Optimal Transport: Lecture 1

- Let us recall some analysis basics.
- Recall $L^p(\Omega)$ for $\Omega \subset \mathbb{R}^d$ is the set of functions f (up to almost everywhere equivalence) such that $|f|^p$ is integrable: $L^p(\Omega) := \{f : \Omega \to \mathbb{C} \mid \int_\Omega |f|^p < \infty\}$.

Here, integration is w.r.t. the Lebesgue measure but the class could be defined w.r.t. other measures μ. When we wish to emphasize the measure μ, we will write $L^p(\Omega, \mu)$.

- So, $L^1(\mathbb{R}^d)$ is the set of functions f s.t. $\int_{\mathbb{R}^d} |f| < \infty$. While $L^2(\mathbb{R}^d)$ is the classical subject of functional and Fourier analysis owing to its Hilbert space structure, $L^1(\mathbb{R}^d)$ will be more salient to us.
- Let $f \in L^1(\mathbb{R}^d)$. Then $\tilde{f} := \frac{|f|}{\int |f|}$ defines a non-negative function that integrates to 1. This provides a natural association of $L^1(\mathbb{R}^d)$ with probability measures in $\mathbb{R}^d$.
- Indeed, let $\mathcal{M}(\mathbb{R}^d)$ denote the space of finite, signed measures over $\mathbb{R}^d$. If $\mu \in \mathcal{M}(\mathbb{R}^d)$ has the extra properties that (i) $\mu(\mathbb{R}^d) = 1$

(ii) $\mu(A) \geq 0$ for all A a Borel subset of $\mathbb{R}^d$,

we say μ is a probability measure. We denote the subset of probability measures as $\mathcal{P}(\mathbb{R}^d)$.

Exercise: Show that $\mathcal{M}(\mathbb{R}^d)$ is generated by closing $\mathcal{P}(\mathbb{R}^d)$ under scalar

multiplication and addition (i.e. under the usual vector space operations for L^p).

Remark: An important subclass of $\mathcal{P}(\mathbb{R}^d)$ will be those measures that can be written as Lebesgue integration with respect to $f \in L^1(\mathbb{R}^d)$:

$$\mu(A) = \int_A f .$$

We call such μ absolutely continuous (a.c.) w.r.t. the Lebesgue measure. The function f is the density of μ, and for the subclass of a.c. probability measures $\mathcal{P}_{ac}(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$, it is natural to identify $\mu \leftrightarrow f$.

Importantly, not all measures are a.c. The easiest example is the Dirac mass (and combinations thereof), i.e. $\delta_{x_0}(A) = \begin{cases} 1, & x_0 \in A \\ 0, & \text{else} \end{cases}$

There are even stranger measures (see Lebesgue decomposition Theorem and its refinements).

• Note that we can compare probability measures using the total variation metric:

$TV(\mu, \nu) = \sup_{A \text{ Borel}} |\mu(A) - \nu(A)|$. One can also, in the case of a.c measures $\mu \leftrightarrow f_\mu$, $\nu \leftrightarrow f_\nu$, consider the Kullback-Leibler divergence

$$KL(\mu, \nu) = \int_{\mathbb{R}^d} f_\mu(x) \log\left(\frac{f_\mu(x)}{f_\nu(x)}\right) dx .$$

Exercise: (i) Show TV is a metric on $\mathcal{P}(\mathbb{R}^d)$

(ii) Show KL is not a metric on $\mathcal{P}_{ac}(\mathbb{R}^d)$.

• One interpretation of the theory of optimal transport is that it provides the "right" way of comparing probability measures. What do we mean by right? Essentially, we want convergence in a certain metric/divergence to imply useful properties and be geometrically intuitive. We will see Wasserstein distances (Chapter 7 of Villani) achieves these desired properties.

• So, what is optimal transport? The basic idea is to find a map (or generalization thereof) which carries one probability measure to another in a "cost-minimizing manner."

• More precisely, for two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we say a map $T: \mathbb{R}^d \to \mathbb{R}^d$ pushes μ on ν / is a pushforward of μ onto ν if

$$\forall A \subset \mathbb{R}^d \text{ Borel measurable}, \quad \nu(A) = \mu(T^{-1}(A))$$

Notationally, $T_{\#}\mu = \nu$.

ex: Let $d=1$, μ be the uniform measure on $[0,1]$. Then for any smooth bijection $T: [0,1] \to \mathbb{R}$, $T_{\#}\mu$ is the measure with the property that

$$\left[T_{\#}\mu\right](A) = \mu\left(T^{-1}(A)\right)$$

$$= \int_{T^{-1}(A)} \mathbb{1}_{[0,1]}(x)\, dx$$

Letting $u = T(x)$ so that $du = T'(x)\,dx$

$x = T^{-1}(u)$ $\quad = T'(T^{-1}(u))\,dx,$

$$[T_{\#}\mu](A) = \int_A \mathbb{1}_{[0,1]}(T^{-1}(u)) \cdot \frac{1}{T'(T^{-1}(u))}\,du.$$ We conclude that

the pushforward measure $T_{\#}\mu$ (i) is a.c.
(ii) has density $\mathbb{1}_{[0,1]}(T^{-1}(u)) \cdot \frac{1}{T'(T^{-1}(u))}$

Note, this could have been done for an arbitrary base measure μ, not just the uniform distribution.

- Our analysis of OT will start by considering optimal pushforwards, in the following sense. Let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be measurable. We think of c as a "cost function" which quantifies how expensive it is to move probability mass at point x to point y. Common choices are $c(x,y) = \|x-y\|_p^p$, especially $p = 1, 2$. We will see certain aspects of the cost function have huge influence.

(abbreviation for optimal transport)

Monge Formulation (MP):

$$T^* = \operatorname*{arg\,min}_{T \text{ s.t. } T_{\#}\mu = \nu} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, T(x))\,d\mu(x)$$

- This idea - dating to pre-revolutionary France - seems simple enough. Among all

pushforwards $T_{\#}\mu = \nu$, find the one which minimizes the "average c-cost", with averaging done w.r.t. μ.

- While natural to formulate, the Monge problem is (i) difficult to analyze head-on and (ii) may behave poorly (eg no solutions, non-uniqueness). Surprisingly perhaps, this problem lay somewhat fallow until its relaxation by Soviet mathematician Kantorovich. This led to a re-formulation of the OT problem in an idiom closer to operations research and linear programming.

- The relaxation requires us to optimize not over maps $T_{\#}\mu = \nu$, but over plans $\pi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that have marginals given by μ, ν. More precisely, let

$$\Pi(\mu,\nu) = \left\{ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ s.t. } \pi[A \times Y] = \mu(A) \quad \forall A, B \text{ Borel sets} \right\}$$
$$\pi[X \times B] = \nu(B)$$

Remark: $\Pi(\mu,\nu)$ is non-empty: $\hat{\pi}(x,y) := \mu(x)\,\nu(y)$ is always in $\Pi(\mu,\nu)$. We shall see that lying in $\Pi(\mu,\nu)$ is essentially a linear constraint, which will give us some basic tools in the discrete case.

Kantorovich Formulation (KP):

$$\pi^* = \operatorname*{argmin}_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\pi(x,y).$$

- Next time, we will begin to analyze (KP) and discuss how it relates to (MP)