The Annals of Probability
1989, Vol. 17, No. 3, 1264–1276

NOTES ON THE WASSERSTEIN METRIC IN HILBERT SPACES[1]

BY JUAN ANTONIO CUESTA AND CARLOS MATRAN

Universidad de Cantabria and Universidad de Valladolid

Let (X, Y) be a pair of Hilbert-valued random variables for which the Wasserstein distance between the marginal distributions is reached. We prove that the mapping $\omega \to (X(\omega), Y(\omega))$ is increasing in a certain sense. Moreover, if Y satisfies a nondegeneration condition, we can take $X = T(Y)$ with T monotone in the sense of Zarantarello.

We apply these results to obtain a proof of the central limit theorem (CLT) in Hilbert spaces which does not make use of the CLT for real-valued random variables.

1. Introduction. Let $\mathbb{M}$ be a Polish space and β its Borel σ-algebra. Let $\sigma: \mathbb{M} \times \mathbb{M} \to R$ be a measurable map. If P and Q are probabilities defined on β, the Wasserstein distance between P and Q with respect to σ is related, through a suitable increasing function, to the minimum value

$$\sigma(P, Q) = \inf\left\{\int \sigma \, d\lambda, \ \lambda \in M(P, Q)\right\}, \tag{1}$$

where $M(P, Q)$ is the set of all probabilities defined on $\beta \times \beta$ with marginal probabilities P and Q, respectively.

In spite of the fact that the priority in the definition of these metrics belongs to Kantorovich (1942), we use the terminology "Wasserstein metric" as do most articles on this subject. The history of this question can be read in the survey by Rachev (1984).

The interest in the Wasserstein metrics relies on the fact that convergence with respect to one of them is not exactly equivalent to the weak convergence of probability measures but, rather roughly speaking, to the weak convergence and the equiintegrability of σ [see Bickel and Freedman (1981) and Shorack and Wellner (1986)].

In this article we are concerned with the study of a pair of $\mathbb{M}$-valued random variables (X, Y) defined on a probability space (Ω, α, μ), whose distributions are, respectively, P and Q, such that the minimum in (1) is given by $\int \sigma(X, Y) \, d\mu$. Therefore, for simplicity, we use the expression in (1) as the definition of the Wasserstein distance between P and Q.

For the case where $\mathbb{M}$ is a separable Banach space and $\sigma(x, y) = \|x - y\|^p$ a proof of the existence of the pair where the minimum in (1) is reached appears in Bickel and Freedman (1981).

Received January 1988; revised August 1988.

[1]Research partially supported by "Dirección General de Investigación Científica y Técnica" of Spain with project number PB87-0905-C02-00.

AMS 1980 *subject classifications.* Primary 60E05; secondary 60B12.

Key words and phrases. Wasserstein distance, representation theorem, Hilbert spaces, central limit theorem.

For the construction of this optimum pair in the real case it is known that if $\sigma(x, y) \equiv \sigma(x - y)$, σ a convex function, and F and G are the distribution functions of P and Q, respectively, then the optimum pair is $X = F^{-1}(U)$ and $Y = G^{-1}(U)$, where U is uniformly distributed on $(0, 1)$ [see Kantorovich and Rubinstein (1958), Vallender (1973) and Major (1978)].

In the multidimensional case, to our best knowledge, the situation is quite different. It appears that the only available result is that in Rüschendorf (1985) which only treats the case where $\sigma(x, y) = \tau[h(x) - g(y)]$, τ is convex and f and g are real-valued functions. This result does not cover, for example, the important case $\sigma(x, y) = \|x - y\|^2$ for which some results are available if P and Q are p-dimensional normal random variables (r.v.'s) [Dowson and Landau (1982) and Olkin and Pukelsheim (1982)].

It is perhaps interesting to say that the article by Rüschendorf provides a construction for the optimum pair of a problem which is similar to but different from (1).

Our article intends to study the construction of the optimum pair in the case $\sigma(x, y) = \|x - y\|^2$.

First, let us study the solution in the real case. The construction in Major (1978) referred to above can be formulated in the following terms.

"Let Y be a r.v. such that its distribution function, G is continuous. If T is an increasing function such that the distribution of $T(Y)$ is P, then the optimal arrangement is obtained for

$$X = T(Y)." \tag{2}$$

A particular possible choice (in fact the only right-continuous choice) for T is $F^{-1} \circ G$, where F is the distribution function of P.

Our first intention was to extend this idea to Hilbert spaces, but it is not too difficult to see that, even in the p-dimensional case, the usual growth concept in R^p is not adequate for this problem. An alternative notion which turns out to be suitable for our purposes is that monotone operator in the sense of Zarantarello [see, e.g., Brezis (1973)]. Namely:

Given the Hilbert space $(H, \langle \cdot, \cdot \rangle)$, and a function $T: H \to H$, we say that T is increasing if $\langle T(x) - T(y), x - y \rangle \geq 0$ for every x, y in H.

This alternative definition of increasing function leads to the desired results. However, arrangements are not well defined in this case as in the real case, because the mappings verifying (2) in Hilbert spaces, or even in R^p, are not as easily obtained as in the real case. Moreover, even if we place additional continuity conditions, there is no unique map verifying (2) for two given distributions. Theorems 2.3 and 2.8 show that, under appropriate hypotheses, any representation of the optimum pair verifies (2) for an increasing T.

The usefulness of our results is shown by a proof of the central limit theorem (CLT) in Hilbert spaces which does not resort to the CLT for the real-valued case or to any result related to the Fourier transformation. On the other hand, we think that Theorem 2.3 can be useful in building an algorithm to find the optimal arrangement.

Propositions 2.9 and 2.10 exhibit some "interesting" distributions verifying the hypotheses which appear in Theorem 2.8.

Our results were inspired by those obtained by Tanaka (1973) and our proof goes along the same lines as that of this author. However, the proofs of the key steps are quite different to those employed in the real case. For instance the proof of step 1 in Tanaka would be valid in a Hilbert space if there existed only one (or perhaps a finite number) hyperplane through the origin, which is not clearly the case and we therefore need to find another technique for the proof of this step. This proof is carried out in Propositions 2.1 (which solves the problem for a special kind of simple r.v.) and 2.2 (which allows one to extend Proposition 2.1 to a general r.v. in Theorem 2.3).

Once more the proof in Tanaka for step 2 relies on the fact that R is unidimensional. In Propositions 2.5 and 2.6 we get around this difficulty.

The difference between the situation studied in the article of Tanaka and that in Hilbert spaces is also made apparent by the fact that the inverse of Theorem 2.8 is true in Tanaka's case but not in our framework, as it is shown in comment 1 following Theorem 2.8.

Our proof of the central limit theorem is similar to that of Tanaka for the real case but here we do not know whether T is continuous (or right-continuous), so that the condition $T(x + y) = T(x) + T(y)$ is not enough to guarantee that T is linear. Proposition 3.3 is included in this connection.

2. Representation theorem. From now on we denote by H a separable Hilbert space and by $\langle \cdot, \cdot \rangle$ and $\|-\|$ its inner product and norm, respectively. $X, Y, \ldots$ (with sub/super-index or not) represent H-valued random variables (r.v.'s) defined on the same probability space (Ω, α, μ), and $P_X, P_Y, \ldots$ denote their probability distributions.

The usual product probability spaces are called $(\Omega \times \Omega, \alpha \otimes \alpha, \mu \otimes \mu)$. Note that, with an abuse of notation, we can also consider the r.v.'s defined on $(\Omega \times \Omega, \alpha \otimes \alpha, \mu \otimes \mu)$.

Given P, Q, two probability measures defined on $\mathscr{B}$, the Borel σ-algebra in H, we define the Wasserstein distance, $W(P, Q)^{1/2}$, between them by

$$W(P, Q) = \inf\left\{ \int \|X - Y\|^2 \, d\mu,\ P_X = P \text{ and } P_Y = Q \right\}.$$

We have noted in the Introduction that for every pair of probabilities in β, there exist X, Y r.v.'s such that [see, e.g., Bickel and Freedman (1981)]:

(a) $P_X = P$ and $P_Y = Q$.
(b) $W(P, Q) = \int \|X - Y\|^2 \, d\mu$.

This section is devoted to obtaining some properties of this optimum pair.

PROPOSITION 2.1. *Let X, Y be a pair of simple random variables such that:*

$$\text{(3)} \qquad \begin{aligned} &\text{If } (a, b), (a', b') \text{ belong to } (X, Y)(\Omega), \text{ then} \\ &\mu\{(X, Y) = (a, b)\} = \mu\{(X, Y) = (a', b')\}. \end{aligned}$$

(Note that this condition does not entail that X and Y are uniformly distributed.)

There exist $\gamma, \delta > 0$ *such that*

$$\mu \otimes \mu\{(\omega, \omega')/\langle X(\omega) - X(\omega'), Y(\omega) - Y(\omega')\rangle < -\gamma\} \geq \delta. \tag{4}$$

Then there exist X^*, Y^* *r.v.'s such that*:

(a) $P_X = P_{X^*}$ *and* $P_Y = P_{Y^*}$.

(b) $\int \|X - Y\|^2 \, d\mu \geq \int \|X^* - Y^*\|^2 \, d\mu + \gamma \cdot \delta$.

PROOF. Let X' and Y' be independent copies of X and Y, respectively. We can choose X, X', Y and Y' defined on $(\Omega \times \Omega, \alpha \otimes \alpha, \mu \otimes \mu)$ in the standard way.

Let us denote: $\mathbb{D} = \{\langle X - X', Y - Y'\rangle < -\gamma\}$. Our r.v.'s being simple, we can obtain a finite set $\mathscr{A} = \{(a_i, b_i, c_i, d_i),\ i = 1, \ldots, k\}$ (contained in H^4), such that if we denote

$$A_i = \{X = a_i\} \cap \{X' = b_i\} \cap \{Y = c_i\} \cap \{Y' = d_i\}, \qquad i = 1, \ldots, k,$$

then $\{A_i, i = 1, \ldots, k\}$ is a partition of $\mathbb{D}$.

From (3) and (4) it is easy to derive the following properties of these sets, stated here for further reference:

(i) There exists p in $(0, 1)$ such that $\mu(A_i) = p$, for every i. From this it is obvious that $\mu \otimes \mu(\mathbb{D}) = p \cdot k \geq \delta$.

(ii) X, X', Y, Y' are constant on each A_i.

(iii) If a belongs to $X(\Omega)$ and there exist b, c and d in H such that (a, b, c, d) belongs to $\mathscr{A}$, then (b, a, d, c) also belongs to $\mathscr{A}$.

From (iii) we conclude that for each $i \in \{1, \ldots, k\}$, there exists i^* in $\{1, \ldots, k\}$, such that $a_i = b_{i^*}$, $b_i = a_{i^*}$, $c_i = d_{i^*}$ and $d_i = c_{i^*}$ and (by definition of $\mathbb{D}$) $a_i \neq b_i$ and $c_i \neq d_i$. Hence we can assure that $i \neq i^*$ and that k is an even number.

Now we are in position to define

$$(X^*, Y^*)(\omega, \omega') = \begin{cases} (X(\omega'), Y(\omega)) & \text{if } (\omega, \omega') \in \cup A_i, \\ (X(\omega), Y(\omega)) & \text{if } (\omega, \omega') \in [\cup A_i]^c. \end{cases}$$

From (iii) and this construction it is easy to verify that:

(iv) Let a in $X(\Omega)$ and i in $\{1, \ldots, k\}$ be. Then A_i is contained in $X^{-1}(a)$ if and only if A_{i^*} is contained in $[X^*]^{-1}(a)$.

It is obvious that Y and Y^* are identically distributed. We show next that X and X^* are also identically distributed.

For a in $X(\Omega)$ we can write

$$\begin{aligned} \mu\{X = a\} &= \mu\big[\{X = a\} \cap (\cup A_i)^c\big] + \sum \mu[\{X = a\} \cap A_i] \\ &= \mu\big[\{X = a\} \cap (\cup A_i)^c\big] + p \cdot \#\{i/A_i \cap \{X = a\} \neq \varnothing\} \\ &= \mu\big[\{X^* = a\} \cap (\cup A_i)^c\big] + p \cdot \#\{i^*/A_{i^*} \cap \{X^* = a\} \neq \varnothing\} \\ &= \mu\{X^* = a\}. \end{aligned}$$

To finish the proof, we need to show that (X^*, Y^*) verifies (b).

Since k is an even number we have (possibly after reordering)

$$\begin{aligned}\sum \int_{A_i} \|X - Y\|^2 \, d\mu &= \sum_{j \le k/2} \left[\int_{A_j} \|X - Y\|^2 \, d\mu + \int_{A_{j^*}} \|X - Y\|^2 : d\mu \right] \\ &= \sum_{j \le k/2} \left[\|a_j - c_j\|^2 + \|b_j - d_j\|^2 \right] \cdot p \\ &= p \cdot \sum_{j \le k/2} \left[\|b_j - c_j\|^2 + \|a_j - d_j\|^2 - 2 \cdot \langle a_j - b_j, c_j - d_j \rangle \right] \\ &\ge p \cdot k \cdot \gamma + \sum_{j \le k/2} \left[\int_{A_j} \|X^* - Y^*\|^2 \, d\mu + \int_{A_{j^*}} \|X^* - Y^*\|^2 \, d\mu \right] \\ &= \delta \cdot \gamma + \sum \int_{A_i} \|X^* - Y^*\|^2 \, d\mu.\end{aligned}$$

And finally

$$\begin{aligned}\int \|X - Y\|^2 \, d\mu &= \int_{(\cup A_i)^c} \|X - Y\|^2 \, d\mu + \sum_i \int_{A_i} \|X - Y\|^2 \, d\mu \\ &\ge \int_{(\cup A_i)^c} \|X^* - Y^*\|^2 \, d\mu + \sum_i \int_{A_i} \|X^* - Y^*\|^2 \, d\mu + \delta \cdot \gamma \\ &= \int \|X^* - Y^*\|^2 \, d\mu + \delta \cdot \gamma. \qquad \square\end{aligned}$$

The following proposition is useful in extending to a general r.v. the previous one. Note that it is valid in every Hilbert space, in particular in $H \times H$.

PROPOSITION 2.2. *Let X be a r.v. with values in a Hilbert space such that $\int \|X\|^2 \, d\mu < \infty$. Then there exists a sequence, $\{X_n\}$, of simple r.v.'s [possibly defined on a different probability space (Ω', α', μ')] such that:*

(a) *X_n is uniformly distributed on a finite set $M_n \subset H$.*
(b) *$\{X_n\}$ converges in distribution to X.*
(c) *$\int \|X_n\|^2 \, d\mu' \to \int \|X\|^2 \, d\mu$ as $n \to \infty$.*

PROOF (Sketched). Other alternative proofs can be given, but it suffices to consider a "good" sequence of empirical distributions (from the Glivenko–Cantelli theorem in Banach spaces and the strong law of large numbers, almost every sequence of empirical distributions is "good") and one obtains the result by making slight changes to avoid repetitions. □

Now, as indicated in the Introduction, we apply the preceding propositions to prove that if (X, Y) is the optimal pair, then this pair can be seen as an increasing map.

THEOREM 2.3. *Let X, Y be two r.v.'s such that*

$$\int \|X - Y\|^2 \, d\mu = W(P_X, P_Y).$$

Then

$$\mu \otimes \mu\{(\omega, \omega')/\langle X(\omega) - X(\omega'), Y(\omega) - Y(\omega')\rangle < 0\} = 0.$$

PROOF. Let $\{X_n, Y_n\}$ be a sequence of simple r.v.'s obtained by applying the preceding proposition to the r.v. (X, Y) in the Hilbert space $H \times H$ and let X', Y' and X_n', Y_n' be independent copies of the corresponding r.v.'s in H.

Without loss of generality, in the sequel, we can assume that all the r.v.'s are defined on the same probability space.

We denote $G = \langle X - X', Y - Y'\rangle$ and $G_n = \langle X_n - X_n', Y_n - Y_n'\rangle$.

If the theorem is not true, then there exist δ in R^+ and $-\gamma$ in the continuity set of the distribution function of the (real-valued) r.v. G such that $\mu \otimes \mu\{G < -\gamma\} > \delta$.

Note that the sequence $\{G_n\}$ converges in distribution to G. Therefore there exists a natural number, N_0, such that if $n \geq N_0$, then $\mu \otimes \mu\{G_n < -\gamma\} > \delta$. So we are in a situation to apply Proposition 2.1 for each n. Let $\{X_n^*, Y_n^*\}$ be the sequence obtained.

As the r.v. X_n^* (resp. Y_n^*) and X_n (resp. Y_n), are identically distributed, then

$$X_n^* \to X \quad \text{and} \quad Y_n^* \to Y \quad \text{in distribution.} \tag{5}$$

Hence the sequence of distributions of the r.v.'s $\{(X_n^*, Y_n^*)\}$ is tight. Therefore there exists a subsequence, which we still denote with the same notation, which converges in distribution to some r.v. (X^*, Y^*).

The continuous mapping theorem implies that the sequence $\{\|X_n^* - Y_n^*\|\}$ converges in distribution to the r.v. $\|X^* - Y^*\|$. From this and a well-known property of weak convergence [see, for example, Billingsley (1968), page 32], we have

$$\int \|X^* - Y^*\|^2 \, d\mu \leq \liminf \int \|X_n^* - Y_n^*\|^2 \, d\mu$$

$$\leq \liminf \int \|X_n - Y_n\|^2 \, d\mu - \gamma \cdot \delta = \int \|X - Y\|^2 \, d\mu - \gamma \cdot \delta.$$

But (5) and, once more, the continuous mapping theorem, imply that the r.v.'s X^* and X (resp. Y^* and Y) are identically distributed, which contradicts that $W(P_X, P_Y) = \int \|X - Y\|^2 \, d\mu$. □

The previous theorem suggests that we are on the right path to prove that if the pair (X, Y) is the optimal one, then it is possible to write each r.v. as an increasing function of the other one. But the following example shows that this is not possible in general even in the R^2 case.

EXAMPLE 2.4. Let P and Q be the uniform distributions in $[0,1] \times [0,1]$ and in $\{0\} \times [0,1]$, respectively.

If the distribution of (X_1, X_2) [resp. (Y_1, Y_2)] is P (resp. Q), then their components are independent. So the distance between them is minimum if and only if the same thing happens between their components. But this is attained if we choose (X_1, X_2, Y_1, Y_2) in such a way that Y_1 is the constant zero, Y_2 has a uniform distribution and the conditional distribution of (X_1, X_2) given $(Y_1, Y_2) = (0, y)$ is the uniform distribution in $[0,1] \times \{y\}$.

Evidently it is impossible to write (X_1, X_2) as a function of (Y_1, Y_2).

Therefore we have a simple example in which it is not possible to write each r.v. as a function of the other. However, this pair exhibits a kind of degeneration. If we exclude these degenerate cases, we obtain the existence of the desired functions.

PROPOSITION 2.5. *Let P be a probability measure on β such that there exists a complete orthonormal system, $\{V_n\}$, such that for each n and almost everywhere ω in $\langle V_n \rangle^+$ (the orthogonal subspace to V_n) the regular conditional probability on $\langle V_n \rangle$ (the subspace generated by V_n) given ω is atomless.*

Now let A in β such that $\mu(A) > 0$.

Then for every n, there exists y_n in A such that the set $\{r \in R / y_n + r \cdot V_n \in A\}$ is nondenumerable infinite.

PROOF. Let n be a natural number. Since $H = \langle V_n \rangle \oplus \langle V_n \rangle^+$ we have

$$\mu(A) = \int_{\langle V_n \rangle^+} P[A_\omega / \omega] \mu_1(d\omega),$$

where A_ω is the section of A on ω, $P[\cdot / \omega]$ is the regular conditional probability on $\langle V_n \rangle$ given ω and μ_1, is the marginal probability on $\langle V_n \rangle^+$.

But the atomless character of $P[\cdot / \omega]$ and $\mu(A) > 0$ imply that there exists ω such that A_ω is an infinite, nondenumerable set. Or, in other words,

$$\#\{r / \omega + r \cdot V_n \in A\} \geq \chi_1$$

and now the result is immediate. □

Our representation theorem is a simple consequence of the following proposition. We first give some additional notation.

Given y in H and the r.v. (on H^2) (X, Y), we denote by $S_y(X)$ (or S_y if no confusion is possible) the support of the regular conditional probability of X given $Y = y$.

If x, v are in H, $v \neq 0$, let $\Pi_v(x) = \langle x, v \rangle / \|v\|$ the coordinate of the projection of x on $\langle v \rangle$. Finally, $I_v(y)$ denotes the smallest interval (of real numbers, of course) containing the numbers $\Pi_v(x)$ for x in S_y.

PROPOSITION 2.6. *Let (X, Y) be a r.v. (on H^2) such that $\int \|X - Y\|^2 \, d\mu = W(P_X, P_Y)$ and let us suppose that P_Y verifies the hypothesis in Proposition 2.5.*

Then there exists A in β such that $P_Y(A) = 1$ and for every y in A, v in H,

$$\langle x - x', v\rangle = 0 \quad \text{for every } x, x' \in S_y.$$

PROOF. Let $\{V_n\}$ be the complete orthonormal system in Proposition 2.5 and let n be a natural number. We set

$$A_n = \{y: \text{there exists } x, x' \in S_y: \langle x - x', V_n\rangle \neq 0\}.$$

As a first stage, we prove that $P_Y(A_n) = 0$.

From Theorem 2.4 and a well-known property of regular conditional probabilities [see, e.g., Parthasarathy (1967), Theorem 8.1], we can suppose that (possibly after redefining these probabilities on a set of P_Y-probability zero)

(6) For every y, y' in H: $S_y \times S_{y'} \subset \{(x, x')/\langle x - x', y - y'\rangle \geq 0\}$.

Now let y be in A_n. We call $D_y = \{y' \in A_n/y' = y + r \cdot V_n, r \in R\}$.

By the preceding proposition there exists y^* such that D_{y^*} is a nondenumerable infinite set.

But, by definition, if $y \in A_n$, then there exist x, x' in S_y such that $\Pi_{Vn}(x) \neq \Pi_{Vn}(x')$. Therefore there exist y, y' in D_{y^*}, $y \neq y'$ such that the interior of $I_v(y) \cap I_v(y')$ is not empty. Or, in other words, there exist x_1, x_2 in S_y and x' in $S_{y'}$ (or vice versa) such that

$$\Pi_{Vn}(x_1) < \Pi_{Vn}(x') < \Pi_{Vn}(x_2).$$

Moreover, y and y' are in D_{y^*} so that there exists $\delta \neq 0$ such that $y - y' = \delta \cdot V_n$ whence

$$\langle x_1 - x', y - y'\rangle = [\Pi_{Vn}(x_1) - \Pi_{Vn}(x')] \cdot \delta,$$
$$\langle x_2 - x', y - y'\rangle = [\Pi_{Vn}(x_2) - \Pi_{Vn}(x')] \cdot \delta$$

and one of these quantities is strictly negative which is not possible by (6). Therefore $P_Y(A_{Vn}) = 0$.

Finally, if $A = \bigcap_n A_n$, it is trivial that A verifies the proposition. □

From the preceding proposition one immediately obtains the following corollary.

COROLLARY 2.7. *Let (X, Y) be r.v.'s such that $\int \|X - Y\|^2 \, d\mu = W(P_X, P_Y)$ and suppose that P_Y verifies the hypothesis in Proposition 2.5. Then*

$$P_Y\{y/\#S_y = 1\} = 1.$$

From this corollary it is clear that we can define the sought mapping. More precisely, from this corollary and Theorem 2.3 we conclude:

THEOREM 2.8 (Representation). *Let P, Q be two probabilities defined on β and suppose that Q verifies the conditions in Proposition 2.5. Let Y be a r.v. such that $P_Y = Q$.*

Then there exists a mapping $T: H \to H$ *such that*:

(a) *For every* y, y' *in* H: $\langle T(y) - T(y'), y - y' \rangle \geq 0$.
(b) $P_{T(Y)} = P$.
(c) $\int \|T(Y) - Y\|^2 \, d\mu = W(P, Q)$.

Before we close this section, we would like to make two remarks.

1. In the real case ($H = R$) every application verifying (a) and (b) in the last theorem verifies also (c) but this is not so in the general situation.

Take Y to be a r.v. with normal distribution (in R^2) with covariance matrix identity. Then if T is a rotation of less than 90° it is clear that T verifies (a) and (b) but it is not necessarily true that $\int \|T(Y) - Y\|^2 \, d\mu = 0$.

2. An important question is that of the existence of some "interesting" distributions verifying the condition in Proposition 2.5. Now we prove that this is the case with every distribution in R^n which is absolutely continuous with respect to the Lebesgue measure and with every Gaussian distribution on a Hilbert space.

More precisely, we prove that every distribution in R^n which is absolutely continuous with respect to the Lebesgue measure verifies the conclusions in Proposition 2.5, and then, trivially, for such distributions there would hold a proof similar to that given above for Theorem 2.8.

PROPOSITION 2.9. *Let us suppose that* $H = R^p$ *and that* A *is in* β *with* $\lambda_p(A) > 0$. *Let* V *in* $R^p - \{0\}$. *Then there exists* y *in* A *such that* $\{r \in R / y + r \cdot V \in A\}$ *is a nondenumerable infinite set.*

PROOF. Let V in $R^p - \{0\}$; we can suppose w.l.o.g. that $\|V\| = 1$. Let us suppose that the proposition is not true.

Let z in R^p, we set $L_z = \{r \in R / z + r \cdot V \in A\}$.

By Fubini's theorem we have

$$\lambda_p(A) = \int_{\langle V \rangle^+} \lambda_1(L_x) \, dx \tag{7}$$

(where λ_1 is the Lebesgue measure in R).

But if $L_x \neq \varnothing$, then there exists y_0 in A such that to each λ in R we can associate a number λ^* in R such that $x + \lambda \cdot V = y_0 + \lambda^* \cdot V$ and this map is injective.

But, by assumption, since y is in A, L_y is a denumerable set. Therefore also L_x is a denumerable set for every x in $\langle V_n \rangle^+$, hence $\lambda_1(L_x) = 0$ for every x in $\langle V_n \rangle^+$.

From the last fact and (7) we, finally, obtain the contradiction $\lambda_p(A) = 0$. □

PROPOSITION 2.10. *Let* P *a Gaussian distribution nondegenerated in a point. Then there exist a subspace,* F, *such that* $P(F) = 1$ *and a complete orthonormal*

system in F, $\{V_n\}$, such that for each n and for every ω in $\langle V_n\rangle^+$ the conditional probability on $\langle V_n\rangle$ given ω is atomless.

PROOF. Let S be the covariance operator of P and $\{e_n\}$, a complete orthonormal system formed with eigenvectors of S, $\{\lambda_n\}$ being the sequence of their associated eigenvalues [see, for example, Vakhania (1981)].

Note that if $\lambda_n = 0$, then $P[\langle e_n\rangle^+] = 1$.

Therefore, if $F = \bigcap_{n/\lambda_n=0}\langle e_n\rangle^+$ it is $P(F) = 1$.

By construction F is the closed subspace generated by the eigenvectors associated with nonzero eigenvalues of S and we can take a complete orthonormal system for F, $\{V_n\}$, formed by the vectors in $\{e_n\}$ associated to these eigenvalues.

The coordinates, $\{x_n\}$, in F are independent normal variables with variance λ_n^2, $\lambda_n \neq 0$. Then, for almost everywhere ω in $\langle V_n\rangle^+$ the conditional distribution of x_n given ω is the same as of x_n which is nondegenerated normal and, therefore, atomless. □

Moreover, from the preceding proof one obtains a result which will be of interest in Section 3.

PROPOSITION 2.11. *Let P be a probability distribution with covariance operator S and let Q be a Gaussian distribution with the same covariance operator.*

Then there exists a subspace, F, which only depends on S such that $P(F) = Q(F) = 1$ and Q verifies the conditions of Proposition 2.5 in F.

3. Application: The central limit theorem. As we indicated in the Introduction, our proof of the CLT follows essentially that developed in Tanaka (1973) for the real case. Therefore we only describe the necessary adjustments of Tanaka's proof.

The general idea is to use Theorem 2.8 to prove that the Wasserstein distance between the sequence of partial sums and a Gaussian distribution converges to 0.

If P (resp. Y) is a distribution (resp. r.v.) such that $\int \|x\|^2\, dP$ is finite, we denote by $e(P)$ [resp. $e(Y)$] the Wasserstein distance between P (resp. P_Y) and the Gaussian distribution with the same vector of means and covariance operator as P.

Moreover, by Proposition 2.11, we can suppose that P (or Y) is defined on a Hilbert space where it is possible to use the representation obtained in Theorem 2.8 for P (or Y) in terms of a Gaussian distribution.

Note that if $\{X_n\}$ is a sequence of independent identically r.v.'s (i.i.d.r.v.'s) the same space that for X_1 is valid for the whole sequence $\{(X_1 + \cdots + X_n)/n^{1/2}\}$.

Hence from now on we can suppose that P (or Y) is defined on this space and we will make reference to the representation of P (or Y) in terms of this Gaussian distribution, simply, as the representation of P (or, of Y).

The theorem we want to prove is the following.

THEOREM 3.1. *Let $\{X_n\}$ be a sequence of i.i.d.r.v.'s such that they are centered and $E\|X_1\|^4 < \infty$. Let S be their covariance operator and let Q be a centered Gaussian distribution with S as covariance operator. Then the sequence $\{(X_1 + \cdots + X_n)/(n^{1/2})\}$ converges in distribution to Q.*

To outline the proof of this theorem (in the real case), let us denote by T_n the r.v. $(X_1 + \cdots + X_n)/(n^{1/2})$.

The condition $E\|X_1\|^4 < \infty$ is used to prove the equiintegrability of the function $\|x\|^2$ with respect to the sequence $\{P_{T_n}\}$ and, therefore can be weakened, but we retain it to follow the line in Tanaka.

This equiintegrability guarantees the existence of some subsequence of $\{T_n\}$ which converges in distribution to some r.v. Y which, in turn, verifies that if Z is a independent copy of Y, then

$$e\left[(Y + Z)/2^{1/2}\right] = e(Y). \tag{8}$$

Now it is proved that every distribution verifying (8) is normal.

Once more the equiintegrability of $\|x\|^2$ implies that Y is centered with covariance operator S.

Finally, it is not too difficult to prove that the preceding can be extended to the whole sequence.

This proof can be developed in the Hilbert case with some minor technical complications excepting the following.

In the outlined proof the next result (which is proved for Hilbert spaces in a similar way to that for the real case) has an important role.

PROPOSITION 3.2. *Let X, Y be i.i.d.r.v.'s such that $E\|X\|^2$ is finite and*

$$e\left[(X + Y)/2^{1/2}\right] = e(X).$$

Then the representation of X (in terms of the corresponding Gaussian) can be chosen such that

$$T(x + y) = T(x) + T(y) \quad \text{for every } x, y \text{ in } H.$$

Note that this equality, by itself, does not guarantee the linearity of T except if the coefficients are rationals.

In the real case every increasing function satisfying (b) in Theorem 2.8 can be taken as a representation for X. Then we can choose the right-continuous one and then the linearity of T is an easy consequence of Proposition 3.2.

But for Hilbert spaces this cannot be carried out because we do not know whether we can choose T with some continuity property. Then it is necessary to prove the following proposition in which we obtain the linearity of T from Proposition 3.2 and its growth.

PROPOSITION 3.3. *Let $T: H \to H$ be a map such that $\langle T(x) - T(y), x - y\rangle \geq 0$ and $T(X + y) = T(x) + T(y)$; for every x, y in H. Then T is linear.*

PROOF. We must only prove that $T(\delta x) = \delta T(x)$ for every δ in R, x in H.

Let us suppose that there exist x in H and δ in R such that $T(\delta x) \neq \delta T(x)$.

By standard techniques it is possible to prove that x is not necessarily zero and δ is an irrational number.

Let us denote V to be the vector $T(\delta x) - \delta T(x)$ and we prove, at a first stage, that $\langle V, x\rangle = 0$.

If $\langle V, x\rangle \neq 0$, we develop the proof in the case $\langle V, x\rangle > 0$, the other case being analogous.

Let $\{\delta_n\}$ be a sequence of rational numbers such that $\delta_n \downarrow \delta$. Now

$$\lim_n \langle T(\delta x) - T(\delta_n x), \delta x - \delta_n x\rangle = \lim_n (\delta - \delta_n)\langle T(\delta x) - \delta_n T(x), x\rangle.$$

Since $\delta - \delta_n < 0$, and $\lim_n [T(\delta x) - \delta_n T(X)] = V$, it is evident that from some index forward these scalar products are negative which is not possible from the hypotheses.

Thus $\langle V, x\rangle = 0$.

Now let $\{\delta_n\}$ and $\{\gamma_n\}$ be two sequences of rational numbers such that both $(\delta_n - \delta)$ and γ_n are of the same order than $1/n$ and, for every n, δ_n is a positive number.

We denote x_n to be the vector $\delta_n \cdot x + \gamma_n V$.

Recall that T is additive and that $T(ax) = aT(x)$ for a rational.

$$\begin{aligned}\langle T(x_n) - T(\delta x), x_n - \delta x\rangle &= \langle T(x_n) - \delta T(x) + \delta T(x) - T(\delta x), x_n - \delta x\rangle \\ &= \langle (\delta_n - \delta)T(x) + \gamma_n T(V) - V, (\delta_n - \delta)x + \gamma_n V\rangle \\ &= o(1/n) - 0 - \gamma_n\langle V, V\rangle,\end{aligned}$$

which is negative (if V is not zero) for large n. But this is not possible by assumption. Therefore we have $V = 0$. □

So we have proved that in the conditions of the preceding proposition the representation of P_X in terms of a Gaussian distribution is linear and, therefore, P_X is also Gaussian.

From this point the proof parallels that of the real case.

Acknowledgment. The authors would like to thank a referee for pointing out the articles of Kantorovich (1942), Kantorovich and Rubinstein (1958) and Rachev (1984).

REFERENCES

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BREZIS, H. (1973). *Operateurs maximaux monotones*. North-Holland, Amsterdam.

DOWSON, D. C. and LANDAU, B. V. (1982). The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.* **12** 450–455.

KANTOROVICH, L. V. (1942). On the transfer of masses. *Dokl. Akad. Nauk SSSR* **37** 7–8.

KANTOROVICH, L. V. and RUBINSTEIN, G. SH. (1958). On a space of completely additive functions. *Vestnik Leningrad Univ. Math.* **13** 52–59.

MAJOR, P. (1978). On the invariance principle for sums of independent identically distributed random variables. *J. Multivariate Anal.* **8** 487–517.

OLKIN, I. and PUKELSHEIM, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.* **43** 257–263.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. Academic, New York.

RACHEV, S. T. (1984). The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory Probab. Appl.* **29** 647–676.

RÜSCHENDORF, L. (1985). The Wasserstein distance and approximation theorems. *Z. Wahrsch. verw. Gebiete* **70** 117–129.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

TANAKA, H. (1973). An inequality for a functional of probability distributions and its application to Kac's one-dimensional model of a Maxwellian gas. *Z. Wahrsch. verw. Gebiete* **27** 47–52.

VAKHANIA, N. N. (1981). *Probability Distributions on Linear Spaces*. North-Holland, Amsterdam.

VALLENDER, S. S. (1973). Calculation of the Wasserstein distance between distributions on the line. *Theory Probab. Appl.* **18** 784–786.

DEPARTAMENTO DE MATEMÁTICAS,
ESTADÍSTICA Y COMPUTACIÓN
FACULTAD DE CIENCIAS
UNIVERSIDAD DE CANTABRIA
39005 SANTANDER
SPAIN

DEPARTAMENTO DE ESTADÍSTICA
E INVESTIGACIÓN OPERATIVA
FACULTAD DE CIENCIAS
UNIVERSIDAD DE VALLADOLID
47002 VALLADOLID
SPAIN