Chapter 2
The Wasserstein Space

The Kantorovich problem described in the previous chapter gives rise to a metric structure, the *Wasserstein distance*, in the space of probability measures $P(\mathscr{X})$ on a space $\mathscr{X}$. The resulting metric space, a subspace of $P(\mathscr{X})$, is commonly known as the *Wasserstein space* $\mathscr{W}$ (although, as Villani [125, pages 118–119] puts it, this terminology is "very questionable"; see also Bobkov and Ledoux [25, page 4]). In Chap. 4, we shall see that this metric is in a sense canonical when dealing with warpings, that is, deformations of the space $\mathscr{X}$ (for example, in Theorem 4.2.4). In this chapter, we give the fundamental properties of the Wasserstein space. After some basic definitions, we describe the topological properties of that space in Sect. 2.2. It is then explained in Sect. 2.3 how $\mathscr{W}$ can be endowed with a sort of infinite-dimensional Riemannian structure. Measurability issues are dealt with in the somewhat technical Sect. 2.4.

2.1 Definition, Notation, and Basic Properties

Let $\mathscr{X}$ be a separable Banach space. The *p-Wasserstein space* on $\mathscr{X}$ is defined as

$$\mathscr{W}_p(\mathscr{X}) = \left\{ \mu \in P(\mathscr{X}) : \int_{\mathscr{X}} \|x\|^p \, \mathrm{d}\mu(x) < \infty \right\}, \qquad p \geq 1.$$

We will sometimes abbreviate and write simply $\mathscr{W}_p$ instead of $\mathscr{W}_p(\mathscr{X})$.

Recall that if $\mu, \nu \in P(\mathscr{X})$, then $\Pi(\mu, \nu)$ is defined to be the set of measures $\pi \in P(\mathscr{X}^2)$ having μ and ν as marginals in the sense of (1.2). The *p-Wasserstein distance* between μ and ν is defined as the minimal total transportation cost between

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-3-030-38438-8_2) contains supplementary material.

V. M. Panaretos, Y. Zemel, *An Invitation to Statistics in Wasserstein Space*,
SpringerBriefs in Probability and Mathematical Statistics,
https://doi.org/10.1007/978-3-030-38438-8_2

μ and ν in the Kantorovich problem with respect to the cost function $c_p(x,y) = \|x-y\|^p$:

$$W_p(\mu,\nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} C_p(\pi)\right)^{1/p} = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} \|x_1-x_2\|^p \,\mathrm{d}\pi(x_1,x_2)\right)^{1/p}.$$

The Wasserstein distance between μ and ν is finite when both measures are in $\mathscr{W}_p(\mathscr{X})$, because

$$\|x_1-x_2\|^p \le 2^p\|x_1\|^p + 2^p\|x_2\|^p.$$

Thus, W_p is finite on $[\mathscr{W}_p(\mathscr{X})]^2 = \mathscr{W}_p(\mathscr{X})\times\mathscr{W}_p(\mathscr{X})$; it is nonnegative and symmetric and it is easy to see that $W_p(\mu,\nu)=0$ if and only if $\mu=\nu$. A proof that W_p is a metric (satisfies the triangle inequality) on $\mathscr{W}_p$ can be found in Villani [124, Chapter 7].

The aforementioned setting is by no means the most general one can consider. Firstly, one can define W_p and $\mathscr{W}_p$ for $0<p<1$ by removing the power $1/p$ from the infimum and the limit case $p=0$ yields the total variation distance. Another limit case can be defined as $W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$. Moreover, W_p and $\mathscr{W}_p$ can be defined whenever $\mathscr{X}$ is a complete and separable metric space (or even only separable; see Clément and Desch [36]): one fixes some x_0 in $\mathscr{X}$ and replaces $\|x\|$ by $d(x,x_0)$. Although the topological properties below still hold at that level of generality (except when $p=0$ or $p=\infty$), for the sake of simplifying the notation we restrict the discussion to Banach spaces. It will always be assumed without explicit mention that $1\le p<\infty$.

The space $\mathscr{W}_p(\mathscr{X})$ is defined as the collection of measures μ such that $W_p(\mu,\delta_0)<\infty$ with δ_x being a Dirac measure at x. Of course, $W_p(\mu,\nu)$ can be finite even if $\mu,\nu\notin\mathscr{W}_p(\mathscr{X})$. But if $\mu\in\mathscr{W}_p(\mathscr{X})$ and $\nu\notin\mathscr{W}_p(\mathscr{X})$, then $W_p(\mu,\nu)$ is always infinite. This can be seen from the triangle inequality

$$\infty = W_p(\nu,\delta_0) \le W_p(\mu,\delta_0) + W_p(\mu,\nu).$$

In the sequel, we shall almost exclusively deal with measures in $\mathscr{W}_p(\mathscr{X})$.

The Wasserstein spaces are ordered in the sense that if $q\ge p$, then $\mathscr{W}_q(\mathscr{X})\subseteq\mathscr{W}_p(\mathscr{X})$. This property extends to the distances in the form:

$$q\ge p\ge 1 \quad\Longrightarrow\quad W_q(\mu,\nu)\ge W_p(\mu,\nu). \tag{2.1}$$

To see this, let $\pi\in\Pi(\mu,\nu)$ be optimal with respect to q. Jensen's inequality for the convex function $z\mapsto z^{q/p}$ gives

$$W_q^q(\mu,\nu) = \int_{\mathscr{X}^2}\|x-y\|^q\,\mathrm{d}\pi(x,y) \ge \left(\int_{\mathscr{X}^2}\|x-y\|^p\,\mathrm{d}\pi(x,y)\right)^{q/p} \ge W_p^q(\mu,\nu).$$

The converse of (2.1) fails to hold in general, since it is possible that W_p be finite while W_q is infinite. A converse can be established, however, if μ and ν are bounded:

$$q \geq p \geq 1, \quad \mu(K) = \nu(K) = 1 \quad \Longrightarrow \quad W_q(\mu,\nu) \leq W_p^{p/q}(\mu,\nu) \left(\sup_{x,y \in K} \|x-y\| \right)^{1-p/q}. \tag{2.2}$$

Indeed, if we denote the supremum by d_K and let π be now optimal with respect to p, then $\pi(K \times K) = 1$ and

$$W_q^q(\mu,\nu) \leq \int_{K^2} \|x-y\|^q \, \mathrm{d}\pi(x,y) \leq d_K^{q-p} \int_{K^2} \|x-y\|^p \, \mathrm{d}\pi(x,y) = d_K^{q-p} W_p^p(\mu,\nu).$$

Another useful property of the Wasserstein distance is the upper bound

$$\mathscr{W}_p(\mathbf{t}\#\mu, \mathbf{s}\#\mu) \leq \left(\int_{\mathscr{X}} \|\mathbf{t}(x) - \mathbf{s}(x)\|^p \, \mathrm{d}\mu(x) \right)^{1/p} = \| \, \|\mathbf{t}-\mathbf{s}\|_{\mathscr{X}} \, \|_{L_p(\mu)} \tag{2.3}$$

for any pair of measurable functions $\mathbf{t}, \mathbf{s} : \mathscr{X} \to \mathscr{X}$. Situations where this inequality holds as equality and $\mathbf{t}$ and $\mathbf{s}$ are optimal maps are related to *compatibility* of the measures μ, $\nu = \mathbf{t}\#\mu$ and $\rho = \mathbf{s}\#\mu$ (see Sect. 2.3.2) and will be of conceptual importance in the context of Fréchet means (see Sect. 3.1).

We also recall the notation $B_R(x_0) = \{x : \|x - x_0\| < R\}$ and $\overline{B}_R(x_0) = \{x : \|x - x_0\| \leq R\}$ for open and closed balls in $\mathscr{X}$.

2.2 Topological Properties

2.2.1 Convergence, Compact Subsets

The topology of a space is determined by the collection of its closed sets. Since $\mathscr{W}_p(\mathscr{X})$ is a metric space, whether a set is closed or not depends on which sequences in $\mathscr{W}_p(\mathscr{X})$ converge. The following characterisation from Villani [124, Theorem 7.12] will be very useful.

Theorem 2.2.1 (Convergence in Wasserstein Space) *Let $\mu, \mu_n \in \mathscr{W}_p(\mathscr{X})$. Then the following are equivalent:*

1. *$W_p(\mu_n, \mu) \to 0$ as $n \to \infty$;*
2. *$\mu_n \to \mu$ weakly and $\int_{\mathscr{X}} \|x\|^p \, \mathrm{d}\mu_n(x) \to \int_{\mathscr{X}} \|x\|^p \, \mathrm{d}\mu(x)$;*
3. *$\mu_n \to \mu$ weakly and*

$$\sup_n \int_{\{x : \|x\| > R\}} \|x\|^p \, \mathrm{d}\mu_n(x) \to 0, \qquad R \to \infty; \tag{2.4}$$

4. *for any $C > 0$ and any continuous $f : \mathscr{X} \to \mathbb{R}$ such that $|f(x)| \leq C(1 + \|x\|^p)$ for all x,*

$$\int_{\mathscr{X}} f(x) \, \mathrm{d}\mu_n(x) \to \int_{\mathscr{X}} f(x) \, \mathrm{d}\mu(x).$$

5. *(Le Gouic and Loubes [87, Lemma 14])* $\mu_n \to \mu$ *weakly and there exists* $\nu \in \mathscr{W}_p(\mathscr{X})$ *such that* $W_p(\mu_n,\nu) \to W_p(\mu,\nu)$.

Consequently, the Wasserstein topology is finer than the weak topology induced on $\mathscr{W}_p(\mathscr{X})$ from $P(\mathscr{X})$. Indeed, let $\mathscr{A} \subseteq \mathscr{W}_p(\mathscr{X})$ be weakly closed. If $\mu_n \in \mathscr{A}$ converge to μ in $\mathscr{W}_p(\mathscr{X})$, then $\mu_n \to \mu$ weakly, so $\mu \in \mathscr{A}$. In other words, the Wasserstein topology has more closed sets than the induced weak topology. Moreover, each $\mathscr{W}_p(\mathscr{X})$ is a weakly closed subset of $P(\mathscr{X})$ by the same arguments that lead to (1.3). In view of Theorem 2.2.1, a common strategy to establish Wasserstein convergence is to first show tightness and obtain weak convergence, hence a candidate limit, and then show that the stronger Wasserstein convergence actually holds. In some situations, the last part is automatic:

Corollary 2.2.2 *Let* $K \subset \mathscr{X}$ *be a bounded set and suppose that* $\mu_n(K) = 1$ *for all* $n \geq 1$. *Then* $W_p(\mu_n,\mu) \to 0$ *if and only if* $\mu_n \to \mu$ *weakly.*

Proof. This is immediate from (2.4).

The fact that convergence in $\mathscr{W}_p$ is stronger than weak convergence is exemplified in the following result. If $\mu_n \to \mu$ and $\nu_n \to \nu$ in $\mathscr{W}_p(\mathscr{X})$, then it is obvious that $W_p(\mu_n,\nu_n) \to W_p(\mu,\nu)$. But if the convergence is only weak, then the Wasserstein distance is still lower semicontinuous:

$$\liminf_{n\to\infty} W_p(\mu_n,\nu_n) \geq W_p(\mu,\nu). \tag{2.5}$$

This follows from Theorem 1.7.2 and (1.3).

Before giving some examples, it will be convenient to formulate Theorem 2.2.1 in probabilistic terms. Let X, X_n be random elements on $\mathscr{X}$ with laws $\mu, \mu_n \in \mathscr{W}_p(\mathscr{X})$. Assume without loss of generality that X, X_n are defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and write $W_p(X_n, X)$ to denote $W_p(\mu_n,\mu)$. Then $W_p(X_n, X) \to 0$ if and only if $X_n \to X$ weakly and $\mathbb{E}\|X_n\|^p \to \mathbb{E}\|X\|^p$.

An early example of the use of Wasserstein metric in statistics is due to Bickel and Freedman [21]. Let X_n be independent and identically distributed random variables with mean zero and variance 1 and let Z be a standard normal random variable. Then $Z_n = \sum_{i=1}^n X_i/\sqrt{n}$ converge weakly to Z by the central limit theorem. But $\mathbb{E}Z_n^2 = 1 = \mathbb{E}Z^2$, so $W_2(Z_n, Z) \to 0$. Let Z_n^* be a bootstrapped version of Z_n constructed by resampling the X_n's. If $W_2(Z_n^*, Z_n) \to 0$, then $W_2(Z_n^*, Z) \to 0$ and in particular Z_n^* has the same asymptotic distribution as Z_n.

Another consequence of Theorem 2.2.1 is that (in the presence of weak convergence) convergence of moments automatically yields convergence of smaller moments (there are, however, more elementary ways to see this). In the previous example, for instance, one can also conclude that $\mathbb{E}|Z_n|^p \to \mathbb{E}|Z|^p$ for any $p \leq 2$ by the last condition of the theorem. If in addition $\mathbb{E}X_1^4 < \infty$, then

$$\mathbb{E}Z_n^4 = 3 - \frac{3}{n} + \frac{\mathbb{E}X_1^4}{n} \to 3 = \mathbb{E}Z^4$$

(see Durrett [49, Theorem 2.3.5]) so $W_4(Z_n, Z) \to 0$ and all moments up to order 4 converge.

Condition (2.4) is called *uniform integrability* of the function $x \mapsto \|x\|^p$ with respect to the collection (μ_n). Of course, it holds for a single measure $\mu \in \mathscr{W}_p(\mathscr{X})$ by the dominated convergence theorem. This condition allows us to characterise compact sets in the Wasserstein space. One should beware that when $\mathscr{X}$ is infinite-dimensional, (2.4) alone is not sufficient in order to conclude that μ_n has a convergent subsequence: take μ_n to be Dirac measures at e_n with (e_n) an orthonormal basis of a Hilbert space $\mathscr{X}$ (or any sequence with $\|e_n\| = 1$ that has no convergent subsequence, if $\mathscr{X}$ is a Banach space). The uniform integrability (2.4) must be accompanied with tightness, which is a consequence of (2.4) only when $\mathscr{X} = \mathbb{R}^d$.

Proposition 2.2.3 (Compact Sets in $\mathscr{W}_p$) *A weakly tight set $\mathscr{K} \subseteq \mathscr{W}_p$ is Wasserstein-tight (has a compact closure in $\mathscr{W}_p$) if and only if*

$$\sup_{\mu \in \mathscr{K}} \int_{\{x : \|x\| > R\}} \|x\|^p \, \mathrm{d}\mu(x) \to 0, \qquad R \to \infty. \tag{2.6}$$

Moreover, (2.6) is equivalent to the existence of a monotonically divergent function $g : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\sup_{\mu \in \mathscr{K}} \int_{\mathscr{X}} \|x\|^p g(\|x\|) \, \mathrm{d}\mu(x) < \infty.$$

The proof is on page 41 of the supplement.

Remark 2.2.4 *For any sequence (μ_n) in $\mathscr{W}_p$ (tight or not) there exists a monotonically divergent g with $\int_{\mathscr{X}} \|x\|^p g(\|x\|) \, \mathrm{d}\mu_n(x) < \infty$ for all n.*

Corollary 2.2.5 (Measures with Common Support) *Let $K \subseteq \mathscr{X}$ be a compact set. Then*

$$\mathscr{K} = \mathscr{W}_p(K) = \{\mu \in P(\mathscr{X}) : \mu(K) = 1\} \subseteq \mathscr{W}_p(\mathscr{X})$$

is compact.

Proof. This is immediate, since $\mathscr{K}$ is weakly tight and the supremum in (2.6) vanishes when R is larger than the finite quantity $\sup_{x \in K} \|x\|$. Finally, K is closed, so $\mathscr{K}$ is weakly closed, hence Wasserstein closed, by the portmanteau Lemma 1.7.1.

For future reference, we give another consequence of uniform integrability, called *uniform absolute continuity*

$$\forall \varepsilon \; \exists \delta \; \forall n \; \forall A \subseteq \mathscr{X} \text{ Borel} : \qquad \mu_n(A) \le \delta \quad \Longrightarrow \quad \int_A \|x\|^p \, \mathrm{d}\mu_n(x) < \varepsilon. \tag{2.7}$$

To show that (2.4) implies (2.7), let $\varepsilon > 0$, choose $R = R_\varepsilon > 0$ such that the supremum in (2.4) is smaller than $\varepsilon/2$, and set $\delta = \varepsilon/(2R^p)$. If $\mu_n(A) \le \delta$, then

$$\int_A \|x\|^p \, \mathrm{d}\mu_n(x) \le \int_{A\cap\overline{B}_R(0)} \|x\|^p \, \mathrm{d}\mu_n(x) + \int_{A\setminus\overline{B}_R(0)} \|x\|^p \, \mathrm{d}\mu_n(x) < \delta R^p + \varepsilon/2 \le \varepsilon.$$

2.2.2 Dense Subsets and Completeness

If we identify a measure $\mu \in \mathscr{W}_p(\mathscr{X})$ with a random variable X (having distribution μ), then X has a finite p-th moment in the sense that the real-valued random variable $\|X\|$ is in L_p. In view of that, it should not come as a surprise that $\mathscr{W}_p(\mathscr{X})$ enjoys topological properties similar to L_p spaces. In this subsection, we give some examples of useful dense subsets of $\mathscr{W}_p(\mathscr{X})$ and then "show" that like $\mathscr{X}$ itself, it is a complete separable metric space. In the next subsection, we describe some of the negative properties that $\mathscr{W}_p(\mathscr{X})$ has, again in similarity with L_p spaces.

We first show that $\mathscr{W}_p(\mathscr{X})$ is separable. The core idea of the proof is the feasibility of approximating any measure with discrete measures as follows.

Let μ be a probability measure on $\mathscr{X}$, and let $X_1, X_2, \dots$ be a sequence of independent random elements in $\mathscr{X}$ with probability distribution μ. Then the *empirical measure* μ_n is defined as the random measure $(1/n)\sum_{i=1}^n \delta\{X_i\}$. The law of large numbers shows that for any (measurable) bounded or nonnegative $f : \mathscr{X} \to \mathbb{R}$, almost surely

$$\int_{\mathscr{X}} f(x) \, \mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^n f(X_i) \to \mathbb{E} f(X_1) = \int_{\mathscr{X}} f(x) \, \mathrm{d}\mu(x).$$

In particular when $f(x) = \|x\|^p$, we obtain convergence of moments of order p. Hence by Theorem 2.2.1, if $\mu \in \mathscr{W}_p(\mathscr{X})$, then $\mu_n \to \mu$ in $\mathscr{W}_p(\mathscr{X})$ if and only if $\mu_n \to \mu$ weakly. We know that integrals of bounded functions converge with probability one, but the null set where convergence fails may depend on the chosen function and there are uncountably many such functions. When $\mathscr{X} = \mathbb{R}^d$, by the portmanteau Lemma 1.7.1 we can replace the collection $C_b(\mathscr{X})$ by indicator functions of rectangles of the form $(-\infty, a_1] \times \cdots \times (-\infty, a_d]$ for $a = (a_1, \dots, a_d) \in \mathbb{R}^d$. It turns out that the countable collection provided by rational vectors a suffices (see the proof of Theorem 4.4.1 where this is done in a more complicated setting). For more general spaces $\mathscr{X}$, we need to find another countable collection $\{f_j\}$ such that convergence of the integrals of f_j for all j suffices for weak convergence. Such a collection exists, by using bounded Lipschitz functions (Dudley [47, Theorem 11.4.1]); an alternative construction can be found in Ambrosio et al. [12, Section 5.1]. Thus:

Proposition 2.2.6 (Empirical Measures in $\mathscr{W}_p$) *For any $\mu \in P(\mathscr{X})$ and the corresponding sequence of empirical measures μ_n, $W_p(\mu_n, \mu) \to 0$ almost surely if and only if $\mu \in \mathscr{W}_p(\mathscr{X})$.*

Indeed, if $\mu \notin \mathscr{W}_p(\mathscr{X})$, then $W_p(\mu_n, \mu)$ is infinite for all n, since μ_n is compactly supported, hence in $\mathscr{W}_p(\mathscr{X})$.

Proposition 2.2.6 is the basis for constructing dense subsets of the Wasserstein space.

Theorem 2.2.7 (Dense Subsets of $\mathscr{W}_p$) *The following collections of measures are dense in $\mathscr{W}_p(\mathscr{X})$:*

1. *finitely supported measures with rational weights;*
2. *compactly supported measures;*
3. *finitely supported measures with rational weights on a dense subset $A \subseteq \mathscr{X}$;*
4. *if $\mathscr{X} = \mathbb{R}^d$, the collection of absolutely continuous and compactly supported measures;*
5. *if $\mathscr{X} = \mathbb{R}^d$, the collection of absolutely continuous measures with strictly positive and bounded analytic densities.*

In particular, $\mathscr{W}_p$ is separable (the third set is countable as $\mathscr{X}$ is separable).

This is a simple consequence of Proposition 2.2.6 and approximations, and the details are given on page 43 in the supplement.

Proposition 2.2.8 (Completeness) *The Wasserstein space $\mathscr{W}_p(\mathscr{X})$ is complete.*

One may find two different proofs in Villani [125, Theorem 6.18] and Ambrosio et al. [12, Proposition 7.1.5]. On page 43 of the supplement, we sketch an alternative argument based on completeness of the weak topology.

2.2.3 Negative Topological Properties

In the previous subsection, we have shown that $\mathscr{W}_p(\mathscr{X})$ is separable and complete like L_p spaces. Just like them, however, the Wasserstein space is neither locally compact nor σ-compact. For this reason, existence proofs of Fréchet means in $\mathscr{W}_p(\mathscr{X})$ require tools that are more specific to this space, and do not rely upon local compactness (see Sect. 3.1).

Proposition 2.2.9 ($\mathscr{W}_p$ is Not Locally Compact) *Let $\mu \in \mathscr{W}_p(\mathscr{X})$ and let $\varepsilon > 0$. Then the Wasserstein ball*

$$\overline{B}_\varepsilon(\mu) = \{\nu \in \mathscr{W}_p(\mathscr{X}) : W_p(\mu, \nu) \leq \varepsilon\}$$

is not compact.

Ambrosio et al. [12, Remark 7.1.9] show this when μ is a Dirac measure, and we extend their argument on page 43 of the supplement.

From this, we deduce:

Corollary 2.2.10 *The Wasserstein space $\mathscr{W}_p(\mathscr{X})$ is not σ-compact.*

Proof. If $\mathscr{K}$ is a compact set in $\mathscr{W}_p(\mathscr{X})$, then its interior is empty by Proposition 2.2.9. A countable union of compact sets has an empty interior (hence cannot equal the entire space $\mathscr{W}_p(\mathscr{X})$) by the Baire property, which holds on the complete metric space $\mathscr{W}_p(\mathscr{X})$ by the Baire category theorem (Dudley [47, Theorem 2.5.2]).

2.2.4 Covering Numbers

Let $\mathscr{K} \subset \mathscr{W}_p(\mathscr{X})$ be compact and assume that $\mathscr{X} = \mathbb{R}^d$. Then for any $\varepsilon > 0$ the covering number

$$N(\varepsilon;\mathscr{K}) = \min\left\{ n : \exists \mu_1,\dots,\mu_n \in \mathscr{W}_p(\mathscr{X}) \text{ such that } \mathscr{K} \subseteq \bigcup_{i=1}^{n} \{\mu : W_p(\mu,\mu_i) < \varepsilon\} \right\}$$

is finite. These numbers appear in statistics in various ways, particularly in empirical processes (see, for instance, Wainwright [126, Chapter 5]) and the goal of this subsection is to give an upper bound for $N(\varepsilon;\mathscr{K})$. Invoking Proposition 2.2.3, introduce a continuous monotone divergent $f : [0,\infty) \to [0,\infty]$ such that

$$\sup_{\mu\in\mathscr{K}} \int_{\mathbb{R}^d} \|x\|^p f(\|x\|)\, \mathrm{d}\mu(x) \le 1.$$

The function f provides a certain measure of how compact $\mathscr{K}$ is. If $\mathscr{K} = \mathscr{W}_p(K)$ is the set of measures supported on a compact $K \subseteq \mathbb{R}^d$, then $f(L)$ can be taken infinite for L large, and L can be treated as a constant in the theorem. Otherwise L increases as $\varepsilon \searrow 0$, at a speed that depends on f: the faster f diverges, the slower L grows with decreasing ε and the better the bound becomes.

Theorem 2.2.11 *Let $\varepsilon > 0$ and $L = f^{-1}(1/\varepsilon^p)$. If $d\varepsilon \le L$, then*

$$\log N(\varepsilon) \le C_1(d) \left(\frac{L}{\varepsilon}\right)^d \left[(p+d)\log\frac{L}{\varepsilon} + C_2(d,p) \right],$$

with $C_1(d) = 3^d e\theta_d$, $C_2(d,p) = (p+d)\log 3 + (p+2)\log 2 + \log\theta_d$ and $\theta_d = d[5 + \log d + \log\log d]$.

Since $\varepsilon > 0$ is small and L is increasing in ε, the restriction that $d\varepsilon \le L$ is typically not binding. We provide some examples before giving the proof.

Example 1: if all the measures are supported on the d-dimensional unit ball, then L can be taken equal to one, independently of ε. We obtain

$$\widetilde{N}(\varepsilon) := \frac{\log N(\varepsilon)}{\log 1/\varepsilon} \le (d+p)C_1(d)\varepsilon^{-d} + \text{ smaller order terms.}$$

Example 2: if all the measures in $\mathscr{K}$ have uniform exponential moments, then $f(L) = e^L$ and $\widetilde{N}(\varepsilon)$ is a constant times $\varepsilon^{-d}[\log 1/\varepsilon]^d$. The exponent p appears only in the constant.

Example 3: suppose that $\mathscr{K}$ is a Wasserstein ball of order $p+\delta$, that is, $f(L) = L^\delta$. Then $L \sim \varepsilon^{-p/\delta}$ and

$$\widetilde{N}(\varepsilon) \le C_1(d)(p+d)(1+p/\delta)\varepsilon^{-d[1+p/\delta]}$$

up to smaller order terms. Here (when $0 < \delta < \infty$) the behaviour of $\widetilde{N}(\varepsilon)$ depends more strongly upon p: if $p' < p$, then we can replace δ by $\delta' = \delta + p - p' > \delta$, leading to a smaller magnitude of $\widetilde{N}(\varepsilon)$.

Example 4: if $f(L)$ is only $\log L$, then $\widetilde{N}$ behaves like $\varepsilon^{-(d+p)}\exp(\varepsilon^{-pd})$, so p has a very dominant effect.

Proof. The proof is divided into four steps.

Step 1: Compact support. Let $P_L : \mathbb{R}^d \to \mathbb{R}^d$ be the projection onto $\overline{B}_L(0) = \{x \in \mathbb{R}^d : \|x\| \le L\}$ and let $\mu \in \mathscr{K}$. Then

$$W_p^p(\mu, P_L\#\mu) \le \int_{\mathbb{R}^d} \|x - P_L(x)\|^p \, \mathrm{d}\mu(x) = \int_{\|x\|>L} \|x - P_L(x)\|^p \, \mathrm{d}\mu(x)$$
$$\le \int_{\|x\|>L} \|x\|^p \, \mathrm{d}\mu(x) \le \frac{1}{f(L)} \int_{\|x\|>L} \|x\|^p f(\|x\|) \, \mathrm{d}\mu(x) \le \frac{1}{f(L)},$$

and this vanishes as $L \to \infty$.

Step 2: n-Point measures. Let $n = N(\varepsilon; B_L(0))$ be the covering number of the Euclidean ball in $\mathbb{R}^d$. There exists a set $x_1, \ldots, x_n \in \mathbb{R}^d$ such that $B_L(0) \subseteq \cup B_\varepsilon(x_i)$. If $\mu \in \mathscr{W}_p(B_L(0))$, there exists a measure μ_n supported on the x_i's and such that

$$W_p(\mu, \mu_n) \le \varepsilon.$$

Indeed let $C_1 = B_\varepsilon(x_1)$, $C_i = B_\varepsilon(x_i) \setminus \cup_{j<i} B_\varepsilon(x_j)$ and define $\mu_n(\{x_i\}) = \mu(C_i)$. The transport map defined by $\mathbf{t}(x) = x_i$ for $x \in C_i$ pushes μ forward to μ_n and

$$W_p^p(\mu_n, \mu) \le \sum_{i=1}^n \int_{C_i} \|x - x_i\|^p \, \mathrm{d}\mu(x) \le \sum_{i=1}^n \varepsilon^p \mu(C_i) = \varepsilon^p.$$

According to Rogers [114], we have the bound

$$n \le e\theta_d [L/\varepsilon]^d, \qquad \theta_d = d[5 + \log d + \log\log d],$$

whenever $\varepsilon \le L/d$.

Step 3: Common weights. If $\mu = \sum a_k \delta_{x_k}$ and $\nu = \sum b_k \delta_{x_k}$, then $W_p^p(\mu, \nu) \le \sum_k |a_k - b_k| \sup_{i,j} \|x_i - x_j\|^p$. Let

$$\mu_{n,\varepsilon,\delta} = \left\{ \sum_{k=1}^n a_k \delta_{x_k} : a_k \in \{0, \delta, 2\delta, \ldots, \lceil 1/\delta \rceil \delta\}; \sum a_k = 1 \right\}.$$

This set contains fewer than $(2 + 1/\delta)^{n-1}$ elements, and any measure supported on $\{x_1, \ldots, x_n\}$ can be approximated by a measure in $\mu_{n,\varepsilon,\delta}$ with error $2L(n\delta)^{1/p}$.

Step 4: Conclusion. Let $L = f^{-1}(1/\varepsilon^p)$, $n = N(\varepsilon; B_L(0))$ and $\delta = [\varepsilon/(2L)]^p/n$. Combining the previous three steps, we obtain in the case $L \ge \varepsilon d$ that

$$N(3\varepsilon) \le (2 + 1/\delta)^{n-1} \le \left[2 + \left(\frac{L}{\varepsilon}\right)^{p+d} 2^p e\theta_d \right]^{e\theta_d [L/\varepsilon]^d} \le \left[\left(\frac{L}{\varepsilon}\right)^{p+d} 2^{p+2} \theta_d \right]^{e\theta_d [L/\varepsilon]^d},$$

because $L/\varepsilon \geq 1$ and $\theta_d \geq 5$. Conclude that

$$N(\varepsilon) \leq \left[3^{p+d} \left(\frac{L}{\varepsilon} \right)^{p+d} 2^{p+2} \theta_d \right]^{3^d e \theta_d [L/\varepsilon]^d}.$$

2.3 The Tangent Bundle

Although the Wasserstein space $\mathscr{W}_p(\mathscr{X})$ is non-linear in terms of measures, it *is* linear in terms of maps. Indeed, if $\mu \in \mathscr{W}_p(\mathscr{X})$ and $T_i : \mathscr{X} \to \mathscr{X}$ are such that $\|T_i\| \in L_p(\mu)$, then $(\alpha T_1 + \beta T_2)\#\mu \in \mathscr{W}_p(\mathscr{X})$ for all $\alpha, \beta \in \mathbb{R}$. Later, in Sect. 2.4, we shall see that $\mathscr{W}_p(\mathscr{X})$ is in fact homeomorphic to a subset of the space of such functions. The goal of this section is to exploit the linearity of the latter in order to define the tangent bundle of $\mathscr{W}_p$. This in particular will be used for deriving differentiability properties of the Wasserstein distance in Sect. 3.1.6. However, the latter can be understood at a purely analytic level, and readers uncomfortable with differential geometry can access most of the rest of the monograph without reference to this section.

We assume here that $\mathscr{X}$ is a Hilbert space and that $p = 2$; the results extend to any $p > 1$. Absolutely continuous measures are assumed to be so with respect to Lebesgue measure if $\mathscr{X} = \mathbb{R}^d$ and otherwise refer to Definition 1.6.4.

2.3.1 Geodesics, the Log Map and the Exponential Map in $\mathscr{W}_2(\mathscr{X})$

Let $\gamma \in \mathscr{W}_2(\mathscr{X})$ be absolutely continuous and $\mu \in \mathscr{W}_2(\mathscr{X})$ arbitrary. From Sect. 1.6.1, we know that there exists a unique solution to the Monge–Kantorovich problem, and that solution is given by a transport map that we denote by $\mathbf{t}_\gamma^\mu$. Recalling that $\mathbf{i} : \mathscr{X} \to \mathscr{X}$ is the identity map, we can define a curve

$$\gamma_t = \left[\mathbf{i} + t(\mathbf{t}_\gamma^\mu - \mathbf{i}) \right] \#\gamma, \qquad t \in [0,1].$$

This curve is known as McCann's [93] interpolant. As hinted in the introduction to this section, it is constructed via classical linear interpolation of the transport maps $\mathbf{t}_\gamma^\mu$ and the identity. Clearly $\gamma_0 = \gamma$, $\gamma_1 = \mu$ and from (2.3),

$$\begin{aligned} W_2(\gamma_t, \gamma) &\leq \sqrt{\int_{\mathscr{X}} \left[t(\mathbf{t}_\gamma^\mu - \mathbf{i}) \right]^2 \mathrm{d}\gamma} &&= t W_2(\gamma, \mu); \\ W_2(\gamma_t, \mu) &\leq \sqrt{\int_{\mathscr{X}} \left[(1-t)(\mathbf{t}_\gamma^\mu - \mathbf{i}) \right]^2 \mathrm{d}\gamma} &&= (1-t) W_2(\gamma, \mu). \end{aligned}$$

It follows from the triangle inequality in $\mathscr{W}_2$ that these inequalities must hold as equalities. Taking this one step further, we see that

$$W_2(\gamma_t,\gamma_s)=(t-s)W_2(\gamma,\mu),\qquad 0\le s\le t\le 1.$$

In other words, McCann's interpolant is a *constant-speed geodesic* in $\mathscr{W}_2(\mathscr{X})$.

In view of this, it seems reasonable to define the *tangent space* of $\mathscr{W}_2(\mathscr{X})$ at μ as (Ambrosio et al. [12, Definition 8.5.1])

$$\mathrm{Tan}_\mu=\overline{\{t(\mathbf{t}-\mathbf{i}):\mathbf{t}=\mathbf{t}_\mu^\nu \text{ for some } \nu\in\mathscr{W}_2(\mathscr{X});t>0\}}^{L_2(\mu)}.$$

It follows from the definition that $\mathrm{Tan}_\mu\subseteq L_2(\mu)$. (Strictly speaking, Tan_μ is a subset of the space of functions $f:\mathscr{X}\to\mathscr{X}$ such that $\|f\|\in L_2(\mu)$ rather than $L_2(\mu)$ itself, as in Definition 2.4.3, but we will write L_2 for simplicity.)

Although not obvious from the definition, this is a linear space. The reason is that, in $\mathbb{R}^d$, Lipschitz functions are dense in $L_2(\mu)$, and for $\mathbf{t}$ Lipschitz the negative of a tangent element

$$-t(\mathbf{t}-\mathbf{i})=s(\mathbf{s}-\mathbf{i}),\qquad s>t\|\mathbf{t}\|_{\mathrm{Lip}},\qquad \mathbf{s}=\mathbf{i}+\frac{t}{s}(\mathbf{i}-\mathbf{t})$$

lies in the tangent space, since $\mathbf{s}$ can be seen to belong to the subgradient of a convex function by definition of s. This also shows that Tan_μ can be seen to be the $L_2(\mu)$-closure of all gradients of C_c^∞ functions. We refer to [12, Definition 8.4.1 and Theorem 8.5.1] for the proof and extensions to other values of $p>1$ and to infinite dimensions, using cylindrical functions that depend on finitely many coordinates [12, Definition 5.1.11]. The alternative definition highlights that it is essentially the inner product in Tan_μ, but not the elements themselves, that depends on μ.

The tangent space definition is valid for arbitrary measures in $\mathscr{W}_2(\mathscr{X})$. The exponential map at $\gamma\in\mathscr{W}_2(\mathscr{X})$ is the restriction to Tan_γ of the mapping that sends $\mathbf{r}\in L_2(\gamma)$ to $[\mathbf{r}+\mathbf{i}]\#\gamma\in\mathscr{W}_2(\mathscr{X})$. More explicitly, $\exp_\gamma:\mathrm{Tan}_\gamma\to\mathscr{W}_2$ takes the form

$$\exp_\gamma(t(\mathbf{t}-\mathbf{i}))=\exp_\gamma([t\mathbf{t}+(1-t)\mathbf{i}]-\mathbf{i})=[t\mathbf{t}+(1-t)\mathbf{i}]\#\gamma\quad(t\in\mathbb{R}).$$

Thus, when γ is absolutely continuous, $\exp_\gamma$ is surjective, as can be seen from its right inverse, the log map

$$\log_\gamma:\mathscr{W}_2\to\mathrm{Tan}_\gamma\qquad \log_\gamma(\mu)=\mathbf{t}_\gamma^\mu-\mathbf{i},$$

defined throughout $\mathscr{W}_2$ (by virtue of Theorem 1.6.2). In symbols,

$$\exp_\gamma(\log_\gamma(\mu))=\mu,\quad \mu\in\mathscr{W}_2,\quad \text{and}\quad \log_\gamma(\exp_\gamma(t(\mathbf{t}-\mathbf{i})))=t(\mathbf{t}-\mathbf{i})\quad(t\in[0,1]),$$

because convex combinations of optimal maps are optimal maps as well. In particular, McCann's interpolant $\left[\mathbf{i}+t(\mathbf{t}_\gamma^\mu-\mathbf{i})\right]\#\gamma$ is mapped bijectively to the line segment $t(\mathbf{t}_\gamma^\mu-\mathbf{i})\in\mathrm{Tan}_\gamma$ through the log map.

It is also worth mentioning that McCann's interpolant can also be defined as

$$[tp_2+(1-t)p_1]\#\pi,\qquad p_1(x,y)=x,\quad p_2(x,y)=y,$$

where $p_1, p_2 : \mathscr{X}^2 \to \mathscr{X}$ are projections and π is any optimal transport plan between γ and μ. This is defined for arbitrary measures $\gamma, \mu \in \mathscr{W}_2$, and reduces to the previous definition if γ is absolutely continuous. It is shown in Ambrosio et al. [12, Chapter 7] or Santambrogio [119, Proposition 5.32] that these are the only constant-speed geodesics in $\mathscr{W}_2$.

2.3.2 Curvature and Compatibility of Measures

Let $\gamma, \mu, \nu \in \mathscr{W}_2(\mathscr{X})$ be absolutely continuous measures. Then by (2.3)

$$W_2^2(\mu,\nu) \le \int_{\mathscr{X}} \|\mathbf{t}_\gamma^\mu(x) - \mathbf{t}_\gamma^\nu(x)\|^2 \,\mathrm{d}\gamma(x) = \|\log_\gamma(\mu) - \log_\gamma(\nu)\|^2.$$

In other words, the distance between μ and ν is smaller in $\mathscr{W}_2(\mathscr{X})$ than the distance between the corresponding vectors $\log_\gamma(\mu)$ and $\log_\gamma(\nu)$ in the tangent space Tan_γ. In the terminology of differential geometry, this means that the Wasserstein space has *nonnegative sectional curvature* at any absolutely continuous γ.

It is instructive to see when equality holds. As $\mathbf{t}_\nu^\gamma = (\mathbf{t}_\gamma^\nu)^{-1}$, a change of variables gives

$$W_2^2(\mu,\nu) \le \int_{\mathscr{X}} \|\mathbf{t}_\gamma^\mu(\mathbf{t}_\nu^\gamma(x)) - x\|^2 \,\mathrm{d}\nu(x).$$

Since the map $\mathbf{t}_\gamma^\mu \circ \mathbf{t}_\nu^\gamma$ pushes forward ν to μ, equality holds if and only if $\mathbf{t}_\gamma^\mu \circ \mathbf{t}_\nu^\gamma = \mathbf{t}_\nu^\mu$. This motivates the following definition.

Definition 2.3.1 (Compatible Measures) *A collection of absolutely continuous measures $\mathscr{C} \subseteq \mathscr{W}_2(\mathscr{X})$ is* compatible *if for all $\gamma, \mu, \nu \in \mathscr{C}$, we have $\mathbf{t}_\gamma^\mu \circ \mathbf{t}_\nu^\gamma = \mathbf{t}_\nu^\mu$ (in $L_2(\nu)$).*

Remark 2.3.2 *The absolute continuity is not necessary and was introduced for notational simplicity. A more general definition that applies to general measures is the following: every finite subcollection of $\mathscr{C}$ admits an optimal multicoupling whose relevant projections are simultaneously pairwise optimal; see the paragraph preceding Theorem 3.1.9.*

A collection of two (absolutely continuous) measures is always compatible. More interestingly, if $\mathscr{X} = \mathbb{R}$, then the entire collection of absolutely continuous (or even just continuous) measures is compatible. This is because of the simple geometry of convex functions in $\mathbb{R}$: gradients of convex functions are nondecreasing, and this property is stable under composition. In a more probabilistic way of thinking, one can always push-forward μ to ν via the uniform distribution $\mathrm{Leb}|_{[0,1]}$ (see Sect. 1.5). Letting F_μ^{-1} and F_ν^{-1} denote the quantile functions, we have seen that

$$W_2(\mu,\nu) = \|F_\mu^{-1} - F_\nu^{-1}\|_{L_2(0,1)}.$$

(As a matter of fact, in this specific case, the equality holds for all $p \geq 1$ and not only for $p = 2$.) In other words, $\mu \mapsto F_\mu^{-1}$ is an *isometry* from $\mathscr{W}_2(\mathbb{R})$ to the subset of $L_2(0,1)$ formed by (equivalence classes of) left-continuous nondecreasing functions on $(0,1)$. Since this is a convex subset of a Hilbert space, this property provides a very simple way to evaluate Fréchet means in $\mathscr{W}_2(\mathbb{R})$ (see Sect. 3.1). If $\gamma = \mathrm{Leb}|_{[0,1]}$, then $F_\mu^{-1} = \mathbf{t}_\gamma^\mu$ for all μ, so we can write the above equality as

$$W_2^2(\mu,\nu) = \|F_\mu^{-1} - F_\nu^{-1}\|_{L_2(0,1)} = \|\log_\gamma(\mu) - \log_\gamma(\nu)\|^2,$$

so that if $\mathscr{X} = \mathbb{R}$, the Wasserstein space is essentially *flat* (has zero sectional curvature).

The importance of compatibility can be seen as mimicking the simple one-dimensional case in terms of a Hilbert space embedding. Let $\mathscr{C} \subseteq \mathscr{W}_2(\mathscr{X})$ be compatible and fix $\gamma \in \mathscr{C}$. Then for all $\mu, \nu \in \mathscr{C}$

$$W_2^2(\mu,\nu) = \int_{\mathscr{X}} \|\mathbf{t}_\gamma^\mu(x) - \mathbf{t}_\gamma^\nu(x)\|^2 \,\mathrm{d}\gamma(x) = \|\log_\gamma(\mu) - \log_\gamma(\nu)\|_{L_2(\gamma)}^2.$$

Consequently, once again, $\mu \mapsto \mathbf{t}_\gamma^\mu$ is an isometric embedding of $\mathscr{C}$ into $L_2(\gamma)$. Generalising the one-dimensional case, we shall see that this allows for easy calculations of Fréchet means by means of averaging transport maps (Theorem 3.1.9).

Example: Gaussian compatible measures. The Gaussian case presented in Sect. 1.6.3 is helpful in shedding light on the structure imposed by the compatibility condition. Let $\gamma \in \mathscr{W}_2(\mathbb{R}^d)$ be a standard Gaussian distribution with identity covariance matrix. Let Σ_μ denote the covariance matrix of a measure $\mu \in \mathscr{W}_2(\mathbb{R}^d)$. When μ and ν are centred nondegenerate Gaussian measures,

$$\mathbf{t}_\gamma^\mu = \Sigma_\mu^{1/2}; \qquad \mathbf{t}_\gamma^\nu = \Sigma_\nu^{1/2}; \qquad \mathbf{t}_\mu^\nu = \Sigma_\mu^{-1/2}[\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2}]^{1/2}\Sigma_\mu^{-1/2},$$

so that γ, μ, and ν are compatible if and only if

$$\mathbf{t}_\mu^\nu = \mathbf{t}_\gamma^\nu \circ \mathbf{t}_\mu^\gamma = \Sigma_\nu^{1/2}\Sigma_\mu^{-1/2}.$$

Since the matrix on the left-hand side must be symmetric, it must necessarily be that $\Sigma_\nu^{1/2}$ and $\Sigma_\mu^{-1/2}$ commute (if A and B are symmetric, then AB is symmetric if and only if $AB = BA$), or equivalently, if and only if Σ_ν and Σ_μ commute. We see that a collection $\mathscr{C}$ of Gaussian measures on $\mathbb{R}^d$ that includes the standard Gaussian distribution is compatible if and only if all the covariance matrices of the measures in $\mathscr{C}$ are *simultaneously diagonalisable*. In other words, there exists an orthogonal matrix U such that $D_\mu = U\Sigma_\mu U^t$ is diagonal for all $\mu \in \mathscr{C}$. In that case, formula (1.6)

$$\mathscr{W}_2^2(\mu,\nu) = \mathrm{tr}[\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}] = \mathrm{tr}[\Sigma_\mu + \Sigma_\nu - 2\Sigma_\mu^{1/2}\Sigma_\nu^{1/2}]$$

simplifies to

$$\mathscr{W}_2^2(\mu,\nu) = \mathrm{tr}[D_\mu + D_\nu - 2D_\mu^{1/2}D_\nu^{1/2}] = \sum_{i=1}^{d}(\sqrt{\alpha_i}-\sqrt{\beta_i})^2,\ \ \alpha_i = [D_\mu]_{ii};\ \ \beta_i = [D_\nu]_{ii},$$

and identifying the (nonnegative) number $a \in \mathbb{R}$ with the map $x \mapsto ax$ on $\mathbb{R}$, the optimal maps take the "orthogonal separable" form

$$\mathbf{t}_\mu^\nu = \Sigma_\nu^{1/2}\Sigma_\mu^{-1/2} = UD_\nu^{1/2}D_\mu^{-1/2}U^t = U \circ \left(\sqrt{\beta_1/\alpha_1},\ldots,\sqrt{\beta_d/\alpha_d}\right) \circ U^t.$$

In other words, up to an orthogonal change of coordinates, the optimal maps take the form of d nondecreasing real-valued functions. This is yet another crystallisation of the one-dimensional-like structure of compatible measures.

With the intuition of the Gaussian case at our disposal, we can discuss a more general case. Suppose that the optimal maps are continuously differentiable. Then differentiating the equation $\mathbf{t}_\mu^\nu = \mathbf{t}_\gamma^\nu \circ \mathbf{t}_\mu^\gamma$ gives

$$\nabla\mathbf{t}_\mu^\nu(x) = \nabla\mathbf{t}_\gamma^\nu(\mathbf{t}_\mu^\gamma(x))\nabla\mathbf{t}_\mu^\gamma(x).$$

Since optimal maps are gradients of convex functions, their derivatives must be symmetric and positive semidefinite matrices. A product of such matrices stays symmetric if and only if they commute, so in this differentiable setting, compatibility is equivalent to commutativity of the matrices $\nabla\mathbf{t}_\gamma^\nu(\mathbf{t}_\mu^\gamma(x))$ and $\nabla\mathbf{t}_\mu^\gamma(x)$ for μ-almost all x. In the Gaussian case, the optimal maps are linear functions, so x does not appear in the matrices.

Here are some examples of compatible measures. It will be convenient to describe them using the optimal maps from a reference measure $\gamma \in \mathscr{W}_2(\mathbb{R}^d)$. Define $\mathscr{C} = \mathbf{t}\#\gamma$ with $\mathbf{t}$ belonging to one of the following families. The first imposes the one-dimensional structure by varying only the behaviour of the norm of x, while the second allows for separation of variables that splits the d-dimensional problem into d one-dimensional ones.

Radial transformations. Consider the collection of functions $\mathbf{t} : \mathbb{R}^d \to \mathbb{R}^d$ of the form $\mathbf{t}(x) = xG(\|x\|)$ with $G : \mathbb{R}_+ \to \mathbb{R}$ differentiable. Then a straightforward calculation shows that

$$\nabla\mathbf{t}(x) = G(\|x\|)I + [G'(\|x\|)/\|x\|]\ xx^t.$$

Since both I and xx^t are positive semidefinite, the above matrix is so if both G and G' are nonnegative. If $\mathbf{s}(x) = xH(\|x\|)$ is a function of the same form, then $\mathbf{s}(\mathbf{t}(x)) = xG(\|x\|)H(\|x\|G(\|x\|))$ which belongs to that family of functions (since G is nonnegative). Clearly

$$\nabla\mathbf{s}(\mathbf{t}(x)) = H\left[\|x\|G(\|x\|)\right]I + \left[G(\|x\|)H'(\|x\|G(\|x\|))/\|x\|\right]\ xx^t$$

commutes with $\nabla\mathbf{t}(x)$, since both matrices are of the form $aI + bxx^t$ with a, b scalars (that depend on x). In order to be able to change the base measure γ, we need to

check that the inverses belong to the family. But if $y = \mathbf{t}(x)$, then $x = ay$ for some scalar a that solves the equation

$$aG(a\|y\|) = 1.$$

Such a is guaranteed to be unique if $a \mapsto aG(a)$ is strictly increasing and it will exist (for y in the range of $\mathbf{t}$) if it is continuous. As a matter of fact, since the eigenvalues of $\nabla\mathbf{t}(x)$ are $G(a)$ and

$$G(a) + G'(a)a = (aG(a))', \qquad a = \|x\|,$$

the condition that $a \mapsto aG(a)$ is strictly increasing is sufficient (this is weaker than G itself increasing). Finally, differentiability of G is not required, so it is enough if G is continuous and $aG(a)$ is strictly increasing.

Separable variables. Consider the collection of functions $\mathbf{t} : \mathbb{R}^d \to \mathbb{R}^d$ of the form

$$\mathbf{t}(x_1, \ldots, x_d) = (T_1(x_1), \ldots, T_d(x_d)), \qquad T_i : \mathbb{R} \to \mathbb{R}, \tag{2.8}$$

with T_i continuous and strictly increasing. This is a generalisation of the compatible Gaussian case discussed above in which all the T_i's were linear. Here, it is obvious that elements in this family are optimal maps and that the family is closed under inverses and composition, so that compatibility follows immediately.

This family is characterised by measures having a *common dependence structure*. More precisely, we say that $C : [0,1]^d \to [0,1]$ is a *copula* if C is (the restriction of) a distribution function of a random vector having uniform margins. In other words, if there is a random vector $V = (V_1, \ldots, V_d)$ with $\mathbb{P}(V_i \le a) = a$ for all $a \in [0,1]$ and all $j = 1, \ldots, d$, and

$$\mathbb{P}(V_1 \le v_1, \ldots, V_d \le v_d) = C(v_1, \ldots, v_d), \qquad u_i \in [0,1].$$

Nelsen [97] provides an overview on copulae. To any d-dimensional probability measure μ, one can assign a copula $C = C_\mu$ in terms of the distribution function G of μ and its marginals G_j as

$$G(a_1, \ldots, a_d) = \mu((-\infty, a_1] \times \cdots \times (-\infty, a_d]) = C(G_1(a_1), \ldots, G_d(a_d)).$$

If each G_j is surjective on $(0,1)$, which is equivalent to it being continuous, then this equation defines C uniquely on $(0,1)^d$, and consequently on $[0,1]^d$. If some marginal G_j is not continuous, then uniqueness is lost, but C still exists [97, Chapter 2]. The connection of copulae to compatibility becomes clear in the following lemma, proven on page 51 in the supplement.

Lemma 2.3.3 (Compatibility and Copulae) *The copulae associated with absolutely continuous measures $\mu, \nu \in \mathscr{W}_2(\mathbb{R}^d)$ are equal if and only if $\mathbf{t}_\mu^\nu$ takes the separable form (2.8).*

Composition with linear functions. If $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex with gradient $\mathbf{t}$ and A is a $d \times d$ matrix, then the gradient of the convex function $x \mapsto \phi(Ax)$ at x is $\mathbf{t}_A = A^t \mathbf{t}(Ax)$.

Suppose ψ is another convex function with gradient $\mathbf{s}$ and that compatibility holds, i.e., $\nabla\mathbf{s}(\mathbf{t}(x))$ commutes with $\nabla\mathbf{t}(x)$ for all x. Then in order for

$$\nabla\mathbf{s}_A(\mathbf{t}_A(x)) = A^t\nabla\mathbf{s}(AA^t\mathbf{t}(Ax))A \qquad \text{and} \qquad \nabla\mathbf{t}_A(x) = A^t\nabla\mathbf{t}(Ax)A$$

to commute, it suffices that $AA^t = I$, i.e., that A be orthogonal. Consequently, if $\{\mathbf{t}\#\mu\}_{\mathbf{t}\in\mathbf{T}}$ are compatible, then so are $\{\mathbf{t}_U\#\mu\}_{\mathbf{t}\in\mathbf{T}}$ for any orthogonal matrix U.

2.4 Random Measures in Wasserstein Space

Let μ be a fixed absolutely continuous probability measure in $\mathscr{W}_2(\mathscr{X})$. If $\Lambda \in \mathscr{W}_2(\mathscr{X})$ is another probability measure, then the transport map $\mathbf{t}_\mu^\Lambda$ and the convex potential are functions of Λ. If Λ is now random, then we would like to be able to make probability statements about them. To this end, it needs to be shown that $\mathbf{t}_\mu^\Lambda$ and the convex potential are *measurable* functions of Λ. The goal of this section is to develop a rigorous mathematical framework that justifies such probability statements. We show that all the relevant quantities are indeed measurable, and in particular establish a Fubini-type result in Proposition 2.4.9. This technical section may be skipped at first reading.

Here is an example of a measurability result (Villani [125, Corollary 5.22]). Recall that $P(\mathscr{X})$ is the space of Borel probability measures on $\mathscr{X}$, endowed with the topology of weak convergence that makes it a metric space. Let $\mathscr{X}$ be a complete separable metric space and $c : \mathscr{X}^2 \to \mathbb{R}_+$ a continuous cost function. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $\Lambda, \kappa : \Omega \to P(\mathscr{X})$ be measurable maps. Then there exists a *measurable selection* of optimal transference plans. That is, a measurable $\pi : \Omega \to P(\mathscr{X}^2)$ such that $\pi(\omega) \in \Pi(\Lambda(\omega), \kappa(\omega))$ is optimal for all $\omega \in \Omega$.

Although this result is very general, it only provides information about π. If π is induced from a map T, it is not obvious how to construct T from π in a measurable way; we will therefore follow a different path. In order to (almost) have a self-contained exposition, we work in a somewhat simplified setting that nevertheless suffices for the sequel. At least in the Euclidean case $\mathscr{X} = \mathbb{R}^d$, more general measurability results in the flavour of this section can be found in Fontbona et al. [53]. On the other hand, we will not need to appeal to abstract measurable selection theorems as in [53, 125].

2.4.1 Measurability of Measures and of Optimal Maps

Let $\mathscr{X}$ be a separable Banach space. (Most of the results below hold for any complete separable metric space but we will avoid this generality for brevity and simpler notation). The Wasserstein space $\mathscr{W}_p(\mathscr{X})$ is a metric space for any $p \geq 1$. We can thus define:

Definition 2.4.1 (Random Measure) *A random measure Λ is any measurable map from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $\mathscr{W}_p(\mathscr{X})$, endowed with its Borel σ-algebra.*

In what follows, whenever we call something random, we mean that it is measurable as a map from some generic unspecified probability space.

Lemma 2.4.2 *A random measure Λ is measurable if and only if it is measurable with respect to the induced weak topology.*

Since both topologies are Polish, this follows from abstract measure-theoretic results (Fremlin [57, Paragraph 423F]). We give an elementary proof on page 53 of the supplement.

Optimal maps are functions from $\mathscr{X}$ to itself. In order to define random optimal maps, we need to define a topology and a σ-algebra on the space of such functions.

Definition 2.4.3 (The Space $\mathscr{L}_p(\mu)$) *Let $\mathscr{X}$ be a Banach space and μ a Borel measure on $\mathscr{X}$. Then the space $\mathscr{L}_p(\mu)$ is the space of measurable functions $f : \mathscr{X} \to \mathscr{X}$ such that*

$$\|f\|_{\mathscr{L}_p(\mu)} = \left(\int_{\mathscr{X}} \|f(x)\|_{\mathscr{X}}^p \, \mathrm{d}\mu(x) \right)^{1/p} < \infty.$$

When $\mathscr{X}$ is separable, $\mathscr{L}_p(\mu)$ is an example of a *Bochner space*, though we will not use this terminology.

It follows from the definition that $\|f\|_{\mathscr{L}_p(\mu)}$ is the L_p norm of the map $x \mapsto \|f(x)\|_{\mathscr{X}}$ from $\mathscr{X}$ to $\mathbb{R}$:

$$\|f\|_{\mathscr{L}_p(\mu)} = \| \ \|f\|_{\mathscr{X}} \ \|_{L_p(\mu)}.$$

As usual we identify functions that coincide almost everywhere. Clearly, $\mathscr{L}_p(\mu)$ is a normed vector space. It enjoys another property shared by L_p spaces—completeness:

Theorem 2.4.4 (Riesz–Fischer) *The space $\mathscr{L}_p(\mu)$ is a Banach space.*

The proof, a simple variant of the classical one, is given on page 53 of the supplement.

Random maps lead naturally to random measures:

Lemma 2.4.5 (Push-Forward with Random Maps) *Let $\mu \in \mathscr{W}_p(\mathscr{X})$ and let $\mathbf{t}$ be a random map in $\mathscr{L}_p(\mu)$. Then $\Lambda = \mathbf{t}\#\mu$ is a continuous mapping from $\mathscr{L}_p(\mu)$ to $\mathscr{W}_p(\mathscr{X})$, hence a random measure.*

Proof. That Λ takes values in $\mathscr{W}_p$ follows from a change of variables

$$\int_{\mathscr{X}} \|x\|^p \, \mathrm{d}\Lambda(x) = \int_{\mathscr{X}} \|\mathbf{t}(x)\|^p \, \mathrm{d}\mu(x) = \|\mathbf{t}\|_{\mathscr{L}_p(\mu)} < \infty.$$

Since $W_p(\mathbf{t}\#\mu, \mathbf{s}\#\mu) \le \| \ \|\mathbf{t} - \mathbf{s}\|_{\mathscr{X}} \ \|_{L_p(\mu)} = \|\mathbf{t} - \mathbf{s}\|_{\mathscr{L}_p(\mu)}$ (see (2.3)), Λ is a continuous (in fact, 1-Lipschitz) function of $\mathbf{t}$.

Conversely, $\mathbf{t}$ is a continuous function of Λ:

Lemma 2.4.6 (Measurability of Transport Maps) *Let Λ be a random measure in $\mathscr{W}_p(\mathscr{X})$ and let $\mu \in \mathscr{W}_p(\mathscr{X})$ such that $(\mathbf{i}, \mathbf{t}_\mu^\Lambda)\#\mu$ is the unique optimal coupling of μ and Λ. Then $\Lambda \mapsto \mathbf{t}_\mu^\Lambda$ is a continuous mapping from $\mathscr{W}_p(\mathscr{X})$ to $\mathscr{L}_p(\mu)$, so $\mathbf{t}_\mu^\Lambda$ is a random element in $\mathscr{L}_p(\mu)$. In particular, the result holds if $\mathscr{X}$ is a separable Hilbert space, $p > 1$, and μ is absolutely continuous.*

Proof. This result is more subtle than Lemma 2.4.5, since $\Lambda \mapsto \mathbf{t}_\mu^\Lambda$ is not necessarily Lipschitz. We give here a self-contained proof for the Euclidean case with quadratic cost and μ absolutely continuous. The general case builds on Villani [125, Corollary 5.23] and is given on page 54 of the supplement.

Suppose that $\Lambda_n \to \Lambda$ in $\mathscr{W}_2(\mathbb{R}^d)$ and fix $\varepsilon > 0$. For any $S \subseteq \mathbb{R}^d$,

$$\|\mathbf{t}_\mu^{\Lambda_n} - \mathbf{t}_\mu^\Lambda\|_{\mathscr{L}_2(\mu)}^2 = \int_S \|\mathbf{t}_\mu^{\Lambda_n} - \mathbf{t}_\mu^\Lambda\|^2 \,\mathrm{d}\mu + \int_{\mathbb{R}^d \setminus S} \|\mathbf{t}_\mu^{\Lambda_n} - \mathbf{t}_\mu^\Lambda\|^2 \,\mathrm{d}\mu.$$

Since $\|a-b\|^p \le 2^p\|a\|^p + 2^p\|b\|^p$, the last integral is no larger than

$$4\int_{\mathbb{R}^d\setminus S} \|\mathbf{t}_\mu^{\Lambda_n}\|^2 \,\mathrm{d}\mu + 4\int_{\mathbb{R}^d\setminus S} \|\mathbf{t}_\mu^{\Lambda}\|^2 \,\mathrm{d}\mu = 4\int_{(\mathbf{t}_\mu^{\Lambda_n})^{-1}(\mathbb{R}^d\setminus S)} \|x\|^2 \,\mathrm{d}\Lambda_n(x) + 4\int_{(\mathbf{t}_\mu^{\Lambda})^{-1}(\mathbb{R}^d\setminus S)} \|x\|^2 \,\mathrm{d}\Lambda(x).$$

Since (Λ_n) and Λ are tight in the Wasserstein space, they must satisfy the absolute uniform continuity (2.7). Let $\delta = \delta_\varepsilon$ as in (2.7), and notice that by the measure preserving property of the optimal maps, the last two integrals are taken on sets of measures $1 - \mu(S)$. Since μ is absolutely continuous, we can find a compact set S of μ-measure at least $1-\delta$ and on which Proposition 1.7.11 applies (see Corollary 1.7.12), yielding

$$\int_S \|\mathbf{t}_\mu^{\Lambda_n} - \mathbf{t}_\mu^\Lambda\|^2 \,\mathrm{d}\mu \le \|\mathbf{t}_\mu^{\Lambda_n} - \mathbf{t}_\mu^\Lambda\|_\infty^2 \to 0, \qquad n \to \infty,$$

so that

$$\limsup_{n\to\infty} \|\mathbf{t}_\mu^{\Lambda_n} - \mathbf{t}_\mu^\Lambda\|_{\mathscr{L}_2(\mu)} \le 8\varepsilon,$$

and this completes the proof upon letting $\varepsilon \to 0$.

In Proposition 5.3.7, we show under some conditions that $\|\mathbf{t}_\mu^\Lambda\|_{\mathscr{L}_2(\mu)}$ is a continuous function of μ.

2.4.2 Random Optimal Maps and Fubini's Theorem

From now on, we assume that $\mathscr{X}$ is a separable Hilbert space and that $p = 2$. The results can most likely be generalised to all $p > 1$ (see Ambrosio et al. [12, Section 10.2]), but we restrict to the quadratic case for simplicity.

Theorem 3.2.13 below requires the application of Fubini's theorem in the form

$$\mathbb{E}\int_{\mathscr{X}}\left\langle \mathbf{t}^{\Lambda}_{\theta_0}-\mathbf{i},\mathbf{t}^{\theta}_{\theta_0}-\mathbf{i}\right\rangle \mathrm{d}\theta_0=\int_{\mathscr{X}}\mathbb{E}\left\langle \mathbf{t}^{\Lambda}_{\theta_0}-\mathbf{i},\mathbf{t}^{\theta}_{\theta_0}-\mathbf{i}\right\rangle \mathrm{d}\theta_0=\int_{\mathscr{X}}\left\langle \mathbb{E}\mathbf{t}^{\Lambda}_{\theta_0}-\mathbf{i},\mathbf{t}^{\theta}_{\theta_0}-\mathbf{i}\right\rangle \mathrm{d}\theta_0.$$

In order for this to even make sense, we need to have a meaning for "expectation" in the spaces $\mathscr{L}_2(\theta_0)$ and $L_2(\theta_0)$, both of which are Banach spaces. There are several (nonequivalent) definitions for integrals in such spaces (Hildebrant [69]); the one which will be the most convenient for our needs is the Bochner integral.

Definition 2.4.7 (Bochner Integral) *Let B be a Banach space and let $f:(\Omega,\mathscr{F},\mathbb{P})\to B$ be a simple random element taking values in B:*

$$f(\omega)=\sum_{j=1}^{n}f_j\mathbf{1}\{\omega\in\Omega_j\},\qquad \Omega_j\in\mathscr{F},\qquad f_j\in B.$$

Then the Bochner integral (or expectation) of f is defined by

$$\mathbb{E}f=\sum_{j=1}^{n}\mathbb{P}(\Omega_j)f_j\in B.$$

If f is measurable and there exists a sequence f_n of simple random elements such that $\|f_n-f\|\to 0$ almost surely and $\mathbb{E}\|f_n-f\|\to 0$, then the Bochner integral of f is defined as the limit

$$\mathbb{E}f=\lim_{n\to\infty}\mathbb{E}f_n.$$

The space of functions for which the Bochner integral is defined is the *Bochner space* $L_1(\Omega;B)$, but we will use neither this terminology nor the notation. It is not difficult to see that Bochner integrals are well-defined: the expectations do not depend on the representation of the simple functions nor on the approximating sequence, and the limit exists in B (because it is complete). More on Bochner integrals can be found in Hsing and Eubank [71, Section 2.6] or Dunford et al. [48, Chapter III.6]. A major difference from the real case is that there is no clear notion of "infinity" here: the Bochner integral is always an element of B, whereas expectations of real-valued random variables can be defined in $\mathbb{R}\cup\{\pm\infty\}$. It turns out that separability is quite important in this setting:

Lemma 2.4.8 (Approximation of Separable Functions) *Let $f:\Omega\to B$ be measurable. Then there exists a sequence of simple functions f_n such that $\|f_n(\omega)-f(\omega)\|\to 0$ for almost all ω if and only if $f(\Omega\setminus\mathscr{N})$ is separable for some $\mathscr{N}\subseteq\Omega$ of probability zero. In that case, f_n can be chosen so that $\|f_n(\omega)\|\le 2\|f(\omega)\|$ for all $\omega\in\Omega$.*

A proof can be found in [48, Lemma III.6.9], or on page 55 of the supplement. Functions satisfying this approximation condition are sometimes called *strongly measurable* or *Bochner measurable*. In view of the lemma, we will call them *separately valued*, since this is the condition that will need to be checked in order to define their integrals.

Two remarks are in order. Firstly, if B itself is separable, then $f(\Omega)$ will obviously be separable. Secondly, the set $\mathcal{N}' \subset \Omega \setminus \mathcal{N}$ on which (g_{n_k}) does not converge to f may fail to be measurable, but must have outer probability zero (it is included in a measurable set of measure zero) [48, Lemma III.6.9]. This can be remedied by assuming that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete. It will not, however, be necessary to do so, since this measurability issue will not alter the Bochner expectation of f.

Proposition 2.4.9 (Fubini for Optimal Maps) *Let Λ be a random measure in $\mathcal{W}_2(\mathcal{X})$ such that $\mathbb{E}W_2(\delta_0, \Lambda) < \infty$ and let $\theta_0, \theta \in \mathcal{W}_2(\mathcal{X})$ such that $\mathbf{t}_{\theta_0}^{\Lambda}$ and $\mathbf{t}_{\theta_0}^{\theta}$ exist (and are unique) with probability one. (For example, if θ_0 is absolutely continuous.) Then*

$$\mathbb{E}\int_{\mathcal{X}} \left\langle \mathbf{t}_{\theta_0}^{\Lambda} - \mathbf{i}, \mathbf{t}_{\theta_0}^{\theta} - \mathbf{i} \right\rangle \mathrm{d}\theta_0 = \int_{\mathcal{X}} \mathbb{E}\left\langle \mathbf{t}_{\theta_0}^{\Lambda} - \mathbf{i}, \mathbf{t}_{\theta_0}^{\theta} - \mathbf{i} \right\rangle \mathrm{d}\theta_0 = \int_{\mathcal{X}} \left\langle \mathbb{E}\mathbf{t}_{\theta_0}^{\Lambda} - \mathbf{i}, \mathbf{t}_{\theta_0}^{\theta} - \mathbf{i} \right\rangle \mathrm{d}\theta_0. \tag{2.9}$$

This holds by linearity when Λ is a simple random measure. The general case follows by approximation: the Wasserstein space is separable and so is the space of optimal maps, by Lemma 2.4.6, so we may apply Lemma 2.4.8 and approximate $\mathbf{t}_{\theta_0}^{\Lambda}$ by simple maps for which the equality holds by linearity. On page 56 of the supplement, we show that these simple maps can be assumed optimal, and give the full details.

2.5 Bibliographical Notes

Our proof of Theorem 2.2.11 borrows heavily from Bolley et al. [29]. A similar result was obtained by Kloeckner [81], who also provides a lower bound of a similar order.

The origins of Sect. 2.3 can be traced back to the seminal work of Jordan et al. [74], who interpret the Fokker–Planck equation as a gradient flow (where functionals defined on $\mathcal{W}_2$ can be differentiated) with respect to the 2-Wasserstein metric. The Riemannian interpretation was (formally) introduced by Otto [99], and rigorously established by Ambrosio et al. [12] and others; see Villani [125, Chapter 15] for further bibliography and more details.

Compatible measures (Definition 2.3.1) were implicitly introduced by Boissard et al. [28] in the context of *admissible optimal maps* where one defines families of gradients of convex functions (T_i) such that $T_j^{-1} \circ T_i$ is a gradient of a convex

function for any i and j. For (any) fixed measure $\gamma \in \mathscr{C}$, compatibility of $\mathscr{C}$ is then equivalent to admissibility of the collection of maps $\{\mathbf{t}_\gamma^\mu\}_{\mu\in\mathscr{C}}$. The examples we gave are also taken from [28].

Lemma 2.3.3 is from Cuesta-Albertos et al. [38, Theorem 2.9] (see also Zemel and Panaretos [135]).