Statistical OT Lecture 6: More Duality

1 Uniqueness

Using the fundamental theorem from last lecture, it is quite simple to argue that the OT coupling must be (essentially) unique.

Theorem 1. Suppose that μ, ν are measures with two bounded moments such that, μ has a density, and $X \sim \mu$. Then there exists a convex function φ_0 such that $(X, \nabla\varphi_0(X))$ is an optimal coupling. Furthermore, $\nabla\varphi_0$ is unique, i.e. if $(X, \nabla\psi(X))$ is a valid coupling with ψ convex, then $\nabla\varphi_0 = \nabla\psi$, μ almost surely.

Proof. By Brenier's theorem we already know that there exists a convex function φ_0 such that $(X, \nabla\varphi_0)$ is an optimal coupling. By the fundamental theorem, we know that $(X, \nabla\psi(X))$ is also an OT coupling. Now let $T_1 = \nabla\varphi_0$, $\gamma_1 = (X, \nabla\varphi_0(X))$, $T_2 = \nabla\psi$ and $\gamma_2 = (X, \nabla\psi(X))$. Since both γ_1, γ_2 are optimal, we know that the coupling $\gamma = (\gamma_1 + \gamma_2)/2$ is also optimal:

$$\int \|x - y\|^2 d\gamma = \frac{1}{2}\int \|x - y\|^2 d\gamma_1 + \frac{1}{2}\int \|x - y\|^2 d\gamma_2.$$

By Brenier's theorem, we then conclude that $\gamma = (X, \nabla\phi(X))$ for some convex function ϕ. But now, we see that the conditional distribution of $Y|X$ for γ should take value $T_1(X)$ with probability $1/2$ and $T_2(X)$ with probability $1/2$, but since γ is also realized by an OT map this is impossible unless $T_1 = T_2$ almost surely. □

2 Implications

Under the conditions for the uniqueness theorem, we know that if a valid transport map is the gradient of a convex function, then it must be the unique OT map.

1D setting: In the 1D setting, when μ has a density, this gives an alternate direct proof of the optimality of the map $T_0(X) = F_\nu^\dagger(F_\mu(X))$. We observe that this map is a valid transport map, and is an increasing function. It is cyclically monotone, and so induced by the gradient of a convex function, and hence is the unique OT map.

Gaussians: Suppose that $\mu = N(m_1, \Sigma_1)$ and $\nu = N(m_2, \Sigma_2)$. A natural transport map to consider is to take:

$$T(x) = \Sigma_2^{1/2}\Sigma_1^{-1/2}(x - m_1) + m_2.$$

However, this is not the gradient of a convex function, so cannot be the OT map. If an affine map is the gradient of a convex function, we'd want the scaling matrix to be PSD. One can verify that taking,

$$T(x) = \ \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\ \Sigma_1^{1/2})^{1/2}\ \Sigma_1^{-1/2}(x - m_1) + m_2,$$

leads to a valid map which is the gradient of a convex function. Appealing to our uniqueness theorem, we then conclude that this is in fact the (unique) OT map.

3 The Semi Dual

From the dual OT program one can learn many nice structural properties. Recall the Kantorovich dual,

$$\sup_{\substack{f\in L^1(\mu), g\in L^1(\nu),\\ f(x)+g(y)\le \|x-y\|^2}} \left[\int f d\mu + \int g d\nu\right].$$

Suppose we held f fixed, and computed for this fixed f what the best possible g would be. We would see that the choice:

$$g(y) = \inf_x \{\|x-y\|^2 - f(x)\},$$

would be the best function satisfying all the inequality constraints. Similar reasoning as before will show that this is an $L^1(\nu)$ function. The function g above is called the c-transform of f and is denoted f^c. Then our reasoning shows that we could rewrite the dual as:

$$\sup_{f\in L^1(\mu)} \left[\int f d\mu + \int f^c d\nu\right].$$

This reasoning (of replacing one of the two Kantorovich potentials by the c-transform of the other) works for any cost c, under appropriate moment assumptions. For the quadratic cost, one can go even further and see that up to a simple transformation, we could equivalently solve the so-called semi-dual program:

$$\inf_{\phi\in L^1(\mu)} \left[\int \phi d\mu + \int \phi^* d\nu\right],$$

where ϕ^* is the Fenchel conjugate of ϕ. The relationship between the semi-dual and the Kantorovich dual is summarized in the following theorem:

Theorem 2. Let μ, ν be two measures with finite second moments. Then:

1. A pair of functions (f_0, g_0) is optimal for the Kantorovich dual, if and only if, $f_0(x) = \|x\|^2 - 2\varphi_0(x)$ and $g_0(y) = \|y\|^2 - 2\varphi_0^*(y)$, where φ_0 is an optimizer of the semi-dual.
2. The optimal values of the two programs are related as:

$$\text{dual-opt} = \int \|x\|^2 d\mu + \int \|y\|^2 d\nu - 2\text{semi-dual-opt}.$$

Proof. To begin with, let us re-write the Kantorovich dual with the reparametrization of $f(x) = \|x\|^2 - 2\phi(x)$, and $g(y) = \|y\|^2 - 2\psi(y)$. Then the objective:

$$\int f d\mu + \int g d\nu = \int \|x\|^2 d\mu + \int \|y\|^2 d\nu - 2\left(\int \phi d\mu + \int \psi d\nu\right).$$

Similarly, the constraint that $f(x) + g(y) \le \|x-y\|^2$ translates to:

$$\phi(x) + \psi(y) \ge x^T y,$$

and we obtain the equivalent (up to reparametrization) program:

$$\inf_{\substack{\phi\in L^1(\mu), \psi\in L^1(\nu)\\ \phi(x)+\psi(y)\ge x^T y}} \left[\int \phi d\mu + \int \psi d\nu\right].$$

Now, we can once again apply the reasoning we used to introduce the c-transform of holding ϕ fixed, and computing the optimal ψ, and this reasoning leads us to the semi-dual program. □

We can also go further and extract some more structural properties of the optimizer of the semi-dual. In particular, iterating the c-transform argument will show that we can start with any function ϕ, conclude that we can replace it with ϕ^{**} (which is a closed convex function). This argument thus shows that when solving the semi-dual we can restrict our attention to closed convex functions.

4 Duality when $p = 1$

The c-transform idea also gives us important structural insights into the optimal transport program when $p = 1$. To begin with we need a strong duality result for general ℓ_p costs which we state without proof:

Theorem 3. Fix $p \geq 1$ and let μ, ν be two measures with finite p-th moment. Then:

$$W_p^p(\mu,\nu) = \inf_{\gamma \in \Gamma_{\mu,\nu}} \int \|x-y\|^p d\gamma = \sup_{\substack{f \in L^1(\mu), g \in L^1(\nu), \\ f(x)+g(y) \leq \|x-y\|^p}} \left[\int f d\mu + \int g d\nu \right].$$

We now need two structural properties of the c-transform when $c(x,y) = \|x-y\|$.

1. The c-transform of any function is 1-Lipschitz.
2. The c-transform of any 1-Lipschitz function f is $-f$..

Proof. **Claim 1:** Define,

$$g(y) = \inf_x \{\|x-y\| - f(x)\},$$

and we would like to show that this is 1-Lipschitz. First, notice that for a fixed x_0, the function $\|x_0 - y\| - f(x_0)$ is clearly 1-Lipschitz. It is also easy to verify that the infimum of a collection of 1-Lipschitz functions is 1-Lipschitz, so we conclude that g is 1-Lipschitz.

Claim 2: Let f be a 1-Lipschitz function. Then we know that $-f$ is also 1-Lipschitz and we have:

$$-f(y) \leq \|x-y\| - f(x).$$

Taking the infimum over x we obtain that, $-f \leq f^c$. On the other hand,

$$f^c(y) = \inf_x \{\|x-y\| - f(x)\} \leq -f(y),$$

by taking $x = y$. So we conclude that, $f^c = -f$.

□

Now, we are ready to prove the following theorem:

Theorem 4. Let μ, ν be measures with bounded 1-st absolute moment. Then:

$$W_p(\mu,\nu) = \sup_{f, 1-\text{Lipschitz}} \int f d\mu - f d\nu.$$

Proof. By strong duality we know that,

$$W_1(\mu,\nu) = \sup_{\substack{f \in L^1(\mu), g \in L^1(\nu), \\ f(x)+g(y) \leq \|x-y\|}} \left[\int f d\mu + \int g d\nu \right].$$

For a fixed function f, we may replace g by f^c, and then replace f by f^{cc}, and finally negate both functions, and use the structural properties above to obtain that:

$$W_1(\mu,\nu) = \sup_{f\in L^1(\mu)} \left[\int f^{cc} d\mu + \int f^c d\nu\right] = \sup_{f\in L^1(\mu), 1-\text{Lipschitz}} \left[\int f d\mu - \int f d\nu\right].$$

Now, any 1-Lipschitz function is integrable since for any y:

$$\int |f| d\mu \leq |f(y)| + \int |f(x) - f(y)| d\mu \leq |f(y)| + \int \|x-y\| d\mu < \infty,$$

and we obtain the theorem as a consequence. □

5 Estimation of and under the Wasserstein distance

We will begin our journey by considering the following question: suppose we are given $X_1, \ldots, X_n \sim \mu$, where μ is supported on $[0,1]^d$. How well can we estimate μ in the Wasserstein sense. This is a question of density estimation where our loss is measured using the Wasserstein distance.

A first question to ponder is – why is this task even feasible/sensible? We have made no smoothness-type assumptions, and it certainly not the case that we can estimate μ in more classical distances (TV, Hellinger, χ^2 etc.) without these assumptions.

A second question to answer is – what are sensible estimators of the distribution? Again, since we have made no smoothness assumptions our favorite non-parametric estimators (kernel/wavelet-type estimators) should seem like bad ideas. However, we're still solving a non-parametric density estimation task so what other options can we consider?

Recall, that we've made this remark earlier – Wasserstein distances behave quite differently from classical distances. In particular, we have noted that the Wasserstein distance is well-defined between a discrete and continuous distribution. Perhaps, it makes sense to simply use the empirical measure:

$$\widehat{\mu} := \mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i},$$

as an estimate of the underlying distribution μ.

We will give a couple of different proofs of the following result:

Theorem 5. Suppose that μ is supported on $[0,1]^d$ then:

$$\mathbb{E}W_1(\mu_n,\mu) \lesssim \sqrt{d}\begin{cases} n^{-1/2} & \text{if d} = 1 \\ \sqrt{\frac{\log n}{n}} & \text{if d} = 2 \\ n^{-1/d} & \text{if } d \geq 3. \end{cases}$$

It is worth noting that (a) surprisingly the empirical measure is consistent without smoothing and without smoothness assumptions, (b) the rate suffers from the curse-of-dimensionality (i.e. in high-dimensions the empirical measure converges very slowly to μ). We will eventually also show complementary lower bounds, i.e. we will show that this rate is unimprovable in general.