Appendix B

Optimal transportation distances

Science is what we understand well enough to explain to a computer. Art is verythings else we do.

Donald Knuth

In Section B.1 the general, probabilistic setting is introduced with which we work in the following. Section B.2 introduces the optimal transportation problem which is used to define a distance in Section B.3.

B.1 The setting

Recall the setting introduced in Section 1.1: A complex system S is measured by a measuring device D. The system S is an element of an abstract space of systems $\mathcal{S}$, and a measuring device is a function that maps $S \in \mathcal{S}$ into a space of measurements M. Since we are interested in quantitative measurements, the space M will be a metric space (M, d), equipped with a distance d. For example, we could take (M, d) to be some Euclidean space E_n or, more generally, a manifold with distance induced by geodesics (shortest paths). However, to account for random influences in the measurement process, we will more generally consider spaces of probability measures on M.

Let (M, d) be a metric space. For simplicity of exposition, let us also assume that M is complete, path-connected and has continuous distance function, such that it is Hausdorff in the induced topology. A *curve* on M is a continuous function $\gamma : [0, 1] \to M$. It is a curve from x to y if $\gamma(0) = x$ and $\gamma(1) = y$. The *arc length* of γ is defined by

$$L_\gamma = \sup_{0=t_0<t_1<\cdots<t_n=1} \sum_{i=0}^{n-1} d(\gamma(t_i), \gamma(t_{i+1})), \tag{B.1}$$

where the supremum is taken over all possible partitions of $[0, 1]$, for all $n \in \mathbb{N}$. Note that L_γ can be infinite; the curve γ is then called non-rectifiable.

Let us define a new metric d_I on M, by letting the value of $d_I(x, y)$ be the infimum of the lengths of all paths from x to y. This is called the *induced intrinsic metric* of M. If $d_I(x, y) = d(x, y)$ for all points $x, y \in M$, then (M, d) is a *length space* and d is called *intrinsic*. Euclidean space E_n and Riemannian manifolds are examples of

length spaces. Since M is path-connected, it is a *convex metric space*, i.e., for any two points $x, y \in M$ there exists a point $z \in M$ between x and y in the intrinsic metric.

Let μ be a probability measure on M with σ-algebra $\mathcal{B}$. We will assume μ to be a Radon measure, i.e., a tight locally-finite measure on the Borel σ-algebra of M, and denote the space of all such measures by $\mathcal{P}(M)$. Most of the time, however, we will be working in the much simpler setting of a discrete probability space: Let μ be a singular measure on M that is finitely presentable, i.e., such that there exists a representation

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \tag{B.2}$$

where δ_{x_i} is the Dirac measure at point $x_i \in M$, and the norming constraint $\sum_{i=1}^{n} a_i = 1$ is fulfilled. We further assume that $x_i \neq x_j$ if $i \neq j$, which makes the representation (B.2) unique (up to permutation of indices). Denote the space of all such measures by $\mathcal{P}_{\mathrm{F}}(M)$. Measures in $\mathcal{P}_{\mathrm{F}}$ correspond to the notion of a *weighted point set* from the literature on classification. In our setting they represent a finite amount of information obtained from a complex system.

In particular, let a probability measure $\mu_0 \in \mathcal{P}(M)$ represent the possible measurements on a system S. Each *elementary* measurement corresponds to a point of M, and if the state of the system S is repeatedly measured, we obtain a finite sequence $X_1, X_2, \ldots, X_n$ of iid random variables (with respect to the measure μ_0) taking values in M. These give rise to an *empirical measure*

$$\mu_n[A] = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}[A], \quad A \in \mathcal{B}. \tag{B.3}$$

The measure μ_n is itself a random variable, but fixing the outcomes, i.e., considering a realization $(x_1, x_2, \ldots, x_n) \in M^n$, a measure $\mu \in \mathcal{P}_{\mathrm{F}}(M)$ is obtained,

$$\mu = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}, \tag{B.4}$$

which we call a *realization* of the measure μ_0. Denote the space of all probability measures (B.4) for fixed $n \in \mathbb{N}$ and $\mu_0 \in \mathcal{P}(M)$ by $\mathcal{P}_n(\mu_0)$.

B.2 Discrete optimal transportation

In this secion we will motivate the notion of distance with which we will be concerned in the rest of the thesis. The starting point is the question of how to define a useful distance for the measures in $\mathcal{P}_F$.

Example 10 (Total variation)**.** The *distance in variation* between two measures μ and

ν is

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{A\in\mathcal{B}} |\mu[A] - \nu[A]|. \tag{B.5}$$

It is obviously reflexive and symmetric. For the triangle inequality, let $\epsilon > 0$ and consider $A \in \mathcal{B}$ such that $d_{\mathrm{TV}}(\mu,\nu) < |\mu[A] - \nu[A]| + \epsilon$. Then

$$\begin{aligned} d_{\mathrm{TV}}(\mu,\nu) &< |\mu[A] - \rho[A]| + |\rho[A] - \nu[A]| + \epsilon \\ &< \sup_{A\in\mathrm{M}} |\mu[A] - \rho[A]| + \sup_{A\in\mathrm{M}} |\rho[A] - \nu[A]| + 2\epsilon. \end{aligned} \tag{B.6}$$

Since this holds for all ϵ, the triangle inequality is established. Total variation distance metrizes the strong topology on the space of measures, and can be interpreted easily: If two measures μ and ν have total variation $p = d_{\mathrm{TV}}(\mu,\nu)$, then for any set $A \in \mathcal{F}$ the probability assigned to it by μ and ν differs by at most p. For two measures $\mu, \nu \in \mathcal{P}_F$ concentrated on a countable set $x_1, x_2, \ldots$, it simplifies to

$$d_{\mathrm{TV}}(\mu,\nu) = \sum_i |\mu[x_i] - \nu[x_i]|. \tag{B.7}$$

Unfortunately, total variation needs further effort to be usable in practice. Consider an absolutely continuous $\mu_0 \in \mathcal{P}(M)$ with density $f : M \to [0,1]$. For two realizations $\mu, \nu \in \mathcal{P}_n(\mu_0)$ we have that $\mathrm{pr}(\operatorname{supp}\mu \cap \operatorname{supp}\nu \neq \emptyset) = 0$, so $d_{\mathrm{TV}}(\mu,\nu) = 0$ almost surely. In practice, therefore, we will need to use some kind of density estimation to achieve a non-trivial value $d_{\mathrm{TV}}(\mu,\nu)$; confer (Schmid and Schmidt, 2006).

Example 11. The Hausdorff metric is a distance of subsets of a metric space (Example 5). It can be turned into a distance for probability measures by "forgetting" the probabilistic weights, i.e.,

$$d_{\mathrm{HD}}(\mu,\nu) \stackrel{\mathrm{def}}{=} d_{\mathrm{H}}(\operatorname{supp} f, \operatorname{supp} g), \tag{B.8}$$

If M is a normed vector space, then a subset $A \subset M$ and its translation $x + A = \{x + a \mid a \in A\}$ have Hausdorff distance $d_H(A, x + A) = ||x||$, which seems natural. However, Hausdorff distance is unstable against outliers. For example, consider the family of measures defined by $P_0 = \delta_0$ and $P_n = \frac{1}{n}\delta_n + (1 - \frac{1}{n})\delta_0$ for all $n > 0$. Then $d_{\mathrm{HD}}(P_0, P_n) = n$. □

Example 12 (Symmetric pullback distance)**.** Let $f : M^n \to N$ be the projection of an ordered n-tuple from M into a single point of a metric space (N, d'). Call f *symmetric* if its value does not depend on the order of its arguments, i.e., if $f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for all permutations σ from the symmetric group $\Sigma(n)$ on n elements. Then

$$d_f(X,Y) \stackrel{\mathrm{def}}{=} d'(f(X), f(Y)) \tag{B.9}$$

defines a distance between n-element subsets $X, Y \subset M$ (the symmetric pullback of the distance in N).

In particular, if M has the structure of a vector space, then each function $f : M^n \to N$ can be symmetrized, yielding a symmetric function

$$f_\sigma(x_1, \dots, x_n) \stackrel{\text{def}}{=} \frac{1}{n!} \sum_{\sigma \in \Sigma(n)} f(x_{\sigma(1)}, \dots, x_{\sigma(n)}). \tag{B.10}$$

For the projection to the first factor,

$$f : M^n \to M, \quad (x_1, \dots, x_n) \mapsto x_1, \tag{B.11}$$

this yields the *centroid*

$$f_\sigma(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{B.12}$$

with centroid distance $d_f(X, Y) = d(\bar{X}, \bar{Y})$. This construction generalizes in the obvious way to finite probability measures $\mu, \nu \in \mathcal{P}_n(\mu_0)$.

Note however, that the symmetric pullback distance is pseudo-metric: There usually exist many n-subsets X, Y of M with the same pullback distance, i.e., $d_f(X, Y) = 0$ does not imply that $X = Y$.

All the above distances have various shortcomings that are not exhibited by the following distance. Let μ, ν be two probability measures on M and consider a cost function $c : M \times M \to \mathbb{R}_+$. The value $c(x, y)$ represents the cost to transport one unit of (probability) mass from location $x \in M$ to some location $y \in M$. We will model the process of transforming measure μ into ν, relocating probability mass, by a probability measure π on $M \times M$. Informally, $\mathrm{d}\pi(x, y)$ measures the amount of mass transferred from location x to y. To be admissible, the transference plan π has to fulfill the conditions

$$\pi[A \times M] = \mu[A], \quad \pi[M \times B] = \nu[B] \tag{B.13}$$

for all measurable subsets $A, B \subseteq M$. We say that π has marginals μ and ν if (B.13) holds, and denote by $\Pi(\mu, \nu)$ the set of all admissible transference plans.

Kantorovich's *optimal transportation problem* is to minimize the functional

$$I[\pi] = \int_{M \times M} c(x, y) \, \mathrm{d}\pi(x, y) \quad \text{for} \quad \pi \in \Pi(\mu, \nu) \tag{B.14}$$

over all transference plans $\Pi(\mu, \nu)$.

The optimal transportation cost between μ and ν is the value

$$T_c(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} I[\pi], \tag{B.15}$$

and transference plans $\pi \in \Pi(\mu,\nu)$ that realize this optimum are called *optimal transference plans*.

Since (B.14) is a convex optimalization problem it admits a dual formulation. Assume that the cost function c is lower semi-continuous, and define

$$J(\varphi,\psi) = \int_M \varphi \,\mathrm{d}\mu + \int_M \psi \,\mathrm{d}\nu \tag{B.16}$$

for all integrable functions $(\varphi,\psi) \in \mathcal{L} = L^1(\,\mathrm{d}\mu) \times L^1(\,\mathrm{d}\nu)$. Let Φ_c be the set of all measurable functions $(\varphi,\psi) \in \mathcal{L}$ such that

$$\varphi(x) + \psi(y) \le c(x,y) \tag{B.17}$$

for $\,\mathrm{d}\mu$-almost all $x \in M$ and $\,\mathrm{d}\nu$-almost all $y \in M$. Then (Villani, 2003, Th. 1.3)

$$\inf_{\Pi(\mu,\nu)} I[\pi] = \sup_{\Phi_c} J(\varphi,\psi). \tag{B.18}$$

For measures $\mu,\nu \in \mathcal{P}_F$ with representations

$$\mu = \sum_{i=1}^m a_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{j=1}^n b_j \delta_{y_j} \tag{B.19}$$

any measure in $\Pi(\mu,\nu)$ can be represented as a bistochastic $m \times n$ matrix $\pi = (\pi_{ij})_{i,j}$, where the source and sink conditions

$$\sum_{i=1}^m \pi_{ij} = b_j, \quad j = 1,2,\dots,n \quad \text{and} \quad \sum_{j=1}^n \pi_{ij} = a_i, \quad i = 1,2,\dots,m, \tag{B.20}$$

are the discrete analog of (B.13), and the problem is to minimize the objective function

$$\sum_{ij} \pi_{ij} c_{ij}, \tag{B.21}$$

where $c_{ij} = c(x_i,y_j)$ is the cost matrix.

Its dual formulation is to maximize

$$\sum_i \varphi_i a_i + \sum_j \psi_j b_j \tag{B.22}$$

under the constraint $\varphi_i + \psi_j \le c_{ij}$.

Example 13 (Discrete distance)**.** Consider the special cost $c(x, y) = 1_{x \neq y}$, i.e., the distance induced by the discrete topology. Then the total transportation cost is

$$T_c(\mu, \nu) = d_{\mathrm{TV}}(\mu, \nu). \tag{B.23}$$

The Kantorovich problem (B.14) is actually a relaxed version of Monge's transportation problem. In the latter, it is further required that no mass be split, so the transference plan π has the special form

$$\mathrm{d}\pi(x, y) = \mathrm{d}\mu(x)\delta[y = T(x)] \tag{B.24}$$

for some measurable map $T : M \to M$. The associated total transportation cost is then

$$I[\pi] = \int_M c(x, T(x))\, \mathrm{d}\mu(x), \tag{B.25}$$

and the condition (B.13) on the marginals translates as

$$\nu[B] = \mu[T^{-1}(B)] \quad \text{for all measurable} \quad B \subseteq M. \tag{B.26}$$

If this condition is satisfied, we call ν the *push-forward* of μ by T, denoted by $\nu = T\#\mu$. For measures $\mu, \nu \in \mathcal{P}_F$, the optimal transference plans in Kantorovich's problem (transportation problem) coincide with solutions to Monge's problem.

A further relaxation is obtained when the cost $c(x, y)$ is a distance. The dual (B.18) of the Kantorovich problem then takes the following form:

Theorem 9 (Kantorovich-Rubinstein (Villani, 2003)[ch. 1.2)**.**] Let $X = Y$ be a Polish space[1], and let c be lower semi-continuous. Then:

$$T_c(\mu, \nu) = \sup \left\{ \int_X \varphi \;\; \mathrm{d}(\mu - \nu); \quad \text{where} \right.$$
$$\left. \varphi \in L^1(d|\mu - \nu|) \quad \text{and} \quad \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{c(x, y)} \leq 1 \right\} \tag{B.27}$$

The Kantorovich-Rubinstein theorem implies that $T_d(\mu + \sigma, \nu + \sigma) = T_d(\mu, \nu)$, i.e., the invariance of the Kantorovich-Rubinstein distance under subtraction of mass (Villani, 2003, Corollary 1.16). In other words, the total cost only depends on the difference $\mu - \nu$. The Kantorovich problem is then equivalent to the Kantorovich-Rubinstein *transshipment problem*: Minimize $I[\pi]$ for all product measures $\pi : M \times M \to \mathbb{R}_+$, such that

$$\pi[A \times M] - \pi[M \times A] = (\mu - \nu)[A]$$

[1] A topological space is a Polish space if it is homeomorphic to a complete metric space that has a countable dense subset. This is a general class of spaces that are convenient to work with. Many spaces of practical interest fall into this category.

for all measureable sets $A \subseteq \mathcal{B}(M)$. This transshipment problem is a strongly relaxed version of the optimal transportation problem. For example, if $p > 1$ then the transshipment problem with cost $c(x, y) = ||x - y||^p$ has optimal cost zero (Villani, 2003). For this reason, the general transshipment problem is not investigated her.e

Example 14 (Assignment and transportation problem)**.** The discrete Kantorovich problem (B.19-B.21) is also known as the (Hitchcock) *transportation problem* in the literature on combinatorial optimization (Korte and Vygen, 2007). The special case where $m = n$ in the representation (B.19) is the *assignment problem*. Interestingly, as a consequence of the Birkhoff theorem, the latter is solved by a permutation σ mapping each source a_i to a unique sink $b_{\sigma(i)}$ $(i = 1, \ldots, n)$; confer (Bapat and Raghavan, 1997).

B.3 Optimal transportation distances

Let (M, d) be a metric space and consider the cost function $c(x, y) = d(x, y)^p$, if $p > 0$ and $c(x, y) = 1_{x \neq y}$ if $p = 0$. Recall that $T_c(\mu, \nu)$ denotes the cost of an optimal transference plan between μ and ν.

Definition 18 (Wasserstein distances)**.** Let $p \geq 0$. The *Wasserstein distance of order p* is $W_p(\mu, \nu) = T_{d^p}(\mu, \nu)^{1/p}$ if $p \in [1, \infty)$, and $W_p(\mu, \nu) = T_{d^p}(\mu, \nu)$ if $p \in [0, 1)$.

Denote by $\mathcal{P}_p$ the space of probability measures with finite moments of order p, i.e., such that

$$\int d(x_0, x)^p \, \mathrm{d}\mu(x) < \infty$$

for some $x_0 \in M$. The following is proved in (Villani, 2003, Th. 7.3):

Theorem 10. The Wasserstein distance W_p, $p \geq 0$, is a metric on $\mathcal{P}_p$.

The Wasserstein distances W_p are ordered: $p \geq q \geq 1$ implies, by Hölder's inequality, that $W_p \geq W_q$. On a normed space, the Wasserstein distances are minorized by the distance in means, such that

$$W_p(\mu, \nu) \geq \left|\left| \int_X x \, \mathrm{d}(\mu - \nu) \right|\right|_p \tag{B.28}$$

and behave well under rescaling:

$$W_p(\alpha\mu, \alpha\nu) = |\alpha| W_p(\mu, \nu),$$

where $\alpha\mu$ indicates the measure $m_\alpha \# \mu$, obtained by push-forward of multiplication by α. If $p = 2$ we have the additional subadditivity property

$$W_2(\alpha_1\mu_1 + \alpha_2\mu_2, \alpha_1\nu_1 + \alpha_2\nu_2) \leq \left(\alpha_1^2 W_2(\mu_1, \nu_1)^2 + \alpha_2^2 W_2(\mu_2, \nu_2)^2\right)^{1/2}.$$