

POLSKA AKADEMIA NAUK. INSTYTUT MATEMATYCZNY

DISSERTATIONES
MATHematicAE
(ROZPRAWY MATEMATYCZNE)

KOMITET REDAKCYJNY

BOGDAN BOJARSKI redaktor

WIESŁAW ŻELAZKO zastępca redaktora

ANDRZEJ BIAŁYNICKI-BIRULA, ZBIGNIEW CIESIELSKI,

JERZY ŁOŚ, ZBIGNIEW SEMADENI

CCXCIX

S. T. RACHEV and R. M. SHORTT

Duality theorems for
Kantorovich–Rubinstein and Wasserstein functionals

WARSZAWA 1990

PAŃSTWOWE WYDAWNICTWO NAUKOWE

S.7133



PRINTED IN POLAND

© Copyright by Instytut Matematyczny PAN, Warszawa 1990

Published by PWN-Polish Scientific Publishers

ISBN 83-01-09970-4

ISSN 0012-3862

W R O C Ł A W S K A D R U K A R N I A N A U K O W A

BUW-EO- 1 f

CONTENTS

§0. Introduction	5
§1. Notation and terminology	6
§2. A generalization of the Kantorovich–Rubinstein theorem	8
§3. Application: explicit representations for a class of probability metrics	14
§4. Topology of the Kantorovich–Rubinstein norm	18
§5. Dual representation for the Wasserstein functional	21
§6. Comparison of Wasserstein functional and Kantorovich Rubinstein norm; completeness	27
§7. Convergence of empirical measures; results of Fortet–Mourier type	30
§8. The convex set of optimal measures	32
References	34

§0. Introduction*

The Kantorovich–Rubinstein duality theorem has a long and colourful history, apparently originating in 18th century work of Monge on the transport of mass problem. For a detailed survey, we refer the reader to the article of Rachev [18]. Roughly speaking, this duality theorem has two basic forms. Given probabilities P_1 and P_2 on a space S and a measurable cost function $c(x, y)$ on $S \times S$, one considers two functionals

$$\hat{\mu}_c(P_1, P_2) = \inf \int c(x, y) db(x, y),$$

where the infimum is taken over all probabilities b on $S \times S$ with marginals $b_1 = P_1$ and $b_2 = P_2$; also

$$\check{\mu}_c(P_1, P_2) = \inf \int c(x, y) db(x, y),$$

where the infimum is over all finite measures b on $S \times S$ with marginal difference $b_1 - b_2 = P_1 - P_2$. These are sometimes called the Wasserstein and Kantorovich–Rubinstein functionals, respectively.

Duality theorems for these functionals are of the general form

$$(1) \quad \hat{\mu}_c(P_1, P_2) = \sup \int f dP_1 + \int g dP_2,$$

where the supremum is taken over a class of functions f, g on S such that $f(x) + g(y) \leq c(x, y)$; also,

$$(2) \quad \check{\mu}_c(P_1, P_2) = \sup \int f d(P_1 - P_2),$$

with the supremum taken over a class of $f: S \rightarrow \mathbf{R}$ satisfying the “Lipschitz” condition $f(x) - f(y) \leq c(x, y)$. When the probabilities in question have a finite support, these become linear programming results. (See [13], [14].)

Such duality theorems have absorbed the attention of a great number of researchers in a correspondingly large number of papers. (See [5]–[8], [10], [11], [15], [16], [18], [21].)

Results for (2) were obtained by Kantorovich and Rubinstein [10] for (S, d) a compact metric space with $c(x, y) = d(x, y)$. An attempt to generalize this result to separable metric spaces was made by Dudley [6]. (See [5] for details.) A proof for Polish (separable and topologically complete, metrizable)

* This research was partially supported by an NSF grant.

spaces has been given by Fernique [7], who uses some results from linear programming. Another argument may be found in Szulga [23], [24], who assumes that every probability on S is the law of some S -valued random variable. See Theorem 1.2 *infra*.

In §2, we follow the basic idea of Dudley to its logical conclusion, obtaining duality (2) for separable metric spaces and cost functions $c(x, y)$ which are not necessarily metrics. The supremum in (2) is shown to be attained for some optimal function f .

This level of generality enables us to obtain an explicit representation of μ with respect to a certain form of $c(x, y)$ for distributions on the line (Theorem 3.1). For $c(x, y) = |x - y|$, this yields a well-known result stated by Vallander [25]. This representation is later used to obtain a formula for a class of metrics on $\mathcal{P}(\mathbf{R})$ first considered by Fortet and Mourier [8]. See our Theorem 7.7.

Even for rather general cost functions, the Kantorovich–Rubinstein norm μ is a metric on an appropriate sub-class of $\mathcal{P}(S)$. The topology of this metric is explored in §4, where issues of convergence and compactness are considered.

The history of dualities of the form (1) is quite complex; again, see [18]. Recently, sophisticated duality theorems for $\hat{\mu}$ have been obtained by H. Kellerer [11], [13]. His results apply to very general cost functions $c(x, y)$, though they do require tightness of the measures involved. In [12], he proves a duality theorem with no tightness required, under the assumption that $c(x, y)$ is a metric. In §5, we improve these findings, requiring no tightness conditions, allowing fairly general $c(x, y)$, and restricting the class of functions f in the supremum (1) to those satisfying a sort of generalized Lipschitz condition (Theorem 5.5).

Following an idea of Dudley and Neveu [17], we show in §6 that $\hat{\mu} = \hat{\mu}$ just in case $c(x, y)$ is a metric. For compact spaces, this result was proved by Kantorovich and Rubinstein [10] and Levin and Milyutin [15], for Polish spaces by de Acosta [5], and for universally measurable spaces by Levin [16]. Also, we consider the completeness of the metric $\hat{\mu}$ in this case.

Next, we prove convergence of empirical measures to their theoretical distribution for the metric μ . This proof, together with a result on the measurability of $\mu(P_n, P)$ (P_n empirical) and applications to a class of metrics associated with the work of Fortet and Mourier [8] constitutes the substance of §7.

Finally, in §8, we consider the geometry of the convex set of optimal measures in (1) and (2) for $c(x, y)$ a metric.

§ 1. Notation and terminology

We work exclusively in the context of separable metric spaces. All of the measures considered will be defined on the Borel σ -field of such spaces. If (S, d) is a metric space, denote by $\mathcal{B}(S)$ and $\mathcal{M}(S)$ the Borel σ -field of S and the

collection of (non-negative) finite Borel measures on S , respectively. Let $\mathcal{P}(S)$ be the set of probability measures in $\mathcal{M}(S)$. We will consider the usual "weak topology" on these spaces induced by integration against continuous bounded functions on S . If $x \in S$, then δ_x is the point mass at x .

If A is a set, then I_A is the indicator function of A defined by

$$I_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

If ϕ is a logical predicate, then

$$I(\phi(x)) = I\{x: \phi(x)\} = I_A,$$

where $A = \{x: \phi(x) \text{ is true}\}$. Thus $\delta_x(A) = I_A(x)$.

If (S, d) is a separable metric space, then each finite Borel measure b on $S \times S$ induces "marginal" measures b_1, b_2 in $\mathcal{M}(S)$ defined by

$$b_1(A) = b(A \times S), \quad b_2(A) = b(S \times A).$$

We shall require the following result, which generalizes a well-known theorem of Strassen [22] (also cf. [21]).

1.1. THEOREM (Strassen). *Suppose that (S, d) is a separable metric space and that $P_n \rightarrow P$ weakly in $\mathcal{P}(S)$. Then for each $\varepsilon, \delta > 0$ there is some N such that whenever $n \geq N$, there is some probability b_n on $S \times S$ with marginals P_n and P such that*

$$b_n\{(x, y): d(x, y) > \delta\} < \varepsilon.$$

Indication. See [6].

Now suppose that $(\Omega, \mathcal{A}, \text{Pr})$ is an abstract probability space. An S -valued random variable is a Borel measurable function $X: \Omega \rightarrow S$, where S is a separable metric space. Each such random variable induces a probability $\mathcal{L}(X)$ in $\mathcal{P}(S)$ (the law of X) defined by $\mathcal{L}(X)(A) = \text{Pr}(X \in A)$. We note

1.2. THEOREM. *There is a probability space $(\Omega, \mathcal{A}, \text{Pr})$ such that for each separable metric space (S, d) and each $P \in \mathcal{P}(S)$, there is a random variable $X: \Omega \rightarrow S$ with $\mathcal{L}(X) = P$.*

Proof. Let \mathcal{A} be the collection of all pairs (T, Q) , where $T \subseteq \mathbf{R}$ and Q is a probability measure on $(T, \mathcal{B}(T))$. Then the cardinality of \mathcal{A} is 2^c . Define

$$(\Omega, \mathcal{A}, \text{Pr}) = \prod \{(T, \mathcal{B}(T), Q): (T, Q) \in \mathcal{A}\}$$

as the measure-theoretic (von Neumann) product. For each $(T, Q) \in \mathcal{A}$, let $\pi(T, Q)$ be projection from \mathcal{A} to the factor indexed by (T, Q) .

The theorem follows upon noting that every separable metric space is Borel-isomorphic with some subset of the real line. See [3; p. 7]. ■

Unless otherwise specified, we shall demand that the probability space underlying our work is as rich as indicated in the theorem, *viz.* if $\mathcal{X}(S)$ is the collection of all S -valued random variables on $(\Omega, \mathcal{A}, \text{Pr})$, then we assume

$$P(S) = \{\mathcal{L}(X): X \in \mathcal{X}(S)\}.$$

In the papers of Szulga [23], [24] this hypothesis is tacitly assumed.

§2. A generalization of the Kantorovich–Rubinstein theorem

Let (S, d) be a separable metric space. Suppose that $c: S \times S \rightarrow [0, +\infty)$ and $\lambda: S \rightarrow [0, +\infty)$ are measurable functions such that

(C1) $c(x, y) = 0$ if and only if $x = y$;

(C2) $c(x, y) = c(y, x)$ for x, y in S ;

(C3) $c(x, y) \leq \lambda(x) + \lambda(y)$ for $x, y \in S$;

(C4) λ maps bounded sets to bounded sets;

(C5) $\sup\{c(x, y); x, y \in B(a; R), d(x, y) \leq \delta\}$ tends to 0 as $\delta \rightarrow 0$ for each $a \in S$ and $R > 0$.

Here, $B(a; R) = \{x \in S: d(x, a) < R\}$. Given a real-valued function $f: S \rightarrow \mathbf{R}$, we define

$$\|f\|_c = \sup\{|f(x) - f(y)|/c(x, y): x \neq y\}$$

and set

$$L = \{f: \|f\|_c < +\infty\}.$$

It is easy to see that $\|\cdot\|_c$ is a semi-norm on the linear space L . Notice that for $f \in L$ we have

$$|f(x) - f(y)| \leq \|f\|_c c(x, y)$$

for all $x, y \in S$. It follows from condition (C5) on c that each function in L is continuous and hence measurable. Note also that $\|f\|_c = 0$ if and only if f is constant. Define L_0 to be the quotient of L modulo the constant functions. Then $\|\cdot\|$ is naturally defined on L_0 , and $(L_0, \|\cdot\|)$ is a normed linear space.

Now suppose that $M = M_\lambda(S)$ denotes the linear space of all finite signed measures m on S such that

$$m(S) = 0 \quad \text{and} \quad \int \lambda d|m| < \infty.$$

Here $|m| = m^+ + m^-$, where $m = m^+ - m^-$ is the Jordan decomposition of m .

For each $m \in M$, let $B(m)$ be the set of all finite measures b on $S \times S$ such that

$$b(A \times S) - b(S \times A) = m(A)$$

for each Borel $A \subseteq S$. Note that $B(m)$ is always non-empty, since it contains $(m^+ \otimes m^-)/m^+(S)$.

Define a function $m \rightarrow \|m\|_w$ on M by

$$\|m\|_w = \inf \left\{ \int c(x, y) db(x, y) : b \in B(m) \right\}.$$

We have

$$\begin{aligned} \|m\|_w &\leq \int c(x, y) d(m^+ \otimes m^-)(x, y) / m^+(S) \\ &\leq \int \lambda(x) dm^+(x) + \int \lambda(y) dm^-(y) = \int \lambda d|m| < \infty. \end{aligned}$$

For $c(x, y) = d(x, y)$, this quantity is sometimes called the Kantorovich–Rubinstein or Wasserstein norm of m .

We shall demonstrate that for probabilities P and Q on S with $P - Q \in M$, we have

$$\|P - Q\|_w = \sup \left\{ \left| \int f d(P - Q) \right| : \|f\|_c \leq 1 \right\}.$$

When $c(x, y) = d(x, y)$ and $\lambda(x) = d(x, a)$, a some fixed point of S , this is a straightforward generalization of the classical Kantorovich–Rubinstein duality theorem. See [18], [10], [6].

2.1. LEMMA. $\|\cdot\|_w$ is a semi-norm on M .

PROOF. Clearly, $\|m\|_w \geq 0$ for each $m \in M$. If $\alpha > 0$, then

$$B(\alpha m) = \{ab : b \in B(m)\},$$

so that $\|\alpha m\|_w = \alpha \|m\|_w$. Also,

$$B(-m) = \{\tilde{b} : b \in B(m)\},$$

where $\tilde{b}(A \times B) = b(B \times A)$. Also, $\|0\|_w = 0$, since $0 \in B(0)$. Therefore, $\|\alpha m\|_w = |\alpha| \|m\|_w$ for each $m \in M$ and $\alpha \in \mathbf{R}$.

To prove subadditivity, suppose that $b \in B(m)$ and $b' \in B(m')$. Then $b + b' \in B(m + m')$. It follows that $\|m + m'\|_w \leq \|m\|_w + \|m'\|_w$ for any m, m' in M . ■

Now given $m \in M$, $f \in L$, and a fixed $a \in S$, we have

$$\begin{aligned} |f(x)| &\leq |f(x) - f(a)| + |f(a)| \leq \|f\|_c c(x, a) + |f(a)| \\ &\leq \|f\|_c (\lambda(x) + \lambda(a)) + |f(a)| = K_1 \lambda(x) + K_2 \quad \text{all } x \in S, \end{aligned}$$

for constants $K_1, K_2 \geq 0$. Thus, each $f \in L$ is $|m|$ -integrable and induces a linear form $\varphi_f: M \rightarrow \mathbf{R}$ defined by

$$\varphi_f(m) = \int f dm.$$

Note that if f and g differ by a constant, then $\varphi_f = \varphi_g$. Given $b \in B(m)$, we have

$$\begin{aligned} |\varphi_f(m)| &= \left| \int f dm \right| = \left| \int (f(x) - f(y)) db(x, y) \right| \\ &\leq \int |f(x) - f(y)| db(x, y) \leq \|f\|_c \int c(x, y) db(x, y). \end{aligned}$$

Taking the infimum over all $b \in B(m)$ yields $|\varphi_f(m)| \leq \|f\|_c \|m\|_w$, so that φ_f is a continuous linear functional with

$$\|\varphi_f\|_w^* \leq \|f\|_c.$$

Thus, we may define a continuous linear transformation

$$(L_0, \|\cdot\|_c) \xrightarrow{D} (M^*, \|\cdot\|_w^*)$$

by $D(f) = \varphi_f$.

2.2. LEMMA. *The map D is an isometry.*

Proof. Note first that if $m_{xy} = \delta_x - \delta_y$ (some $x, y \in S$), then

$$\|m_{xy}\|_w \leq \int c(s, t) d(\delta_x \otimes \delta_y)(s, t) = c(x, y).$$

Then for each $f \in L$,

$$\begin{aligned} \|f\|_c &= \sup\{|f(x) - f(y)|/c(x, y) : x \neq y\} \\ &= \sup\{|\varphi_f(m_{xy})|/c(x, y) : x \neq y\} \\ &\leq \|\varphi_f\|_w^* \sup\{\|m_{xy}\|_w/c(x, y) : x \neq y\} \leq \|\varphi_f\|_w^*, \end{aligned}$$

so that $\|f\|_c = \|\varphi_f\|_w^*$, as claimed.

We now set about proving that the map D is surjective and hence an isometric isomorphism of Banach spaces. We need some preliminary facts. Let M_0 be the set of signed measures of the form $m = m_1 - m_2$, where m_1 and m_2 are finite measures on S with bounded support such that $m_1(S) = m_2(S)$. Condition (C4) on λ implies that $M_0 \subseteq M$.

2.3. LEMMA. *M_0 is a dense subspace of $(M, \|\cdot\|_w)$.*

Proof. Given $m \in M$ ($m \neq 0$), fix $a \in S$ and put $B_n = B(a; n)$ for $n = 1, 2, 3, \dots$. For all sufficiently large n , we have $m^+(B_n)m^-(B_n) > 0$. For such n ,

$$m_n(A) = m^+(S) \left[\frac{m^+(A \cap B_n)}{m^+(B_n)} - \frac{m^-(A \cap B_n)}{m^-(B_n)} \right].$$

We may also assume that $|m|(B_n^c) > 0$ for all $n \geq 1$. Now define $\delta_n \geq 0$ and $\varepsilon_n \geq 0$ by

$$\delta_n = \frac{m^-(S)}{m^-(B_n)} - 1, \quad \varepsilon_n = \frac{m^+(S)}{m^+(B_n)} - 1.$$

Then $\delta_n, \varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$. Also,

$$(m - m_n)(A) = m(A - B_n) - \varepsilon_n m^+(A \cap B_n) + \delta_n m^-(A \cap B_n).$$

Define finite measures μ_n and ν_n on S by

$$\mu_n(A) = m^+(A - B_n) + \delta_n m^-(A \cap B_n), \quad \nu_n(A) = m^-(A - B_n) + \varepsilon_n m^+(A \cap B_n).$$

Then $m - m_n = \mu_n - \nu_n$. Moreover, μ_n and ν_n are absolutely continuous with respect to $|m|$. Letting (P, N) be a Hahn decomposition for m , we determine the Radon-Nikodym derivatives

$$\frac{d\mu_n}{d|m|}(x) = \begin{cases} 1 & x \in P - B_n, \\ \delta_n & x \in N \cap B_n, \\ 0 & \text{otherwise,} \end{cases} \quad \frac{d\nu_n}{d|m|}(y) = \begin{cases} 1 & y \in N - B_n, \\ \varepsilon_n & y \in P \cap B_n, \\ 0 & \text{otherwise.} \end{cases}$$

Then the measure $b_n = (\mu_n \otimes \nu_n) / \mu_n(S)$ belongs to $B(m - m_n)$. Noting that

$$\begin{aligned} \nu_n(S) &= \mu_n(S) = m^+(S - B_n) + \delta_n m^-(B_n) \\ &= m^+(S - B_n) + (m^-(S) - m^-(B_n)) = |m|(S - B_n) = |m|(B_n^c), \end{aligned}$$

we write the Radon-Nikodym derivative

$$f_n(x, y) = \frac{db_n}{d(|m| \oplus |m|)}(x, y) = \frac{1}{|m|(B_n^c)} \frac{d\mu_n}{d|m|}(x) \frac{d\nu_n}{d|m|}(y).$$

Then we

CLAIM. *The function $g(x, y) = \sup_n f_n(x, y)c(x, y)$ is $|m| \otimes |m|$ -integrable.*

Proof of claim. We show that g is integrable over various subsets of $S \times S$.

(i) g is integrable over $P \times N$: We suppose that $x \in P$ and $y \in N$. Then

$$g(x, y) = \sum_{n=1}^{\infty} \frac{c(x, y)}{|m|(B_n^c)} I_{C_n}(x, y),$$

where $C_n = (B_n^c \times B_n^c) - (B_{n+1}^c \times B_{n+1}^c)$. So

$$\begin{aligned} \int_{P \times N} g d|m| \otimes |m| &\leq \sum_{n=1}^{\infty} \frac{1}{|m|(B_n^c)} \int_{C_n} (\lambda(x) + \lambda(y)) d|m| \otimes |m|(x, y) \\ &\leq \sum_{n=1}^{\infty} \frac{2}{|m|(B_n^c)} \int_{(B_n^c - B_{n+1}^c) \times B_n^c} \lambda(x) d|m| \otimes |m|(x, y) \\ &= 2 \sum_{n=1}^{\infty} \int_{B_n^c - B_{n+1}^c} \lambda(x) d|m|(x) = 2 \int_{B_1^c} \lambda d|m| < +\infty. \end{aligned}$$

(ii) $g(x, y) \leq Kc(x, y)$ for some $K \geq 0$ on $P \times P$: We suppose $x, y \in P$. Then

$$\begin{aligned} g(x, y) &\leq \frac{\varepsilon_n c(x, y)}{|m|(B_n^c)} = \frac{c(x, y)}{m^+(B_n^c)} (m(S) - m(B_n^c)) \frac{1}{|m|(B_n^c)} \\ &= \frac{m^+(B_n^c) c(x, y)}{|m|(B_n^c) m^+(B_n)} \leq \frac{c(x, y)}{m^+(B_1)}. \end{aligned}$$

Very similar arguments serve to demonstrate

(iii) $g(x, y) \leq Kc(x, y)$ for some $K \geq 0$ on $N \times N$.

(iv) $g(x, y) \leq Kc(x, y)$ for some $K \geq 0$ on $N \times P$.

Combining (i)–(iv) establishes the claim.

Now $f_n(x, y) \rightarrow 0$ as $n \rightarrow \infty$ for all $x, y \in S$. In view of the claim, Lebesgue's dominated convergence theorem implies that

$$\begin{aligned} \|m - m_n\|_w &\leq \int c(x, y) db_n(x, y) \\ &= \int c(x, y) f_n(x, y) d(|m| \otimes |m|)(x, y) \rightarrow 0 \end{aligned}$$

as $n \rightarrow \infty$. ■

Call a signed measure on S *simple* if it is a finite linear combination of signed measures of the form $\delta_x - \delta_y$. M contains all the simple measures.

2.4. LEMMA. *The simple measures are dense in $(M, \|\cdot\|_w)$.*

PROOF. In view of Lemmas 2.1 and 2.3, it is no loss of generality to assume that $m = P - Q$, where P and Q are probabilities on S supported on a bounded set $S_0 \subseteq S$. Then there are probabilities $P_n \rightarrow P$, $Q_n \rightarrow Q$ weakly such that for each n , we have $P_n(S_0) = Q_n(S_0) = 1$ and $P_n - Q_n$ simple. To prove the lemma, it is enough to show $\|P_n - P\|_w \rightarrow 0$ as $n \rightarrow \infty$.

Given $\varepsilon > 0$, use the boundedness of S_0 and condition (C5) on c to find $\delta > 0$ such that $c(x, y) < \varepsilon/2$ whenever $x, y \in S_0$ with $d(x, y) \leq \delta$. Put $K = \sup\{\lambda(x) : x \in S_0\}$. By Strassen's Theorem (1.1), for all large n , there is a probability b_n on $S \times S$ with marginals P_n and P such that

$$b_n\{(x, y) : d(x, y) > \delta\} < \varepsilon/4K.$$

Put $A = \{(x, y) : d(x, y) > \delta\}$ and $B = S - A$. Then

$$\begin{aligned} \|P_n - P\|_w &\leq \int c(x, y) db_n(x, y) = \int_A c(x, y) db_n(x, y) + \int_B c(x, y) db_n(x, y) \\ &\leq \int_A (\lambda(x) + \lambda(y)) db_n(x, y) + \varepsilon/2 \leq 2Kb_n(A) + \varepsilon/2 < \varepsilon \end{aligned}$$

for all large n . ■

2.5. LEMMA. *The linear transformation D is an isometric isomorphism of $(L_0, \|\cdot\|_c^*)$ onto $(M^*, \|\cdot\|_w^*)$.*

PROOF. Suppose that $\varphi : M \rightarrow \mathbf{R}$ is a continuous linear functional on M . Fix $a \in S$ and define $f : S \rightarrow \mathbf{R}$ by

$$f(x) = \varphi(\delta_x - \delta_a).$$

For any $x, y \in S$,

$$|f(x) - f(y)| = |\varphi(\delta_x - \delta_y)| \leq \|\varphi\|_w^* \|\delta_x - \delta_y\|_w \leq \|\varphi\|_w^* c(x, y),$$

so that $\|f\|_c \leq \|\varphi\|_w^* < \infty$. We see that $\varphi(m) = \varphi_f(m)$ for $m = \delta_x - \delta_y$, and hence for all simple $m \in M$. Lemma 2.4 implies that $\varphi(m) = \varphi_f(m)$ for all $m \in M$. Thus $\varphi = D(f)$.

We have shown that D is surjective. Earlier results now apply to complete the argument. ■

Now we consider the adjoint of the transformation D . As usual, the Hahn-Banach Theorem applies to show that

$$(L_0^*, \|\cdot\|_c^*) \xleftarrow{D^*} (M^{**}, \|\cdot\|_w^{**})$$

is an isometric isomorphism. Let

$$(M^{**}, \|\cdot\|_w^{**}) \xleftarrow{T} (M, \|\cdot\|_w)$$

be the canonical isometric embedding. Then

$$(L_0^*, \|\cdot\|_c^*) \xleftarrow{D^* \circ T} (M, \|\cdot\|_w)$$

is an isometry. A routine diagram chase shows that

$$\|m\|_w = \sup \left\{ \int f dm : \|f\|_c \leq 1 \right\}.$$

We summarize by stating the following, the main result of this section:

2.6. THEOREM. *Let m be a measure in M . Then*

$$\|m\|_w = \sup \left\{ \int f dm : f(x) - f(y) \leq c(x, y) \right\}.$$

We now show that the supremum in Theorem 2.6 is attained for some optimal f .

2.7. THEOREM. *Let m be a measure in M . Then there is some $f \in L$ with $\|f\|_c = 1$ such that*

$$\|m\|_w = \int f dm.$$

Proof. Using the Hahn-Banach Theorem, choose a linear functional φ in M^* with $\|\varphi\|_w^* = 1$ and such that $\varphi(m) = \|m\|_w$. By Lemma 2.5, we have $\varphi = \varphi_f$ for some $f \in L$ with $\|f\|_c = \|\varphi\|_w^* = 1$. ■

Given probability measures P_1, P_2 on S , define

$$\hat{\mu}_c(P_1, P_2) = \inf \left\{ \int c(x, y) db(x, y) : b \in B(P_1 - P_2) \right\}.$$

Let $\mathcal{P}_\lambda(S)$ be the set of all probabilities P on S such that λ is P -integrable. Then $\hat{\mu}_c(P_1, P_2)$ defines a pseudo-metric on $\mathcal{P}_\lambda(S)$. In a later section, we shall analyze the topological properties of these pseudo-metrics.

It should also be noted that if X and Y are random variables taking values in S , then it is natural to make the definition

$$\hat{\mu}_c(X, Y) = \hat{\mu}_c(\mathcal{L}(X), \mathcal{L}(Y)).$$

We shall freely use both notations in the next section.

2.8. EXAMPLE. Suppose that $c(x, y) = d(x, y)$ and put $\lambda(x) = d(x, a)$ for some $a \in S$. Then conditions (C1)–(C5) are satisfied, and Theorem 2.6 yields

$$\inf\left\{\int d(x, y)db(x, y): b \in B(P_1 - P_2)\right\} = \sup\left\{\int f d(P_1 - P_2): \|f\|_L \leq 1\right\},$$

where $P_1, P_2 \in \mathcal{P}_\lambda(S)$ and $\|f\|_L$ is the Lipschitz norm of f . In this case, $\hat{\mu}_c(P_1, P_2)$ is an actual metric. This, the classical situation, has been much studied: [18], [10], [6].

2.9. EXAMPLE. Take $S = \mathbf{R}$ with the usual metric $d(x, y) = |x - y|$. Suppose that

$$\begin{aligned} c(x, y) &= |x - y|^p, \quad p > 1, \\ \lambda(x) &= 2^{p-1}|x|^p. \end{aligned}$$

Again, conditions (C1)–(C5) are satisfied, but the space L contains only constant functions. Thus

$$\inf\left\{\int |x - y|^p db(x, y): b \in B(P_1 - P_2)\right\} = \sup\left\{\int f d(P_1 - P_2): \|f\|_c \leq 1\right\} = 0.$$

In this case, $\hat{\mu}_c(P_1, P_2)$ is identically zero (the trivial pseudo-metric). It is an amusing exercise to prove directly that the infimum that defines $\|P_1 - P_2\|_w$ is zero.

For example, suppose that $P_1 = \delta_0$ and $P_2 = \delta_1$. For each $n = 1, 2, \dots$ consider a measure b_n on $\mathbf{R} \times \mathbf{R}$ with total mass $2n + 1$ and

$$\begin{aligned} b_n\{(i/n, i/n)\} &= b_n\{(i/n, (i+1)/n)\} = 1, \\ b_n\{(1, 1)\} &= 1, \quad i = 0, 1, \dots, n-1. \end{aligned}$$

Then $b_n \in B(P_1 - P_2)$ and

$$\int_{\mathbf{R}^2} |x - y|^p db_n(x, y) = \sum_{i=0}^{n-1} \left(\frac{1}{n}\right)^p = n^{1-p}.$$

So $\hat{\mu}(P_1, P_2) \leq n^{1-p}$ for each n , and $\hat{\mu}(P_1, P_2) = 0$. It is easy to see that there is no optimal b in $B(P_1 - P_2)$.

This example shows that $B(\delta_0 - \delta_1)$ contains many measures other than $\delta_0 \otimes \delta_1$. This was one of the difficulties in [6].

§ 3. Application: explicit representations for a class of probability metrics

Throughout this section, we take $S = \mathbf{R}$, $d(x, y) = |x - y|$ and define $c: \mathbf{R} \times \mathbf{R} \rightarrow [0, +\infty)$ by

$$c(x, y) = |x - y| \max(h(|x - a|), h(|y - a|)),$$

where a is a fixed point of \mathbf{R} and $h: [0, +\infty) \rightarrow [0, +\infty)$ is a continuous non-decreasing function such that $h(x) > 0$ for $x > 0$. Define $\lambda: \mathbf{R} \rightarrow [0, +\infty)$ by

$$\lambda(x) = 2|x|h(|x-a|).$$

It is not hard to verify that c and λ satisfy the conditions (C1)–(C5) specified in §2. As in §2, the normed space $(L_0, \|\cdot\|_c)$ and the set M , comprising all finite signed measures m on \mathbf{R} such that

$$m(S) = 0 \quad \text{and} \quad \int \lambda d|m| < +\infty$$

are to be investigated.

We consider random variables X and Y in $\mathcal{X} = \mathcal{X}(\mathbf{R})$ with $E(\lambda(X)) + E(\lambda(Y)) < \infty$. Then $m = \mathcal{L}(X) - \mathcal{L}(Y)$ is an element of M , and Theorem 2.6 implies that

$$\begin{aligned} \hat{\mu}_c(X, Y) &= \inf\{\alpha E(c(X', Y')): X', Y' \in \mathcal{X}, \alpha > 0, \alpha(\mathcal{L}(X') - \mathcal{L}(Y')) = m\} \\ &= \sup\left\{\left|\int f dm\right|: |f(x) - f(y)| \leq c(x, y), \text{ all } x, y \in \mathbf{R}\right\}. \end{aligned}$$

An explicit representation is given in

3.1. THEOREM. *Suppose $X, Y \in \mathcal{X}$ with $E(\lambda(X)) + E(\lambda(Y)) < \infty$. Then*

$$\hat{\mu}_c(X, Y) = \int_{-\infty}^{\infty} h(|x-a|) |F_X(x) - F_Y(x)| dx.$$

Proof. We begin by proving the theorem in the special case where X and Y are bounded. Suppose that $|X| \leq N$ and $|Y| \leq N$ for some N . Application of Theorem 2.6 with $S = S_N = [-N, N]$ yields

$$\hat{\mu}_c(X, Y) = \sup\left\{\left|\int f dm\right|: f: S_N \rightarrow \mathbf{R}, f(x) - f(y) \leq c(x, y), \text{ all } x, y \in S_N\right\},$$

where $m = \mathcal{L}(X) - \mathcal{L}(Y)$.

Now given disjoint intervals $(s_i, t_i) \subseteq [-N, N]$ and $f(x) - f(y) \leq c(x, y)$ as above, we have

$$\begin{aligned} \sum |f(s_i) - f(t_i)| &\leq \sum c(s_i, t_i) \\ &= \sum |s_i - t_i| \max(h(|s_i - a|), h(|t_i - a|)) \leq h(N + |a|) \sum |s_i - t_i|. \end{aligned}$$

It follows that such an f is absolutely continuous on $[-N, N]$. Thus, f is differentiable a.e. on $[-N, N]$, and $|f'(x)| \leq h(|x-a|)$ wherever f' exists. So

$$\begin{aligned} \hat{\mu}_c(X, Y) &\leq \sup\left\{\left|\int f dm\right|: f: S_N \rightarrow \mathbf{R} \text{ differentiable a.e. with } |f'(x)| \leq h(|x-a|)\right\} \\ &= \sup\left\{\left|\int_{-\infty}^{\infty} (F_X(x) - F_Y(x)) f'(x) dx\right|: f: S_N \rightarrow \mathbf{R}, |f'(x)| \leq h(|x-a|) \text{ a.e.}\right\} \\ &\leq \int_{-\infty}^{\infty} h(|x-a|) |F_X(x) - F_Y(x)| dx, \end{aligned}$$

using integration by parts.

On the other hand, if f is absolutely continuous with $|f'(x)| \leq h(|x-a|)$ a.e., then

$$|f(x) - f(y)| = \left| \int_x^y f'(t) dt \right| \leq |x-y| \max(h(|x-a|), h(|y-a|)) = c(x, y).$$

Define $f_*: \mathbf{R} \rightarrow \mathbf{R}$ by

$$f_*(x) = \int_0^x h(|t-a|) \operatorname{sgn}(F_X(t) - F_Y(t)) dt.$$

Then

$$f'_*(x) = h(|x-a|) \operatorname{sgn}(F_X(x) - F_Y(x)) \quad \text{a.e.},$$

and

$$\begin{aligned} \hat{\mu}_c(X, Y) &\geq \sup \{ \left| \int f dm \right| : |f'(x)| \leq h(|x-a|) \text{ a.e.} \} \\ &= \sup \{ \left| \int (F_X(x) - F_Y(x)) f'(x) dx \right| : |f'(x)| \leq h(|x-a|) \text{ a.e.} \} \\ &\geq \left| \int (F_X(x) - F_Y(x)) f'_*(x) dx \right| = \int h(|x-a|) |F_X(x) - F_Y(x)| dx. \end{aligned}$$

We have shown that whenever X and Y are bounded random variables

$$\hat{\mu}_c(X, Y) = \int h(|x-a|) |F_X(x) - F_Y(x)| dx.$$

Now define $H: \mathbf{R} \rightarrow \mathbf{R}$ by

$$H(t) = \int_0^t h(|x-a|) dx.$$

Note that H is an odd function and that for $t \geq 0$,

$$\begin{aligned} H(t) &= \int_{-a}^0 h(|x|) dx + \int_0^{t-a} h(|x|) dx \\ &\leq h(|a|)|a| + |t-a|h(|t-a|) = \text{constant} + \lambda(t)/2, \end{aligned}$$

so that $E(\lambda(X)) + E(\lambda(Y)) < \infty$ implies that $E|H(X)| + E|H(Y)| < \infty$. Under this assumption, we have for each $t_0 > 0$

$$\int_{t_0}^{\infty} H(t) dF_X(t) \geq \Pr(X > t_0) H(t_0) = (1 - F_X(t_0)) H(t_0)$$

and

$$\int_{-\infty}^{-t_0} H(t) dF_X(t) \geq \Pr(X \leq -t_0) H(-t_0).$$

Taking $t_0 \rightarrow +\infty$ we obtain

$$\lim_{t_0 \rightarrow \infty} (1 - F_X(t_0)) H(t_0) = 0, \quad \lim_{t_0 \rightarrow -\infty} F_X(t_0) H(t_0) = 0.$$

Then define $\bar{F}(t) = 1 - F_X(t)$ and apply integration by parts

$$E|H(X)| = \int_0^{\infty} H(t) dF_X(t) + \int_{-\infty}^0 H(-t) dF_X(t) = I_1 + I_2,$$

where

$$\begin{aligned} I_1 &= - \int_0^{\infty} H(t) d\bar{F}(t) = -\bar{F}(t)H(t)\Big|_{t=0}^{+\infty} + \int_0^{\infty} \bar{F}(t)h(|t-a|)dt \\ &= \int_0^{\infty} \bar{F}(t)h(|t-a|)dt, \\ I_2 &= - \int_{-\infty}^0 H(t) dF_X(t) = -F_X(t)H(t)\Big|_{t=-\infty}^0 + \int_{-\infty}^0 F_X(t)h(|t-a|)dt \\ &= \int_{-\infty}^0 F_X(t)h(|t-a|)dt, \end{aligned}$$

whence

$$E|H(X)| = \int_0^{\infty} h(|x-a|)(1 - F_X(x))dx + \int_{-\infty}^0 h(|x-a|)F_X(x)dx.$$

An analogous equality holds for the variable Y . These imply that

$$\int_{-\infty}^{\infty} h(|x-a|)|F_X(x) - F_Y(x)|dx < \infty.$$

For $n \geq 1$, define random variables X_n, Y_n by

$$X_n = \begin{cases} n & \text{if } X > n, \\ X & \text{if } -n \leq X \leq n, \\ -n & \text{if } X < -n, \end{cases} \quad Y_n = \begin{cases} n & \text{if } Y > n, \\ Y & \text{if } -n \leq Y \leq n, \\ -n & \text{if } Y < -n. \end{cases}$$

Then for $X_n \rightarrow X, Y_n \rightarrow Y$ in law, and for $n \geq |a|$,

$$\begin{aligned} \hat{\mu}_c(X_n, X) &\leq E(c(X_n, X)) = E(|X_n - X| \max(h(|X_n - a|), h(|X - a|))) \\ &= E(|X_n - X| \cdot h(|X - a|)) \quad (|X_n - a| \leq |X - a|) \\ &\leq E(|X| I\{|X| \geq n\} h(|X - a|)), \end{aligned}$$

which tends to 0 as $n \rightarrow \infty$ ($E(\lambda(X)) < \infty$). Similarly, $\hat{\mu}_c(Y_n, Y) \rightarrow 0$. Then $\hat{\mu}_c(X_n, Y_n) \rightarrow \hat{\mu}_c(X, Y)$ as $n \rightarrow \infty$. Also, we have

$$|F_{X_n}(x) - F_{Y_n}(x)| = \begin{cases} |F_X(x) - F_Y(x)| & \text{for } -n \leq x < n, \\ 0 & \text{otherwise.} \end{cases}$$

Applying dominated convergence, we see that

$$\int h(|x-a|)|F_{X_n}(x) - F_{Y_n}(x)|dx \rightarrow \int h(|x-a|)|F_X(x) - F_Y(x)|dx.$$



Combining this with $\hat{\mu}_c(X_n, Y_n) \rightarrow \hat{\mu}_c(X, Y)$ and the result for bounded random variables yields

$$\hat{\mu}_c(X, Y) = \int_{-\infty}^{\infty} h(|x-a|) |F_X(x) - F_Y(x)| dx. \quad \blacksquare$$

For $h(x) = 1$, this yields a well-known formula present in [10] and [25]. We also note the following formulation, which is not hard to derive from the strict monotonicity of H :

3.2. COROLLARY. *Suppose $X, Y \in \mathcal{X}$ with $E(\lambda(X)) + E(\lambda(Y)) < \infty$ and put $P = \mathcal{L}(X)$, $Q = \mathcal{L}(Y)$. Then*

$$\hat{\mu}_c(P, Q) = \int_{-\infty}^{\infty} |F_{H(X)}(x) - F_{H(Y)}(x)| dx.$$

For $h(x) \equiv 1$, we see that $H(t) = t$ and that $\hat{\mu}_c$ gives the L^1 distance between X and Y .

3.3. COROLLARY. *In this context, $\hat{\mu}_c(P_1, P_2)$ defines a metric on $\mathcal{P}_\lambda(\mathbf{R})$.*

§4. Topology of the Kantorovich–Rubinstein norm

We continue under the conditions (C1)–(C5) of §2 with respect to the functions c and λ . In addition, we consider two further assumptions as follow.

(C6) $\alpha_1 = \sup\{d(x, y) \wedge 1/c(x, y): x \neq y\}$ is finite.

(C7) $\alpha_2 = \sup\{|\lambda(x) - \lambda(y)|/c(x, y): x \neq y\}$ is finite.

These conditions will aid us in our attempts to study the topology on the space $\mathcal{P}_\lambda(S)$ induced by the metric $\hat{\mu}_c(P_1, P_2)$. First, we develop a preliminary result.

Given a real-valued function $f: S \rightarrow \mathbf{R}$, define

$$\|f\|_x = \sup\{|f(x)|: x \in S\}, \quad \|f\|_L = \sup\{|f(x) - f(y)|/d(x, y): x \neq y\},$$

$$\|f\|_{BL} = \|f\|_\infty + \|f\|_L.$$

For P, Q in $\mathcal{P}(S)$, define

$$\beta(P, Q) = \sup\{|\int f d(P-Q)|: \|f\|_{BL} \leq 1\},$$

$$\beta_0(P, Q) = \sup\{|\int f d(P-Q)|: |f(x) - f(y)| \leq d(x, y) \wedge 1\}.$$

It is shown in [6; Chap. 8] that β is a metric on $\mathcal{P}(S)$ which metrizes weak convergence. An elementary argument establishes

4.1. LEMMA. *Given P, Q in $\mathcal{P}(S)$, we have*

$$\beta(P, Q)/2 \leq \beta_0(P, Q) \leq 2\beta(P, Q).$$

Thus, β_0 also metrizes weak convergence in $\mathcal{P}(S)$.

4.2. THEOREM. Assume (C1)–(C7) and that P, P_n are probabilities in $\mathcal{P}_\lambda(S)$ for $n \geq 1$. The following are equivalent:

- (A) $\hat{\mu}_c(P_n, P) \rightarrow 0$;
- (B) $P_n \rightarrow P$ weakly, and $\int \lambda d(P_n - P) \rightarrow 0$ as $n \rightarrow \infty$.

Proof. B \Rightarrow A: From Kantorovich-Rubinstein duality result (Theorem 2.6), we have, for P, Q in $\mathcal{P}_\lambda(S)$,

$$\hat{\mu}_c(P, Q) = \sup \left\{ \left| \int f d(P - Q) \right| : f \in \mathcal{F}_c \right\},$$

where

$$\mathcal{F}_c = \{ f \in L : \|f\|_c \leq 1 \text{ and } f(a) = 0 \},$$

and a is a fixed point of S . Define $G(x) = \lambda(x) + \lambda(a)$.

CLAIM 1. $\sup \{ |f(x)| : f \in \mathcal{F}_c \} \leq G(x)$.

CLAIM 2. $\limsup_{y \rightarrow x} \{ |f(y) - f(x)| : f \in \mathcal{F}_c \} = 0$ for each $x \in S$.

CLAIM 3. The functions $f \in \mathcal{F}_c$ and G are continuous.

CLAIM 4. The family \mathcal{F}_c is equicontinuous.

Claim 1 follows from (C3) of §2, whilst condition (C5) implies Claims 2 and 4. Lastly, note that $|\lambda(x) - \lambda(y)| \leq \alpha_2 c(x, y)$ (condition (C7)) in conjunction with (C5) yields continuity of λ and hence of G .

We now apply a theorem of Ranga Rao [20; p. 663] to the effect that if $P_n \rightarrow P$ weakly and $\int G d(P_n - P) \rightarrow 0$ as $n \rightarrow \infty$, then

$$\limsup_{n \rightarrow \infty} \left\{ \left| \int f d(P_n - P) \right| : f \in \mathcal{F}_c \right\} = 0.$$

By Theorem 2.6, this is the statement

$$\lim_{n \rightarrow \infty} \mu_c(P_n, P) = 0.$$

Thus B \Rightarrow A.

A \Rightarrow B: Using condition (C6) and Theorem 2.6, we obtain, for $P, Q \in \mathcal{P}_\lambda(S)$,

$$\beta_0(P, Q) \leq \sup \left\{ \left| \int f d(P - Q) \right| : |f(x) - f(y)| \leq \alpha_1 \cdot c(x, y) \right\} \leq \alpha_1 \hat{\mu}_c(P, Q).$$

Thus, $\hat{\mu}_c(P_n, P) \rightarrow 0$ implies $P_n \rightarrow P$ weakly as $n \rightarrow \infty$.

Also, if $\hat{\mu}_c(P_n, P) \rightarrow 0$, we may choose measures $b_n \in B(P_n - P)$ such that

$$\int c(x, y) db_n(x, y) \rightarrow 0$$

as $n \rightarrow \infty$. However

$$\begin{aligned} \left| \int \lambda d(P_n - P) \right| &= \left| \int (\lambda(x) - \lambda(y)) db_n(x, y) \right| \\ &\leq \int |\lambda(x) - \lambda(y)| db_n(x, y) \leq \alpha_2 \int c(x, y) db_n(x, y) \end{aligned} \quad (C7)$$

which tends to zero, as desired. ■

4.3. COROLLARY. Under conditions (C6) and (C7), $\hat{\mu}_c(P_1, P_2)$ is actually a metric on $\mathcal{P}_\lambda(S)$.

We now direct attention to the notion of compactness in $\mathcal{P}_\lambda(S)$ for the metric $\hat{\mu}_c$. The main result is

4.4. THEOREM. Suppose (C1)–(C7) and that $\mathcal{A} \subseteq \mathcal{P}_\lambda(S)$. The following are equivalent:

(A) \mathcal{A} is relatively compact for $\hat{\mu}_c$ (i.e. each sequence P_n in \mathcal{A} has a subsequence converging to an element of $\mathcal{P}_\lambda(S)$).

(B) \mathcal{A} is weakly compact in $\mathcal{P}(S)$, and

$$\sup \{ \int \lambda(x) I(d(x, a) > N) dP(x) : P \in \mathcal{A} \} \rightarrow 0$$

as $N \rightarrow \infty$. (Here, $a \in S$ is fixed but arbitrary.)

Proof. A \Rightarrow B: From Theorem 4.2, $\hat{\mu}_c(P_n, P) \rightarrow 0$ implies $P_n \rightarrow P$ weakly as $n \rightarrow \infty$. It follows that \mathcal{A} is weakly compact in $\mathcal{P}(S)$. Now suppose that the supremum in (B) does not go to zero. Then there is a sequence $N_1 < N_2 < \dots$ of positive real numbers with $N_k \rightarrow \infty$ such that $P\{x: d(x, a) = N_k\} = 0$ for each k and such that

$$\sup \{ \int \lambda(x) I\{x: d(x, a) > N_k\} dP(x) : P \in \mathcal{A} \} \geq \delta$$

for some $\delta > 0$ and all N_k . For each k , choose $P_k \in \mathcal{A}$ such that

$$\int \lambda(x) I\{x: d(x, a) > N_k\} dP_k(x) \geq \delta.$$

By hypothesis, there is a subsequence of the P_k converging with respect to $\hat{\mu}_c$ to some $P \in \mathcal{P}_\lambda(S)$. In order not to confuse the notation, we shall assume that $\hat{\mu}_c(P_k, P) \rightarrow 0$ as $k \rightarrow \infty$. By Theorem 4.2,

$$\int \lambda d(P_k - P) \rightarrow 0$$

as $k \rightarrow \infty$. Also, because $\int \lambda dP < \infty$, we have

$$\int \lambda(x) I\{x: d(x, a) > N_k\} dP(x) \rightarrow 0$$

as $k \rightarrow \infty$. Now for each $k \geq 1$, define $\lambda_k: S \rightarrow [0, \infty)$ by

$$\lambda_k(x) = \lambda(x) I\{x: d(x, a) > N_k\}.$$

Each λ_k is bounded and continuous off a set of P -measure zero. Since $P_k \rightarrow P$ weakly, we have [2; Theorem 5.1]

$$\int \lambda_k d(P_k - P) \rightarrow 0$$

as $k \rightarrow \infty$. Thus

$$\begin{aligned} \int \lambda(x) I\{x: d(x, a) > N_k\} dP_k(x) \\ \leq \left| \int \lambda_k d(P_k - P) \right| + \int \lambda(x) I\{x: d(x, a) > N_k\} dP(x), \end{aligned}$$

which tends to zero as $k \rightarrow \infty$: a contradiction which establishes that A \Rightarrow B.

B \Rightarrow **A: Given any sequence P_n in \mathcal{A} , use weak compactness to extract a convergent subsequence $P_{n(k)}$ converging weakly to some $P \in \mathcal{P}(S)$. Let $0 < N_1 < N_2 < \dots$ be a sequence of reals tending to infinity such that $P\{x: d(x, a) = N_k\} = 0$. Then from our hypotheses in **(B)**,**

$$\begin{aligned} \int \lambda(x) I\{x: d(x, a) \leq N_k\} dP(x) &= \lim_{k \rightarrow \infty} \int \lambda(x) I\{x: d(x, a) \leq N_k\} dP_{n(k)}(x) \\ &\leq \liminf_{k \rightarrow \infty} \int \lambda(x) dP_{n(k)}(x) \\ &\leq \sup\left\{ \int \lambda(x) I\{x: d(x, a) \geq \varepsilon\} dP(x): P \in \mathcal{A} \right\} \\ &\quad + \sup\{\lambda(x): d(x, a) \leq \varepsilon\}, \end{aligned}$$

where $\varepsilon > 0$ is such that

$$\sup\left\{ \int \lambda(x) I\{x: d(x, a) \geq \varepsilon\} dP(x): P \in \mathcal{A} \right\}$$

is finite. (Note that λ maps bounded sets to bounded sets (C4), so that

$$\sup\{\lambda(x): d(x, a) \leq \varepsilon\}$$

is finite.) Letting $k \rightarrow \infty$ shows that $P \in \mathcal{P}_\lambda(S)$.

Continuing this train of thought, we have

$$\begin{aligned} \left| \int \lambda d(P_{n(k)} - P) \right| &\leq \left| \int \lambda(x) I\{x: d(x, a) \leq N_k\} d(P_{n(k)} - P)(x) \right| \\ &\quad + \int \lambda(x) I\{x: d(x, a) > N_k\} d(P_{n(k)} + P)(x). \end{aligned}$$

Given $\delta > 0$, choose k_0 such that for $k \geq k_0$ we have

$$\begin{aligned} \int \lambda(x) I\{x: d(x, a) > N_k\} dP(x) &< \delta/4, \\ \sup\left\{ \int \lambda(x) I\{x: d(x, a) > N_k\} dQ(x): Q \in \mathcal{A} \right\} &< \delta/4, \\ \left| \int \lambda(x) I\{x: d(x, a) \leq N_k\} d(P_{n(k)} - P)(x) \right| &< \delta/2. \end{aligned}$$

Then $\left| \int \lambda d(P_{n(k)} - P) \right| < \delta$ for all $k \geq k_0$. It follows from Theorem 4.2 that $\hat{\mu}_c(P_{n(k)}, P) \rightarrow 0$ as $k \rightarrow \infty$. ■

§5. Dual representation for the Wasserstein functional

Let (S, d) be a separable metric space and let $c: S \times S \rightarrow [0, \infty)$ be a measurable function. Let P_1 and P_2 be probabilities on S and denote by $V(P_1, P_2)$ the set of all probabilities on $S \times S$ with marginals P_1 and P_2 , i.e. all $b \in \mathcal{P}(S \times S)$ such that

$$b(A \times S) = P_1(A) \quad \text{and} \quad b(S \times A) = P_2(A)$$

for each Borel $A \subseteq S$. Define

$$\hat{\mu}_c(P_1, P_2) = \inf\left\{ \int c(x, y) db(x, y): b \in V(P_1, P_2) \right\}.$$

Using notation of §2, we see that $V(P_1, P_2) \subseteq B(P_1 - P_2)$, so that

$$\hat{\mu}_c(P_1, P_2) \leq \hat{\mu}_c(P_1, P_2)$$

for any P_1, P_2 in $\mathcal{P}(S)$.

Now define $G(S)$ as the set of all pairs (f, g) of Borel measurable functions $f: S \rightarrow \mathbf{R}$ and $g: S \rightarrow \mathbf{R}$ such that

$$f(x) + g(y) \leq c(x, y)$$

for all $x, y \in S$. Let $G_B(S)$ [resp. $G_C(S)$] be the set of all pairs $(f, g) \in G(S)$ with f and g bounded [resp. continuous].

5.1. LEMMA. *Let P_1 and P_2 be probabilities on S . Then*

$$\hat{\mu}_c(P_1, P_2) \geq \sup \left\{ \int f dP_1 + \int g dP_2 : (f, g) \in G_B(S) \right\}.$$

Proof. Given $Q \in V(P_1, P_2)$ and $(f, g) \in G_B(S)$, we have

$$\int c dQ \geq \int (f(x) + g(y)) dQ(x, y) = \int f dP_1 + \int g dP_2.$$

The lemma follows. ■

The crucial question is whether the inequality in Lemma 5.1 can be replaced by equality. In view of this, we shall be concerned throughout the section with variations and extensions of the following basic duality result, which seems the strongest currently available.

5.2. THEOREM. *Let P_1, P_2 be tight probabilities on (S, d) . Then*

$$\hat{\mu}_c(P_1, P_2) = \sup \left\{ \int f dP_1 + \int g dP_2 : (f, g) \in G(S), f \in L^1(P_1), g \in L^1(P_2) \right\}.$$

Also, there exists an optimal pair $(f, g) \in G(S)$, $f \in L^1(P_1)$, $g \in L^1(P_2)$ for which this supremum is attained.

If the function c is lower semi-continuous, then there is some optimal $Q \in V(P_1, P_2)$ such that

$$\hat{\mu}_c(P_1, P_2) = \int c(x, y) dQ(x, y).$$

If the function c is continuous we have

$$\hat{\mu}_c(P_1, P_2) = \sup \left\{ \int f dP_1 + \int g dP_2 : (f, g) \in G_C(S), f \in L^1(P_1), g \in L^1(P_2) \right\}.$$

Indication. This result is the central theorem (2.14) in Kellerer [11]. The existence of the optimal pair (f, g) and the optimal Q are (2.21) and (2.19) in the same reference. The last sentence is (1.33). Compare also Levin [16] and Rachev [19]. ■

Put $G_{BC}(S) = G_C(S) \cap G_B(S)$.

5.3. COROLLARY. *Let $P_1, P_2 \in \mathcal{P}(S)$ be tight. Then*

$$\hat{\mu}_c(P_1, P_2) = \sup \left\{ \int f dP_1 + \int g dP_2 : (f, g) \in G_{BC}(S) \right\}.$$

Proof. Apply Theorem 5.2. Note that if $f \in L^1(P_1)$ and $g \in L^1(P_2)$ with $f(x) + g(y) \leq c(x, y)$, then $f_n(x) + g_n(y) \leq c(x, y)$, where

$$f_n(x) = \begin{cases} n & \text{if } f(x) > n, \\ f(x) & \text{if } n \geq f(x) \geq -n, \\ -n & \text{if } -n > f(x), \end{cases} \quad g_n(x) = \begin{cases} n & \text{if } g(x) > n, \\ g(x) & \text{if } n \geq g(x) \geq -n, \\ -n & \text{if } -n > g(x). \end{cases}$$

Then $f_n \rightarrow f$ and $g_n \rightarrow g$ as $n \rightarrow \infty$ in $L^1(P_1)$ and $L^1(P_2)$, respectively. ■

We now attempt to remove the tightness conditions on P_1 and P_2 . In analogy with §2, we introduce a measurable function $\lambda: S \rightarrow [0, \infty)$ and a metric D for $S \times S$ defined by

$$D((x, y), (x', y')) = d(x, x') + d(y, y').$$

We consider the conditions

(C3) $c(x, y) \leq \lambda(x) + \lambda(y)$ for all $x, y \in S$;

(C8) c is D -uniformly continuous on D -bounded subsets of $S \times S$.

Note that (C8) implies (C5) in §2. Define $\mathcal{P}_\lambda^0(S)$ as the set of all $P \in \mathcal{P}_\lambda(S)$ such that

(*) for some $a \in S$, we have $P\{x: d(x, a) \geq N\} \sup\{\lambda(x): d(x, a) < N\} \rightarrow 0$ as $N \rightarrow \infty$.

Note. Suppose that λ is “increasing” in the sense that for some $a \in S$, $d(x, a) \leq d(x', a)$ implies $\lambda(x) \leq \lambda(x')$. Then condition (*) follows from P -integrability of λ by a Chebyshev inequality argument; in this case, as for example when

$$c(x, y) = d^p(x, y) \quad \text{if } p \geq 1, \\ \lambda(x) = 2^{p-1} d^p(x, a) \quad \text{if some } a \in S,$$

we have $\mathcal{P}_\lambda^0(S) = \mathcal{P}_\lambda(S)$. Compare Example 2.9.

5.4. THEOREM. Suppose that c and λ satisfy conditions (C3) and (C8). Then

$$\hat{\mu}_c(P_1, P_2) = \sup\left\{\int f dP_1 + \int g dP_2: (f, g) \in G_{BC}(S)\right\}$$

for any P_1, P_2 in $\mathcal{P}_\lambda^0(S)$.

Proof. Let (\bar{S}, d) be the completion of (S, d) . We see that c extends uniquely to a non-negative function \bar{c} on $\bar{S} \times \bar{S}$ which is D -uniformly continuous on D -bounded subsets of $\bar{S} \times \bar{S}$. Take probability measures \bar{P}_1 and \bar{P}_2 on \bar{S} defined by

$$\bar{P}_1(B) = P_1(B \cap S), \quad \bar{P}_2(B) = P_2(B \cap S).$$

Then \bar{P}_1 and \bar{P}_2 are automatically tight. We apply Theorem 5.2 to produce some $Q \in V(\bar{P}_1, \bar{P}_2)$ such that

$$\hat{\mu}_c(\bar{P}_1, \bar{P}_2) = \int_{\bar{S} \times \bar{S}} \bar{c}(x, y) dQ(x, y) = \sup\left\{\int f d\bar{P}_1 + \int g d\bar{P}_2: (f, g) \in G_{BC}(\bar{S})\right\}.$$

Define now

$$\begin{aligned}\bar{B}_N &= \{x \in \bar{S}: d(x, a) < N\}, & \bar{L}_N &= \bar{S} - \bar{B}_N, \\ B_N &= \bar{B}_N \cap S, & L_N &= \bar{L}_N \cap S.\end{aligned}$$

for N positive.

Given $\varepsilon > 0$ arbitrary, choose N large so that

$$\begin{aligned}P_i(L_N) \sup\{\lambda(x): x \in B_N\} &< \varepsilon \quad \text{if } i = 1, 2, \\ \int_{L_N} \lambda dP_i &< \varepsilon \quad \text{if } i = 1, 2.\end{aligned}$$

Then for any $R \in V(P_1, P_2)$, one has

$$\begin{aligned}\int_{S \times S - B_N \times B_N} c(x, y) dR(x, y) &\leq \int_{L_N \times L_N} (\lambda(x) + \lambda(y)) dR(x, y) \\ &\quad + \int_{B_N \times L_N} (\lambda(x) + \lambda(y)) dR(x, y) \\ &\quad + \int_{L_N \times B_N} (\lambda(x) + \lambda(y)) dR(x, y) \\ &\leq \int_{L_N} \lambda(x) dP_1(x) + \int_{L_N} \lambda(y) dP_2(y) \\ &\quad + P_2(L_N) \sup\{\lambda(x): x \in B_N\} + \int_{L_N} \lambda(y) dP_2(y) \\ &\quad + \int_{L_N} \lambda(x) dP_1(x) + P_1(L_N) \sup\{\lambda(y): y \in B_N\} \\ &< 6\varepsilon.\end{aligned}$$

We now exploit uniform continuity of \bar{c} on $\bar{B}_N \times \bar{B}_N$. Choose $\delta > 0$ so that

$$|\bar{c}(x, y) - \bar{c}(x', y')| < \varepsilon$$

whenever $D((x, y), (x', y')) < \delta$ with x, y, x', y' in \bar{B}_N . Let A_1, A_2, \dots be a partition of B_N into Borel sets of diameter less than δ . Put $A_0 = L_N$ and choose $x_n \in A_n$ for $n = 1, 2, 3, \dots$. Then there is a partition of \bar{B}_N into Borel sets $\bar{A}_1, \bar{A}_2, \bar{A}_3, \dots$ of diameter $< \delta$ such that $A_n = \bar{A}_n \cap S$ for $n \geq 1$. Put $\bar{A}_0 = \bar{L}_N$. For each $m \geq 0$, define finite measures P_1^m and P_2^m on S by

$$P_1^m(B) = P_1(B \cap A_m), \quad P_2^m(B) = P_2(B \cap A_m).$$

Define $Q_{mn} = c_{mn}(P_1^m \otimes P_2^n)$ on $S \times S$, where each c_{mn} is a non-negative real constant chosen so that $Q_{mn}(A_m \times A_n) = Q(\bar{A}_m \times \bar{A}_n)$. Put

$$Q_\varepsilon = \sum_{mn} Q_{mn},$$

so that

$$\begin{aligned} Q_\varepsilon(B \times S) &= \sum c_{mn} P_1^m(B) P_2^n(S) = \sum c_{mn} P_1(B \cap A_m) P_2(A_n) \\ &= \sum' \frac{Q(\bar{A}_m \times \bar{A}_n)}{P_1(A_m) P_2(A_n)} P_1(B \cap A_m) P_2(A_n), \end{aligned}$$

where \sum' indicates summation over all m, n such that $P_1(A_m) P_2(A_n) > 0$. Note that if $P_1(A_m) > 0$, we have

$$\sum_n \frac{Q(\bar{A}_m \times \bar{A}_n)}{P_1(A_m)} = \frac{Q(\bar{A}_m \times \bar{S})}{P_1(A_m)} = \frac{\bar{P}_1(\bar{A}_m)}{P_1(A_m)} = 1,$$

so that

$$Q_\varepsilon(B \times S) = \sum_m P_1(B \cap A_m) = P_1(B).$$

This, together with an analogous calculation for $Q_\varepsilon(S \times B)$, shows that $Q_\varepsilon \in V(P_1, P_2)$. Then

$$\begin{aligned} \int c(x, y) dQ_\varepsilon(x, y) &= \int_{S \times S - B_N \times B_N} c(x, y) dQ_\varepsilon(x, y) + \sum_{m, n \geq 1} \int_{A_m \times A_n} c(x, y) dQ_{mn}(x, y) \\ &\leq 6\varepsilon + \sum_{A_m \times A_n} \int c dQ_{mn} \quad (\text{sums over } m, n \geq 1) \\ &\leq 6\varepsilon + \sum_{A_m \times A_n} \int c(x_m, x_n) dQ_{mn} \\ &\quad + \sum_{A_m \times A_n} \int |c(x_m, x_n) - c(x, y)| dQ_{mn}(x, y) \\ &\leq 6\varepsilon + \sum_{A_m \times A_n} \int c(x_m, x_n) dQ + \varepsilon \\ &\leq \sum_{\bar{A}_m \times \bar{A}_n} \int \bar{c}(x, y) dQ(x, y) \\ &\quad + \sum_{\bar{A}_m \times \bar{A}_n} \int |\bar{c}(x, y) - \bar{c}(x_m, x_n)| dQ(x, y) + 7\varepsilon \\ &\leq \int \bar{c}(x, y) dQ(x, y) + \varepsilon + 7\varepsilon. \end{aligned}$$

Letting $\varepsilon \rightarrow 0$, we obtain

$$\begin{aligned} \hat{\mu}_c(P_1, P_2) &\leq \hat{\mu}_c(\bar{P}_1, \bar{P}_2) = \sup \{ \int f d\bar{P}_1 + \int g d\bar{P}_2 : (f, g) \in G_{BC}(\bar{S}) \} \\ &\leq \sup \{ \int f dP_1 + \int g dP_2 : (f, g) \in G_{BC}(S) \}. \end{aligned}$$

Application of Lemma 5.1 yields the desired equality. ■

We now attempt to sharpen this duality theorem by restricting $G_{BC}(S)$ to functions satisfying a Lipschitz-like condition. Define $G_{BL}(S)$ as the set of pairs $(f, g) \in G_B(S)$ such that

$$|f(x) - f(y)| \leq \sup \{ |c(x, u) - c(y, u)| : u \in S \},$$

$$|g(x) - g(y)| \leq \sup\{|c(u, x) - c(u, y)| : u \in S\}$$

for all $x, y \in S$ and such that f, g are upper semi-continuous.

Note. In case $c(x, y)$ is a metric, these reduce to the Lipschitz conditions $|f(x) - f(y)| \leq c(x, y)$ and $|g(x) - g(y)| \leq c(x, y)$.

Note. Suppose that $S = \mathbf{R}$ and that $c(x, y) = |x - y|^p$ for $p > 1$. Then the suprema above are infinite, so that $G_{BL}(S) = G_{BC}(S)$ in this case.

The following is our strongest duality result for the functional $\hat{\mu}$.

5.5. THEOREM. Suppose that c and λ satisfy conditions (C3) and (C8). Then

$$\hat{\mu}_c(P_1, P_2) = \sup\{\int f dP_1 + \int g dP_2 : (f, g) \in G_{BL}(S)\}$$

for any P_1, P_2 in $\mathcal{P}_\lambda^0(S)$.

Proof. Suppose now that $(f, g) \in G_{BC}(S)$. Define

$$f^*(x) = \inf\{c(x, u) - g(u) : u \in S\}, \quad g^*(x) = \inf\{c(u, x) - f^*(u) : u \in S\}.$$

Since c is continuous, it follows that f^* and g^* are upper semi-continuous, hence Borel measurable. Also,

$$f^*(x) + g^*(y) \leq c(x, y)$$

for all $x, y \in S$. Furthermore,

$$\begin{aligned} f^*(x) - f^*(y) &= \inf_u\{c(x, u) - g(u)\} - \inf_v\{c(y, v) - g(v)\} \\ &= \inf_u\{c(x, u) - g(u)\} + \sup_v\{g(v) - c(y, v)\} \\ &= \sup_v\{\inf_u\{c(x, u) - g(u)\} + g(v) - c(y, v)\} \\ &\leq \sup_v\{c(x, v) - g(v) + g(v) - c(y, v)\} \\ &\leq \sup\{|c(x, v) - c(y, v)| : v \in S\} \end{aligned}$$

for all $x, y \in S$. A similar argument proves that

$$g^*(x) - g^*(y) \leq \sup\{|c(u, x) - c(u, y)| : u \in S\}.$$

Thus $(f^*, g^*) \in G_{BL}(S)$. Given $x \in S$, we have $f(x) \leq c(x, y) - g(y)$ for all $y \in S$. Thus $f(x) \leq f^*(x)$. Also,

$$\begin{aligned} g^*(x) &= \inf_u\{c(u, x) - \inf_v\{c(u, v) - g(v)\}\} \\ &\geq \inf_u\{c(u, x) - c(u, x) + g(x)\} = g(x). \end{aligned}$$

Hence

$$\int f dP_1 + \int g dP_2 \leq \int f^* dP_1 + \int g^* dP_2.$$

It follows that

$$\begin{aligned} \hat{\mu}_c(P_1, P_2) &= \sup\{\int f dP_1 + \int g dP_2 : (f, g) \in G_B(S)\} \\ &\geq \sup\{\int f dP_1 + \int g dP_2 : (f, g) \in G_{BL}(S)\} \\ &\geq \sup\{\int f dP_1 + \int g dP_2 : (f, g) \in G_{BC}(S)\} = \hat{\mu}_c(P_1, P_2). \quad \blacksquare \end{aligned}$$

5.6. EXAMPLE. Let (S, d) be a separable metric space and suppose that $H: [0, \infty) \rightarrow [0, \infty)$ is a continuous, non-decreasing, convex function. Put

$$c(x, y) = H(d(x, y)), \quad \lambda(x) = H(2d(x, a)),$$

where $a \in S$ is some arbitrarily chosen point. Then conditions (C3) and (C8) are satisfied, and condition (*) holds for every $P \in \mathcal{P}(S)$ for which λ is P -integrable. The case of $H(t) = t^p$ ($p \geq 1$) is of particular interest. See e.g. [6; p. 20.1], [18], [19], [9]. Compare Example 2.9.

§ 6. Comparison of Wasserstein functional and Kantorovich–Rubinstein norm; completeness

We begin this section by proving the following form of an unpublished result of Neveu and Dudley.

6.1. THEOREM. *Suppose that conditions (C1)–(C8) hold for the functions c and λ . Then*

$$(**) \quad \hat{\mu}_c(P_1, P_2) = \hat{\mu}_\lambda(P_1, P_2)$$

for all P_1, P_2 in $\mathcal{P}_\lambda(S)$ if and only if c is a metric (i.e. satisfies the triangle inequality).

PROOF. Suppose (**) holds and put $P_1 = \delta_x$ and $P_2 = \delta_y$ for $x, y \in S$. Then $V(P_1, P_2)$ contains only $P_1 \otimes P_2 = \delta_{(x,y)}$ and

$$\begin{aligned} \hat{\mu}_c(P_1, P_2) = c(x, y) &= \hat{\mu}_c(P_1, P_2) = \sup \{ \int f d(P_1 - P_2) : \|f\|_c \leq 1 \} \\ &= \sup \{ |f(x) - f(y)| : \|f\|_c \leq 1 \}. \end{aligned}$$

From this it is clear that $c(x, y)$ satisfies the triangle inequality.

Now suppose that $c(x, y)$ is a metric and that $(f, g) \in G_B(S)$. Define

$$h(x) = \inf \{ c(x, y) - g(y) : y \in S \}.$$

As the infimum of a family of continuous functions, h is upper semi-continuous. For each $x \in S$, we have $f(x) \leq h(x) \leq -g(x)$. Then

$$\begin{aligned} h(x) - h(x') &= \inf_u (c(x, u) - g(u)) - \inf_v (c(x', v) - g(v)) \\ &= \inf_u (c(x, u) - g(u)) + \sup_v (g(v) - c(x', v)) \\ &= \sup_v (g(v) - c(x', v) + \inf_u (c(x, u) - g(u))) \\ &\leq \sup_v (g(v) - c(x', v) + c(x, v) - g(v)) \\ &= \sup_v (c(x, v) - c(x', v)) = c(x, x'), \end{aligned}$$

so that $\|h\|_c \leq 1$. Then for $P_1, P_2 \in \mathcal{P}_\lambda^0(S)$

$$\int f dP_1 + \int g dP_2 \leq \int h d(P_1 - P_2),$$

so that (according to the duality results 2.6 and 5.4) we have

$$\begin{aligned}\hat{\mu}_c(P_1, P_2) &= \sup\{\int f dP_1 + \int g dP_2: (f, g) \in G_B(S)\} \\ &\leq \sup\{\int h d(P_1, P_2): \|h\|_c \leq 1\} = \hat{\mu}_c(P_1, P_2).\end{aligned}$$

Thus $\hat{\mu}_c(P_1, P_2) = \hat{\mu}_c(P_1, P_2)$. ■

6.2. COROLLARY. *Let (S, d) be a separable metric space and $a \in S$. Then*

$$\hat{\mu}_d(P_1, P_2) = \hat{\mu}_d(P_1, P_2) = \sup\{\int f d(P_1 - P_2): \|f\|_L \leq 1\}$$

whenever

$$\int d(x, a) d(P_1 + P_2)(x) < \infty.$$

The supremum is attained for some optimal f_0 with $\|f_0\|_L = 1$.

If P_1 and P_2 are tight, there is some $b_0 \in V(P_1, P_2)$ such that

$$\hat{\mu}_d(P_1, P_2) = \int d(x, y) db_0(x, y).$$

Then $f_0(x) - f_0(y) = d(x, y)$ for b_0 -almost every (x, y) in $S \times S$.

Proof. Put $c(x, y) = d(x, y)$ and $\lambda(x) = d(x, a)$. Conditions (C1)–(C8) and (*) obtain, with $\mathcal{P}_\lambda^0(S) = \mathcal{P}_\lambda(S)$. Application of the theorem proves the first sentence. The second (existence of f_0) follows from Theorem 2.7.

For each $n \geq 1$, choose $b_n \in V(P_1, P_2)$ with

$$\int d(x, y) db_n(x, y) < \hat{\mu}_d(P_1, P_2) + 1/n.$$

If P_1 and P_2 are tight, the collection $V(P_1, P_2)$ is uniformly tight and hence compact. Hence the b_n have a subsequence $b_{n(k)}$ converging weakly to some $b_0 \in V(P_1, P_2)$. Then [2; Theorem 5.3]

$$\liminf_{k \rightarrow \infty} \int d(x, y) db_{n(k)}(x, y) \geq \int d(x, y) db_0(x, y).$$

It follows that

$$\hat{\mu}_d(P_1, P_2) = \int d(x, y) db_0(x, y),$$

i.e. that b_0 is optimal. Integrating both sides of

$$f_0(x) - f_0(y) \leq d(x, y)$$

with respect to b_0 yields

$$\int f_0 d(P_1 - P_2) \leq \int d(x, y) db_0(x, y).$$

However, we know that we have equality of these integrals. This forces

$$f_0(x) - f_0(y) = d(x, y)$$

b_0 -a.e. ■

Continuing in the context where $c(x, y) = d(x, y)$ and $\lambda(x) = d(x, a)$, we examine the question of completeness for $\mathcal{P}_\lambda(S)$ under the metric $\hat{\mu} = \hat{\mu}$. The following is a useful fact, which we shall employ.

6.3. LEMMA. *Suppose that $P_1, P_2 \in \mathcal{P}(S)$ with $\hat{\mu}(P_1, P_2) < \infty$. Then $P_1 \in \mathcal{P}_\lambda(S)$ if and only if $P_2 \in \mathcal{P}_\lambda(S)$.*

Proof. Choose $b \in V(P_1, P_2)$ such that $d(x, y)$ is b -integrable. Then

$$\int |d(x, a) - d(y, a)| db(x, y) \leq \int d(x, y) db(x, y) < \infty,$$

and $d(x, a) - d(y, a)$ is b -integrable. The statement that $P_1 \in \mathcal{P}_\lambda(S)$ means that $k(x, y) = d(x, a)$ is b -integrable. Then $d(y, a)$ is b -integrable, so that $P_2 \in \mathcal{P}_\lambda(S)$. ■

6.4. THEOREM. *The metric space $(\mathcal{P}_\lambda(S), \hat{\mu})$ is complete if and only if (S, d) is complete.*

Proof. The “only if” direction is clear: If x_n is a Cauchy sequence in (S, d) , put $P_n = \delta_{x_n}$. Then $\mu(P_n, P_m) = d(x_n, x_m)$, so that P_n is a Cauchy sequence in $\mathcal{P}_\lambda(S)$. If $\hat{\mu}(P_n, P) \rightarrow 0$ for some $P \in \mathcal{P}_\lambda(S)$, we have $P_n \rightarrow P$ weakly. Then $P = \delta_x$ for some $x \in S$, and $d(x_n, x) \rightarrow 0$.

Now suppose that (S, d) is complete and that $P_n \in \mathcal{P}_\lambda(S)$ is a Cauchy sequence for $\hat{\mu} = \hat{\mu}$. As in the proof of Theorem 4.2, we note that

$$\beta(P_n, P_m) \leq 2\beta_0(P_n, P_m) \leq 2\hat{\mu}(P_n, P_m).$$

Thus P_n is a Cauchy sequence for β , which is a complete metric on $\mathcal{P}(S)$. See [6; Chap. 8]. Thus $P_n \rightarrow P$ weakly as $n \rightarrow \infty$ for some $P \in \mathcal{P}(S)$. We shall demonstrate that $P \in \mathcal{P}_\lambda(S)$ and that $\hat{\mu}(P_n, P) \rightarrow 0$ as $n \rightarrow \infty$.

Using Corollary 6.2, choose optimal b_{nm} in $V(P_n, P_m)$ with

$$\hat{\mu}(P_n, P_m) = \int d(x, y) db_{nm}(x, y).$$

Since the P_n are uniformly tight on S , so also are the b_{nm} on $S \times S$. Fix n . There is a subsequential index $m(k)$ such that $b_{nm(k)} \rightarrow b_n$ weakly as $k \rightarrow \infty$ for some $b_n \in V(P_n, P)$. Then

$$\begin{aligned} \hat{\mu}(P_n, P) &\leq \int d(x, y) db_n(x, y) \leq \liminf_{k \rightarrow \infty} \int d(x, y) db_{nm(k)}(x, y) \\ &= \liminf_{k \rightarrow \infty} \mu(P_n, P_{m(k)}). \end{aligned}$$

Given $\varepsilon > 0$, choose N so that $\mu(P_n, P_m) < \varepsilon/2$ for $n, m \geq N$. Choose k so that $m(k) \geq N$. Then for each $n \geq N$ we have $\mu(P_n, P) < \varepsilon$. It follows that $\mu(P_n, P) \rightarrow 0$ as $n \rightarrow \infty$.

Application of Lemma 6.3 shows that $P \in \mathcal{P}_\lambda(S)$, concluding the argument. ■

§7. Convergence of empirical measures; results of Fortet–Mourier type

Suppose that c and λ satisfy conditions (C1)–(C7) of §2 and §4 and suppose $P \in \mathcal{P}_\lambda(S)$. Let X_1, \dots, X_n be i.i.d. random variables in $\mathcal{X}(S)$ defined on some probability space $(\Omega, \mathcal{A}; \Pr)$ with $\mathcal{L}(X_i) = P$. Then

$$P_n(B) = \frac{I(A)(X_1) + \dots + I(A)(X_n)}{n}$$

defines the n th empirical measure $P_n = P_n^\omega$ based on P .

7.1. THEOREM. *Suppose that c, λ, P, P_n are as above. Then $\hat{\mu}_c(P_n, P) \rightarrow 0$ a.s. as $n \rightarrow \infty$.*

Remark. Although we shall prove that $\{\omega: \hat{\mu}_c(P_n^\omega, P) \rightarrow 0\}$ is measurable and of unit probability, we are not claiming measurability for the mapping $\omega \rightarrow \hat{\mu}_c(P_n, P)$. In this regard, see Theorem 7.2 *infra*.

Proof. Following the technique of Varadarajan [6; Theorem 9.1], we note the existence of bounded continuous functions $f_k: S \rightarrow \mathbf{R}$ ($k \geq 1$) such that $Q_n \rightarrow Q$ weakly in $\mathcal{P}(S)$ if and only if

$$\int f_k dQ_n \rightarrow \int f_k dQ \quad \text{as } n \rightarrow \infty$$

for each $k \geq 1$. Put

$$A_k = \{\omega: \int f_k dP_n^\omega \rightarrow \int f_k dP \text{ as } n \rightarrow \infty\}$$

for $k \geq 1$ and

$$A_0 = \{\omega: \int \lambda dP_n^\omega \rightarrow \int \lambda dP \text{ as } n \rightarrow \infty\}.$$

Then A_0, A_1, \dots are measurable, and by Theorem 4.2, we have $\hat{\mu}_c(P_n^\omega, P) \rightarrow 0$ if and only if $\omega \in \bigcap A_k$. The strong law of large numbers implies that $\Pr(A_k) = 1$ for each $k \geq 0$, establishing the result. ■

We now consider the measurability of the map $\omega \rightarrow \hat{\mu}_c(P_n, P)$. For this, we shall insist that

$$\Omega = S \times S \times \dots, \quad \mathcal{A} = \mathcal{B}(S) \otimes \mathcal{B}(S) \otimes \dots, \quad \Pr = P \otimes P \otimes \dots,$$

with $X_n(\omega_1, \omega_2, \dots) = \omega_n$ projection to the n th factor; such an assumption seems usual for empirical measures.

7.2. THEOREM. *Let (S, d) be a Polish space (i.e. complete and separable). In the situation just described, $\hat{\mu}_c(P_n, P)$ is measurable for the completion of \Pr on Ω .*

Remark. The hypothesis that S be complete may be dropped in case that P is tight. The proof is similar.

Proof. We note that $\mathcal{M}(S \times S)$ is a Polish space under the topology of weak convergence [1; Theorem B] and that the mappings $b \rightarrow b(A)$ are Borel measurable on $\mathcal{M}(S \times S)$ for each fixed $A \in \mathcal{B}(S \times S)$. Let A_1, A_2, \dots be a countable base for the topology of S . Then for each n ,

$$B_n = \{(\omega, b): b(A_n \times S) - b(S \times A_n) = (P_n^\omega - P)(A)\}$$

is a Borel subset of $\Omega \times \mathcal{M}(S \times S)$. Put $B_\infty = B_1 \cap B_2 \cap \dots$. Then the section of B_∞ over $\omega \in \Omega$ is $B(P_n^\omega - P)$. Given $a \in \mathbf{R}$, put

$$B^a = \{b \in \mathcal{M}(S \times S): \int c(x, y)db(x, y) < a\},$$

a Borel subset of $\mathcal{M}(S \times S)$. Then

$$\{\omega: \hat{\mu}_c(P_n^\omega, P) < a\}$$

is the projection to Ω of $B_\infty \cap (\Omega \times B^a)$, hence an analytic set, hence completion-measurable for any probability measure, Pr in particular. (See [4; Chapter 8.4].) ■

Given any $f: S \rightarrow \mathbf{R}$, $t \geq 0$, $p \geq 1$ and some fixed point $a \in S$, we define the quantities

$$L(f, t) = \sup\{|f(x) - f(y)|/d(x, y): d(x, a) \leq t, d(y, a) \leq t, x \neq y\},$$

$$L_p(f) = \sup\{L(f, t)/\max(1, t^{p-1}): t > 0\}.$$

Put $\lambda(x) = d^p(x, a)$ and define $\mathcal{P}_p(S) = \mathcal{P}_\lambda(S)$. For P_1, P_2 in $\mathcal{P}_p(S)$, define

$$FM_p(P_1, P_2) = \sup\{\int f d(P_1 - P_2): L_p(f) \leq 1\}.$$

For $p = 1$, this quantity was studied by R. Fortet and E. Mourier in their well-known paper on empirical measures [8].

We hold $p \geq 1$ fixed throughout what follows. Define

$$c(x, y) = d(x, y)\max(1, d^{p-1}(x, a), d^{p-1}(y, a)), \quad \mu(x) = 2d(x, a) + 2d^p(x, a).$$

Then $\mathcal{P}_p(S) = \mathcal{P}_\mu(S)$. Verification of the following fact is routine and therefore omitted:

7.3. LEMMA. *The functions c and μ satisfy hypotheses (C1)–(C7) of §2 and §4.*

Now given $f: S \rightarrow \mathbf{R}$, define

$$\|f\|_c = \sup\{|f(x) - f(y)|/c(x, y): x \neq y\}.$$

Then

$$7.4. \text{ LEMMA. } \|f\|_c = L_p(f).$$

Proof. Given $x \neq y$, put $t_0 = t_0(x, y) = \max(d(x, a), d(y, a)) > 0$. Then

$$f(x) - f(y) \leq L(f, t_0)d(x, y) \leq L_p(f)\max(1, t_0^{p-1})d(x, y) = L_p(f)c(x, y),$$

If, additionally, there is a family of continuous projections $P_{G,k}: X \rightarrow G$ such that

$$\|P_{G,k}\|_{\varphi(i),i} \leq C_i \quad \text{for } i = 1, \dots, k,$$

then $(X, (\|\cdot\|_k)_{k \in \mathbb{N}})$ is called a *strongly locally \mathcal{L}_p -space*.

Remark. The above definition does not depend on the choice of a sequence of seminorms defining the topology of X .

The next lemma (irreversible in general for any p — Ex. 8.24) clarifies relations between these two introduced notions.

LEMMA 5.1. *Every (strongly) globally \mathcal{L}_p -space X is also a (strongly) locally \mathcal{L}_p -space.*

Proof. We define φ and ψ in such a way that for every 0-neighbourhood U in X there is a family $(a_i)_{i \in \varphi(U)}$ of positive numbers such that

$$T_j^{-1}(V) \subseteq U \quad \text{for every } j \in J,$$

where

$$V := \prod_{i \in \varphi(U)} a_i B_{l_p} \times \prod_{i \notin \varphi(U)} l_p$$

and if

$$W = \prod_{i \in I_0} B_{l_p} \times \prod_{i \notin I_0} l_p,$$

then

$$T_j(\psi(I_0)) \subseteq W.$$

Now, let $H \subseteq X$, $\dim H < \infty$ and $\ker \mathcal{U}_0 \cap H = \{0\}$. Then H is contained in X_j for suitably chosen $j \in J$. The natural projection

$$P: l_p^I \rightarrow \prod_{i \in \varphi(\mathcal{U}_0)} l_p(k_{ji}) =: S$$

is an algebraic isomorphism if restricted to $T_j(H)$. Moreover, if $I_0 \subseteq \varphi(\mathcal{U}_0)$, then

$$\|Py\|_{I_0} = \|y\|_{I_0} \quad \text{for } y \in \prod_{i \in I} l_p(k_{ji}).$$

If H_1 is an arbitrary algebraic complement of $P \circ T_j(H)$ in S and $F = H_1 + T_j(H)$, then $P|_F$ is an algebraic isomorphism onto! Now,

$$G := T_j^{-1}(F); \quad T := T_j^{-1} \circ (P|_F)^{-1}$$

are the space and the mapping we are looking for.

For strongly \mathcal{L}_p -spaces the proof is similar.

EXAMPLE 5.2. It is easily seen that a Banach space is a (strongly) locally or globally \mathcal{L}_p -space iff it is an \mathcal{L}_p -space in the sense of Lindenstrauss and Pełczyński.

PROPOSITION 5.3. *Every product of globally (strongly globally, locally, strongly locally, resp.) \mathcal{L}_p -spaces is a globally (strongly globally, locally, strongly locally, resp.) \mathcal{L}_p -space.*

EXAMPLE 5.4. Every product of Banach \mathcal{L}_p -spaces is a strongly globally \mathcal{L}_p -space.

By Cor. 8.22 below, every complete barrelled locally \mathcal{L}_2 -space with a bornological dual is isomorphic to a product of Hilbert spaces, in particular, every complete strongly globally \mathcal{L}_2 -space and every Fréchet \mathcal{L}_2 -space is isomorphic to a product of Hilbert spaces.

For $p = 1$ or ∞ , the class of Fréchet globally \mathcal{L}_p -spaces is not exhausted by complemented subspaces of products of Banach \mathcal{L}_p -spaces (comp. Cor. 10.2). Before we present the corresponding example we need a few results.

LEMMA 5.5. *Let $p = 1$ or ∞ and let X be a Banach $\mathcal{L}_{p,\lambda}$ -space, $\lambda > 1$. If Y is an $\mathcal{L}_{p,\lambda}$ -subspace of X such that X/Y is also an $\mathcal{L}_{p,\lambda}$ -space, $q: X \rightarrow X/Y$ is the quotient map, then for every finite-dimensional subspace E of X there exists a finite-dimensional subspace $K \subseteq X$, a continuous onto projection $P: X \rightarrow K$ and a linear section $s: q(K) \rightarrow K$ such that:*

- (i) $\ker P \supseteq Y$;
- (ii) $\|P\| \leq 2\lambda^2$;
- (iii) $E \subseteq K + Y$;
- (iv) $d(K, l_p(\dim K)) \leq 2\lambda^2$;
- (v) $\|s\| \leq 2\lambda$.

Proof. Obviously, $q(E)$ is contained in a λ -isomorphic and λ -complemented (a projection $R: X/Y \rightarrow W$) copy W of a finite-dimensional l_p -space. Moreover, there is a linear section $s: W \rightarrow X$ for q :

$$q \circ s(x) = x \quad \text{for } x \in W \quad \text{and} \quad \|s\| \leq 2\lambda.$$

Indeed, this is obvious for $p = 1$, for $p = \infty$ we can find a projection $s_0: q^{-1}(W) \rightarrow Y$ of norm $\leq 2\lambda - 1$ (see [41, Lemma 3.2]). Thus we define

$$s(q(x)) := x - s_0(x), \quad K := s(W), \quad P := s \circ R \circ q,$$

and our lemma follows easily.

COROLLARY 5.6. *Let Y be a subspace of a Banach space X such that $Y, X, X/Y$ are $\mathcal{L}_{p,\lambda}$ -spaces, $\lambda > 1$, $p = 1$ or ∞ , and let $q: X \rightarrow X/Y$ be the natural quotient map. For every finite-dimensional subspace E of X and C_1 of X/Y , $d(C_1, l_p(\dim C_1)) \leq \lambda$, $q(E) \subseteq C_1$, there are: a map $s: C_1 \rightarrow X$, $\|s\| \leq 2\lambda$, $q \circ s = \text{id}_{C_1}$, a finite-dimensional subspace $B \subseteq Y$ and a projection $P: B + C \rightarrow B$, $\|P\| \leq 2\lambda^2$, $P(C) = \{0\}$, $C := s(C_1)$, such that B and C are $2\lambda^2$ -isomorphic to finite-dimensional l_p -spaces and $E \subseteq B + C$.*

References

- [1] R. Bartoszyński, *A characterization of weak convergence of measures*, Ann. Math. Stat. 32 (1961), 561–576.
- [2] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York 1968.
- [3] J. P. R. Christensen, *Topology and Borel structure*, North-Holland–Elsevier, Amsterdam 1974.
- [4] D. L. Cohn, *Measure Theory*, Birkhäuser, Boston 1980.
- [5] A. de Acosta, *Invariance principles for triangle arrays of B -valued random vectors*, Ann. Probab. 10 (1982), 346–373.
- [6] R. M. Dudley, *Probabilities and metrics*, Aarhus Universitet, Aarhus 1976.
- [7] X. Fernique, *Sur le Théorème de Kantorovich–Rubinstein dans les espaces polonaises*, Lecture Notes in Math. 850, Springer–Verlag, 1981, pp. 6–10.
- [8] R. Fortet and E. Mourier, *Convergence de la repartition empirique vers la repartition théorique*, Ann. Sci. École Norm. Sup. 70 (3) (1953), 267–285.
- [9] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, Michigan Math. J. 31 (1984), 231–240.
- [10] L. V. Kantorovich and G. S. Rubinstein, *On a space of completely additive functions* (in Russian), Vestnik Leningrad. Univ. 13 (7) (1958), 52–59.
- [11] H. G. Kellerer, *Duality theorems for marginal problems*, Z. Wahrsch. Verw. Gebiete 67 (1984), 399–432.
- [12] – *Duality theorems and probability metrics*, in: Proc. 7th Brasov Conf. 1982, Bucuresti 1984, pp. 211–220.
- [13] – *Measure-theoretic versions of linear programming*, Math. Z. 198 (1988), 367–400.
- [14] J. H. B. Kemperman, *On the role of duality in the theory of moments*, in: Proc. Semi-infinite Programming and Applications 1981, Lecture Notes in Economics and Math. Systems 215, Springer-Verlag, New York 1983, pp. 63–92.
- [15] V. L. Levin and A. A. Milyutin, *The problem of mass transfer with a discontinuous cost function*, Russian Math. Surveys 34 (3) (1979), 1–78.
- [16] V. L. Levin, *The problem of mass transfer in a topological space*, Soviet Math. Dokl. (1984), 638–643.
- [17] J. Neveu and R. M. Dudley, *On Kantorovich–Rubinstein theorems*, typescript.
- [18] S. T. Rachev, *The Monge–Kantorovich mass transfer problem and its stochastic applications*, Theory Prob. Appl. 29 (1984), 647–676.
- [19] – *Extreme functionals in the space of probability measures*, in: Proc. Stability problems for stochastic models 1984, Lecture Notes in Math. 1155, Springer-Verlag, New York 1985, pp. 320–348.
- [20] R. Ranga Rao, *Relations between weak and uniform convergence of measures with applications*, Ann. Math. Stat. 33 (1962), 659–680.
- [21] R. M. Shortt, *Strassen’s marginal problem in two or more dimensions*, Z. Wahrsch. Verw. Gebiete 64 (1983), 313–325.
- [22] V. Strassen, *The existence of probability measures with given marginals*, Ann. Math. Stat. 36 (1965), 423–439.

- [23] A. Szulga, *On the Wasserstein metric*, in: *Transactions 8th Prague Conf. on Information Theory*, Akademia, Prague 1978, pp. 267–273.
- [24] – *On minimal metrics in the space of random variables*, *Theory Prob. Appl.* 27 (1982), 424–430.
- [25] S. S. Vallander, *Calculation of the Wasserstein distance between probability distributions on the line*, *Theory Prob. Appl.* 18 (1973), 784–786.
- [26] R. G. Douglas, *On extremal measures and subspace density*, *Michigan Math. J.* 11 (1964), 243–246.
- [27] J. Lindenstrauss, *A remark on doubly-stochastic measures*, *Amer. Math. Monthly* 72 (1965), 379–382.
- [28] R. M. Shortt, *The singularity of extremal measures*, *Real Analysis Exchange* 12 (1986–7), 205–215.

UNIVERSITY OF CALIFORNIA
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
SANTA BARBARA, CALIFORNIA 93106, USA