Z. Wahrscheinlichkeitstheorie verw. Gebiete
70, 117-129 (1985)

Zeitschrift für
**Wahrscheinlichkeitstheorie**
und verwandte Gebiete


# The Wasserstein Distance and Approximation Theorems

Ludger Rüschendorf

Institut für Mathematische Stochastik, Universität Freiburg, Hebelstr. 27, D-7800 Freiburg, Federal Republic of Germany

**Summary.** By an extension of the idea of the multivariate quantile transform we obtain an explicit formula for the Wasserstein distance between multivariate distributions in certain cases. For the general case we use a modification of the definition of the Wasserstein distance and determine optimal 'markov-constructions'. We give some applications to the problem of approximation of stochastic processes by simpler ones, as e.g. weakly dependent processes by independent sequences and, finally, determine the optimal martingale approximation to a given sequence of random variables; the Doob decomposition gives only the 'one-step optimal' approximation.

## 1. Calculation of the Wasserstein Distance

For a polish space $(M, \mathfrak{A})$ with Borel $\sigma$-algebra $\mathfrak{A}$ and a non-negative, product-measurable function $\sigma$: $M \times M \to R$ define the (generalized) Wasserstein metric w.r.t. $\sigma$ for probability measures $P$, $Q$ on $\mathfrak{A}$ by:

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

where $M(P, Q)$ is the set of probability measures on $\mathfrak{A} \otimes \mathfrak{A}$ with marginals $P$, $Q$. There are good historical reasons to call $\sigma(P, Q)$ the Kantorovic, Rubinstein distance (cf. the survey article of Zolotarev (1982)) but we would like to follow the notation "Wasserstein distance" as is done in the most papers on coupling of distributions.

There are not many explicit results for the determination of the Wasserstein distance $\sigma(P, Q)$. If $M = R^1$ and $\sigma(x, y) = |x - y|$, then $\sigma(P, Q) = \int |F(x) - G(x)|\, d\lambda^1(x) = \int |F^{-1}(u) - G^{-1}(u)|\, du$ where $F$, $G$ are the $df$s of $P$, $Q$ (cf. Dall'Aglio 1956; Kantorovic and Rubinstein 1958; Vallander 1973). If for general $M$ $\sigma$ is the discrete metric and is measurable, then $\sigma(P, Q) = \sup\{|P(A) - Q(A)|, A \in \mathfrak{A}\} = \|P - Q\|$, so $\sigma$ is up to a factor 1/2 the total variation distance (cf. Dobrushin 1970). For multivariate normal distributions and $\sigma(x, y) = |x - y|^2$,

$x, y \in R^n$, the Wasserstein distance was calculated by Dowson and Landau (1982) and Olkin and Pukelsheim (1982). Furthermore, there is the well-known connection between the Prohorov distance and the Wasserstein metric due to Strassen. For some related results concerning Levy-type and Hausdorff-type distances we refer to Rachev (1982) and Zolotarev (1982).

Let now $(M, \mathfrak{A}) = (R^n, \mathfrak{B}^n)$, let $h: R^n \to R^m$ be measurable and let $P \in M^1(R^n, \mathfrak{B}^n)$ – the set of distributions on $(R^n, \mathfrak{B}^n)$ – have the *df.* $F$. Let, furthermore, $S, U_1, \ldots, U_n$ be independent random variables on a space $(M', \mathfrak{A}', R)$ such that $S$ and $h$ have identical distributions, i.e. $R^S = P^h$ and the $U_i$ are $R(0,1)$-distributed, $R(0,1)$ denoting the uniform distribution on $(0,1)$. With $F_i(x_i | x_1, \ldots, x_{n-1}, s)$ we denote a regular conditional *df* w.r.t. $F$ of the $i$-th component given the first $i-1$ components are $x_1, \ldots, x_{i-1}$ and given the condition $h = s$, in other words

$$F_i(x_i | x_1, \ldots, x_{i-1}, s) = P^{\pi_i | \pi_1 = x_1, \ldots, \pi_{i-1} = x_{i-1}, h = s}(-\infty, x_i],$$

where $\pi_i$: $R^n \to R$ denotes the $i$-th projection, $1 \leqq i \leqq n$. Let $H^{-1}(u)$ be the generalized inverse of a right-continuous *df* $H$, i.e. $H^{-1}(u) = \inf\{y, H(y) \geqq u\}$.

The following construction generalizes the multivariate quantile transform. Define inductively the vector $X = (X_1, \ldots, X_n)$ by:

$$X_1 = F_1^{-1}(U_1 | S), \quad X_2 = F_2^{-1}(U_2 | X_1, S), \ldots, X_n = F_n^{-1}(U_n | X_1, \ldots, X_{n-1}, S). \quad (2)$$

**Proposition 1.** *The random variable $X$ on $(M', \mathfrak{A}', R)$ has the following properties:*

a) $R^X = P$ b) $h(X) = S\,[R]$

*Proof.* By our independence assumption

$$R^{X_1 | S = s} = R^{F_1^{-1}(U_1 | S) | S = s} = R^{F_1^{-1}(U_1 | s)} = P^{\pi_1 | h = s}.$$

Similarly,

$$R^{X_2 | X_1 = x_1, S = s} = R^{F_2^{-1}(U_2 | X_1, S) | X_1 = x_1, S = s} = R^{F_2^{-1}(U_2 | x_1, s)} = P^{\pi_2 | \pi_1 = x_1, h = s},$$

implying that

$$R^{(X_1, X_2) | S = s}(A \times B) = \int_A R^{X_2 | x_1, s}(B)\, dR^{X_1 | s}(x_1)$$

$$= \int_A P^{\pi_2 | \pi_1 = x_1, h = s}(B)\, dP^{\pi_1 | h = s}(x_1) = P^{(\pi_1, \pi_2) | h = s}(A \times B).$$

Inductively, we obtain $R^{X | S = s} = P^{\pi | h = s}$. Therefore,

$$R^X(A) = \int R^{X | S = s}(A)\, dR^S(s) = \int P^{\pi | h = s}(A)\, dP^h(s) = P(A), \qquad A \in \mathfrak{B}^n.$$

Since almost surely w.r.t. $P^h$ holds $P^{\pi | h = s}\{x; h(x) = s\} = 1$, we obtain $R^{X | S = s}\{x; h(x) = s\} = 1\ [R^S]$ and so

$$R\{h(X) = S\} = \int R^{X | S = s}\{x : h(x) = s\}\, dR^S(s) = 1. \quad \square$$

Returning to the Wasserstein distance let

$$h, g : (R^n, \mathfrak{B}^n) \to (R^m, \mathfrak{B}^m), \ \varphi : (R^{2m}, \mathfrak{B}^{2m}) \to (R_+, \mathfrak{B}_+)$$

and $\sigma(x, y) = \varphi(h(x), g(y))$, $x, y \in R^n$.

**Theorem 2.** *For $P, Q \in M^1(R^n, \mathfrak{B}^n)$ and $\sigma(x,y) = \varphi(h(x), g(y))$, $x, y \in R^n$, holds:*

a) $\sigma(P,Q) = \varphi(P^h, Q^g)$

b) *If $m=1$ and $F_h, G_g$ are the df's of $P^h, Q^g$ and $\varphi(h,g) = \varphi(h-g)$, $\varphi$ convex, then $\sigma(P,Q) = \int_0^1 \varphi(F_h^{-1}(u) - G_g^{-1}(u))\,du$.*

*Proof.* a) Clearly, $\sigma(P,Q) \geqq \varphi(P^h, Q^g)$. Conversely, let $S, \tilde{S}$ be random variables on a probability space (which is rich enough) with distributions $P^h$, $Q^g$. By Proposition 1, we can construct random variables $X \sim P$, i.e. $X$ has distribution $P$, and $Y \sim Q$ such that almost surely $h(X) = S$ and $g(Y) = \tilde{S}$. Since $E\sigma(X,Y) = E\varphi(h(X), g(Y)) = E\varphi(S, \tilde{S})$, we get the converse direction.

b) Follows from a) and a well-known one dimensional coupling result (cf. Cambanis et al. 1976; Major 1978; Rüschendorf 1983. □

*Remark.* a) Theorem 2 could be proved for more general spaces, since only essential use has been made upon regular conditional distributions. But no explicit construction of random variables could be given in this case.

b) A similar idea as in Theorem 2b) is implicitely contained in the paper of Major (1978) for $h(x) = g(x) = \sum_{i=1}^{n} x_i$. □

Generally, one can not expect explicit results for the Wasserstein metric since its determination leads already in the most simple discrete cases to a difficult and unsolved rearrangement problem. We consider, therefore, the following modification of the definition, allowing to use inductive arguments.

Let $(M, \mathfrak{A}) = (M_1, \mathfrak{A}_1) \otimes (M_2, \mathfrak{A}_2)$ and let $P, Q \in M^1(M, \mathfrak{A})$ with factorization $P = P_x \times P_1$, $Q = Q_y \times Q_1$, where $P_1, Q_1$ are the marginals on $\mathfrak{A}_1$ and $P_x$, $Q_y$ are (fixed) conditional distributions.

Define the following subclass $M_{1,2}(P,Q)$ of $M(P,Q)$:

$$\begin{aligned} &M_{1,2}(P,Q) = \{R^{(X,Y)};\ R^{(X_1,Y_1)} \in M(P_1, Q_1), \\ &R^{(X_2,Y_2)|x_1,y_1} \in M(P_{x_1}, Q_{y_1}),\ x_1, y_1 \in M_1\} = M(P_x, Q_y) \times M(P_1, Q_1). \end{aligned} \tag{3}$$

So $Y_2$ is conditionally independent of $X_1$ given $Y_1$ and $X_2$ is conditionally independent of $Y_1$ given $X_1$. For this reason we call elements of $M_{1,2}(P,Q)$ markov-constructions. Clearly, this definition extends to higher products of spaces.

Define for $\sigma\colon M \times M \to R_+$

$$\sigma_{1,2}(P,Q) = \inf\{\textstyle\int \sigma\,d\lambda;\ \lambda \in M_{1,2}(P,Q)\},$$

and the section of $\sigma$ in $(x_1, y_1)$ as

$$\sigma_{x_1,y_1}(x_2, y_2) = \sigma((x_1, x_2), (y_1, y_2)).$$

**Theorem 3.** *For $\lambda = \lambda_{(x,y)} \times \mu \in M_{1,2}(P,Q)$ holds: $\sigma_{1,2}(P,Q) = \int \sigma\,d\lambda < \infty$ if and only if*

a) *$h(x,y) = \int \sigma_{x,y}\,d\lambda_{(x,y)} = \sigma_{x,y}(P_x, Q_y) < \infty$ for $\mu$ almost all $(x,y) \in M_1 \times M_1$*

b) *$h(P_1, Q_1) = \int h\,d\mu < \infty$.*

*Proof.* If $\lambda \in M_{1,2}(P,Q)$ satisfies a), b), then for any $\tilde{\lambda}=\tilde{\lambda}_{(x,y)} \times \tilde{\mu} \in M_{1,2}(P,Q)$ holds

$$\begin{aligned}\int \sigma\, d\lambda &= \int \left(\int \sigma_{x,y}\, d\lambda_{(x,y)}\right) d\mu(x,y)\\ &= \int h\, d\mu \leqq \int h\, d\tilde{\mu} \leqq \int \left(\int \sigma_{x,y}\, d\tilde{\lambda}_{(x,y)}\right) d\tilde{\mu}(x,y) = \int \sigma\, d\tilde{\lambda}.\end{aligned}$$

Let now, conversely, $\lambda \in M_{1,2}(P,Q)$ satisfy $\sigma_{1,2}(P,Q)=\int \sigma\, d\lambda$. Define the function $T$ from $M_1 \times M_1$ into the compact, convex subsets of $M^1(M_2 \times M_2, \mathfrak{A}_2 \otimes \mathfrak{A}_2)$ by $T(x,y)=M(P_x, Q_y)$. $T$ defines a multifunction whose graph belongs to $\hat{\mathfrak{A}}_1 \otimes \hat{\mathfrak{A}}_1 \otimes \mathfrak{B}$, where $\hat{\mathfrak{A}}_1$ is the universal completion of $\mathfrak{A}_1$ and $\mathfrak{B}$ is the Borel $\sigma$-algebra on $M^1(M_2 \times M_2, \mathfrak{A}_2 \otimes \mathfrak{A}_2)$ supplied with weak topology (cf. Th. III.30 of Castaing and Valadier 1977) and observe that $T$ is lower semicontinuous). Therefore, by Lemma III.39 of Castaing and Valadier (1977) there exists a markov kernel $\tilde{\lambda}$ from $(M_1 \times M_1, \hat{\mathfrak{A}}_1 \otimes \hat{\mathfrak{A}}_1)$ to $(M_2 \times M_2, \mathfrak{A}_2 \otimes \mathfrak{A}_2)$ such that

$$\int \sigma_{x,y}\, d\tilde{\lambda}_{(x,y)} = \sigma_{x,y}(P_x, Q_y) \qquad \text{for all } (x,y) \in M_1 \times M_1.$$

If a) and b) would not hold true, we could construct a measure $\tilde{\lambda}=\tilde{\lambda}_{(x,y)} \times S$ on $\hat{\mathfrak{A}}_1 \otimes \mathfrak{A}_2 \otimes \hat{\mathfrak{A}}_1 \otimes \mathfrak{A}_2$ with $\int \sigma\, d\tilde{\lambda} < \int \sigma\, d\lambda = \sigma_{1,2}(P,Q)$. With $\lambda^*$ – the restriction of $\tilde{\lambda}$ on $\mathfrak{A} \otimes \mathfrak{A}$ – we would obtain a contradiction. □

The idea of Theorem 3 also works under certain additional restrictions which are motivated by strong approximation results (cf. Schwarz 1980, Lemma 2).

Let e.g. $P_1 = Q_1$ and let

$$\tilde{M}(P,Q)=\{\lambda \in M(P,Q);\ \lambda\{\pi_1=\pi_3\}=1\}. \tag{5}$$

$\pi_i$, $i=1,3$, denoting the projections on the $i$'th components of $M_1 \times M_2 \times M_1 \times M_2$; $\tilde{M}(P,Q) \subset M_{12}(P,Q)$.

**Proposition 4.** *Let $\lambda=\lambda_{(x,x)} \times \mu \in \tilde{M}(P,Q)$, then $\beta(P,Q)=\inf\{\int \sigma\, d\tilde{\lambda};\ \tilde{\lambda} \in \tilde{M}(P,Q)\} = \int \sigma\, d\lambda$ if and only if $\sigma_{x,x}(P_x, Q_x)=\int \sigma_{x,x}\, d\lambda_{(x,x)}\ [P_1]$.*

*Examples and Remarks.* a) Let $M_1 = M_2$ and $\sigma$ be the discrete metric on $M$.

**Corollary 1.** a) $\sigma_{1,2}(P,Q)=\|P_1-Q_1\|+\int \|P_x-Q_x\|\, dP_1 \wedge Q_1(x)$ *where* $P_1 \wedge Q_1(A) = \inf\{P_1(A_1)+Q_1(A_2);\ A_1+A_2=A\}$

b) $\|P-Q\| \leqq \sigma_{1,2}(P,Q)$.

c) If $P_1=Q_1$ *then* $\|P-Q\|=\sigma(P,Q)=\sigma_{1,2}(P,Q)=\beta(P,Q)$.

*Proof.* a) From Theorem 3, $\sigma_{1,2}(P,Q)=\inf\{\int h\, d\mu;\ \mu \in M(P_1,Q_1)\}$, where

$$h(x,y)=\sigma_{x,y}(P_x,Q_y)=\begin{cases} 1 & \text{if } x \neq y \\ \|P_x-Q_x\|, & \text{if } x=y \end{cases}$$

using Dobrushin's result. Therefore,

$$\begin{aligned}\sigma_{1,2}(P,Q) &= \inf\Big\{\mu\{x \neq y\} + \int_{\{x=y\}} \|P_x-Q_x\|\, d\mu;\ \mu \in M(P_1,Q_1)\Big\}\\ &= \inf\Big\{1+\int_{\{x=y\}} (\|P_x-Q_x\|-1)\, d\mu;\ \mu \in M(P_1,Q_1)\Big\}.\end{aligned}$$

Let $\tilde{R} \in M(P_1, Q_1)$ satisfy

$$\tilde{R}(\Delta(A)) = \max\{R(\Delta(A)); R \in M(P_1, Q_1)\} \\ = P_1 \wedge Q_1(A), \qquad \text{for all } A \in \mathfrak{A}$$

where $\Delta(A) = \{(x,x); x \in A\}$ (for existence and construction of $\tilde{R}$ cf. Rüschendorf 1981, Prop. 3).

Since $\|P_x - Q_x\| - 1 \leqq 0$, $\tilde{R}$ solves the inf problem, implying a).

b) Follows from Dobrushin's result saying $\|P - Q\| = \sigma$.

c) Let $P = f\nu$, $Q = g\nu$ and let $P = P_x \times P_1$, $Q = Q_x \times Q_1$, $\nu = \nu_x \times \nu_1$; then as is well-known from the theory of conditional tests

$$\frac{dP_1}{d\nu_1}(x) = \int f(x, y')\, \nu_x(dy'), \qquad P_x \ll \nu_x [\nu_1]$$

and

$$\frac{dP_x}{d\nu_x}(y) = \frac{f(x,y)}{\frac{dP_1}{d\nu_1}(x)}.$$

Therefore,

$$\begin{aligned} \|P - Q\| &= \tfrac{1}{2} \int |f - g|\, d\nu \\ &= \tfrac{1}{2} \int \left| \frac{dP_x}{d\nu_x}(y) - \frac{dQ_x}{d\nu_x}(y) \right| \frac{dP_1}{d\nu_1}(x)\, d\nu(x,y) \\ &= \tfrac{1}{2} \int \left[ \int \left| \frac{dP_x}{d\nu_x}(y) - \frac{dQ_x}{d\nu_x}(y) \right| d\nu_x(y) \right] dP_1(x) \\ &= \int \|P_x - Q_x\|\, dP_1(x) = \sigma_{12}(P, Q). \end{aligned}$$

The identity with $\beta(P, Q)$ is immediate from Proposition 4. □

For $Q = Q_1 \otimes Q_2$ in part c) of Corollary 1 cf. Schwarz (1980), Lemma 2, and Volkonskii and Rozanov (1961), Lemma 4.1.

b) If $(M, \mathfrak{A}) = \prod_{i=1}^{n} (M_i, \mathfrak{A}_i)$ and $\sigma(x,y) = \sum_{i=1}^{n} \sigma_i(x_i, y_i)$, then for $P = \bigotimes_{i=1}^{n} P_i$, $Q = \bigotimes_{i=1}^{n} Q_i$ holds

$$\sigma(P, Q) = \sum_{i=1}^{n} \sigma_i(P_i, Q_i).$$

c) Let $P = N\left(0, \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}\right)$, $Q = N\left(0, \begin{pmatrix} \tau_1^2 & \nu \\ \nu & \tau_2^2 \end{pmatrix}\right)$, then

$$P_x = N\left(\frac{\rho}{\sigma_2^2} x, \sigma_2^2 - \frac{\rho}{\sigma_1^2}\right), \qquad Q_y = N\left(\frac{\nu}{\tau_2^2} y, \tau_2^2 - \frac{\nu^2}{\tau_1^2}\right).$$

Therefore, with $\sigma(x,y) = |x - y|^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2$

$$\sigma_{x,y}(P_x, Q_y) = (x - y)^2 + \left(\frac{\rho}{\sigma_2^2} x - \frac{\nu}{\tau_2^2} y\right)^2 + A,$$

where $A = \left(\left(\sigma_2^2 - \frac{\rho^2}{\sigma_1^2}\right)^{1/2} - \left(\tau_2^2 - \frac{\nu^2}{\tau_1^2}\right)^{1/2}\right)^2.$

By Theorem 3 it holds that $\sigma_{1,2}(P,Q)=A+h(N(0,\sigma_2^2), N(0,\tau_2^2))$, where

$$h(x,y)=\left(\frac{\rho^2}{\sigma_2^4}+1\right)x^2-2\left(\frac{\rho\nu}{\sigma_2^2\tau_2^2}+1\right)xy+\left(\frac{\nu^2}{\tau_2^4}+1\right)y^2.$$

By simple calculation

$$\begin{aligned}&h(N(0,\sigma_2^2), N(0,\tau_2^2))\\&=\left(\frac{\rho^2}{\sigma_2^4}+1\right)\sigma_2^2-2\left|\left(\frac{\rho\nu}{\sigma_2^2\tau_2^2}+1\right)\rho_2\tau_2\right|+\left(\frac{\nu^2}{\tau_2^4}+1\right)\tau_2^2. \quad \square\end{aligned}$$

d) Let $P$, $Q$ be distributions on $(R^2,\mathfrak{B}^2)$ with first marginals $P_1,Q_1$ and conditional distributions $P_x,Q_x$. Let $\sigma(x,y)=\varphi(x_1-y_1)+\psi(x_2-y_2)$, where $\varphi,\psi$ are convex and let $(X_1^*,X_2^*)$, $(Y_1^*,Y_2^*)$ be the two-dimensional quantile transforms, i.e.

$$X_1^*=F_1^{-1}(U_1),\qquad Y_1^*=G_1^{-1}(U_1),\qquad X_2^*=F_{X_1^*}^{-1}(U_2),\qquad Y_2^*=G_{Y_1^*}^{-1}(U_2),$$

$F_1,G_1$ being the $df$'s of $P_1,Q_1$; $F_x$, $G_y$ the $df$'s of $P_x,Q_y$ and $U_1,U_2$ are independent and uniformly distributed on $[0,1]$. Under the assumption of monotone regression dependence the quantile transforms yield the best markov constructions.

**Corollary 2.** *If $F_x$, $G_y$ are both monotonically nondecreasing (or nonincreasing) in $x$, $y$, then $\sigma_{1,2}(P,Q)=E\sigma((X_1^*,X_2^*),(Y_1^*,Y_2^*))$.*

*Proof.* By Theorem 3 it is sufficient to show that

a) $\sigma_{x,y}(P_x,Q_y)=\int\sigma_{x,y}(F_x^{-1}(u_2), G_y^{-1}(u_2))\,du_2$ and

b) $\inf\{\int\sigma_{x,y}(P_x,Q_y)\,dR(x,y);\ R\in M(P_1,Q_1)\}$
$=\int\sigma_{x,y}(P_x,Q_y)\,dR^{X_1^*,Y_1^*}(x,y)$.

Condition a) is implied by Cambanis et al. (1976) or Rüschendorf (1983).
Similarly,

$$\begin{aligned}&\inf\{\textstyle\int\sigma_{x,y}(P_x,Q_y)\,dR(x,y);\ R\in M(P_1,Q_1)\}\\&=\inf_R\{\textstyle\int(\int\sigma_{x,y}(F_x^{-1}(u_2), G_y^{-1}(u_2))\,du_2)\,dR(x,y)\}\\&\geqq\textstyle\int\inf_R\{\int\sigma_{x,y}(F_x^{-1}(u_2), G_y^{-1}(u_2))\,dR(x,y)\}\,du_2.\end{aligned}$$

But

$$\sigma_{x,y}(F_x^{-1}(u_2), G_y^{-1}(u_2))=\varphi(x-y)+\psi(F_x^{-1}(u_2)-G_y^{-1}(u_2))$$

is for each fixed $u_2$ a $L$-superadditive function of $(x,-y)$ (cf. Marshall and Olkin 1979, p. 151–152) implying as above that the distribution of $(F_1^{-1}(U_1), G_1^{-1}(U_1))$ minimizes the inner integral for each $u_2$. $\square$

By induction the optimality of the quantile transform under all markov constructions (for similar distances and under monotone regression dependence) extends to $R^n$, $n\geqq 2$.

## 2. Some Approximation Results

Using the inductive idea of Theorem 3 we obtain several approximation results, which are useful e.g. for the proof of invariance principles for weakly dependent random variables (cf. Berkes and Philipp 1979; Eberlein 1983). We give some results under different conditions on the dependence. The simplicity of the proof is a consequence of an adaption of an idea of Schwarz (1980).

a) Consider the situation of Berkes and Philipp (1979), Theorem 2, i.e. let $(X_k)$ be a sequence of random variables with values in complete seprable metric spaces $(S_k, \sigma_k)$, $k \in \mathbb{N}$, satisfying a $\varphi$-mixing condition

$$|P(X_k \in A_k, X_{(k-1)} \in B_k) - P(X_k \in A_k) P(X_{(k-1)} \in B_k)| \leqq \varphi_k P(X_{(k-1)} \in B_k) \tag{6}$$

for all $A_k \in \mathfrak{B}_k$, the Borel $\sigma$-algebra on $S_k$ and

$$B_k \in \mathfrak{B}_1 \oplus \ldots \oplus \mathfrak{B}_{k-1}, \qquad X_{(k-1)} = (X_1, \ldots, X_{k-1}), \quad k \in \mathbb{N}.$$

**Proposition 5.** *Under assumption* (6) *there exist stochastic processes* $Y = (Y_k)$, $Z = (Z_k)$ *with:*

a) $X \sim Y$, $X = (X_k)$ ($\sim$ *denotes: "same distribution"*)
b) $\{Z_k\}$ *independent*, $Z_k \sim X_k$, $k \in \mathbb{N}$,
c) $P(Z_k \neq Y_k) \leqq \varphi_k$, *for all* $k \in \mathbb{N}$.

*Proof.* For $k \in \mathbb{N}$ let (as in the proof of Theorem 3) $\lambda^k$ be a markov kernel from

$$\left(\prod_{i=1}^{k-1} S_i, \bigotimes_{i=1}^{k-1} \mathfrak{B}_i\right) \quad \text{to} \quad (S_k, \mathfrak{B}_k)$$

with

$$\lambda^k_{x(k-1)} \in M(P^{X_k | x(k-1)}, P^{X_k})$$

such that

$$\|P^{X_k | x(k-1)} - P^{X_k}\| = \int \sigma(x_k, y_k)\, \lambda^k_{x(k-1)}(d(x_k, y_k)) \tag{7}$$

$\sigma$ denoting the discrete metric ($P^{X_1 | x(0)} = P^{X_1}$, $\lambda^1_{x(0)} = \lambda^1$). By Ionescu-Tulcea's theorem we can construct a probability measure $\lambda$ on $\bigotimes_{k=1}^{\infty} (\mathfrak{B}_k \otimes \mathfrak{B}_k)$ with $\lambda^{(Y_k, Z_k) | (y(k-1), z(k-1))} = \lambda^k_{y(k-1)}$, $(Y_k, Z_k)$ denoting the projection on $S_k \times S_k$, implying that $Y = (Y_k) \sim X$, $\{Z_k\}$ independent and $Z_k \sim X_k$, $k \in \mathbb{N}$.
By formula 17.2.10, p. 308 of Ibragimov and Linnik (1971) the mixing assumption (6) implies

$$\|P^{X_k | x(k-1)} - P^{X_k}\| \leqq \varphi_k [P^{X(k-1)}]. \tag{8}$$

Therefore, $P(Y_k \neq Z_k) \leqq \varphi_k$. $\square$

*Remark.* 1) Proposition 5 sharpens the result of Berkes and Philipp (1979), Theorem 2, saying that an approximation is possible with $P(\sigma_k(Y_k, Z_k) \geqq 6\varphi_k) \leqq 6\varphi_k$.

2) With $Q_1 = P^{X(k)}$, $Q_2 = P^{X(k-1)} \otimes P^{X_k}$, Corollary 1, c) and (8) imply that a consequence of the $\varphi$-mixing condition (6) is

$$\|P^{X(k)} - P^{X(k-1)} \otimes P^{X_k}\| \leqq \varphi_k, \qquad k \in \mathbb{N}. \tag{9}$$

So for Proposition 5 the mixing assumption (6) could be weakened to condition (9). That (9) is a consequence of (6) was already noted by Eberlein (1979) and has useful applications for the proof of the central limit theorem.

b) In the situation of example a) replace the $\varphi$-mixing assumption (6) by

$$E\sigma_k(P^{X_k|X_{(k-1)}}, P^{X_k}) \leqq \varphi_k, \qquad k \in \mathbb{N}. \tag{10}$$

This is a kind of weak Bernoulli condition. Similarly to Proposition 5 we obtain:

**Proposition 6.** *Under assumption* (10) *there exist processes* $Y=(Y_k)$, $Z=(Z_k)$ *with* a) $Y \sim X$, b) $\{Z_k\}$ *independent*, $Z_k \sim X_k$, $k \in \mathbb{N}$, c) $E\sigma_k(Y_k, Z_k) \leqq \varphi_k$, $k \in \mathbb{N}$. □

A similar result was given (for stationary processes) by Strittmatter (1982), Theorem E.

c) Assume that $0 \leqq \varphi_k, \eta_k, \psi_k \leqq 1$, $l \in \mathbb{N}$,

$$P(\sigma_k(P^{X_k|X_{(k-1)}}, P^{X_k}) \geqq \varphi_k) \leqq \eta_k, \qquad k \in \mathbb{N} \tag{11}$$

which is a very weak Bernoulli-type condition and was considered by Eberlein (1983) (in a somewhat modified but essentially equivalent form). The following proposition corresponds to his Theorem 1.

**Proposition 7.** *Under condition* (11) *there exist processes* $Y=(Y_k)$, $Z=(Z_k)$ *with* a) $Y \sim X$, b) $\{Z_k\}$ *independent*, $Z_k \sim X_k$, $k \in \mathbb{N}$, *c)* $P(\sigma_k(Y_k, Z_k) \geqq \psi_k) \leqq \dfrac{\varphi_k + \eta_k}{\psi_k}$, $k \in \mathbb{N}$.

*Proof.* For the proof of Proposition 7 we may assume that $\sigma_k \leqq 1$; then we obtain $E\sigma_k(P^{X_k|X_{(k-1)}}, P^{X_k}) \leqq \varphi_k + \eta_k$. Therefore, Proposition 7 follows from Proposition 6 and the Tschebycheff-Markov inequality. □

*Remark.* If we consider more generally also approximations by non independent sequences, the problem arises how to replace the assumption (10) on the conditional distributions by a different workable hypothesis.

If $P$, $Q$ are distributions of infinite sequences then the meaning of the corresponding condition

$$\text{“}E\sigma_k(P^{X_k|X_{(k-1)}}, Q^{W_k|W_{(k-1)}}) \leqq \varphi_k\text{”}$$

is unclear, $X_k$, $W_k$ denoting the corresponding projections. But one can construct as in the proofs of Propositions 6, 7 a probability measure $\lambda$ on $\bigoplus_{k=1}^{\infty} (S_k, \mathfrak{B}_k)$ with

$$\int \sigma_k(x_k, w_k)\, \lambda^{(X_k, W_k)|(x_{(k-1)}, w_{(k-1)})}(d(x_k, w_k)) = \sigma_k(P^{X_k|x_{(k-1)}}, Q^{W_k|w_{(k-1)}}) \tag{12}$$

Now the inductive condition

$$E_\lambda \sigma_k(P^{X_k|X_{(k-1)}}, Q^{W_k|W_{(k-1)}}) \leqq \varphi_k, \qquad k \in \mathbb{N}$$

is well defined and implies the existence of processes

$$Y=(Y_k) \sim P, \qquad Z=(Z_k) \sim Q, \tag{13}$$

and

$$E\sigma_k(Y_k, Z_k) \leqq \varphi_k, \qquad k \in \mathbb{N}.$$

*Example.* Let $P$, $Q$ be the distributions of two real random walks with initial $df$'s $H_1$, $L_1$ and conditional transition $df$'s

$$F_k(x_k | x_{k-1}) = H_k(x_k - x_{k-1}), \qquad G_k(x_k | x_{k-1}) = L_k(x_k - x_{k-1}).$$

For $\sigma_k(x_k, y_k) = (x_k - y_k)^2$, the proposed construction of $Y$, $Z$ leads to:

$$Y_n = \sum_{i=1}^{n} H_i^{-1}(U_i), \qquad Z_n = \sum_{i=1}^{n} L_i^{-1}(U_i), \quad n \in \mathbb{N}, \tag{14}$$

where $U_i$ are independent, $R(0,1)$-distributed.

So our sufficient condition of (13) reads:

$$E(Y_n - Z_n)^2 = E\left[\sum_{i=1}^{n} (H_i^{-1}(U_i) - L_i^{-1}(U_i))\right]^2 \leqq \varphi_n, \qquad n \in \mathbb{N}$$

## 3. Martingale Approximation

Let $X = (X_1, \ldots, X_n)$ be $n$ real random variables on a probability space $(M, \mathfrak{A}, P)$ and let $\mathfrak{A}_1 \subset \mathfrak{A}_2 \subset \ldots \subset \mathfrak{A}_n$ be the sub $\sigma$-algebras of $\mathfrak{A}$ such that $X_k$ is $\mathfrak{A}_k$-measurable. We consider the problem of finding the optimal approximation of $X$ by a martingale $(Y_k, \mathfrak{A}_k)$, $1 \leqq k \leqq n$, w.r.t. the 'Wasserstein distance' generated by $\sigma(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$, where $EX_i$, $1 \leqq i \leqq n$, are assumed to exist, i.e. $E \sum_{i=1}^{n} (X_i - Y_i)^2$ is minimal w.r.t. all martingales.

This problem is interesting in connection with a method of proving central limit theorems due to Gordin (1969) and Statulevičius (1969) and worked out by Philipp and Stout (1975). In a first step one considers approximations by a martingale sequence and then applies martingale embedding theorems.

The prominent candidate for a good approximation is the martingale arising from the Doob-decomposition (w.r.t. $\mathfrak{A}_k$) $X_k = M_k + Z_k$, $1 \leqq k \leqq n$, in a martingale $M$ and a predictable process $Z$ with normalization

$$M_1 = X_1, \qquad \text{i.e. } M_k = \sum_{l=2}^{k} (X_l - E(X_l | \mathfrak{A}_{l-1})) + X_1, \qquad 2 \leqq k \leqq n. \tag{15}$$

But this construction has only a restricted optimality property.

**Proposition 8.** *Let $X_k = M_k + Z_k$, $1 \leqq k \leqq n$, be the (unique) Doob-decomposition with $M_1 = X_1$. Then $M$ is the optimal one-step approximation to $X$* w.r.t. $\sigma$, i.e. *for all $0 \leqq k \leqq n-1$ holds: $E(X_{k+1} - M_{k+1})^2 \leqq E(X_{k+1} - Y_{k+1})^2$ for all $Y_{k+1}$ such that $M_1, \ldots, M_k$, $Y_{k+1}$ is a martingale* w.r.t. $\mathfrak{A}_1, \ldots, \mathfrak{A}_{k+1}$.

*Proof.* For $k = 0$ the statement is trivial. While for $k \leqq n-1$ by Jensen's inequality

$$E(X_{k+1} - Y_{k+1})^2 \geqq E(E(X_{k+1} | \mathfrak{A}_k) - M_k)^2$$

and equality holds iff

$$Y_{k+1}=X_{k+1}-E(X_{k+1}|\mathfrak{A}_k)+M_k=M_{k+1}. \quad \square$$

Clearly Proposition 8 holds true also for distances of the form $\sigma(x,y)=\sum_{i=1}^{n}\varphi_i(x_i-y_i)$, $\varphi_i$ convex, with a different normalization for $M_1$. The weakness of the Doob-decomposition is that it is blind for the further future of the process $X_k$.

**Lemma 9.** *Let* $Z_1,\ldots,Z_k\in L^2(\mathfrak{A},P)$ *and define for* $\mathfrak{B}\subset\mathfrak{A}$ $F^\perp(\mathfrak{B})=\{Y\in L^2(\mathfrak{A},P);\ E(Y|\mathfrak{B})=0\}$.

*Then* a) $Y^*=\frac{1}{k}\sum_{i=1}^{k}(Z_i-E(Z_i|\mathfrak{B}))\in F^\perp(\mathfrak{B})$

b) *For all* $Y\in F^\perp(\mathfrak{B})$ *holds*

$$E\sum_{i=1}^{k}(Z_i-Y)^2\geqq E\sum_{i=1}^{k}(Z_i-Y^*)^2$$

c) $E\sum_{i=1}^{k}(Z_i-Y^*)^2=E\sum_{i=1}^{k}Z_i^2-kE(Y^*)^2$

*Proof.* a) is obvious

b) An element $\tilde{Y}\in F^\perp(\mathfrak{B})$ is the 'projection'

$$\text{iff } E\sum_{i=1}^{k}(Z_i-\tilde{Y})\,Y=0 \qquad \text{for all } Y\in F^\perp(\mathfrak{B}).$$

(For the proof consider the Hilbertspace $\{(Y,\ldots,Y);\ Y\in F^\perp(\mathfrak{B})\}$ and project $Z=(Z_1,\ldots,Z_k)$ on it w.r.t. $\langle X,Y\rangle=E\sum_{i=1}^{k}X_iY_i$.) Since

$$\sum_{i=1}^{k}(Z_i-Y^*)=\sum_{i=1}^{k}E(Z_i|\mathfrak{B})$$

and for

$$Y\in F^\perp(\mathfrak{B})\quad E\sum_{i=1}^{k}E(Z_i|\mathfrak{B})\,Y=E\sum_{i=1}^{k}E(Z_i|\mathfrak{B})\,E(X\,|\mathfrak{B})=0,$$

$Y^*$ is the projection.

c) From the orthogonality condition

$$E\sum_{i=1}^{k}(Z_i-Y^*)^2=E\sum_{i=1}^{k}(Z_i-Y^*)\,Z_i=E\sum_{i=1}^{k}Z_i^2-E\sum_{i=1}^{k}Z_i\,Y^*=E\sum_{i=1}^{k}Z_i^2-kE(Y^*)^2.$$

$\square$

Define now for $k,l\leqq n$, $m_{k,l}=E(X_k|\mathfrak{A}_l)$ and $Y_1=\frac{1}{n}\left(X_1+\sum_{l=2}^{n}m_{l,1}\right)$

$$\begin{aligned}Y_k&=\frac{1}{n-k+1}\left(X_k+\sum_{l=k+1}^{n}m_{l,k}-\sum_{l=k}^{n}m_{l,k-1}\right)+Y_{k-1}\\&=\frac{1}{n-k+1}\sum_{l=k}^{n}(m_{l,k}-m_{l,(k-1)})+Y_{k-1},\qquad 2\leqq k\leqq n.\end{aligned}\tag{16}$$

**Theorem 10.** *Let $(X_k, \mathfrak{A}_k)$, $1 \leqq k \leqq n$, be a stochastic square integrable sequence; the optimal martingale approximation $(Y_k, \mathfrak{A}_k)$ to $X$ w.r.t. to the "Wasserstein distance" generated by $\sigma(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2$ is given by (16).*

*Proof.* First note that $(Y_k, \mathfrak{A}_k)$ is a martingale, since

$$E(Y_k | \mathfrak{A}_{k-1}) = Y_{k-1} + \frac{1}{n-k+1} \sum_{l=k}^{n} (E(m_{l,k} | \mathfrak{A}_{k-1}) - m_{l,k-1}) = Y_{k-1}.$$

We prove by induction that an optimal martingale $(Z_k, \mathfrak{A}_k)$ satisfies (16), starting with $k=n$. By Jensen's inequality

$$E \sum_{r=1}^{n} (X_r - Z_r)^2 \geqq E(m_{n,n-1} - Z_{n-1})^2 + E \sum_{r=1}^{n-1} (X_r - Z_r)^2$$

while equality holds iff $Z_n = X_n - m_{n,n-1} + Z_{n-1}$.

Assume now that an optimal martingale $(Z_k, \mathfrak{A}_k)$ satisfies (16) for indices $r \geqq k+1$, i.e.

$$Z_r = \frac{1}{n-r+1} \left( \sum_{l=r}^{n} m_{l,r} - \sum_{l=r}^{n} m_{l,r-1} \right) + Z_{r-1}, \qquad r \geqq k+1$$

and use that

$$\sum_{r=1}^{n} (X_r - Z_r)^2 = \sum_{r=1}^{k-1} (X_r - Z_r)^2 + \sum_{r=k}^{n} (X_r - Z_r + Z_k - Z_{k-1} - W_k)^2,$$

where $W_k = Z_k - Z_{k-1} \in F^{\perp}(\mathfrak{A}_{k-1})$.

By Lemma 9 we get a lower bound for the expectation of this expression by the choice

$$W_k = \frac{1}{n-k+1} \sum_{l=k}^{n} [X_l - Z_l + Z_k - Z_{k-1} - (m_{l,k-1} - Z_{k-1} + Z_{k-1} - Z_{k-1})]$$
$$= \frac{1}{n-k+1} \left\{ \sum_{l=k}^{n} m_{l,l} - \sum_{l=k+1}^{n} Z_l - \sum_{l=k}^{n} m_{l,k-1} + (n-k) Z_k \right\}.$$

By induction hypothesis

$$\sum_{l=k+1}^{n} Z_l = \sum_{l=k+1}^{n} m_{ll} - \sum_{l=k+1}^{n} m_{l,k} + (n-k) Z_k \tag{17}$$

implying that $W_k = \frac{1}{n-k+1} \left( X_k + \sum_{l=k+1}^{n} m_{l,k} - \sum_{l=k}^{n} m_{l,k-1} \right)$, i.e. (16) holds also for the index $k$.

In the final step we have to minimize

$$E \sum_{r=1}^{n} (X_r - Z_r)^2 = E \sum_{r=1}^{n} (X_r - (Z_r - Z_1) - Z_1)^2$$

as function of $Z_1$, since $Z_r - Z_1 = \sum_{l=2}^{r} (Z_l - Z_{l-1})$ depends only on functions of $X$ and its conditional expectations. But it is well known that this expression is minimized by

$$\begin{aligned} Z_1 &= \frac{1}{n}\sum_{r=1}^{n} (X_r - (Z_r - Z_1)) \\ &= \frac{1}{n}\left(X_1 + \sum_{r=2}^{n} X_r - \sum_{r=2}^{n} Z_r - (n-1) Z_1\right) \\ &= \frac{1}{n}\left(X_1 + \sum_{l=2}^{n} m_{l,1}\right). \end{aligned}$$

So we obtain that an optimal martingale fulfills (16). □

*Remark.* a) The construction of an optimal martingale approximation for $\sigma(x,y) = \sum \varphi(x_i - y_i)$, $\varphi$ convex, can be given along similar lines, but does not allow to get explicit general terms. For $n=2$ we get e.g. $Y_2 = X_2 - m_{2,1} + Y_1$, where $Y_1$ is a minimum point of $y \to \varphi(X_1 - y) + \varphi(m_{2,1} - y)$.

b) For $n=2$ the best martingale approximation has the distance $\frac{1}{2}E(m_{2,1} - X_1)^2$. The Doob-decomposition martingale has the distance $E(m_{2,1} - X_1)^2$, while the best approximation by random variables $(W_1, \mathfrak{A}_1)$, $(W_2, \mathfrak{A}_2)$ with $EW_1 = EW_2$ has the distance $\frac{1}{2}(EX_1 - EX_2)^2$.

c) If $\mathfrak{A}(X_k) \subset \mathfrak{B}_k \subset \mathfrak{A}_k$, $k \leqq n$, $\mathfrak{A}_k$, $\mathfrak{B}_k$ increasing, then each $\mathfrak{A}_k$-martingale $Y_k$ can be improved by a $\mathfrak{B}_k$-martingale, namely $\tilde{Y}_k = E(Y_k | \mathfrak{B}_k)$. So the best choice in this case is $\mathfrak{A}_k = \mathfrak{A}(X_1, \ldots, X_k)$.

For general increasing sequence $\mathfrak{A}_k$ and any $\mathfrak{A}_k$-martingale $(Y_k)$ holds

$$E\sum_{k=1}^{n} (X_k - Y_k)^2 = \sum_{k=1}^{n} E(X_k - E(X_k|\mathfrak{A}_k))^2 + \sum_{k=1}^{n} E(E(X_k|\mathfrak{A}_k) - Y_k)^2,$$

so the optimal approximation can be read off from Theorem 10 (replacing $X_k$ by $E(X_k|\mathfrak{A}_k)$). But it seems to be difficult to find the optimal sequence $\mathfrak{A}_k$. One problem arising in this connection is to determine for

$X, Y \in L^2(\mathfrak{A}, P)$, $\quad \inf\{E(X - E(Y|\mathfrak{B}))^2; \mathfrak{B} \subset \mathfrak{A}\}$.

*Acknowledgement.* I would like to thank a referee for some critical remarks on the notion of "Wasserstein distance" and for pointing out the relevance of the papers of Dall'Aglio (1956); Kantorovic and Rubinstein (1958) and a survey article of Rachev (1984) (forthcoming in Theory Prob. Appl. Vol. 30, 1985)

## References

Berkes, I., Philipp, W.: Approximation theorems for independent and weakly dependent random vectors. Ann. Probab. **7**, 29–54 (1979)

Cambanis, S., Simons, G., Stout, W.: Inequalities for $Ek(X,Y)$ when the marginals are fixed. Z. Wahrscheinlichkeitstheor. verw. Geb. **36**, 285–294 (1976)

Castaing, C., Valadier, M.: Convex Analysis and Measurable Multifunctions. Lecture Notes in Mathematics **580**. New York-Heidelberg-New York: Springer 1977

Dall'Aglio, G.: Sugli estremi di momenti delle funzioni di ripartizione doppia. Annali Scuola Normale Superiore di Pisa, Vol. **10**, 35–74 (1956)

Dobrushin, R.L.: Prescribing a system of random variables by conditional distributions. Theory Probab. Appl. **15**, 458–486 (1970)

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

Eberlein, E.: Strong approximation of very weak Bernoulli processes. Z. Wahrscheinlichkeitstheor. Verw. Gebiete **62**, 17–37 (1983)

Eberlein, E.: An invariance principle for lattices of dependent random variables. Z. Wahrscheinlichkeitstheor. Verw. Gebiete **50**, 119–133 (1979)

Gordin, M.I.: The central limit theorem for stationary processes. Sov. Math. Dokl. **10**, 1174–1176 (1969)

Ibragimov, I.A., Linnik, Yu.V.: Independent and stationary sequences of Random Variables. Groningen: Wolks Noordhoff 1971

Kantorovic, L., Rubinstein, G.: On a space of completely additive functions. Vestn. Leningr. Univ. Math. 13, **7**, 52–59 (1958)

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

Marshall, A.W., Olkin, I.: Inequalities: Theory of Majorization and its Applications. New York: Academic Press 1979

Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrizes. Linear Algebra Appl. **43**, 257–263 (1982)

Philipp, W., Stout, W.: Almost sure invariance principles for sums of weakly dependent random variables. Am. Math. Soc., Mémoir **161**.

Rachev, S.T.: Minimal metrices in the random variable space. Publ. Inst. Stat. Univ. Paris **28**, 1–26 (1982)

Rüschendorf, L.: Solution of a statistical optimization problem by rearrangement methods. Metrika **30**, 55–62 (1983)

Rüschendorf, L.: Robust tests for independence. Preprint 1981

Schwarz, G.: Finitely determined processes – an indiscrete approach. J. Math. Anal. Appl. **76**, 146–158 (1980)

Strittmatter, W.: Metriken für stochastische Prozesse und schwache Abhängigkeit. Diplomarbeit, Freiburg 1982

Statulevicius, V.A.: Limit theorems for sums of random variables related to a markov chain II. Litov. Mat. Sb. **9**, 635–672 (1969)

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

Volkonskii, V.A., Rozanov, J.A.: Some limit theorems for random functions II. Theory Probab. Appl. **6**, 186–198 (1961)

Zolotarev, V.M.: Probability metrics. Theory Probab. Appl. **28**, 278–302 (1983)


Received June 11, 1983; in revised form March 25, 1985