Notes on Optimal Transport
ENSAE
3A

Gabriel Romon

Last updated: Thursday 7th February, 2019 at 09:12

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_θ 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 ν_{data} and f_θ are not absolutely continuous with respect to the same measure (the former is discrete), minimizing the Kullback-Leibler divergence between the two writes as

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

A weakness of likelihood maximization is its poor scalability to high-dimensional settings, which are commonplace. For instance, a 100×100 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 $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$ (e.g deconvolution networks). We define the pushforward measure $g_{\theta\sharp}\mu$ by $\forall B \in \mathcal{B}(\mathbb{R}^d), g_{\theta\sharp}\mu(B) := \mu(g_\theta \in B)$ and we look for θ such that $g_{\theta\sharp}\mu$ matches ν_{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 μ and ν be probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ a cost function. Monge's optimal transportation problem is the following:

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

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. $T_\sharp\mu = \nu$ 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 (μ, ν) 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:

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

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

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

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

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 π_1 and π_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 φ and ψ are indicators, and a standard limit argument shows that it remains 0 for arbitrary integrable φ and ψ. 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

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

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 $\lambda\to\infty$ 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). □

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 φ as

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

and the the c-transform of ψ as

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

However, φ^c and ψ^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 φ^c is ν-integrable and the pair (φ,φ^c) increases the objective function. The same argument applies to (φ^{cc},φ^c), so (KP-dual) rewrites as

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

and

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

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

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

□

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

$$\sup_{\substack{\varphi \in L_1(\mu) \\ \varphi\ c-\text{concave}}} \int \varphi d\mu + \int \varphi^c d\nu \tag{KP-dual 4}$$

The following result provides conditions under which the gap between Monge problem and its relaxation is 0.

Theorem 2.2. Let $\|\cdot\|$ a norm on $\mathbb{R}^d$ and $h : \mathbb{R}^d \to \mathbb{R}$ a strictly convex and superlinear function (meaning that $\frac{h(z)}{\|z\|} \xrightarrow[\|z\| \to \infty]{} \infty$ and Assumption (H2) of [2] holds). Let $c : (x, y) \mapsto h(x - y)$ and assume that μ is absolutely continuous with respect to Lebesgue measure. Then there is a unique solution to (KP-primal) and it has the form $(\text{Id}, T)_\sharp \mu$ where T is uniquely defined μ-almost everywhere by

$$\begin{cases} T_\sharp \mu = \nu \\ \exists \varphi\ c - \text{concave } / T = \text{Id} - \nabla h^* \circ \nabla \varphi \end{cases}$$

Remark 1.
• h^* denotes the Legendre transform of h. Since φ is c-concave, it is locally Lipschitz (see Proposition 2.43 in [1]) hence differentiable almost-everywhere (Rademacher's theorem), so its gradient makes sense.
• We proved earlier that (KP-primal) provides a lower bound for (MP). Given the form of the solution, it is clear that T solves (MP) and that (MP) and (KP-dual) share the same optimal value.
• Any $h : z \mapsto \|z\|_p^p$ with $p > 1$ satisfies the conditions of the theorem.
• With $h : z \mapsto \frac{1}{2}\|z\|_2^2$, $T(x) = x - \nabla\varphi(x) = \nabla(\frac{1}{2}\|\cdot\|_2^2 - \varphi)$. In this case, $h \in \mathcal{C}^{1,1}_{\text{loc}}(\mathbb{R}^d)$, hence φ is semi-concave (Proposition 2.43 in [1]) and under condition (iii) of Lemma 2.42 in [1], $\frac{1}{2}\|\cdot\|_2^2 - \varphi$ is convex, and we have therefore recovered Brenier's theorem.

2.3 Wasserstein distances

The Wasserstein distance creates a metric on the space of probability measures from a metric on the data space. Instead of $\mathbb{R}^d$, let us consider a separable complete metric space (X, D) with the cost function $c(x, y) = D(x, y)^p$ where $p \geq 1$:

$$W_p(\mu, \nu) := \left(\inf_{P \in \Pi(\mu,\nu)} \int\int D(x, y)^p dP(x, y) \right)^{1/p}$$

It is shown in [1] that W_p is a metric on the space $\mathcal{W}_p(X)$ of probability measures with finite moments of order p, i.e. such that there exists x_0 with $\int D(x_0, x)^p d\mu(x) < \infty$. Note that $\mathcal{W}_p(X)$ is the set of all probability measures when D is bounded (which happens when X is bounded for example).
The choice of p is of much importance in practice. While $p = 2$ is easier to deal with computationally, $p = 1$ yields special properties.

Proposition 2.1. If c is a metric D, then φ is c-concave iff $\text{Lip}\,\varphi \leq 1$. Moreover, $\text{Lip}\,\varphi \leq 1 \implies \varphi^c = -\varphi$.

Proof. $\implies$ Let us prove the following lemma: if $f, g : X \to \mathbb{R}$ are bounded below, then $|\inf f - \inf g| \leq \sup |f - g|$. Indeed for any $x \in X$, $\inf f - \sup |f - g| \leq f(x) - |f(x) - g(x)| \leq g(x)$, hence $\inf g \geq \inf f - \sup |f - g|$ and $\inf f - \inf g \leq \sup |f - g|$. Switching f with g finishes the proof.

Since φ is c-concave, $\varphi^{cc} = \varphi$, hence

$$\begin{aligned}
|\varphi(x) - \varphi(y)| &= |\inf_z [D(x,z) - \varphi^c(z)] - \inf_z [D(y,z) - \varphi^c(z)]| \\
&\leq \sup_z |D(x,z) - D(y,z)| \\
&\leq D(x,y)
\end{aligned}$$

$\Longleftarrow$ If $\operatorname{Lip}\varphi \leq 1$, for all $x, y \in X$,

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

Next, note that $(-\varphi)^c(y) \leq D(y,y) + \varphi(y) = \varphi(y)$, hence $(-\varphi)^c = \varphi$ and φ is c-concave. Moreover,

$$\begin{aligned}
&\varphi(z) - \varphi(x) \leq D(z,x) \\
&\implies -\varphi(x) \leq D(z,x) - \varphi(z) \quad \forall z \\
&\implies -\varphi(x) \leq \inf_z D(z,x) - \varphi(z) \\
&\implies -\varphi(x) \leq \varphi^c(x)
\end{aligned}$$

Finally, $\varphi^c(x) \leq D(x,x) - \varphi(x) = -\varphi(x)$, hence $\varphi^c = -\varphi$. □

This result relates W_1 to integral probability metrics since (KP-Dual 4) turns into

$$W_1(\mu,\nu) = \sup_{\varphi \ 1-\mathrm{Lip}} \int \varphi d\mu - \int \varphi d\nu$$

References

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

[2] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.