Lecture 13

**Announcements:**
- Select Wiki Article for revision by Mon 2/21.
- Make up class tomorrow:
  Friday 2/18, 1:30-2:45pm, SH 6635

From Kantorovich back to Monge...

**Thm (Knott-Smith Optimality Criterion):**
Fix $X \subset \mathbb{R}^d$, $\mu, \nu \in \mathcal{P}(X)$. Let $C(x, y) = |x - y|^2$.

(i) There exists $f_\ast \in L^2(\mu)$ proper, lsc, convex s.t.

\[
\sup_{\psi \in C_b(\mathbb{R}^d)} \int \psi \, d\mu + \int \psi \, d\nu = -P_0 = \int |x|^2 - 2f_\ast(x) \, d\mu(x) + \int |x|^2 - 2f_\ast(x) \, d\nu(x)
\]

(i.b) For any optimal transport plan $\delta_\ast$, we have $x^2 \in \mathcal{D}f_\ast(x)$ for $\delta_\ast$-a.e. $(x^1, x^2)$

(ii) Conversely, if $x \in \mathcal{P}(\mu, \nu)$ and $f \in L^2(\mu)$ proper, lsc, convex for which $x^2 \in \mathcal{D}f(x^1)$ for $\delta$-a.e. $(x^1, x^2)$ then...

(ii.a) $x$ is optimal

(ii.b) $-P_0 = \int |x|^2 - f(x) \, d\mu(x) + \int |x|^2 - 2f_\ast(x) \, d\nu(x)$.
Now, we have what we need to not only solve the Monge problem, but also to characterize its unique solution.

**Remark:** Recall that $t#\mu = \nu$

$$\Rightarrow \int \varphi(t(x)) d\mu(x) = \int \varphi(y) d\nu(y) \quad \forall \varphi \in L^1(\nu)$$

Note that if $t = s \mu$-a.e., then $t#\mu = s#\mu$.

**Thm (Brenier):** Given $\mu, \nu \in P_2(\mathbb{R}^d)$, $\mu \ll L^d$

1. For any optimal transport plan $\pi$, there exists an optimal transport map $t_*$ s.t. $\nu = (\text{Id} \times t_*)#\mu$

   In particular, $\inf M_2(t) = \inf \{K_2(\nu) \quad \text{for } \nu \in P(\mathbb{R}^d)\}$

2. Given $t$ s.t. $t#\mu = \nu$, $t$ is optimal $\iff t = \nabla \varphi$ for $\varphi \in L^1(\mu)$ convex, lsc defined $\mu$-a.e. differentiable $\mu$-a.e.

3. The OT map is unique, up to $\mu$-a.e. equiv.

The proof relies strongly on following theorem.
Thm: Given $\mu \in P_2(\mathbb{R}^d), \mu \ll \lambda^d, \Phi \in L^1(\mu)$ convex then
- $\Phi$ is differentiable $\mu$-a.e.
- where it is differentiable, $D\Phi = \{D\Phi\}$
- $\nabla \Phi$ coincides with the distributional gradient.

Now, we can prove Brenier's theorem!

\textbf{Pf:}

\textbf{CLAIM:} If $\tilde{\sigma}$ is an optimal plan and $\tilde{\sigma} = (id \times t)^* \# \mu$ for some $t$ s.t. $t^* \# \mu = \nu$, then $t$ is an optimal transport map.

Suppose $\hat{\sigma}$ is another transport map from $\mu$ to $\nu$, i.e., $\hat{\sigma} \# \mu = \nu$. Then
$$\hat{\sigma} = (id \times \hat{t})^* \# \mu \in \Pi(\mu, \nu).$$

$$|M_2(t)| = |M_2(\tilde{t})| = \int x^2 \, d\mu(x) = \int |x^2 - x_1^2| \, d\tilde{\sigma}$$
$$\leq \int |x^2 - x_1^2| \, d\tilde{\hat{\sigma}} = \int |\hat{t}^* (x) \cdot \hat{t} - x_1^2| \, d\mu(x) = |M_2(\hat{t})|.$$
Now prove 1:
Let $\delta_\#$ be an OT plan. By K-S Thm, exist $f \in L^1(\mu)$ proper, lsc, convex s.t.

\[ x^2 \in df_\#(x^1) \quad \delta_\#\text{-a.e.} \]

By prev Thm

\[ x^2 = \nabla f_\#(x^1) \quad \delta_\#\text{-a.e.} \]

Consequently, \( \forall \varphi \in L^1(\delta_\#) \)

\[ \int \varphi(x^1, x^2) \delta_\#(x^1, x^2) = \int \varphi(x^1, \nabla f_\#(x^1)) d\delta_\#(x^1, x^2) \]

\[ = \int \varphi(x^1, \nabla f_\#(x^1)) d\mu(x^2) \]

Thus \( \delta_\# = (id \times \nabla f) \# \mu \).

In particular, \( \nabla = \nabla_2 \# \delta_\# = \nabla f \# \mu \), so \( t = \nabla f \) is a transport map from \( \mu \) to \( \nu \).

By our claim, it is an optimal transport map.

Now, prove 2:
"\leq^\prime\" 
Given \( t = \nabla \Phi \) for such a \( \Phi \), define 
\( \gamma = (\text{id} \times t)^\# \mu \in \mathcal{P}(\mu) \).

Then 
\[
\int x^2 - \nabla \Phi(x^1)^2 \, d\gamma(x^2,x^2) = \int h(x^2) - \nabla \Phi(x^1)^2 \, d\mu(x^2) = 0
\]
so \( x^2 = \nabla \Phi(x^1) \in \partial \Phi(x^1) \) \( \gamma \)-a.e.

By K-S thm, \( \gamma \) is optimal. 
So CLAIM ensures \( t \) optimal.

"\Rightarrow" 
Given \( t \) optimal, \( \gamma = (\text{id} \times t)^\# \mu \) is an optimal plan.

By K-S thm, \( \exists \Phi^\star \) satisfying hypotheses s.t. \( x^2 = \nabla \Phi^\star(x^1) \) \( \gamma \)-a.e.

Thus, 
\[
0 = \int x^2 - \nabla \Phi^\star(x^1)^2 \, d\gamma(x^2,x^2) = \int h(x^2) - \nabla \Phi^\star(x^1)^2 \, d\mu(x^2)
\]
Therefore, \( t = \mathcal{V} \Phi \mu \)-a.e.

Now show \( \Theta \):

Suppose \( t, \tilde{t} \) are OT maps.

By (2), \( \exists \, \phi, \tilde{\Phi} \in L^1(\mu) \) convex, l.s.c., proper s.t. \( t = \mathcal{V} \phi \), \( \tilde{t} = \mathcal{V} \tilde{\Phi} \).

Arguing as before,

\[
\gamma := (\text{id} \times t) \# \mu, \quad \tilde{\gamma} := (\text{id} \times \tilde{t}) \# \mu
\]

then \( x_2 \in d \phi(x^1) \) \( \gamma \)-a.e., \( x_2 \in d \tilde{\Phi}(x^1) \) \( \tilde{\gamma} \)-a.e.

and

\[
\int x_2^2 - 2\Phi(x) + x_2^2 - 2\Phi(x) \, d\tilde{\gamma}(x^4, x^2) \\
\int x_2^2 - 2\tilde{\Phi}(x) \, d\mu(x) + \int x_2^2 - 2\Phi(x) \, d\gamma(x) \\
\int x_2^2 - 2\tilde{\Phi}(x) \, d\mu(x) + \int x_2^2 - 2\Phi(x) \, d\gamma(x) \\
\int x_2^2 - 2\tilde{\Phi}(x) + x_2^2 - 2\Phi(x) \, d\delta(x^2, x^4)
\]
Rearranging and recalling that $x^2 \in \mathcal{D}(x^2)$ iff equality holds in Young's inequality,
\[
\int \tilde{\Phi}(x^2) + \Phi^*(x^2) \, d\tilde{\gamma}(x^2, x^4) = \int \tilde{\Phi}(x^2) + \Phi^*(x^2) \, d\tilde{\gamma}(x^2, x^4)
\]
\[
= \int x^2 \cdot x^2 \, d\tilde{\gamma}(x^2, x^4)
\]
Rearranging,
\[
\int \tilde{\Phi}(x^2) + \Phi^*(x^2) - x^2 \cdot x^2 \, d\tilde{\gamma}(x^2, x^4) = 0
\]
\[
\int \tilde{\Phi}(x^2) + \Phi^*(\nabla \tilde{\Phi}(x^2)) - x^2 \cdot \nabla \tilde{\Phi}(x^2) \, d\tilde{\gamma}(x^2, x^4)
\]
\[
\int \tilde{\Phi}(x^2) + \Phi^*(\nabla \tilde{\Phi}(x^2)) - x^2 \cdot \nabla \tilde{\Phi}(x^2) \, d\mu(x^2)
\]
\[\geq 0 \text{ by Young} \]
Thus $\tilde{\Phi}(x^2) + \Phi^*(\nabla \tilde{\Phi}(x^2)) - x^2 \cdot \nabla \tilde{\Phi}(x^2) = 0$ $\mu$-a.e.
\[
\nabla \tilde{\Phi}(x^2) \in \partial \tilde{\Phi}(x^2) \quad \mu$-a.e.
\[
\nabla \tilde{\Phi}(x^2) = \nabla \Phi(x^2) \quad \mu$-a.e.

This shows OT maps are unique $\mu$-a.e.
In this way, the dual of the Kantorovich problem, with \( c(x_1, x_2) = |x_1^2 - x_2|^2 \) on \( \mathbb{R}^d \), helped us solve the Monge problem in this case.

The solution of the Monge problem will then be a key component in proving

\[
W_2(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \int |x-y|^2 \, d\gamma(x, y)
\]

is a metric on the space \( \mathcal{P}_2(\mathbb{R}^d) \).

But, before we leave duality behind, one last important application to the case \( c(x_1, x_2) = |x_1^2 - x_2|^2 \).

Recall...

**Thm:** For all \( \mu, \nu \in \mathcal{P}(\mathbb{R}) \), \( c: \mathbb{R} \times \mathbb{R} \to [0, \infty) \) cts,

\[
\inf_{\gamma \in \Pi(\mu, \nu)} \int |x-y|^2 \, d\gamma(x, y) = \sup_{(\varphi, \psi) \in C(\mathbb{R}) \times C(\mathbb{R})} \int \varphi(x) \, d\mu(x) + \int \psi(y) \, d\nu(y)
\]
Furthermore, the maximum is attained.

In the particular case $\mathbb{R} \times \mathbb{R}$, $c(x^1, x^2) = |x^1 - x^2|$, we can rewrite this in a nice way.

**Thm:** Given $\mathbb{R} \times \mathbb{R}$, $\mu, \nu \in \mathcal{P}(\mathbb{R})$,

$$
\inf_{\delta \in \mathcal{P}(\mu, \nu)} \int_{\mathbb{R}} |x - y| \, d\delta(x, y) = \sup_{\phi \in \mathcal{C}(\mathbb{R})} \phi \left( \mathbb{R} \times \mathbb{R} \right),
$$

$\phi \in \mathcal{C}(\mathbb{R})$, $\|\phi\|_{\text{Lip}} \leq 1$

and $\exists \phi^*$ that achieves maximum.

This gives another important way to compare probability measures, known as the 1-Wasserstein or Earth Movers Distance,

$$
\mathcal{W}_1(\mu, \nu) := \inf_{\delta \in \mathcal{P}(\mu, \nu)} \int_{\mathbb{R}} |x - y| \, d\delta(x, y).
$$
Remark: The first part of the theorem continues to hold on any Polish space, under the additional constraint that \( \Phi \in L^1(\mu - \nu) \).

\textbf{Pf: } By our duality theorem, it suffices to show

\[
\sup_{\Phi, \Psi \in \mathcal{C}(x)} \int \Phi d\mu + \int \Psi d\nu = \sup_{\Phi \in \mathcal{C}(x), \|\Phi\|_{\text{Lip}} \leq 1} \int \Phi d(\mu - \nu)
\]

First "\( \geq \)"

Note that \( \Phi \in \mathcal{C}(x), \|\Phi\|_{\text{Lip}} \leq 1 \), then \( \Phi(x^2) - \Phi(x^1) \leq |x^1 - x^2| \), so taking \( \Psi = -\Phi \) gives \((\Phi, \Psi)\) satisfying constraints on LHS w/ values of objective functions equal.

Next "\( \leq \)"

Take \((\Phi^*, \Psi^*)\) that attain maximum on LHS.
Double convexification trick

Define $\hat{\psi}(x^2) = \inf_{x^1} x^2 \in \mathcal{X} \{x^1 - x^2 - \Phi(x^1)\}$

- $\hat{\psi} \geq \psi$
- For all $\varepsilon > 0$, $\exists \tilde{x}^2 \in \mathcal{X}$ s.t.

$$-\hat{\psi}(\tilde{x}^2) \leq -|x^1 - \tilde{x}^2|$$
We now have obtained characterizations of the dual problem for both $W_1, W_2$.

**Exercise:** Consider

$$W_p(\mu, \nu) := \min_{\gamma \in \Pi(\mu, \nu)} \left( \int d\gamma(x, x') \, dx' \right)^{1/p}.$$  

Use Hölder's inequality to prove that

$$p \leq q \Rightarrow W_p(\mu, \nu) \leq W_q(\mu, \nu).$$

Furthermore, if $\text{supp}(\gamma)$ is bounded, prove that $\mathbb{E} C$ s.t.
\[ p \leq q \Rightarrow W_{q}(\mu, \nu) \leq C W_{p}(\mu, \nu). \]

We will now prove our earlier claim that \( W_2 \) is a metric on \( \mathcal{P}_2(\mathbb{R}^d) \).

Our proof will use Brenier's theorem... which requires \( \mu \ll \lambda^d \).

We need a way to approximate \( \mu \in \mathcal{P}_2(\mathbb{R}^d) \) by \( \mu_k \) that are absolutely continuous.

Prop: \( W_2 \) jointly lsc wrt narrow convergence.

Suppose \( \mu_n, \nu \in \mathcal{P}(\mathbb{R}^d) \) satisfy \( \mu_n \rightharpoonup \mu, \ \nu_n \rightharpoonup \nu \) narrowly. Then

\[ \lim_{n \to \infty} W_2(\mu_n, \nu_n) = W_2(\mu, \nu). \]
Approximation by Convolution

Def: (mollifiers, convolution)

mollifier: \( \Phi: \mathbb{R}^d \to \mathbb{R} \), bdd, measures s.t. \( \Phi \geq 0 \),
\( \Phi(x) = \Phi(-x) \), \( \int \Phi(x) dx = 1 \)
\( \Phi_\varepsilon(x) = \frac{1}{\varepsilon^d} \Phi(\frac{x}{\varepsilon}) \)

\( \int \Phi_\varepsilon = 1 \)

convolution: Given \( \mu \in \mathcal{P}(\mathbb{R}^d) \), define \( \Phi_\varepsilon \ast \mu \in \mathcal{P}(\mathbb{R}^d) \)
by,
\( \Phi_\varepsilon \ast \mu(x) = \int \Phi_\varepsilon(x-y) d\mu(y) \)
**Lemma:** Suppose \( \mu \in \mathcal{P}(\mathbb{R}^d) \)

(i) For any \( f \) measurable, bounded below,

\[
\int f \, d(\rho_\varepsilon \ast \mu) = \int f \ast \rho_\varepsilon \, d\mu
\]

"\( \rho_\varepsilon \) jumps from \( \mu \) to \( f \)"

associativity of convolution

(ii) \( \rho_\varepsilon \ast \mu \to \mu \) narrowly as \( \varepsilon \to 0 \).

**Pf:** Beginning with (i), since \( \exists \, m \in \mathbb{R} \) s.t. \( f = m \) a.e.,

\[
\int f \, d(\rho_\varepsilon \ast \mu) = \int (f-m) \, d(\rho_\varepsilon \ast \mu) + \int m \, d(\rho_\varepsilon \ast \mu)
\]

\[
\int f \ast \rho_\varepsilon \, d\mu = \int (f-m) \ast \rho_\varepsilon \, d\mu + \int m \ast \rho_\varepsilon \, d\mu
\]

remains to show

there are equal

Since \( f-m \geq 0 \), by Tonelli,

\[
\int (f(x)-m) \, d(\rho_\varepsilon \ast \mu)(x) = \int \int (f(x)-m) \rho_\varepsilon(x-y) \, d\mu(y) \, dx
\]

\[
= \int \int (f(x)-m) \rho_\varepsilon(x-y) \, dx \, d\mu(y)
\]

\[
= \int (f-m) \ast \rho_\varepsilon \, d\mu
\]

This completes proof of (i).
Now, show (ii). For any \( f \in C_0(\mathbb{R}^d) \), 
\( \phi + f \to f \) uniformly on compact subsets of \( \mathbb{R}^d \).

Goal: use that \( \phi + f \to f \) uniformly on compact sets to prove \( \int \phi d(\mathbb{E} + \mu) \to \int f d\mu \). Since \( f \) is arbitrary, this will give narrow convergence of \( \mathbb{E} + \mu \to \mu \).

"inner regular"

Since \( \mathbb{E} \) is tight (by Prokhorov), so \( \forall \delta > 0 \)
\( \exists \, K_8 \subseteq \mathbb{R}^d \) s.t.
\[
2 \| \phi \|_{\text{lip}} (\mathbb{R}^d \setminus K_8) < \frac{\delta}{2}.
\]

Furthermore, \( \exists \, \varepsilon_8 \) s.t. \( \varepsilon < \varepsilon_8 \) we have
\[
\| \mathbb{E} \varepsilon + f - f \|_{L^\infty(K_8)} < \frac{\varepsilon}{2}.
\]

Thus, \( \forall \varepsilon > 0 \), \( \exists \, \varepsilon_8 > 0 \) s.t. \( \varepsilon < \varepsilon_8 \) ensures,
\[
\left| \int \phi d(\mathbb{E} + \mu) - \int f d\mu \right|
= \left| \int (\mathbb{E} \varepsilon + f - f) d\mu \right|
\leq \left| \int (\mathbb{E} \varepsilon + f - f) d\mu \right|_{K_8} + \left| \int (\mathbb{E} \varepsilon + f - f) d\mu \right|_{\mathbb{R}^d \setminus K_8}
\leq 2 \| \phi \|_{\text{lip}} \mathbb{E} \varepsilon_8 + \frac{\varepsilon}{2}.
\]
\[ \| \varphi_\varepsilon \ast f - f \ast \lambda_\varepsilon \|_2 + 2 \| f \|_{L_\infty} \mu(\mathbb{R}^d \setminus K_\varepsilon) < \delta. \]

Since \( f \in C_b(\mathbb{R}^d) \) was arbitrary, \( \varphi_\varepsilon \ast \mu \rightarrow \mu \) narrowly.

**Lemma:** Given \( \mu \in \mathcal{P}(\mathbb{R}^d) \) and \( \varphi_\varepsilon \) as above,

(i) \( W_2(\mu, \varphi_\varepsilon \ast \mu) \leq \varepsilon (M_2(\varphi))^\frac{1}{2} \)

(ii) \( W_2(\varphi_\varepsilon \ast \mu, \varphi_\varepsilon \ast \nu) \leq W_2(\mu, \nu) \)

(iii) \( \lim_{\varepsilon \to 0} W_2(\varphi_\varepsilon \ast \mu, \varphi_\varepsilon \ast \nu) = W_2(\mu, \nu) \)

**Proof:** We begin with (i). Consider \( \delta \in \mathcal{M}(\mathbb{R}^d \times \mathbb{R}^d) \) defined by

\[ d\delta(x_1, x_2) = \varphi_\varepsilon(x_2 - x_1) \, dx_2 \, d\mu(x_1) \]

Then \( \forall f \in \mathcal{B}(\mathbb{R}^d) \),

\[ \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x_1) \, d\delta(x_1, x_2) = \int f(x_1) \varphi_\varepsilon(x_2 - x_1) \, dx_2 \, d\mu(x_1) \]

\[ = \int_{\mathbb{R}^d} f(x_1) \, d\mu(x_1) \]

\[ = \int_{\mathbb{R}^d} f(x_2) \, d\delta(x_1, x_2) = \int f(x_2) \varphi_\varepsilon(x_2 - x_1) \, dx_2 \, d\mu(x_1) \]
Thus, $\gamma \in \Gamma (\mu, \varphi_{\delta} \ast \mu)$. So by defn of $W_2$,

$$W_2^2(\mu, \varphi_{\delta} \ast \mu) = \int x_2 - x_1^2 \, d\gamma_{\delta} (x_1, x_2)$$

\[
\begin{align*}
\gamma &= \frac{x_2 - x_1}{\varepsilon} \\
\gamma &= \frac{1}{\varepsilon^2} \delta_{x_2}
\end{align*}
\]

$$= \int x_2 - x_1^2 \varphi_{\delta} (x_2 - x_1) \, dx_2 \, d\mu(x_1)$$

$$= \varepsilon^2 \int |z|^2 \varphi(z) \, dz \, d\mu(x_1)$$

$$= \varepsilon^2 M_2(\varphi).$$

Now prove (ii). Define $\overline{\varphi_{\delta}} (z, z') = \{ \varphi_{\delta} (z) \text{ if } z = z' \}$

Consider $\varphi_{\delta} \in M(R^d \times R^d)$ defined by

$$\delta \varepsilon := \overline{\varphi_{\delta}} \ast \gamma,$$

where $\gamma$ is an optimal plan from $\mu$ to $\nu$. 
For any $f: \mathbb{R}^d \to \mathbb{R}$ bdd, meas, define $f(x_1, x_2) = f(x_1)$.

Then,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} f(x_1) \, d\delta_\epsilon(x_1, x_2) = \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x_1, x_2) \, d((\bar{\epsilon} \ast \delta_\epsilon))(x_1, x_2)$$

$$= \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x_1, x_2) \, d\delta_\epsilon(x_1, x_2)$$

$$= \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x_1, x_2) \, d\delta_\epsilon(x_1, x_2)$$

$$= \int_{\mathbb{R}^d} f(x_1) \, d\delta_\epsilon(x_1)$$

Similarly, $\int_{\mathbb{R}^d} f(x_1) \, d\delta_\epsilon(x_1, x_2) = \int_{\mathbb{R}^d} f(x_2) \, d((\bar{\epsilon} \ast \nu)(\epsilon \ast \mu))$

Thus, we conclude $\delta_\epsilon \in \Gamma(\nu, \mu, \epsilon \ast \nu)$. Therefore,
\[ W_2^2 (\Phi^e + \mu, \Phi^e + \nu) \]
\[ \leq \int |x_1 - x_2|^2 \, d\overline{\Phi^e} + \chi(x_1, x_2) \]
\[ = \int \int (x_1 - z_1) - (x_2 - z_2)^2 \overline{\Phi^e}(z_1, z_2) \, dz_1 \, dz_2 \, \chi(x_1, x_2) \]
\[ = \int \int (x_1 - z_1) - (x_2 - z_2)^2 \overline{\Phi^e}(z_1, z_2) \, dz_1 \, dz_2 \, \chi(x_1, x_2) \]
\[ = \int \int |x_1 - x_2|^2 \, \overline{\Phi^e}(z_1, z_2) \, dz_1 \, dz_2 \, \chi(x_1, x_2) \]
\[ = W_2^2 (\mu, \nu) \]

Thus, \( W_2 (\Phi^e + \mu, \Phi^e + \nu) \leq W_2 (\mu, \nu) \).

Now part (iii).

By part (ii), \( \limsup_{\varepsilon \to 0} W_2 (\Phi^e + \mu, \Phi^e + \nu) \leq W_2 (\mu, \nu) \).

Since \( \Phi^e + \mu \to \mu \), \( \Phi^e + \nu \to \nu \) narrowly and \( W_2 \) is jointly \( \varepsilon \) wrt narrow convergence,

\[ \lim_{\varepsilon \to 0} W_2 (\Phi^e + \mu, \Phi^e + \nu) \leq W_2 (\mu, \nu) \].
taking these two limits gives the result.