STUDIA MATHEMATICA 200 (1) (2010)

A simple proof in Monge–Kantorovich duality theory

by

D. A. Edwards (Oxford)

Abstract. A simple proof is given of a Monge–Kantorovich duality theorem for a lower bounded lower semicontinuous cost function on the product of two completely regular spaces. The proof uses only the Hahn–Banach theorem and some properties of Radon measures, and allows the case of a bounded continuous cost function on a product of completely regular spaces to be treated directly, without the need to consider intermediate cases. Duality for a semicontinuous cost function is then deduced via the use of an approximating net. The duality result on completely regular spaces also allows us to extend to arbitrary metric spaces a well known duality theorem on Polish spaces, at the same time simplifying the proof. A deep investigation by Kellerer [Z. Warsch. Verw. Gebiete 67 (1984)] yielded a wide range of conditions sufficient for duality to hold. The more limited aims of the present paper make possible simpler, very direct, proofs which also offer an alternative to some recent accounts of duality.

1. Introduction. This paper is inspired by Chapter 1 of Villani's beautiful book [12]. Using a minimax approach, Villani employs the Fenchel–Rockafellar duality theorem of convexity theory to prove a version for Polish spaces of the Monge–Kantorovich duality theorem of optimal transport theory. His proof thus depends indirectly on the Hahn–Banach theorem, since the Fenchel–Rockafellar theorem is a non-trivial consequence of the latter. (See also [8] for another saddle-point proof, and [13] for a radically different proof.) The object of the present paper is to show that a more direct application of the Hahn–Banach theorem provides a simpler treatment, and also leads to somewhat more general statements. We prove a Monge–Kantorovich duality theorem for a lower bounded lower semicontinuous cost function on the product of two completely regular spaces (Theorem 4.1), and use it to extend to arbitrary metric spaces (in Theorem 5.1) a well known duality theorem [12, Theorem 1.3] on Polish spaces, at the same time simplifying the proof. Kellerer, in his fundamental paper [7] concerning duality, devel-

2010 *Mathematics Subject Classification*: Primary 90C46, 49N15; Secondary 46A22, 28C05, 28C15.
Key words and phrases: Monge–Kantorovich duality, Hahn–Banach theorem, Radon measures.

DOI: 10.4064/sm200-1-4

ops an apparatus that allows him to obtain a very wide variety of duality theorems, but his proofs are for that reason very different and also considerably more difficult. For a comprehensive survey of the literature on Monge–Kantorovich duality see [13, pp. 97–104].

The paper is organized as follows. In §2 we recall some basic facts about Radon probability measures defined on a product $Z = X \times Y$ of completely regular spaces and having prescribed marginals. Using the Hahn–Banach theorem, we obtain in §3 functionals on $\mathcal{C}_b(Z)$ which satisfy a certain tightness condition that allows them to be represented as integrals with respect to Radon probability measures on Z having the appropriate marginals. Exploiting such functionals, we arrive immediately in §4 at the Monge–Kantorovich duality for a cost function belonging to $\mathcal{C}_b(Z)$. Duality for a lower bounded lower semicontinuous cost function is then deduced easily via the use of an approximating net. In §5 we take X and Y to be metric spaces. Here the approximation argument requires only sequential convergence, and with the help of a standard generalization of convex conjugacy we obtain a sharper result.

For technical reasons, we in general formulate everything below that involves a cost function $c(x, y)$ in terms of its *negative* $h(x, y) = -c(x, y)$.

2. Image measures and marginal probabilities. Throughout what follows it will be tacitly assumed that all topological spaces mentioned are Hausdorff. Given a completely regular topological space S, we denote by $\mathcal{M}_b(S)$ the space of bounded real Radon measures on S and by $\mathcal{P}(S)$ the set of all probability Radon measures on S. (For the theory of Radon measures see, for instance, [1, 2, 3, 5].) We denote by $\mathfrak{B}(S)$ the set of all Borel subsets of S, by $\mathfrak{K}(S)$ the set of all compact subsets of S, and by $\mathcal{C}_b(S)$ the space of bounded real continuous functions on S. The norm $\|\cdot\|_\infty$ in $\mathcal{C}_b(S)$ is taken to be the supremum norm. By the *weak topology* for $\mathcal{M}_b(S)$ we mean the topology $\sigma(\mathcal{M}_b(S), \mathcal{C}_b(S))$.

A measure $\sigma \in \mathcal{M}_b(S)$ is *positive* if $\sigma(B) \geq 0$ for all $B \in \mathfrak{B}(S)$; this is the case if and only if $\int_S f \, d\sigma \geq 0$ for all $f \in \mathcal{C}_b^+(S)$. Suppose now that T is another completely regular space and that $\phi : S \to T$ is a continuous surjection. Then $\phi^{-1}(E) \in \mathfrak{B}(S)$ for all $E \in \mathfrak{B}(T)$. Given $\sigma \in \mathcal{P}(S)$, we define the *image measure* $\phi(\sigma)$ by postulating that $\phi(\sigma)(E) = \sigma(\phi^{-1}(E))$ for all $E \in \mathfrak{B}(T)$, and we have $\phi(\sigma) \in \mathcal{P}(T)$. We shall frequently use functional notation for integrals. For example, in the following proposition $\tau(f)$ denotes the integral $\int_T f \, d\tau$.

PROPOSITION 2.1. *Let S, T be completely regular spaces and suppose that $\phi : S \to T$ is a continuous surjection. Let $\sigma \in \mathcal{P}(S)$ and $\tau \in \mathcal{P}(T)$. Then the following assertions are equivalent:*

(i) $\tau = \phi(\sigma)$;
(ii) $\tau(f) = \sigma(f \circ \phi)$ *for all* $f \in \mathcal{C}_b(T)$;
(iii) *for every Borel function* $f : T \to [-\infty, \infty]$ *in* $\mathcal{L}^1(\tau)$ *we have* $f \circ \phi \in \mathcal{L}^1(\sigma)$ *and* $\sigma(f \circ \phi) = \tau(f)$.

Proof. The implication (i)⇒(iii) is an elementary consequence of the definition of $\phi(\sigma)$. That (iii)⇒(ii) is trivial. To prove that (ii)⇒(i), let G be an open subset of T. Then its indicator function $\mathbf{1}_G$ is lower semicontinuous and by the complete regularity of T there exists an increasing net (f_α) in $\mathcal{C}_b^+(T)$ that converges pointwise to $\mathbf{1}_G$. Passing to the limit in the equation $\tau(f_\alpha) = \sigma(f_\alpha \circ \phi)$, and noting the fact that $\mathbf{1}_G \circ \phi = \mathbf{1}_{\phi^{-1}(G)}$, we obtain $\tau(G) = \sigma(\phi^{-1}(G))$. This shows that the two measures τ and $\phi(\sigma)$ agree on open sets, and hence on Borel sets, and are thus equal. ∎

Now let X, Y be completely regular spaces, and let Z be the topological product $X \times Y$. Note that Z is completely regular. We denote by pr_X and pr_Y the natural projections of Z onto X and Y respectively. We suppose given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, and we denote by $\Pi(\mu, \nu)$ the convex set consisting of all $\pi \in \mathcal{P}(Z)$ such that $\mathrm{pr}_X(\pi) = \mu$ and $\mathrm{pr}_Y(\pi) = \nu$. Note that $\Pi(\mu, \nu) \neq \emptyset$, because the product measure $\theta = \mu \otimes \nu$ belongs to $\Pi(\mu, \nu)$. Given functions $u : X \to (-\infty, \infty]$ and $v : Y \to (-\infty, \infty]$, we shall denote by $u \oplus v$ the function $(x, y) \mapsto u(x) + v(y)$. The following result is an immediate consequence of Prop. 2.1

Corollary 2.2. *Suppose that* $\pi \in \mathcal{P}(Z)$. *Then the following assertions are equivalent:*

(i) $\pi \in \Pi(\mu, \nu)$;
(ii) *for all* $A \in \mathfrak{B}(X)$ *and* $B \in \mathfrak{B}(Y)$ *we have* $\pi(A \times Y) = \mu(A)$ *and* $\pi(X \times B) = \nu(B)$;
(iii) *for all* $u \in \mathcal{C}_b(X)$ *and* $v \in \mathcal{C}_b(Y)$ *we have* $\pi(u \oplus v) = \mu(u) + \nu(v)$;
(iv) *whenever* $u : X \to (-\infty, \infty]$ *and* $v : Y \to (-\infty, \infty]$ *are Borel functions in* $\mathcal{L}^1(\mu)$ *and* $\mathcal{L}^1(\nu)$ *respectively we have* $u \oplus v \in \mathcal{L}^1(\pi)$ *and* $\pi(u \oplus v) = \mu(u) + \nu(v)$.

Now suppose that $\epsilon > 0$. Since μ, ν are Radon measures, there exist $K \in \mathfrak{K}(X)$ and $L \in \mathfrak{K}(Y)$ such that

$$\mu(X \setminus K) < \epsilon/2 \quad \text{and} \quad \nu(Y \setminus L) < \epsilon/2. \tag{2.1}$$

Then for all $\pi \in \Pi(\mu, \nu)$ we have

$$\begin{aligned} \pi(Z \setminus (K \times L)) &\leq \pi((X \setminus K) \times Y) + \pi(X \times (Y \setminus L)) \\ &= \mu(X \setminus K) + \nu(Y \setminus L) < \epsilon. \end{aligned}$$

Thus the (bounded) set of measures $\Pi(\mu, \nu)$ is uniformly tight and so, by Prokhorov's theorem (see, for example, [3, IX, §5, No. 5, Theorem 1] or [2,

Theorem 8.6.7]), it is relatively $\sigma(\mathcal{M}_b(Z), \mathcal{C}_b(Z))$-compact. Now let ρ belong to the weak closure of $\Pi(\mu,\nu)$, and let (ρ_α) be a net in $\Pi(\mu,\nu)$ that converges weakly to ρ. Then $\rho \in \mathcal{P}(Z)$ and $\lim_\alpha \rho_\alpha(f) = \rho(f)$ for all $f \in \mathcal{C}_b(Z)$. In particular, $\lim_\alpha \rho_\alpha(u \circ \mathrm{pr}_X) = \rho(u \circ \mathrm{pr}_X)$ for all $u \in \mathcal{C}_b(X)$. But for all $u \in \mathcal{C}_b(X)$ and all α we have $\rho_\alpha(u \circ \mathrm{pr}_X) = \mu(u)$, and hence $\rho(u \circ \mathrm{pr}_X) = \mu(u)$. Thus $\mathrm{pr}_X(\rho) = \mu$; similarly $\mathrm{pr}_Y(\rho) = \nu$. This shows that $\rho \in \Pi(\mu,\nu)$, hence that $\Pi(\mu,\nu)$ is weakly closed and therefore weakly compact. We have thus proved the following well known theorem.

THEOREM 2.3. *The set $\Pi(\mu,\nu)$ is a non-empty $\sigma(\mathcal{M}_b(Z), \mathcal{C}_b(Z))$-compact convex subset of $\mathcal{P}(Z)$.*

3. Representation of certain functionals. Let $\mathcal{F}(Z)$ denote the set of all upper semicontinuous functions $f : Z \to [-\infty, \infty)$ and, for $f \in \mathcal{F}(Z)$, let $\Phi(f)$ be the set of pairs (u, v) of Borel functions $u : X \to (-\infty, \infty]$ and $v : Y \to (-\infty, \infty]$ such that $u \in \mathcal{L}^1(\mu)$ and $v \in \mathcal{L}^1(\nu)$ and which satisfy

$$f(x,y) \le u(x) + v(y) \quad \text{for all } (x,y) \in Z.$$

We define $p(f)$ by the formula

$$p(f) = \begin{cases} \inf\{\mu(u) + \nu(v) : (u,v) \in \Phi(f)\} & \text{if } \Phi(f) \neq \emptyset, \\ \infty & \text{if } \Phi(f) = \emptyset. \end{cases}$$

It is easy to see that the map $f \mapsto p(f)$ is isotone, and that $p(0) = 0$, $p(-1) = -1$, $p(1) = 1$. If $f \in \mathcal{F}(Z)$, $(u,v) \in \Phi(f)$, and $\pi \in \Pi(\mu,\nu)$ then $\pi(f) \le \pi(u \oplus v) = \mu(u) + \nu(v)$. It follows that

$$\pi(f) \le p(f) \quad \text{for all } \pi \in \Pi(\mu,\nu) \text{ and } f \in \mathcal{F}(Z). \tag{3.1}$$

PROPOSITION 3.1. *On $\mathcal{C}_b(Z)$ the functional $p(\cdot)$ has the following properties:*

(i) $-\|f\|_\infty \le \inf f \le p(f) \le \sup f \le \|f\|_\infty$ *for all* $f \in \mathcal{C}_b(Z)$;
(ii) *the map* $\mathcal{C}_b(Z) \ni f \mapsto p(f)$ *is sublinear;*
(iii) $p(u \oplus v) = \mu(u) + \nu(v)$ *for all* $u \in \mathcal{C}_b(X)$ *and* $v \in \mathcal{C}_b(Y)$.

Proof. (i) This is obvious.

(ii) Obviously $p(\lambda f) = \lambda p(f)$ for all $f \in \mathcal{C}_b(Z)$ and all constants $\lambda \ge 0$. Suppose next that $f_i \in \mathcal{C}_b(Z)$ and $(u_i, v_i) \in \Phi(f_i)$ for $i = 1, 2$. Then

$$(u_1 + u_2, v_1 + v_2) \in \Phi(f_1 + f_2)$$

and hence

$$p(f_1 + f_2) \le \mu(u_1 + u_2) + \nu(v_1 + v_2) = (\mu(u_1) + \nu(v_1)) + (\mu(u_2) + \nu(v_2)).$$

Hence $p(f_1 + f_2) \le p(f_1) + p(f_2)$.

(iii) Since $(u,v) \in \Phi(u \oplus v)$ we have, using the inequality (3.1),

$$\mu(u) + \nu(v) = \theta(u \oplus v) \le p(u \oplus v) \le \mu(u) + \nu(v),$$

and so must have equality throughout. ∎

Let S be a completely regular space and $I : \mathcal{C}_b(S) \to \mathbb{R}$ a linear functional. We shall say that I is *tight* if for each $\epsilon > 0$ there exists $Q(\epsilon) \in \mathfrak{K}(S)$ such that $|I(f)| < \epsilon$ for each $f \in \mathcal{C}_b(S)$ that satisfies: (i) f vanishes identically on $Q(\epsilon)$, and (ii) $\|f\|_\infty \leq 1$. We shall need the following representation theorem, for which see [3, IX, §5, No. 2, Proposition 5] or [2, Theorem 7.10.6].

THEOREM 3.2. *Let S be a completely regular space and $I : \mathcal{C}_b(S) \to \mathbb{R}$ a continuous tight linear functional. Then there exists a unique measure $\sigma \in \mathcal{M}_b(S)$ such that $I(f) = \int_S f \, d\sigma$ for all $f \in \mathcal{C}_b(S)$.*

If $\sigma \in \Pi(\mu, \nu)$ and $I(f) = \int_Z f \, d\sigma$ then, as we have seen, $I(f) \leq p(f)$ for all $f \in \mathcal{C}_b(Z)$. We now prove the following converse statement.

THEOREM 3.3. *Let $I : \mathcal{C}_b(Z) \to \mathbb{R}$ be a linear functional such that $I(f) \leq p(f)$ for all $f \in \mathcal{C}_b(Z)$. Then there exists a unique measure $\sigma \in \Pi(\mu, \nu)$ such that $I(f) = \int_Z f \, d\sigma$.*

Proof. We shall use Theorem 3.2 for the case $S = Z$.

We have $I(f) \leq p(f) \leq \|f\|_\infty$ for all $f \in \mathcal{C}_b(Z)$. Replacing f by $-f$ we see that $-I(f) \leq \|f\|_\infty$ and hence that $|I(f)| \leq \|f\|_\infty$. Thus I is continuous.

Next, we show that I is tight. To prove this, suppose that $\epsilon > 0$ and choose compact subsets K, L of X, Y respectively to satisfy the inequalities (2.1). Let $f \in \mathcal{C}_b(Z)$ and suppose that f vanishes identically on the compact set $K \times L$ and that $\|f\|_\infty \leq 1$. Now let

$$u(x) = \begin{cases} 1 & \text{if } x \in X \setminus K, \\ 0 & \text{if } x \in K, \end{cases} \qquad v(y) = \begin{cases} 1 & \text{if } y \in Y \setminus L, \\ 0 & \text{if } y \in L. \end{cases}$$

Then $(u, v) \in \Phi(f)$, so $I(f) \leq p(f) \leq \mu(u) + \nu(v) < \epsilon$. Similarly, $(u, v) \in \Phi(-f)$, so $-I(f) = I(-f) < \epsilon$. Hence $|I(f)| < \epsilon$, and thus the functional I is tight. By Theorem 3.2, it follows that there exists a unique measure $\sigma \in \mathcal{M}_b(Z)$ such that $I(f) = \int_Z f \, d\sigma$ for all $f \in \mathcal{C}_b(Z)$.

We next show that $\sigma \in \mathcal{P}(Z)$. Observe first that if $f \in \mathcal{C}_b^+(Z)$ then

$$-I(f) = I(-f) \leq p(-f) \leq 0$$

and so $I(f) \geq 0$. Therefore $\sigma \in \mathcal{M}_b^+(Z)$. Next,

$$I(1) \leq p(1) = 1 \quad \text{and} \quad -I(1) = I(-1) \leq p(-1) = -1,$$

So $\int_Z d\sigma = I(1) = 1$, and thus $\sigma \in \mathcal{P}(Z)$.

Finally, let $\Lambda = \{u \oplus v : u \in \mathcal{C}_b(X), v \in \mathcal{C}_b(Y)\}$. Then, for all $f \in \Lambda$, $p(f) = \theta(f)$, by Proposition 3.1, and hence $I(f) \leq \theta(f)$ and $I(-f) \leq \theta(-f)$. Therefore $I(f) = \theta(f)$ for all $f \in \Lambda$; in other words

$$\sigma(u \oplus v) = \mu(u) + \nu(v)$$

for all $u \in \mathcal{C}_b(X)$ and $v = \mathcal{C}_b(Y)$, and thus $\sigma \in \Pi(\mu, \nu)$. ∎

4. Monge–Kantorovich duality in completely regular spaces. We continue to suppose that X and Y are completely regular topological spaces.

THEOREM 4.1. *Let $h : Z \to [-\infty, \infty)$ be an upper semicontinuous function and suppose that there exist real lower semicontinuous functions $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$ such that $h(x,y) \le a(x) + b(y)$ for all $(x,y) \in Z$. Then*

$$\max_{\pi \in \Pi(\mu,\nu)} \pi(h) = \inf\{\mu(u) + \nu(v) : (u,v) \in \Phi(h)\}.$$

(*The case in which both terms equal $-\infty$ is not excluded.*)

Kellerer [7, Theorem 2.6] (see also [7, Theorem 2.19]) obtains a duality theorem that is stronger than Theorem 4.1 in a number of respects, but the proof is much more difficult. Note that Kellerer's theorem, unlike Theorem 4.1, does not stipulate that h is to satisfy an inequality of the type $h \le a \oplus b$. On the other hand, for lower semicontinuous h (not treated in the present paper) duality may fail unless $a \oplus b \le h$ for suitable $a \in \mathcal{L}^1(\mu)$ and $b \in \mathcal{L}^1(\nu)$ (see [7, Theorem 2.2 and Example 2.5]).

In the optimal transport literature it is customary to reformulate statements such as Theorem 4.1 in terms of the function $c(x,y) = -h(x,y)$, known as the *cost function*. The total cost of the *transport plan* π is then $\int_Z c\, d\pi$ and the recast Theorem 4.1 evaluates the minimum possible total cost, namely $\min_{\pi \in \Pi(\mu,\nu)} \int_Z c\, d\pi$. (See, for instance, [12, 13].)

Proof of Theorem 4.1. We treat first the case in which $h \in \mathcal{C}_b(Z)$. By the Hahn–Banach theorem there exists a linear functional $I : \mathcal{C}_b(Z) \to \mathbb{R}$ such that

$$I(f) \le p(f) \quad \text{for all } f \in \mathcal{C}_b(Z), \quad \text{and} \quad I(h) = p(h).$$

By Theorem 3.3 there exists $\sigma \in \Pi(\mu,\nu)$ such that $I(f) = \sigma(f)$ for all $f \in \mathcal{C}_b(Z)$. The equation $I(h) = p(h)$ can now be expressed as $\sigma(h) = p(h)$. But by the inequality (3.1) we have $\pi(h) \le p(h)$ for all $\pi \in \Pi(\mu,\nu)$. This shows that

$$\max_{\pi \in \Pi(\mu,\nu)} \pi(h) = \sigma(h) = p(h) = \inf\{\mu(u) + \nu(v) : (u,v) \in \Phi(h)\}$$

as desired.

To deal with semicontinuous h we shall use the following lemma, a close relative of Dini's theorem.

LEMMA 4.2. *Let Ω be a compact space and let $(g_\alpha)_{\alpha \in A}$ be a decreasing net in $\mathcal{C}(\Omega)$ with pointwise limit $g : \Omega \to [-\infty, \infty)$. Then*

$$\lim_\alpha \max_{\omega \in \Omega} g_\alpha(\omega) = \inf_\alpha \max_{\omega \in \Omega} g_\alpha(\omega) = \max_{\omega \in \Omega} g(\omega).$$

Proof. Here g is upper semicontinuous, and hence it attains its supremum. Now let Δ be a real constant such that $\max_{\omega\in\Omega} g(\omega) < \Delta$, and for each α let

$$F(\alpha) = \{\omega \in \Omega : g_\alpha(\omega) \geq \Delta\}.$$

Then if $\alpha_1, \dots, \alpha_n \in A$ we can find $\beta \in A$ such that $\alpha_r \leq \beta$ for $r = 1, \dots, n$, and hence we have

$$F(\alpha_1) \cap \cdots \cap F(\alpha_n) \supseteq F(\beta).$$

It follows, if $F(\alpha) \neq \emptyset$ for all α, that $\{F(\alpha)\}$ is a family of closed sets with the finite intersection property, and consequently that $\bigcap_\alpha F(\alpha) \neq \emptyset$. In that case let $\omega_0 \in \bigcap_\alpha F(\alpha)$. Then $g_\alpha(\omega_0) \geq \Delta$ for all α. This contradicts the assumption that $(g_\alpha(\omega_0))$ tends to the limit $g(\omega_0)$, so we are obliged to conclude that there is some α_0 for which $F(\alpha_0) = \emptyset$. Then $\max_{\omega\in\Omega} g(\omega) \leq \max_{\omega\in\Omega} g_\alpha(\omega) < \Delta$ for all $\alpha \geq \alpha_0$. ∎

Returning to our proof of Theorem 4.1, we suppose for the time being that $h \leq 0$. We need one further lemma.

LEMMA 4.3. *Let $z_0 \in Z$ and let t be a real constant such that $h(z_0) < t$. Then there exists $g \in \mathcal{C}_b(Z)$ such that $h \leq g \leq 0$ and $g(z_0) < t$.*

Proof. If $t > 0$, let g be identically zero. If $t \leq 0$, choose a constant s such that $h(z_0) < s < t$ and let $V = \{\, z : h(z) < s \,\}$. Then V is open and, by the complete regularity of Z, there exists a continuous function $\chi : Z \to [0,1]$ such that $\chi(z_0) = 1$ and $\chi(z) = 0$ for all $z \in \complement V$. It now suffices to take $g = s\chi$. ∎

From this lemma it follows immediately that there exists a decreasing net (h_α) in $\mathcal{C}_b(Z)$ that has pointwise limit h and is such that $h_\alpha \leq 0$ for all α. Then $(\pi(h_\alpha))$ is for each $\pi \in \Pi(\mu,\nu)$ a decreasing net in $(-\infty, 0]$ that, by [1, Theorem 1.5], has the limit $\pi(h) \in [-\infty, 0]$. Write $\hat{h}_\alpha(\pi) = \pi(h_\alpha)$, $\hat{h}(\pi) = \pi(h)$ and recall that $\Pi(\mu,\nu)$ is a non-empty $\sigma(\mathcal{M}_b(Z), \mathcal{C}_b(Z))$-compact set. Thus $(\hat{h}_\alpha)$ is a decreasing net in $\mathcal{C}(\Pi(\mu,\nu))$ with pointwise limit $\hat{h}$. By Lemma 4.2 we have

$$\lim_\alpha \max_{\pi\in\Pi(\mu,\nu)} \hat{h}_\alpha(\pi) = \inf_\alpha \max_{\pi\in\Pi(\mu,\nu)} \hat{h}_\alpha(\pi) = \max_{\pi\in\Pi(\mu,\nu)} \hat{h}(\pi). \tag{4.1}$$

By (3.1) we have $\hat{h}(\pi) \leq p(h)$; and $\max_{\pi\in\Pi(\mu,\nu)} \hat{h}_\alpha(\pi) = p(h_\alpha)$ for all α by the first part of the present proof. Thus

$$\hat{h}(\pi) \leq p(h) \leq p(h_\alpha) = \max_{\rho\in\Pi(\mu,\nu)} \hat{h}_\alpha(\rho)$$

for all $\pi \in \Pi(\mu,\nu)$ and all α. We deduce, using (4.1), that

$$\max_{\pi\in\Pi(\mu,\nu)} \hat{h}(\pi) \leq p(h) \leq \inf_\alpha p(h_\alpha) = \inf_\alpha \max_{\rho\in\Pi(\mu,\nu)} \hat{h}_\alpha(\rho) = \max_{\rho\in\Pi(\mu,\nu)} \hat{h}(\rho).$$

We therefore have equality throughout. Reverting to our former notation, we see in particular that

$$\max_{\pi\in\Pi(\mu,\nu)} \pi(h) = p(h) = \inf\{\,\mu(u) + \nu(v) : (u,v) \in \Phi(h)\}$$

as desired.

Finally we must consider the general case, in which the condition $h \le a \oplus b$ replaces the assumption that $h \le 0$. Let us write $k = h - a \oplus b$. Then k is an upper semicontinuous map $Z \to [-\infty, 0]$ and, by what we have already proved, $\max_{\pi\in\Pi(\mu,\nu)} \pi(k) = p(k)$. Moreover

$$p(h) = p(k) + \mu(a) + \nu(b),$$

and for all $\pi \in \Pi(\mu,\nu)$ we have

$$\pi(h) = \pi(k) + \mu(a) + \nu(b).$$

Hence

$$\max_{\pi\in\Pi(\mu,\nu)} \pi(h) = \max_{\pi\in\Pi(\mu,\nu)} \pi(k) + \mu(a) + \nu(b) = p(k) + \mu(a) + \nu(b) = p(h). \ \blacksquare$$

5. Monge–Kantorovich duality in metric spaces. In this section we assume that X, Y are metric spaces with metrics d_X and d_Y respectively. Then Z is metrizable and we may take its metric d_Z to be defined by

$$d_Z(z,z') = d_X(x,x') + d_Y(y,y'),$$

where $z = (x,y)$ and $z' = (x',y')$. Given $f \in \mathcal{F}(Z)$, we define $\Psi(f)$ to be the set of all pairs (u,v) of functions $u \in \mathcal{C}_{\rm bu}(X)$ and $v \in \mathcal{C}_{\rm bu}(Y)$ such that

$$f(x,y) \le u(x) + v(y) \text{ for all } (x,y) \in Z,$$

where, for instance, $\mathcal{C}_{\rm bu}(X)$ denotes the space of all bounded real *uniformly* continuous functions on X.

THEOREM 5.1. *Let $h : Z \to [-\infty, \infty)$ be an upper semicontinuous function that is bounded above. Then*

$$\max_{\pi\in\Pi(\mu,\nu)} \pi(h) = \inf\{\mu(u) + \nu(v) : (u,v) \in \Psi(h)\}.$$

This theorem is proved for Polish spaces by Villani [12, Theorem 1.3]. Our proof differs from his because we use Theorem 4.1; this simplifies the argument and allows us to drop the assumption that the metric spaces X, Y are Polish.

Proof of Theorem 5.1. For $f \in \mathcal{F}(Z)$ we define $q(f)$ by the formula

$$q(f) = \begin{cases} \inf\{\mu(u) + \nu(v) : (u,v) \in \Psi(f)\} & \text{if } \Psi(f) \ne \emptyset, \\ \infty & \text{if } \Psi(f) = \emptyset. \end{cases}$$

LEMMA 5.2. *Let h be a bounded real uniformly continuous function on Z. Then $q(h) = p(h)$.*

Outline of proof. We are indebted here to Villani's account (see [12, pp. 29–30]) of a standard generalization of convex conjugacy. We can assume that $h \le 0$. Evidently $\inf h \le p(h) \le q(h) \le 0$. Suppose that $0 < \epsilon < 1$. For suitable $(u_0, v_0) \in \Phi(h)$ we now have

$$p(h) \le \mu(u_0) + \nu(v_0) < p(h) + \epsilon < 1.$$

Thus $\int_Z (u_0 \oplus v_0)\, d\theta < 1$, and hence $u_0(x_0) + v_0(y_0) < 1$ for some $(x_0, y_0) \in Z$. Since $\mu(u_0 + t) + \nu(v_0 - t) = \mu(u_0) + \nu(v_0)$ for $t \in \mathbb{R}$, we can suppose that

$$u_0(x_0) < 1/2 \quad \text{and} \quad v_0(y_0) < 1/2. \tag{5.1}$$

Now for $x \in X$ and $y \in Y$ let

$$\overline{u}_0(x) = \sup_{y \in Y} (h(x,y) - v_0(y)), \qquad \overline{v}_0(y) = \sup_{x \in X} (h(x,y) - \overline{u}_0(x)). \tag{5.2}$$

From (5.1) and (5.2) it follows that $|\overline{u}_0| \le \|h\|_\infty + 1/2$ and $h \le \overline{u}_0 \oplus v_0$. Moreover $\overline{u}_0 \le u_0$, and hence $\overline{u}_0(x_0) < 1/2$.

Next, we prove that $\overline{u}_0$ is uniformly continuous. Let $k(x,y) = h(x,y) - v_0(y)$ and let $N = \{y \in Y : v_0(y) = \infty\}$. Let $Y_0 = Y \setminus N$ and observe that $\overline{u}_0(x) = \sup_{y \in Y_0} k(x,y)$. Suppose that $\eta > 0$. By the uniform continuity of h we can find $\delta > 0$ such that

$$|k(x,y) - k(x',y)| = |h(x, y - h(x',y))| < \eta \quad \text{for all } y \in Y_0$$

whenever $d_X(x,x') < \delta$. For such x, x' and all $y \in Y_0$ we thus have

$$k(x,y) < k(x',y) + \eta.$$

Hence

$$\overline{u}_0(x) = \sup_{y \in Y_0} k(x,y) \le \sup_{y \in Y_0} k(x',y) + \eta = \overline{u}_0(x') + \eta.$$

By symmetry, $\overline{u}_0(x') \le \overline{u}_0(x) + \eta$ and hence

$$|\overline{u}_0(x) - \overline{u}_0(x')| \le \eta$$

whenever $d_X(x,x') < \delta$. Thus $\overline{u}_0$ is uniformly continuous, and we see that $\overline{u}_0 \in \mathcal{C}_{\mathrm{bu}}(X)$ and $(\overline{u}_0, v_0) \in \Phi(h)$.

Starting from the pair $(\overline{u}_0, v_0)$, and recalling that $v_0(y_0) < 1/2$, we can now prove, by reasoning which parallels exactly that for the pair (u_0, v_0), that (i) $\overline{v}_0$ is bounded, (ii) $\overline{v}_0$ is uniformly continuous, (iii) $\overline{v}_0 \le v_0$, and (iv) $(\overline{u}_0, \overline{v}_0) \in \Psi(h)$. We thus have $\overline{u}_0 \oplus \overline{v}_0 \le u_0 \oplus v_0$ and so

$$\begin{aligned} q(h) &\le \mu(\overline{u}_0) + \nu(\overline{v}_0) = \theta(\overline{u}_0 \oplus \overline{v}_0) \\ &\le \theta(u_0 \oplus v_0) = \mu(u_0) + \nu(v_0) < p(h) + \epsilon. \end{aligned}$$

Hence $q(h) \le p(h)$, and consequently $q(h) = p(h)$. ∎

LEMMA 5.3. *Let h be a bounded real uniformly continuous function on Z. Then the conclusion of Theorem* 5.1 *is satisfied by h.*

Proof. By the preceding lemma and Theorem 4.1 we have

$$\max_{\pi\in\Pi(\mu,\nu)} \pi(h) = p(h) = q(h). \ \blacksquare$$

In the final part of the proof we use the following well known result.

LEMMA 5.4. *Let E be a metric space with metric d and let $f : E \to [0,\infty]$ be a lower semicontinuous function. Then there exists an increasing sequence (f_n) in $\mathcal{C}^+_{\mathrm{bu}}(E)$ that has pointwise limit f.*

Outline of proof. If $f \equiv \infty$, we may take $f_n(x) = n$ for all $x \in E$ and $n \geq 1$. If $f \not\equiv \infty$, let

$$g_n(x) = \inf_{y\in E}(f(y) + nd(x,y)) \quad \text{for all } x \text{ and } n.$$

Then (g_n) is an increasing sequence of real uniformly continuous functions that has pointwise limit f, and $g_n \geq 0$ for all n. Writing $f_n = \min(n, g_n)$, we obtain a sequence (f_n) that satisfies our requirements. ∎

Equipped with Lemmas 5.3 and 5.4, we could now finish the proof of Theorem 5.1 by Villani's method (see [12, pp. 31–33]); but it is quicker for us to argue as follows. Let h satisfy the hypotheses of Theorem 5.1. We can assume that $h \leq 0$. Then, since Z is a metric space, there exists, by Lemma 5.4, a decreasing sequence (h_n) in $\mathcal{C}_{\mathrm{bu}}(Z)$ having pointwise limit h and such that $h_n \leq 0$ for all n. Then, by Lemma 5.3, $\max_{\tau\in\Pi(\mu,\nu)} \hat{h}_n(\tau) = q(h_n)$ for all n, and so, for $\pi \in \Pi(\mu,\nu)$,

$$\pi(h) \leq p(h) \leq q(h) \leq q(h_n) = \max_{\tau\in\Pi(\mu,\nu)} \hat{h}_n(\tau).$$

By the monotone convergence theorem or by [1, Theorem 1.5], the decreasing sequence $(\hat{h}_n)$ in $\mathcal{C}(\Pi(\mu,\nu))$ has pointwise limit $\hat{h}$. Hence, by Lemma 4.2,

$$\max_{\pi\in\Pi(\mu,\nu)} \pi(h) \leq p(h) \leq q(h) \leq \inf_n q(h_n) = \inf_n \max_{\tau\in\Pi(\mu,\nu)} \hat{h}_n(\tau) = \max_{\tau\in\Pi(\mu,\nu)} \hat{h}(\tau).$$

So we must have equality throughout, and in particular $\max_{\pi\in\Pi(\mu,\nu)} \pi(h) = q(h)$, as desired. ∎

6. Concluding remarks. The use of the Hahn–Banach theorem in the study of measures with prescribed marginals is not new (see for instance [4, 6, 7, 8, 9, 11, 12]). From the standpoint of linear programming, the above discussion is incomplete in that it throws no light upon the question whether the function $\Phi(h) \ni (u,v) \mapsto \mu(u) + \nu(v)$ attains its infimum. A number of authors have studied the problem. Kellerer [7, Theorem 2.21], for a much wider class of Monge–Kantorovich problems than we have considered here, shows that under appropriate conditions the infimum is attained. See also [9, 10, 13] for the case in which the topological spaces X, Y are Polish. The three

books [9, 12, 13] together give an extremely comprehensive account of the very large subject that optimal transport theory has become.

References

[1] C. Berg, J. P. R. Christensen and P. Ressel, *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions*, Grad. Texts Math. 100, Springer, New York, 1984.

[2] V. I. Bogachev, *Measure Theory*, Vol. II, Springer, Berlin, 2007.

[3] N. Bourbaki, *Elements of Mathematics*: *Integration*, Vol. II, Springer, Berlin, 2004.

[4] D. A. Edwards, *On the existence of probability measures with given marginals*, Ann. Inst. Fourier (Grenoble) 28 (1978), no. 4, 53–78.

[5] D. H. Fremlin, *Measure Theory*, Vol. IV, Torres Fremlin, Colchester, 2003.

[6] R. Haydon and V. Shulman, *On a measure-theoretic problem of Arveson*, Proc. Amer. Math. Soc. 124 (1996), 497–503.

[7] H. G. Kellerer, *Duality theorems for marginal problems*, Z. Wahrsch. Verw. Gebiete 67 (1984), 399–432.

[8] C. Léonard, *A saddle-point approach to the Monge–Kantorovich optimal transport problem*, preprint, 2007.

[9] S. T. Rachev and L. Rüschendorf, *Mass transportation problems*, Vol. I, Springer, Berlin, 1998.

[10] W. Schachermayer and J. Teichmann, *Characterization of optimal transport plans for the Monge–Kantorovich problem*, Proc. Amer. Math. Soc. 137 (2009), 519–529.

[11] V. Strassen, *The existence of probability measures with given marginals*, Ann. Math. Statist. 36 (1965), 423–439.

[12] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, Amer. Math. Soc., Providence RI, 2003.

[13] —, *Optimal Transport*, *Old and New*, Grundlehren Math. Wiss. 338, Springer, Berlin, 2009.

D. A. Edwards
Mathematical Institute
24–29 St Giles'
Oxford OX1 3LB, United Kingdom
E-mail: edwardsd@maths.ox.ac.uk

Received December 7, 2009
Revised version May 22, 2010 (6775)