Optimal Transport: Lecture #4

- Let us recall the Kantorovich problem. Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we let

$$\Pi(\mu,\nu) := \{ \gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \mid \forall A \subset \mathbb{R}^d,\ \gamma(A \times \mathbb{R}^d) = \mu(A) \}$$
$$\forall B \subset \mathbb{R}^d,\ \gamma(\mathbb{R}^d \times B) = \nu(B).$$

We call $\Pi(\mu,\nu)$ the space of couplings between μ, ν. The Kantorovich formulation of the OT problem with quadratic cost solves

$$\pi^* = \operatorname*{argmin}_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, d\pi(x,y). \qquad (KP)$$

- We saw last time that for general forms of this problem (i.e. $\mathbb{R}^d$ replaced with abstract Polish spaces, $\|x-y\|_2^2$ replaced with an abstract cost function satisfying certain properties) that (KP) admits a dual formulation:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, d\pi(x,y) = \sup_{(\ell,\psi) \in C_b(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d} \ell(x)\, d\mu(x) + \int_{\mathbb{R}^d} \psi(y)\, d\nu(y).$$
(bounded)
$$\text{s.t. } \ell(x) + \psi(y) \le \|x-y\|_2^2$$

- The "dual" formulation (RHS) says: maximize a linear functional

$$(\ell, \psi) \mapsto \int_{\mathbb{R}^d} \ell(x)\, d\mu(x) + \int_{\mathbb{R}^d} \psi(y)\, d\nu(y)$$

subject to the constraint $\ell(x) + \psi(y) \le \|x-y\|_2^2$. Note that the constraint is linear in ℓ, ψ, so in fact we have a linear objective subject to linear constraints.

Q: Can we address this with practical algorithms in the discrete case?

- First, let's see how this carries out in the discrete setting.
- Suppose instead of arbitrary measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we instead have discrete measures, i.e. those supported on a finite number of points in $\mathbb{R}^d$, i.e. sums of (potentially non-uniformly weighted) Dirac masses:

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j} \quad \text{s.t.} \quad \sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j = 1.$$

In this case, (KP) shakes out as follows:

$$\operatorname*{argmin}_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, d\pi(x,y)$$ can be reduced with μ, ν

to a matrix problem by noting that any $\pi \in \Pi(\mu,\nu)$ as above must be supported on points in the Cartesian product $\{x_i\}_{i=1}^n \times \{y_j\}_{j=1}^m =$

$$\{(x,y) \mid x = x_i \text{ for some } i = 1,\dots,n \text{ and } y = y_j \text{ for some } j = 1,\dots,m\}.$$

So, any $\hat{\pi} \in \Pi(\mu,\nu)$ is supported on only (at most!) $m \cdot n$ points in $\mathbb{R}^d \times \mathbb{R}^d$. This means

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, d\pi(x,y)$$

$$= \sum_{i,j} \|x_i - y_j\|_2^2 \cdot \hat{\pi}_{ij},$$

where $\hat{\pi}_{ij} \in [0,1]$ are how much mass gets sent from x_i to y_j (intuitively). This allows us to rewrite (KP) in a more familiar form for those

with background in discrete optimization:

$$\underset{P \in U(a,b)}{\arg\min} \sum_{i,j} C_{ij} \cdot P_{ij}, \quad \text{where: } C_{ij} := \|x_i - y_j\|_2^2 \text{ are costs}$$

P is an "assignment" matrix

$$U(a,b) = \{ P \in \mathbb{R}_{\geq 0}^{n \times m} \mid P \mathbb{1}_m = a, \; P^T \mathbb{1}_n = b \}$$

is the space of admissible transport matrices

$\mathbb{1}_n^T = (1, 1, \ldots, 1)$ n times

- This is evidently a linear program, but we can write it in standard form (linear objective, equality constraints defined with a matrix and constant vector, and non-negativity constraints on the variables).

- Let $I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$ (n times)

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ \vdots & & & \\ a_{n1}B & \cdots & & a_{nm}B \end{bmatrix} \in \mathbb{R}^{np \times mq}$$

$A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{p \times q}$

be the Kronecker product of A, B

- Let $A = \begin{bmatrix} \mathbb{1}_n^T \otimes I_m \\ I_n \otimes \mathbb{1}_m^T \end{bmatrix} \in \mathbb{R}^{(n+m) \times (nm)}$

$\in \mathbb{R}^{n \times nm}$

Exercise: $P \in U(a,b) \iff p \in \mathbb{R}_{\geq 0}^{nm}$ and $A p = \begin{bmatrix} a \\ b \end{bmatrix}$, where

p is the "vectorized" format of P with $p_{i+n(j-1)} = P_{ij}$.

$p \in \mathbb{R}^{n \cdot m \times 1}$

Letting $L_c(a,b) = \min_{P \in U(a,b)} \sum_{\substack{i=1,\dots,n \\ j=1,\dots,m}} C_{ij} P_{ij}$ be the OT distance

(using that word informally), we thus have

$$L_c(a,b) = \min_{p \in \mathbb{R}^{nm}_{\geq 0}} c^T p, \qquad \text{(KP-LP)}$$

$$\pi^{mn \times 1} \text{ s.t. } Ap = \begin{bmatrix} a \\ b \end{bmatrix}$$

where again c is the vectorized form of the cost matrix $C \in \mathbb{R}^{n \times m}$, namely $c_{i+n\cdot(j-1)} = C_{ij}$. This has the standard form of a linear program, so there are many off-the-shelf tools to handle it.

Remark: Formulating the transport as a (discrete) linear program is what Kantorovich's original work and applications centered on. Note, Kantorovich won the Nobel (Memorial) Prize in Economics, indicating the huge impact his work had on the applied side of things.

As we might hope (KP-LP) admits a dual formulation that matches what we saw in the general setting:

$$L_c(a,b) = \max_{h \in \mathbb{R}^{n+b}} \begin{bmatrix} a \\ b \end{bmatrix}^T h.$$

$$A^T h \leq c$$

Remark: Suppose $m = n$, so that there are the same number of support

points in μ and ν, and suppose moreover $a = b = \frac{1}{n}\mathbb{1}_n$, i.e. the weights are uniform on both measures. Then

$$U(a,b) = U\left(\tfrac{1}{n}\mathbb{1}_n, \tfrac{1}{n}\mathbb{1}_n\right)$$

$$= \left\{ P \in \mathbb{R}^{n\times n}_{\geq 0} \text{ s.t. } P\mathbb{1}_n = P^T\mathbb{1}_n = \tfrac{1}{n}\mathbb{1}_n \right\}$$

So, $U(a,b)$ is the set of non-negative (entrywise) matrices with rows and columns that sum up to $\frac{1}{n}\mathbb{1}_n$.

An important point is that, $P \in U\left(\frac{1}{n}\mathbb{1}_n, \frac{1}{n}\mathbb{1}_n\right)$ could be very diffuse, (e.g., $P = \begin{pmatrix} \frac{1}{n^2} & \frac{1}{n^2} & \cdots & \frac{1}{n^2} \\ \frac{1}{n^2} & & & \\ \vdots & & & \frac{1}{n^2} \\ \frac{1}{n^2} & \cdots & & \frac{1}{n^2} \end{pmatrix} \in U$). This is because KP allows us to "spread mass" far and wide, at least in principal. In contrast, the Monge problem specifies a map from $\mathrm{supp}(\mu)$ to $\mathrm{supp}(\nu)$. In the discrete setting, we can capture this idea through permutation matrices: let $\sigma : \{1,\dots,n\} \to \{1,\dots,n\}$ be a permutation. Then we have an associated matrix

$$[P_\sigma]_{ij} = \begin{cases} \frac{1}{n}, & j = \sigma(i) \\ 0, & \text{else} \end{cases}$$

Note $P_\sigma \in U\left(\frac{1}{n}\mathbb{1}_n, \frac{1}{n}\mathbb{1}_n\right)$ for all permutations σ, and in fact $\sum_{i,j=1}^n [P_\sigma]_{ij} C_{ij} = \sum_{i=1}^n C_{i\sigma(i)}$, which is the discrete analogue of the "$= C_{x T(x)}$" objective in the Monge problem, $\int_{\mathbb{R}^d} \|x - T(x)\|_2^2 \, d\mu(x)$.

- So, the discrete Monge problem involves optimizing the usual cost functional over permutation matrices rather than ∞ of bi-stochastic matrices:

$$\min_{\sigma : \{1,\dots,n\} \circlearrowright \text{permutation}} \sum_{i=1}^{n} C_{i\sigma(i)} \qquad \text{(discrete MP)}.$$

- Note, since (discrete KP) $= \min_{U \in \Pi(\frac{1}{n}\mathbb{1}_n, \frac{1}{n}\mathbb{1}_n)} \sum_{i,j=1}^{n} C_{ij} U_{ij}$ optimizes over a larger set, (discrete KP) $\leq$ (discrete MP). But, (discrete MP) is a combinatorial optimization (hard to work with permutation matrices directly), so (discrete KP) is preferable. In this sense, KP relaxes MP.

Note: A focus of the course going on will be to understand in both narrow and general settings when the relaxed KP coincides with MP.

Next time: More on (KP-LP) and discussion of approximations that improve computational complexity.