).*

Proof Idea: For separability let x_i be a countable dense subset of $\mathcal{X}$. Then:

$$\left\{ \sum_{i=1}^{n} a_i \delta_{x_i} \; : a_i \in \mathbb{Q}, \; n \in \mathbb{N} \right\} \subseteq P_p(\mathcal{X})$$

is countable and dense.

For completeness, show that Cauchy in W_p implies tightness, and hence by Prokorhov's theorem there exists a weakly convergent subsequence. This gives a candidate for the limit. Fatou's lemma shows this is in $P_p(\mathcal{X})$ so we just must show weak convergence in $P_p(\mathcal{X})$. We just use the lower semicontinuity of W_p to show $W_p(\mu_k, \mu) \to 0$. $\square$

3 Displacement Interpolation