OT Lecture 1: Introduction to Optimal Transport

1 The Monge problem (1781)

Monge was interested in the following question – how can one transport a pile of sand to fill a ditch while minimizing the cost of transporting the sand?

It's easy to see that the pile and ditch have to have the same volume, and we can normalize them to have volume 1. Slightly more abstractly one can formulate the question as that of transporting a probability measure μ to another probability measure ν, while minimizing some cost. We'll need to be more precise about what exactly "transporting" means.

A *transport map* is a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that the push-forward constraint

$$T_{\#}\mu = \nu, \qquad \text{is satisfied.} \tag{1}$$

This constraint is shorthand for the constraints that,

$$\mu\Big(T^{-1}(B)\Big) = \nu(B) \ \text{ for all Borel } B.$$

Then, given a cost $c(x, y)$, the *Monge problem* seeks

$$\inf_{T:\, T_{\#}\mu=\nu} \int c(x, T(x))\, d\mu(x), \tag{2}$$

for example $c(x, y) = \|x - y\|$ or perhaps $c(x, y) = \|x - y\|^p$ for $p \geq 1$.

Non-existence/ill-posedness. In general a map cannot *split* mass, so a feasible T need not exist (e.g. when μ is a point mass and ν is a sum of two separated point masses).

In general, even when a solution exists, trying to study Monge's problem is quite challenging. Even when the cost is "nice", the constraints are highly non-linear/non-convex. One way to see this is to observe that if μ, ν had nice densities (and we could assert that the map was smooth) then the constraints encode the change-of-variables formula:

$$\mu(x) = \nu(T(x))\det(\nabla T(x)).$$

2 Kantorovich relaxation (1941)

Define the set of *couplings* (or transport plans)

$$\Gamma(\mu, \nu) = \Big\{\gamma \text{ prob. on } \mathbb{R}^d \times \mathbb{R}^d :\ \gamma(A \times \mathbb{R}^d) = \mu(A),\ \gamma(\mathbb{R}^d \times B) = \nu(B)\Big\}. \tag{3}$$

The Kantorovich problem is the infinite-dimensional linear program

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int c(x, y)\, d\gamma(x, y). \tag{4}$$

The feasible set $\Gamma(\mu, \nu)$ is nonempty and convex; when a minimizer is *induced by a map* $\gamma = (\mathrm{id}, T)_{\#}\mu$, it solves Monge as well.

3 Discrete case and the Birkhoff polytope

Suppose $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}$. A coupling is a nonnegative matrix $\gamma \in \mathbb{R}^{n\times n}$ with $\sum_j \gamma_{ij} = \frac{1}{n}$ and $\sum_i \gamma_{ij} = \frac{1}{n}$ (doubly stochastic). Let $C_{ij} = c(x_i, y_j)$. Then Kantorovich reduces to

$$\min_{\gamma\in\Gamma} \langle C, \gamma\rangle \qquad \text{subject to } \gamma \text{ doubly stochastic.} \tag{5}$$

By the Birkhoff–von Neumann theorem, the feasible polytope is the convex hull of permutation matrices, hence a linear objective attains an optimum at a permutation: one solution is induced by a map (Monge solution).

4 Wasserstein distances

For $p \geq 1$ and measures with finite p-th moments, define

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^p \, d\gamma(x, y), \qquad W_p(\mu, \nu) = \left(W_p^p(\mu, \nu)\right)^{1/p}. \tag{6}$$

Remark 1 (TV distance as a special cost)**.** With $c(x, y) = \mathbb{1}\{x \neq y\}$,

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{P}(X \neq Y) = \tfrac{1}{2}\|\mu - \nu\|_{\mathrm{TV}}, \tag{7}$$

the optimal value equals the total variation distance and ignores the ground metric.

5 Metric properties

Proposition 1. For $p \geq 1$, W_p defines a metric on $\mathcal{P}_p(\mathbb{R}^d)$.

Sketch. Nonnegativity and symmetry are immediate. If $\mu = \nu$, $T(x) = x$ yields $W_p(\mu, \nu) = 0$; conversely $W_p(\mu, \nu) = 0$ implies $\mu = \nu$. For the triangle inequality, let $\gamma_{XY} \in \Gamma(\mu, \rho)$ and $\gamma_{YZ} \in \Gamma(\rho, \nu)$ be optimal couplings for random variables (X, Y) and (Y, Z). By the *gluing lemma* there exists a joint law of (X, Y, Z) with these marginals. Then by Minkowski,

$$\begin{aligned} W_p(\mu, \nu) &\leq \left(\mathbb{E}\|X - Z\|^p\right)^{1/p} \leq \left(\mathbb{E}\|X - Y\|^p\right)^{1/p} + \left(\mathbb{E}\|Y - Z\|^p\right)^{1/p} \\ &= W_p(\mu, \rho) + W_p(\rho, \nu). \end{aligned} \tag{8}$$

□