6

The Wasserstein distances

Assume, as before, that you are in charge of the transport of goods between producers and consumers, whose respective spatial distributions are modeled by probability measures. The farther producers and consumers are from each other, the more difficult will be your job, and you would like to summarize the degree of difficulty with just one quantity. For that purpose it is natural to consider, as in (5.27), the **optimal transport cost** between the two measures:

$$C(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y)\, d\pi(x, y), \tag{6.1}$$

where $c(x, y)$ is the cost for transporting one unit of mass from x to y. Here we do not care about the shape of the optimizer but only the *value* of this optimal cost.

One can think of (6.1) as a kind of distance between μ and ν, but in general it does not, strictly speaking, satisfy the axioms of a distance function. However, when the cost is defined in terms of a distance, it is easy to cook up a distance from (6.1):

Definition 6.1 (Wasserstein distances). *Let $(\mathcal{X}, d)$ be a Polish metric space, and let $p \in [1, \infty)$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by the formula*

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^p\, d\pi(x, y) \right)^{1/p} \tag{6.2}$$

$$= \inf \left\{ \Big[\mathbb{E}\, d(X, Y)^p \Big]^{\frac{1}{p}}, \quad \text{law}\,(X) = \mu, \quad \text{law}\,(Y) = \nu \right\}.$$

C. Villani, *Optimal Transport*, Grundlehren der mathematischen Wissenschaften 338,

Particular Case 6.2 (Kantorovich–Rubinstein distance). The distance W_1 is also commonly called the Kantorovich–Rubinstein distance (although it would be more proper to save the the terminology Kantorovich–Rubinstein for the *norm* which extends W_1; see the bibliographical notes).

Example 6.3. $W_p(\delta_x, \delta_y) = d(x, y)$. In this example, the distance does not depend on p; but this is not the rule.

At the present level of generality, W_p is still not a distance in the strict sense, because it might take the value $+\infty$; but otherwise it does satisfy the axioms of a distance, as will be checked right now.

Proof that W_p satisfies the axioms of a distance. First, it is clear that $W_p(\mu, \nu) = W_p(\nu, \mu)$.

Next, let μ_1, μ_2 and μ_3 be three probability measures on $\mathcal{X}$, and let (X_1, X_2) be an optimal coupling of (μ_1, μ_2) and (Z_2, Z_3) an optimal coupling of (μ_2, μ_3) (for the cost function $c = d^p$). By the Gluing Lemma (recalled in Chapter 1), there exist random variables (X_1', X_2', X_3') with $\text{law}\,(X_1', X_2') = \text{law}\,(X_1, X_2)$ and $\text{law}\,(X_2', X_3') = \text{law}\,(Z_2, Z_3)$. In particular, (X_1', X_3') is a coupling of (μ_1, μ_3), so

$$
\begin{aligned}
W_p(\mu_1, \mu_3) \le \big(\mathbb{E}\, d(X_1', X_3')^p\big)^{\frac{1}{p}} &\le \Big(\mathbb{E}\,\big(d(X_1', X_2') + d(X_2', X_3')\big)^p\Big)^{\frac{1}{p}} \\
&\le \Big(\mathbb{E}\, d(X_1', X_2')^p\Big)^{\frac{1}{p}} + \Big(\mathbb{E}\, d(X_2', X_3')^p\Big)^{\frac{1}{p}} \\
&= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3),
\end{aligned}
$$

where the inequality leading to the second line is an application of the Minkowski inequality in $L^p(\mathbb{P})$, and the last equality follows from the fact that (X_1', X_2') and (X_2', X_3') are optimal couplings. So W_p satisfies the triangle inequality.

Finally, assume that $W_p(\mu, \nu) = 0$; then there exists a transference plan which is entirely concentrated on the diagonal $(y = x)$ in $\mathcal{X} \times \mathcal{X}$. So $\nu = \text{Id}\,_{\#}\mu = \mu$. □

To complete the construction it is natural to restrict W_p to a subset of $P(\mathcal{X}) \times P(\mathcal{X})$ on which it takes finite values.

Definition 6.4 (Wasserstein space). *With the same conventions as in Definition 6.1, the Wasserstein space of order p is defined as*

$$P_p(\mathcal{X}) := \left\{ \mu \in P(\mathcal{X}); \qquad \int_{\mathcal{X}} d(x_0, x)^p \, \mu(dx) < +\infty \right\},$$

where $x_0 \in \mathcal{X}$ is arbitrary. This space does not depend on the choice of the point x_0. Then W_p defines a (finite) distance on $P_p(\mathcal{X})$.

In words, the Wasserstein space is the space of probability measures which have a *finite moment of order p*. In this course, it will always be equipped with the distance W_p.

Proof that W_p is finite on P_p. Let π be a transference plan between two elements μ and ν in $P_p(\mathcal{X})$. Then the inequality

$$d(x, y)^p \le 2^{p-1} \left[d(x, x_0)^p + d(x_0, y)^p \right]$$

shows that $d(x, y)^p$ is $\pi(dx\, dy)$-integrable as soon as $d(\,\cdot\,, x_0)^p$ is μ-integrable and $d(x_0, \,\cdot\,)^p$ is ν-integrable. □

Remark 6.5. Theorem 5.10(i) and Particular Case 5.4 together lead to the useful **duality formula for the Kantorovich–Rubinstein distance**: For any μ, ν in $P_1(\mathcal{X})$,

$$W_1(\mu, \nu) = \sup_{\|\psi\|_{\mathrm{Lip}} \le 1} \left\{ \int_{\mathcal{X}} \psi \, d\mu - \int_{\mathcal{X}} \psi \, d\nu \right\}. \tag{6.3}$$

Among many applications of this formula I shall just mention the following covariance inequality: if f is a probability density with respect to μ then

$$\left(\int f \, d\mu \right) \left(\int g \, d\mu \right) - \int (fg) \, d\mu \le \|g\|_{\mathrm{Lip}} \, W_1(f\,\mu, \mu).$$

Remark 6.6. A simple application of Hölder's inequality shows that

$$p \le q \Longrightarrow W_p \le W_q. \tag{6.4}$$

In particular, the Wasserstein distance of order 1, W_1, is the weakest of all. The most useful exponents in the Wasserstein distances are $p = 1$ and $p = 2$. As a general rule, the W_1 distance is more flexible and easier to bound, while the W_2 distance better reflects geometric features (at least for problems with a Riemannian flavor), and is better adapted when there is more structure; it also scales better with the dimension. Results in W_2 distance are usually stronger, and more difficult to establish, than results in W_1 distance.

Remark 6.7. On the other hand, under adequate regularity assumptions on the cost function and the probability measures, it is possible to control W_p in terms of W_q even for $q < p$; these reverse inequalities express a certain rigidity property of optimal transport maps which comes from c-cyclical monotonicity. See the bibliographical notes for more details.

Convergence in Wasserstein sense

Now we shall study a characterization of convergence in the Wasserstein space. The notation $\mu_k \longrightarrow \mu$ means that μ_k converges weakly to μ, i.e. $\int \varphi \, d\mu_k \to \int \varphi \, d\mu$ for any bounded continuous φ.

Definition 6.8 (Weak convergence in P_p). *Let $(\mathcal{X}, d)$ be a Polish space, and $p \in [1, \infty)$. Let $(\mu_k)_{k \in \mathbb{N}}$ be a sequence of probability measures in $P_p(X)$ and let μ be another element of $P_p(\mathcal{X})$. Then (μ_k) is said to converge weakly in $P_p(\mathcal{X})$ if any one of the following equivalent properties is satisfied for some (and then any) $x_0 \in \mathcal{X}$:*

(i) $\mu_k \longrightarrow \mu$ *and* $\displaystyle\int d(x_0, x)^p \, d\mu_k(x) \longrightarrow \int d(x_0, x)^p \, d\mu(x)$;

(ii) $\mu_k \longrightarrow \mu$ *and* $\displaystyle\limsup_{k \to \infty} \int d(x_0, x)^p \, d\mu_k(x) \leq \int d(x_0, x)^p \, d\mu(x)$;

(iii) $\mu_k \longrightarrow \mu$ *and* $\displaystyle\lim_{R \to \infty} \limsup_{k \to \infty} \int_{d(x_0, x) \geq R} d(x_0, x)^p \, d\mu_k(x) = 0$;

(iv) For all continuous functions φ with $|\varphi(x)| \leq C\big(1 + d(x_0, x)^p\big)$, $C \in \mathbb{R}$, one has

$$\int \varphi(x) \, d\mu_k(x) \longrightarrow \int \varphi(x) \, d\mu(x).$$

Theorem 6.9 (W_p metrizes P_p). *Let $(\mathcal{X}, d)$ be a Polish space, and $p \in [1, \infty)$; then the Wasserstein distance W_p metrizes the weak convergence in $P_p(\mathcal{X})$. In other words, if $(\mu_k)_{k \in \mathbb{N}}$ is a sequence of measures in $P_p(\mathcal{X})$ and μ is another measure in $P(\mathcal{X})$, then the statements*

$$\mu_k \text{ converges weakly in } P_p(\mathcal{X}) \text{ to } \mu$$

and

$$W_p(\mu_k, \mu) \longrightarrow 0$$

are equivalent.

Remark 6.10. As a consequence of Theorem 6.9, convergence in the p-Wasserstein space implies convergence of the moments of order p. There is a stronger statement that the map $\mu \longmapsto (\int d(x_0, x)^p \, \mu(dx))^{1/p}$ is 1-Lipschitz with respect to W_p; in the case of a locally compact length space, this will be proven in Proposition 7.29.

Below are two immediate corollaries of Theorem 6.9 (the first one results from the triangle inequality):

Corollary 6.11 (Continuity of W_p). *If $(\mathcal{X}, d)$ is a Polish space, and $p \in [1, \infty)$, then W_p is continuous on $P_p(\mathcal{X})$. More explicitly, if μ_k (resp. ν_k) converges to μ (resp. ν) weakly in $P_p(\mathcal{X})$ as $k \to \infty$, then*

$$W_p(\mu_k, \nu_k) \longrightarrow W_p(\mu, \nu).$$

Remark 6.12. On the contrary, if these convergences are only usual weak convergences, then one can only conclude that $W_p(\mu, \nu) \leq \liminf W_p(\mu_k, \nu_k)$: the Wasserstein distance is lower semicontinuous on $P(\mathcal{X})$ (just like the optimal transport cost C, for any lower semicontinuous nonnegative cost function c; recall the proof of Theorem 4.1).

Corollary 6.13 (Metrizability of the weak topology). *Let $(\mathcal{X}, d)$ be a Polish space. If $\widetilde{d}$ is a bounded distance inducing the same topology as d (such as $\widetilde{d} = d/(1 + d)$), then the convergence in Wasserstein sense for the distance $\widetilde{d}$ is equivalent to the usual weak convergence of probability measures in $P(\mathcal{X})$.*

Before starting the proof of Theorem 6.9, it will be good to make some more comments. The short version of that theorem is that *Wasserstein distances metrize weak convergence*. This sounds good, but after all, there are many ways to metrize weak convergence. Here is a list of some of the most popular ones, defined either in terms of measures μ, ν, or in terms of random variables X, Y with $\text{law}\,(X) = \mu$, $\text{law}\,(Y) = \nu$:

- the **Lévy–Prokhorov distance** (or just Prokhorov distance):

$$d_P(\mu, \nu) = \inf\Big\{\varepsilon > 0; \;\; \exists\, X, Y; \;\; \inf \mathbb{P}\left[d(X, Y) > \varepsilon\right] \leq \varepsilon\Big\}; \qquad (6.5)$$

- the **bounded Lipschitz distance** (also called Fortet–Mourier distance):

$$d_{bL}(\mu, \nu) = \sup\left\{\int \varphi \, d\mu - \int \varphi \, d\nu; \quad \|\varphi\|_\infty + \|\varphi\|_{\text{Lip}} \leq 1\right\}; \quad (6.6)$$

- the **weak-$*$ distance** (on a locally compact metric space):

$$d_{w*}(\mu,\nu) = \sum_{k\in\mathbb{N}} 2^{-k} \left| \int \varphi_k \, d\mu - \int \varphi_k \, d\nu \right|, \tag{6.7}$$

  where $(\varphi_k)_{k\in\mathbb{N}}$ is a dense sequence in $C_0(\mathcal{X})$;

- the **Toscani distance** (on $P_2(\mathbb{R}^n)$):

$$d_T(\mu,\nu) = \sup_{\xi\in\mathbb{R}^n\setminus\{0\}} \left(\frac{\left| \int e^{-ix\cdot\xi} \, d\mu(x) \; - \; \int e^{-ix\cdot\xi} \, d\nu(x) \right|}{|\xi|^2} \right). \tag{6.8}$$

  (Here I implicitly assume that μ, ν have the same mean, otherwise $d_T(\mu,\nu)$ would be infinite; one can also introduce variants of d_T by changing the exponent 2 in the denominator.)

So why bother with Wasserstein distances? There are several answers to that question:

1. Wasserstein distances are rather strong, especially in the way they take care of large distances in $\mathcal{X}$; this is a definite advantage over, for instance, the weak-$*$ distance (which in practice is so weak that I advise the reader to never use it). It is not so difficult to combine information on convergence in Wasserstein distance with some smoothness bound, in order to get convergence in stronger distances.
2. The definition of Wasserstein distances makes them convenient to use in problems where optimal transport is naturally involved, such as many problems coming from partial differential equations.
3. The Wasserstein distances have a rich duality; this is especially useful for $p = 1$, in view of (6.3) (compare with the definition of the bounded Lipschitz distance). Passing back and forth from the original to the dual definition is often technically convenient.
4. Being defined by an infimum, Wasserstein distances are often relatively easy to bound from above: The construction of *any* coupling between μ and ν yields a bound on the distance between μ and ν. In the same line of ideas, any C-Lipschitz mapping $f : \mathcal{X} \to \mathcal{X}'$ induces a C-Lipschitz mapping $P_1(\mathcal{X}) \to P_1(\mathcal{X}')$ defined by $\mu \longmapsto f_{\#}\mu$ (the proof is obvious).

5. Wasserstein distances incorporate a lot of the geometry of the space. For instance, the mapping $x \longmapsto \delta_x$ is an *isometric* embedding of $\mathcal{X}$ into $P_p(\mathcal{X})$; but there are much deeper links. This partly explains why $P_p(\mathcal{X})$ is often very well adapted to statements that combine weak convergence and geometry.

To prove Theorem 6.9 I shall use the following lemma, which has interest on its own and will be useful again later.

Lemma 6.14 (Cauchy sequences in W_p are tight). *Let $\mathcal{X}$ be a Polish space, let $p \geq 1$ and let $(\mu_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $(P_p(\mathcal{X}), W_p)$. Then (μ_k) is tight.*

The proof is not so obvious and one might skip it at first reading.

Proof of Lemma 6.14. Let $(\mu_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $P_p(\mathcal{X})$: This means that

$$W_p(\mu_k, \mu_\ell) \longrightarrow 0 \qquad \text{as } k, \ell \to \infty.$$

In particular,

$$\int d(x_0, x)^p \, d\mu_k(x) = W_p\big(\delta_{x_0}, \mu_k\big)^p \leq \Big[W_p(\delta_{x_0}, \mu_1) + W_p(\mu_1, \mu_k)\Big]^p$$

remains bounded as $k \to \infty$.

Since $W_p \geq W_1$, the sequence (μ_k) is also Cauchy in the W_1 sense. Let $\varepsilon > 0$ be given, and let $N \in \mathbb{N}$ be such that

$$k \geq N \Longrightarrow \ W_1(\mu_N, \mu_k) < \varepsilon^2. \tag{6.9}$$

Then for any $k \in \mathbb{N}$, there is $j \in \{1, \ldots, N\}$ such that $W_1(\mu_j, \mu_k) < \varepsilon^2$. (If $k \geq N$, this is (6.9); if $k < N$, just choose $j = k$.)

Since the finite set $\{\mu_1, \ldots, \mu_N\}$ is tight, there is a compact set K such that $\mu_j[\mathcal{X} \setminus K] < \varepsilon$ for all $j \in \{1, \ldots, N\}$. By compactness, K can be covered by a finite number of small balls: $K \subset B(x_1, \varepsilon) \cup \ldots \cup B(x_m, \varepsilon)$.

Now write

$$U := B(x_1, \varepsilon) \bigcup \ldots \bigcup B(x_m, \varepsilon);$$

$$U_\varepsilon := \Big\{x \in \mathcal{X};\ d(x, U) < \varepsilon\Big\} \ \subset \ \ B(x_1, 2\varepsilon) \bigcup \ldots \bigcup B(x_m, 2\varepsilon);$$

$$\phi(x) := \left(1 - \frac{d(x, U)}{\varepsilon}\right)_+.$$

Note that $1_U \le \phi \le 1_{U_\varepsilon}$ and ϕ is $(1/\varepsilon)$-Lipschitz. By using these bounds and the Kantorovich–Rubinstein duality (6.3), we find that for $j \le N$ and k arbitrary,

$$\begin{aligned}\mu_k[U_\varepsilon] &\ge \int \phi \, d\mu_k \\ &= \int \phi \, d\mu_j + \left(\int \phi \, d\mu_k - \int \phi \, d\mu_j \right) \\ &\ge \int \phi \, d\mu_j - \frac{W_1(\mu_k, \mu_j)}{\varepsilon} \\ &\ge \mu_j[U] - \frac{W_1(\mu_k, \mu_j)}{\varepsilon}.\end{aligned}$$

On the one hand, $\mu_j[U] \ge \mu_j[K] \ge 1 - \varepsilon$ if $j \le N$; on the other hand, for each k we can find $j = j(k)$ such that $W_1(\mu_k, \mu_j) \le \varepsilon^2$. So in fact

$$\mu_k[U_\varepsilon] \ge 1 - \varepsilon - \frac{\varepsilon^2}{\varepsilon} = 1 - 2\varepsilon.$$

At this point we have shown the following: For each $\varepsilon > 0$ there is a finite family $(x_i)_{1 \le i \le m}$ such that all measures μ_k give mass at least $1 - 2\varepsilon$ to the set $Z := \cup B(x_i, 2\varepsilon)$. The point is that Z might not be compact. There is a classical remedy: Repeat the reasoning with ε replaced by $2^{-(\ell+1)}\varepsilon$, $\ell \in \mathbb{N}$; so there will be $(x_i)_{1 \le i \le m(\ell)}$ such that

$$\mu_k \left[\mathcal{X} \setminus \bigcup_{1 \le i \le m(\ell)} B(x_i, 2^{-\ell}\varepsilon) \right] \le 2^{-\ell}\varepsilon.$$

Thus

$$\mu_k[\mathcal{X} \setminus S] \le \varepsilon,$$

where

$$S := \bigcap_{1 \le p \le \infty} \bigcup_{1 \le i \le m(p)} \overline{B(x_i, \varepsilon 2^{-p})}.$$

By construction, S can be covered by finitely many balls of radius δ, where δ is arbitrarily small (just choose ℓ large enough that $2^{-\ell}\varepsilon < \delta$, and then $\overline{B(x_i, 2^{-\ell}\varepsilon)}$ will be included in $B(x_i, \delta)$). Thus S is *totally bounded*, i.e. it can be covered by finitely many balls of arbitrarily small radius. It is also closed, as an intersection of finite unions of closed sets. Since $\mathcal{X}$ is a complete metric space, it follows from a classical result in topology that S is compact. This concludes the proof of Lemma 6.14. □

Proof of Theorem 6.9. Let $(\mu_k)_{k\in\mathbb{N}}$ be such that $\mu_k \to \mu$ in distance W_p; the goal is to show that μ_k converges to μ in $P_p(\mathcal{X})$. First, by Lemma 6.14, the sequence $(\mu_k)_{k\in\mathbb{N}}$ is tight, so there is a subsequence $(\mu_{k'})$ such that $\mu_{k'}$ converges weakly to some probability measure $\widetilde{\mu}$. Then by Lemma 4.3,

$$W_p(\widetilde{\mu}, \mu) \le \liminf_{k'\to\infty} W_p(\mu_{k'}, \mu) = 0.$$

So $\widetilde{\mu} = \mu$, and the whole sequence (μ_k) has to converge to μ. This only shows the weak convergence in the usual sense, not yet the convergence in $P_p(\mathcal{X})$.

For any $\varepsilon > 0$ there is a constant $C_\varepsilon > 0$ such that for all nonnegative real numbers a, b,

$$(a+b)^p \le (1+\varepsilon)\, a^p + C_\varepsilon\, b^p.$$

Combining this inequality with the usual triangle inequality, we see that whenever x_0, x and y are three points in X, one has

$$d(x_0, x)^p \le (1+\varepsilon)\, d(x_0, y)^p + C_\varepsilon\, d(x, y)^p. \tag{6.10}$$

Now let (μ_k) be a sequence in $P_p(\mathcal{X})$ such that $W_p(\mu_k, \mu) \longrightarrow 0$, and for each k, let π_k be an optimal transport plan between μ_k and μ. Integrating inequality (6.10) against π_k and using the marginal property, we find that

$$\int d(x_0, x)^p\, d\mu_k(x) \le (1+\varepsilon) \int d(x_0, y)^p\, d\mu(y) + C_\varepsilon \int d(x, y)^p\, d\pi_k(x, y).$$

But of course,

$$\int d(x, y)^p\, d\pi_k(x, y) = W_p(\mu_k, \mu)^p \xrightarrow[k\to\infty]{} 0;$$

therefore,

$$\limsup_{k\to\infty} \int d(x_0, x)^p\, d\mu_k(x) \le (1+\varepsilon) \int d(x_0, x)^p\, d\mu(x).$$

Letting $\varepsilon \to 0$, we see that Property (ii) of Definition 6.8 holds true; so μ_k does converge weakly in $P_p(\mathcal{X})$ to μ.

Conversely, assume that μ_k converges weakly in $P_p(\mathcal{X})$ to μ; and again, for each k, introduce an optimal transport plan π_k between μ_k and μ, so that

$$\int d(x,y)^p \, d\pi_k(x,y) \longrightarrow 0.$$

By Prokhorov's theorem, (μ_k) forms a tight sequence; also $\{\mu\}$ is tight. By Lemma 4.4, the sequence (π_k) is itself tight in $P(\mathcal{X} \times \mathcal{X})$. So, up to extraction of a subsequence, still denoted by (π_k), one may assume that

$$\pi_k \longrightarrow \pi \qquad \text{weakly in } P(\mathcal{X} \times \mathcal{X}).$$

Since each π_k is optimal, Theorem 5.20 guarantees that π is an optimal coupling of μ and μ, so this is the (completely trivial) coupling $\pi = (\mathrm{Id}\,, \mathrm{Id}\,)_\# \mu$ (in terms of random variables, $Y = X$). Since this is independent of the extracted subsequence, actually π is the limit of the whole sequence (π_k).

Now let $x_0 \in \mathcal{X}$ and $R > 0$. If $d(x,y) > R$, then the largest of the two numbers $d(x,x_0)$ and $d(x_0,y)$ has to be greater than $R/2$, and no less than $d(x,y)/2$. In a slightly pedantic form,

$$\begin{aligned} &1_{d(x,y) \geq R} \\ &\leq 1_{[d(x,x_0) \geq R/2 \text{ and } d(x,x_0) \geq d(x,y)/2]} + 1_{[d(x_0,y) \geq R/2 \text{ and } d(x_0,y) \geq d(x,y)/2]}. \end{aligned}$$

So, obviously

$$\begin{aligned} \big[d(x,y)^p - R^p\big]_+ &\leq d(x,y)^p \, 1_{[d(x,x_0) \geq R/2 \text{ and } d(x,x_0) \geq d(x,y)/2]} \\ &\quad + d(x,y)^p \, 1_{[d(x_0,y) \geq R/2 \text{ and } d(x_0,y) \geq d(x,y)/2]} \\ &\leq 2^p d(x,x_0)^p \, 1_{d(x,x_0) \geq R/2} + 2^p d(x_0,y)^p \, 1_{d(x_0,y) \geq R/2}. \end{aligned}$$

It follows that

$$\begin{aligned} W_p(\mu_k,\mu)^p &= \int d(x,y)^p \, d\pi_k(x,y) \\ &= \int \big[d(x,y) \wedge R\big]^p \, d\pi_k(x,y) + \int \big[d(x,y)^p - R^p\big]_+ \, d\pi_k(x,y) \\ &\leq \int \big[d(x,y) \wedge R\big]^p \, d\pi_k(x,y) + 2^p \int_{d(x,x_0) \geq R/2} d(x,x_0)^p \, d\pi_k(x,y) \\ &\qquad + 2^p \int_{d(x_0,y) > R/2} d(x_0,y)^p \, d\pi_k(x,y) \\ &= \int \big[d(x,y) \wedge R\big]^p \, d\pi_k(x,y) + 2^p \int_{d(x,x_0) \geq R/2} d(x,x_0)^p \, d\mu_k(x) \\ &\qquad + 2^p \int_{d(x_0,y) \geq R/2} d(x_0,y)^p \, d\mu(y). \end{aligned}$$

Since π_k converges weakly to π, the first term goes to 0 as $k \to \infty$. So

$$\begin{aligned}\limsup_{k\to\infty} W_p(\mu_k,\mu)^p \le & \lim_{R\to\infty} 2^p \limsup_{k\to\infty} \int_{d(x,x_0)\ge R/2} d(x,x_0)^p \, d\mu_k(x) \\ & + \lim_{R\to\infty} 2^p \limsup_{k\to\infty} \int_{d(x_0,y)\ge R/2} d(x_0,y)^p \, d\mu(y) \\ & = 0.\end{aligned}$$

This concludes the argument. □

Control by total variation

The total variation is a classical notion of distance between probability measures. There is, by the way, a classical probabilistic representation formula of the total variation:

$$\|\mu-\nu\|_{TV} \;=\; 2 \inf \mathbb{P}\,[X \neq Y], \tag{6.11}$$

where the infimum is over all couplings (X,Y) of (μ,ν); this identity can be seen as a very particular case of Kantorovich duality for the cost function $1_{x\neq y}$.

It seems natural that a control in Wasserstein distance should be weaker than a control in total variation. This is not completely true, because total variation does not take into account large distances. But one can control W_p by *weighted* total variation:

Theorem 6.15 (Wasserstein distance is controlled by weighted total variation). *Let μ and ν be two probability measures on a Polish space $(\mathcal{X},d)$. Let $p \in [1,\infty)$ and $x_0 \in \mathcal{X}$. Then*

$$W_p(\mu,\nu) \le 2^{\frac{1}{p'}} \left(\int d(x_0,x)^p \, d|\mu-\nu|(x) \right)^{\frac{1}{p}}, \qquad \frac{1}{p}+\frac{1}{p'} = 1. \tag{6.12}$$

Particular Case 6.16. In the case $p = 1$, if the diameter of $\mathcal{X}$ is bounded by D, this bound implies $W_1(\mu,\nu) \le D\,\|\mu-\nu\|_{TV}$.

Remark 6.17. The integral in the right-hand side of (6.12) can be interpreted as the Wasserstein distance W_1 for the particular cost function $[d(x_0,x)+d(x_0,y)]1_{x\neq y}$.

Proof of Theorem 6.15. Let π be the transference plan obtained by keeping fixed all the mass shared by μ and ν, and distributing the rest uniformly: this is

$$\pi = (\mathrm{Id}\,, \mathrm{Id}\,)_{\#}(\mu \wedge \nu) + \frac{1}{a}\,(\mu-\nu)_+ \otimes (\mu-\nu)_-,$$

where $\mu \wedge \nu = \mu - (\mu-\nu)_+$ and $a = (\mu-\nu)_-[X] = (\mu-\nu)_+[X]$. A more sloppy but probably more readable way to write π is

$$\pi(dx\,dy) = (\mu \wedge \nu)(dx)\,\delta_{y=x} + \frac{1}{a}\,(\mu-\nu)_+(dx)\,(\mu-\nu)_-(dy).$$

By using the definition of W_p, the definition of π, the triangle inequality for d, the elementary inequality $(A+B)^p \le 2^{p-1}(A^p + B^p)$, and the definition of a, we find that

$$\begin{aligned}
W_p(\mu,\nu)^p &\le \int d(x,y)^p\,d\pi(x,y)\\
&= \frac{1}{a}\int d(x,y)^p\,d(\mu-\nu)_+(x)\,d(\mu-\nu)_-(y)\\
&\le \frac{2^{p-1}}{a}\int \big[d(x,x_0)^p + d(x_0,y)^p\big]\,d(\mu-\nu)_+(x)\,d(\mu-\nu)_-(y)\\
&\le 2^{p-1}\left[\int d(x,x_0)^p\,d(\mu-\nu)_+(x) + \int d(x_0,y)^p\,d(\mu-\nu)_-(y)\right]\\
&= 2^{p-1}\int d(x,x_0)^p\,d\big[(\mu-\nu)_+ + (\mu-\nu)_-\big](x)\\
&= 2^{p-1}\int d(x,x_0)^p\,d|\mu-\nu|(x).
\end{aligned}$$

□

Topological properties of the Wasserstein space

The Wasserstein space $P_p(\mathcal{X})$ inherits several properties of the base space $\mathcal{X}$. Here is a first illustration:

Theorem 6.18 (Topology of the Wasserstein space). *Let $\mathcal{X}$ be a complete separable metric space and $p \in [1,\infty)$. Then the Wasserstein space $P_p(\mathcal{X})$, metrized by the Wasserstein distance W_p, is also a complete separable metric space. In short: The Wasserstein space over a*

Polish space is itself a Polish space. Moreover, any probability measure can be approximated by a sequence of probability measures with finite support.

Remark 6.19. If $\mathcal{X}$ is compact, then $P_p(\mathcal{X})$ is also compact; but if $\mathcal{X}$ is only locally compact, then $P_p(\mathcal{X})$ is *not* locally compact.

Proof of Theorem 6.18. The fact that $P_p(\mathcal{X})$ is a metric space was already explained, so let us turn to the proof of **separability**. Let $\mathcal{D}$ be a dense sequence in $\mathcal{X}$, and let $\mathcal{P}$ be the space of probability measures that can be written $\sum a_j \delta_{x_j}$, where the a_j are rational coefficients, and the x_j are finitely many elements in $\mathcal{D}$. It will turn out that $\mathcal{P}$ is dense in $P_p(\mathcal{X})$.

To prove this, let $\varepsilon > 0$ be given, and let x_0 be an arbitrary element of $\mathcal{D}$. If μ lies in $P_p(\mathcal{X})$, then there exists a compact set $K \subset \mathcal{X}$ such that

$$\int_{\mathcal{X}\setminus K} d(x_0, x)^p \, d\mu(x) \le \varepsilon^p.$$

Cover K by a finite family of balls $B(x_k, \varepsilon/2)$, $1 \le k \le N$, with centers $x_k \in \mathcal{D}$, and define

$$B'_k = B(x_k, \varepsilon) \setminus \bigcup_{j<k} B(x_j, \varepsilon).$$

Then all B'_k are disjoint and still cover K.

Define f on $\mathcal{X}$ by

$$f(B'_k \cap K) = \{x_k\}, \qquad f(\mathcal{X} \setminus K) = \{x_0\}.$$

Then, for any $x \in K$, $d(x, f(x)) \le \varepsilon$. So

$$\begin{aligned} \int d(x, f(x))^p \, d\mu(x) &\le \varepsilon^p \int_K d\mu(x) \; + \; \int_{\mathcal{X}\setminus K} d(x, x_0)^p \, d\mu(x) \\ &\le \varepsilon^p + \varepsilon^p = 2\,\varepsilon^p. \end{aligned}$$

Since $(\mathrm{Id}\,, f)$ is a coupling of μ and $f_\# \mu$, $W_p(\mu, f_\# \mu) \le 2\varepsilon^p$.

Of course, $f_\# \mu$ can be written as $\sum a_j \delta_{x_j}$, $0 \le j \le N$. This shows that μ might be approximated, with arbitrary precision, by a finite combination of Dirac masses. To conclude, it is sufficient to show that the coefficients a_j might be replaced by rational coefficients, up to a very small error in Wasserstein distance. By Theorem 6.15,

$$W_p\left(\sum_{j\le N} a_j\delta_{x_j},\ \sum_{j\le N} b_j\delta_{x_j}\right)\ \le\ 2^{\frac{1}{p'}}\left[\max_{k,\ell} d(x_k,x_\ell)\right]\sum_{j\le N}|a_j-b_j|^{\frac{1}{p}},$$

and obviously the latter quantity can be made as small as possible for some well-chosen rational coefficients b_j.

Finally, let us prove the **completeness**. Again let $(\mu_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $P_p(\mathcal{X})$. By Lemma 6.14, it admits a subsequence $(\mu_{k'})$ which converges weakly (in the usual sense) to some measure μ. Then,

$$\int d(x_0,x)^p\,d\mu(x)\le\liminf_{k'\to\infty}\int d(x_0,x)^p\,d\mu_{k'}(x)<+\infty,$$

so μ belongs to $P_p(\mathcal{X})$. Moreover, by lower semicontinuity of W_p,

$$W_p(\mu,\mu_{\ell'})\le\liminf_{k'\to\infty} W_p(\mu_{k'},\mu_{\ell'}),$$

so in particular

$$\limsup_{\ell'\to\infty} W_p(\mu,\mu_{\ell'})\le\limsup_{k',\ell'\to\infty} W_p(\mu_{k'},\mu_{\ell'})=0,$$

which means that $\mu_{\ell'}$ converges to μ in the W_p sense (and not just in the sense of weak convergence). Since (μ_k) is a Cauchy sequence with a converging subsequence, it follows by a classical argument that the whole sequence is converging. □

Bibliographical notes

The terminology of Wasserstein distance (apparently introduced by Dobrushin) is very questionable, since (a) these distances were discovered and rediscovered by several authors throughout the twentieth century, including (in chronological order) Gini [417, 418], Kantorovich [501], Wasserstein [803], Mallows [589] and Tanaka [776] (other contributors being Salvemini, Dall'Aglio, Hoeffding, Fréchet, Rubinstein, Ornstein, and maybe others); (b) the explicit definition of this distance is not so easy to find in Wasserstein's work; and (c) Wasserstein was only interested in the case $p=1$. By the way, also the spelling of Wasserstein is

doubtful: the original spelling was Vasershtein. (Similarly, Rubinstein was spelled Rubinshtein.) These issues are discussed in a historical note by Rüschendorf [720], who advocates the denomination of "minimal L^p-metric" instead of "Wasserstein distance". Also Vershik [808] tells about the discovery of the metric by Kantorovich and stands up in favor of the terminology "Kantorovich distance".

However, the terminology "Wasserstein distance" (or "Wasserstein metric") has been extremely successful: at the time of writing, about 30,000 occurrences can be found on the Internet. Nearly all recent papers relating optimal transport to partial differential equations, functional inequalities or Riemannian geometry (including my own works) use this convention. I will therefore stick to this by-now well-established terminology. After all, even if this convention is a bit unfair since it does not give credit to all contributors, not even to the most important of them (Kantorovich), at least it does give credit to somebody.

As I learnt from Bernot, terminological confusion was enhanced in the mid-nineties, when a group of researchers in image processing introduced the denomination of "Earth Mover's distance" [713] for the Wasserstein (Kantorovich–Rubinstein) W_1 distance. This terminology was very successful and rapidly spread by the high rate of growth of the engineering literature; it is already starting to compete with "Wasserstein distance", scoring more than 15,000 occurrences on Internet.

Gini considered the special case where the random variables are discrete and lie on the real line; like Mallows later, he was interested by applications in statistics (the "Gini distance" is often used to roughly quantify the inequalities of wealth or income distribution in a given population). Tanaka discovered applications to partial differential equations. Both Mallows and Tanaka worked with the case $p = 2$, while Gini was interested both in $p = 1$ and $p = 2$, and Hoeffding and Fréchet worked with general p (see for instance [381]). A useful source on the point of view of Kantorovich and Rubinstein is Vershik's review [809].

Kantorovich and Rubinstein [506] made the important discovery that the original Kantorovich distance (W_1 in my notation) can be extended into a *norm* on the set $M(\mathcal{X})$ of signed measures over a Polish space $\mathcal{X}$. It is common to call this extension the **Kantorovich–Rubinstein norm**, and by abuse of language I also used the denomination Kantorovich–Rubinstein metric for W_1. (It would be more proper to call it just the Kantorovich metric, but more or less everything in this subject should be called after Kantorovich.) This norm property is a

particular feature of the exponent $p = 1$, and should be taken seriously because it has strong implications in functional analysis. For one thing, the Kantorovich–Rubinstein norm provides an explicit isometric embedding of an arbitrary Polish space in a Banach space.

As pointed out to me by Vershik, the Kantorovich–Rubinstein norm on a metric space $(\mathcal{X}, d)$ can be intrinsically characterized as the maximal norm $\| \cdot \|$ on $M(\mathcal{X})$ which is "compatible" with the distance, in the sense that $\|\delta_x - \delta_y\| = d(x, y)$ for all $x, y \in \mathcal{X}$; this maximality property is a consequence of the duality formula.

Here are a few words about the other probability metrics mentioned in this chapter. The Toscani metric is useful in the theory of the Boltzmann equation, see [812, Section 4.2] and references quoted therein. Together with its variants, it is also handy for studying rates of convergence in the central limit theorem, or certain stable limit theorems [424]. The Lévy–Prokhorov metric appears in a number of textbooks, such as Dudley [318, p. 394]. For the taxonomy of probability metrics and their history, the unavoidable reference is the monograph by Rachev [695], which lists dozens and dozens of metrics together with their main properties and applications. (Many of them are variants, particular cases or extensions of the Wasserstein and Lévy–Prokhorov metrics.) The more recent set of notes by Carrillo and Toscani [216] also presents applications of various probability metrics to problems arising in partial differential equations (in particular the inelastic Boltzmann equation).

Here as in all the rest of this course, I only considered complete separable metric spaces. However, Wasserstein distances also make sense in noncomplete separable metric spaces: The case $p = 1$ was treated by Dudley [318, Lemma 11.8.3], while the general case was recently considered by Clement and Desch [237]. In this reference the triangular inequality is proven by approximation by countable spaces.

The equivalence between the four statements in Definition 6.8 is proven in [814, Theorem 7.12]. I borrowed the proof of Lemma 6.14 from Bolley [136]; and the scheme of proof of Theorem 6.9 from Ambrosio, Gigli and Savaré [30]. There are alternative proofs of Theorem 6.9 in the literature, for instance in [814, Chapter 7]. Similar convergence results had been obtained earlier by various authors, at least in particular cases, see e.g. [260, 468].

In dimension 1, Theorem 6.9 can be proven by simpler methods, and interpreted as a quantitative refinement of Skorokhod's representation theorem, as noticed in [795] or [814, Section 7.3].

The ∞-Wasserstein distance, $W_\infty = \lim_{p\to\infty} W_p$, does not fit in the setting considered in this chapter, in particular because the induced topology is quite stronger than the weak topology of measures. This distance however is useful in a surprising number of problems [208, 212, 222, 466, 617].

The representation formula (6.11) for the total variation distance is a particular case of Strassen's duality theorem, see for instance [814, Section 1.4]. Remark 6.17 is extracted from [427, comments following Remark VI.5].

Theorem 6.15 is a copy–paste from [814, Proposition 7.10], which itself was a slight adaptation of [696, Lemma 10.2.3]. Other upper bounds for the Wasserstein distances are available in the literature; see for instance [527] for the case of the Hamming distance on discrete product spaces.

Results of lower bounds for the Wasserstein distance (in terms of moments for instance) are not so common. One example is Proposition 7.29 in the next chapter. In the particular case of the 2-Wasserstein distance on a Hilbert space, there are lower bounds expressed in terms of moments and covariance matrices [258, 407].

In relation with the ∞-Wasserstein distance, Bouchitté, Jimenez and Rajesh [151] prove the following estimate: If Ω is a bounded Lipschitz open subset of $\mathbb{R}^n$, equipped with the usual Euclidean distance, $\mu(dx) = f(x)\,dx$ and $\nu(dy)$ are probability measures on $\overline{\Omega}$, and the density f is uniformly bounded below, then for any $p > 1$,

$$W_\infty(\mu,\nu)^{p+n} \le \left(\frac{C}{\inf f}\right) W_p(\mu,\nu)^p,$$

where $C = C(p,n,\Omega)$. As mentioned in Remark 6.7, this "converse" estimate is related to the fact that the optimal transport map for the cost function $|x-y|^p$ enjoys some monotonicity properties which make it very rigid, as we shall see again in Chapter 10. (As an analogy: the Sobolev norms $W^{1,p}$ are all topologically equivalent when applied to C-Lipschitz convex functions on a bounded domain.)

Theorem 6.18 belongs to folklore and has probably been proven many times; see for instance [310, Section 14]. Other arguments are due to Rachev [695, Section 6.3], and Ambrosio, Gigli and Savaré [30]. In the latter reference the proof is very simple but makes use of the deep Kolmogorov extension theorem. Here I followed a much more elementary argument due to Bolley [136].

The statement in Remark 6.19 is proven in [30, Remark 7.1.9].

In a Euclidean or Riemannian context, the Wasserstein distance W_2 between two very close measures, say $(1+h_1)\,\nu$ and $(1+h_2)\,\nu$ with h_1, h_2 very small, is approximately equal to the $H^{-1}(\nu)$-norm of $h_1 - h_2$; see [671, Section 7], [814, Section 7.6] or Exercise 22.20. (One may also take a look at [567, 569].) There is in fact a natural one-parameter family of distances interpolating between $H^{-1}(\nu)$ and W_2, defined by a variation on the Benamou–Brenier formula (7.34) (insert a factor $(d\mu_t/d\nu)^{1-\alpha}$, $0 \leq \alpha \leq 1$ in the integrand of (7.33); this construction is due to Dolbeault, Nazaret and Savaré [312]).

Applications of the Wasserstein distances are too numerous to be listed here; some of them will be encountered again in the sequel. In [150] Wasserstein distances are used to study the best approximation of a measure by a finite number of points. Various authors [700, 713] use them to compare color distributions in different images. These distances are classically used in statistics, limit theorems, and all kinds of problems involving approximation of probability measures [254, 256, 257, 282, 694, 696, 716]. Rio [704] derives sharp quantitative bounds in Wasserstein distance for the central limit theorem on the real line, and surveys the previous literature on this problem. Wasserstein distances are well adapted to study rates of fluctuations of empirical measures, see [695, Theorem 11.1.6], [696, Theorem 10.2.1], [498, Section 4.9], and the research papers [8, 307, 314, 315, 479, 771, 845]. (The most precise results are those in [307]: there it is shown that the average W_1 distance between two independent copies of the empirical measure behaves like $(\int \rho^{1-1/d})/N^{1-1/d}$, where N is the size of the samples, ρ the density of the common law of the random variables, and $d \geq 3$; the proofs are partly based on subadditivity, as in [150].) Quantitative Sanov-type theorems have been considered in [139, 742]. Wasserstein distances are also commonly used in statistical mechanics, most notably in the theory of propagation of chaos, or more generally the mean behavior of large particle systems [768] [757, Chapter 5]; the original idea seems to go back to Dobrushin [308, 309] and has been applied in a large number of problems, see for instance [81, 82, 221, 590, 624]. The original version of the Dobrushin–Shlosman uniqueness criterion [308, 311] in spin systems was expressed in terms of optimal transport distance, although this formulation was lost in most subsequent developments (I learnt this from Ollivier).

Wasserstein distances are also useful in the study of mixing and convergence for Markov chains; the original idea, based on a contraction property, seems to be due to Dobrushin [310], and has been rediscovered since then by various authors [231, 662, 679]. Tanaka proved that the W_2 distance is contracting along solutions of a certain class of Boltzmann equations [776, 777]; these results are reviewed in [814, Section 7.5] and have been generalized in [138, 214, 379, 590].

Wasserstein distances behave well with increasing dimension, and therefore have been successfully used in large or infinite dimension; for instance for the large-time behavior of stochastic partial differential equations [455, 458, 533, 605], or hydrodynamic limits of systems of particles [444].

In a Riemannian context, the W_2 distance is well-adapted to the study of Ricci curvature, in relation with diffusion equations; these themes will be considered again in Part II.

Here is a short list of some more surprising applications. Werner [836] suggested that the W_1 distance is well adapted to quantify some variants of the uncertainty principle in quantum physics. In a recent note, Melleray, Petrov and Vershik [625] use the properties of the Kantorovich–Rubinstein norm to study spaces which are "linearly rigid", in the sense that, roughly speaking, there is only one way to embed them in a Banach space. The beautiful text [809] by Vershik reviews further applications of the Kantorovich–Rubinstein distance to several original topics (towers of measures, Bernoulli automorphisms, classification of metric spaces); see also [808] and the older contribution [807] by the same author.