

## Chapter 2

# The Nested Distance

In the present context of stochastic optimization we are interested in approximations of stochastic processes. To quantify the quality of an approximation, a concept of distance between stochastic processes is necessary. This is accomplished by the nested distance, which was introduced in Chap. 1 and is systematically treated in what follows. To this end we review different concepts of distances for probability measures first. The Wasserstein distance will be generalized to the nested distance between discrete time stochastic processes.

The distance of stochastic processes is based on the distance of the induced probability measures. There exists a broad variety of different concepts of distances on probability spaces in the literature. Some of them metricize convergence in probability or other variants of different topologies on random variables or probability measures. Rachev [105, 108] lists 76 metrics for measures, and many of them are adapted to concrete and particular problems.

A useful distance, which is adapted to stochastic optimization, should comprise various properties: it

- should measure distances of distributions and be independent of different, underlying probability spaces,
- should allow reasonable computational implementations,
- should represent a version of the weak\* topology<sup>1</sup> for random variables to enable approximations by discrete measures and, above all,
- should extend to general stochastic processes.

The Wasserstein distance, which is a solution of an optimization problem itself, covers the desired properties in a natural way. As an extra, there is a close, almost intimate relation between the Wasserstein distance and risk functionals. In addition,

---

<sup>1</sup>Recall that  $P_n \rightarrow P$  in the weak\* topology, if  $\int h dP_n \rightarrow \int h dP$  for all bounded and continuous functions  $h$ .

this distance is the basis for its multistage generalization, the nested distance and is therefore discussed in more detail below.

## 2.1 Distances of Probability Measures

In this section we work with ordinary probability distributions  $P$  on  $\mathbb{R}^m$ , say. When replacing a probability model  $P$  by another (typically simpler) model  $\tilde{P}$ , the basic question arises: how close is  $\tilde{P}$  to  $P$ ? Obviously, distances quantify the notion of closeness. We review here some ways of dealing with the concept of closeness for probability measures.

Let  $\mathcal{P}$  be a set of probability measures on  $\mathbb{R}^m$ .

**Definition 2.1.** A *semi-distance*  $d$  on  $\mathcal{P} \times \mathcal{P}$  satisfies the following three conditions:

(i) *Nonnegativity*: for all  $P_1, P_2 \in \mathcal{P}$ ,

$$d(P_1, P_2) \geq 0;$$

(ii) *Symmetry*: for all  $P_1, P_2 \in \mathcal{P}$ ,

$$d(P_1, P_2) = d(P_2, P_1);$$

(iii) *Triangle Inequality*: for all  $P_1, P_2, P_3 \in \mathcal{P}$ ,

$$d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2).$$

A semi-distance  $d(\cdot, \cdot)$  is called a *distance* if it satisfies the strictness property:

(iv) *Strictness*: if  $d(P_1, P_2) = 0$ , then  $P_1 = P_2$ .

### 2.1.1 Semi-Distances Generated by a Class of Test Functions

A general principle for defining semi-distances and distances consists in choosing a family of integrable functions  $\mathcal{H}$  (i.e., a family of functions such that the integral  $\int h(w) P(dw)$  exists for all  $P \in \mathcal{P}$ ) and defining

$$d_{\mathcal{H}}(P_1, P_2) := \sup_{h \in \mathcal{H}} \left| \int h dP_1 - \int h dP_2 \right|.$$

$d_{\mathcal{H}}$  is called the (semi-)distance *generated by*  $\mathcal{H}$ .

In general,  $\mathbf{d}_{\mathcal{H}}$  is only a semi-distance. If  $\mathcal{H}$  is *separating*, i.e., if for every pair  $P_1, P_2 \in \mathcal{P}$  there is a function  $h \in \mathcal{H}$  such that  $\int h \, \mathbf{d}P_1 \neq \int h \, \mathbf{d}P_2$ , then  $\mathbf{d}_{\mathcal{H}}$  is strict and thus is a distance.

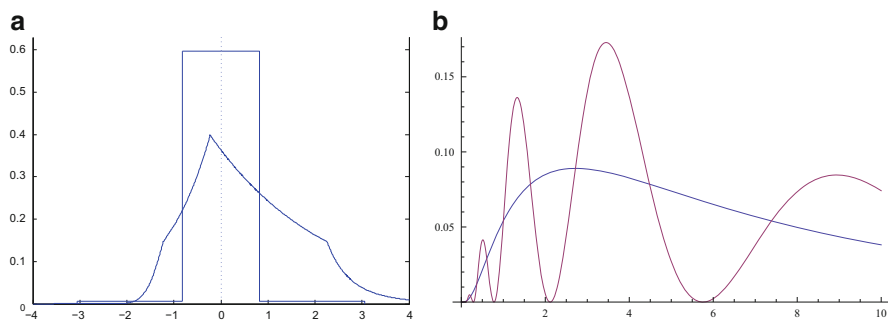
**The Moment Matching Semi-Distance.** Let  $\mathcal{P}_q$  be the set of all probability measures on  $\mathbb{R}^1$  which possess the  $q$ -th moment, i.e., for which  $\int \max\{1, |w|^q\} P(dw) < \infty$ . The moment matching semi-distance on  $\mathcal{P}_q$  is

$$\mathbf{d}_{M_q}(P_1, P_2) = \sup \left\{ \left| \int w^s P_1(dw) - \int w^s P_2(dw) \right| : s \in \{1, 2, \dots, q\} \right\}. \quad (2.1)$$

**The Moment Matching Caveat.** The moment matching semi-distance is not a distance, even if  $q$  is chosen to be large or even infinity. In fact, there are examples of different probability measures on  $\mathbb{R}^1$ , which have the same moments of all orders. For instance, there is a manifold of probability measures, which have all moments equal to those of the lognormal distribution, but are not lognormal (the lognormal distribution is often present in mathematical finance). Figure 2.1b displays two (of infinitely many) distributions with all moments coinciding, cf. Heyde [58]. Indeed, there are also distributions taking values on the negative axis having the same moments as the lognormal distribution, which itself has nonnegative support.

Ignoring these facts it is a widespread method in applications to match the first four moments, i.e., to work with  $\mathbf{d}_{M_4}$ . The following example displays two further densities, coinciding in their first four moments, but exhibiting very different properties in a drastic way.

*Example 2.2 (See [93]).* Let  $P_1$  and  $P_2$  be the two probability measures on  $\mathbb{R}$  with densities  $g_1$  and  $g_2$ , where



**Fig. 2.1** The moment matching caveat. (a) Two densities with identical first four moments. (b) Two densities with *all* moments coinciding

$$\begin{aligned}
g_1(w) &= 0.3988 [\exp(-|w + 0.2297|^3) \cdot \mathbb{1}_{\{w \leq -1.2297\}} \\
&\quad + \exp(-|w + 0.2297|) \cdot \mathbb{1}_{\{-1.2297 < w \leq -0.2297\}} \\
&\quad + \exp(-0.4024 \cdot (w + 0.2297)) \cdot \mathbb{1}_{\{-0.2297 < w \leq 2.2552\}} \\
&\quad + 1.0985 \cdot (0.4024w + 0.2925)^{-6} \cdot \mathbb{1}_{\{2.2552 < w\}}] \text{ and} \\
g_2(w) &= 0.5962 \cdot \mathbb{1}_{\{|w| \leq 0.8163\}} + 0.00595 \cdot \mathbb{1}_{\{0.8163 < |w| \leq 3.0588\}}
\end{aligned}$$

(see Fig. 2.1a). Both densities are unimodal and coincide in the first four moments, which are  $m_1 = 0$ ,  $m_2 = 0.3275$ ,  $m_3 = 0$ , and  $m_4 = 0.7230$  ( $m_q(P) = \int w^q dP(w)$ ). Their fifth moment, however, could not differ more: while the fifth moment of  $P_2$  is zero,  $P_1$  has infinite fifth moment. The density  $g_1$  is asymmetric, has a sharp cusp at  $-0.2297$  and unbounded support; in contrast,  $g_2$  is symmetric around 0, has a flat density there, has finite support, and possesses all moments. The distribution functions and quantiles differ drastically as well: we have that  $G_{P_1}(0.81) = 0.6257$  and  $G_{P_1}(-0.81) = 0.1098$ , while  $G_{P_2}(0.81) = 0.9807$  and  $G_{P_2}(-0.81) = 0.0133$ . Thus the probability of the interval  $[-0.81, 0.81]$  is only 51 % under  $P_1$ , while it is 95 % under  $P_2$ .

Summarizing, matching moments do not match the distributions. The moment matching semi-distance is not well suited for approximating probability distributions, since it is not fine enough to capture the relevant quality of an approximation (cf. also the additional Example 2.22 below.)

**Variational Distance.** The other extreme would be to choose as the generating class  $\mathcal{H}$  all measurable functions  $h$  such that  $|h| \leq 1$ . This class generates a distance, which is called the variational distance (more precisely, twice the variational distance). It is easy to see that if  $P_1$  (resp.  $P_2$ ) has density  $g_1$  (resp.  $g_2$ ), then

$$\begin{aligned}
&\sup \left\{ \left| \int h dP_1 - \int h dP_2 \right| : |h| \leq 1, h \text{ measurable} \right\} \\
&= \int |g_1(w) - g_2(w)| dw \\
&= 2 \cdot \sup \{ |P_1(A) - P_2(A)| : A \text{ a measurable set} \}.
\end{aligned}$$

The distance

$$d_V(P_1, P_2) := \sup \{ |P_1(A) - P_2(A)| : A \text{ a measurable set} \} \quad (2.2)$$

is called the *variational distance* between  $P_1$  and  $P_2$ .

The variational distance is a very fine distance, too fine for our applications: if  $P_1$  has a density and  $P_2$  sits on at most countably many points, then  $d_V(P_1, P_2) = 1$ , independently of the number of mass points of  $P_2$ . Thus there is no hope to approximate any continuous distribution by a discrete one with respect to the variational distance.

**Uniform Distance.** One may restrict the class of sets in (2.2) to a certain subclass. If one employs the class of half-unbounded rectangles in  $\mathbb{R}^m$  of the form  $(-\infty, w_1] \times (-\infty, w_2] \times \cdots \times (-\infty, w_m]$  one obtains the *uniform distance*, also called *Kolmogorov distance*

$$d_U(P_1, P_2) := \sup \{|G_{P_1}(w) - G_{P_2}(w)| : w \in \mathbb{R}^m\},$$

where  $G_P(\cdot)$  is the distribution function of  $P$ ,

$$G_P(w) = P\{(-\infty, w_1] \times \cdots \times (-\infty, w_m]\}.$$

Notice that a unit mass at point  $x$  and at point  $y$  are at a distance 1 both in the  $d_V$  distance and in the  $d_U$  distance, irrespective of how close  $x$  is to  $y$ . Especially when dealing with continuous baseline models and approximating discrete models, these distances are too fine.

**Bounded Lipschitz Distance.** Reducing the class  $\mathcal{H}$  to the class of all bounded, Lipschitz functions leads to the bounded Lipschitz metric, which metricizes the weak convergence of probability measures.

The *bounded Lipschitz distance* is defined as

$$d_{BL}(P_1, P_2) := \sup \left\{ \int h dP_1 - \int h dP_2 : |h(w)| \leq 1, |h(w) - h(v)| \leq \|w - v\| \right\},$$

it involves the class  $\mathcal{H}$  of functions  $h$  which are uniformly bounded by 1, and which are Lipschitz continuous with Lipschitz constant 1.

**Kantorovich Distance.** The *Kantorovich distance* (also *Wasserstein distance of order 1*, cf. Definition 2.4 below) is the bounded Lipschitz distance, where the requirement of boundedness of  $h$  is dropped:

$$d_1(P_1, P_2) := \sup \left\{ \int h dP_1 - \int h dP_2 : h(w) - h(v) \leq \|w - v\| \right\}.$$

This distance metricizes weak convergence on sets of probability measures which possess uniformly a first moment, as is elaborated in Theorem 2.23 below. On the real line, the Kantorovich metric may also be written as

$$d_1(P_1, P_2) = \int_{-\infty}^{\infty} |G_{P_1}(w) - G_{P_2}(w)| dw = \int_0^1 |G_{P_1}^{-1}(p) - G_{P_2}^{-1}(p)| dp, \quad (2.3)$$

where  $G_P^{-1}(p) = \inf\{w : G_P(w) \geq p\}$  (see Vallander [135]).

**Fortet–Mourier Distance.** If  $\mathcal{H}$  is the class of Lipschitz functions of order  $q$  ( $q$ -Lipschitz), the *Fortet–Mourier distance* is obtained:

$$\mathbf{d}_{\text{FM}_q}(P_1, P_2) := \sup \left\{ \int h dP_1 - \int h dP_2 : L_q(h) \leq 1 \right\}, \quad (2.4)$$

where the Lipschitz constant of order  $q$  is defined as

$$L_q(h) = \inf \left\{ L : |h(w) - h(v)| \leq L \cdot \|w - v\| \cdot \max(1, \|w\|^{q-1}, \|v\|^{q-1}) \right\}. \quad (2.5)$$

Notice that  $L_{q'}(h) \leq L_q(h)$  for  $q \leq q'$ ; in particular,  $L_q(h) \leq L_1(h)$  for all  $q \geq 1$  and therefore

$$\mathbf{d}_1(P_1, P_2) \leq \mathbf{d}_{\text{FM}_q}(P_1, P_2) \leq \mathbf{d}_{\text{FM}_{q'}}(P_1, P_2) \quad \text{for } 1 \leq q \leq q'.$$

The Fortet–Mourier distance metricizes weak convergence on sets of probability measures possessing uniformly a  $q$ -th moment. Notice that the function  $w \mapsto \|w\|^q$  is  $q$ -Lipschitz with Lipschitz constant  $L_q = q$ . On  $\mathbb{R}^1$ , the Fortet–Mourier distance may be equivalently written as

$$\mathbf{d}_{\text{FM}_q}(P_1, P_2) = \int \max \{1, |u|^{q-1}\} \cdot |G_{P_1}(u) - G_{P_2}(u)| du$$

(see Rachev [105, page 93]). For  $q = 1$ , the Fortet–Mourier distance coincides with the Kantorovich distance.

Further distances on probability measures and their relations can be found, e.g., in the review of Gibbs and Su [45].

## 2.2 The Wasserstein Distance

The Wasserstein distance generalizes the Kantorovich distance, although it is not generated by a set of test functions  $\mathcal{H}$  (except in special cases).

Importantly, the Wasserstein distance allows a generalization for stochastic processes. This generalization, the nested distance, is of particular interest in multistage stochastic optimization, and addressed in Sect. 2.10 below.

We adapt and augment the common concept of the Wasserstein distance here to prepare it for multistage stochastic optimization. For this we consider a general, real valued and measurable function

$$c: \Omega \times \tilde{\Omega} \rightarrow \mathbb{R} \quad (2.6)$$

linking two sample spaces  $\Omega$  and  $\tilde{\Omega}$ .

The function  $c$  is often associated with the interpretation that moving a particle  $\omega \in \Omega$  to  $\tilde{\omega} \in \tilde{\Omega}$  costs  $c(\omega, \tilde{\omega})$ , therefore  $c$  is often called a *cost function*.

The common definition of the Wasserstein distance considers the cost function

$$c(\cdot, \cdot) := \mathbf{d}(\cdot, \cdot)^r : \Omega \times \Omega \rightarrow \mathbb{R},$$

where  $\mathbf{d}$  is a distance on  $\Omega$  and  $r \geq 1$ . The notable difference is that the function  $c$  in (2.6) deals with two *different* spaces  $\Omega$  and  $\tilde{\Omega}$ , whereas the distance function  $\mathbf{d}$  involves just a single space, i.e.,  $\tilde{\Omega} = \Omega$ . In this situation the transportation costs  $c(\omega, \tilde{\omega})$  are assumed to be proportional to the transported distance  $\mathbf{d}(\omega, \tilde{\omega})$ , or to  $\mathbf{d}(\omega, \tilde{\omega})^r$ .

**Inheriting a Distance from Random Variables.** Typically, the probability space  $(\Omega, \mathcal{F}, P)$  does not carry a topology or distance. However, in all applications in this book, we assume that a distance or semi-distance is inherited on  $\Omega$  from a random variable  $\xi : \Omega \rightarrow \mathbb{R}^m$  by

$$\mathbf{d}(\omega_1, \omega_2) := \|\xi(\omega_1) - \xi(\omega_2)\|,$$

where  $\|\cdot\|$  is some norm in  $\mathbb{R}^m$ .<sup>2</sup> The notion can be extended to the case of two different probability spaces  $\Omega$  and  $\tilde{\Omega}$ :

**Definition 2.3.** If  $\xi$  is an  $\mathbb{R}^m$ -valued random variable on  $\Omega$  and  $\tilde{\xi}$  is an  $\mathbb{R}^m$ -valued random variable on  $\tilde{\Omega}$ , then the *inherited distance* between elements of  $\Omega$  and  $\tilde{\Omega}$  can be defined by the transportation cost function  $c(\omega, \tilde{\omega})$

$$\mathbf{d}(\omega, \tilde{\omega}) := c(\omega, \tilde{\omega}) = \mathbf{d}(\xi(\omega), \tilde{\xi}(\tilde{\omega})) \quad (2.7)$$

for some distance  $\mathbf{d}$  in  $\mathbb{R}^m$ ; often  $\mathbf{d}(w, v) = \|w - v\|$  for some norm  $\|\cdot\|$  in  $\mathbb{R}^m$ .

The transportation costs or distances between elements of  $\Omega$  and  $\tilde{\Omega}$  can be extended to transportation costs or distances between probabilities  $P$  on  $\Omega$  and  $\tilde{P}$  on  $\tilde{\Omega}$ .

**Definition 2.4 (Optimal Transportation Cost).** Given two probability spaces  $(\Omega, \mathcal{F}, P)$  and  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  and a transportation cost function  $c$ , the optimal transportation cost is

$$\inf_{\pi} \iint_{\Omega \times \tilde{\Omega}} c(\omega, \tilde{\omega}) \pi(d\omega, d\tilde{\omega}), \quad (2.8)$$

where the infimum is taken over all (bivariate) probability measures  $\pi$  on  $\Omega \times \tilde{\Omega}$  having the marginals  $P$  and  $\tilde{P}$ , that is

---

<sup>2</sup>Notice that it might happen that two different elements  $\omega_1$  and  $\omega_2$  are at distance 0, namely if  $\xi(\omega_1) = \xi(\omega_2)$ . In this case the distance is only a semi-distance, but it can as well be taken as the basis of a Wasserstein distance construction, which will then also turn out to be a semi-distance.

$$\pi(A \times \tilde{\Omega}) = P(A) \text{ and } \pi(\Omega \times B) = \tilde{P}(B) \quad (2.9)$$

for all measurable sets  $A \in \mathcal{F}$  and  $B \in \tilde{\mathcal{F}}$ . The optimal measure  $\pi$  is called the *optimal transport plan*. It exists under the conditions of Remark 2.6 below.

Specializing to the case where the costs are given by an inherited distance between elements of  $\Omega$  and  $\tilde{\Omega}$  one obtains the Wasserstein distance.

**Definition 2.5 (Wasserstein Distance).** The *Wasserstein distance* of order  $r$  ( $r \geq 1$ ) is

$$d_r(P, \tilde{P}) := \left( \inf_{\pi} \iint_{\Omega \times \tilde{\Omega}} d(\omega, \tilde{\omega})^r \pi(d\omega, d\tilde{\omega}) \right)^{1/r}, \quad (2.10)$$

where the infimum is among all joint probability measures  $\pi$  on  $\Omega \times \tilde{\Omega}$  (more precisely: on the product  $\mathcal{F} \otimes \tilde{\mathcal{F}}$  of the  $\sigma$ -algebras) which satisfy (2.9).

*Remark 2.6.* The infimum in (2.8) is attained, if both measures  $P$  and  $\tilde{P}$  are tight, i.e., for every  $\epsilon > 0$  there are compact sets  $K$  and  $\tilde{K}$  such that  $P(K^c) \leq \epsilon$  and  $\tilde{P}(\tilde{K}^c) \leq \epsilon$ .<sup>3</sup> Under this condition, the family of all measures  $\pi$  with marginals  $P$  and  $\tilde{P}$  is uniformly tight, since for all these measures

$$\pi((K \times \tilde{K})^c) \leq \pi(K^c \times \tilde{\Omega}) + \pi(\Omega \times \tilde{K}^c) \leq 2\epsilon,$$

i.e., is arbitrarily small if  $K$  and  $\tilde{K}$  are chosen appropriately. Closed families of uniformly tight probability measures are compact (Prohorov's Theorem, see, e.g., Parthasarathy and Kalyanapuram [85]). Since the integrand of (2.10) is continuous in  $\pi$ , the infimum is attained.

Definition 2.5 is used in two different situations:

- either there are given two abstract probability spaces  $(\Omega, \mathcal{F}, P)$  and  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  and two random variables  $\xi : \Omega \rightarrow \mathbb{R}^m$  and  $\tilde{\xi} : \tilde{\Omega} \rightarrow \mathbb{R}^m$  such that the distance is the induced distance according to Definition 2.3. In this case one may write

$$d_r(P, \tilde{P}) := \left( \inf_{\pi} \iint_{\Omega \times \tilde{\Omega}} d(\xi(\omega), \tilde{\xi}(\tilde{\omega}))^r \pi(d\omega, d\tilde{\omega}) \right)^{1/r},$$

where the infimum is over all joint probability measures with marginals  $P$  ( $\tilde{P}$ , resp.);

- or the two probabilities  $P$  and  $\tilde{P}$  are defined on  $\mathbb{R}^m$  endowed with a distance  $d$  (e.g.,  $d(u, v) = \|u - v\|$  for  $u, v \in \mathbb{R}^m$ ). In the latter case one may write

---

<sup>3</sup>  $K^c$  denotes the complement of the set  $K$ .



$$\mathbf{d}_r(P, \tilde{P}) := \left( \inf_{\pi} \iint_{\mathbb{R}^m \times \mathbb{R}^m} \mathbf{d}(u, v)^r \pi(du, dv) \right)^{1/r},$$

where the infimum is over all probability measures on  $\mathbb{R}^m \times \mathbb{R}^m$  with marginals  $P$  ( $\tilde{P}$ , resp.).

The second case can be seen as a special case of the first one considering the identical random variables  $\xi = \text{id}$  and  $\tilde{\xi} = \text{id}$ . Both cases are considered in the following. The context always makes clear whether we consider probabilities on abstract spaces endowed with the induced distance or their image measures on  $\mathbb{R}^m$ .

The collection of all probability measures  $P$ , which satisfy for some—and thus for any  $\omega_0 \in \Omega$ —the moment-like condition

$$\int_{\Omega} \mathbf{d}(\omega, \omega_0)^r P(d\omega) < \infty$$

is denoted by  $\mathcal{P}_r(\Omega; \mathbf{d})$ . It is immediate from the inequality

$$\mathbf{d}(\omega, \tilde{\omega})^r \leq 2^{r-1} (\mathbf{d}(\omega, \omega_0)^r + \mathbf{d}(\tilde{\omega}, \omega_0)^r)$$

(this is the triangle inequality when  $r = 1$ ) that the problem (2.10) is feasible and well defined whenever  $P \in \mathcal{P}_r(\Omega; \mathbf{d})$  and  $\tilde{P} \in \mathcal{P}_r(\tilde{\Omega}; \mathbf{d})$ , because the product measure<sup>4</sup>

$$\pi := P \otimes \tilde{P}$$

has the required marginals and

$$\mathbf{d}_r(P, \tilde{P})^r \leq \int_{\Omega} \int_{\tilde{\Omega}} \mathbf{d}(\omega, \tilde{\omega})^r P(d\omega) \tilde{P}(d\tilde{\omega}) < \infty.$$

Notice that if  $\mathbf{d}$  is inherited from  $\xi$  and the distance on  $\mathbb{R}^m$  is given by a norm  $\|\cdot\|$ , then  $P \in \mathcal{P}_r(\Omega; \mathbf{d})$  iff  $\int \|\xi(\omega)\|^r P(d\omega) < \infty$ , i.e., if  $\xi$  has finite  $r$ -th moment.

*Remark 2.7.* A comprehensive and intensive discussion of the Wasserstein distance is provided in the books by Rachev and Rüschendorf [107] and the book by Villani [137]. We shall use the properties that the infimum in (2.10) is actually attained, and  $\mathbf{d}_r(\cdot, \cdot)$  turns out to be a metric on the space  $\mathcal{P}_r(\Omega; \mathbf{d})$ .

---

<sup>4</sup> $(P \otimes \tilde{P})(A \times B) := P(A) \cdot \tilde{P}(B)$  defines a  $\sigma$ -additive measure due to the Hahn–Kolmogorov theorem.

*Remark 2.8 (Remark on Naming).* The terms in Definition 2.4 are not used consistently in the literature: in honor of G. Monge<sup>5</sup> (cf. [79]) and Leonid Kantorovich<sup>6</sup> (cf. [64]) the distance  $\mathbf{d}_r$  is sometimes called *Monge–Kantorovich distance* of order  $r$ . The term *Vasershtein distance*<sup>7</sup> appears the first time in Dobrushin [29].  $\mathbf{d}_2$  is sometimes called *quadratic Wasserstein distance*. Moreover, the distance  $\mathbf{d}_1$  is also called *Kantorovich–Rubinstein distance* and sometimes denoted by  $\mathbf{d}_{KA} := \mathbf{d}_1$ . In Russian literature the term Kantorovich distance (cf. Vershik [136]) is used instead of Wasserstein distance.

The terms in Definition 2.4 apparently became accepted in recent years, particularly due to Villani’s before-mentioned book [137] and other authors. We follow this general trend, in particular we reserve the term Kantorovich distance for  $\mathbf{d}_{KA} = \mathbf{d}_1$  ( $r = 1$ ).

**Notational Convenience.** We are using the symbol  $\mathbf{d}$  for the distance in the original space  $\Omega$ , and the same symbol  $\mathbf{d}_r(\cdot, \cdot)$  with subscript  $r$  to account for the distance on probabilities in  $\mathcal{P}_r(\Omega; \mathbf{d})$  induced by  $\mathbf{d}$ . This is justified in view of the following proposition, which identifies  $(\Omega, \mathbf{d})$  as a closed subspace of  $(\mathcal{P}_r, \mathbf{d}_r)$ .

**Proposition 2.9 (Embedding).** *It holds that*

$$\mathbf{d}_r(P, \delta_{\omega_0})^r = \int_{\Omega} \mathbf{d}(\omega, \omega_0)^r P(d\omega),$$

and the mapping

$$\begin{aligned} i: (\Omega, \mathbf{d}) &\rightarrow (\mathcal{P}_r(\Omega; \mathbf{d}), \mathbf{d}_r), \\ \omega &\mapsto \delta_{\omega} \end{aligned}$$

assigning to each point  $\omega \in \Omega$  its point measure  $\delta_{\omega}$  (Dirac measure<sup>8</sup>) is an isometric embedding for all  $1 \leq r < \infty$  ( $(\Omega, \mathbf{d}) \hookrightarrow \mathcal{P}_r(\Omega; \mathbf{d})$ ).

*Proof.* There is just one single measure with marginals  $P$  and  $\delta_{\omega_0}$ , which is the transport plan  $\pi = P \otimes \delta_{\omega_0}$ . Hence

$$\mathbf{d}_r(P, \delta_{\omega_0})^r = \int_{\Omega} \int_{\Omega} \mathbf{d}(\omega, \tilde{\omega})^r \delta_{\omega_0}(d\tilde{\omega}) P(d\omega) = \int_{\Omega} \mathbf{d}(\omega, \omega_0)^r P(d\omega),$$

the first assertion.

<sup>5</sup>Gaspard Monge (1746–1818) investigated how to efficiently construct dugouts.

<sup>6</sup>L. Kantorovich was awarded the price in Economic Sciences in Memory of Alfred Nobel in 1975.

<sup>7</sup>In honor of Leonid N. Vaserstein.

<sup>8</sup> $\delta_{\omega}(A) := \mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$  is the usual Dirac measure.

For the particular choice  $P = \delta_{\tilde{\omega}_0}$  the latter formula simplifies to

$$\mathbf{d}_r(\delta_{\tilde{\omega}_0}, \delta_{\omega_0})^r = \int_{\Omega} \mathbf{d}(\omega, \omega_0)^r \delta_{\tilde{\omega}_0}(d\omega) = \mathbf{d}(\tilde{\omega}_0, \omega_0)^r,$$

and hence  $\omega \mapsto \delta_{\omega}$  is an isometry.  $\square$

Notice that if  $\mathbf{d}$  is inherited by  $\xi$ , then  $\mathbf{d}_r(P, \delta_{\omega_0})^r = \int_{\Omega} \|\xi(\omega) - \xi(\omega_0)\|^r P(d\omega)$ .

## 2.3 Elementary Properties of the Wasserstein Distance

The following properties of the Wasserstein distance  $\mathbf{d}_r$  will be employed frequently.

**Lemma 2.10 (Monotonicity and Convexity).**

- (i) If  $r_1 \leq r_2$ , then  $\mathbf{d}_{r_1}(P, \tilde{P}) \leq \mathbf{d}_{r_2}(P, \tilde{P})$ .
- (ii) The Wasserstein distance is  $r$ -convex<sup>9</sup> in any of its components, that is for  $0 \leq \lambda \leq 1$  it holds that

$$\mathbf{d}_r(P, (1 - \lambda)P_0 + \lambda P_1)^r \leq (1 - \lambda)\mathbf{d}_r(P, P_0)^r + \lambda\mathbf{d}_r(P, P_1)^r,$$

and

$$\begin{aligned} \mathbf{d}_r(P, (1 - \lambda)P_0 + \lambda P_1) &\leq (1 - \lambda)^{\frac{1}{r}} \mathbf{d}_r(P, P_0) + \lambda^{\frac{1}{r}} \mathbf{d}_r(P, P_1) \\ &\leq \max\{\lambda, 1 - \lambda\}^{\frac{1}{r}-1} \cdot \left((1 - \lambda)\mathbf{d}_r(P, P_0) + \lambda\mathbf{d}_r(P, P_1)\right). \end{aligned} \quad (2.11)$$

- (iii)  $\mathbf{d}_r$  is a distance, it satisfies the triangle inequality  $\mathbf{d}_r(P, \tilde{P}) \leq \mathbf{d}_r(P, \tilde{\tilde{P}}) + \mathbf{d}_r(\tilde{\tilde{P}}, \tilde{P})$ .

*Remark 2.11.* Convexity in the traditional sense is actually achieved for the Kantorovich distance ( $r = 1$ ), it follows from (2.11) that

$$\mathbf{d}_1(P, (1 - \lambda)P_0 + \lambda P_1) \leq (1 - \lambda)\mathbf{d}_1(P, P_0) + \lambda\mathbf{d}_1(P, P_1).$$

For the general Wasserstein distance ( $r > 1$ ), however, a correction factor

$$1 \leq \max\{\lambda, 1 - \lambda\}^{\frac{1}{r}-1} \leq 2^{\frac{r-1}{r}} < 2$$

has to be accepted in (2.11).

---

<sup>9</sup>For the notion of  $r$ -concavity ( $r$ -convexity) see Dentcheva [129].

*Proof.* Observe that  $\frac{1}{\frac{r_2}{r_1}} + \frac{1}{\frac{r_2}{r_2-r_1}} = 1$ . By use of Hölder's inequality

$$\int \mathbf{d}^{r_1} d\pi = \int \mathbf{d}^{r_1} \cdot 1 d\pi \leq \left( \int \mathbf{d}^{r_1 \frac{r_2}{r_1}} d\pi \right)^{\frac{r_1}{r_2}} \cdot \left( \int 1^{\frac{r_2}{r_2-r_1}} d\pi \right)^{\frac{r_2-r_1}{r_2}} = \left( \int \mathbf{d}^{r_2} d\pi \right)^{\frac{r_1}{r_2}}.$$

Thus,  $\left( \int \mathbf{d}^{r_1} d\pi \right)^{\frac{1}{r_1}} \leq \left( \int \mathbf{d}^{r_2} d\pi \right)^{\frac{1}{r_2}}$  for every measure  $\pi$ , which proves the first assertion.

As for the second let  $\pi_0$  and  $\pi_1$  be measures chosen with adequate marginals in such way that the infimum is attained,

$$\mathbf{d}_r(P, P_0)^r = \int \mathbf{d}(\omega, \tilde{\omega})^r \pi_0(d\omega, d\tilde{\omega}) \text{ and } \mathbf{d}_r(P, P_1)^r = \int \mathbf{d}(\omega, \tilde{\omega})^r \pi_1(d\omega, d\tilde{\omega}).$$

The probability measure  $\pi_\lambda := (1-\lambda)\pi_0 + \lambda\pi_1$  then has the marginals  $P$  and  $P_\lambda := (1-\lambda)P_0 + \lambda P_1$ , and

$$\begin{aligned} \mathbf{d}_r(P, (1-\lambda)P_0 + \lambda P_1)^r & \\ & \leq \int \mathbf{d}(\omega, \tilde{\omega})^r \pi_\lambda(d\omega, d\tilde{\omega}) \\ & = (1-\lambda) \int \mathbf{d}(\omega, \tilde{\omega})^r \pi_0(d\omega, d\tilde{\omega}) + \lambda \int \mathbf{d}(\omega, \tilde{\omega})^r \pi_1(d\omega, d\tilde{\omega}) \\ & = (1-\lambda) \mathbf{d}_r(P, P_0)^r + \lambda \mathbf{d}_r(P, P_1)^r. \end{aligned}$$

The assertion follows from monotonicity and concavity of  $x \mapsto x^{\frac{1}{r}}$  and as  $(x+y)^{\frac{1}{r}} \leq x^{\frac{1}{r}} + y^{\frac{1}{r}}$ .

The other statements follow by employing Hölder's  $L^1 - L^\infty$  inequality. For (iii) we refer to the proof involving the gluing lemma in Villani [137].  $\square$

*Remark 2.12.* To note an important consequence: all functions are continuous with respect to  $\mathbf{d}_r$ , provided they are continuous with respect to  $\mathbf{d}_1 = \mathbf{d}_{KA}$ , the Kantorovich distance. A simple and useful example is provided by the following well-known lemma.

**Lemma 2.13.** *If the distance  $\mathbf{d}$  is inherited from  $\xi$  and  $\tilde{\xi}$  and based on a norm  $\|\cdot\|$  (see (2.7)), then*

$$\left\| \mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi}) \right\| \leq \mathbf{d}_r(P, \tilde{P}) \quad (2.12)$$

for  $r \geq 1$ .

In an alternative notation, let  $P_1$  ( $\tilde{P}_1$ , resp.) be probability measure on  $\mathbb{R}^m$  (for instance  $P_1 = P^\xi$ , the image or pushforward measure of  $P$ ) and let the point

$\mu_{P_1} := \mathbb{E}_{P_1}(\text{id}) = \int \xi P_1(d\xi)$  be the expectation (barycenter) of measure  $P_1$ <sup>10</sup> (provided it exists) and the same for  $\tilde{P}_1$ , then

$$\|\mu_{P_1} - \mu_{\tilde{P}_1}\| \leq \mathbf{d}_r(P_1, \tilde{P}_1).$$

*Proof.* The proof for the Kantorovich distance ( $r = 1$ ) is an application of Jensen's inequality as the norm is a convex function:

$$\begin{aligned} \|\mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi})\| &= \left\| \int \xi(\omega) P(d\omega) - \int \tilde{\xi}(\tilde{\omega}) \tilde{P}(d\tilde{\omega}) \right\| \\ &= \left\| \int (\xi - \tilde{\xi}) \pi(d\omega, d\tilde{\omega}) \right\| \leq \int \|\xi - \tilde{\xi}\| \pi(d\omega, d\tilde{\omega}). \end{aligned}$$

Taking the infimum over all measures  $\pi$  with appropriate marginals  $P$  and  $\tilde{P}$  gives the assertion, as  $\|\mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi})\| \leq \mathbf{d}_1(P, \tilde{P}) \leq \mathbf{d}_r(P, \tilde{P})$ .  $\square$

*Remark 2.14.* Formula (2.12) gives rise to the interpretation, that particles have to be transported—on average—at least the distance of the barycenters  $\mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi})$ .

### 2.3.1 The Wasserstein Distance on the Real Line

The Wasserstein distance for probability measures on the real line allows a closed form representation, which turns out to be useful in many situations. We cite the statement from Ambrosi et al. [3, Theorem 6.0.2], see also Vallander [135] and (2.3).

**Theorem 2.15.** *The Wasserstein distance of order  $r \geq 1$  for measures  $P$  and  $\tilde{P}$  on the real line  $\mathbb{R}$  is*

$$\mathbf{d}_r(P, \tilde{P})^r = \int_0^1 \left| G_P^{-1}(\alpha) - G_{\tilde{P}}^{-1}(\alpha) \right|^r d\alpha,$$

where  $G_P(y) = P((-\infty, y])$  is the associated cumulative distribution function and  $G_P^{-1}(\alpha) = \inf \{y : G_P(y) \geq \alpha\}$  its generalized inverse.

*Example 2.16 (Normal Distribution).* If  $P = N(\mu, \sigma^2)$  and  $\tilde{P} = N(\tilde{\mu}, \tilde{\sigma}^2)$  are normally distributed, then the explicit value for the Wasserstein distance of order  $r = 2$  is

$$\mathbf{d}_2(P, \tilde{P})^2 = (\mu - \tilde{\mu})^2 + (\sigma - \tilde{\sigma})^2,$$

---

<sup>10</sup> $\text{id}(\xi) := \xi$  is the identity.

while the Kantorovich distance is bounded by

$$d_1(P, \tilde{P}) \leq |\mu - \tilde{\mu}| + \sqrt{\frac{2}{\pi}} |\sigma - \tilde{\sigma}|.$$

The statement follows by considering  $G_P^{-1}(u) = \mu + \sigma \cdot \Phi^{-1}(u)$ , where  $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{1}{2}v^2} dv$  is the cdf of the standard normal distribution. It follows from Theorem 2.15 that

$$\begin{aligned} d_2(P, \tilde{P})^2 &= \int_0^1 (\mu - \tilde{\mu} + (\sigma - \tilde{\sigma}) \Phi^{-1}(u))^2 du \\ &= (\mu - \tilde{\mu})^2 + 2(\mu - \tilde{\mu})(\sigma - \tilde{\sigma}) \int_0^1 \Phi^{-1}(u) du \\ &\quad + (\sigma - \tilde{\sigma})^2 \int_0^1 (\Phi^{-1}(u))^2 du \\ &= (\mu - \tilde{\mu})^2 + (\sigma - \tilde{\sigma})^2. \end{aligned}$$

Moreover,

$$\begin{aligned} d_1(P, \tilde{P}) &= \int_0^1 |\mu + \sigma \Phi^{-1}(u) - \tilde{\mu} - \tilde{\sigma} \Phi^{-1}(u)| du \\ &\leq |\mu - \tilde{\mu}| + |\sigma - \tilde{\sigma}| \cdot \int_0^1 |\Phi^{-1}(u)| du \end{aligned}$$

provides the second assertion, as  $\int_0^1 |\Phi^{-1}(u)| du = \int_{-\infty}^{\infty} |u| \Phi'(u) du = \sqrt{\frac{2}{\pi}}$ .

*Example 2.17.* Evidently, explicit expressions are also available for even integer orders, an example is

$$d_4(P, \tilde{P})^4 = (\mu - \tilde{\mu})^4 + 6(\mu - \tilde{\mu})^2(\sigma - \tilde{\sigma})^2 + 3(\sigma - \tilde{\sigma})^4,$$

etc.

A further, general upper bound is provided by the following example.

*Example 2.18.* For two real valued random variables  $\xi \sim P$  and  $\tilde{\xi} \sim \tilde{P}$  with finite second moments, means  $\mu$  ( $\tilde{\mu}$ , resp.) and variances  $\sigma^2$  ( $\tilde{\sigma}^2$ , resp.), it follows from the elementary expansion

$$\begin{aligned} (x - y)^2 &= (\mu - \tilde{\mu})^2 + (x - \mu)^2 + (y - \tilde{\mu})^2 \\ &\quad - 2(xy - \mu\tilde{\mu}) + 2\mu(x - \mu) + 2\tilde{\mu}(y - \tilde{\mu}), \end{aligned}$$

together with (2.12), that

$$(\mu - \tilde{\mu})^2 \leq d_2(P, \tilde{P})^2 \leq (\mu - \tilde{\mu})^2 + \sigma^2 + \tilde{\sigma}^2, \quad (2.13)$$

because  $\pi := P \otimes \tilde{P}$  is a feasible bivariate measure. The upper bound (2.13) is rather conservative, although attained if one of the measures is a Dirac measure.

## 2.4 Alternative Distances as Basis for the Wasserstein Distance

### 2.4.1 The Role of the Distance on the Underlying Space

To every metric  $d$  on  $\mathbb{R}^m$  there corresponds a Wasserstein distance according to Definition 2.5. A special situation occurs for the discrete metric

$$d_0(u, v) := \begin{cases} 0 & \text{if } u = v \\ 1 & \text{if } u \neq v. \end{cases}$$

The set of all Lipschitz functions with respect to the discrete metric  $d_0$  coincides with the set of all measurable functions  $h$  such that  $0 \leq h \leq 1$  or its translates. Consequently the pertaining Kantorovich distance coincides with the variational distance (see (2.2))

$$d_1(P, \tilde{P} | d_0) = d_V(P, \tilde{P}),$$

(we write  $d_r(P, \tilde{P} | d_0)$  to emphasize the dependency on the metric  $d_0$  of the basic space).

### 2.4.2 Transformation of the Axis, and Fortet–Mourier Distances

Alternative metrics on  $\mathbb{R}^1$  are obtained by a nonlinear transform of the axis. Let  $\chi$  be any bijective, monotone transformation, which maps  $\mathbb{R}$  into  $\mathbb{R}$ . Then  $d_\chi(u, v) := |\chi(u) - \chi(v)|$  defines a new metric on  $\mathbb{R}^1$ . Notice that the family of functions, which are Lipschitz with respect to the distance  $d_\chi$  and to the Euclidean distance  $|u - v|$ , may be quite different.

To establish a relation between the Fortet–Mourier distance and a transformation on the real line we consider the bijective transformation (for  $q > 0$ )

$$\chi_q(u) = \begin{cases} u & \text{if } |u| \leq 1 \\ |u|^q \cdot \text{sign}(u) & \text{otherwise,} \end{cases} \quad (2.14)$$

which introduces the metric  $\mathbf{d}_{\chi_q}(u, v) = |\chi_q(u) - \chi_q(v)|$ . On bounded intervals the distances  $\chi_1(u, v) = |u - v|$  and  $\mathbf{d}_{\chi_q}$  are equivalent, since

$$|u - v| \leq |\chi_q(u) - \chi_q(v)| \leq q \cdot K^{q-1} |u - v| \quad \text{whenever } |u| \leq K \text{ and } |v| \leq K.$$

Denote by  $\mathbf{d}_1(\cdot, \cdot | \mathbf{d}_{\chi_q})$  the Kantorovich distance based on the distance  $\mathbf{d}_{\chi_q}$ ,

$$\mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) = \sup \left\{ \int h dP - \int h d\tilde{P} : |h(u) - h(v)| \leq \mathbf{d}_{\chi_q}(u, v) \right\}.$$

Notice that  $\mathbf{d}_{\chi_{q'}}(u, v) \leq \mathbf{d}_{\chi_q}(u, v)$  for  $q' < q$  and therefore

$$\mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_{q'}}) \leq \mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) \quad \text{for } q' < q. \quad (2.15)$$

Let  $P^{\chi_q}$  be the image measure of  $P$  under  $\chi_q$ , that is  $P^{\chi_q}(A) = P(\chi_{1/q}(A))$ , as  $\chi_q^{-1}(u) = \chi_{1/q}(u)$ , and note that  $P^{\chi_q}$  has distribution function

$$G_{P^{\chi_q}}(x) = G_P(\chi_{1/q}(x)),$$

where  $G_P$  is the distribution function of  $P$ . This leads to the identity

$$\mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) = \mathbf{d}_1(P^{\chi_q}, \tilde{P}^{\chi_q} | \mathbf{d}_{\chi_1}),$$

as  $d_{\chi_1}(u, v) = |u - v|$ .

To relate the Fortet–Mourier distance  $d_{M_q}$  to the distance  $\mathbf{d}_1(\cdot, \cdot | \mathbf{d}_{\chi_q})$  we show first the relations

$$L_q(h \circ \chi_q) \leq q \cdot L_1(h) \quad (2.16)$$

and

$$L_1(h \circ \chi_{1/q}) \leq 2 \cdot L_q(h) \quad (2.17)$$

for the Lipschitz constants of order  $q$  defined in (2.5).

Indeed, if  $L_1(h) < \infty$ , then

$$\begin{aligned} |h(\chi_q(u)) - h(\chi_q(v))| &\leq L_1(h) \cdot |\chi_q(u) - \chi_q(v)| \\ &\leq L_1(h) \cdot q \cdot \max\{1, |u|^{q-1}, |v|^{q-1}\} \cdot |u - v|, \end{aligned}$$



which implies (2.16). On the other hand, if  $L_q(h) < \infty$ , then (2.17) holds by

$$\begin{aligned} |h(\chi_{1/q}(u)) - h(\chi_{1/q}(v))| &\leq L_q(h) \cdot \max \{1, |\chi_{1/q}(u)|^{q-1}, |\chi_{1/q}(v)|^{q-1}\} \\ &\quad \cdot |\chi_{1/q}(u) - \chi_{1/q}(v)| \\ &\leq 2 \cdot L_q(h) \cdot |u - v|, \end{aligned}$$

where we have used that

$$\max \{1, |\chi_{1/q}(u)|^{q-1}, |\chi_{1/q}(v)|^{q-1}\} \cdot \frac{|\chi_{1/q}(u) - \chi_{1/q}(v)|}{|u - v|} \leq 2. \quad (2.18)$$

This latter inequality is clear if  $|v| \leq |u| \leq 1$ . If  $|v| \leq |u|$  and  $|u| > 1$ , then the left-hand side of (2.18) is bounded by 2.

As a consequence of (2.16) and (2.17) the relations

$$\begin{aligned} \frac{1}{q} \mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) &= \frac{1}{q} \mathbf{d}_1(G_P \circ \chi_{1/q}, G_{\tilde{P}} \circ \chi_{1/q}) \\ &\leq \mathbf{d}_{\text{FM}_q}(P, \tilde{P}) \\ &\leq 2 \mathbf{d}_1(G_P \circ \chi_{1/q}, G_{\tilde{P}} \circ \chi_{1/q}) = 2 \mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) \end{aligned} \quad (2.19)$$

are obtained.

One thus sees that the Fortet–Mourier distance of order  $q$  and the Kantorovich distance (i.e., the Fortet–Mourier distance of order 1) with the alternative metric  $\mathbf{d}_{\chi_q}$  are topologically equivalent.

A further relation can be based on the function  $\psi_r(u) = |u|^r \cdot \text{sign}(u)$  and the distance  $\mathbf{d}_{\psi_r}(u, v) = |\psi_r(u) - \psi_r(v)|$ . Notice that by an easy geometric consideration, for  $r \geq 1$ ,

$$|\psi_r(u) - \psi_r(v)| \geq 2 \left( \frac{|u - v|}{2} \right)^r$$

and therefore  $|u - v|^r \leq 2^{r-1} |\psi(u) - \psi(v)|$ , which implies that

$$\mathbf{d}_r(P, \tilde{P})^r \leq 2^{r-1} \cdot \mathbf{d}_1(P, \tilde{P} | \psi_r). \quad (2.20)$$

**Lemma 2.19.** *On the set of probability distributions, which have uniformly bounded  $r$ -th moments, the topologies generated by the distances  $\mathbf{d}_r$  and  $\mathbf{d}_1(\cdot, \cdot | \mathbf{d}_{\psi_r})$  are equivalent.*

*Proof.* Inequality (2.20) shows that  $\mathbf{d}_1(\cdot, \cdot | \psi_r)$  is finer than  $\mathbf{d}_r$ . For the inverse relation, let  $\xi \sim P$  and  $\tilde{\xi} \sim \tilde{P}$ . Notice that the Lipschitz constant of order  $r$  of  $\psi_r$  is  $L_r(\psi_r) = r$  and therefore

$$|\psi_r(\xi) - \psi_r(\tilde{\xi})| \leq r \cdot |\xi - \tilde{\xi}| \cdot \max \left\{ 1, |\xi|^{r-1}, |\tilde{\xi}|^{r-1} \right\}.$$

Using Hölder's inequality, we get<sup>11</sup>

$$\begin{aligned} \mathbb{E}|\psi_r(\xi) - \psi_r(\tilde{\xi})| &\leq r \mathbb{E}^{1/r} |\xi - \tilde{\xi}|^r \cdot \mathbb{E}^{\frac{r-1}{r}} \left[ (1 + |\xi|^{r-1} + |\tilde{\xi}|^{r-1})^{\frac{r}{r-1}} \right] \\ &\leq r \mathbb{E}^{1/r} |\xi - \tilde{\xi}|^r \cdot \left[ 1 + \mathbb{E}^{\frac{r-1}{r}} (|\xi|^r) + \mathbb{E}^{\frac{r-1}{r}} (|\tilde{\xi}|^r) \right] \end{aligned}$$

and consequently considering the minima with respect to the joint distribution of  $\xi$  and  $\tilde{\xi}$  one gets

$$\mathbf{d}_1(P, \tilde{P} | \psi_r) \leq r \cdot \mathbf{d}_r(P, \tilde{P}) \left( 1 + \mathbb{E}^{\frac{r-1}{r}} (|\xi|^r) + \mathbb{E}^{\frac{r-1}{r}} (|\tilde{\xi}|^r) \right),$$

which shows that  $\mathbf{d}_r$  is finer than  $\mathbf{d}_1(\cdot, \cdot | \psi_r)$ . □

## 2.5 Estimates Involving the Wasserstein Distance

In this section we ask the question: How close are some important statistical parameters, if the Wasserstein distances are small? Suppose that a probability distribution  $P$  on  $\mathbb{R}^m$  and some (typically discrete) approximation  $\tilde{P}$ , which is close to  $P$  in Wasserstein distance, are given. One may ask the following questions:

- Do  $P$  and  $\tilde{P}$  have a similar mean?
- Do  $P$  and  $\tilde{P}$  have a similar variance?
- If  $P$  and  $\tilde{P}$  are multidimensional, do  $P$  and  $\tilde{P}$  have a similar covariance matrix?
- Are the higher moments of  $P$  and  $\tilde{P}$  similar?

Precise answers to these questions are given below in Proposition 2.20. Some authors argue that a close approximation  $\tilde{P}$  to  $P$  should have at least the same first and second moments. Since we aim at approximating the distribution as a whole, there is not much reason in trying to match some specific moments (as it is done by *moment matching*, cf. Example 2.2 and (2.1)). In some applications one might be interested in the median and mean matching would not help. Also matching some Pearson correlation (product-moment correlation) would not help in matching Spearman's or Kendall's correlation.

If  $\xi \sim P$  and  $\tilde{\xi} \sim \tilde{P}$ , it is evident that for functions  $h$  with Lipschitz constant  $L$  the Wasserstein distance controls the distance of their integrals,

$$\left| \mathbb{E}[h(\xi)] - \mathbb{E}[h(\tilde{\xi})] \right| \leq L \cdot \mathbf{d}_1(P, \tilde{P}).$$

---

<sup>11</sup>We use the shorthand notation  $\mathbb{E}^P [\xi]$  for  $(\mathbb{E} [\xi])^P$ .

The following proposition collects results involving distances of probability measures and Lipschitz constants.

**Proposition 2.20 (Bounds Involving Lipschitz Constants).** *Assume that  $\xi \sim P$  and  $\tilde{\xi} \sim \tilde{P}$ . Then*

- (i)  $\left| \mathbb{E}\xi - \mathbb{E}\tilde{\xi} \right| \leq \mathbf{d}_1(P, \tilde{P}),$
- (ii)  $\left| \mathbb{E}|\xi| - \mathbb{E}|\tilde{\xi}| \right| \leq \mathbf{d}_1(P, \tilde{P}),$
- (iii)  $\left| \mathbb{E}(\xi - a)_+ - \mathbb{E}(\tilde{\xi} - a)_+ \right| \leq \mathbf{d}_1(P, \tilde{P}),$
- (iv)  $\left| \mathbb{E}(\xi^q) - \mathbb{E}(\tilde{\xi}^q) \right| \leq q \cdot \mathbf{d}_{\text{FM}_q}(P, \tilde{P})$  for integer  $q$  and
- (v)  $\left| \mathbb{E}(|\xi|^q) - \mathbb{E}(|\tilde{\xi}|^q) \right| \leq q \cdot \mathbf{d}_{\text{FM}_q}(P, \tilde{P}).$

*Proof.* The functions  $u \mapsto u$ ,  $u \mapsto |u|$  and  $u \mapsto (u - a)_+$  are Lipschitz continuous (with Lipschitz constant 1). For the proof of (iv) and (v) recall the definition of the Fortet–Mourier distance  $\mathbf{d}_{\text{FM}_q}(P, \tilde{P})$  in (2.4) and use the fact that the Lipschitz constant of order  $q$  (see (2.5)) of  $x \mapsto x^q$  is  $L_q = q$ . The same is true for the function  $x \mapsto |x|^q$ .  $\square$

The following proposition collects examples to demonstrate how the Wasserstein distance controls also higher moments, provided that they exist.

**Proposition 2.21 (Wasserstein Distance Controls All Moments).** *Assume that  $\xi \sim P$  and  $\tilde{\xi} \sim \tilde{P}$ . Then*

- (i)  $\left| \mathbb{E}|\xi|^p - \mathbb{E}|\tilde{\xi}|^p \right| \leq p \cdot \mathbf{d}_r(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{r-1}{r}} \left[ |\xi|^{r \cdot \frac{p-1}{r-1}} \right], \mathbb{E}^{\frac{r-1}{r}} \left[ |\tilde{\xi}|^{r \cdot \frac{p-1}{r-1}} \right] \right\},$
- (ii)  $\left| \mathbb{E}(\xi^p) - \mathbb{E}(\tilde{\xi}^p) \right| \leq p \cdot \mathbf{d}_r(P, \tilde{P}) \cdot \left\{ \mathbb{E}^{\frac{r-1}{r}} \left[ |\xi|^{r \cdot \frac{p-1}{r-1}} \right] + \mathbb{E}^{\frac{r-1}{r}} \left[ |\tilde{\xi}|^{r \cdot \frac{p-1}{r-1}} \right] \right\}$  for  $p$  an integer,
- (iii)  $\left| \mathbb{E}\xi^2 - \mathbb{E}\tilde{\xi}^2 \right| \leq 2 \cdot \mathbf{d}_2(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{1}{2}} \left[ \xi^2 \right], \mathbb{E}^{\frac{1}{2}} \left[ \tilde{\xi}^2 \right] \right\},$
- (iv)  $\left| \mathbb{E}|\xi|^r - \mathbb{E}|\tilde{\xi}|^r \right| \leq r \cdot \mathbf{d}_r(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{r-1}{r}} \left[ |\xi|^r \right], \mathbb{E}^{\frac{r-1}{r}} \left[ |\tilde{\xi}|^r \right] \right\}$  and
- (v)  $\left| \mathbb{E}|\xi|^p - \mathbb{E}|\tilde{\xi}|^p \right| \leq p \cdot \mathbf{d}_2(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{1}{2}} \left[ |\xi|^{2(p-1)} \right], \mathbb{E}^{\frac{1}{2}} \left[ |\tilde{\xi}|^{2(p-1)} \right] \right\},$

where  $p \geq 1$  and  $r > 1$ .

*Proof.* By convexity of the function  $x \mapsto |x|^p$  for  $p \geq 1$  it holds that

$$|\tilde{x}|^p \geq |x|^p + p \cdot \text{sign}(x) |x|^{p-1} (\tilde{x} - x),$$

and consequently

$$|\xi|^p - |\tilde{\xi}|^p \leq p \cdot \text{sign}(\xi) |\xi|^{p-1} (\xi - \tilde{\xi}) \leq p \left| \xi - \tilde{\xi} \right| |\xi|^{p-1}.$$

Taking expectations with respect to  $\pi$ , where  $\pi$  has marginals  $P$  and  $\tilde{P}$ , and employing Hölder's inequality for the conjugate parameters  $\frac{1}{r} + \frac{1}{r'} = 1$  (i.e.,  $r' = \frac{r}{r-1}$ ) reveals that

$$\mathbb{E} |\xi|^p - \mathbb{E} |\tilde{\xi}|^p \leq p \cdot \left\| \xi - \tilde{\xi} \right\|_r \cdot \left\| |\xi|^{p-1} \right\|_{r'} = p \cdot \left\| \xi - \tilde{\xi} \right\|_r \cdot \left\| \xi \right\|_{r, \frac{p-1}{r-1}}^{p-1}.$$

Taking the infimum with of all bivariate probability measures  $\pi$  with marginals  $P$  and  $\tilde{P}$  it follows that

$$\mathbb{E} |\xi|^p - \mathbb{E} |\tilde{\xi}|^p \leq p \cdot \mathbf{d}_r(P, \tilde{P}) \cdot \mathbb{E}^{r-1} \left[ |\xi|^{r \cdot \frac{p-1}{r-1}} \right].$$

The assertion (i) follows now for general  $r > 1$  and  $p > 1$  by interchanging  $\xi$  and  $\tilde{\xi}$ .

The second inequality has only to be proved for odd  $p$  since for even  $p$  it is a consequence of (i). Using the monotonicity of the odd function  $x \mapsto p \cdot x^{p-1}$  one gets

$$\xi^p - \tilde{\xi}^p \leq p \cdot \left( |\xi|^{p-1} + |\tilde{\xi}|^{p-1} \right) \left| \xi - \tilde{\xi} \right|,$$

and in analogy to the proof of (i) the inequality (ii) follows.

The other assertions can be derived from (i) as special cases ( $r = p = 2$ ,  $r = p$ , etc.).  $\square$

Further inequalities of the type addressed in Proposition 2.20 and in Proposition 2.21 can be derived, if one considers the Wasserstein norms with alternative distances on  $\mathbb{R}$ . Again, let  $\xi \sim P$  and  $\tilde{\xi} \sim \tilde{P}$ . Using the functions (see (2.14))

$$\chi_q(u) = \begin{cases} u & \text{if } |u| \leq 1 \\ |u|^q \operatorname{sign}(u) & \text{otherwise} \end{cases}$$

and noticing (2.15) one gets that for  $q' \leq q$

$$\mathbb{E} \left[ \max \left\{ |\xi|, |\xi|^{q'} \right\} \right] \leq \mathbf{d}_1(P, \tilde{P} | d_{\chi_q}).$$

Also, using the Fortet–Mourier metric (2.4) we get the basic inequality

$$\left| \mathbb{E}[h(\xi)] - \mathbb{E}[h(\tilde{\xi})] \right| \leq L_q(h) \cdot \mathbf{d}_{\text{FM}_q}(P, \tilde{P}),$$

where  $L_q$  is the Lipschitz constant of order  $q$  (see (2.5)). A special case is

$$\left| \mathbb{E} |\xi|^{q'} - \mathbb{E} |\tilde{\xi}|^{q'} \right| \leq q \cdot \mathbf{d}_{\text{FM}_q}(P, \tilde{P}) \leq 2q \cdot \mathbf{d}_1(P, \tilde{P} | d_{\chi_q})$$

for  $q' < q$ , since the Lipschitz constant of order  $q$  of  $x \mapsto |x|^q$  equals  $q$  (cf. also inequality (2.19)).

*Example 2.22 (Approximation of a Bivariate Normal Distribution).* As an illustration consider the best approximation of a normal distribution

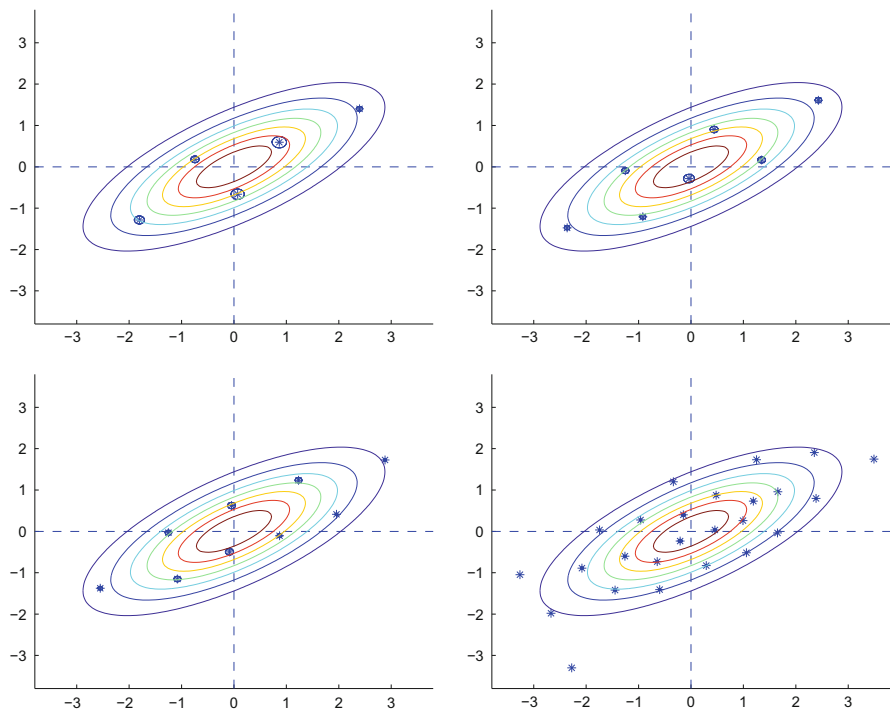
$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right) \quad (2.21)$$

by a discrete distribution located at  $s$  points in  $\mathbb{R}^2$ . Notice that we may generate random variates from this distribution by

$$\xi_1 = Z_1 + Z_2,$$

$$\xi_2 = Z_2,$$

where  $Z_1$  and  $Z_2$  are independent standard normals. We approximate these distributions by discrete ones, sitting on  $s = 5, 7, 9$ , and  $25$  points. The discrete approximations are displayed in Fig. 2.2, where the little circles around each point



**Fig. 2.2** Discrete approximations of a two-dimensional normal distribution with 5, 7, 9, and 25 points. Some statistical parameters of these approximations are given in Table 2.1

**Table 2.1** Approximations of a bivariate normal distribution (2.21) by 5, 7, 9, and 25 points. Their Wasserstein distance (of order 1) is shown in the first column

	Distance	$\mathbb{E}(\xi_1)$	$\mathbb{E}(\xi_2)$	$\text{Var}(\xi_1)$	$\text{Var}(\xi_2)$	Cov	$\mathbb{E}(\xi_1^3)$	$\mathbb{E}(\xi_1^4)$
True value		0.	0.	2.	1.	1.	0.	12.
5 points	0.587	0.056	−0.064	1.59	0.76	0.96	0.54	6.06
7 points	0.454	0.043	0.004	1.95	0.87	1.105	0.47	8.63
9 points	0.359	−0.03	−0.009	1.90	0.83	1.02	0.22	9.52
25 points	0.147	0.01	0.01	2.04	0.99	1.07	0.41	11.6

(a) Approximation quality of selected moments. Cov is the covariance between  $\xi_1$  and  $\xi_2$

	Distance	$\text{Med}(\xi_1)$	$\text{Med}(\xi_2)$	$\mathbb{E} \xi_1 $	$\mathbb{E} \xi_2 $	Spm	$P(\xi_1 > \xi_2)$
True value		0.	0.	1.128	0.797	0.695	0.50
5 points	0.587	0.07	0.18	1.005	0.773	0.88	0.642
7 points	0.454	−0.03	0.094	1.11	0.73	0.83	0.60
9 points	0.359	−0.08	−0.03	1.07	0.75	0.76	0.502
25 points	0.147	−0.14	0.03	1.14	0.81	0.76	0.53

(b) Continuation of Table 2.1a: here the medians, the first absolute moments, the Spearman correlation coefficient (Spm) as well as the probability of a particular event are shown

symbolize the respective probability mass (these approximations were found by the Stochastic Approximation (SA) Algorithm 4.5, which is explained in Chap. 4). The results for comparison are collected in Table 2.1.

It is not claimed that these approximations are optimal, however, they are good ones. As one can see, they approximate the true distribution in many aspects, not just for the first two or four moments. While it is not possible to derive from closeness with respect to some moments the closeness with respect to other aspects, the closeness in the Wasserstein distance implies closeness for moments and other statistics in a natural way (cf. also Example 2.2).

## 2.6 Approximations in the Wasserstein Metric

This subsection provides the foundations for approximations of probability measures by probability measures with finite support. This is quite relevant, because only probability measures with finite support are eligible for numerical computations and algorithmic treatment.

Suppose that the supports of all considered probabilities are contained in some closed set  $\Xi \subseteq \mathbb{R}^m$ , which is endowed with some metric  $d$ . The elements of  $\Xi$  are denoted by  $\xi$  (these are here points and not random variables). We discuss the important and necessary theorems. The precise proofs of some results of this subsection are beyond the scope of this book, we give the references instead.

The crucial tool to identify the topology induced by the metric  $d_r$  with the topology of weak\* convergence is the uniform tightness condition (2.22) below.

**Theorem 2.23 (Wasserstein Metricizes the Weak\* Topology).** *Let  $(P_n)_{n \geq 1}$  be a sequence of measures in  $\mathcal{P}_r(\Xi)$ , and let  $P \in \mathcal{P}_r(\Xi)$ . Then the following are equivalent:*

- (i)  $d_r(P_n, P) \xrightarrow{n \rightarrow \infty} 0$ ,
- (ii)  $P_n \xrightarrow{n \rightarrow \infty} P$  in weak\* sense, and  $P_n$  satisfies the following uniform tightness condition: for some (and thus any)  $\xi_0 \in \Omega$ ,

$$\limsup_{n \rightarrow \infty} \int_{\{d(\xi_0, \xi) \geq R\}} d(\xi_0, \xi)^r P_n(d\xi) \xrightarrow{R \rightarrow \infty} 0. \quad (2.22)$$

*Proof.* For a proof we refer to Theorem 7.12 in Villani [137].  $\square$

**Remark 2.24.** One may always replace the metric  $d$  by the uniformly bounded distance  $d'(\xi, \tilde{\xi}) := \frac{d(\xi, \tilde{\xi})}{1+d(\xi, \tilde{\xi})}$  or  $d'(\xi, \tilde{\xi}) := \min\{1, d(\xi, \tilde{\xi})\}$  without changing the topology of  $\Xi$ . In this situation, however, the uniform tightness condition (2.22) is trivial, and  $d'_r$  thus metricizes weak\* convergence on the whole of  $\mathcal{P}_r(\Xi)$ .

The following theorem is essential for our intentions to approximate probability measures by measures with finite support.

**Theorem 2.25.** *If  $(\Xi, d)$  is separable, then  $(\mathcal{P}_r(\Xi), d_r)$  is separable and all measures  $\sum_{\xi \in \tilde{\Xi}} P_\xi \cdot \delta_\xi$  with finite support  $\tilde{\Xi} \subset \Xi$  ( $P_\xi \geq 0$  and  $\sum_{\xi \in \tilde{\Xi}} P_\xi = 1$ ) are dense.*

*Proof.* A proof by elementary means is contained in Bolley [13]. Initial proofs of the statement, however, involve the weaker Prohorov distance and deep results of Kolmogorov; cf. Ambrosi et al. [3].  $\square$

To complete the essential characteristics we mention that the space  $(\mathcal{P}_r(\Xi; d), d_r)$  is not only separable and metrizable, but also complete, hence a Polish space.

**Theorem 2.26.** *Let  $(\Xi, d)$  be a Polish space, then  $(\mathcal{P}_r(\Xi; d), d_r)$  is a Polish space again.*

*Proof.* The space is metrizable and separability is established by Theorem 2.25. Completeness is proved in Bolley [13].  $\square$

## 2.7 The Wasserstein Distance in a Discrete Framework

In many applications and in implementations the measures considered are discrete measures (measures with finite support) of the form  $P = \sum_{i=1}^n P_i \delta_{\xi_i}$  (where  $P_i \geq 0$ ,  $\sum_{i=1}^n P_i = 1$  and the support  $\{\xi_i: i = 1, 2, \dots, n\} \subset \Xi$  is finite).

Given two discrete measures  $P = \sum_{i=1}^n P_i \delta_{\xi_i}$  and  $\tilde{P} = \sum_{j=1}^{\tilde{n}} \tilde{P}_j \delta_{\tilde{\xi}_j}$  the computation of the Wasserstein distance (2.10) corresponds to solving the linear program

$$\begin{aligned} & \text{minimize} && \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\ & (\text{in } \pi) \end{aligned} \tag{2.23}$$

$$\begin{aligned} & \text{subject to} && \sum_{j=1}^{\tilde{n}} \pi_{i,j} = P_i \quad (i = 1, 2, \dots, n), \\ & && \sum_{i=1}^n \pi_{i,j} = \tilde{P}_j \quad (j = 1, 2, \dots, \tilde{n}), \\ & && \pi_{i,j} \geq 0, \end{aligned} \tag{2.24}$$

where  $d_{i,j} = d(\xi_i, \tilde{\xi}_j)$  is an  $n \times \tilde{n}$ -matrix carrying the distances. The  $n \times \tilde{n}$ -matrix  $\pi_{i,j}$  in (2.23) corresponds to the bivariate probability measure

$$\pi = \sum_{i,j} \pi_{i,j} \cdot \delta_{(\xi_i, \tilde{\xi}_j)}$$

on the product  $\Xi \times \tilde{\Xi}$ ;  $\pi$  is a probability measure as  $\pi_{i,j} \geq 0$  and  $\sum_{i,j} \pi_{i,j} = \sum_i \sum_j \pi_{i,j} = \sum_i P_i = 1$ .

Figure 2.3 exhibits the structure of this linear program (2.23), where the matrix can be written in the form

$$\begin{pmatrix} \mathbb{1}_{\tilde{n}} \otimes I_n \\ I_{\tilde{n}} \otimes \mathbb{1}_n \end{pmatrix} \begin{pmatrix} \pi_{1,1} \\ \vdots \\ \pi_{n,1} \\ \pi_{1,2} \\ \vdots \\ \pi_{n,\tilde{n}} \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_n \\ \tilde{P}_1 \\ \vdots \\ \tilde{P}_{\tilde{n}} \end{pmatrix}$$

$\underbrace{\begin{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{pmatrix}}_n & \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{pmatrix}}_n & \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{pmatrix}}_n \\ \vdots & \vdots & \vdots \\ \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{pmatrix}}_n & \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{pmatrix}}_n & \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \end{pmatrix}}_n \end{pmatrix}_{\tilde{n}} \begin{pmatrix} \pi_{1,1} \\ \vdots \\ \pi_{n,1} \\ \pi_{1,2} \\ \vdots \\ \pi_{n,\tilde{n}} \end{pmatrix}$

**Fig. 2.3** Structure of the linear constraints of the linear program (2.23). The  $(n + \tilde{n}) \times (n \cdot \tilde{n})$  matrix is totally unimodular



( $\otimes$  denotes the Kronecker product,  $\mathbb{1}_n = \underbrace{(1, 1, \dots, 1)}_{n \text{ times}}$  and  $I_n$  is the  $n \times n$ -identity matrix). From this figure it becomes evident that the constraints are *linearly dependent*, because the sum of the first  $n$  lines equals the sum of the following  $\tilde{n}$  lines. As a consequence, one of all  $n + \tilde{n}$  constraints in (2.23) can be removed. For efficiency reasons in numerical implementations a line *should* be removed for most numerical solvers.

It follows from complementary slackness conditions of linear programs that the optimal transport plan  $\pi$  in (2.23) is sparse, it has at most  $n + \tilde{n} - 1$  nonzero entries, because (2.23) has not more than  $n + \tilde{n} - 1$  linearly independent equality constraints.

*Remark 2.27 (Transport Plans and Their Relation to Bipartite Graphs).* One may define the bipartite graph  $G = (U \cup V, E)$  with distinct nodes

$$U = \{\xi_i : i = 1, \dots, n\} \text{ and } V = \{\tilde{\xi}_j : j = 1, \dots, \tilde{n}\}$$

and vertices  $E = \{(\xi_i, \tilde{\xi}_j) : \pi_{i,j} > 0, i = 1, \dots, n, j = 1, \dots, \tilde{n}\}$ . The linear constraints in (2.23) correspond to the incidence matrix of this graph  $G$ , which is a *totally unimodular* matrix (i.e., every square non-singular submatrix is invertible over the integers, cf. Hoffman and Krukskal [60]). It follows from Cramer's rule that each entry of the matrix  $\pi$  has the specific form

$$\pi_{i,j} = \sum_{k=1}^n \epsilon_{i,j}^k P_k + \sum_{\ell=1}^{\tilde{n}} \tilde{\epsilon}_{i,j}^\ell \tilde{P}_\ell,$$

where  $\epsilon_{i,j}^k, \tilde{\epsilon}_{i,j}^\ell \in \{-1, 0, 1\}$ .

## 2.8 Duality for the Wasserstein Metric

The linear program (2.23) to compute  $d_r(P, \tilde{P})$  naturally—as any linear program—has a dual linear program. It is given by

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^n P_i \lambda_i + \sum_{j=1}^{\tilde{n}} \tilde{P}_j \mu_j \\ & (\text{in } \lambda, \mu) \end{aligned} \tag{2.25}$$

$$\text{subject to } \lambda_i + \mu_j \leq d_{i,j}^r \quad \text{for all } i = 1, \dots, n \text{ and } j = 1, \dots, \tilde{n}. \tag{2.26}$$

By the vanishing duality gap of the primal (2.23) and its dual (2.25) it follows that

$$\sum_{i=1}^n \sum_{j=1}^{\tilde{n}} \pi_{i,j} d_{i,j}^r \leq \sum_{i=1}^n P_i \lambda_i + \sum_{j=1}^{\tilde{n}} \tilde{P}_j \mu_j = \sum_{i=1}^n \sum_{j=1}^{\tilde{n}} \pi_{i,j} (\lambda_i + \mu_j) \leq \sum_{i=1}^n \sum_{j=1}^{\tilde{n}} \pi_{i,j} d_{i,j}^r,$$

from which further follows, by (2.24) and (2.26), that

$$\pi_{i,j} (\lambda_i + \mu_j) = \pi_{i,j} \cdot d_{i,j}^r,$$

which is the complementary slackness condition.

Recalling the fact that  $P = \sum_{i=1}^n P_i \delta_{\xi_i}$  ( $\tilde{P} = \sum_{j=1}^{\tilde{n}} \tilde{P}_j \delta_{\tilde{\xi}_j}$ , resp.) one may extend the dual variables

$$\lambda(\xi) := \begin{cases} \lambda_i & \text{if } \xi = \xi_i \\ -\infty & \text{else} \end{cases} \quad \text{and} \quad \mu(\tilde{\xi}) := \begin{cases} \mu_j & \text{if } \tilde{\xi} = \tilde{\xi}_j \\ -\infty & \text{else.} \end{cases}$$

Then the dual program (2.25) can be rewritten as

$$\begin{aligned} & \text{maximize (in } \lambda, \mu) \mathbb{E}_P \lambda + \mathbb{E}_{\tilde{P}} \mu \\ & \text{subject to } \lambda(\xi) + \mu(\tilde{\xi}) \leq \mathbf{d}(\xi, \tilde{\xi})^r \text{ for all } \xi \in \Xi \text{ and } \tilde{\xi} \in \tilde{\Xi}, \end{aligned} \tag{2.27}$$

and the complementary slackness reads

$$\pi \left( \left\{ (\xi, \tilde{\xi}) : \lambda(\xi) + \mu(\tilde{\xi}) = \mathbf{d}(\xi, \tilde{\xi})^r \right\} \right) = 1.$$

This means that

$$\lambda(\xi) + \mu(\tilde{\xi}) = \mathbf{d}(\xi, \tilde{\xi})^r \quad \pi \text{ almost everywhere,}$$

the inequality in (2.27) is thus replaced by equality on the support set of the optimal measure  $\pi$ .

A pair  $(\lambda, \mu)$  of feasible dual variables can moreover be replaced by  $(\lambda, \lambda^*)$  or  $(\mu^*, \mu)$ , where

$$\lambda^*(\tilde{\xi}) := \inf_{\xi \in \Xi} \mathbf{d}(\xi, \tilde{\xi})^r - \lambda(\xi)$$

and

$$\mu^*(\xi) := \inf_{\tilde{\xi} \in \tilde{\Xi}} \mathbf{d}(\xi, \tilde{\xi})^r - \mu(\tilde{\xi}),$$

because

$$\lambda(\xi) + \mu(\tilde{\xi}) \leq \lambda(\xi) + \lambda^*(\tilde{\xi}) \leq \mathbf{d}(\xi, \tilde{\xi})^r$$

and

$$\lambda(\xi) + \mu(\tilde{\xi}) \leq \mu^*(\xi) + \mu(\tilde{\xi}) \leq \mathbf{d}(\xi, \tilde{\xi})^r.$$

For an arbitrary function  $\lambda$  the pair  $(\lambda, \lambda^*)$  is feasible. By the same reasoning, given  $\mu$ , the pair  $(\mu^*, \mu)$  is feasible. This gives an improved objective, as

$$\mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\mu) \leq \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) \leq \mathbf{d}_r(P, \tilde{P})^r, \quad (2.28)$$

and analogously for the pair  $(\mu, \mu^*)$ .

**Rapid Computation of the Wasserstein Distance.** The cascading property (2.28) can be exploited in algorithms to quickly compute the Wasserstein distance of discrete probability measures. By duality the objective of both problems,

$$\text{maximize (in } \lambda) \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) \quad \text{and} \quad \text{maximize (in } \mu) \mathbb{E}_P(\mu^*) + \mathbb{E}_{\tilde{P}}(\mu), \quad (2.29)$$

is  $\mathbf{d}_r(P, \tilde{P})^r$ , but the dimension of the vector  $\lambda$  (or  $\mu$ ) in (2.29) is much smaller than the dimension of the matrix  $\pi$  in the primal (2.23). The problems (2.29) are unconstrained, nonlinear, and the objectives

$$\lambda \mapsto \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) \quad \text{and} \quad \mu \mapsto \mathbb{E}_P(\mu^*) + \mathbb{E}_{\tilde{P}}(\mu)$$

are moreover concave. In addition a subdifferential (an element of the subgradient) of the objective with respect to  $\lambda$  and  $\mu$  is available, as

$$\frac{\partial}{\partial \lambda_i} \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) = P_i - \tilde{P}_j \quad \text{and} \quad \frac{\partial}{\partial \mu_j} \mathbb{E}_P(\mu^*) + \mathbb{E}_{\tilde{P}}(\mu) = \tilde{P}_j - P_i,$$

where for the first equation  $i \in \operatorname{argmin}_k \{d_{k,j}^r - \lambda_k\}$  and for the second one  $j \in \operatorname{argmin}_k \{d_{i,k}^r - \mu_k\}$ , so that equality holds in the duality equations  $\lambda_j^* = d_{i,j}^r - \lambda_i$  and  $\mu_i^* = d_{i,j}^r - \mu_j$ . The nonlinear conjugate gradient method (cf. Ruszczyński [119]) is an appropriate choice to compute successive improvements of the unconstrained problems (2.29).

This algorithmic approach to compute  $\mathbf{d}_r(P, \tilde{P})$  notably provides the dual variables  $\lambda$  and  $\mu$  and the distance, but not the primal solution  $\pi$ . However, the primal  $\pi$  is supported only at points  $(i, j)$  with  $\lambda_i^* + \mu_j^* = d_{i,j}^r$ . This can be exploited to determine the primal variable  $\pi$  in a dual–primal step.

*Example 2.28.* As an example we consider two discrete distributions on  $\mathbb{R}^3$ , whose probability mass functions on the vectors

$$\begin{pmatrix} x_{i,0} \\ x_{i,1} \\ x_{i,2} \end{pmatrix}$$

are given by

$$P = \left[ \begin{array}{cccccccccccc} 0.02 & 0.04 & 0.08 & 0.06 & 0.21 & 0.09 & 0.15 & 0.06 & 0.09 & 0.08 & 0.12 \\ \begin{pmatrix} 10 \\ 13 \\ 15 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 14 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 13 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 11 \end{pmatrix} & \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix} & \begin{pmatrix} 10 \\ 11 \\ 9 \end{pmatrix} & \begin{pmatrix} 10 \\ 8 \\ 10 \end{pmatrix} & \begin{pmatrix} 10 \\ 8 \\ 8 \end{pmatrix} & \begin{pmatrix} 10 \\ 8 \\ 6 \end{pmatrix} & \begin{pmatrix} 10 \\ 6 \\ 7 \end{pmatrix} & \begin{pmatrix} 10 \\ 6 \\ 5 \end{pmatrix} \end{array} \right]$$

and

$$\tilde{P} = \left[ \begin{array}{cccccc} 0.12 & 0.18 & 0.30 & 0.16 & 0.16 & 0.08 \\ \begin{pmatrix} 10 \\ 13 \\ 14 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 12 \end{pmatrix} & \begin{pmatrix} 10 \\ 11 \\ 10 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 9 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 8 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 5 \end{pmatrix} \end{array} \right].$$

The matrix in Table 2.2 collects the distances  $\mathbf{d}_{i,j} = \sum_{t=0}^2 |x_{i,t} - x_{j,t}|$ . Later, these vectors will be interpreted as the values on the paths of a tree (see Fig. 2.12), but for the simple Wasserstein distance as we discuss it here, the treestructure is irrelevant.

The solutions of the Wasserstein problem (2.23) and its dual (2.25) are displayed in the Table 2.3.

**Table 2.2** The distance matrix with entries  $\mathbf{d}_{i,j}$  from the Example 2.28. The optimal transportation plan sits only on the 16 pairs which are italicized

Distance $\mathbf{d}_{i,j}$	1	2	3	4	5	6
1	<b><i>1</i></b>	3	7	12	13	16
2	<b><i>0</i></b>	2	6	11	12	15
3	<b><i>1</i></b>	<b><i>1</i></b>	5	10	11	14
4	3	<b><i>1</i></b>	3	8	9	12
5	4	<b><i>2</i></b>	<b><i>2</i></b>	7	8	11
6	7	5	<b><i>1</i></b>	4	5	8
7	9	7	<b><i>3</i></b>	<b><i>2</i></b>	3	6
8	11	9	5	2	<b><i>1</i></b>	4
9	13	11	7	<b><i>4</i></b>	3	2
10	14	12	8	<b><i>3</i></b>	<b><i>2</i></b>	3
11	16	14	10	5	<b><i>4</i></b>	<b><i>1</i></b>

**Table 2.3** The solutions of the primal and the dual Wasserstein problem

Probabilities, $\pi_{i,j}$	0.12	0.18	0.30	0.16	0.16	0.08
0.02	<b>0.02</b>	0	0	0	0	0
0.04	<b>0.04</b>	0	0	0	0	0
0.08	<b>0.06</b>	<b>0.02</b>	0	0	0	0
0.06	0	<b>0.06</b>	0	0	0	0
0.21	0	<b>0.10</b>	<b>0.11</b>	0	0	0
0.09	0	0	<b>0.09</b>	0	0	0
0.15	0	0	<b>0.10</b>	<b>0.05</b>	0	0
0.06	0	0	0	0	<b>0.06</b>	0
0.09	0	0	0	<b>0.09</b>	0	0
0.08	0	0	0	<b>0.02</b>	<b>0.06</b>	0
0.12	0	0	0	0	<b>0.04</b>	<b>0.08</b>

(a) The transportation plan  $\pi$  solving the primal Wasserstein problem (2.23) for the two distributions given in Example 2.28,  $d_1(P, \tilde{P}) = \sum_{i,j} \pi_{i,j} d_{i,j} = 1.91$

Dual variables	$\lambda_i + \mu_j$	$\mu$					
		6	6	6	5	4	1
$\lambda$	−5	<b>I</b>	1	1	0	−1	−4
	−6	<b>0</b>	0	0	−1	−2	−5
	−5	<b>I</b>	<b>I</b>	1	0	−1	−4
	−5	1	<b>I</b>	1	0	−1	−4
	−4	2	<b>2</b>	<b>2</b>	1	0	−3
	−5	1	1	<b>I</b>	0	−1	−4
	−3	3	3	<b>3</b>	<b>2</b>	1	−2
	−3	3	3	3	2	<b>I</b>	−2
	−1	5	5	5	<b>4</b>	3	0
	−2	4	4	4	<b>3</b>	2	−1
	0	6	6	6	0	<b>4</b>	<b>I</b>

(b) The variables  $\lambda$  (leftmost column) and  $\mu$  (upmost row) solving the dual Wasserstein problem (2.27) for the two distributions given in Example 2.28. Their sum  $\lambda_i + \mu_j$  is shown as matrix elements. They satisfy  $\lambda_i + \mu_j \leq d_{i,j}$ , with equality in the shaded cells (cf. Table 2.2), and  $\sum_i P_i \lambda_i + \sum_j \tilde{P}_j \mu_j = 1.91$

## 2.9 Continuity of the Dual Variables, and the Kantorovich–Rubinstein Theorem

To investigate the continuity of the dual variables define the diameter  $\Delta := \sup_{\xi \in \Xi, \tilde{\xi} \in \tilde{\Xi}} d(\xi, \tilde{\xi})$  ( $\Delta$  may be unbounded, but is bounded for discrete measures and even by 1 for the distances discussed in Remark 2.24).

By convexity of the function  $x \mapsto x^r$  it holds that

$$d(\xi_2, \tilde{\xi})^r \geq d(\xi_1, \tilde{\xi})^r + r d(\xi_1, \tilde{\xi})^{r-1} (d(\xi_2, \tilde{\xi}) - d(\xi_1, \tilde{\xi})),$$

from which follows that

$$\begin{aligned} d(\xi_1, \tilde{\xi})^r - \mu(\tilde{\xi}) - (d(\xi_2, \tilde{\xi})^r - \mu(\tilde{\xi})) &\leq r d(\xi_1, \tilde{\xi})^{r-1} (d(\xi_1, \tilde{\xi}) - d(\xi_2, \tilde{\xi})) \\ &\leq r d(\xi_1, \tilde{\xi})^{r-1} d(\xi_1, \xi_2) \end{aligned} \quad (2.30)$$

by the triangle inequality,  $d(\xi_1, \tilde{\xi}) \leq d(\xi_2, \tilde{\xi}) + d(\xi_1, \xi_2)$ . As one may assume by (2.27) that  $\lambda(\xi) = \inf_{\tilde{\xi}} d(\tilde{\xi}, \xi)^r - \mu(\tilde{\xi})$ , it follows that

$$\lambda(\xi_1) - (d(\xi_2, \tilde{\xi})^r - \mu(\tilde{\xi})) \leq r \Delta^{r-1} d(\xi_1, \xi_2),$$

and thus

$$\lambda(\xi_1) - \lambda(\xi_2) \leq r \Delta^{r-1} d(\xi_1, \xi_2).$$

By interchanging the roles of  $\xi_1$  and  $\xi_2$  it follows that  $\lambda$  is continuous with Lipschitz constant  $r \Delta^{r-1}$ —provided that the diameter is bounded,  $\Delta < \infty$ . The same reasoning as above can be repeated to verify that  $\mu$  is Lipschitz continuous as well with the same Lipschitz constant.

**Kantorovich–Rubinstein Theorem.** A particular situation arises for the Kantorovich distance (i.e., the Wasserstein of order  $r = 1$ ). It follows from (2.30) directly that

$$\lambda(\xi_1) - \lambda(\xi_2) \leq d(\xi_1, \xi_2),$$

that is the dual functions  $\lambda$  and  $\mu$  are Lipschitz continuous with constant 1, irrespective of the diameter (notice as well that  $r \Delta^{r-1} = \Delta^0 = 1$  whenever  $r = 1$ ).

Moreover,

$$-\lambda(\xi) \leq \inf_{\tilde{\xi}} d(\tilde{\xi}, \xi) - \lambda(\tilde{\xi}) \leq -\lambda(\xi)$$

by Lipschitz-1 continuity and by choosing  $\tilde{\xi} = \xi$ , hence  $\mu(\xi) = -\lambda(\xi)$ . This is the content of the Kantorovich–Rubinstein Theorem.

**Theorem 2.29 (Kantorovich–Rubinstein Theorem).** *Let  $(\Xi, d)$  be a Polish space, then*

$$d_1(P, \tilde{P}) = \sup_{\lambda} \mathbb{E}_P \lambda - \mathbb{E}_{\tilde{P}} \lambda,$$

where the supremum is among all Lipschitz continuous functions  $\lambda$ , i.e.,

$$\sup_{\xi \neq \tilde{\xi}} \frac{\lambda(\xi) - \lambda(\tilde{\xi})}{d(\xi, \tilde{\xi})} \leq 1,$$

which are integrable with respect to  $P$  and  $\tilde{P}$ .

## 2.10 Multistage Generalization: The Nested Distance

Multistage optimization problems do not consider just one single stage, but, as its name indicates and as was outlined in the introduction, multiple and subsequent stages. In mathematical terms it is not a single random variable which has to be considered, but an entire stochastic process instead. To this end, let  $(\Omega, \mathcal{F}, P)$  and  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  be two probability spaces and let  $\xi : \Omega \rightarrow \Xi$  and  $\tilde{\xi} : \tilde{\Omega} \rightarrow \Xi$  be two random variables with common image space  $\Xi \subseteq \mathbb{R}^m$ , which is endowed with a metric  $d$ . We assume that  $(\Xi, d)$  is a Polish space and the Wasserstein distance  $d_r$  on  $\mathcal{P}_r(\Xi)$  is well defined. This distance will now be extended for stochastic processes  $\xi_t$  defined on a filtered probability space  $(\Omega, \mathfrak{F} = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_T), P)$ <sup>12</sup> and another process  $\tilde{\xi}_t$  defined on  $(\tilde{\Omega}, \tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_0, \tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_T), \tilde{P})$ .

### 2.10.1 The Inherited Distance

Consider a stochastic process

$$\xi_t : \Omega \rightarrow (\Xi_t, d_t), \quad t = 0, 1, \dots, T,$$

with possibly different state spaces  $(\Xi_t, d_t)$  for every  $t = 1, \dots, T$ . The value  $\xi_0$  is considered as being deterministic. These random variables  $(\xi_t)_{t=0}^T$  can be compounded to a single random variable  $\xi$  via

$$\begin{aligned} \xi : \Omega &\rightarrow \Xi_0 \times \Xi_1 \times \dots \times \Xi_T \\ \omega &\mapsto (\xi_0(\omega), \dots, \xi_T(\omega)), \end{aligned} \quad (2.31)$$

where each  $\omega$  is mapped to its path (the trajectory) in the state space  $\Xi := \Xi_0 \times \Xi_1 \times \dots \times \Xi_T$ . This setting generalizes the usual definition of a stochastic process as the state spaces of the partial observations

---

<sup>12</sup>Often also called a *stochastic basis*.

$$\xi_t = \text{proj}_t \circ \xi : \Omega \rightarrow \Xi_t \quad t = 0, 1, \dots, T$$

may differ at different times ( $\text{proj}_t : \Xi \rightarrow \Xi_t$  is the natural projection).

For  $\xi$  a process as in (2.31),  $P^\xi$  can be considered again.  $P^\xi = P \circ \xi^{-1}$  is called the *law of the process*  $\xi$ , it is a probability measure on the product  $\Xi := \Xi_0 \times \Xi_1 \times \dots \times \Xi_T$ .

Now note that any of the spaces  $\Xi_t$  are equipped with a distance function  $\mathbf{d}_t$ , and there are many metrics  $\mathbf{d}$  such that  $(\Xi, \mathbf{d})$  is a metric space. Given two processes  $\xi$  resp.  $\tilde{\xi}$  (with the same state spaces  $\Xi_t$ ) on  $\Omega$  ( $\tilde{\Omega}$ , resp.), a (semi-)distance is inherited to  $\Omega \times \tilde{\Omega}$  in an analogous way as in Definition 2.3, for example by

$$\mathbf{d}(\omega, \tilde{\omega}) := \sum_{t=0}^T w_t \mathbf{d}_t(\xi_t(\omega), \tilde{\xi}_t(\tilde{\omega})), \quad (2.32)$$

the weighted  $\ell^1$ -distance (with weights  $w_t > 0$ ), or

$$\mathbf{d}(\omega, \tilde{\omega}) := \left( \sum_{t=0}^T w_t \mathbf{d}_t(\xi_t(\omega), \tilde{\xi}_t(\tilde{\omega}))^2 \right)^{\frac{1}{2}}, \quad (2.33)$$

the  $\ell^2$ -distance, or

$$\mathbf{d}(\omega, \tilde{\omega}) := \max_{t=0, \dots, T} w_t \mathbf{d}_t(\xi_t(\omega), \tilde{\xi}_t(\tilde{\omega})),$$

the  $\ell^\infty$ -distance.

For any of these choices  $\mathbf{d}$  is a cost function or (semi-)distance on  $\Omega \times \tilde{\Omega}$ , and the Wasserstein distance

$$\mathbf{d}_r(P^\xi, P^{\tilde{\xi}}) \quad (2.34)$$

of the laws is available.

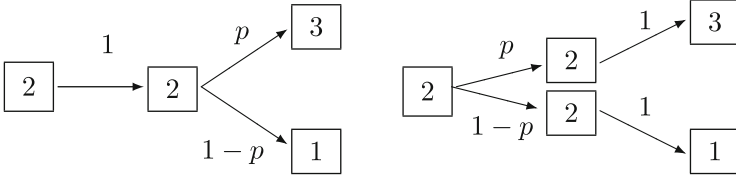
The following example elaborates that this simple application of the Wasserstein distance (2.34) is not suitable yet to distinguish stochastic processes.

*Example 2.30.* To observe the hidden caveat for the (final) Wasserstein distance consider the example depicted in Fig. 2.4. Two processes are shown there, which have the same states. The paths of successive observations, for both processes, are (2, 2, 3) or (2, 2, 1). Each path has the same probability in both processes ( $p$ , and  $1 - p$ , resp.).

The Wasserstein distance of these processes is simply 0: indeed, the state space is

$$\Xi = \tilde{\Xi} = \{(2, 2, 1), (2, 2, 3)\}$$





**Fig. 2.4** Two processes with identical final probabilities and identical states. The second process reveals already at an earlier stage that the final observation will be 3 (1, resp.). The nested distance of the trees is  $4p(1-p)$

for both processes. The distance matrix then is  $d = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$  and  $\pi = \begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}$  is a feasible transport plan. Hence, the Wasserstein distance of the laws  $P$  and  $\tilde{P}$  of the processes is  $d_r(P, \tilde{P}) = \sum_{i,j} d_{i,j} \pi_{i,j} = 0$ .

However, the processes  $P$  and  $\tilde{P}$  depicted in Fig. 2.4 are certainly not the same processes: having observed the partial path  $(2, 2)$  in the second process, we already know whether the final observation will be 1 or 3. This knowledge (information) is *not* available for the first process. So as the distance was identified to be  $d_r(P, \tilde{P}) = 0$ , the Wasserstein distance, in its genuine form, does not qualify as a distance for filtered stochastic processes.

The reason why the Wasserstein distance does not detect this difference is because it does not take conditional probabilities into account ( $\mathcal{F}_t$  for  $t = 0, 1, \dots, T-1$ ), but only final probabilities, where the sigma algebras coincide,  $\mathcal{F}_T = \tilde{\mathcal{F}}_T$ . But the sigma algebras differ at stage 1 (cf. (1.21)),

$$\mathcal{F}_1 = \sigma(\{(2, 2, 1), (2, 2, 3)\}) \subsetneq \sigma(\{(2, 2, 1)\}, \{(2, 2, 3)\}) = \tilde{\mathcal{F}}_1.$$

**Definition 2.31 (The Filtration Induced by the Process).** The *history process* is

$$\xi_{0:t} := \text{proj}_{0:t} \circ \xi := (\text{proj}_0 \circ \xi, \dots, \text{proj}_t \circ \xi) = (\xi_0, \dots, \xi_t),$$

that is  $\xi_{0:t}(\omega) := (\xi_0(\omega), \dots, \xi_t(\omega)) \in \Xi_0 \times \dots \times \Xi_t$ .

The history process generates the *natural filtration* of the process  $\xi$ ,

$$\mathfrak{F}^\xi = \left( \mathcal{F}_t^\xi \right)_{t=0}^T, \quad \mathcal{F}_t^\xi := \sigma(\{\xi_{0:t}^{-1}(A_0 \times \dots \times A_t) : A_s \in \mathcal{B}(\Xi_s)\}),$$

where  $\mathcal{B}(\Xi_s)$  denotes the Borel sets on  $\Xi_s$ . Notice that the relation  $\xi \triangleleft \mathfrak{F}$  implies that the filtration  $\mathfrak{F}$  is finer than  $\mathfrak{F}^\xi$ .

### 2.10.2 The Nested Distance

The nested distance is based on the Wasserstein distance. Extending the Wasserstein distance to stochastic processes the nested distance takes notice of all sigma algebras contained in the filtrations of the filtered probability spaces.

**Definition 2.32 (The Nested Distance).** The *nested distance* of order  $r \geq 1$  of two filtered probability spaces  $\mathbb{P} = (\Omega, (\mathcal{F}_t), P)$  and  $\tilde{\mathbb{P}} = (\tilde{\Omega}, (\tilde{\mathcal{F}}_t), P)$ , for which a distance  $\mathbf{d} : \Omega \times \tilde{\Omega} \rightarrow \mathbb{R}$  is defined, is the optimal value of the optimization problem

$$\begin{aligned} & \underset{(\text{in } \pi)}{\text{minimize}} && \left( \int \mathbf{d}(\omega, \tilde{\omega})^r \pi(d\omega, d\tilde{\omega}) \right)^{\frac{1}{r}} \\ & \text{subject to} && \pi(A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t) \quad (A \in \mathcal{F}_t, t \in \mathbf{T}), \\ & && \pi(\Omega \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \tilde{P}(B \mid \tilde{\mathcal{F}}_t) \quad (B \in \tilde{\mathcal{F}}_t, t \in \mathbf{T}), \end{aligned} \quad (2.35)$$

where the infimum in (2.35) is among all bivariate probability measures  $\pi \in \mathcal{P}(\Omega \times \tilde{\Omega})$  which are defined on  $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$ <sup>13</sup> and  $\mathbf{T} = \{0, 1 \dots T\}$ . Its optimal value, the nested distance, is denoted by

$$\mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}}).$$

*Remark 2.33.* The nested distance is often called *multistage distance* or *process distance* as well. A feasible measure  $\pi$  is called a nested transport plan.

The nested distance was initially constructed on nested distributions (cf. Definition 1.7), both were introduced by Pflug in [92]. The definition given here notably applies for continuous time,  $\mathbf{T} = \{t \in \mathbb{R} : t \geq 0\}$  as well.

The Markov-constructions contained in Rüschendorf [116] can be compared with the nested distribution for two stages, such that the distance on Markov-constructions can be considered as a special case of the definition provided here.

The multistage formulation presented here is based on filtrations. The following discussion of the nested distance is adapted from [94].

**Discussion of the Nested Distance.** We recall first that the conditional probability is defined by the conditional expectation by  $P(A \mid \mathcal{F}_t) = \mathbb{E}(\mathbb{1}_A \mid \mathcal{F}_t)$  for every  $A \in \mathcal{F}_T$ , it is thus a random variable itself,

$$P(A \mid \mathcal{F}_t) = \mathbb{E}(\mathbb{1}_A \mid \mathcal{F}_t) : \Omega \rightarrow [0, 1],$$

which is measurable with respect to  $\mathcal{F}_t$ . Its characterizing property is

<sup>13</sup>  $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$  is the smallest sigma-algebra on the product space  $\Omega \times \tilde{\Omega}$ , which contains all rectangles  $A \times \tilde{A}$  for  $A \in \mathcal{F}$ ,  $\tilde{A} \in \tilde{\mathcal{F}}$ .

$$\int_B P(A | \mathcal{F}_t) dP = \int_B P(A | \mathcal{F}_t)(\omega) P(d\omega) = P(A \cap B) \quad (A \in \mathcal{F}_T, B \in \mathcal{F}_t).$$

The identity

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t)$$

thus expresses that

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)(\omega, \tilde{\omega}) = P(A | \mathcal{F}_t)(\omega) \quad \pi \text{ almost everywhere}$$

for every  $A \in \mathcal{F}_T$ . The right-hand side of this equation is notably independent of  $\tilde{\omega}$ . It is sometimes helpful to make this independence explicit by using the notations

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) \circ \text{id} = P(A | \mathcal{F}_t)(\text{id}),$$

where  $\text{id}$  is the projection  $\text{id} : \Omega \times \tilde{\Omega} \rightarrow \Omega$ ,  $\text{id}(\omega, \tilde{\omega}) = \omega$  and  $\tilde{\text{id}}(\omega, \tilde{\omega}) = \tilde{\omega}$ , respectively.

*Remark 2.34.* Two stochastic processes  $\xi_t : \Omega \rightarrow \Xi_t$  and  $\tilde{\xi}_t : \Omega \rightarrow \tilde{\Xi}_t$  on the same probability space  $(\Omega, \mathcal{F}; P)$  induce the filtered probability spaces  $\mathbb{P}^\xi := (\Omega, \mathfrak{F}^\xi, P^\xi)$  and  $\mathbb{P}^{\tilde{\xi}} := (\Omega, \tilde{\mathfrak{F}}^\xi, P^{\tilde{\xi}})$ , for which the nested distance is available, provided that there is a cost function

$$\mathbf{d} : (\Xi_0 \times \dots \times \Xi_T) \times (\tilde{\Xi}_0 \times \dots \times \tilde{\Xi}_T) \rightarrow \mathbb{R}. \quad (2.36)$$

This justifies the name *process distance*. In addition it should be repeated that the state spaces  $\Xi_t$  and  $\tilde{\Xi}_t$  do not necessarily have to coincide. Then  $\mathbf{d}$  in (2.36) is more a cost function than a distance function. Notice that the inherited distance defined in (2.7) is rather a cost function too.

*Remark 2.35 (The Initial Stage,  $t = 0$ ).* For the trivial sigma-algebra  $\mathcal{F}_0 = \{\emptyset, \Omega\}$ ,  $P(A | \mathcal{F}_0)$  is deterministic (a constant) and satisfies  $P(A | \mathcal{F}_0) = P(A)$  (almost everywhere).

*Remark 2.36 (The Final Stage,  $t = T$ ).* For  $A \in \mathcal{F}_T$  it holds that  $P(A | \mathcal{F}_T) = \mathbb{E}(\mathbb{1}_A | \mathcal{F}_T) = \mathbb{1}_A$  and  $\pi(A \times \tilde{\Omega} | \mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) = \mathbb{1}_{A \times \tilde{\Omega}}$ . But as  $\mathbb{1}_A \circ \text{id} = \mathbb{1}_{A \times \tilde{\Omega}}$  always holds true it follows that the constraints in (2.35) are redundant for  $t = T$ , they can be omitted.

**Lemma 2.37.** *Let  $(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}$  and  $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi}) \sim \tilde{\mathbb{P}}$  be nested distributions with  $\mathcal{F}_0 = \{\emptyset, \Omega\}$  and  $\tilde{\mathcal{F}}_0 = \{\emptyset, \tilde{\Omega}\}$ .<sup>14</sup> The product measure  $\pi := P \otimes \tilde{P}$  is*

<sup>14</sup>If not otherwise specified,  $\mathbf{d}$  is always the distance inherited from  $\xi$  and  $\tilde{\xi}$ .

feasible for the multistage distance (i.e., the nested distance is well defined). It holds moreover that

$$\mathbf{d}_r(P, \tilde{P})^r \leq \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r \leq \mathbb{E}_{P \otimes \tilde{P}}(\mathbf{d}^r). \quad (2.37)$$

*Proof.* The first inequality follows in view of Remark 2.35. We shall verify that all constraints in (2.35) are satisfied for  $\pi := P \otimes \tilde{P}$ . For this choose  $C \in \mathcal{F}_I$  and  $D \in \tilde{\mathcal{F}}_I$  and observe that, for  $\pi = P \otimes \tilde{P}$ ,

$$\begin{aligned} & \int_{C \times D} P(A | \mathcal{F}_I)(\text{id}) \cdot \tilde{P}(B | \mathcal{F}_I)(\tilde{\text{id}}) d\pi \\ &= \int_C P(A | \mathcal{F}_I)(\text{id}) dP \cdot \int_D \tilde{P}(B | \mathcal{F}_I)(\tilde{\text{id}}) d\tilde{P} \\ &= P(A \cap C) \cdot \tilde{P}(B \cap D) \\ &= \pi((A \cap C) \times (B \cap D)) = \pi((A \times B) \cap (C \times D)) \\ &= \int_{C \times D} \pi(A \times B | \mathcal{F}_I \otimes \tilde{\mathcal{F}}_I) d\pi. \end{aligned}$$

It follows that the conditional probabilities  $P(A | \mathcal{F}_I) \circ \text{id} \cdot \tilde{P}(B | \mathcal{F}_I) \circ \tilde{\text{id}}$  and  $\pi(A \times B | \mathcal{F}_I \otimes \tilde{\mathcal{F}}_I)$  are ( $\pi$ -almost everywhere) identical, as equality holds for any sets  $C \in \mathcal{F}_I$  and  $D \in \tilde{\mathcal{F}}_I$ . By choosing  $A = \Omega$  ( $B = \tilde{\Omega}$ , resp.) it follows that  $\pi = P \otimes \tilde{P}$  is feasible and hence  $\mathbf{d}_r(P, \tilde{P})^r \leq \mathbb{E}_{P \otimes \tilde{P}}(\mathbf{d}^r)$ .  $\square$

The following example demonstrates that convergence with respect to the distance of the multivariate distributions  $\mathbf{d}_r$  is quite different from convergence of the nested distributions (i.e., with respect to the nested distance  $\mathbf{d}_r$ ).

*Example 2.38 (See Heitsch et al. [55]).* Consider the following nested distributions

$$\mathbb{P}_\epsilon = \left[ \begin{array}{cc} 0.5 & 0.5 \\ 2 & 2 + \epsilon \\ \left[ \frac{1.0}{3} \right] & \left[ \frac{1.0}{1} \right] \end{array} \right] \quad \text{and} \quad \mathbb{P}_0 = \left[ \begin{array}{c} 1.0 \\ 2 \\ \left[ \frac{0.5 \ 0.5}{3 \ 1} \right] \end{array} \right].$$

Notice that the pertaining multivariate distribution of  $\mathbb{P}_\epsilon$  on  $\mathbb{R}^2$  converges weakly to the one of  $\mathbb{P}_0$ , if  $\epsilon \rightarrow 0$ . However, the nested distributions do not converge to  $\mathbb{P}_0$ : The nested distance is  $\mathbf{d}(\mathbb{P}_\epsilon, \mathbb{P}_0) = 1 + \epsilon$  for all  $\epsilon$ . The limit  $\mathbb{P}_\epsilon$  as  $\epsilon \rightarrow 0$  in the sense of nested distances is

$$\tilde{\mathbb{P}} = \left[ \begin{array}{c} \frac{0.5}{\frac{2}{\left[ \frac{1.0}{3} \right]}} \quad \frac{0.5}{\frac{2}{\left[ \frac{1.0}{1} \right]}} \end{array} \right]$$

which is different from  $\mathbb{P}_0$ . To put it differently: the topology of the tree is not changed by going to the limit in the nested distance sense. The filtration of the limiting nested distribution  $\tilde{\mathbb{P}}$  is larger than the one generated by the scenario values. The concept of nested distributions can handle this.

The following proposition shows that the left side of inequality (2.37) can be refined by considering finer filtrations, which are between the original filtration and the full clairvoyant filtration.

**Proposition 2.39.** *Let  $\mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)$  and  $\tilde{\mathbb{P}} \sim (\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi})$  be filtered spaces with filtrations  $\mathfrak{F} = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_T)$  and  $\tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_0, \tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_T)$ , respectively. Denote by  $\mathbb{P}^t$  the pertaining nested distribution made clairvoyant from time  $t$  onwards, that is*

$$\mathbb{P}^t \sim (\Omega, \mathfrak{F}^t, P, \xi^t) \quad \text{with} \quad \mathfrak{F}^t = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_{t-1}, \mathcal{F}_T, \dots, \mathcal{F}_T, \mathcal{F}_T).$$

*In a similar manner we define  $\tilde{\mathbb{P}}^t$ . Then, for  $1 \leq t \leq T$ ,*

$$\mathbf{d}_r(P, \tilde{P}) = \mathbf{d}_r(\mathbb{P}^1, \tilde{\mathbb{P}}^1) \leq \dots \leq \mathbf{d}_r(\mathbb{P}^t, \tilde{\mathbb{P}}^t) \leq \dots \leq \mathbf{d}_r(\mathbb{P}^T, \tilde{\mathbb{P}}^T) = \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}}).$$

*Proof.* The proof follows from the fact that the multivariate distance is always not larger than the nested distance. Arguing this way for the subtrees at stage  $t$  and considering the recursive structure of the nested distance, the assertion is obvious.  $\square$

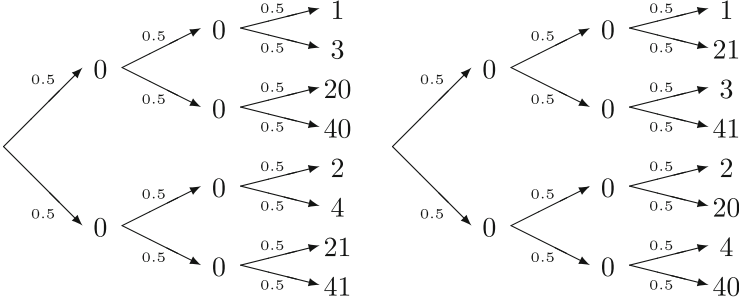
**Example 2.40.** Figure 2.5 shows two trees (nested distributions)  $\mathbb{P}$  and  $\tilde{\mathbb{P}}$ . Their nested distance is  $\mathbf{dl}(\mathbb{P}, \tilde{\mathbb{P}}) = 8.75$ . Figure 2.6 shows the same trees, but both are made clairvoyant from time 2 onwards. Their distance is reduced to  $\mathbf{dl}(\mathbb{P}^2, \tilde{\mathbb{P}}^2) = 0.5$ . Finally, in Fig. 2.7, the same trees are now made totally clairvoyant. Their distance is  $\mathbf{dl}(\mathbb{P}^1, \tilde{\mathbb{P}}^1) = 0$ , since their set of trajectories is identical.

The following lemma recovers the properties of the Wasserstein distance for the nested distance.

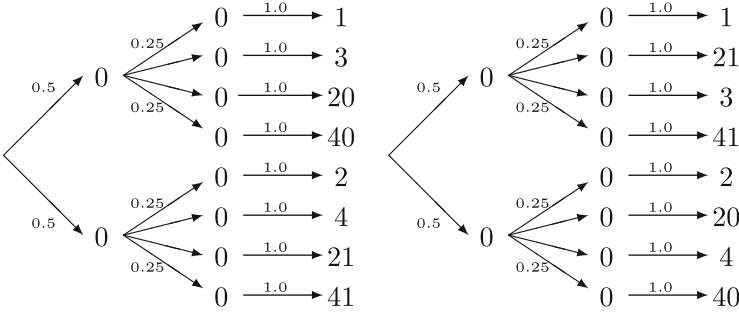
**Lemma 2.41 (Monotonicity and Convexity).**

(i) *Suppose that  $r_1 \leq r_2$ , then*

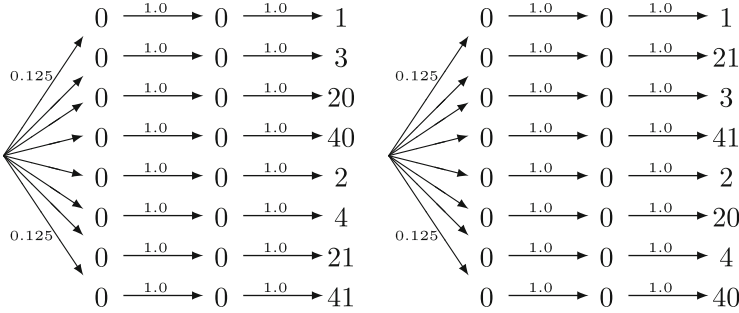
$$\mathbf{dl}_{r_1}(\mathbb{P}, \tilde{\mathbb{P}}) \leq \mathbf{dl}_{r_2}(\mathbb{P}, \tilde{\mathbb{P}}).$$



**Fig. 2.5** The two original trees  $\mathbb{P}$  (left) and  $\tilde{\mathbb{P}}$  (right). Their nested distance is  $\mathbf{dl}(\mathbb{P}, \tilde{\mathbb{P}}) = 8.75$



**Fig. 2.6** The two trees of Fig. 2.5 have been made clairvoyant from time 2 onwards leading to the new trees  $\mathbb{P}^2$  and  $\tilde{\mathbb{P}}^2$ . Their distance is  $\mathbf{dl}(\mathbb{P}^2, \tilde{\mathbb{P}}^2) = 0.5$



**Fig. 2.7** The two trees of Fig. 2.5 have now been made further clairvoyant, namely from time 1 onwards leading to the new trees  $\mathbb{P}^1$  and  $\tilde{\mathbb{P}}^1$ . Their distance is  $\mathbf{dl}(\mathbb{P}^1, \tilde{\mathbb{P}}^1) = 0.0$ , i.e., they are identical

(ii) The nested distance is  $r$ -convex in any of its components, that is for  $0 \leq \lambda \leq 1$  it holds that

$$\mathbf{dl}_r \left( \mathbb{P}, \mathcal{C}(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1, \lambda) \right)^r \leq \lambda \mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}}_0 \right)^r + (1 - \lambda) \mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}}_1 \right)^r,$$

and

$$\begin{aligned}
& \mathbf{dl}_r \left( \mathbb{P}, \mathcal{C}(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1, \lambda) \right) \\
& \leq \lambda^{\frac{1}{r}} \mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}}_0 \right) + (1 - \lambda)^{\frac{1}{r}} \mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}}_1 \right) \\
& \leq \max \{ \lambda, 1 - \lambda \}^{\frac{1}{r}-1} \cdot \left( \lambda \mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}}_0 \right) + (1 - \lambda) \mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}}_1 \right) \right).
\end{aligned}$$

Here  $\mathcal{C}(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1, \lambda)$  is the compound distribution defined in (1.24) of the introduction.

(iii)  $\mathbf{dl}_r$  satisfies the triangle inequality,  $\mathbf{dl}_r \left( \mathbb{P}, \tilde{\mathbb{P}} \right) \leq \mathbf{dl}_r \left( \mathbb{P}, \tilde{\tilde{\mathbb{P}}} \right) + \mathbf{dl}_r \left( \tilde{\tilde{\mathbb{P}}}, \tilde{\mathbb{P}} \right)$ .

*Proof.* The proof of Lemma 2.10 applies. As for the triangle inequality we refer to the proof contained in Villani [137] involving the gluing lemma. A similar proof applies here.  $\square$

### 2.10.3 The Nested Distance for Trees

The Wasserstein distance between discrete probability measures can be calculated by solving the linear program (2.23). To establish the corresponding linear program for the nested distance we use trees that model the whole space and filtration. Recall that we denote by  $m \prec i$  that node  $m$  is a predecessor of node  $i$ , not necessarily the immediate predecessor. Problem (2.35) reads

$$\begin{aligned}
& \text{minimize} && \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\
& \text{(in } \pi) && \\
& \text{subject to} && \sum_{j \succ_n} \pi(i, j | k, l) = P(i | k) \quad (k \prec i, l), \\
& && \sum_{i \succ_m} \pi(i, j | k, l) = \tilde{P}(j | l) \quad (l \prec j, k), \\
& && \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1,
\end{aligned} \tag{2.38}$$

where again  $\pi_{i,j}$  is a matrix defined on the terminal nodes ( $i \in \mathcal{N}_T, j \in \tilde{\mathcal{N}}_T$ ) and  $k \in \mathcal{N}_t, l \in \tilde{\mathcal{N}}_t$  are arbitrary nodes on the same stage  $t$ . The conditional probabilities  $\pi(i, j | k, l)$  are given by

$$\pi(i, j | k, l) = \frac{\pi_{i,j}}{\sum_{i' \succ k, j' \succ l} \pi_{i',j'}}. \tag{2.39}$$

In view of this quotient it becomes evident that the constraint  $\sum_{i,j} \pi_{i,j} = 1$  is necessary in (2.38) to specify a probability measure, as otherwise every multiple of any feasible  $\pi$  would be feasible as well.

**Formulation as a Linear Program.** The constraints in (2.38) can be rewritten as

$$P(i) \cdot \sum_{i' > m, j' > n} \pi_{i',j'} = P(m) \cdot \sum_{j' > n} \pi_{i,j'} \quad (m < i, n) \quad \text{and}$$

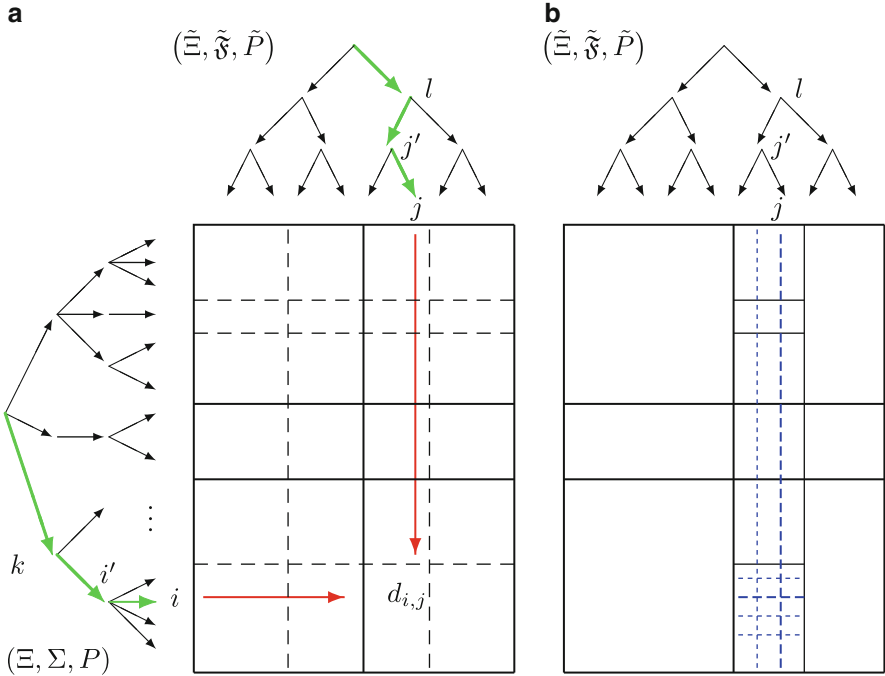
$$\tilde{P}(j) \cdot \sum_{i' > m, j' > n} \pi_{i',j'} = \tilde{P}(n) \cdot \sum_{i' > m} \pi_{i',j} \quad (m, n < j).$$

As  $P$  and  $\tilde{P}$  are given, the latter equations show that (2.38) is equivalent to

$$\begin{aligned} & \text{minimize} \quad \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\ & (\text{in } \pi) \\ & \text{subject to} \quad P(i) \cdot \sum_{i' > k, j' > l} \pi_{i',j'} = P(k) \cdot \sum_{j' > l} \pi_{i,j'} \quad (k < i), \\ & \quad \tilde{P}(j) \cdot \sum_{i' > k, j' > l} \pi_{i',j'} = \tilde{P}(l) \cdot \sum_{i' > k} \pi_{i',j} \quad (l < j), \\ & \quad \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1, \end{aligned}$$

which is indeed a *linear* program. (This is not immediate in the formulation (2.38), as it involves quotients.)

The nested structure of the transportation plan  $\pi$ , which is induced by the trees, is schematically depicted in Fig. 2.8a.



**Fig. 2.8** Schematic structure of the distance matrix  $d$  and the transport matrix  $\pi$ , as it is imposed by the structures of the trees and the respective constraints



*Remark 2.42.* As is the case for the Wasserstein distance, many constraints in (2.38) are linearly dependent. For computational reasons (loss of significance during numerical evaluations, which can impact linear dependencies and the feasibility) it is advisable to remove linear dependencies. This is partially accomplished by the simpler program

$$\begin{aligned}
 & \text{minimize} && \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\
 & \text{(in } \pi) && \\
 & \text{subject to} && \sum_{\{j': j' - = l\}} \pi(i', j' | k, l) = Q(i') = P(i' | k) \quad (i' \in \mathcal{N} \setminus \{1\}, k = i' -, l = j' -), \\
 & && \sum_{\{i': i' - = k\}} \pi(i', j' | k, l) = \tilde{Q}(j') = \tilde{P}(j' | l) \quad (j' \in \mathcal{N} \setminus \{1\}, k = i' -, l = j' -), \\
 & && \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1,
 \end{aligned} \tag{2.40}$$

where the conditional probabilities are

$$\begin{aligned}
 Q(i') &= P(i' | i' -) = P(i' | k) \\
 \tilde{Q}(j') &= \tilde{P}(j' | j' -) = \tilde{P}(j' | l)
 \end{aligned}$$

for  $k = i' -$ ,  $l = j' -$ . Here only one-step conditional transportation measures are required. They are defined as

$$\pi(i', j' | k, l) = \frac{\sum_{i > i', j > j'} \pi_{i,j}}{\sum_{i > k, j > l} \pi_{i,j}}. \tag{2.41}$$

Equation (2.40) is equivalent to (2.38) by the following lemma, and which can be reformulated as an LP as above (recall that  $\mathcal{N} \setminus \{1\}$  denotes all nodes except the root). A computational advantage of (2.40) is given by the fact that the conditional probabilities involved are considered only at successive stages.

Further constraints can be removed from (2.40) by taking into account that  $\sum_{\{i': i' \in k+\}} Q(i') = 1$ . Hence, for each node  $k$  it is possible to drop one constraint out of all equations related to  $\{i' : i' \in k\}$  (cf. Sect. 2.7 and Fig. 2.3 for the Wasserstein distance).

**Lemma 2.43 (Tower Property).** *To compute the nested distance it is enough to condition on the immediately following sigma algebra: the conditions*

$$\pi(A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t) \text{ for all } A \in \mathcal{F}_T$$

in (2.35) may be replaced by

$$\pi(A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t) \text{ for all } A \in \mathcal{F}_{t+1}.$$

*Proof.* The result follows from the tower property of the conditional expectation.

Let  $A \in \mathcal{F}_T$  and observe first that

$$\begin{aligned}\mathbb{E}_\pi (\mathbb{1}_A (\text{id}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) &= \mathbb{E}_\pi (\mathbb{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\ &= \pi (A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) (\text{id}) = \mathbb{E}_P (\mathbb{1}_A | \mathcal{F}_t) (\text{id}),\end{aligned}$$

such that by linearity

$$\mathbb{E}_\pi (\lambda \circ \text{id} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \mathbb{E}_P (\lambda | \mathcal{F}_t) \circ \text{id}$$

for every  $\lambda \triangleleft \mathcal{F}_t$ . It follows then that

$$\begin{aligned}\pi (A \times \tilde{\Omega} | \mathcal{F}_{T-2} \otimes \tilde{\mathcal{F}}_{T-2}) &= \pi (\pi (A \times \tilde{\Omega} | \mathcal{F}_{T-1} \otimes \tilde{\mathcal{F}}_{T-1}) | \mathcal{F}_{T-2} \otimes \tilde{\mathcal{F}}_{T-2}) \\ &= \pi (P(A | \mathcal{F}_{T-1}) | \mathcal{F}_{T-2} \otimes \tilde{\mathcal{F}}_{T-2}) = P(A | \mathcal{F}_{T-2}),\end{aligned}$$

which is the assertion for  $t = T - 2$ . The assertion for general  $t$  follows by repeated application of the previous argument.  $\square$

Notice that the nested distance may be defined not only between trees, but also between a filtered stochastic process and a tree. In the following example, we intend to approximate a simple stochastic process (with only two periods) by a discrete process sitting on a tree. It is crucial that the approximation aims at minimizing the nested distance and not the multivariate distance.

*Example 2.44.* Consider correlated normal variables  $\xi_1 \sim N(0, 1)$  and  $\xi_2 \sim N(\xi_1, 1)$ , that is the joint distribution is

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

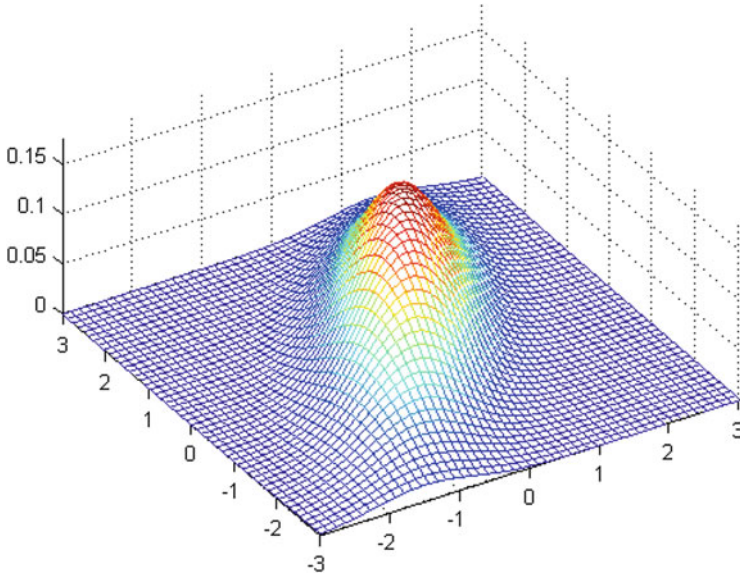
Figure 2.9 displays the density of this distribution.

Let  $\mathbb{P}$  be the nested distribution pertaining to the two-stage process  $(\xi_1, \xi_2)$ . We approximate this distribution by a tree with 9 leaves. First, we consider a tree of height 2 with bushiness 3. To this end the first stage variable  $\xi_1$  is approximated by a discrete distribution sitting on 3 points. It is well known that the optimal approximation of an  $N(\mu, \sigma^2)$  distribution in the  $\mathbf{d}_1$  (Kantorovich) sense by a 3-point distribution is

$$\left[ \begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ \mu - 1.029\sigma & \mu & \mu + 1.029\sigma \end{array} \right]$$

The distance is  $0.3397 \sigma$ . Therefore, the best approximation of  $\xi_1$  is

$$\left[ \begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ -1.029 & 0.0 & 1.029 \end{array} \right]$$



**Fig. 2.9** A bivariate normal distribution

Then, conditional on the first coordinate, the second coordinate is approximated, resulting in the following nested distribution  $\tilde{\mathbb{P}}^*$

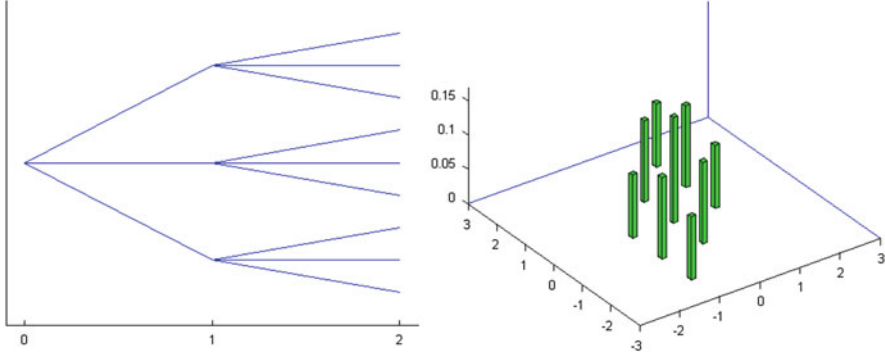
$$\tilde{\mathbb{P}}^* = \left[ \begin{array}{c} \frac{0.3035}{-1.029} \quad \frac{0.3930}{0.0} \quad \frac{0.3035}{1.029} \\ \left[ \begin{array}{ccc} 0.3035 & 0.393 & 0.3035 \\ -2.058 & -1.029 & 0.0 \end{array} \right] \left[ \begin{array}{ccc} 0.3035 & 0.393 & 0.3035 \\ -1.029 & 0.0 & 1.029 \end{array} \right] \left[ \begin{array}{ccc} 0.3035 & 0.393 & 0.3035 \\ 0.0 & 1.029 & 2.058 \end{array} \right] \end{array} \right].$$

The resulting nested distance is  $\mathbf{dl}_1(\mathbb{P}, \tilde{\mathbb{P}}^*) = 0.76$ . The example is illustrated in Fig. 2.10.

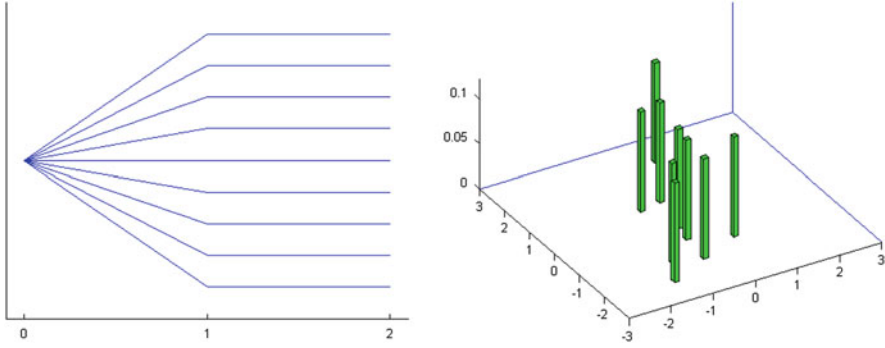
As a comparison we have calculated the best approximation of the two-dimensional distribution  $(\xi_1, \xi_2)$  by a discrete probability sitting on 9 points. Notice that this approximation does not respect the tree structure, it can be seen as a fan with 9 leaves. The calculated approximate distribution is

$$\tilde{\mathbb{P}} = \left[ \begin{array}{c} \frac{0.114}{1.205} \quad \frac{0.108}{0.277} \quad \frac{0.152}{-1.068} \quad \frac{0.148}{0.088} \quad \frac{0.078}{-0.577} \quad \frac{0.046}{-1.855} \quad \frac{0.188}{-0.412} \quad \frac{0.114}{0.894} \quad \frac{0.052}{0.052} \\ \left( \begin{array}{c} 1.205 \\ 1.205 \end{array} \right) \left( \begin{array}{c} 0.277 \\ 1.601 \end{array} \right) \left( \begin{array}{c} -1.068 \\ -0.855 \end{array} \right) \left( \begin{array}{c} 0.088 \\ -0.660 \end{array} \right) \left( \begin{array}{c} -0.577 \\ -2.074 \end{array} \right) \left( \begin{array}{c} -1.855 \\ -2.522 \end{array} \right) \left( \begin{array}{c} -0.412 \\ 0.397 \end{array} \right) \left( \begin{array}{c} 0.894 \\ 0.132 \end{array} \right) \left( \begin{array}{c} 0.052 \\ 1.673 \end{array} \right) \end{array} \right].$$

These points are shown in Fig. 2.11. While the multivariate distance is smaller than before, the nested distance is  $\mathbf{dl}_1(\mathbb{P}, \tilde{\mathbb{P}}) = 1.12$  which is much larger than before because  $\tilde{\mathbb{P}}$  does not respect the filtration structure.



**Fig. 2.10** An “optimal” discrete approximation  $\tilde{\mathbb{P}}$  to the process  $\mathbb{P}$  of Fig. 2.9. *Left*: the prespecified tree structure. *Right*: the visualization of the values and probabilities



**Fig. 2.11** A discrete approximation  $\tilde{\mathbb{P}}$  to the bivariate distribution of Fig. 2.9. *Left*: the structure of the fan, i.e., a trivial tree. *Right*: the visualization of the values and probabilities

**Rapid Computation of the Nested Distance.** In view of the tower property it should be noted that instead of solving the full problem (2.40), the nested distance can be calculated in a recursive way.<sup>15</sup> For this observe that  $\mathbf{dl}_r$  is the nested distance of both trees, starting at their roots. For a rapid computation we extend the distance  $\mathbf{dl}_r$  to subtrees, starting at given nodes of the trees.

For two leaves  $i \in \mathcal{N}_T$ ,  $j \in \tilde{\mathcal{N}}_T$  at the final stage of the tree define first

$$\mathbf{dl}_T^r(i, j) := \mathbf{d}(\xi_i, \tilde{\xi}_j)^r.$$

Given  $\mathbf{dl}_{t+1}^r(i', j')$  for  $i' \in \mathcal{N}_{t+1}$  and  $j' \in \tilde{\mathcal{N}}_{t+1}$ , set

<sup>15</sup>Notice that this recursive calculation corresponds to the way the nested distance was introduced in Sect. 1.4.1.

$$\mathbf{dl}_t^r(k, l) := \sum_{i' \in k_+, j' \in l_+} \pi_t(i', j' | k, l) \cdot \mathbf{dl}_{t+1}^r(i', j') \quad (k \in \mathcal{N}_t, l \in \tilde{\mathcal{N}}_t),$$

where the one-step conditional probabilities  $\pi(\cdot, \cdot | k, l)$  solve the usual Wasserstein problem, conditioned on  $k$  and  $l$ , that is

$$\begin{aligned} & \text{minimize} && \sum_{i' \in k_+, j' \in l_+} \pi_t(i', j' | k, l) \cdot \mathbf{dl}_{t+1}^r(i', j') \\ & \text{in } \pi_t(\cdot, \cdot | k, l) && \\ & \text{subject to} && \sum_{j' \in l_+} \pi_t(i', j' | k, l) = \mathcal{Q}(i') = P(i' | k) \quad (i' \in k_+), \\ & && \sum_{i' \in m_+} \pi_t(i', j' | k, l) = \tilde{\mathcal{Q}}(j') = \tilde{P}(j' | l) \quad (j' \in l_+), \\ & && \pi_t(i', j' | k, l) \geq 0. \end{aligned}$$

The values  $\mathbf{dl}_t^r(k, l)$  can be interpreted as the nested distances of the subtrees starting in nodes  $k$  and  $l$ . Finally the transport plan  $\pi$  on the leaves is recomposed by

$$\pi(i, j) = \pi(i_1, j_1 | i_0, j_0) \cdots \pi(i_{T-1}, j_{T-1} | i_{T-2}, j_{T-2}) \cdot \pi(i, j | i_{T-1}, j_{T-1}) \quad (2.42)$$

with  $i_t = \text{pred}_t(i)$ ,  $j_t = \text{pred}_t(j)$ . The nested distance is given by  $\mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \mathbf{dl}_0^r(1, 1)$ , where  $(i_0, j_0) = (1, 1)$  is the pair of root nodes of both trees.

Algorithm 2.1 summarizes this procedure in order to efficiently compute the nested distance for tree processes in a nested, recursive manner.

*Example 2.45.* As an example for Algorithm 2.1 to efficiently compute the nested distance we consider the nested distributions  $\mathbb{P}$  (a tree with 11 leaves) and  $\tilde{\mathbb{P}}$  (a tree with 6 leaves) shown below and depicted in Fig. 2.12.

$$\begin{aligned} \mathbb{P} &= \left[ \begin{array}{c} \frac{0.2}{13} \quad \frac{0.3}{11} \quad \frac{0.3}{8} \quad \frac{0.2}{6} \\ \left[ \begin{array}{c} \frac{0.1 \ 0.2 \ 0.4 \ 0.3}{15 \ 14 \ 13 \ 11} \end{array} \right] \left[ \begin{array}{c} \frac{0.7 \ 0.3}{12 \ 9} \end{array} \right] \left[ \begin{array}{c} \frac{0.5 \ 0.2 \ 0.3}{10 \ 8 \ 6} \end{array} \right] \left[ \begin{array}{c} \frac{0.4 \ 0.6}{7 \ 5} \end{array} \right] \end{array} \right], \\ \tilde{\mathbb{P}} &= \left[ \begin{array}{c} \frac{0.3}{13} \quad \frac{0.3}{11} \quad \frac{0.4}{7} \\ \left[ \begin{array}{c} \frac{0.4 \ 0.6}{14 \ 12} \end{array} \right] \left[ \begin{array}{c} \frac{1.0}{10} \end{array} \right] \left[ \begin{array}{c} \frac{0.4 \ 0.4 \ 0.2}{9 \ 8 \ 5} \end{array} \right] \end{array} \right]. \end{aligned}$$

The pertaining multivariate distributions and their Wasserstein distances were already considered in Example 2.28.

**Initialization.** Table 2.2 collects the distances of two paths of the trees (the state space). Here, the  $\ell^1$ -distance is employed.

---

**Algorithm 2.1** Nested computation of the nested distance  $\mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})$  of two tree-processes  $\mathbb{P}$  and  $\tilde{\mathbb{P}}$

---

**Initialization** ( $t = T$ ):

For all combinations of *leaf* nodes  $i \in \mathcal{N}_T$  and  $j \in \tilde{\mathcal{N}}_T$  with predecessors  $(i_0, i_1, \dots, i_{T-1}, i)$  and  $(j_0, j_1, \dots, j_{T-1}, j)$  define

$$\mathbf{dl}_T^r(i, j) := \mathbf{d}\left((\xi_{i_0}, \xi_{i_1}, \dots, \xi_i), (\tilde{\xi}_{j_0}, \tilde{\xi}_{j_1}, \dots, \tilde{\xi}_j)\right)^r.$$

**Iteration, backwards:**

For  $t = T - 1$  down to 0, and

for every combination of *inner* nodes  $k \in \mathcal{N}_t$  and  $l \in \mathcal{N}_t$  solve the LP (cf. (2.23))

$$\begin{aligned} \mathbf{dl}_t^r(k, l) := & \underset{(\text{in } \pi)}{\text{minimize}} \quad \sum_{i' \in k+, j' \in l+} \pi(i', j' | k, l) \cdot \mathbf{dl}_{t+1}^r(i', j') \\ & \text{subject to} \quad \sum_{j' \in l+} \pi(i', j' | k, l) = \underline{Q}(i') \quad i' \in k+ \\ & \quad \quad \quad \sum_{i' \in k+} \pi(i', j' | k, l) = \tilde{Q}(j') \quad j' \in l+ \\ & \quad \quad \quad \pi(i', j' | k, l) \geq 0 \end{aligned} \quad (2.43)$$

**Final Assignment:**

The nested distance of the trees is the distance of the trees at their roots 1,

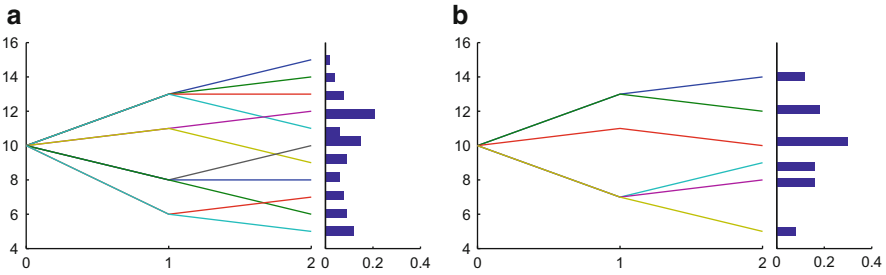
$$\mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \mathbf{dl}_0^r(1, 1).$$

The optimal transport plan at the leaf nodes  $i \in \mathcal{N}_T$  and  $j \in \tilde{\mathcal{N}}_T$  is

$$\pi(i, j) := \pi_1(i_1, j_1 | i_0, j_0) \cdots \pi_{T-1}(i, j | i_{T-1}, j_{T-1}).$$

$\pi_t$  is the optimal transport plan obtained in (2.43) at stage  $t$ .

---



**Fig. 2.12** Depicted are two trees in three stages, annotated is the histogram of their final probability distribution. (a) Initial tree  $\mathbb{P}$ . (b) Target tree  $\tilde{\mathbb{P}}$

**Iteration, backwards.** Based on the distance matrix and the structure of the trees the respective subproblems are computed for each combination of nodes at the same stage. For stage 1, the result of the above two subtrees is displayed in Table 2.4a; the result (0.8) is the corresponding, new entry in the distance Table 2.4b at the earlier stage 0.

**Table 2.4** The distance at different levels: distance  $\mathbf{dl}$ , probabilities  $\pi$ , and dual variables ( $\lambda$  and  $\mu$ ; cf. Table 2.3)

Distance of states, $\mathbf{dl}_2(i, j)$	0	0	Probabilities, $\pi_2(i, j)$	0.4	0.6
1	<b>1</b>	3	0.1	<b>0.1</b>	0
0	<b>0</b>	2	0.2	<b>0.2</b>	0
1	<b>1</b>	<b>1</b>	0.4	<b>0.1</b>	<b>0.3</b>
1	3	<b>1</b>	0.3	0	<b>0.3</b>

(a) Stage 1, the first two subtrees, primal and dual solutions. The Wasserstein distance of the subtrees is  $\sum_{i,j} \pi_2(i, j) \mathbf{dl}_2(i, j) = 0.8$ —the first entry in Table 2.4b

Distance of subtrees, $\mathbf{dl}_1(i, j)$	7.6	5.6	3	Probabilities, $\pi_1(i, j)$	0.3	0.3	0.4
−6.8	<b>0.8</b>	4.8	11	0.2	<b>0.2</b>	0	0
−3.9	<b>3.7</b>	<b>1.7</b>	7.3	0.3	<b>0.1</b>	<b>0.2</b>	0
−1	9.4	<b>4.6</b>	<b>2</b>	0.3	0	<b>0.1</b>	<b>0.2</b>
0	14	9.2	<b>3</b>	0.2	0	0	<b>0.2</b>

(b) Stage 0, the combination of all subtrees. The nested distance is  $\sum_{i,j} \pi_1(i, j) \mathbf{dl}_1(i, j) = 2.33$

**Table 2.5** The conditional probabilities of the tree, and the probabilities of the nested distance,  $\sum_{i,j} \pi_{i,j} d_{i,j} = 2.33$ . The unconditional probabilities for the nested distance notably do not coincide with Wasserstein distance (Table 2.3)

Conditional probabilities			1					
			0.3		0.3	0.4		
			0.4	0.6	1	0.4	0.4	0.2
1	0.2	0.1	<b>0.02</b>	0	0	0	0	0
		0.2	<b>0.04</b>	0	0	0	0	0
		0.4	<b>0.02</b>	<b>0.06</b>	0	0	0	0
		0.3	0	<b>0.06</b>	0	0	0	0
	0.3	0.7	<b>0.04</b>	<b>0.03</b>	<b>0.14</b>	0	0	0
		0.3	0	<b>0.03</b>	<b>0.06</b>	0	0	0
	0.3	0.5	0	0	<b>0.05</b>	<b>0.08</b>	<b>0.02</b>	0
		0.2	0	0	<b>0.02</b>	0	<b>0.04</b>	0
		0.3	0	0	<b>0.03</b>	0	<b>0.02</b>	<b>0.04</b>
	0.2	0.4	0	0	0	<b>0.04</b>	<b>0.04</b>	0
		0.6	0	0	0	<b>0.04</b>	<b>0.04</b>	<b>0.04</b>

**Final Assignment.** The nested distance finally is the distance of the subtrees at level 0,  $\mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}}) = 2.33$ . Moreover, the transport plan  $\pi$  can be computed. The resulting transport plan is displayed in Table 2.5.

It should be noted that the final probabilities of any of the sub-problems can be recovered in the probability at an earlier stage. For example, the probabilities  $\pi$  from Table 2.4a are in the upper-left section of the matrix in Table 2.5, but multiplied with its conditional probability 10 %, which is the result from Table 2.4a.

*Example 2.46.* The nested distance of the trees in Fig. 1.13 (Chap. 1) is 0, a minimal transport plan is

$$\pi = \begin{pmatrix} 0.42 & 0 \\ 0 & 0.18 \\ 0.28 & 0 \\ 0 & 0.12 \end{pmatrix}.$$

This example demonstrates that the nested distance correctly identifies equivalent processes.

*Example 2.47.* The nested distance of the trees addressed in Example 2.30 (Fig. 2.4) is  $4p(1-p) > 0$ , the corresponding nested transport plan is given by

$$\pi = \begin{pmatrix} p^2 & p(1-p) \\ (1-p)p & (1-p)^2 \end{pmatrix}.$$

The nested distance thus correctly identifies two different trees in this situation (except for the degenerate cases  $p = 0$  or  $p = 1$ , where they coincide again). This is in notable contrast to the final Wasserstein distance of these trees, which was found to be 0.

## 2.11 Dual Representation of the Nested Distance

The duality for the Wasserstein distance was established in Sect. 2.8 by considering LP duality for the defining optimization problem. This section provides a martingale representation of the nested distance. The martingale corresponds to successively solving Wasserstein problems on subsequent stages, the main result is Theorem 2.49 below. This (dual) representation is adapted from [94].

To prepare for the dual representation of the nested distance it is helpful to observe that one may state the dual problem (2.25) in the equivalent form

$$\begin{aligned} & \text{maximize} \\ & \quad (\text{in } M_0, \lambda, \mu) M_0 \\ & \text{subject to } \mathbb{E}\lambda = 0, \tilde{\mathbb{E}}\mu = 0, \\ & \quad M_0 + \lambda(\xi) + \mu(\tilde{\xi}) \leq d(\xi, \tilde{\xi})^r \text{ for all } \xi \text{ and } \tilde{\xi}. \end{aligned}$$

The defining equations for the nested distance in Definition 2.32 involve constraints for each of the stages. As the variables in the dual program correspond to



constraints in the primal, the dual program for the nested distance involves variables for each stage.

To establish the dual representation of the nested distance it is necessary to incorporate the constraints (2.35) in the Lagrangian function. To this end we define projections, which act on one component while leaving the other unaffected, by

$$\begin{aligned} \text{proj}_t : L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) &\rightarrow L^1(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_T) \\ \lambda(\text{id}) \cdot \mu(\tilde{\text{id}}) &\mapsto \mathbb{E}(\lambda|\mathcal{F}_t)(\text{id}) \cdot \mu(\tilde{\text{id}}) \end{aligned}$$

and

$$\begin{aligned} \tilde{\text{proj}}_t : L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) &\rightarrow L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_t) \\ \lambda(\text{id}) \cdot \mu(\tilde{\text{id}}) &\mapsto \lambda(\text{id}) \cdot \mathbb{E}(\mu|\tilde{\mathcal{F}}_t)(\tilde{\text{id}}). \end{aligned}$$

The one-sided projections  $\text{proj}$  and  $\tilde{\text{proj}}$  are well defined by linearity, because functions of the form  $(x, y) \mapsto \mathbb{1}_A(x) \mathbb{1}_B(y)$  form a basis for  $L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T)$ .

**Proposition 2.48 (Characterization of the Projection).** *The measure  $\pi$  satisfies the marginal condition*

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) \quad \text{for all } A \in \Omega \quad (2.44)$$

if and only if

$$\mathbb{E}_\pi \lambda = \mathbb{E}_\pi \text{proj}_t \lambda \quad \text{for all } \lambda \triangleleft \mathcal{F}_T \otimes \tilde{\mathcal{F}}_t. \quad (2.45)$$

Moreover,  $\text{proj}_t(\lambda) = \mathbb{E}_\pi(\lambda | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$  if  $\pi$  has marginal  $P$ .

*Proof.* Note first that the left-hand side and the right-hand side of (2.44) are probability measures, it is thus sufficient to show that

$$\int_{C \times D} \pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) d\pi = \int_{C \times D} P(A | \mathcal{F}_t)(\text{id}) d\pi$$

for all sets  $C \in \mathcal{F}_t$  and  $D \in \tilde{\mathcal{F}}_t$ .

To this end observe that

$$\begin{aligned} \int_{C \times D} \pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) d\pi &= \mathbb{E}_\pi(\mathbb{1}_{C \times D} \mathbb{E}_\pi(\mathbb{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)) \\ &= \mathbb{E}_\pi \mathbb{E}_\pi(\mathbb{1}_{C \times D} \mathbb{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\ &= \mathbb{E}_\pi \mathbb{1}_{(C \times D) \cap (A \times \tilde{\Omega})} = \pi((A \cap C) \times D). \end{aligned}$$

Next,

$$\begin{aligned}
 \int_{C \times D} P(A|\mathcal{F}_t) d\pi &= \mathbb{E}_\pi \mathbb{1}_{C \times D} \mathbb{E}_P(\mathbb{1}_A|\mathcal{F}_t) \circ \text{id} \\
 &= \mathbb{E}_\pi \mathbb{1}_C(\text{id}) \cdot \mathbb{1}_D(\tilde{\text{id}}) \cdot \mathbb{E}_P(\mathbb{1}_A|\mathcal{F}_t)(\text{id}) \\
 &= \mathbb{E}_\pi \mathbb{E}_P(\mathbb{1}_C \mathbb{1}_A|\mathcal{F}_t)(\text{id}) \cdot \mathbb{1}_D(\tilde{\text{id}}) \\
 &= \mathbb{E}_\pi \mathbb{E}_P(\mathbb{1}_{C \cap A}|\mathcal{F}_t)(\text{id}) \cdot \mathbb{1}_D(\tilde{\text{id}}) \\
 &= \mathbb{E}_\pi \text{proj}_t(\mathbb{1}_{C \cap A}(\text{id}) \cdot \mathbb{1}_D(\tilde{\text{id}})),
 \end{aligned}$$

and by the assertion (2.45) thus

$$\begin{aligned}
 \int_{C \times D} P(A|\mathcal{F}_t) d\pi &= \mathbb{E}_\pi \mathbb{1}_{A \cap C}(\text{id}) \cdot \mathbb{1}_D(\tilde{\text{id}}) \\
 &= \mathbb{E}_\pi \mathbb{1}_{(A \cap C) \times D} = \pi((A \cap C) \times D),
 \end{aligned}$$

from which the first assertion (2.44) follows.

As for the converse it is enough to prove (2.45) for functions of the form  $\mu \circ \text{id} \cdot \tilde{\mu} \circ \tilde{\text{id}}$  for  $\mu \in \mathcal{F}_T$  and  $\tilde{\mu} \in \tilde{\mathcal{F}}_t$ , as these products form a basis.

For the function  $\mathbb{1}_A$  ( $A \in \mathcal{F}_T$ )

$$\begin{aligned}
 \mathbb{E}_P(\mathbb{1}_A|\mathcal{F}_t)(\text{id}) &= P(A|\mathcal{F}_t) = \pi(A \times \tilde{\Omega}|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\
 &= \mathbb{E}(\mathbb{1}_{A \times \tilde{\Omega}}|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \mathbb{E}_\pi(\mathbb{1}_A(\text{id})|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t),
 \end{aligned}$$

and by linearity thus  $\mathbb{E}_P(\mu|\mathcal{F}_t)(\text{id}) = \mathbb{E}_\pi(\mu(\text{id})|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$ . With this identity it follows further that

$$\mathbb{E}_P(\mu|\mathcal{F}_t)(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}}) = \mathbb{E}_\pi(\mu(\text{id})|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \cdot \tilde{\mu}(\tilde{\text{id}}) = \mathbb{E}_\pi(\mu(\text{id}) \tilde{\mu}(\tilde{\text{id}})|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t).$$

Taking expectations gives that

$$\begin{aligned}
 \mathbb{E}_\pi \text{proj}_t(\mu(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}})) &= \mathbb{E}_\pi \mathbb{E}_P(\mu|\mathcal{F}_t)(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}}) \\
 &= \mathbb{E}_\pi \mathbb{E}_\pi(\mu(\text{id}) \tilde{\mu}(\tilde{\text{id}})|\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\
 &= \mathbb{E}_\pi \mu(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}}),
 \end{aligned}$$

which is the assertion for the basis functions  $\mu(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}})$ .  $\square$

### 2.11.1 Martingale Representation of the Nested Distance

Proposition 2.48 is the essential tool to describe the dual representation of the nested distance.

**Theorem 2.49 (Duality for the Nested Distance).** *The infimum of problem (2.35) to compute the nested distance  $\mathbf{dl}_r^r(\mathbb{P}, \tilde{\mathbb{P}})$  equals the supremum of all numbers  $M_0$  such that*

$$M_T(\omega, \tilde{\omega}) \leq \mathbf{d}(\omega, \tilde{\omega})^r \quad (\omega, \tilde{\omega}) \in \Omega \times \tilde{\Omega},$$

where  $M_t$  is an  $\mathbb{R}$ -valued process on  $\Omega \times \tilde{\Omega}$  of the form

$$M_t = M_0 + \sum_{s=1}^t (\lambda_s + \mu_s)$$

and the measurable functions  $\lambda_t \triangleleft \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$  and  $\mu_t \triangleleft \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_{t-1}$  satisfy  $\text{proj}_{t-1}(\lambda_t) = 0$  and  $\tilde{\text{proj}}_{t-1}(\mu_t) = 0$ .

The process  $M_t$ , for which the supremum is attained, is a martingale with respect to the optimal measure  $\pi$ ,

$$M_t = \mathbb{E}(\mathbf{d}^r | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \quad \pi\text{-a.e. in } \Omega \times \tilde{\Omega}$$

and

$$\mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \mathbb{E}_\pi(\mathbf{d}^r) = M_0 = \mathbb{E}_\pi M_T.$$

Before we conclude this section with the proof of the theorem we give an example, which displays the martingale process.

*Example 2.50 (Continuation of Example 2.45).* Table 2.6 collects the final stage of the process  $M_T$ . Comparing  $M_T$  with the initial distance (Table 2.2) it becomes apparent that  $M_T \leq \mathbf{d}$ . Moreover it holds that  $M_T = \mathbf{d}$  for all entries for which  $\pi_{i,j} > 0$  (these entries are bold in the table. Cf. Table 2.5 to compare this representation by duality with the primal nested distance).

*Proof of Theorem 2.49.* Evoking Proposition 2.48 the constraints in (2.35) can be encoded in the Lagrangian as

$$\inf_{\pi \geq 0} \sup_{M_0, f_t, g_t} \mathbb{E}_\pi \mathbf{d}^r + M_0 \cdot (1 - \mathbb{E}_\pi \mathbb{1}) + \\ - \sum_{s=0}^{T-1} (\mathbb{E}_\pi f_{s+1} - \mathbb{E}_\pi \text{proj}_s(f_{s+1})) +$$

**Table 2.6** The final state of the dual martingale process  $M_T$  corresponding to Example 2.45

Conditional probabilities			1					
			0.3		0.3	0.4		
			0.4	0.6	1	0.4	0.4	0.2
1	0.2	0.1	<b>I</b>	1	1	-2.8	-1.8	1.2
		0.2	<b>0</b>	0	0	-3.8	-2.8	0.2
		0.4	<b>I</b>	<b>I</b>	-1	-4.8	-3.8	-0.8
		0.3	1	<b>I</b>	-3	-6.8	-5.8	-2.8
	0.3	0.7	<b>4</b>	<b>2</b>	<b>2</b>	-1.2	-0.2	2.8
		0.3	7	<b>5</b>	<b>I</b>	-4.2	-3.2	-0.2
	0.3	0.5	6.2	4.2	<b>3</b>	<b>2</b>	<b>3</b>	2
		0.2	8.2	6.2	<b>5</b>	0	<b>I</b>	0
		0.3	10.2	8.2	<b>7</b>	2	<b>3</b>	<b>2</b>
	0.2	0.4	7.6	5.6	4.4	<b>3</b>	<b>2</b>	-1
		0.6	9.6	7.6	6.4	<b>5</b>	<b>4</b>	<b>I</b>

$$- \sum_{s=0}^{T-1} (\mathbb{E}_\pi g_{s+1} - \mathbb{E}_\pi \tilde{\text{proj}}_s(g_{s+1})),$$

the infimum being among positive measures  $\pi \geq 0$ , not only probability measures; the functions in the inner supremum satisfy  $f_t \triangleleft \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$  and  $g_t \triangleleft \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t$ . According to Sion's minimax theorem (cf. Sion [132]) this saddle point has the same objective value as

$$\sup_{M_0, f_t, g_t} M_0 + \inf_{\pi \geq 0} \mathbb{E}_\pi \left[ \begin{array}{l} \mathbf{d}^r - M_0 \cdot \mathbb{1} \\ - \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) \\ - \sum_{s=0}^{T-1} (g_{s+1} - \tilde{\text{proj}}_s(g_{s+1})) \end{array} \right]. \quad (2.46)$$

Now notice that the infimum over all  $\pi \geq 0$  is  $-\infty$  unless the integrand is positive for every measure  $\pi \geq 0$ , which means that

$$M_0 + \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) + \sum_{s=0}^{T-1} (g_{s+1} - \tilde{\text{proj}}_s(g_{s+1})) \leq \mathbf{d}^r$$

necessarily has to hold. For a positive integrand in (2.46), the infimum over positive measures  $\inf_{\pi \geq 0} \mathbb{E}_\pi$  is 0. Equation (2.46) thus can be reformulated as

$$\begin{array}{ll} \text{maximize} & M_0 \\ \text{in } M_0, f_t, g_t & \\ \text{subject to} & M_0 + \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s f_{s+1}) + \sum_{s=0}^{T-1} (g_{s+1} - \tilde{\text{proj}}_s g_{s+1}) \leq \mathbf{d}^r, \\ & f_t \triangleleft \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}, \quad g_t \triangleleft \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t. \end{array}$$

Using the setting  $\lambda_s := f_s - \text{proj}_{s-1}(f_s)$  and  $\mu_s := g_s - \text{proj}_{s-1}(g_s)$  allows rewriting the latter equation as

$$\begin{aligned} & \text{maximize (in } M_0, \lambda_t, \mu_t) \quad M_0 \\ & \text{subject to} \quad M_0 + \sum_{s=1}^T (\lambda_s + \mu_s) \leq \mathbf{d}^r \\ & \quad \text{proj}_{t-1}(\lambda_t) = 0, \text{proj}_{t-1}(\mu_t) = 0, \end{aligned}$$

which is the desired formulation.

To accept the martingale property observe that

$$\mathbb{E}_\pi (M_T | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \mathbb{E}_\pi \left( M_0 + \sum_{s=0}^{T-1} \lambda_s + \mu_s | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right) = M_0 + \sum_{s=0}^t \lambda_s + \mu_s,$$

provided that  $\mathbb{E}_\pi (\lambda_s + \mu_s | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = 0$  whenever  $s > t$ . This can be seen as follows: recall that  $\lambda(\text{id}) \cdot \mathbb{1}_B(\tilde{\text{id}})$  are base functions. Now if

$$\text{proj}_t \lambda(\text{id}) \cdot \mathbb{1}_B(\tilde{\text{id}}) = \mathbb{E}(\lambda | \mathcal{F}_t)(\text{id}) \cdot \mathbb{1}_B(\tilde{\text{id}}) = 0,$$

then  $\mathbb{E}(\lambda | \mathcal{F}_t) = 0$  and consequently  $\mathbb{E}(\lambda(\text{id}) \tilde{\lambda}(\tilde{\text{id}}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = 0$ .

It holds moreover that

$$M_t = \mathbb{E}_\pi (M_T | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \leq \mathbb{E}_\pi (\mathbf{d}^r | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$$

due to the constraints. But the vanishing duality gap forces

$$M_0 = \mathbb{E}_\pi \mathbf{d}^r,$$

such that we conclude in addition that

$$M_t = \mathbb{E}_\pi (\mathbf{d}^r | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \quad \pi - \text{a.e.},$$

which completes the proof. □

Multistage Stochastic Optimization

Pflug, G.C.; Pichler, A.

2014, XIV, 301 p. 81 illus., Hardcover

ISBN: 978-3-319-08842-6