

# Notes on Optimal Transport

## ENSAE

### 3A

Gabriel Romon

Last updated: Wednesday 6<sup>th</sup> February, 2019 at 23:21

## 1 Introduction

These are notes I gathered from a seminar session given by Shuangjian Zhang at ENSAE and from lectures given by Marco Cuturi at MLSS 2019 in South Africa.

Likelihood maximization is an instance of generative modeling: given data points  $x_1, \dots, x_N \in \mathbb{R}^d$  and a family of densities  $(f_\theta)_{\theta \in \Theta}$ , we look for the  $f_\theta$  that matches the most the empirical data distribution defined as  $\nu_{\text{data}} := \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ . This is done by maximizing the log-likelihood  $\frac{1}{N} \sum_{i=1}^N \log f_\theta(x_i)$ . This quantity exists only if  $f_\theta(x_i) > 0$  for all  $i$ , which forces the densities to have the whole of  $\mathbb{R}^d$  as support. Likelihood maximization can be given a geometric interpretation: if one overlooks that  $\nu_{\text{data}}$  and  $f_\theta$  are not absolutely continuous with respect to the same measure (the former is discrete), minimizing the Kullback-Leibler divergence between the two writes as  $\operatorname{argmin}_{\theta \in \Theta} \operatorname{KL}(\nu_{\text{data}} || f_\theta) = \operatorname{argmin}_{\theta \in \Theta} E_{X \sim \nu_{\text{data}}} [\log(f_{\text{data}}(X)) - \log(f_\theta(X))]$

$$\begin{aligned} &= \operatorname{argmin}_{\theta \in \Theta} -E_{X \sim \nu_{\text{data}}} \log(f_\theta(X)) \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N \log f_\theta(x_i) \end{aligned}$$

A weakness of likelihood maximization is its poor scalability to high-dimensional settings, which are commonplace. For instance, a  $100 \times 100$  image with 3 color channels lives in  $\mathbb{R}^{30.000}$ . Instead of working directly in the data space  $\mathbb{R}^d$ , we may rather consider a latent space  $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mu)$  with  $p \ll d$ , and measurable functions  $g_\theta : \mathbb{R}^p \rightarrow \mathbb{R}^d$  (e.g deconvolution networks). We define the pushforward measure  $g_{\theta\#}\mu$  by  $\forall B \in \mathcal{B}(\mathbb{R}^d), g_{\theta\#}\mu(B) := \mu(g_\theta \in B)$  and we look for  $\theta$  such that  $g_{\theta\#}\mu$  matches  $\nu_{\text{data}}$ . This requires setting a metric on the space of probability measures, and fortunately, many exist: Hellinger, Kantorovitch, MMD, Wasserstein... Some of these metrics arise from the theory of optimal transport.

## 2 Optimal transport

### 2.1 Monge problem and its Kantorovitch relaxation

Let  $\mu$  and  $\nu$  be probability measures on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$  and  $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$  a cost function. Monge's optimal transportation problem is the following:

$$\inf_{\substack{T \text{ measurable} \\ T_{\#}\mu = \nu}} \int c(x, T(x)) d\mu(x) \tag{MP}$$

In words,  $T$  is a transportation mapping that assigns to each  $x$  exactly one location  $T(x)$ , implying that the mass at  $x$  cannot be split to several locations.  $T_{\#}\mu = \nu$  is the transportation constraint, and the objective  $\int c(x, T(x)) d\mu(x)$  measures the total transportation cost.

The Kantorovitch relaxation allows for mass splitting by considering couplings. A probability measure  $P$  on  $\mathbb{R}^d \times \mathbb{R}^d$  is said to be a coupling of  $(\mu, \nu)$  if for any  $A, B \in \mathcal{B}(\mathbb{R}^d)$ ,  $P(A \times \mathbb{R}^d) = \mu(A)$  and  $P(\mathbb{R}^d \times B) = \nu(B)$ . Let  $\Pi(\mu, \nu)$  denote the set of such couplings.  $\Pi(\mu, \nu)$  is clearly non-empty since it contains the product measure  $\mu \otimes \nu$ .

Kantorovitch's optimal transportation problem is the following:

$$\inf_{P \in \Pi(\mu, \nu)} \int \int c(x, y) dP(x, y) \quad (\text{KP-primal})$$

It is important to note that (KP-primal) provides a lower bound for (MP). Indeed, if  $T$  verifies  $T_{\#}\mu = \nu$  and we let  $P = (\text{Id}, T)_{\#}\mu$ , it is easy to check that  $P$  is a coupling. A theorem of integration with respect to pushforward measures yields  $\int \int c(x, y) dP(x, y) = \int c \circ (\text{Id}, T)(x) d\mu(x) = \int c(x, T(x)) d\mu(x)$ , hence the claim.

The dual of (KP-primal) is defined as follows:

$$\sup_{\substack{\varphi \in L_1(\mu) \\ \psi \in L_1(\nu) \\ \forall (x, y), \varphi(x) + \psi(y) \leq c(x, y)}} \int \varphi d\mu + \int \psi d\nu \quad (\text{KP-dual})$$

**Theorem 2.1.** *If  $c$  is lower semi-continuous, then (KP-primal) and (KP-dual) share the same optimal value. Moreover, the infimum in (KP-primal) is attained.*

*Proof.* We give a partial proof of the first part of the theorem. Let  $\varphi \oplus \psi : (x, y) \mapsto \varphi(x) + \psi(y)$  and

$$\iota_{\Pi}(P) := \sup_{\substack{\varphi \in L_1(\mu) \\ \psi \in L_1(\nu)}} \int \varphi d\mu + \int \psi d\nu - \int \varphi \oplus \psi dP$$

Let  $P \in \Pi(\mu, \nu)$ . If we let  $\pi_1$  and  $\pi_2$  denote the projections respectively on the first and last  $d$  coordinates, then  $\int \varphi \oplus \psi dP = \int \varphi \circ \pi_1 dP + \int \psi \circ \pi_2 dP$ . For  $\varphi = 1_A$ ,

$$\int \varphi \circ \pi_1 dP = \int 1_{\pi_1^{-1}(A)} dP = P(\pi_1^{-1}(A)) = P(A \times \mathbb{R}^d) = \mu(A)$$

Therefore  $\int \varphi d\mu + \int \psi d\nu - \int \varphi \oplus \psi dP = 0$  when  $\varphi$  and  $\psi$  are indicators, and a standard limit argument shows that it remains 0 for arbitrary integrable  $\varphi$  and  $\psi$ . Thus  $\iota_{\Pi}(P) = 0$  when  $P$  is a coupling.

If  $P \notin \Pi(\mu, \nu)$ , WLOG there exists some  $A \in \mathcal{B}(\mathbb{R}^d)$  such that  $P(A \times \mathbb{R}^d) \neq \mu(A)$ . If  $P(A \times \mathbb{R}^d) < \mu(A)$ , consider  $\varphi = \lambda 1_A$  with  $\lambda > 0$  and  $\psi = 0$ . Then

$$\int \varphi d\mu + \int \psi d\nu - \int \varphi \oplus \psi dP = \lambda(\mu(A) - P(A \times \mathbb{R}^d)) \xrightarrow{\lambda \rightarrow \infty} \infty$$

Hence  $\iota_{\Pi}(P) = \infty$ . The other case is similar.

If we let  $\mathcal{M}^+$  denote the set of positive measures on  $\mathbb{R}^d \times \mathbb{R}^d$ , then (KP-primal) turns into

$$\begin{aligned} \inf_{P \in \mathcal{M}^+} \int \int c dP + \iota_{\Pi}(P) &= \inf_{P \in \mathcal{M}^+} \sup_{\substack{\varphi \in L_1(\mu) \\ \psi \in L_1(\nu)}} \int c dP + \int \varphi d\mu + \int \psi d\nu - \int \varphi \oplus \psi dP \\ &= \inf_{P \in \mathcal{M}^+} \sup_{\substack{\varphi \in L_1(\mu) \\ \psi \in L_1(\nu)}} \int c - \varphi \oplus \psi dP + \int \varphi d\mu + \int \psi d\nu \\ &= \sup_{\substack{\varphi \in L_1(\mu) \\ \psi \in L_1(\nu)}} \inf_{P \in \mathcal{M}^+} \left[ \int c - \varphi \oplus \psi dP \right] + \int \varphi d\mu + \int \psi d\nu \end{aligned}$$

Switching the inf and the sup is the technical hurdle of the proof. In [1] Villani provides a rigorous justification that is quite involved.

Next,  $\inf_{P \in \mathcal{M}^+} \int c - \varphi \oplus \psi \, dP$  can be rewritten more simply. If  $c - \varphi \oplus \psi \geq 0$ , 0 is a lower bound for the integral, and it is attained for  $P = 0$ . Hence  $c - \varphi \oplus \psi \geq 0 \implies \inf_{P \in \mathcal{M}^+} \int c - \varphi \oplus \psi \, dP = 0$ . Otherwise, if there exists  $(x_0, y_0)$  such that  $c(x_0, y_0) - \varphi \oplus \psi(x_0, y_0) < 0$ , consider  $P = \lambda \delta_{(x_0, y_0)}$  and let  $\lambda \rightarrow \infty$  to get  $\inf_{P \in \mathcal{M}^+} \int c - \varphi \oplus \psi \, dP = -\infty$ . Thus (KP-primal) can be written as

$$\sup_{\substack{\varphi \in L_1(\mu) \\ \psi \in L_1(\nu) \\ \varphi \oplus \psi \leq c}} \int \varphi \, d\mu + \int \psi \, d\nu$$

which is exactly (KP-dual).  $\square$

## 2.2 $c$ -transforms

The constraint in (KP-dual) rewrites as  $\forall x, y \in \mathbb{R}^d, \psi(y) \leq c(x, y) - \varphi(x)$  or equivalently  $\forall y, \psi(y) \leq \inf_x [c(x, y) - \varphi(x)]$ . This motivates the definition of the  $c$ -transform of  $\varphi$  as

$$\varphi^c(y) := \inf_x c(x, y) - \varphi(x)$$

and the  $c$ -transform of  $\psi$  as

$$\psi^c(x) := \inf_y c(x, y) - \psi(y)$$

However,  $\varphi^c$  and  $\psi^c$  may not be integrable without additional hypotheses on  $c$ , so we assume as in Exercise 2.36 of [1] that there exist  $c_X \in L_1(\mu)$  and  $c_Y \in L_1(\nu)$  such that  $\forall x, y \in \mathbb{R}^d, c(x, y) \leq c_X(x) + c_Y(y)$ . With this assumption it is easily checked that  $\varphi^c$  is  $\nu$ -integrable and the pair  $(\varphi, \varphi^c)$  increases the objective function. The same argument applies to  $(\varphi^{cc}, \varphi^c)$ , so (KP-dual) rewrites as

$$\sup_{\varphi \in L_1(\mu)} \int \varphi \, d\mu + \int \varphi^c \, d\nu \quad (\text{KP-dual 2})$$

and

$$\sup_{\varphi \in L_1(\mu)} \int \varphi^{cc} \, d\mu + \int \varphi^c \, d\nu \quad (\text{KP-dual 3})$$

It is important to note that  $\varphi^{ccc} = \varphi^c$ .

*Proof.* Note that  $\varphi \leq \varphi^{cc}$ . Indeed for  $x, y \in \mathbb{R}^d$ ,

$$\begin{aligned} \varphi^c(y) &\leq c(x, y) - \varphi(x) \\ \implies \varphi(x) &\leq c(x, y) - \varphi^c(y) \quad \forall y \\ \implies \varphi(x) &\leq \inf_y c(x, y) - \varphi^c(y) \\ \implies \varphi(x) &\leq \varphi^{cc}(x) \end{aligned}$$

Next, for  $x, y \in \mathbb{R}^d$ ,

$$\begin{aligned} \varphi^{ccc}(y) &\leq c(x, y) - \varphi^{cc}(x) \\ \implies \varphi^{ccc}(y) &\leq c(x, y) - \varphi(x) \quad \forall x \\ \implies \varphi^{ccc}(y) &\leq \inf_x c(x, y) - \varphi(x) \\ \implies \varphi^{ccc}(y) &\leq \varphi^c(y) \end{aligned}$$

and

$$\begin{aligned} \varphi^{cc}(x) &\leq c(x, y) - \varphi^c(y) \\ \implies \varphi^c(y) &\leq c(x, y) - \varphi^{cc}(x) \quad \forall x \\ \implies \varphi^c(y) &\leq \inf_x c(x, y) - \varphi^{cc}(x) \\ \implies \varphi^c(y) &\leq \varphi^{ccc}(y) \end{aligned}$$

$\square$

Besides,  $\varphi$  is said to be  $c$ -concave if  $\varphi^{cc} = \varphi$ . In [1], Villani states that (KP-dual) is solved by a pair  $(\varphi, \varphi^c)$  where  $\varphi$  is  $c$ -concave. As a result, (KP-dual) may be also rewritten as

$$\sup_{\substack{\varphi \in L_1(\mu) \\ \varphi \text{ } c\text{-concave}}} \int \varphi d\mu + \int \varphi^c d\nu \quad (\text{KP-dual 2})$$

## 2.3 Wasserstein distances

## References

- [1] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.