Chapter 9

Probability metrics

9.1 Total variation distance

For probability measures μ_1 and μ_2 on a measurable space $(S, \mathcal{S})$, the total variation distance is defined by

$$d_{\rm tv}(\mu_1, \mu_2) \;=\; \sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_2(A)|. \tag{9.1.1}$$

Proposition 9.1.1. *$d_{\rm tv}$ is a metric on the space of probability measures on $(S, \mathcal{S})$.*

Proof. (i) Obviously $d_{\rm tv}(\mu_1, \mu_1) = 0$. On the other hand, if $d_{\rm tv}(\mu_1, \mu_2) = 0$, then $|\mu_1(A) - \mu_2(A)| = 0$ for all $A \in \mathcal{S}$, so that $\mu_1 = \mu_2$.

(ii) Obviously $d_{\rm tv}(\mu_1, \mu_2) = d_{\rm tv}(\mu_1, \mu_2)$.

(iii) Let μ_1, μ_2, μ_3 be probability measures on $(S, \mathcal{S})$. The triangle inequality for the Euclidean norm on the real line implies that

$$\begin{aligned}
\sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_3(A)| \;&\le\; \sup_{A \in \mathcal{S}} \Big(|\mu_1(A) - \mu_2(A)| + |\mu_2(A) - \mu_3(A)| \Big) \\
&\le\; \sup_{A \in \mathcal{S}} |\mu_1(A) - \mu_2(A)| + \sup_{A \in \mathcal{S}} |\mu_2(A) - \mu_3(A)|,
\end{aligned}$$

so that $d_{\rm tv}(\mu_1, \mu_3) \le d_{\rm tv}(\mu_1, \mu_2) + d_{\rm tv}(\mu_2, \mu_3)$. □

The following result provides a helpful symmetry property for densities of probability measures. Remember that by Radon–Nikodym theorem, any pair of probability measures admit density functions with respect to some reference measure.

Lemma 9.1.2. *Let μ_1, μ_2 be probability measures admitting density functions $f_1, f_2 \colon S \to \mathbb{R}_+$ with respect to a measure ν on $(S, \mathcal{S})$. Then*

$$\int_S (f_1 - f_2)_+ \, d\nu \ = \ \int_S (f_2 - f_1)_+ \, d\nu \ = \ \frac{1}{2} \int_S |f_1 - f_2| \, d\nu \tag{9.1.2}$$

and

$$\int_S (f_1 \wedge f_2) \, d\nu \ = \ 1 - \frac{1}{2} \int_S |f_1 - f_2| \, d\nu.$$

Draw a picture.

Proof. Observe that $|x - y| = (x - y)_+ + (y - x)_+$ where $a_+ = \max\{a, 0\}$ denotes the positive part of a real number a. Then

$$\int_S |f_1 - f_2| \, d\nu \ = \ \int_S (f_1 - f_2)_+ \, d\nu + \int_S (f_2 - f_1)_+ \, d\nu. \tag{9.1.3}$$

Denoting $A_1 = \{x \colon f_1(x) > f_2(x)\}$, we see that

$$\begin{aligned} (f_1 - f_2)_+ \ &= \ (f_1 - f_2) 1_{A_1}, \\ (f_2 - f_1)_+ \ &= \ (f_2 - f_1) 1_{A_1^c}. \end{aligned}$$

Hence

$$\begin{aligned} \int_S (f_1 - f_2)_+ \, d\nu \ &= \ \int_{A_1} (f_1 - f_2) \, d\nu \ = \ \mu_1(A_1) - \mu_2(A_1), \\ \int_S (f_2 - f_1)_+ \, d\nu \ &= \ \int_{A_1^c} (f_2 - f_1) \, d\nu \ = \ \mu_2(A_1^c) - \mu_1(A_1^c). \end{aligned}$$

Because $\mu_1(A_1) = 1 - \mu_1(A_1^c)$ and $\mu_2(A_1) = 1 - \mu_2(A_1^c)$, we find that the above integrals are equal to each other, and we conclude using (9.1.3) that (9.1.2) is valid.

Next, we note that

$$\int_S (f_1 \wedge f_2) \, d\nu \ = \ \int_{A_1} f_2 \, d\nu + \int_{A_1^c} f_1 \, d\nu \ = \ \mu_2(A_1) + \mu_1(A_1^c) \ = \ 1 - \mu_1(A_1) + \mu_2(A_1).$$

It follows that

$$\int_S (f_1 \wedge f_2) \, d\nu \ = \ 1 - \int_S (f_1 - f_2)_+ \, d\nu \ = \ 1 - \frac{1}{2} \int_S |f_1 - f_2| \, d\nu.$$

$\square$

Proposition 9.1.3. *Let μ_1 and μ_2 be probability measures on $(S, \mathcal{S})$ admitting densities*[1] $f_1, f_2 \colon S \to \mathbb{R}_+$ *with respect to a reference measure ν on $(S, \mathcal{S})$. Then*

$$d_{\mathrm{tv}}(\mu_1, \mu_2) \;=\; \frac{1}{2} \int_S |f_1(x) - f_2(x)| \, \nu(dx). \tag{9.1.4}$$

Proof. (i) By Lemma 9.1.2, we see that

$$\frac{1}{2} \int_S |f_1 - f_2| \, d\nu \;=\; \int_S (f_1 - f_2)_+ \, d\nu.$$

By writing $(f_1 - f_2)_+ = (f_1 - f_2) 1_A$ for $A = \{x \colon f_1(x) > f_2(x)\}$, we see that

$$\int_S (f_1 - f_2)_+ \, d\nu \;=\; \int_A f_1 \, d\nu - \int_A f_2 \, d\nu \;=\; \mu_1(A) - \mu_2(A) \;\le\; |\mu_1(A) - \mu_2(A)|.$$

Hence $\frac{1}{2} \int_S |f_1 - f_2| \, d\nu \le d_{\mathrm{tv}}(\mu_1, \mu_2)$.

(ii) Fix a set $A \in \mathcal{S}$. Observe that $(f_1 - f_2) 1_A \le (f_1 - f_2)_+ 1_A \le (f_1 - f_2)_+$ pointwise. Hence

$$\mu_1(A) - \mu_2(A) \;=\; \int_A f_1 \, d\nu - \int_A f_2 \, d\nu \;=\; \int_S (f_1 - f_2) 1_A \, d\nu \;\le\; \int_S (f_1 - f_2)_+ \, d\nu.$$

Similarly, we find that

$$\mu_2(A) - \mu_1(A) \;\le\; \int_S (f_2 - f_1)_+ \, d\nu.$$

In light of Lemma 9.1.2, both of the rightmost integrals appearing in the above inequalities are equal to $\frac{1}{2} \int_S |f_1 - f_2| \, d\nu$. As a consequence,

$$|\mu_1(A) - \mu_2(A)| \;\le\; \frac{1}{2} \int_S |f_1 - f_2| \, d\nu.$$

Because this is true for all $A \in \mathcal{S}$, we see that $d_{\mathrm{tv}}(\mu_1, \mu_2) \le \frac{1}{2} \int_S |f_1 - f_2| \, d\nu$. □

The factor $\frac{1}{2}$ in front of the L_1-distance could be eliminated by normalising the total variation distance differently. The motivation for the current normalisation is that now $d_{\mathrm{tv}}(\mu_1, \mu_2) \in [0, 1]$ always, as confirmed by formula (9.1.1).

[1] Here we need densities to be finite-valued because we compute $f_1 - f_2$.

Example 9.1.4. Denote by $\mathrm{Ber}(p)$ the Bernoulli distribution with parameter $p \in [0,1]$. Determine the total variation distance between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$.

Recall that $\mathrm{Ber}(p)$ is a probability measure with density

$$f_p(x) = \begin{cases} 1-p & x=0, \\ p & x=1, \\ 0 & \text{else}, \end{cases}$$

with respect to the counting measure $\#$ on $(\mathbb{Z}, 2^{\mathbb{Z}})$. By Proposition 9.1.3,

$$\begin{aligned} d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) &= \frac{1}{2}\int_{\mathbb{Z}} |f_p(x) - f_q(x)| \, \#(dx) \\ &= \frac{1}{2}\sum_{x\in\mathbb{Z}} |f_p(x) - f_q(x)| \\ &= \frac{1}{2}\Big(|(1-p)-(1-q)| + |p-q|\Big) \\ &= |p-q|. \end{aligned}$$

9.2 Couplings

A coupling of probability measures μ_1 on $(S_1, \mathcal{S}_1)$ and μ_2 on $(S_2, \mathcal{S}_2)$ is a probability measure λ on $(S_1 \times S_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$ with marginal distributions μ_1, μ_2, that is,

$$\begin{aligned} \lambda(B_1 \times S_2) &= \mu_1(B_1) \quad \text{for all } B_1 \in \mathcal{S}_1, \\ \lambda(S_1 \times B_2) &= \mu_2(B_2) \quad \text{for all } B_2 \in \mathcal{S}_2. \end{aligned} \tag{9.2.1}$$

This is related to mass transportation.

Equivalently, a coupling is a pair (X_1, X_2) of random variables $X_1 \colon \Omega \to S_1$ and $X_2 \colon \Omega \to S_2$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathrm{Law}(X_1) = \mu_1$ and $\mathrm{Law}(X_2) = \mu_2$.

Proposition 9.2.1. $d_{\mathrm{tv}}(\mu_1, \mu_2) = \inf_{\lambda\in\Gamma(\mu_1,\mu_2)} \lambda\{(x_1, x_2) \colon x_1 \neq x_2\}$, *where* $\Gamma(\mu_1, \mu_2)$ *denotes the set of couplings of* μ_1 *and* μ_2*, and the infimum is attained by a coupling* λ_**.*

Proof. (i) Assume that λ is a coupling of μ_1 and μ_2. Then λ is a probability measure on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ with marginals μ_1 and μ_2. Then for any $A \in \mathcal{S}$,

$$\begin{aligned} \mu_1(A) - \mu_2(A) &= \lambda(A \times S) - \lambda(S \times A) \\ &= \int_{S\times S} \Big(1_{A\times S}(x_1,x_2) - 1_{S\times A}(x_1,x_2)\Big)\, \lambda(dx_1, dx_2) \\ &= \int_{S\times S} \Big(1_A(x_1) - 1_A(x_2)\Big)\, \lambda(dx_1, dx_2). \end{aligned}$$

We note that $1_A(x_1) - 1_A(x_2) = 0$ whenever $x_1 = x_2$. Therefore,

$$|1_A(x_1) - 1_A(x_2)| \ \le \ 1_D(x_1, x_2)$$

where $D = \{(x_1, x_2) \in S \times S \colon x_1 \neq x_2\}$. It follows that

$$\begin{aligned} |\mu_1(A) - \mu_2(A)| \ &\le \ \int_{S\times S} |1_A(x_1) - 1_A(x_2)| \, \lambda(dx_1, dx_2) \\ &\le \ \int_{S\times S} 1_D(x_1, x_2) \, \lambda(dx_1, dx_2) \\ &= \ \lambda(D). \end{aligned}$$

We conclude that

$$d_{\rm tv}(\mu_1, \mu_2) \le \lambda(D) \quad \text{for all couplings } \lambda. \tag{9.2.2}$$

(ii) We will construct[2] a coupling of μ_1 and μ_2. We assume that μ_1 and μ_2 admit[3] density functions $f_1, f_2 \colon S \to \mathbb{R}_+$ for some reference measure ν. Define

$$c \ = \ \int_S (f_1 \wedge f_2) \, d\nu.$$

Assume $0 < c < 1$, and define The case with $c = 0$ and the case with $c = 1$ are homeworks?

$$\begin{aligned} g_0(x) \ &= \ \frac{f_1(x) \wedge f_2(x)}{c}, \\ g_1(x) \ &= \ \frac{(f_1(x) - f_2(x))_+}{1-c}, \\ g_2(x) \ &= \ \frac{(f_2(x) - f_1(x))_+}{1-c}. \end{aligned}$$

With the help of Lemma 9.1.2, we see that $\int_S g_k \, d\nu = 1$ for all k, so that the weighted measures $\mu_k(A) = \int_A g_k \, d\nu$ are probability measures on $(S, \mathcal{S})$. Now define (see Remark 9.2.3 for an intuitive meaning)

$$\lambda_* \ = \ c(\mu_0 \circ \psi^{-1}) + (1-c)(\mu_1 \otimes \mu_2),$$

where $\psi \colon x \mapsto (x, x)$. Being a linear combination of probability measures $\mu_0 \circ \psi^{-1}$ and $\mu_1 \otimes \mu_2$, we see that λ_* is a probability measure on $(S \times S, \mathcal{S} \otimes \mathcal{S})$.

[2] This could be in appendix, not the most important thing.

[3] This is without loss of generality. Let $\nu = \mu_1 + \mu_2$. This is a finite measure that dominates μ_1 and μ_2 in the sense that $\nu(A) = 0 \implies \mu_1(A) = 0$ and $\mu_2(A) = 0$. By the Radon–Nikodym theorem ref there exist densities $f_1, f_2 \colon S \to \mathbb{R}_+$ of μ_1, μ_2 with respect to ν.

(iii) Let us verify that λ_* is a coupling of μ_1 and μ_2. Fix a set $B_1 \in \mathcal{S}$. We note that

$$\psi^{-1}(B_1 \times S) \;=\; \{x \in S : (x,x) \in B_1 \times S\} \;=\; B_1.$$

Hence

$$\begin{aligned}\lambda_*(B_1 \times S) \;&=\; c\,\mu_0(\psi^{-1}(B_1 \times S)) + (1-c)\,(\mu_1 \otimes \mu_2)(B_1 \times S)\\ &=\; c\mu_0(B_1) + (1-c)\mu_1(B_1),\end{aligned}$$

so that by plugging in the density formulas, we see that

$$\lambda_*(B_1 \times S) \;=\; \int_{B_1} \Big((f_1 \wedge f_2) + (f_1 - f_2)_+\Big)\, d\nu \;=\; \int_{B_1} f_1\, d\nu \;=\; \mu_1(B_1).$$

A similar computation shows that $\lambda_*(S \times B_2) = \mu_2(B_2)$ for all $B_2 \in \mathcal{S}$. Hence λ_* is a coupling of μ_1 and μ_2.

(iv) Finally, by noting that $\psi^{-1}(D) = \emptyset$, we find that

$$\lambda_*(D) \;=\; (1-c)(\mu_1 \otimes \mu_2)(D) \;\le\; 1-c \;=\; d_{\mathrm{tv}}(\mu_1, \mu_2).$$

In light of (9.2.2), we conclude that

$$\lambda_*(D) \;=\; \inf_{\lambda \in \Gamma(\mu_1,\mu_2)} \lambda(D) \;=\; d_{\mathrm{tv}}(\mu_1, \mu_2).$$

□

Example 9.2.2 (Coupling two coins)**.** Construct a coupling λ of Bernoulli distributions $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ such that $0 \le p \le q \le 1$, for which the probability $\lambda\{(i,j) : i \ne j\}$ is small.

Define a probability mass function on $\mathbb{Z}^2$ by $h(i,j) = L_{ij}$ for $i,j \in \{0,1\}$ and $f(i,j) = 0$ otherwise, where

$$L \;=\; \begin{bmatrix} 1-q & q-p \\ 0 & p \end{bmatrix}.$$

Then the probability measure $\lambda(A) = \sum_{(i,j)\in A} h(i,j)$ on $(\mathbb{Z}^2, 2^{\mathbb{Z}^2})$ has marginals $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$, and

$$\lambda\{(i,j) : i \ne j\} \;=\; L_{01} + L_{10} \;=\; q - p.$$

Hence by the coupling inequality ref , we find that $d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \le q - p$.

We saw in Example 9.1.4 that $d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) = q - p$. Hence the λ is actually an optimal coupling.

Remark 9.2.3. A probabilistic interpretation of Proposition 9.2.1 is obtained by construction random variables X_1, X_2 whose joint law is the optimal coupling λ_*. Let I, W_0, W_1, W_2 be independent random variables defined on some probability space such that $\mathrm{Law}(I) = \mathrm{Ber}(c)$ and $\mathrm{Law}(W_k) = \mu_k$ for $k = 0, 1, 2$. Define

$$X_1 = \begin{cases} W_0 & I = 1, \\ W_1 & I = 0, \end{cases} \qquad \text{and} \qquad X_2 = \begin{cases} W_0 & I = 1, \\ W_2 & I = 0. \end{cases}$$

Then the joint law of X_1 and X_2 equals the optimal coupling λ_* (homework).

Lindvall [Lin92] points out a subtle thing: To compute $\mathbb{P}(X_1 \neq X_2)$ the diagonal $\{(x, x) \colon x \in S\}$ should be a measurable set in $S \otimes S$. This is ok for Polish spaces.

9.3 Convergence in total variation

Convergence in total variation for discrete probability spaces corresponds to pointwise convergence of probability mass functions. Somewhat surprisingly, pointwise convergence and L_1-convergence are equivalent in this setting.

Proposition 9.3.1. *Let S be countable. Then the following are equivalent for probability measures μ_n, μ on $(S, 2^S)$ with probability mass functions f_n, f:*

(i) $d_{\mathrm{tv}}(\mu_n, \mu) \to 0$.

(ii) $f_n(x) \to f(x)$ *for every* $x \in S$.

(iii) $\sum_{x \in S} |f_n(x) - f(x)| \to 0$.

Proof. (i) $\iff$ (iii) follows by Proposition 9.1.3.

(iii) $\implies$ (ii) is obvious.

(ii) $\implies$ (iii). Assume that $f_n(x) \to f(x)$ for every $x \in S$. Enumerate $S = \{x_1, x_2, \dots\}$. Fix $\epsilon > 0$. Because $\sum_{k=1}^{\infty} f(x_k) = 1$, we may fix an integer $K \geq 1$ such that $\sum_{k>K}^{\infty} f(x_k) \leq \epsilon$. Then

$$\begin{aligned} \sum_{k>K} f_n(x_k) &= \sum_{k>K} f(x_k) + \sum_{k>K} (f_n(x_k) - f(x_k)) \\ &= \sum_{k>K} f(x_k) + \sum_{k \leq K} (f(x_k) - f_n(x_k)) \\ &\leq \sum_{k>K} f(x_k) + \sum_{k \leq K} |f_n(x_k) - f(x_k)|. \end{aligned}$$

Hence

$$\begin{aligned}\sum_{x\in S}|f_n(x)-f(x)| &= \sum_{k\le K}|f_n(x_k)-f(x_k)|+\sum_{k>K}|f_n(x_k)-f(x_k)| \\ &\le \sum_{k\le K}|f_n(x_k)-f(x_k)|+\sum_{k>K}(f_n(x_k)+f(x_k)) \\ &\le 2\sum_{k\le K}|f_n(x_k)-f(x_k)|+2\sum_{k>K}f(x_k) \\ &\le 2K\max_{k\le K}|f_n(x_k)-f(x_k)|+2\epsilon.\end{aligned}$$

By taking limits as $n\to\infty$, we find that

$$\limsup_{n\to\infty}\sum_{x\in S}|f_n(x)-f(x)| \le 2\epsilon.$$

Because the above inequality is true for all $\epsilon>0$, we conclude that (iii) holds. □

9.4 Poisson approximation

Let $X_1,\dots,X_n$ be mutually independent $\mathrm{Ber}(p)$-distributed random variables defined on some probability space $(\Omega,\mathcal{A},\mathbb{P})$. Define $S_n = X_1+\cdots+X_n$. Observe that $\mathbb{E}S_n = \sum_{k=1}^n \mathbb{E}X_k = np$. When np is a small, a classical result, discovered by Siméon Poisson[4], is that S_n is approximately Poisson distributed.

Proposition 9.4.1. *When $p_n = \lambda/n$ for some constant $0<\lambda<\infty$, then $\mathrm{Bin}(n,p_n)\to\mathrm{Poi}(\lambda)$ in total variation as $n\to\infty$.*

Proof. Fix an integer $n\ge 1$. We construct a coupling of $\mathrm{Bin}(n,p_n)$ and $\mathrm{Poi}(\lambda)$ as follows. Let λ be an optimal coupling of $\mathrm{Ber}(p_n)$ and $\mathrm{Poi}(p_n)$, so that $\lambda\{(x_1,\tilde{x}_1)\colon x_1\ne\tilde{x}_1\} = d_{\mathrm{tv}}(\mathrm{Ber}(p_n),\mathrm{Poi}(p_n))$. Define

$$\begin{aligned} S_n &= X_1+\cdots+X_n, \\ \tilde{S}_n &= \tilde{X}_1+\cdots+\tilde{X}_n,\end{aligned}$$

where $(X_1,\tilde{X}_1),\dots,(X_n,\tilde{X}_n)$ are independent λ-distributed random variables in $\mathbb{Z}^2$, defined on some probability space $(\Omega,\mathcal{A},\mathbb{P})$. Then $\mathrm{Law}(S_n) = \mathrm{Bin}(n,p_n)$ and[5] $\mathrm{Law}(\tilde{S}_n)=\mathrm{Poi}(np_n)$. Hence the joint law $\lambda_n = \mathrm{Law}(S_n,\tilde{S}_n)$

[4] 1781 – 1840, PhD École Polytechnique 1800 for Lagrange and Laplace.

[5] This is a preliminary, that the sum of independent Poisson random variables is Poisson.

constitutes a coupling of $\mathrm{Bin}(n, p_n)$ and $\mathrm{Poi}(np_n)$. The construction of the coupling shows that $S_n \neq \tilde{S}_n$ is possible only when $X_k \neq \tilde{X}_k$ for some $k = 1, \dots, n$. Hence the union bound implies that

$$\mathbb{P}(S_n \neq \tilde{S}_n) \ \le \ \sum_{k=1}^n \mathbb{P}(X_k \neq \tilde{X}_k).$$

We conclude by the coupling inequality that

$$d_{\mathrm{tv}}(\mathrm{Bin}(n, p_n), \mathrm{Poi}(np_n)) \ \le \ n\, d_{\mathrm{tv}}(\mathrm{Ber}(p_n), \mathrm{Poi}(p_n)). \tag{9.4.1}$$

Next, with the help of Proposition 9.1.3 we note that (exercise)

$$d_{\mathrm{tv}}(\mathrm{Ber}(p), \mathrm{Poi}(p)) \ = \ p(1 - e^{-p}) \qquad \text{for all } 0 \le p \le 1. \tag{9.4.2}$$

By plugging this into (9.4.1) and applying the bound $1-t \le e^{-t}$, we conclude that

$$d_{\mathrm{tv}}(\mathrm{Bin}(n, p_n), \mathrm{Poi}(np_n)) \ \le \ np_n^2.$$

Recalling that $p_n = \lambda/n$, we see that

$$d_{\mathrm{tv}}(\mathrm{Bin}(n, p_n), \mathrm{Poi}(\lambda)) \ \le \ \lambda^2/n \ \to \ 0 \qquad \text{as } n \to \infty.$$

□

9.5 Wasserstein distances

The Wasserstein distance[6] of order p between probability measures on a metric space (S, d) is defined by

$$W_p(\mu_1, \mu_2) \ = \ \inf_{\lambda \in \Gamma(\mu_1, \mu_2)} \left(\int_{S \times S} d(x_1, x_2)^p \, \lambda(dx_1, dx_2) \right)^{1/p}, \tag{9.5.1}$$

where $\Gamma(\mu_1, \mu_2)$ denotes the set of coupling of μ_1 and μ_2. The Wasserstein distance W_1 is also called *earth mover's distance*, because it can be viewed as a minimum transportation cost in the following setting:

- $\mu_1(dx_1)$ is the amount of mass supplied at x_1,
- $\mu_2(dx_2)$ is the amount of mass demanded at x_2,
- $d(x_1, x_2)$ is the transportation cost from x_1 to x_2.

[6] Named after Leonid Vaserstein (1944–). PhD 1969 @ Moscow State University.

A coupling λ corresponds to a transportation plan in which $\lambda(dx_1, dx_2)$ is the amount of mass transported from x_1 to x_2. The cost of the transportation plan is $\int_{S\times S} d(x_1, x_2)\, \lambda(dx_1, dx_2)$. The constraint $\lambda \in \Gamma(\mu_1, \mu_2)$ means that the transportation plan meets supply and demand.

Example 9.5.1 (Discrete metric)**.** For the metric $d_0(x, y) = 1(x \neq y)$, we see that

$$\int_{S\times S} d_0(x_1, x_2)\, \lambda(dx_1, dx_2) \;=\; \lambda\{(x_1, x_2) \colon x_1 \neq x_2\}.$$

Proposition 9.2.1 tells that the Wasserstein distance W_1 corresponding to the discrete metric equals the total variation distance.

Example 9.5.2 (Euclidian metric)**.** Consider the space $\mathbb{R}^n$ equipped with the metric $d(x, y) = \|x - y\|$ induced by the Euclidean norm $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$. Let $\mathcal{P}_1(\mathbb{R}^n)$ be the space of probability measures μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\int_{\mathbb{R}^n} \|x\|\, \mu(dx) < \infty$. It is possible but not that easy to prove that W_1 is a metric on $\mathcal{P}_1(\mathbb{R}^n)$, see [AGS08, Vil09].

9.6 Wasserstein distances on the real line

Wasserstein distances are in general not easy to compute in analytical form. Neither are optimal coupling achieving a minimum in (9.5.1) easy to find. An exception is the case of univariate probability distributions on the real line, for which an optimal coupling can be formed by a standard simulation method known as inverse transform sampling. In deriving a simple formula for Wasserstein distances for probability distributions on $\mathbb{R}$, the following formulas will turn out useful.

Lemma 9.6.1. *For any (possibly dependent) real-valued random variables X and Y defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$,*

$$\mathbb{E}(Y - X)_+ \;=\; \int_{\mathbb{R}} \mathbb{P}(X \le t < Y)\, dt, \tag{9.6.1}$$

$$\mathbb{E}|Y - X| \;=\; \int_{\mathbb{R}} \Big(\mathbb{P}(X \le t < Y) + \mathbb{P}(Y \le t < X)\Big)\, dt. \tag{9.6.2}$$

Proof. The Lebesgue measure of any real interval $[x, y)$ can be expressed either as the interval length $(y - x)_+$, or as the integral of the indicator $\int_{\mathbb{R}} 1_{[x,y)}(t)\, dt$. As a consequence, we see that

$$(Y(\omega) - X(\omega))_+ \;=\; \int_{\mathbb{R}} 1_{[X(\omega), Y(\omega))}(t)\, dt \;=\; \int_{\mathbb{R}} 1_{A_t}(\omega)\, dt,$$

where $A_t = \{\omega\colon X(\omega) \le t < Y(\omega)\}$. By taking expectations and using Fubini's theorem to interchange the expectation and the integral, we find that

$$\mathbb{E}(Y-X)_+ \;=\; \int_{\mathbb{R}} \mathbb{E}1_{A_t}\,dt \;=\; \int_{\mathbb{R}} \mathbb{P}(A_t)\,dt,$$

which confirms (9.6.1).

A symmetric argument shows that formula (9.6.1) also holds with the roles of X and Y swapped. By writing $|Y-X| = (Y-X)_+ + (X-Y)_+$, and taking expectations, we find that

$$\mathbb{E}|Y-X| \;=\; \mathbb{E}(Y-X)_+ + \mathbb{E}(X-Y)_+.$$

Formula (9.6.2) thens follows by applying (9.6.1) and its symmetric analogue. □

Proposition 9.6.2. *For probability measures on* $(\mathbb{R},\mathcal{B}(\mathbb{R}))$*, the Wasserstein distance of order 1 can be computed by* $W_1(\mu_1,\mu_2) = \int_{\mathbb{R}} |F_1(t)-F_2(t)|\,dt$ *where* $F_i(t) = \mu_i((-\infty,t])$ *is the cumulative distribution function of* μ_i*.*

Proof. (i) We construct a coupling of μ_1 and μ_2 using a method called *inverse transform sampling* that is a standard method to simulate random variables from a given univariate probability distribution. Assume that the F_1, F_2 are invertible[7]. Then define $X_1 = F_1^{-1}(U)$ and $X_2 = F_2^{-1}(U)$ with U being uniformly distributed in $(0,1)$. Then $\mathrm{Law}(X_1,X_2)$ is a coupling of μ_1 and μ_2 (check this yourself), and

$$\mathbb{E}|X_1-X_2| \;=\; \mathbb{E}|F_1^{-1}(U)-F_2^{-1}(U)| \;=\; \int_0^1 |F_1^{-1}(u)-F_2^{-1}(u)|\,du.$$

We claim that

$$\int_0^1 |F_1^{-1}(u)-F_2^{-1}(u)|\,du \;=\; \int_{\mathbb{R}} |F_1(t)-F_2(t)|\,dt.$$

By Lemma 9.6.1, we see that

$$\mathbb{E}|X_1-X_2| \;=\; \int_{\mathbb{R}} \Big(\mathbb{P}(X_1 \le t < X_2) + \mathbb{P}(X_2 \le t < X_1)\Big)\,dt.$$

We note that

$$\begin{aligned}\mathbb{P}(X_1 \le t < X_2) \;&=\; \mathbb{P}(F_1^{-1}(U) \le t < F_2^{-1}(U))\\ &=\; \mathbb{P}(F_2(t) < U \le F_1(t)).\end{aligned}$$

[7]If they are not, we use use a generalised inverse, that is, a quantile function.

Because $\mathbb{P}(U \in B)$ equals the Lebesgue measure of B for any $B \subset [0,1]$, we conclude that

$$\mathbb{P}(X_1 \le t < X_2) \;=\; (F_1(t) - F_2(t))_+.$$

By symmetry, the above formula holds also with the roles of X_1 and X_2 swapped. We conclude that

$$\begin{aligned}\mathbb{E}|X_1 - X_2| \;&=\; \int_{\mathbb{R}} \Big(\mathbb{P}(X_1 \le t < X_2) + \mathbb{P}(X_2 \le t < X_1)\Big)\, dt \\ &=\; \int_{\mathbb{R}} \Big((F_1(t) - F_2(t))_+ + (F_2(t) - F_1(t))_+\Big)\, dt \\ &=\; \int_{\mathbb{R}} |F_1(t) - F_2(t)|\, dt.\end{aligned}$$

Hence $\lambda = \mathrm{Law}(X_1, X_2)$ is a coupling of μ_1 and μ_2, for which

$$\int_{\mathbb{R}^2} |x_1 - x_2|\, \lambda(dx_1, dx_2) \;=\; \int_{\mathbb{R}} |F_1(t) - F_2(t)|\, dt. \tag{9.6.3}$$

(ii) It remains to show that no coupling of μ_1 and μ_2 attains a smaller value for $\int_{\mathbb{R}^2} |x_1 - x_2|\, \lambda(dx_1, dx_2)$ than the right side of (9.6.3). Let $(X_1, X_2) \in \mathbb{R}^2$ be random vector such that $\mathrm{Law}(X_1) = \mu_1$ and $\mathrm{Law}(X_2) = \mu_2$. By Lemma 9.6.1, we see that

$$\mathbb{E}|X_1 - X_2| \;=\; \int_{\mathbb{R}} \Big(\mathbb{P}(X_1 \le t < X_2) + \mathbb{P}(X_1 \le t < X_2)\Big)\, dt.$$

We also note that

$$\begin{aligned}\mathbb{P}(X_1 \le t < X_2) \;&=\; \mathbb{P}(X_1 \le t, X_2 > t) \;=\; F_1(t) - F_{12}(t), \\ \mathbb{P}(X_2 \le t < X_1) \;&=\; \mathbb{P}(X_2 \le t, X_1 > t) \;=\; F_2(t) - F_{12}(t),\end{aligned}$$

where $F_i(t) = \mathbb{P}(X_i \le t)$ and $F_{12}(t) = \mathbb{P}(X_1 \le t,\, X_2 \le t)$. Hence

$$\mathbb{E}|X_1 - X_2| \;=\; \int_{\mathbb{R}} \Big(F_1(t) + F_2(t) - 2F_{12}(t)\Big)\, dt. \tag{9.6.4}$$

Furthermore, $F_{12}(t) \le F_i(t)$ for $i = 1, 2$ implies that $F_{12}(t) \le F_1(t) \wedge F_2(t)$. We also note that the formula $x - (x \wedge y) = (x - y)_+$ implies that

$$x + y - 2(x \wedge y) \;=\; (x - y)_+ + (y - x)_+ \;=\; |x - y|.$$

Therefore, (9.6.4) implies that

$$\begin{aligned}\mathbb{E}|X_1 - X_2| \;&\ge\; \int_{\mathbb{R}} \Big(F_1(t) + F_2(t) - 2(F_1(t) \wedge F_2(t))\Big)\, dt \\ &=\; \int_{\mathbb{R}} |F_1(t) - F_2(t)|\, dt.\end{aligned}$$

Because the above inequality holds for all random vectors (X_1, X_2) with $\mathrm{Law}(X_1) = \mu_1$ and $\mathrm{Law}(X_2) = \mu_2$, we conclude that

$$\int_{\mathbb{R}} |F_1(t) - F_2(t)| \, dt \;\leq\; W_1(\mu_1, \mu_2).$$

$\square$