The Wasserstein distances

March 20, 2011

This document presents the proof of the main results we proved on Wasserstein distances themselves (and not on curves in the Wasserstein space). In particular, triangle inequality and characterization of the topology. These proof are not easy to be found in the same terms.

Definition of the distances and triangle inequality

First, for $\Omega \subset \mathbb{R}^n$ and $p \geq 1$, let us set

$$\mathcal{P}_p(\Omega) := \{\mu \in \mathcal{P}(\Omega) \, : \, \int |x|^p d\mu < +\infty\}.$$

This subset of $\mathcal{P}(\Omega)$ will be the space where we define our distances. Obviously, if Ω is bounded then $\mathcal{P}_p(\Omega) = \mathcal{P}(\Omega)$.

For $\mu, \nu \in \mathcal{P}_p(\Omega)$, let us define

$$W_p(\mu,\nu) := \inf \left\{ \int |x-y|^p d\gamma \, \gamma \in \Pi(\mu,\nu) \right\}^{1/p},$$

i.e. the $p-$th root of the minimal transport cost for the cost $|x-y|^p$. The assumption $\mu, \nu \in \mathcal{P}_p(\Omega)$ guarantees finiteness of this value, since $|x-y|^p \leq C(|x|^p + |y|^p)$ and hence $W_p(\mu,\nu)^p \leq C(\int |x|^p d\mu + \int |x|^p d\nu)$.

Notice that, due to Jensen inequality, since for any $\gamma \in \Pi(\mu,\nu)$ we have $\gamma(\Omega \times \Omega) = 1$, for $p \leq q$ we can infer

$$\left(\int |x-y|^p d\gamma \right)^{1/p} = ||x-y||_{L^p(\gamma)} \leq ||x-y||_{L^q(\gamma)} = \left(\int |x-y|^q d\gamma \right)^{1/q},$$

which implies $W_p(\mu,\nu) \leq W_q(\mu,\nu)$. In particular $W_1(\mu,\nu) \leq W_p(\mu,\nu)$ for every $p \geq 1$. We will not define here W_∞ (as a limit for $p \to \infty$, or, which is the same, as the minimal value of the supremal problem $\min_{\gamma \in \Pi(\mu,\nu)} ||x-y||_{L^\infty(\gamma)}$).

On the other hand, for bounded Ω an opposite inequality holds, since

$$\left(\int |x-y|^p d\gamma \right)^{1/p} \leq \operatorname{diam}(\Omega)^{\frac{p}{p-1}} \left(\int |x-y| \, d\gamma \right)^{1/p},$$

which implies $W_p(\mu,\nu) \leq C W_1(\mu,\nu)^{1/p}$, for $C = \operatorname{diam}(\Omega)^{p'}$ and $p' = \frac{p}{p-1}$.

Proposition 0.1. *The quantity W_p defined above is actually a distance over $\mathcal{P}_p(\Omega)$.*

Proof. First, let us notice that $W_p \geq 0$. Then, we also notice that $W_p(\mu, \nu) = 0$ implies, as a consequence that the minimum in the definition of W_p is attained, that there exists $\gamma \in \pi(\mu, \nu)$ such that $\int |x-y|^p d\gamma = 0$, which means that γ is concentrated on $\{x = y\}$. This implies $\mu = \nu$ since, for any test function ϕ we have

$$\int \phi \, d\mu = \int \phi(x) d\gamma = \int \phi(y) d\gamma = \int \phi \, d\nu.$$

We need now to prove the triangle inequality. For that, let us take μ, ρ and $\nu \in \mathcal{P}_p(\Omega)$, $\gamma^+ \in \Pi(\mu, \rho)$ and $\gamma^- \in \Pi(\mu, \rho)$. We can also choose $\gamma^\pm$ to be optimal. Let us use the Lemma 0.2 below to say that there exists a measure $\sigma \mathcal{P}(\Omega \times \Omega \times \Omega)$ such that $(\pi_{x,y})_\# \sigma = \gamma^+$ and $(\pi_{y,z})_\# \sigma = \gamma^-$, where $\pi_{x,y}$ and $\pi_{y,z}$ denote the projections on the two first and two last variables, respectively. Let us take $\gamma := (\pi_{x,z})_\# \sigma$. By composition of the projections, it is easy to see that $(\pi_x)_\# \gamma = (\pi_x)_\# \sigma = (\pi_x)_\# \gamma^+ = \mu$ and, analogously, $(\pi_z)_\# \gamma = \nu$. This means $\gamma \in \Pi(\mu, \nu)$ and

$$\begin{aligned} W_p(\mu, \nu) &\leq \left(\int |x-z|^p d\gamma \right)^{1/p} = \left(\int |x-z|^p d\sigma \right)^{1/p} = ||x-z||_{L^p(\sigma)} \\ &\leq ||x-y||_{L^p(\sigma)} + ||y-z||_{L^p(\sigma)} = \left(\int |x-z|^p d\sigma \right)^{1/p} + \left(\int |x-z|^p d\sigma \right)^{1/p} \\ &= \left(\int |x-z|^p d\gamma^+ \right)^{1/p} + \left(\int |x-z|^p d\gamma^- \right)^{1/p} = W_p(\mu, \rho) + W_p(\rho, \nu). \quad \square \end{aligned}$$

Lemma 0.2. *Given two measures $\gamma^+ \in \Pi(\mu, \rho)$ and $\gamma^- \in \Pi(\mu, \rho)$ there exists at least a measure $\sigma \mathcal{P}(\Omega \times \Omega \times \Omega)$ such that $(\pi_{x,y})_\# \sigma = \gamma^+$ and $(\pi_{y,z})_\# \sigma = \gamma^-$, where $\pi_{x,y}$ and $\pi_{y,z}$ denote the projections on the two first and two last variables, respectively.*

Proof. Start by taking γ^+ and disintegrate it w.r.t. the projection π_y. We get a family of measures $\gamma_y^+ \in \mathcal{P}(\Omega)$ (we can think of them as measures over Ω, instead of viewing them as measures over $\Omega \times \{y\} \subset \Omega \times \Omega$)/ They satisfy (and they are defined by)

$$\int_{\Omega \times \Omega} \phi(x, y) d\gamma^+(x, y) = \int_\Omega d\rho(y) \int_\Omega \phi(x, y) \, d\gamma_y^+(x),$$

for every measurable function ϕ of two variables. In the same way, one has a family of measures $\gamma_y^- \in \mathcal{P}(\Omega)$ such that for every ψ we have

$$\int_{\Omega \times \Omega} \psi(y, z) d\gamma^-(y, z) = \int_\Omega d\rho(y) \int_\Omega \psi(y, z) \, d\gamma_y^-(z).$$

For every y take now $\gamma_y^+ \otimes \gamma_y^-$, which is a measure over $\Omega \times \Omega$. Define σ through

$$\int_{\Omega^3} \zeta(x, y, z) d\sigma(x, y, z) := \int_\Omega d\rho(y) \int_{\Omega \times \Omega} \zeta(x, y, z) \, d\left(\gamma_y^+ \otimes \gamma_y^-\right)(x, z).$$

It is easy to check that, for ϕ depending only on x and y, we have

$$\int_{\Omega^3} \phi(x,y)d\sigma = \int_\Omega d\rho(y) \int_{\Omega\times\Omega} \phi(x,y)\, d\left(\gamma_y^+ \otimes \gamma_y^-\right)(x,z) = \int_\Omega d\rho(y) \int_\Omega \phi(x,y)\, d\gamma_y^+(x) = \int \phi\, d\gamma^+.$$

This proves $(\pi_{x,y})_\#\sigma = \gamma^+$ and the proof of $(\pi_{y,z})_\#\sigma = \gamma^-$ is completely analogous. $\square$

For the sake of completeness, we also give a proof of the triangle inequality which avoids using disintegrations. We first need the following lemma.

Lemma 0.3. *Given $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^n)$ and χ_ε any usual regularizing kernel in L^1 with $\int \chi_\varepsilon = 1$ and* $\mathrm{spt}(\chi_\varepsilon) \subset B(0,\varepsilon)$*, we have*

$$\lim_{\varepsilon\to 0} W_p(\mu * \chi_\varepsilon, \nu) = W_p(\mu,\nu).$$

Proof. Take an optimal transport plan $\gamma \in \Pi(\mu,\nu)$ and define a measure $\gamma_\varepsilon \in \Pi(\mu * \chi_\varepsilon, \nu)$ through

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} \psi(x,y)d\gamma_\varepsilon := \int_{\mathbb{R}^n\times\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(x-z,y)\chi_\varepsilon(z)dz\, d\gamma(x,y).$$

We need to check that its marginals are actually $\mu * \chi_\varepsilon$ and ν. For that just consider

$$\begin{aligned}\int_{\mathbb{R}^n\times\mathbb{R}^n} \psi(x)d\gamma_\varepsilon &= \int_{\mathbb{R}^n\times\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(x-z)\chi_\varepsilon(z)dz\, d\gamma(x,y) = \int_{\mathbb{R}^n} dz \int_{\mathbb{R}^n\times\mathbb{R}^n} \psi(x-z)d\gamma(x,y)\\ &= \int_{\mathbb{R}^n} dz \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(x-z)d\mu(x) = \int \psi\, d\mu * \chi_\varepsilon\end{aligned}$$

and, more easily

$$\int_{\mathbb{R}^n\times\mathbb{R}^n} \psi(y)d\gamma_\varepsilon = \int_{\mathbb{R}^n\times\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(y)\chi_\varepsilon(z)dz\, d\gamma(x,y) = \int_{\mathbb{R}^n\times\mathbb{R}^n} \psi(y)d\gamma(x,y) = \int \psi\, d\nu.$$

It is then easy to show that $\int |x-y|^p d\gamma_\varepsilon \to \int |x-y|^p d\gamma$, since

$$\left|\int |x-y|^p d\gamma_\varepsilon - \int |x-y|^p d\gamma\right| \le \int d\gamma(x,y) \left| |x-y|^p - \int |x-y-z|^p \chi_\varepsilon(z)dz\right| \le p\varepsilon \int d\gamma(x,y)(|x-y|+1)^{p-1}$$

(we use the fact that $|z| \le \varepsilon$ on $\mathrm{spt}(\chi_\varepsilon)$ and we roughly estimate $(a+\varepsilon)^p - a^p \le \varepsilon p(a+1)^{p-1}$ thanks to the mean value theorem (for $a \ge 0$ and $0 \le \varepsilon \le 1$). The last integral being finite since $\int |x-y|^p d\gamma < +\infty$, letting $\varepsilon \to 0$ we get

$$\limsup_{\varepsilon\to 0} W_p(\mu * \chi_\varepsilon, \nu)^p \le \limsup_{\varepsilon\to 0} \int |x-y|^p d\gamma_\varepsilon = \int |x-y|^p d\gamma.$$

This shows $\limsup_{\varepsilon\to 0} W_p(\mu * \chi_\varepsilon, \nu) \le W_p(\mu,\nu)$.

One cans also obtain the opposite inequality with the liminf in the following way. First fix a sequence $\varepsilon_k \to 0$ such that $\lim_k W_p(\mu*\chi_{\varepsilon_k},\nu) = \liminf_{\varepsilon\to 0} W_p(\mu*\chi_\varepsilon,\nu)$. Then extract a subsequence

ε_{k_j} so as to guarantee that the optimal transport plans $\gamma^{\varepsilon_{k_j}}$ sending $\mu * \chi_{\varepsilon_{k_j}}$ to ν have a weak limit γ^0 (see next section for disambiguations on the meaning of weak convergence). This weak limit must belong to $\Pi(\mu, \nu)$ (the fact that the marginals of γ_0 are μ and ν follows by the properties of composition with continuous functions of the weak convergence). Then we have

$$W_p(\mu,\nu)^p \leq \int |x-y|^p d\gamma^0 \leq \liminf_j \int |x-y|^p d\gamma^{\varepsilon_{k_j}} = \liminf_j W_p(\mu * \chi_{\varepsilon_{k_j}}, \nu)^p = \liminf_{\varepsilon \to 0} W_p(\mu * \chi_\varepsilon, \nu),$$

where the first inequality follows from the fact that γ^0 is not necessarily optimal but is admissible and the second by semicontinuity (since $|x-y|^p$ is a positive and continuous function, which is the increasing limit of positive, continuous and bounded functions). $\square$

Then, we can perform a proof of the triangle inequality based on the use of optimal transport maps.

Proposition 0.4. *Even if we refuse to use disintegrations, the triangle inequality is true for W_p.*

Proof. First consider the case where μ and ρ are absolutely continuous and ν is arbitrary. Let T be the optimal transport from μ to ρ and S from ρ to ν. Then $S \circ T$ is an admissible transport from μ to ν, since $(S \circ T)_\# \mu = S_\#(T_\# \mu) = S_\# \rho = \nu$. Then we have

$$W_p(\mu,\nu) \leq \left(\int |S(T(x)) - x|^p d\mu \right)^{1/p} = ||S \circ T - id||_{L^p(\mu)} \leq ||S \circ T - T||_{L^p(\mu)} + ||T - id||_{L^p(\mu)}.$$

Yet,

$$||S \circ T - T||_{L^p(\mu)} = \left(\int |S(T(x)) - T(x)|^p d\mu \right)^{1/p} = \left(\int |S(y) - y|^p d\rho \right)^{1/p} = W_p(\rho, \nu)$$

and $||T - id||_{L^p(\mu)} = W_p(\mu, \rho)$, hence

$$W_p(\mu,\nu) \leq W_p(\mu,\rho) + W_p(\rho,\nu).$$

This gives the proof when $\mu, \rho << \mathcal{L}^d$. If ρ is arbitrary, take now $\rho * \chi_\varepsilon$ instead, thus obtaining $W_p(\mu,\nu) \leq W_p(\mu, \rho * \chi_\varepsilon) + W_p(\rho * \chi_\varepsilon, \nu)$. By passing to the limit as $\varepsilon \to 0$ and using Lemma 0.3 the inequality follows for arbitrary ρ. Finally, μ may be taken arbitrary as well by considering now $\mu * \chi_\varepsilon$, with arbitrary ρ and ν and letting $\varepsilon \to 0$. $\square$

Topology induced by W_p

First of all, let us clarify that we often use the term "weak convergence", when speaking of probability measures, to denote the convergence in the duality with bounded continuous functions (which is often referred to as narrow convergence), and write $\mu_n \rightharpoonup \mu$ to say that μ_n converges in such a sense to μ. Notice also that, when both μ_n and μ are probability measures, this convergence coincides with the convergence in the duality with functions $\phi \in C_0(\Omega)$, vanishing at infinity. To convince of such a fact, we only need to show that if we take $\phi \in C_b(\Omega)$, $\mu_n, \mu \in \mathcal{P}(\Omega)$ and we suppose $\int \psi d\mu_n \to \int \psi d\mu$ for every $\psi \in C_0(\Omega)$, then we also have $\int \phi d\mu_n \to \int \phi d\mu$. If all the measures are probability, we can add for free a constant C to ϕ and, since ϕ is bounded, we can choose C so that $\phi + C \geq 0$. Hence $\phi + C$ is the sup of an increasing family of functions in C_0 (take $(\phi + C)\chi_n$, χ_n being an increasing family of cut-off functions with $\chi_n = 1$ on $B(0, n)$. Hence, by semicontinuity we have $\int(\phi + C)d\mu \leq \liminf_n \int(\phi + C)d\mu_n$, which implies $\int \phi d\mu \leq \liminf_n \int \phi d\mu_n$. If the same argument is performed with $-\phi$ we have te desired convergence of the integrals.

Once the weak convergence is understood, we can start from the following result.

Theorem 0.5. *If Ω is compact, then $\mu_n \rightharpoonup \mu$ if and only if $W_1(\mu_n, \mu) \to 0$.*

Proof. Let us recall the duality formula, which gives for arbitrary $\mu, \nu \in \mathcal{P}(\Omega)$

$$W_1(\mu, \nu) = \min \left\{ \int |x - y| \, d\gamma, \ \gamma \in \Pi(\mu, \nu) \right\} = \max \left\{ \int \phi \, d(\mu - \nu) \ : \ \phi \in \mathrm{Lip}_1 \right\}.$$

Let us start from a sequence μ_n such that $W_1(\mu_n, \mu) \to 0$. Thanks to the duality formula, for every $\phi \in \mathrm{Lip}_1(\Omega)$ we have $\int \phi \, d(\mu_n - \mu) \to 0$. By linearity, the same will be true for any Lipschitz function. By density, for any function in $C_b(\Omega)$. This shows that the Wasserstein convergence implies the weak convergence.

To prove the opposite implication, let us first fix a subsequence μ_{n_k} such that $\lim_k W_1(\mu_{n_k}, \mu) = \limsup_n W_1(\mu_n, \mu)$. For every k, pick a function $\phi_{n_k} \in Lip_1(\Omega)$ such that $\int \phi_{n_k} \, d(\mu_{n_k} - \mu) = W_1(\mu_{n_k}, \mu)$. Up to adding a constant, which does not affect the integral, we can suppose that ϕ_{n_k} all vanish on a same point, and they are hence uniformly bounded and equicontinuous. By Ascoli's theorem we can extract a sub-subsequence uniformly converging to a certain $\phi \in \mathrm{Lip}_1(\Omega)$. By replacing the original subsequence with this new one we can avoid relabeling. We have now

$$\begin{aligned} W_1(\mu_{n_k}, \mu) = \int \phi_{n_k} \, d(\mu_{n_k} - \mu) \quad &\leq \quad \int |\phi_{n_k} - \phi| \; d(\mu_{n_k} + \mu) + \int \phi \, d(\mu_{n_k} - \mu) \\ &\leq \quad 2||\phi_{n_k} - \phi||_{L^\infty} + \int \phi \, d(\mu_{n_k} - \mu) \to 0, \end{aligned}$$

where the first term goes to 0 by uniform convergence and the second by weak convergence. This shows that $\limsup_n W_1(\mu_n, \mu) \leq 0$ and concludes the proof. □

Theorem 0.6. *If Ω is compact and $p \geq 1$, then $\mu_n \rightharpoonup \mu$ if and only if $W_p(\mu_n, \mu) \to 0$.*

Proof. We have already proved this equivalence for $p = 1$. For the other values of p, just use the inequalities

$$W_1(\mu, \nu) \leq W_p(\mu, \nu) \leq C W_1(\mu, \nu)^{1/p},$$

that give the equivalence between the convergence for W_p and for W_1. □

We can now pass to the case of unbounded domains.

Theorem 0.7. *Consider any $\Omega \subset \mathbb{R}^d$ and $p \geq 1$, then $W_p(\mu_n, \mu) \to 0$ if and only if $\mu_n \rightharpoonup \mu$ and $\int |x|^p d\mu_n \to \int |x|^p d\mu$.*

Proof. Consider first a sequence $\mu_n \in \mathcal{P}_p(\Omega)$ which is converging to μ for the W_p distance. It is still true in this case that

$$\sup \left\{ \int \phi \, d(\mu_n - \mu) \, : \, \phi \in \mathrm{Lip}_1 \right\} \to 0,$$

which gives the weak convergence testing against any Lipschitz function. Notice that Lipschitz functions are dense (for the uniform convergence) in the space $C_0(\Omega)$ (while it is not necessarily the case for $C_b(\Omega)$) and that this is enough to prove $\mu_n \to \mu$.

To obtain the other condition, namely $\int |x|^p d\mu_n \to \int |x|^p d\mu$ (which is not a consequence of the weak convergence, since $|x|^p$ is not bounded), it is sufficient to notice that

$$\int |x|^p d\mu_n = W_p^p(\mu_n, \delta_0) \to W_p^p(\mu, \delta_0) = \int |x|^p d\mu.$$

We need now to prove the opposite implication. Consider a sequence a $\mu_n \rightharpoonup \mu$ satisfying also $\int |x|^p d\mu_n \to \int |x|^p d\mu$. Fix $R > 0$ and consider the function $\phi(x) := (|x| \wedge R)^p$, which is continuous and bounded. We have

$$\int (|x|^p - (|x| \wedge R)^p) \, d\mu_n = \int |x|^p d\mu_n - \int \phi \, d\mu_n \to \int |x|^p d\mu - \int \phi \, d\mu = \int (|x|^p - (|x| \wedge R)^p) \, d\mu.$$

Since $\int (|x|^p - (|x| \wedge R)^p) \, d\mu \leq \int_{B(0,R)^c} |x|^p d\mu$ it is possible to choose R so that

$$\int (|x|^p - (|x| \wedge R)^p) \, d\mu < \varepsilon/2$$

and hence one can also guarantee that $\int (|x|^p - (|x| \wedge R)^p) \, d\mu_n < \varepsilon$ for all n large enough.

We use now the inequality $(|x| - R)^p \leq |x|^p - R^p = |x|^p - (|x| \wedge R)^p$ which is valid for $|x| \geq R$ (see Lemma 0.8 below) to get

$$\int (|x| - R)^p d\mu_n < \varepsilon \text{ for } n \text{ large enough and } \quad \int (|x| - R)^p d\mu < \varepsilon.$$

Consider now $\pi_R : \mathbb{R}^d \to \overline{B(0, R)}$ defined as the projection over $\overline{B(0, R)}$. This map is well defined and continuous and is the identity on $\overline{B(0, R)}$. Moreover, for every $x \notin \overline{B(0, R)}$ we have $|x - \pi_R(x)| = |x| - R$. We can deduce

$$W_p(\mu, (\pi_R)_\# \mu) \leq \left(\int (|x| - R)^p d\mu \right)^{1/p} \leq \varepsilon^{1/p}, \quad W_p(\mu_n, (\pi_R)_\# \mu_n) \leq \left(\int (|x| - R)^p d\mu_n \right)^{1/p} \leq \varepsilon^{1/p}.$$

Notice also that, due to the usual composition of the weak convergence with continuous functions, from $\mu_n \rightharpoonup \mu$ we also infer $(\pi_R)_\# \mu_n \rightharpoonup (\pi_R)_\# \mu$. Yet, these measures are all concentrated on the compact set $\overline{B(0,R)}$ and here we can use the equivalence between weak convergence and W_p convergence. Hence, we get

$$\begin{aligned}\limsup_n W_p(\mu_n,\mu) &\leq \limsup_n \left(W_p(\mu_n,(\pi_R)_\#\mu_n) + W_p((\pi_R)_\#\mu_n,(\pi_R)_\#\mu) + W_p(\mu,(\pi_R)_\#\mu)\right)\\ &\leq 2\varepsilon^{1/p} + \lim_n W_p((\pi_R)_\#\mu_n,(\pi_R)_\#\mu) = 2\varepsilon^{1/p}.\end{aligned}$$

The parameter $\varepsilon > 0$ being arbitrary, we get $\limsup_n W_p(\mu_n,\mu) = 0$ and the proof is concluded. □

Lemma 0.8. *For $a, b \in \mathbb{R}_+$ and $p \geq 1$ we have $a^p + b^p \leq (a+b)^p$.*

Proof. Suppose without loss of generality that $a \geq b$. Then we can write $(a+b)^p = a^p + p\xi^{p-1}b$, for a point $\xi \in [a, a+b]$. Use now $p \geq 1$ and $\xi \geq a \geq b$ to get $(a+b)^p \geq a^p + b^p$. □