Optimal Transport: Lecture #3

Last time: We established the form of the Kantorovich problem in the setting of discrete measures. As usual, let $\Pi(\mu,\nu)$

$$:= \{\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \mid \gamma(A \times \mathbb{R}^d) = \mu(A), \gamma(\mathbb{R}^d \times B) = \nu(B), \text{ f-}$$
$$\text{Borel } A, B\}$$

be the space of couplings between measures μ, ν.

- When we consider the common squared Euclidean cost function $c(x,y) = \|x-y\|_2^2$, the Kantorovich problem is:

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

- When μ, ν are discrete and supported, respectively, on $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$, this reduces to a minimization over bistochastic matrices: let $a :=$ $b :=$

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

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

- We saw last time this can be written as a linear program. This can be solved using off-the-shelf methods (e.g. simplex algorithms, interior point methods), though these scale quite poorly in the problem size.

- More precisely, if $n = m$, so we are optimizing over $U(a,b) = \{P \in \mathbb{R}_{\geq 0}^{n \times n} \mid P\mathbb{1}_n = a,\ P^T\mathbb{1}_n = b\}$, then typical algorithms will on average scale roughly like $\Omega(n^3)$, and could be worse in practice.

- This area is well-developed and there are bespoke algorithms available that achieve essentially cubic complexity in practice, but this is still bad. (A possible presentation could focus on some of these bespoke algorithms)

- A major innovation is due to M. Cuturi in 2013. The idea is to regularize the problem, allowing alternative, simpler approaches to be used for optimization.

=

- Let's establish some preliminaries. Let $P \in \mathbb{R}_{\geq 0}^{n \times m}$ be a coupling matrix. We define the (discrete) entropy of P as

$$H(P) := -\sum_{i,j} P_{ij}\left(\log(P_{ij}) - 1\right),$$

where $H(P) = -\infty$ if any entry of P is 0.

- Note that $H : \mathbb{R}^{n \times n} \to \mathbb{R}$, and we can compute its gradient and Hessian:

$$\partial_{ij} H(P) = -\left[\log(P_{ij}) + 1 - 1\right]$$

$$\Rightarrow \partial_{ij,i'j'}^{H(P)} = \begin{cases} 0, & ij \neq i'j' \\ -\frac{1}{P_{ij}}, & ij = i'j' \end{cases}$$

$$\Rightarrow \partial^2 H(P) = -\operatorname{diag}\left(\frac{1}{P_{ij}}\right)$$

Now, by assumption, $P_{ij} \leq 1 \Rightarrow P_{ij}^{-1} \geq 1 \Rightarrow H$ is strongly-concave. In particular, $-\varepsilon H(P)$ is ε-strongly convex in P.

- Now, the map $P \to \sum_{ij} P_{ij} C_{ij}$ is linear in P, and in particular its Hessian is 0. We conclude that the entropy-regularized cost functional

$$P \to \underbrace{\sum_{ij} P_{ij} C_{ij}}_{= \langle P, C \rangle} - \varepsilon H(P)$$

is ε-strongly convex.

- This implies (⋆)

$$L_c^{\varepsilon}(a,b) := \min_{P \in U(a,b)} \left\{ \langle P, C \rangle - \varepsilon H(P) \right\}$$

has a unique optimum.

- One can see immediately that introducing $-\varepsilon H(P)$ kills any sparsity in the optimizer $P_\varepsilon^* = \operatorname{argmin}_{P \in U(a,b)} \left\{ \langle P, C \rangle - \varepsilon H(P) \right\}$. Indeed, if $\left[P_\varepsilon^*\right]_{ij} = 0$ for any (i,j) index, then $\langle P, C \rangle - \varepsilon H(P) = +\infty$.

- In this sense, regularizing with entropy ensures we have a "diffuse" coupling, rather than a sparse one as in the classical Monge problem.

- So, in some sense we are "giving up" on the (frankly, desirable) property of sparse maps/couplings. So, why do this? What's the gain?

- The key point is, Sinkhorn's algorithm allows (⋆) to be solved much faster than its unregularized cousin.

- Let $K_\varepsilon \in \mathbb{R}^{n \times m}_{>0}$ be the Gibbs/heat kernel associated to the cost matrix $C \in \mathbb{R}^{n \times m}$: $\left[K_\varepsilon\right]_{ij} = \exp\left(-C_{ij}/\varepsilon\right)$. So, if we have

quadratic cost, then

$$[K_\varepsilon]_{ij} = \exp(-\|x_i - x_j\|^2/\varepsilon).$$

Theorem (structure of solution to (⋆)): The solution to (⋆) has the form

$$[P_\varepsilon^*]_{ij} = u_i [K_\varepsilon]_{ij} v_j \quad \text{for two (unknown) vectors } (u,v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$$

Proof: Let $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ be dual variables for the marginal constraint enforcing $P \in U(a,b)$. Then the Lagrangian is

$$\mathcal{L}(P,f,g) = \underbrace{\langle P, C\rangle - \varepsilon H(P)}_{\text{original objective}} - \underbrace{\langle f, P\mathbb{1}_m - a\rangle - \langle g, P^T\mathbb{1}_n - b\rangle}_{\text{constraints}}$$

Differentiating in P yields

$$\frac{\partial \mathcal{L}(P,f,g)}{\partial P_{ij}} = C_{ij} - \varepsilon \log(P_{ij}) - f_i - g_j.$$

Setting equal to 0 and using first order optimality, we have

$$C_{ij} - \varepsilon \log(P_{ij}) - f_i - g_j = 0$$

$$\Leftrightarrow \quad P_{ij} = \exp(f_i/\varepsilon)\exp(-C_{ij}/\varepsilon)\exp(g_j/\varepsilon)$$

$= u_i [K_\varepsilon]_{ij} v_j$. ∎

- One can find the vectors (u,v) through an alternating scheme as follows.
- Note that $P^*_{ij} = u_i [K_\varepsilon]_{ij} v_j \Leftrightarrow P^* = \mathrm{diag}(u) K_\varepsilon \mathrm{diag}(v)$, and the constraints shake out to
$$\begin{cases} \mathrm{diag}(u) K_\varepsilon \mathrm{diag}(v) \mathbb{1}_m = a, \\ \mathrm{diag}(v) K_\varepsilon^T \mathrm{diag}(u) \mathbb{1}_n = b \end{cases}$$
- Let [redacted] us define a sequence of iterates as $v^{(0)} = \mathbb{1}_m$ and
$$u^{(\ell)} = \frac{a}{K_\varepsilon v^{(\ell)}},$$
$$v^{(\ell)} = \frac{b}{K_\varepsilon^T u^{(\ell+1)}}.$$

- One can show (see Peyré & Cuturi) that this sequence of iterates converges to the optimal u, v. Note moreover that an update requires only a matrix multiplication $\to O(mn)/O(n^2)$. If there are not too many iterates, this gives us an essentially quadratic algorithm.