2. GLIVENKO-CANTELLI-convergence: The VAPNIK-CHERVONENKIS-Theory with some extensions.

Let us start with the simplest case: Assume that $(\xi_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. random variables on some p-space $(\Omega,\mathcal{F},\mathbb{P})$ with distribution function (df) F; let F_n be the EMPIRICAL df pertaining to $\xi_1,\dots,\xi_n$, i.e.,

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(\xi_i), \quad t\in\mathbb{R}.$$

Then the classical GLIVENKO-CANTELLI Theorem states:

$$(8) \qquad D_n^F := \sup_{t\in\mathbb{R}}|F_n(t)-F(t)| \to 0 \quad \mathbb{P}\text{-a.s.}$$

(Note that D_n^F is a random variable since $D_n^F = \sup_{t\in\mathbb{Q}}|F_n(t)-F(t)|$, where $\mathbb{Q}$ denotes the rationals.)

The proof of (8) usually runs as follows:

a) One shows that (8) holds true if the ξ_i's are uniformly distributed on (0,1).

b) Using the QUANTILE TRANSFORMATION

$$s \mapsto F^{-1}(s) := \inf\{t\in\mathbb{R}: F(t)\geq s\}, \quad s\in(0,1)$$

and a) one obtains (8) for the SPECIAL VERSIONS $\hat{\xi}_i := F^{-1}(\eta_i)$, where the η_i's are independent and uniformly distributed on (0,1) (and defined on the same p-space as the ξ_i's).
Note that $\mathcal{L}\{\hat{\xi}_i\} = \mathcal{L}\{\xi_i\}$ for each i; even more, by independence, one has $\mathcal{L}\{(\hat{\xi}_i)_{i\in\mathbb{N}}\} = \mathcal{L}\{(\xi_i)_{i\in\mathbb{N}}\}$.

c) Reasoning on the fact that the validity of (8) only dependes on $\mathcal{L}\{(\xi_i)_{i\in\mathbb{N}}\}$ the proof is concluded.

In view of the more general situations we shall consider later on in this section we want to clarify c) a little bit more:

c*) (8) claims that

$$\mathbb{P}(\{\omega\in\Omega: \lim_{n\to\infty} (\sup_{t\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(\xi_i(\omega))-F(t)|)=0\}) = 1;$$

consider $\xi(\omega) := (\xi_1(\omega), \xi_2(\omega),\ldots)\in\mathbb{R}^{\mathbb{N}}$ and put

$$g_n(\xi(\omega)) := \sup_{t\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(\xi_i(\omega))-F(t)|.$$

Now, <u>note</u> (<u>and remember</u>) the fact that in the present situation

$$(9) \qquad g_n : \mathbb{R}^{\mathbb{N}} \to \mathbb{R} \text{ is } \mathcal{B}_{\mathbb{N}},\mathcal{B}\text{-measurable}$$

since $g_n(\underline{x}) = \sup_{t\in\mathbb{Q}} |\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(x_i)-F(t)|$ for $\underline{x} := (x_1,x_2,\ldots)$.

Therefore $A := \{\underline{x}\in\mathbb{R}^{\mathbb{N}} : \lim_{n\to\infty} g_n(\underline{x})=0\} \in \mathcal{B}_{\mathbb{N}}$, whence, putting

$\hat{\xi}(\omega) := (\hat{\xi}_1(\omega), \hat{\xi}_2(\omega),\ldots)$ one obtains

$$\begin{aligned}
&\mathbb{P}(\{\omega\in\Omega : \lim_{n\to\infty} (\sup_{t\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(\xi_i(\omega))-F(t)|)=0\})\\
&= \mathbb{P}(\{\omega\in\Omega : \lim_{n\to\infty} g_n(\xi(\omega))=0\}) = \mathbb{P}(\{\omega\in\Omega: \xi(\omega)\in A\})\\
&= L\{\xi\}(A) \underset{b)}{=} L\{\hat{\xi}\}(A) = \mathbb{P}(\{\omega\in\Omega : \hat{\xi}(\omega)\in A\})\\
&= \mathbb{P}(\{\omega\in\Omega : \lim_{n\to\infty} (\sup_{t\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(\hat{\xi}_i(\omega))-F(t)|)=0\}) \underset{b)}{=} 1.
\end{aligned}$$

When taking $(X,\mathcal{B}) = (\mathbb{R},\mathcal{B})$ and $\mathcal{C} := \{(-\infty,t] : t\in\mathbb{R}\}$, then, in the setting of Section 1, the GLIVENKO-CANTELLI-convergence (8) reads as

$$(8^*) \qquad D_n(\mathcal{C},\mu) \equiv \sup_{C\in\mathcal{C}}|\mu_n(C) - \mu(C)| \to 0 \quad \mathbb{P}\text{-a.s.}$$

Concerning more general situations it turns out however that (8^*) may hold for the empirical measures obtained from one sequence $\xi_1,\xi_2,\ldots$ of independent random elements in $(X,\mathcal{B})$ each having distribution μ but <u>not</u> for the empirical measures obtained from another such sequence, say $\eta_1,\eta_2,\ldots$.

<u>EXAMPLE</u> (cf. D. Pollard (1981), Example (5.1)).

Let $(X,\mathcal{B},\mu)$ be a nonatomic p-space (i.e., $\{x\}\in\mathcal{B}$ and $\mu(\{x\})=0$ for all $x\in X$). Suppose that there exists a subset A of X with inner measure $\mu_*(A)=0$ and outer measure $\mu^*(A)=1$ (cf. P.R. Halmos (1969), Section 16, for an example). Let $\mathcal{B}_A := A\cap\mathcal{B}$ be the trace σ-algebra of $\mathcal{B}$ on A and μ_A be the p-measure defined on $\mathcal{B}_A$ by $\mu_A(A\cap B) := \mu(B)$, $B\in\mathcal{B}$; note that μ_A is well defined since $\mu^*(A)=1$. By the definition of $\mathcal{B}_A$ the embedding ξ_A of A into X is $\mathcal{B}_A,\mathcal{B}$-measurable $(\xi_A^{-1}(B) = \{x\in A: \xi_A(x)=x\in B\} = A\cap B\in\mathcal{B}_A$ for any $B\in\mathcal{B})$ and one has

$$\mu_A(\xi_A^{-1}(B)) = \mu(B) \quad \text{for all } B\in\mathcal{B}.$$

Consider the p-space

$$(\Omega_1,\mathcal{F}_1,\mathbb{P}_1) := (A^{\mathbb{N}}, \bigotimes_{\mathbb{N}} \mathcal{B}_A, \times_{\mathbb{N}} \mu_A)$$

and on it the random elements $\xi_i: \Omega_1 \to X$, defined by

$$\xi_i(\omega_1) := \xi_A(\pi_i(\omega_1)), \quad i\in\mathbb{N}, \ \omega_1\in\Omega_1,$$

where $\pi_i: A^{\mathbb{N}} \to A$ denotes the i-th coordinate projection.

Then, by construction, the ξ_i's are independent having distribution μ $(\mathcal{L}\{\xi_i\}(B) = \mathbb{P}_1(\xi_i^{-1}(B)) = \mathbb{P}_1(\pi_i^{-1}(\xi_A^{-1}(B))) = \mu_A(\xi_A^{-1}(B)) = \mu(B)$ for each $B\in\mathcal{B})$.

Now, let $\mathcal{C}$ be the class of all finite subsets of A; then $\mathcal{C}\subset\mathcal{B}$ and $\mu(C)=0$ for all $C\in\mathcal{C}$ since $(X,\mathcal{B},\mu)$ we assumed to be nonatomic. But since all the ξ_i's take their values in A it follows that for the empirical measures μ_n pertaining to $\xi_1,\ldots,\xi_n$ one has

$$\sup_{C\in\mathcal{C}} |\mu_n(C)-\mu(C)| \equiv 1.$$

Taking instead the p-space

$$(\Omega_2,\mathcal{F}_2,\mathbb{P}_2) := ((\complement A)^{\mathbb{N}}, \bigotimes_{\mathbb{N}} \mathcal{B}_{\complement A}, \times_{\mathbb{N}} \mu_{\complement A})$$

and on it the random elements $\eta_i: \Omega_2 \to X$, defined by

$$\eta_i(\omega_2) := \xi_{\complement A}(\pi_i(\omega_2)), \quad i\in\mathbb{N}, \ \omega_2\in\Omega_2,$$

where here $\pi_i:(\complement A)^{\mathbb{N}} \to \complement A$ is again the i-th coordinate projection and where $\xi_{\complement A}$

denotes the embedding of $\complement A$ into X, it follows as before (noticing that $\mu^*(\complement A)=1$) that the η_i's are i.i.d. with distribution μ (whence $L\{(\eta_i)_{i\in\mathbb{N}}\} = L\{(\xi_i)_{i\in\mathbb{N}}\}$).

But, for the <u>same</u> class $\mathcal{C}$ as before one has now for the empirical measure μ_n pertaining to $\eta_1,\ldots,\eta_n$, $\mu_n(C)=0$ for all $C\in\mathcal{C}$, since the η_i's take their values in $\complement A$, whence

$$\sup_{C\in\mathcal{C}}|\mu_n(C)-\mu(C)| \equiv 0.$$

(Note that in both cases $D_n(\mathcal{C},\mu) \equiv \sup_{C\in\mathcal{C}}|\mu_n(C)-\mu(C)|$ is measurable.)

Finally, taking as underlying p-space the CANONICAL MODEL

$$(\Omega,\mathcal{A},\mathbb{P}) = (X^{\mathbb{N}},\mathcal{B}_{\mathbb{N}}, \underset{\mathbb{N}}{\times}\mu)$$

and on it the coordinate projections ξ_i, $i\in\mathbb{N}$, being again i.i.d. with distribution μ, the above example shows that for the very same class $\mathcal{C}$ one gets e.g.,

$$\sup_{C\in\mathcal{C}}|\mu_1(C,\underline{x})-\mu(C)| = \sup_{C\in\mathcal{C}} \mu_1(C,\underline{x}) = 1_A(x_1)$$

for $\underline{x} = (x_1,x_2,\ldots) \in X^{\mathbb{N}}$, whence, since $A\notin\mathcal{B}$

$$\{\underline{x}\in X^{\mathbb{N}} : \sup_{C\in\mathcal{C}}|\mu_1(C,\underline{x})-\mu(C)| = 1\} = A\times X\times X\times\ldots\notin\mathcal{B}_{\mathbb{N}},$$

i.e., here - <u>in contrast to (9)</u> -

(10) $$\underline{x} \mapsto g_n(x) := \sup_{C\in\mathcal{C}}|\mu_n(C,\underline{x})-\mu(C)|$$

is <u>not</u> $\mathcal{B}_{\mathbb{N}},\mathcal{B}$-measurable.

This indicates already the need for appropriate measurability assumptions to be discussed later.

Let us point out at this stage the usefulness of GLIVENKO-CANTELLI-convergence in statistics by giving only one example concerning CHERNOFF-type estimates of the mode (c.f. H. Chernoff (1964), and E.J. Wegman (1971)). For other examples, see P. Gaenssler and J.A. Wellner (1981). For the moment we anticipate the following GLIVENKO-CANTELLI-Theorem which will be proved later in this section:

(11) Let $\xi_1, \xi_2, \ldots$ be i.i.d. random vectors on some p-space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $X = \mathbb{R}^k$, $k \geq 1$, having distribution μ on $\mathcal{B} = \mathcal{B}_k$. Let $\mathcal{C} \equiv \mathbb{B}_k$ be the class of all closed Euclidean balls in $\mathbb{R}^k$; then

$$\lim_{n\to\infty} \left(\sup_{C\in\mathbb{B}_k} |\mu_n(C) - \mu(C)|\right) = 0 \quad \mathbb{P}\text{-a.s.}$$

Now, consider $(X, \mathcal{B}) = (\mathbb{R}^k, \mathcal{B}_k)$, $k \geq 1$, and suppose that μ is "unimodal" in the following sense:

(*) there exists a $\theta \in \mathbb{R}^k$ such that for some $\delta_o > 0$, $\mu(B^c(\theta, \delta_o)) > \mu(B^c(x, \delta_o))$ for all $x \in \mathbb{R}^k$, $x \neq \theta$, where $B^c(x, \delta_o)$ denotes the closed Euclidean ball with center x and radius δ_o.

Facing the problem of finding a consistent sequence of estimators for the (unknown) θ, one may proceed as follows:

Suppose that $0 < r_n$ with $\lim_{n\to\infty} r_n = \delta_o$ are given; then, given i.i.d. observations $x_i = \xi_i(\omega)$, $i = 1, \ldots, n$, choose as estimate $\theta_n(\omega) \equiv \theta(\xi_1(\omega), \ldots, \xi_n(\omega))$ a center of a closed Euclidean ball with radius r_n which covers most of the observations, i.e. for which

(**) $\quad \mu_n(B^c(\theta_n(\omega), r_n), \omega) \geq \mu_n(B^c(x, r_n), \omega)$ for all $x \in \mathbb{R}^k$.

Then the claim is that $\lim_{n\to\infty} \theta_n = \theta$ $\mathbb{P}$-a.s.

<u>Proof.</u> Choose $M > 0$ such that $\mu(\complement B^c(0, M)) < \mu(B^c(\theta, \delta_o))$. According to (11) we have

$$\mathbb{P}(\Omega_o) = 1 \text{ for } \Omega_o := \{\omega \in \Omega : \lim_{n\to\infty} (\sup_{C\in\mathbb{B}_k} |\mu_n(C) - \mu(C)|) = 0\}.$$

Let $\omega \in \Omega_o$ and suppose that $\theta_n(\omega) \not\to \theta$ as $n \to \infty$; then

<u>either</u> (1) $\limsup_{n\to\infty} |\theta_n(\omega)| = \infty$

<u>or</u> (2) there exists an $x \neq \theta$ such that $\lim_{j\to\infty} \theta_{n_j}(\omega) = x$ for some subsequence (n_j) of $\mathbb{N}$.

We will show that both, (1) and (2), will lead to a contradiction.

<u>ad (1):</u> $\limsup_{n\to\infty} |\theta_n(\omega)| = \infty$ implies that there exists some subsequence (n_j) of $\mathbb{N}$ such that $B^c(\theta_{n_j}(\omega), r_{n_j}) \subset \complement B^c(0, M)$ for all j, whence

$$\limsup_{j\to\infty} \mu_{n_j}(B^c(\theta_{n_j}(\omega),r_{n_j}),\omega) \leq \liminf_{j\to\infty} \mu_{n_j}(\complement B^c(0,M),\omega)$$

$$= 1 - \limsup_{j\to\infty} \mu_{n_j}(\underbrace{B^c(0,M)}_{\in\mathbb{B}_k},\omega) = 1-\mu(B^c(0,M)) = \mu(\complement B^c(0,M))$$

$$< \mu(\underbrace{B^c(\theta,\delta_o)}_{\in\mathbb{B}_k}) = \liminf_{j\to\infty} \mu_{n_j}(B^c(\theta,\delta_o),\omega) \leq \limsup_{j\to\infty} \mu_{n_j}(B^c(\theta,r_{n_j}),\omega)$$

which is in contradiction with the choice of θ_n according to (**).

<u>ad (2):</u> $\lim_{j\to\infty} \theta_{n_j}(\omega) = x \neq \theta$ implies that $\liminf_{j\to\infty} \mu_{n_j}(B^c(\theta_{n_j}(\omega),r_{n_j}),\omega)$

$$\leq \underbrace{\lim_{j\to\infty} |\mu_{n_j}(B^c(\theta_{n_j}(\omega),r_{n_j}),\omega) - \mu(B^c(\theta_{n_j}(\omega),r_{n_j}))|}_{=\,0 \text{ according to (11)}}$$

$$+ \limsup_{j\to\infty} \mu(B^c(\theta_{n_j}(\omega),r_{n_j})) \leq \mu(B^c(x,\delta_o)) \underset{(*)}{<} \mu(B^c(\theta,\delta_o))$$

$$= \liminf_{j\to\infty} \mu_{n_j}(B^c(\theta,\delta_o),\omega) \leq \liminf_{j\to\infty} \mu_{n_j}(B^c(\theta,r_{n_j}),\omega),$$

which again is in contradiction with the choice of θ_n according to (**). □

Before starting with the VAPNIK-CHERVONENKIS Theory we want to add here some remarks concerning the $\mathbb{P}$-a.s. limiting behaviour of so-called weighted discrepancies which are of importance in statistics as well (cf. T.W. Anderson and D.A. Darling (1952), J. Durbin (1953)).

For this, let $(X,\mathcal{B}) = (\mathbb{R},\mathcal{B})$ and $(\xi_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables on some p-space $(\Omega,\mathcal{F},\mathbb{P})$ with df F; let F_n be the empirical df pertaining to $\xi_1,\ldots,\xi_n$ and define the WEIGHTED DISCREPANCY by

$$D_n^F(q) := \sup_{t\in\mathbb{R}} \frac{|F_n(t)-F(t)|}{q(F(t))}$$

where $q\colon [0,1] \to \mathbb{R}_+$ is some given WEIGHT FUNCTION.

(Note that $D_n^F(q) \equiv D_n^F$ for $q \equiv 1$.)

Considering instead of the ξ_i's the special versions $\hat{\xi}_i = F^{-1}(\eta_i)$, the η_i's being independent and uniformly distributed on $(0,1)$, it follows in the same way as pointed out in part c*) of the outline of the proof of (8) that w.r.t. $\mathbb{P}$-a.s. convergence of $D_n^F(q)$ for <u>continuous</u> q's one may consider w.l.o.g. instead of $D_n^F(q)$ its versions

$$\hat{D}_n^F(q) := \sup_{t\in\mathbb{R}} \frac{|\hat{F}_n(t)-F(t)|}{q(F(t))},$$

where $\hat{F}_n$ is the empirical df pertaining to $\hat{\xi}_1,\ldots,\hat{\xi}_n$.
But, due to the identity $\{\hat{\xi}_i\leq t\} = \{\eta_i\leq F(t)\}$ for all $t\in\mathbb{R}$ one has $\hat{F}_n(t) = U_n(F(t))$ for all $t\in\mathbb{R}$, where U_n is the empirical df pertaining to $\eta_1,\ldots,\eta_n$. Therefore

$$\hat{D}_n^F(q) = \sup_{t\in\mathbb{R}} \frac{|U_n(F(t))-F(t)|}{q(F(t))} \leq \sup_{s\in[0,1]} \frac{|U_n(s)-s|}{q(s)} =: D_n(q),$$

where we remark that for continuous F we have even $\hat{D}_n^F(q) = D_n(q)$, whence for continuous q's and F's one has, comparing again with part c*) of the outline of the proof of (8), that

$$D_n^F(q) \overset{L}{=} \hat{D}_n^F(q) = D_n(q)$$

showing that in this case $D_n^F(q)$ is a DISTRIBUTION-FREE STATISTIC. By the way, since for continuous q's and <u>arbitrary</u> F's $D_n^F(q) \overset{L}{=} \hat{D}_n^F(q) \leq D_n(q)$, we obtain in this case that

$$\mathbb{P}(D_n^F(q)\geq d) \leq \mathbb{P}(D_n(q)\geq d) \quad \text{for each } d\geq 0.$$

Also, the above remarks show that for continuous q's we may restrict ourselves w.l.o.g. to the case of finding conditions on q such that

$$(*) \quad \lim_{n\to\infty} D_n(q) = 0 \quad \mathbb{P}\text{-a.s.}$$

in order to get the same GLIVENKO-CANTELLI-convergence for $D_n^F(q)$.

The following theorem gives in a certain sense necessary and sufficient conditions on q for (*) to hold (cf. J.A. Wellner (1977) and (1978)).

<u>THEOREM 1.</u> Let $(\eta_i)_{i\in\mathbb{N}}$ be a sequence of independent random variables on some p-space $(\Omega,\mathcal{F},\mathbb{P})$ being uniformly distributed on [0,1]. Let $D_n(q)$ be defined as above with a weight function q belonging to the set

$\mathcal{Q}_1 := \{q:[0,1] \to \mathbb{R}$, q continuous, $q(0) = q(1) = 0$, $q(t) > 0$ for all $t\in(0,1)$, q monotone increasing on $[0,\delta_o]$ and monotone decreasing on $[1-\delta_1,1]$ for appropriate $\delta_i=\delta_i(q)$, $i=0,1\}$.

Then, putting $\Psi(t) := \frac{1}{q(t)}$, one has:

(i) For any $q\in\mathcal{Q}_1$ with $\int_0^1 \Psi(t)dt < \infty$ it follows that

$$\lim_{n\to\infty} D_n(q) = 0 \quad \mathbb{P}\text{-a.s.}$$

(ii) For any $q\in\mathcal{Q}_1$ with $\int_0^1 \Psi(t)dt = \infty$ it follows that

$$\limsup_{n\to\infty} D_n(q) = \infty \quad \mathbb{P}\text{-a.s.}$$

Proof.(i): For any $\varepsilon>0$ there exist $\theta_i>0$ such that $\theta_i<\delta_i$, $i=0,1$, with

$$\int_0^{\theta_o}\Psi(t)dt < \varepsilon/4 \quad \text{and} \quad \int_{1-\theta_1}^1 \Psi(t)dt < \varepsilon/4.$$

We have $D_n(q) = \sup_{t\in(0,1)} \Psi(t)|U_n(t)-t| \le \sup_{0<t\le\theta_o} \Psi(t)U_n(t)$

$$+ \sup_{0<t\le\theta_o} \Psi(t)t + \sup_{\theta_o\le t\le 1-\theta_1} \Psi(t)|U_n(t)-t| + \sup_{1-\theta_1\le t<1} \Psi(t)|U_n(t)-t|$$

$$=: I_1(n) + I_2 + I_3(n) + I_4(n), \text{ say.}$$

Now, to start with the first summand $I_1(n)$, one has

$$I_1(n) \le \frac{1}{n}\sum_{i=1}^n \sup_{0<t\le\theta_o} \Psi(t)\, 1_{[0,t]}(\eta_i) \le \frac{1}{n}\sum_{i=1}^n \Psi(\eta_i)\, 1_{[0,\theta_o]}(\eta_i)$$

$\to \int_0^{\theta_o} \Psi(t)dt$ $\mathbb{P}$-a.s. by the SLLN, whence $\limsup_{n\to\infty} I_1(n) < \varepsilon/4$ $\mathbb{P}$-a.s.

Concerning I_2, note that for all $0<t\le\theta_o$

$$\Psi(t)t \le \int_0^t\Psi(s)ds \le \int_0^{\theta_o} \Psi(s)ds, \text{ whence } I_2 = \sup_{0<t\le\theta_o} \Psi(t)t < \varepsilon/4.$$

As to $I_3(n)$ we have $I_3(n) = \sup_{\theta_o\le t\le 1-\theta_o} \frac{1}{q(t)} |U_n(t)-t|$

$$\le \underbrace{[\min_{\theta_o\le t\le 1-\theta_1} q(t)]^{-1}}_{=:\, c \text{ with } 0<c<\infty} \sup_{\theta_o\le t\le 1-\theta_1} |U_n(t)-t| \to 0 \quad \mathbb{P}\text{-a.s. according to (8).}$$

Therefore $\lim_{n\to\infty} I_3(n) = 0$ $\mathbb{P}$-a.s.

Finally, $I_4(n) = \sup_{1-\theta_1 \leq t<1} \Psi(t)\, |1 - \frac{1}{n} \sum_{i=1}^{n} 1_{(t,1]}(\eta_i) - t|$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{1-\theta_1 \leq t<1} \Psi(t) 1_{(t,1]}(\eta_i) + \sup_{1-\theta_1 \leq t<1} \Psi(t)(1-t)$$

$\leq \frac{1}{n} \sum_{i=1}^{n} \Psi(\eta_i) 1_{(1-\theta_1,1]}(\eta_i) + \int_{1-\theta_1}^{1} \Psi(t)dt$ where again by the SLLN

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Psi(\eta_i) 1_{(1-\theta_1,1]}(\eta_i) = \int_{1-\theta_1}^{1} \Psi(t)dt \quad \mathbb{P}\text{-a.s.; thus}$$

$$\limsup_{n\to\infty} I_4(n) < \varepsilon/2 \quad \mathbb{P}\text{-a.s.}$$

So we have shown that $\limsup_{n\to\infty} D_n(q) < \varepsilon$ $\mathbb{P}$-a.s. for any $\varepsilon>0$; this proves the assertion in (i).

<u>(ii):</u> Let $N \in \mathbb{N}$ be arbitrary but fixed; we will show

$$(+) \quad \limsup_{n\to\infty} D_n(q) \geq N \quad \mathbb{P}\text{-a.s.}$$

which gives the assertion in (ii).

Now, by assumption,

$$\text{either (a)} \quad \int_0^{\delta_o} \Psi(t)dt = \infty$$

$$\text{or} \quad \text{(b)} \quad \int_{1-\delta_1}^{1} \Psi(t)dt = \infty.$$

Let us consider case (a) (case (b) can be dealt with in an analogous way), i.e., assume

$$\int_0^{\delta_o} \Psi(t)dt = \infty.$$

Then, for n sufficiently large, say $n \geq n_o$,

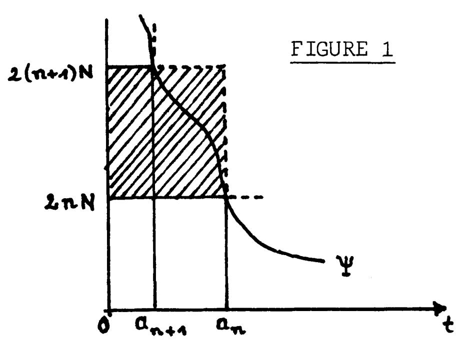

FIGURE 1

$a_n := \max \{t \leq \delta_o : \Psi(t) = 2nN\}$

is well defined (i.e. $\{\ldots\} \neq \emptyset$) and (cf. Figure 1)

$$\sum_{n\geq n_o} a_n = \frac{1}{2N} \sum_{n\geq n_o} a_n[2(n+1)N-2nN] = \infty,$$

in other words one has

$$\sum_{n\geq n_o} \mathbb{P}(\eta_n \leq a_n) = \infty,$$

whence by the BOREL-CANTELLI LEMMA

$$\mathbb{P}(\limsup_{n\to\infty} \{\eta_n \leq a_n\}) = 1.$$

From this we obtain, looking at the first order statistic $\eta_{n:1}$, that $\mathbb{P}$-a.s. one has $\eta_{n:1} \leq a_n$ for infinitely many n and this implies that $\mathbb{P}$-a.s.

$$\sup_{t\in(0,1)} \Psi(t)\ |U_n(t)-t| \geq \frac{1}{2}\,\Psi(\eta_{n:1})\ [U_n(\eta_{n:1})-U_n(\eta_{n:1}-0)]$$

$$\geq \frac{1}{2}\,\Psi(a_n)\,\frac{1}{n} = N \quad \text{for infinitely many n, which implies (+).} \quad \square$$

Remark. The validity of the first inequality at the end of the proof is based on the following fact which is easy to prove:

Given two df's F_1, F_2 on $\mathbb{R}$ and some strictly positive function h on some interval $(a,b)\subset\mathbb{R}$, then for any continuity point $t_o\in(a,b)$ of F_1 and of h one has

$$\sup_{a<t<b} \frac{|F_2(t)-F_1(t)|}{h(t)} \geq \frac{1}{2}\,\frac{F_2(t_o)-F_2(t_o-0)}{h(t_o)}.$$

THE VAPNIK-CHERVONENKIS THEORY:

There are various methods for proving GLIVENKO-CANTELLI-Theorems (i.e., a.s. convergence of empirical $\mathcal{C}$-discrepancies $D_n(\mathcal{C},\mu)$) in cases where a common geometrical structure for the sets in $\mathcal{C}$ is essentially used; see P. Gaenssler and W. Stute (1979), Section 1.1, for a survey on the results and methods of proof.

For arbitrary sample spaces where geometrical arguments are no longer available, perhaps the most striking method based on combinatorial arguments was developed by V.N. Vapnik and A.Ya. Chervonenkis (1971). We are going to present here their main results together with some extensions and applications.

In what follows, if not stated otherwise, let X be an arbitrary nonempty set and denote by $\mathcal{P}(X)$ the power set of X (i.e., the class of all subsets of X). For any set A, $|A|$ denotes its cardinality.

DEFINITION 1. Let $\mathcal{C}$ be an arbitrary subclass of $\mathcal{P}(X)$ and, for any $F\subset X$ with $|F| < \infty$, let

$$\Delta^{\mathcal{C}}(F) := |\{F\cap C\colon C\in\mathcal{C}\}|$$

be the number of different sets of the form $F\cap C$ for $C\in\mathcal{C}$.

Furthermore, for $r=0,1,2,\ldots$ let

$$m^{\mathcal{C}}(r) := \max\{\Delta^{\mathcal{C}}(F): |F| = r\},$$

and

$$V(\mathcal{C}) := \begin{cases} \inf\{r: m^{\mathcal{C}}(r)<2^r\} \\ \infty, \text{ if } m^{\mathcal{C}}(r)=2^r \text{ for all } r. \end{cases}$$

If $m^{\mathcal{C}}(r)<2^r$ for some r, i.e., if $V(\mathcal{C})<\infty$, $\mathcal{C}$ will be called a VAPNIK-CHERVONENKIS CLASS (VCC).

REMARKS.

a) $m^{\mathcal{C}}(\cdot)$ is called the GROWTH FUNCTION pertaining to $\mathcal{C}$. Note that $m^{\mathcal{C}}(r)\leq 2^r$ for all r and $m^{\mathcal{C}}(r)=2^r$ iff there exists an $F\subset X$ with $|F|=r$ such that for all $F'\subset F$ there exists a $C\in\mathcal{C}$ with $F\cap C=F'$; in other words: $m^{\mathcal{C}}(r)=2^r$ iff $\mathcal{C}$ cuts all subsets of some $F\subset X$ with $|F|=r$, saying that F is shattered by $\mathcal{C}$.
On the other hand, $m^{\mathcal{C}}(r)<2^r$ implies $m^{\mathcal{C}}(n)<2^n$ for all $n\geq r$.

b) EXAMPLES.

1) If $X=\mathbb{R}$ and $\mathcal{C} = \{(-\infty,t]: t\in\mathbb{R}\}$, then
$m^{\mathcal{C}}(r) = r+1$, whence $\mathcal{C}$ is a VCC with $V(\mathcal{C}) = 2$.

2) (cf. R.S. Wenocur and R.M. Dudley (1981)): More generally, let X be an arbitrary set with $|X|\geq 2$ and suppose that $\mathcal{C}\subset\mathcal{P}(X)$ fulfills the following condition (*):

(*) $\forall F\subset X$ with $|F|=2$ there exist $C_i\in\mathcal{C}$, $i=1,2$, such that $F\cap C_1=\emptyset$ and $F\cap C_2=F$.

(Note that (*) holds if $\{\emptyset,X\}\subset\mathcal{C}$.)

Then $\mathcal{C}$ is a VCC with $V(\mathcal{C})=2$ iff $\mathcal{C}$ is linearly ordered by inclusion.

PROOF. Assume to the contrary that $\mathcal{C}$ is not linearly ordered by inclusion. Then there exist $C',C''\in\mathcal{C}$ such that $C'\not\subset C''$ and $C''\not\subset C'$; choosing $x_1\in C'\setminus C''$ and $x_2\in C''\setminus C'$, it follows, together with (*), that $F := \{x_1,x_2\}$ is shattered by $\mathcal{C}$ which implies $V(\mathcal{C})>2$.

To prove the other direction, assume (cf. (*)) that $\mathcal{C}$ contains at least two elements and is linearly ordered by inclusion. Then, since $|\mathcal{C}|\geq 2$

implies $V(\mathcal{C})\geq 2$ (note that $|\mathcal{C}|=1$ iff $V(\mathcal{C})=1$), it remains to show that $V(\mathcal{C})\leq 2$:

For this, consider an $F\subset X$ with $|F|=2$; then, since $\mathcal{C}$ is linearly ordered by inclusion, there is at most one $F'\subset F$ with $|F'|=1$ and $F'=F\cap C$ for some $C\in\mathcal{C}$, showing that F is not shattered by $\mathcal{C}$. □

3) If $X=[0,1]$ and $\mathcal{C} := \{C\subset X: |C|<\infty\}$, then $m^{\mathcal{C}}(r)=2^r$ for all r, whence $V(\mathcal{C})=\infty$.

c) Let $\phi(v,r) := \sum_{j=0}^{v} \binom{r}{j}$, where $\binom{r}{j} := 0$ for $j>r$,

i.e.,
$$\phi(v,r) = \begin{cases} \sum_{j=0}^{v} \binom{r}{j}, & \text{if } v<r \\ 2^r, & \text{if } v\geq r. \end{cases}$$

(Note that $\phi(v,r)$ is the number of all subsets of an r-element set with at most v elements.)

Then it is easy to show that the following relations hold true:

$$\phi(v,r) = \phi(v,r-1) + \phi(v-1,r-1), \tag{12}$$

where $\phi(0,r) = 1$ and $\phi(v,0) = 1$;

$$\phi(v,r) \leq r^v+1 \text{ for all } v,r\geq 0. \tag{13}$$

The following remarks d) and e) are taken from R.M. Dudley (1978).

d) Let $\mathcal{H}_k$ be the collection of all open halfspaces in $\mathbb{R}^k$, $k\geq 1$, i.e., all sets of the form

$$\{x\in\mathbb{R}^k : (x,u)>c\} \quad \text{for} \quad 0 \neq u\in\mathbb{R}^k \text{ and some } c\in\mathbb{R},$$

and let $N_k(r)$ be the maximum number of open regions into which $\mathbb{R}^k$ is decomposed by r hyperplanes H_j

$$H_j = \{x\in\mathbb{R}^k : (x,u_j) = c_j\},\ j=1,\dots,r;$$

then the maximum number $N_k(r)$ is attained for $H_1,\dots,H_r$ in "general position" i.e. if any k or fewer of the u_j are linearly independent.

L. Schläfli (1901, posth.) showed that

(14) $$N_k(r) = \phi(k,r).$$

J. Steiner (1826) had proved this for $k \leq 3$.

e) If F is an r-element subset of $\mathbb{R}^k$, then

(15) $$\Delta^{\mathcal{H}_k}(F) \leq 2\phi(k,r-1)$$

and equality is attained if the points of F are in "general position", i.e. no k+1 of them are in any hyperplane (cf. T.M. Cover (1965); E.F. Harding (1967); D. Watson (1969)).

Therefore the growth function $m^{\mathcal{H}_k}(\cdot)$ pertaining to $\mathcal{H}_k$ satisfies

(16) $$m^{\mathcal{H}_k}(r) = 2\phi(k,r-1).$$

Without using (14)-(16), but directly from the definition of $\phi(v,r)$ and the recurrence relation (12), Vapnik and Chervonenkis ((1971), Lemma 1) proved the following lemma:

<u>Lemma 7.</u> If X is any set and if $\mathcal{C} \subset \mathcal{P}(X)$ is a Vapnik-Chervonenkis class (i.e., $V(\mathcal{C}) \leq v < \infty$ for some v), then

$$m^{\mathcal{C}}(r) < \phi(v,r) \text{ for all } r \geq v.$$

In view of (16) this implies that for arbitrary X one gets an upper estimate for the growth function $m^{\mathcal{C}}(\cdot)$ pertaining to a VCC $\mathcal{C} \subset \mathcal{P}(X)$ by the growth function $m^{\mathcal{H}_v}(\cdot)$ pertaining to the class $\mathcal{H}_v$ of all open halfspaces in $X = \mathbb{R}^v$, namely

(17) $$m^{\mathcal{C}}(r) < \frac{1}{2}\, m^{\mathcal{H}_v}(r+1) \quad \text{for all } r \geq v \geq V(\mathcal{C}).$$

Instead of Lemma 7 we shall show here a slightly sharper result whose nice proof (based on a proof of a more general result in J.M. Steele (1975)) I learned from David Pollard on occasion of one of his Seminar talks in Seattle (1982); cf. also N. Sauer (1972).

<u>Lemma 8 (Vapnik-Chervonenkis-Lemma).</u> Let X be any set and $\mathcal{C} \subset \mathcal{P}(X)$ be a Vapnik-Chervonenkis class (i.e., $V(\mathcal{C}) =: s < \infty$), then

$$m^{\mathcal{C}}(r) \leq \phi(s-1,r) \text{ for all } r \geq s.$$

Proof. Let $r \geq s$ be arbitrary but fixed. We have to show that for any $F \subset X$ with $|F|=r$

(a) $$\Delta^{\mathcal{C}}(F) := |\{F \cap C: C \in \mathcal{C}\}| \leq \phi(s-1,r).$$

Let $\{F_1,\ldots,F_p\}$ be the collection of all subsets of F of at least size s

(so $p = \binom{r}{s} + \binom{r}{s+1} + \ldots + \binom{r}{r}$).

Note that (a) is trivially fulfilled if

(b) $$F \cap C \neq F_i \text{ for all } i=1,\ldots,p \text{ and all } C \in \mathcal{C}.$$

Now, by assumption, we have:

(c) For each F_i there exists an $F_i^1 \subset F_i$ such that

$$F_i^1 \neq F_i \cap C \text{ for all } C \in \mathcal{C},$$

implying

(d) $$\{F \cap C: C \in \mathcal{C}\} \subset \mathcal{B}_1 := \{B \subset F: B \cap F_i \neq F_i^1 \text{ for all } i=1,\ldots,p\}.$$

In one special case the result follows readily, namely if $F_i^1 = F_i$ for all $i=1,\ldots,p$ since then $B \neq F_i$ for all $i=1,\ldots,p$ and each $B \in \mathcal{B}_1$ (which means that $\mathcal{B}_1$ cannot contain any subset of F of at least size s), so that (a) follows from (d).

We are going to show that by a successive modification of the F_i^1's the general case will reduce in a finite number of steps to this special case:

If $F_i^1 \neq F_i$ for some i, choose any $x^1 \in F$ and put

$$F_i^2 := (F_i^1 \cup \{x^1\}) \cap F_i,\ i=1,\ldots,p,$$

and define the corresponding class

$$\mathcal{B}_2 := \{B \subset F: B \cap F_i \neq F_i^2 \text{ for all } i=1,\ldots,p\}.$$

We will show below that

(e) $$|\mathcal{B}_1| \leq |\mathcal{B}_2|.$$

If now $F_i^2 = F_i$ for all i=1,...,p, then $\mathcal{B}_2$ cannot contain any subset of F of at least size s in which case (a) follows from (d) and (e).

If $F_i^2 \neq F_i$ for some i we go once more through the same argument, i.e., we choose any $x^2 \in F$, $x^2 \neq x^1$ and put

$$F_i^3 := (F_i^2 \cup \{x^2\}) \cap F_i, \ i=1,\dots,p,$$

$$\mathcal{B}_3 := \{B \subset F : B \cap F_i \neq F_i^3 \text{ for all } i=1,\dots,p\}$$

and show as in (e) that $|\mathcal{B}_2| \leq |\mathcal{B}_3|$.

So, another n-2 ($n \leq r$) repetitions of this argument would generate classes $\mathcal{B}_4,\dots,\mathcal{B}_n$ such that

$$|\mathcal{B}_1| \leq |\mathcal{B}_2| \leq |\mathcal{B}_3| \leq \dots \leq |\mathcal{B}_n|$$

with

$$\mathcal{B}_n = \{B \subset F : B \cap F_i \neq F_i^n \text{ for all } i=1,\dots,p\}$$

and $F_i^n = F_i$ for all i=1,...,p, which is the special case implying (a).

So it remains to prove (e):

For this it suffices to show that there exists a one-to-one map, say T, from $\mathcal{B}_1 \setminus \mathcal{B}_2$ into $\mathcal{B}_2 \setminus \mathcal{B}_1$.

Our claim is that $T(B) := B \setminus \{x^1\}$ is appropriate:

Let $B \in \mathcal{B}_1 \setminus \mathcal{B}_2$; then by definition of $\mathcal{B}_i$, i=1,2,

$$B \cap F_i \neq F_i^1 \text{ for all } i=1,\dots,p$$

$$\text{and} \quad B \cap F_j = F_j^2 \text{ for at least one } j \in \{1,\dots,p\}$$

implying that $x^1 \in \complement F_j^1$ whereas, by construction, $x^1 \in F_j^2$ and therefore $x^1 \in F_j$ and $x^1 \in B$; the last makes T one-to-one. It remains to show that $B \setminus \{x^1\} \in \mathcal{B}_2 \setminus \mathcal{B}_1$ for all $B \in \mathcal{B}_1 \setminus \mathcal{B}_2$. So, let $B \in \mathcal{B}_1 \setminus \mathcal{B}_2$; then

$$(B \setminus \{x^1\}) \cap F_j = (B \cap F_j) \setminus \{x^1\} = F_j^2 \setminus \{x^1\} = F_j^1, \text{ whence}$$

$(B \setminus \{x^1\}) \notin \mathcal{B}_1$; so we must finally show that $B \setminus \{x^1\} \in \mathcal{B}_2$, i.e. that

$$(+) \quad (B \setminus \{x^1\}) \cap F_i \neq F_i^2 \text{ for all } i=1,\dots,p.$$

<u>ad (+):</u> Let $i \in \{1,\dots,p\}$ be arbitrary but fixed; if $x^1 \in F_i$, then $x^1 \in F_i^2$,

but $x^1 \notin (B\setminus\{x^1\}) \cap F_i$, implying (+) in this case. If $x^1 \in \complement F_i$, i.e., $\{x^1\} \cap F_i = \emptyset$, then $F_i^2 = F_i^1$, whence $(B\setminus\{x^1\}) \cap F_i = B \cap F_i \neq F_i^1 = F_i^2$, implying (+) also in this case. This proves Lemma 8. □

The next lemma, being a consequence of Lemma 7 or Lemma 8, respectively, will be one of the key results used below.

Lemma 9. Let X be any set and $\mathcal{C}\subset\mathcal{P}(X)$ be a Vapnik-Chervonenkis class (i.e., $V(\mathcal{C}) =: s < \infty$); then

(i) $m^{\mathcal{C}}(r) \leq r^s$ for all $r\geq 2$, and

(ii) $m^{\mathcal{C}}(r) \leq r^s+1$ for all $r\geq 0$.

Proof. According to Lemma 7 and (13) we have

$m^{\mathcal{C}}(r) < \phi(s,r) \leq r^s+1$ for all $r\geq s$, whence (note that $m^{\mathcal{C}}(\cdot)$ is integer valued)

$$m^{\mathcal{C}}(r) \leq r^s \text{ for all } r\geq s;$$

if $2 \leq r \leq s$, it follows that $m^{\mathcal{C}}(r) \leq 2^r \leq 2^s \leq r^s$; this proves (i). Finally, for r=0 we have $m^{\mathcal{C}}(0) = 1 = 2^0$ (whence $s\geq 1$) $\leq 0^s+1$, and for r=1 we have $m^{\mathcal{C}}(1) \leq 2 = 1^s+1$, proving (ii). □

Besides Lemma 9, the following VAPNIK-CHERVONENKIS-INEQUALITIES are basic for the whole theory. We are going to present this part in a form strengthening the original bounds obtained by Vapnik and Chervonenkis. This will be done in a similar way as in a recent paper by L. Devroye (1981).

For this, let again $(X,\mathcal{B})$ be an arbitrary measurable space (Devroye (1981) considers only $(X,\mathcal{B}) = (\mathbb{R}^k, \mathcal{B}_k)$, $k\geq 1$), and let $(\xi_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random elements in $(X,\mathcal{B})$, defined on some common p-space $(\Omega,\mathcal{F},\mathbb{P})$, with distribution $\mu \equiv \mathcal{L}\{\xi_i\}$ on $\mathcal{B}$.
For $n,n' \in \mathbb{N}$ let μ_n and $\nu_{n'}$ be the empirical measures based on $\xi_1,\ldots,\xi_n$ and $\xi_{n+1},\ldots,\xi_{n+n'}$, respectively.

Let $\mathcal{C}$ be an arbitrary subset of $\mathcal{B}$, and let

$$D_n(\mathcal{C},\mu) := \sup_{C\in\mathcal{C}} |\mu_n(C) - \mu(C)|,$$

$$\overline{D}_{n,n'}(\mathcal{C}) := \sup_{C\in\mathcal{C}} |\mu_n(C) - \nu_{n'}(C)|,$$

where we assume that both $D_n(\mathcal{C},\mu)$ and $\overline{D}_{n,n'}(\mathcal{C})$ are measurable w.r.t. the canonical model (i.e., with $(X^{\mathbb{N}}, \mathcal{B}_{\mathbb{N}}, \times_{\mathbb{N}} \mu)$ as basic p-space and with ξ_i's being the coordinate projections of $X^{\mathbb{N}}$ onto X).

(Note that then $D_n(\mathcal{C},\mu)$ and $\overline{D}_{n,n'}(\mathcal{C})$ are also measurable considered as functions on the initially given p-space $(\Omega,\mathcal{F},\mathbb{P})$, since

$\omega \mapsto \xi(\omega) := (\xi_1(\omega),\xi_2(\omega),\ldots) \in X^{\mathbb{N}}$ is $\mathcal{F},\mathcal{B}_{\mathbb{N}}$-measurable.)

The proof of the following inequalities is patterned on the proof of Vapnik and Chervonenkis (1971). As a corollary we will obtain both, the fundamental Vapnik-Chervonenkis inequality and its improvement by Devroye (1981).

Lemma 10. For any $\varepsilon>0$, any $0<\alpha<1$, and any $n,n' \in \mathbb{N}$ one has

(a) $$\mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \le \left(1 - \frac{1}{4\alpha^2\varepsilon^2 n'}\right)^{-1} \mathbb{P}(\overline{D}_{n,n'}(\mathcal{C}) > (1-\alpha)\varepsilon),$$

and

(b) $$\mathbb{P}(\overline{D}_{n,n'}(\mathcal{C}) > (1-\alpha)\varepsilon) \le m^{\mathcal{C}}(n+n')\cdot 2\cdot\exp\left[-2n\left(\frac{n'}{n+n'}\right)^2(1-\alpha)^2\varepsilon^2\right],$$

where $m^{\mathcal{C}}(\cdot)$ denotes the growth function pertaining to the class $\mathcal{C}$.

Before proving this lemma, let us point out the following facts:
We have that

$$\overline{D}_{n,n'}(\mathcal{C})(\omega) = \sup_{C\in\mathcal{C}} |\mu_n(C,\omega) - \nu_{n'}(C,\omega)|$$

$$= h_{n,n'}(\xi(\omega)) \text{ for } \xi(\omega) = (\xi_1(\omega),\xi_2(\omega),\ldots) \in X^{\mathbb{N}},$$

where $h_{n,n'}: X^{\mathbb{N}} \longrightarrow \mathbb{R}$, defined by

$$h_{n,n'}(\underline{x}) := \sup_{C\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^{n} 1_C(x_i) - \frac{1}{n'}\sum_{i=n+1}^{n+n'} 1_C(x_i)\right|,$$

for $\underline{x} = (x_1,x_2,\ldots) \in X^{\mathbb{N}}$, is, by assumption, $\mathcal{B}_{\mathbb{N}},\mathcal{B}$-measurable, whence

$$A := \{\underline{x} \in X^{\mathbb{N}} : \bar{D}_{n,n'}(C) > (1-\alpha)\varepsilon\} \in \mathcal{B}_{\mathbb{N}},$$

and therefore

$$\mathbb{P}(\{\omega \in \Omega: \bar{D}_{n,n'}(C)(\omega) > (1-\alpha)\varepsilon\}) = L\{\xi\}(A)$$

$$= \int_{X^{\mathbb{N}}} u(h_{n,n'} - (1-\alpha)\varepsilon)\ d(\times_{\mathbb{N}} \mu) \quad \text{where } u(z) := \begin{cases} 1, & \text{if } z>0 \\ 0, & \text{if } z\leq 0. \end{cases}$$

Noticing further that $h_{n,n'}$ depends only on the first n and the following n' coordinates of $\underline{x}$, we obtain using the notation

$$x^1 := (x_1,\dots,x_n),\ x^2 := (x_{n+1},\dots,x_{n+n'}),$$

$$X^1 := \mathop{\times}_{1}^{n} X,\ X^2 := \mathop{\times}_{n+1}^{n+n'} X,\ P^1 := \mathop{\times}_{1}^{n} \mu,\ P^2 := \mathop{\times}_{n+1}^{n+n'} \mu$$

and $P := P^1 \times P^2$, that

(*) $$\mathbb{P}(\bar{D}_{n,n'}(C) > (1-\alpha)\varepsilon) = \int_{X^1\times X^2} u(h_{n,n'}(x^1,x^2)-(1-\alpha)\varepsilon)\ P(dx^1,dx^2).$$

In the very same way one has

(**) $$\mathbb{P}(D_n(C,\mu) > \varepsilon) = P^1(\{x^1=(x_1,\dots,x_n) \in X^1: \sup_{C\in\mathcal{C}} |\tfrac{1}{n} \sum_{i=1}^{n} 1_C(x_i)-\mu(C)| > \varepsilon\}).$$

Proof of inequality (a) in Lemma 10.

According to (*) of our remarks made before, one has

$$\mathbb{P}(\bar{D}_{n,n'}(C) > (1-\alpha)\varepsilon) = \int_{X^1\times X^2} u(h_{n,n'} - (1-\alpha)\varepsilon)\ d(P^1\times P^2)$$

$$\overset{\text{(Fubini)}}{=} \int_{X^1}\Big[\int_{X^2} u(h_{n,n'} - (1-\alpha)\varepsilon)\ dP^2\Big]\ dP^1$$

$$\geq \text{ⓐ} := \int_{A^1}\Big[\int_{X^2} u(h_{n,n'} - (1-\alpha)\varepsilon)\ dP^2\Big]\ dP^1$$

with $A^1 := \{x^1 \in X^1: \sup_{C\in\mathcal{C}} |\mu_n(C,x^1)-\mu(C)| > \varepsilon\}$,

where $$\mu_n(C,x^1) = \frac{1}{n}\sum_{i=1}^{n} 1_C(x_i) \text{ for } x^1 = (x_1,\dots,x_n) \in X^1.$$

But for any $(x^1,x^2) \in X^1\times X^2$ with $x^1 \in A^1$ there exists a $C_{x^1} \in \mathcal{C}$

such that

$$|\mu_n(C_{x^1},x^1) - \mu(C_{x^1})| > \varepsilon,$$

whence (for $\nu_{n'}(C,x^2) = \frac{1}{n'} \sum_{i=n+1}^{n+n'} 1_C(x_i)$ with $x^2 = (x_{n+1},\ldots,x_{n+n'}) \in X^2$)

we have

$$|\mu_n(C_{x^1},x^1) - \nu_{n'}(C_{x^1},x^2)| > (1-\alpha)\varepsilon$$

$$\text{if} \qquad |\nu_{n'}(C_{x^1},x^2) - \mu(C_{x^1})| \leq \alpha\varepsilon;$$

therefore, we obtain for all $x^1 \in A^1$ the following estimate for the inner integral in ⓐ:

$$\int_{X^2} u(h_{n,n'}(x^1,x^2) - (1-\alpha)\varepsilon)\, P^2(dx^2)$$

$$\geq P^2(\{x^2 \in X^2 : |\nu_{n'}(C_{x^1},x^2) - \mu(C_{x^1})| \leq \alpha\varepsilon\}).$$

But, by Tschebyschev's inequality,

$$P^2(\{x^2 \in X^2 : |\nu_{n'}(C_{x^1},x^2) - \mu(C_{x^1})| > \alpha\varepsilon\})$$

$$\leq \frac{1}{\alpha^2\varepsilon^2} \cdot \frac{\mu(C_{x^1})\ (1-\mu(C_{x^1}))}{n'} \leq \frac{1}{4\alpha^2\varepsilon^2 n'};$$

thus, summarizing we obtain

$$\mathbb{P}(\overline{D}_{n,n'}(C) > (1-\alpha)\varepsilon) \geq P^1(A^1)(1 - \frac{1}{4\alpha^2\varepsilon^2 n'})$$

$$\underset{(**)}{=} \mathbb{P}(D_n(C,\mu) > \varepsilon)(1 - \frac{1}{4\alpha^2\varepsilon^2 n'})$$

which proves (a).

Proof of inequality (b) in Lemma 10.

According to our remarks preceeding the proof we may and will consider $\overline{D}_{n,n'}(C)$ as a function $h_{n,n'}$ of

$$x = (x^1,x^2) \in X^1 \times X^2 \quad \text{with} \quad x^1 = (x_1,\ldots,x_n) \text{ and } x^2 = (x_{n+1},\ldots,x_{n+n'}).$$

Due to the symmetry of $P = P^1 \times P^2$ (w.r.t. coordinate permutations) one has for each $f \in L^1(X^1 \times X^2, \bigotimes_1^{n+n'} \mathcal{B}, P)$

$$\int_{X^1 \times X^2} f(x)\, P(dx) = \int_{X^1 \times X^2} f(T_i x) P(dx)$$

for every permutation $T_i x$ of x (which means that $T_i x$ is the image of x when applying a permutation T_i to the $n+n'$ components of x).

Therefore,

$$\mathbb{P}(\bar{D}_{n,n'}(C) > (1-\alpha)\varepsilon) = \int_{X^1 \times X^2} u(h_{n,n'}(x) - (1-\alpha)\varepsilon)\, P(dx)$$

$$= \int_{X^1 \times X^2} \left[\frac{1}{(n+n')!} \sum_i u(h_{n,n'}(T_i x) - (1-\alpha)\varepsilon)\right] P(dx),$$

where the summation w.r.t. i is over all $(n+n')!$ permutations T_i.

We also remark here for later use in the proof that

$$u(h_{n,n'} - (1-\alpha)\varepsilon) = \sup_{C \in \mathcal{C}} u(h^C_{n,n'} - (1-\alpha)\varepsilon),$$

where $h^C_{n,n'}(x) := |\mu_n(C,x^1) - \nu_{n'}(C,x^2)|$ for

$x = (x^1,x^2) \in X^1 \times X^2$.

Next, let $x = (x^1,x^2) \in X^1 \times X^2$ (with $x^1 = (x_1,\ldots,x_n)$ and $x^2 = (x_{n+1},\ldots,x_{n+n'})$) be arbitrary but fixed and put $F_x := \{x_1,\ldots,x_n,x_{n+1},\ldots,x_{n+n'}\}$; then for any $C_1,C_2 \in \mathcal{C}$ one has that

$F_x \cap C_1 = F_x \cap C_2$ implies $h^{C_1}_{n,n'}(T_i x) = h^{C_2}_{n,n'}(T_i x)$ for all T_i.

Hence, denoting with $\mathcal{C}_x$ a subclass of $\mathcal{C}$ such that for any two $C_1,C_2 \in \mathcal{C}_x$, $F_x \cap C_1 \neq F_x \cap C_2$, and such that at the same time for any $C \in \mathcal{C}$ there exists a $C_x \in \mathcal{C}_x$ with $F_x \cap C_x = F_x \cap C$, we obtain for all T_i

$$\sup_{C \in \mathcal{C}} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon) = \sup_{C \in \mathcal{C}_x} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)$$

$$\leq \sum_{C\in\mathcal{C}_x} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon).$$

For later use in the proof we note here also that

$$|\mathcal{C}_x| = \Delta^{\mathcal{C}}(F_x) \quad \text{(cf. DEFINITION 1)}$$

for every $x = (x^1,x^2) \in X^1 \times X^2$.

It follows that

$$\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} u(h_{n,n'}(T_i x) - (1-\alpha)\varepsilon)$$

$$= \frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} \sup_{C\in\mathcal{C}} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)$$

$$\leq \sum_{C\in\mathcal{C}_x} \left[\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)\right].$$

Note that for each fixed x and C

$$\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)$$

is the fraction of all (n+n')! permutations $T_i x$ of x for which

$$|\mu_n(C,(T_i x)^1) - \nu_{n'}(C,(T_i x)^2)| > (1-\alpha)\varepsilon.$$

Now, for x and C being arbitrary but fixed, put

$$\eta_j := \begin{cases} 1, & \text{if } x_j \in C \\ 0, & \text{if } x_j \in \complement C \end{cases}, \quad j=1,\dots,n+n', \text{ and denote by}$$

$(\eta_1^{(i)},\dots,\eta_{n+n'}^{(i)})$ the vector $T_i\eta$ for $\eta = (\eta_1,\dots,\eta_{n+n'})$.

Consider then the p-space $(\Omega_o,\mathcal{A}_o,\mathbb{P}_o)$ with Ω_o being the set of all (n+n')! permutations T_i, $\mathcal{A}_o := \mathcal{P}(\Omega_o)$, and

$$\mathbb{P}_o(A) := \frac{|A|}{(n+n')!}, \quad A \in \mathcal{A}_o.$$

Then, using the random variables $\zeta_j : \Omega_o \longrightarrow \{0,1\}$, defined by $\zeta_j(T_i) := \eta_j^{(i)}$, $T_i \in \Omega_o$, $j=1,\dots,n+n'$, we obtain

$$\frac{1}{(n+n')!} \sum_{i=1}^{n} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)$$

$$= \mathbb{P}_o(|\frac{1}{n} \sum_{j=1}^{n} \zeta_j - \frac{1}{n'} \sum_{j=1}^{n'} \zeta_{n+j}| > (1-\alpha)\varepsilon)$$

$$= \mathbb{P}_o(|\frac{1}{n} \sum_{j=1}^{n} \zeta_j - \frac{1}{n'} [(n+n')\mu_{n+n'}(C,x) - \sum_{j=1}^{n} \zeta_j]| > (1-\alpha)\varepsilon)$$

$$= \mathbb{P}_o(|\frac{n'}{n} \sum_{j=1}^{n} \zeta_j - (n+n')\mu_{n+n'}(C,x) + \frac{n}{n} \sum_{j=1}^{n} \zeta_j| > (1-\alpha)\varepsilon)$$

$$= \mathbb{P}_o(|\frac{1}{n} \sum_{j=1}^{n} \zeta_j - \mu_{n+n'}(C,x)| > \frac{n'}{n+n'} (1-\alpha)\varepsilon)$$

$$\leqq 2 \cdot \exp\,[-2n\,(\frac{n'}{n+n'})^2(1-\alpha)^2\varepsilon^2],$$

using Hoeffding's inequality for sampling without replacement from n+n' binary-valued random variables with sum $(n+n')\mu_{n+n'}(C,x)$; cf. W. Hoeffding (1963) and R.J. Serfling (1974).

Summarizing we thus obtain for every $x = (x^1,x^2) \in X^1 \times X^2$

$$\sum_{C\in\mathcal{C}} [\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)]$$

$$\leqq |\mathcal{C}_x|\cdot(2\cdot\exp\,[-2n(\frac{n'}{n+n'})^2(1-\alpha)^2\varepsilon^2]) = \Delta^{\mathcal{C}}(F_x)\cdot(\dots\dots\dots)$$

$$\leqq m^{\mathcal{C}}(n+n')\cdot(\dots\dots\dots),\ \text{and therefore}$$

$$\mathbb{P}(\overline{D}_{n,n'}(\mathcal{C}) > (1-\alpha)\varepsilon) = \int_{X^1\times X^2} [\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} u(h_{n,n'}(T_i x)-(1-\alpha)\varepsilon)]\, P(dx)$$

$$\leqq \int_{X^1\times X^2} (\sum_{C\in\mathcal{C}_x} [\frac{1}{(n+n')!} \sum_{i=1}^{(n+n')!} u(h^C_{n,n'}(T_i x) - (1-\alpha)\varepsilon)])\, P(dx)$$

$$\leq m^{\mathcal{C}}(n+n')\cdot 2\cdot \exp\,[-2n(\tfrac{n'}{n+n'})^2(1-\alpha)^2\varepsilon^2]$$

which concludes the proof of (b). □

NOTE: We have in fact shown, assuming measurability of

$$\Delta^{\mathcal{C}}(\{\xi_1,\ldots,\xi_n,\xi_{n+1},\ldots,\xi_{n+n'}\}),\ \text{that}$$

$$\mathbb{P}(\overline{D}_{n,n'}(\mathcal{C}) > (1-\alpha)\varepsilon) \leq 2\cdot\exp[-2n(\tfrac{n'}{n+n'})^2(1-\alpha)^2\varepsilon^2]\ \mathbb{E}(\Delta^{\mathcal{C}}(\{\xi_1,\ldots,\xi_{n+n'}\}));$$

in many cases this bound is considerably smaller than the r.h.s. of (b).

COROLLARY.

(i) Vapnik-Chervonenkis (1971).

Taking $\alpha = \frac{1}{2}$ and $n' = n$, one gets

$$\mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \leq 4\cdot m^{\mathcal{C}}(2n)\cdot\exp\,(-\frac{\varepsilon^2 n}{8})\ \text{for all}\ n \geq 2/\varepsilon^2.$$

(ii) Devroye (1981).

Taking $\alpha = \frac{1}{n\varepsilon}$ and $n' = n^2-n$, one gets

$$\mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \leq 4\cdot\exp(4\varepsilon + 4\varepsilon^2)\cdot m^{\mathcal{C}}(n^2)\cdot\exp(-2n\varepsilon^2)$$

for all $n > \max(\frac{1}{\varepsilon}, 2)$.

Proof. (i): It follows from (a) and (b) in Lemma 10 that in the present case

$$\mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \leq (1 - \frac{1}{\varepsilon^2 n})^{-1}\cdot m^{\mathcal{C}}(2n)\cdot 2\cdot\exp\,[-2n\cdot\tfrac{1}{4}\cdot\tfrac{1}{4}\,\varepsilon^2]$$

$$\underset{(n\varepsilon^2\geq 2)}{\leq}\ 4\cdot m^{\mathcal{C}}(2n)\exp(-\frac{\varepsilon^2 n}{8}).$$

(ii): Again (a) and (b) in Lemma 10 yield in the present case

$$\mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \geq (1 - \frac{n^2}{4(n^2-n)})^{-1}\cdot m^{\mathcal{C}}(n^2)2\exp[-2n(\tfrac{n-1}{n})^2(\varepsilon^2 - 2\alpha\varepsilon^2 + \alpha^2\varepsilon^2)]$$

$$\underset{((\frac{n-1}{n})^2 \geq 1-\frac{2}{n})}{\leq} \frac{4(n-1)}{3n-4} \cdot m^{\mathcal{C}}(n^2) \times$$

$$\times\ 2\exp\left[-2n(\varepsilon^2-2\alpha\varepsilon^2+\alpha^2\varepsilon^2)+4(\varepsilon^2-2\alpha\varepsilon^2+\alpha^2\varepsilon^2)\right]$$

$$\underset{(n \geq 2)}{\leq} 4\cdot m^{\mathcal{C}}(n^2)\cdot\exp[-2n\varepsilon^2+4\alpha n\varepsilon^2-2n\alpha^2\varepsilon^2+4\varepsilon^2+4\alpha^2\varepsilon^2]$$

$$\underset{(n \geq 2)}{\leq} 4\cdot m^{\mathcal{C}}(n^2)\exp[-2n\varepsilon^2+4\varepsilon+4\varepsilon^2]$$

$$= 4\cdot\exp(4\varepsilon+4\varepsilon^2)m^{\mathcal{C}}(n^2)\exp(-2n\varepsilon^2). \quad \square$$

Based on Lemma 9 (i) and on part (i) of the corollary to Lemma 10 we now obtain the main result of Vapnik and Chervonenkis concerning almost sure convergence of empirical $\mathcal{C}$-discrepancies in arbitrary sample spaces.

THEOREM 2. Let $(X,\mathcal{B})$ be an arbitrary measurable space and let $(\xi_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random elements in $(X,\mathcal{B})$, defined on some common p-space $(\Omega,\mathcal{F},\mathbb{P})$, with distribution $\mu \equiv \mathcal{L}\{\xi_i\}$ on $\mathcal{B}$.

For $n\in\mathbb{N}$ let μ_n and ν_n be the empirical measures based on $\xi_1,\ldots,\xi_n$ and $\xi_{n+1},\ldots,\xi_{2n}$, respectively.

Let $\mathcal{C}\subset\mathcal{B}$ be a VCC such that both $D_n(\mathcal{C},\mu)$ as well as $\overline{D}_{n,n}(\mathcal{C})$ are measurable w.r.t. the canonical model; then

$$\lim_{n\to\infty} D_n(\mathcal{C},\mu) = 0 \quad \mathbb{P}\text{-a.s.}$$

Proof. Of course, it suffices to show that

$$(*) \qquad \limsup_{n\to\infty} D_n(\mathcal{C},\mu) \leq \varepsilon \quad \mathbb{P}\text{-a.s. for every } \varepsilon>0;$$

according to the Borel-Cantelli Lemma, (*) is implied by

$$(**) \qquad \sum_{n\in\mathbb{N}} \mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) < \infty \text{ for every } \varepsilon>0,$$

whence the proof will be concluded by showing that (**) holds true.

ad (**): Given any $\varepsilon>0$, we obtain from part (i) of the corollary to Lemma 10 that for all $n \geqq 2/\varepsilon^2$

$$\mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \leqq 4m^{\mathcal{C}}(2n)\cdot\exp(-\frac{\varepsilon^2 n}{8}).$$

Since, by assumption, $\mathcal{C}$ is a VCC, we have $V(\mathcal{C}) =: s < \infty$, whence by Lemma 9 (i)

$$\sum_{n\in\mathbb{N}} \mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) \leqq \sum_{n<2/\varepsilon^2} \mathbb{P}(D_n(\mathcal{C},\mu) > \varepsilon) +$$

$$+ 4\cdot \sum_{n\geqq 2/\varepsilon^2} (2n)^s\cdot\exp(-\frac{\varepsilon^2 n}{8}) < \infty. \quad \square$$

The proof shows that the assumption of $\mathcal{C}$ being a VCC was essentially used to the amount that in this case the growth function $m^{\mathcal{C}}(r)$ is majorized by r^s for $r \geqq 2$ (with s being the minimal r for which $m^{\mathcal{C}}(r) < 2^r$); without this assumption, i.e., in case that $m^{\mathcal{C}}(r) = 2^r$ for all r, we would have arrived at

$$4\cdot \sum_{n\geqq 2/\varepsilon^2} 2^{2n} \exp(-\frac{\varepsilon^2 n}{8}) = \infty.$$

Thus, Theorem 2 can be restated as follows:

(18) If for a given $\mathcal{C}\subset\mathcal{B}$ there exists an $s<\infty$ such that $\mathcal{C}$ does not shatter any $F\subset X$ with $|F|=s$ (i.e., for any $F\subset X$ with $|F|=s$ there exists an $F'\subset F$ s.t. $F' \neq F \cap C$ for all $C \in \mathcal{C}$), then $\mathcal{C}$ is a GLIVENKO-CANTELLI-class (i.e., $\lim_{n\to\infty} D_n(\mathcal{C},\mu) = 0$ $\mathbb{P}$-a.s.), provided that the measurability assumptions stated in Theorem 2 are fulfilled.

The following example shows that these measurability assumptions cannot be dispensed with, in general.

(19) EXAMPLE (cf. M. Durst and R.M. Dudley (1980)).

Let $X = (X,<)$ be an uncountable wellordered set such that all its initial segments $\{x\in X: x<y\}$, $y \in X$, are countable (cf. J. Kelley

(1961), p. 29 -).

Then $\mathcal{C} := \{\{x \in X: x < y\}, y \in X\}$ does not shatter any $F \subset X$ with $|F| = 2$ (in fact: for any $F = \{x_1,x_2\} \subset X$ with $x_1 < x_2$ we have $\{x_2\} \neq F \cap C$ for all $C \in \mathcal{C}$, since $x_2 \in C$ would necessarily imply that $x_1 \in C$ for all $C \in \mathcal{C}$).

Note that $\mathcal{C}$ is linearly ordered by inclusion! To complete the example, let $\mathcal{B} := \{A \subset X: A \text{ countable or } \complement A \text{ countable}\}$, and let μ on $\mathcal{B}$ be defined by

$$\mu(A) := \begin{cases} 0, & \text{if } A \text{ is countable} \\ 1, & \text{if } \complement A \text{ is countable.} \end{cases}$$

Then $\mathcal{C} \subset \mathcal{B}$ and $\mu(C) = 0$ for all $C \in \mathcal{C}$; on the other hand, given any observations x_i, $i=1,\ldots,n$, of i.i.d. random elements ξ_i in $(X,\mathcal{B})$ with distribution μ, there exists a $C \in \mathcal{C}$ s.t. $x_i \in C$ for all $i=1,\ldots,n$, whence $D_n(\mathcal{C},\mu) \equiv 1$.

Note that in the present situation $\bar{D}_{n,n}(\mathcal{C})$ <u>fails</u> to be measurable w.r.t. the canonical model (cf. Theorem 2). In fact, consider for simplicity n=1, i.e., $\Omega = X \times X$, $\mathcal{F} = \mathcal{B} \otimes \mathcal{B}$, $\mathbb{P} = \mu \times \mu$ with ξ_1 and ξ_2 being the projections of Ω onto the first and second coordinate, respectively. Then

$$\bar{D}_{1,1}(\mathcal{C}) = \sup_{C \in \mathcal{C}} |\mu_1(C) - \nu_1(C)| = 1_{\complement\Delta},$$

where Δ denotes the diagonal in $X \times X$ which is <u>not</u> contained in $\mathcal{F}$: note that $\Delta \in \mathcal{B} \otimes \mathcal{B}$ iff there exists a countable subsystem $\mathcal{E}$ of $\mathcal{B}$ which is separating in the sense that

(+) for any $x,y \in X$ with $x \neq y$ there exists an $E \in \mathcal{E}$ such that $1_E(x) \neq 1_E(y)$; but in the present situation it can easily be shown that any countable subsystem $\mathcal{E}$ of $\mathcal{B}$ does <u>not</u> have the property (+) which implies that

$$\Delta \notin \mathcal{B} \otimes \mathcal{B} = \mathcal{F}.$$

Thus, although $D_n(\mathcal{C},\mu) \equiv 1$ is measurable w.r.t. the canonical model, $\bar{D}_{n,n}(\mathcal{C})$ is <u>not</u> in the present case.

We will show later (Section 4) that in $X = \mathbb{R}^k$, $k \geq 1$, the class $\mathbb{B}_k$ of all closed Euclidean balls fulfills the measurability assumptions made in Theorem 2. As shown by R.M. Dudley (1979) one has $V(\mathbb{B}_k) = k+2$, implying that $\mathbb{B}_k$ is a VCC; therefore, by (18) with $s = k+2$, we obtain the GLIVENKO-CANTELLI result (11) stated earlier without proof.

We are going to present here an independent nice proof of (11) which I learned from F. Topsøe ((1976), personal communication); this proof is based on the following two auxiliary results (20) and (21).

(20) <u>RADON'S THEOREM</u> (cf. F. Valentine (1964), Theorem 1.26).

Any $F \subset \mathbb{R}^k$, $k \geq 1$, with $|F| \geq k+2$, can be decomposed into two (disjoint) subsets F_i, $i=1,2$, such that

$$co(F_1) \cap co(F_2) \neq \emptyset,$$

where $co(F_i)$ denotes the convex hull of F_i.

(21) <u>(SEPARATION PROPERTY).</u> For any two $C_1, C_2 \in \mathbb{B}_k$ one has

$$co(C_1 \setminus C_2) \cap co(C_2 \setminus C_1) = \emptyset.$$

Now, according to (18), in order to prove (11) it suffices to show that $\mathbb{B}_k$ does not shatter any $F \subset \mathbb{R}^k$ with $|F| = s := k+2$.
Suppose to the contrary that there exists an $F \subset \mathbb{R}^k$ with $|F| = k+2$ such that for every $F_o \subset F$ there exists a $C \in \mathbb{B}_k$ with $F \cap C = F_o$.
This implies that for the F_i's of (20) which decompose a given $F \subset \mathbb{R}^k$ with $|F| = k+2$, there exist $C_i \in \mathbb{B}_k$ such that $F_i = F \cap C_i$, $i=1,2$.
Since $F_1 \cap F_2 = \emptyset$, we have

$$F_1 \subset C_1 \setminus C_2 \text{ and } F_2 \subset C_2 \setminus C_1,$$

and therefore

$$\mathrm{co}(C_1 \setminus C_2) \cap \mathrm{co}(C_2 \setminus C_1) \supset \mathrm{co}(F_1) \cap \mathrm{co}(F_2) \neq \emptyset \text{ (by (20))}$$

which contradicts (21). □

As the proof has shown, the separation property of the class $\mathcal{C} = \mathbb{B}_k$ was essential; at the same time the proof has shown that in general one has the following result (again under appropriate measurability assumptions as in Theorem 2):

(22) If a given class $\mathcal{C} \subset \mathcal{B}_k$ in $X = \mathbb{R}^k$, $k \geqq 1$, fulfills the separation property, then $\mathcal{C}$ is a VCC and therefore also a GLIVENKO-CANTELLI class.

Let me conclude this section with the following

CONJECTURE: The class of all translates of a fixed convex set in $X = \mathbb{R}^k$, $k \geqq k_o$, is, in general, not a VCC; at least it does not fulfill the separation property: in fact, consider the class of all translates of a tetrahedron C in $\mathbb{R}^3$, then the situation looks like this where you (hopefully) can see that for $C_z := C + z$ one has $C \setminus C_z = C \setminus \{x\}$ and $C_z \setminus C = C_z \setminus \{x\}$, whence

$$\mathrm{co}(C \setminus C_z) \cap \mathrm{co}(C_z \setminus C) = \{x\} \text{ (cf. Figure 2);}$$

I am grateful to Professors K. Seebach and R. Fritsch (Munich) for pointing out to me this example.

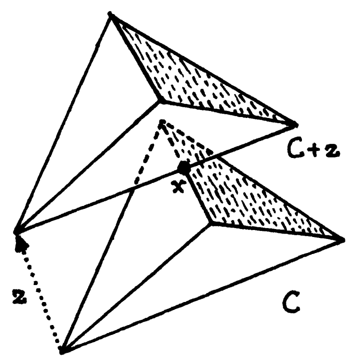

FIGURE 2

Note added in proof: As pointed out by a referee, translates of multiples of a fixed convex set need not be a GCC nor VCC: cf. Elker, Pollard and Stute (1979), Adv. Appl. Prob. 11, p. 830.