An Introduction to Functional Central Limit Theorems for Dependent Stochastic Processes

Donald W.K. Andrews† and David Pollard‡

1 Department of Economics, Yale University, Box 208281 Yale Station, New Haven, CT 06520-8281
2 Department of Statistics, Yale University, Box 208290 Yale Station, New Haven, CT 06520-8290

Summary

This paper shows how the modern machinery for generating abstract empirical central limit theorems can be applied to arrays of dependent variables. It develops a bracketing approximation (closely related to results of Philipp and Massart) based on a moment inequality for sums of strong mixing arrays, in an effort to illustrate the sorts of difficulty that need to be overcome when adapting the empirical process theory for independent variables. Some suggestions for further development are offered. The paper is largely self-contained.

Key words: Strong mixing; Functional central limit theorem; Empirical process.

1 Introduction

Since the landmark paper of Dudley (1978), there have been many generalizations to abstract settings of Donsker's theorem for the empirical distribution function. Much of the generalized theory has treated empirical processes for independent summands, in contrast with the development of the theory in the one-dimensional case, where results for various types of dependence were discovered early. For example, Chapter 4 of Billingsley's (1968) influential book treated $\phi$-mixing sequences.

We embarked upon the work leading to this paper in response to the often posed question, How much of the abstract theory for independent processes carries over to the dependent case? Some subtle difficulties made the task less straightforward than expected. Eventually we developed several techniques that work for various types of mixing, but not without many fruitless excursions into the literature. Both referees later pointed out that part of our efforts was spent in rediscovering ideas already contained in the work of Philipp (1982, and work cited therein) and Massart (1987).

As a guide to others who might want to extend empirical process theory to different types of dependent variables, we present in this paper a self-contained treatment of an illustrative empirical central limit theorem for strong mixing triangular arrays of random processes. The result is close to Philipp's (1982) Theorem 6.1 and Massart's (1987) Theorem 4, although there are small technical differences that might appeal to the connoisseur. We have tried to simplify and streamline the argument as much as we could, to make the basic idea as clear as possible. The formal statement of our limit theorem

† Supported by NSF Grants SES-8812021 and SES-9121914.
‡ Supported by NSF Grants DMS-8806900 and DMS-9001088.
appears in Section 2. An outline of general principles and a discussion of some ways in which our theorem could be modified or strengthened appear in Section 5, together with references to the literature.

Our results apply to an empirical process \( v_n \) indexed by a class of functions \( \mathcal{F} \),

\[
v_n f = n^{-1} \sum_{i=1}^{n} (f(\xi_n) - \mathbb{P}(\xi_n)),
\]

where \( \{\xi_n; i \leq n, n = 1, 2, \ldots\} \) is a strong mixing triangular array. An empirical central limit theorem (Corollary 2.3) gives conditions under which \( v_n \) converges in distribution, as a stochastic process indexed by \( \mathcal{F} \), to an appropriate Gaussian process. The proof of such a theorem consists of the usual two steps: establish convergence of finite-dimensional distributions; then establish stochastic equicontinuity, a close relative of the familiar uniform tightness property. The literature already contains several results that can handle finite-dimensional convergence. Our Theorem 2.2 gives sufficient conditions for a convenient strengthening of stochastic equicontinuity. With an appropriate seminorm \( \rho(\cdot) \) on \( \mathcal{F} \) and an appropriate \( \mathcal{L}^2 \) norm, it gives for each \( \epsilon > 0 \) a \( \delta > 0 \) such that

\[
\lim_{n \to \infty} \sup_{f \in \mathcal{F}, \|f\|_2 \leq \delta} \left\| v_n f - v_n g \right\|_2 < \epsilon.
\]

The proof is largely a repeated application of a moment inequality (Lemma 3.1) for sums of strong mixing sequences.

Stochastic equicontinuity is a most useful property even apart from its role in the functional central limit theorem. It implies that \( |v_n f_n - v_n g_n| \to 0 \) in probability for all sequences \( \{f_n, g_n\} \), possibly random, from \( \mathcal{F} \) such that \( \rho(f_n, g_n) \to 0 \) in probability. Andrews (1994) has shown how this form of stochastic equicontinuity is the key to many semiparametric limit theorems; it was also the main hypothesis in the general central limit theorem for minimization estimators, established in Section VII.1 of Pollard (1984). We present a typical application in Section 4. By establishing conditions for stochastic equicontinuity under strong mixing assumptions, one automatically extends the range of application of all those results.

2 Definitions and Statement of Results

Let \( \{\xi_n; i \leq n, n = 1, 2, \ldots\} \) be a triangular array of random elements of a measurable space \( S \). Define \( \mathcal{A}_n(m) \) to be the \( \sigma \)-field generated by the variables \( \xi_i \) for \( i \leq m \) and \( \mathcal{M}_n(m + d) \) to be the \( \sigma \)-field generated by the variables \( \xi_i \) for \( i > m + d \). We say that \( \{\xi_n\} \) is strong mixing if there is a sequence of numbers \( \{\alpha(d)\} \) converging to zero for which

\[
|\mathbb{P}(A M) - \mathbb{P}(A) \mathbb{P}(M)| = \alpha(d)
\]

for all \( A \in \mathcal{A}_n(m) \), all \( M \in \mathcal{M}_n(m + d) \), all \( m, d, n \).

We define a uniform analogue of the \( \mathcal{L}^2 \) norm by \( \rho(f) = \sup_{\|f\|_2} \|f(\xi_n)\|_2 \). [In general, we write \( \|Z\|_p \) for the \( \mathcal{L}^p \) norm \( \mathbb{P}[|Z|^p]^{1/p} \) of a random variable \( Z \).]

Results for triangular arrays are more powerful than their analogues for a single strong mixing sequence. For example, local power calculations and asymptotic minimax theorems require triangular arrays.

We will establish a maximal inequality for the empirical process indexed by a class of functions \( \mathcal{F} \), with a bound involving a measure of complexity for \( \mathcal{F} \) based on the concept of bracketing. There are several slightly different definitions of bracketing in the literature. We have chosen a form that fits most easily into our method of proof.

Definition 2.1. The bracketing number \( N(\delta) = N(\delta, \mathcal{F}) \) equals the smallest value of \( N \) for which there exist functions \( f_1, \ldots, f_N \) in \( \mathcal{F} \) and \( b_1, \ldots, b_N \) with \( \rho(b_i) = \delta \) for each \( i \) such that: for each \( f \) in \( \mathcal{F} \) there exists an \( i \) for which \( |f - f_i| < b_i \).
Notice that the bracketing functions $b_n$ need not belong to $\mathcal{F}$.

Useful bounds on bracketing numbers can be obtained, for example, if $\mathcal{F}$ is a parametric family, $\mathcal{F} = \{f(\cdot, \theta): \theta \in \Theta\}$, with $\Theta$ a bounded subset of some Euclidean space $\mathbb{R}^d$ and the functions subject to the condition: for some constants $C < \infty$ and $\lambda > 0$, and all $r$ small enough,

$$\sup_{n,i} \sup_{B(0, r)} |f(\xi_{ni}, \theta') - f(\xi_{ni}, \theta)|^2 = C^2 r^{2\lambda} \text{ for all } \theta$$

(2.1)

where $B(\theta, r)$ is the ball of radius $r$ around $\theta$. For example, such an inequality would follow from a Lipschitz condition,

$$|f(x, \theta) - f(x, \theta')| \leq L(x) |\theta - \theta'|^\lambda,$$

if $\sup_{n,i} \|L(\xi_{ni})\| = C < \infty$. If (2.1) holds, one takes the $f_i$ in Definition 2.1 to correspond to the centers of the $O(r^d)$ many balls of radius $r = (\delta/C)^{1/\lambda}$ that are needed to cover the bounded set $\Theta$. This gives bracketing numbers of order $O(\delta^{-k/\lambda})$, which is the sort of geometric bound needed for our theorem.

**Theorem 2.2.** Let $\{\xi_{ni}\}$ be a strong mixing triangular array whose mixing coefficients satisfy

$$\sum_{d=1}^\infty d^{Q-2} \alpha(d)^{\gamma(Q+\gamma)} < \infty$$

for some even integer $Q \geq 2$ and some $\gamma > 0$, and let $\mathcal{F}$ be a uniformly bounded class of real-valued functions whose bracketing numbers satisfy

$$\int_0^1 x^{-\gamma(Q+\gamma)} N(x) \mathcal{F}^{1/Q} dx < \infty$$

for the same $Q$ and $\gamma$. Then for each $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\lim_{n \to \infty} \sup_{\|f - f_i\| < \delta} \|v_n f_i - v_n g\|_Q < \epsilon.$$

**Corollary (Functional Central Limit Theorem)** 2.3. If the conditions of Theorem 2.2 are satisfied and if $(v_n f_1, \ldots, v_n f_k)$ has an asymptotic normal distribution for all choices of $f_1, \ldots, f_k$ from $\mathcal{F}$, then $\{v_n f: f \in \mathcal{F}\}$ converges in distribution to a Gaussian process indexed by $\mathcal{F}$ with $p$-continuous sample paths.

The formal meaning of the Corollary and the general concept of convergence in distribution are explained in Sections 9 and 10 of Pollard (1990)—see Theorem 10.2 in particular.

The conditions of Theorem 2.2 require a balance between the rate of decrease in the mixing coefficients and the rate of growth in the bracketing numbers. For example, if $N(x) = O(x^{-\beta})$ and $\alpha(d) = O(d^{-A})$ for some $\beta > 0$ and $A > 0$, then the requirements would be satisfied with $Q$ equal to the smallest even integer greater than $2\beta$ and $\gamma = 2$, if $A > (Q-1)(1 + Q/2)$. (These are not the best choices possible.) We have required $Q$ to be an even integer merely to simplify the calculations in the Appendix. It is possible that the condition could be relaxed, to allow fractional $Q$, at the cost of a more delicate argument analogous to that of Yorokawa (1980). We have not included explicit conditions for the finite-dimensional convergence as part of the Corollary, because there are many possibilities (Philipp 1969; McLeish 1977, Corollary 2.11; Herrndorf 1984, Corollary 1).
3 Proof of Theorem 2.2

The proof depends upon a moment inequality applied to the increments $v_n f - v_n g$ of the empirical process. For independent summands the inequality is well known. For strong mixing arrays it extends results of Sen (1974) and Yokoyama (1980). It corresponds to Theorem 4 of Doukhan & Portal (1984) and Theorem 10 of Doukhan & Portal (1983).

Because the last two papers offer only sketch proofs, and because typographical errors make the statement of their inequalities slightly confusing, we give a complete, self-contained proof of our inequality in the Appendix. Massart's (1987) results for strong mixing sequences rest upon the same moment inequality.

**Lemma 3.1.** Let $Z(1), Z(2), \ldots$ be a strong mixing sequence of random variables with mixing coefficients $\{\alpha(d)\}$. For some $\tau > 0$, $\gamma > 0$, and even integer $Q \geq 2$, suppose:

(i) $|Z(i)| \leq 1$, $PZ(i) = 0$, and $PZ(i)^2 \leq \tau^{2+\gamma}$ for every $i$;

(ii) $\sum_{n=1}^{\infty} d^{Q-2} \alpha(d)^{1/Q} < \infty$.

Then

$$P \left| \sum_{i=1}^{n} Z(i) \right|^Q \leq C((n\tau^2) + \ldots + (n\tau^2)^Q) \text{ for all } n,$$

for some constant $C$ that depends only on $Q$, $\gamma$, and the mixing coefficients.

Substitution of $\tau^2$ in place of the variance is the price extracted by the dependence. For independent summands one could take $\gamma$ equal to zero.

For empirical processes it will be most convenient to work with a new seminorm, $\tau(h) = \rho(h)^Q/(2+\gamma)$, because then

$$P |h(\xi_m) - h(\xi_m')|^2 \leq \tau(h)^{2+\gamma} \text{ for every } m \text{ and } n.$$

Without loss of generality we assume that $0 \leq f \leq \frac{1}{2}$ for every $f$ in $\mathcal{F}$. All the bracketing bounds $b_i$ and all the differences $|f - f_i|$ that appear in Definition 2.1 can then be assumed less than $\frac{1}{2}$, and the moment bound from Lemma 3.1 will apply directly without the intrusion of extra scaling constants. Indeed, if $|h| \leq \frac{1}{2}$ we may apply the lemma for each fixed $n$ to the random variables $\xi(i) = h(\xi_m) - Ph(\xi_m)$ to get

$$P \left| v_n h \right|^Q \leq n^{-Q/2} C((n\tau^2) + \ldots + (n\tau^2)^Q) \text{ where } \tau = \tau(h).$$

When $\tau(h)$ is bigger than $n^{-1}$ the $(n\tau^2)^Q$ term dominates. Putting $C' = (CQ/2)^{Q^2}$ we deduce

$$\left\| v_n h \right\|_Q \leq C' \max (n^{-1}, \tau(h)). \quad (3.1)$$

We must be precise here with the form of the inequality, because usually we will assume only an upper bound for $\tau(h)$; the actual value might not satisfy the inequality $n\tau(h)^2 \geq 1$.

For $k = 1, 2, \ldots$ invoke the definition of bracketing numbers with $\delta = 2^{-k}$ to find approximating subclasses $\mathcal{F}_k$ and their corresponding bounding classes $\mathcal{B}_k$ with the property that to each $f$ in $\mathcal{F}$ there is an $f_k$ in $\mathcal{F}_k$ and a corresponding $b_k$ in $\mathcal{B}_k$ for which $|f - f_k| \leq b_k$ and $\rho(b_k) \leq 2^{-4k}$. If we define $\tau_k = 2^{-4k(2+\gamma)}$ then $\tau(b_k) \leq \tau_k$. The class $\mathcal{F}_k$ need contain no more than $N_k = N(2^{-k}, \mathcal{F})$ functions.

In outline, the proof of the theorem goes as follows. We first show that $v_n f$ is uniformly well approximated by $v_n f_k(n)$ if $k(n)$ diverges to infinity at a suitable rate. More precisely, we will choose $k(n)$ to ensure that

$$\sup_{\mathcal{F}} \left\| v_n f - v_n f_k(n) \right\|_Q \to 0. \quad (3.2)$$
We then apply a chaining argument to show that, for some fixed \( m \) and \( n \) large enough, \( v_n f_{k(n)} \) is uniformly well approximated by \( v_n f_m \) for some \( f_m \in \mathcal{F}_m \), in the sense that
\[
\max_{f \in \mathcal{F}} \left\| v_n f_{k(n)} - v_n f_m \right\|_Q < 2\varepsilon \quad \text{eventually.} \tag{3.3}
\]

Here we write \( \max \) instead of \( \sup \), to emphasize that \( f_{k(n)} \) and \( f_m \) range over only finitely many functions, even when \( \mathcal{F} \) is infinite. The choice of \( f_m \) depends on \( n \), but that does not disturb the subsequent steps in the proof. Together inequalities (3.2) and (3.3) imply
\[
\left\| \sup_{f \in \mathcal{F}} |v_n f - v_n f_m| \right\|_Q < 2\varepsilon \quad \text{eventually.} \tag{3.4}
\]

Finally, using (3.4) and a subtle argument from Ledoux & Talagrand (1991, Section 11.1), we reduce the comparisons between pairs \( f, g \) from \( \mathcal{F} \) to comparisons between at most \( N_m^2 \) pairs.

Throughout the proof we rely on a simple maximal inequality due to Pisier (1983): for random variables \( Z_1, \ldots, Z_N \),
\[
\left\| \max_{i = 1}^N |Z_i| \right\|_Q \leq N^{1/4} \max_{i = 1}^N \|Z_i\|_Q,
\]
which is a consequence of the trivial bound
\[
\mathbb{P} \max_{i = 1}^N |Z_i|^\alpha \leq \sum_{i = 1}^N \mathbb{P} |Z_i|^\alpha.
\]

When specialized to the empirical process evaluated at functions \( h_1, \ldots, h_N \) with \( |h_i| \leq \frac{\varepsilon}{2} \), Pisier's inequality together with the bound (3.1) gives
\[
\left\| \max_{i = 1}^N |v_n h_i| \right\|_Q \leq C N^{1/4} \max_{i = 1}^N \left( n^{-\frac{1}{2}} \max_{i = 1}^N \tau(h_i) \right). \tag{3.5}
\]

For all applications, the \( n^{-\frac{1}{2}} \) term will be the smaller of the two terms in the max.

Proof of inequality (3.2). Let \( k(n) \) be the largest value of \( k \) for which \( \tau_k \gg n^{-\frac{1}{2}} \). Notice that \( \rho(b_{k(n)}) \leq \tau_{k(n)}^2 = o(n^{-\frac{1}{2}}) \). For each \( f \in \mathcal{F} \),
\[
|v_n f - v_n f_{k(n)}| \leq n^{-\frac{1}{2}} \sum_{i = 1}^n (b_{k(n)}(\xi_{i,n}) + \mathbb{P} b_{k(n)}(\xi_{i,n})) = v_n b_{k(n)} + 2n^{-\frac{1}{2}} \sum_{i = 1}^N \mathbb{P} b_{k(n)}(\xi_{i,n}).
\]

The last sum is less than \( 2n^{1/2}\rho(b_{k(n)}) \), which tends to \( 0 \). Thus
\[
\left\| \sup_{f \in \mathcal{F}} |v_n f - v_n f_{k(n)}| \right\|_Q \leq \left\| \max_{b \in \mathcal{B}(\mathcal{F})} |v_n b| \right\|_Q + o(1).
\]

Invoking inequality (3.5) and the integral condition of the Theorem, we deduce that
\[
\left\| \max_{b \in \mathcal{B}(\mathcal{F})} |v_n b| \right\|_Q \leq C' \frac{N^{1/4}}{k(n)} \tau_{k(n)} \leq C' \int_0^{k(n)} x^{-\gamma(2 + \gamma)} N(x, \mathcal{F})^{1/4} dx \to 0.
\]

Assertion (3.2) follows.

Proof of inequality (3.3). The integer \( m \) will soon be fixed at a value depending only on \( \varepsilon \). Eventually \( k(n) \) will be larger than \( m \). To bridge the gap between \( m \) and \( k(n) \) we argue recursively, relating the approximation via \( \mathcal{F}_k \) to the cruder approximation via
for $k = m + 1, \ldots, k(n)$. A subtle difficulty now arises. If $f$ and $f'$ are functions in $\mathscr{F}$ for which $f_k = f_k'$, there is no guarantee that $f_{k-1} = f_{k-1}'$. Potentially $f_k - f_{k-1}$ could range over as many as $N_k N_{k-1}$ differences as $f$ ranges over $\mathscr{F}$. To reduce the number of differences to $N_k$ we recycle notation by redefining $f_{k-1}$ inductively, for $k < k(n)$, to be the function from $\mathscr{F}_{k-1}$ that best approximates the function $f_k$ in $\mathscr{F}_k$, in the sense of the $\tau$ distance. Certainly

$$\tau(f_k - f_{k-1}) \leq \tau_{k-1},$$

and $\tau_{k-1} \geq n^{-1}$ if $k \leq k(n)$. Invoking inequality (3.5) again we get

$$\left\| \max_{f \in \mathscr{F}} |v_n f_k - v_n f_{k-1}| \right\|_Q \leq C' N_k^{1/Q} \tau_{k-1}.$$

As before, the max emphasizes that the differences range over only finitely many functions, $N_k$ of them, as $f$ ranges over $\mathscr{F}$. It follows that, for $n$ large enough,

$$\left\| \max_{f \in \mathscr{F}} |v_n f_k(n) - v_n f_m| \right\|_Q \leq \sum_{k=m+1}^{k(n)} \max_{f \in \mathscr{F}} |v_n f_k - v_n f_{k-1}| \right\|_Q \leq \sum_{k=m+1}^{k(n)} C' N_k^{1/Q} \tau_{k-1} \leq \sum_{k=m+1}^{k(n)} C'(2^{-k+1})^{2(2+\gamma)} N(2^{-k}, \mathscr{F})^{1/Q}.$$

For some constant $C_\gamma$, the sum is bounded by

$$C_\gamma \int_0^{b_m} x^{-\gamma(2+\gamma)} N(x, \mathscr{F})^{1/Q} \, dx.$$

With $m$ fixed so that the last bound is less than $2\epsilon$, we have (3.3). Notice that $f_m$ depends on $n$, because it is the last function in a chain leading from $f_k(n)$.

**Comparison of pairs.** Define an equivalence relation on $\mathscr{F}$ by: $f \sim f'$ if $f_m = f'_m$. The relation serves to partition $\mathscr{F}$ into $N_m$ equivalence classes $\mathscr{E}[1], \ldots, \mathscr{E}[N_m]$. (The partition actually depends on $n$, because of the way $f_m$ depends on $f_k(n)$.) From (3.4) applied twice,

$$\left\| \sup_{f \neq f'} |v_n f - v_n f'| \right\|_Q < 4\epsilon \quad \text{eventually}. \quad (3.6)$$

Define a distance between the classes by

$$d(\mathscr{E}[i], \mathscr{E}[j]) = \inf \{ \rho(f - f') : f \in \mathscr{E}[i], f' \in \mathscr{E}[j] \}.$$

For a fixed $\delta > 0$ choose functions $\phi_i$ in $\mathscr{E}[i]$ and $\phi_j$ in $\mathscr{E}[j]$ such that

$$\rho(\phi_i - \phi_j) < d(\mathscr{E}[i], \mathscr{E}[j]) + \delta.$$

If $f \in \mathscr{E}[i]$ and $f' \in \mathscr{E}[j]$ and $\rho(f - f') < \delta$, then $\rho(\phi_i - \phi_j) < 2\delta$ and

$$|v_n f - v_n f'| \leq 2 \sup_{g \neq g'} |v_n g - v_n g'| + \max \{ |v_n \phi_i - v_n \phi_j| : \rho(\phi_i - \phi_j) < 2\delta \}.$$

Notice that the last maximum runs over at most $N_m^2$ pairs. Taking norms of both sides, we deduce via (3.5) and (3.6) that

$$\left\| \sup_{|\rho(f - f')| < \delta} |v_n f - v_n f'| \right\|_Q < 8\epsilon + C' N_m^{2(2+\gamma)} (2\delta)^{2(2+\gamma)} \quad \text{eventually}.$$
We have already fixed the value of \( m \). We can therefore choose \( \delta \) small enough to make the right-hand side less than \( 9\varepsilon \).

4 An Application of Stochastic Equicontinuity

We will sketch a typical example of how stochastic equicontinuity can be used to simplify asymptotic arguments.

Suppose \( \mathcal{F} = \{ f(\cdot, \theta); \theta \in \Theta \} \) is a class of \( \mathbb{R}^k \)-valued functions indexed by a subset of \( \mathbb{R}^d \). The \( \rho \) seminorm defines a new distance on \( \Theta \) by

\[
d(\theta, \theta') = \rho(f(\cdot, \theta) - f(\cdot, \theta')).
\]

An \( m \)-estimator \( \hat{\theta}_n \) might be chosen to make the random function

\[
F_n(\theta) = \frac{1}{n} \sum_{i=n} f(\xi_i, \theta)
\]

close to zero, in the sense that

\[
F_n(\hat{\theta}_n) = o_p(n^{-1}). \tag{4.1}
\]

The true \( \theta_0 \) might be identified as the root of the corresponding expected value,

\[
M_n(\theta) = \mathbb{P}F_n(\theta) = \frac{1}{n} \sum_{i=n} \mathbb{P}f(\xi_i, \theta),
\]

in the sense that

\[
M_n(\theta_0) = 0 \quad \text{for all } n.
\]

With a preliminary argument (often based on a uniform law of large numbers) one might be able to establish consistency, \( \hat{\theta}_n \rightarrow \theta_0 \) in probability. With mild continuity and domination conditions on the \( f(\cdot, \theta) \) functions, this can usually be reinterpreted as

\[
d(\hat{\theta}_n, \theta_0) \rightarrow 0 \quad \text{in probability.} \tag{4.2}
\]

With such preliminaries, asymptotic normality of \( n^{1/2}(\hat{\theta}_n - \theta_0) \) can then be deduced from the following three requirements on the processes.

(i) Uniform differentiability of the \( M_n \) functions at \( \theta_0 \); for some nonsingular matrix \( D \),

\[
M_n(\theta) = D(\theta - \theta_0) + o(\|\theta - \theta_0\|),
\]

with the \( o(\cdot) \) term uniform in \( n \). Notice that this requirement is weaker than pointwise differentiability of the \( f(x, \cdot) \) functions, which can be useful in such applications as least absolute deviations estimation.

(ii) Asymptotic normality of \( v_n f(\cdot, \theta_0) \).

(iii) Stochastic equicontinuity of \( v_n \) at \( \theta_0 \); for each \( \epsilon > 0 \) and \( \eta > 0 \) there exists a \( \delta > 0 \) such that

\[
\limsup_{n \to \infty} \mathbb{P}\left( \sup_{d(\theta, \theta_0) < \delta} |v_n f(\cdot, \theta) - v_n f(\cdot, \theta_0)| > \eta \right) < \epsilon.
\]

When reduced to assertions about each of the components of the vector processes, requirement (iii) is weaker than the stochastic equicontinuity property delivered by Theorem 2.2. Together with (4.2) it implies that

\[
v_n f(\cdot, \hat{\theta}_n) = v_n f(\cdot, \theta_0) + o_p(1). \tag{4.3}
\]
In addition, from (i) we get
\[ M_n(\hat{\theta}_n) = D(\hat{\theta}_n - \theta_0) + o(||\hat{\theta}_n - \theta_0||). \]  \hspace{1cm} (4.4)

Substitution into (4.1) then gives
\[ o_p(n^{-1}) = F_n(\hat{\theta}_n) = M_n(\hat{\theta}_n) + n^{-1}v_n \xi f(\cdot, \hat{\theta}_n) = D(\hat{\theta}_n - \theta_0) + o(||\hat{\theta}_n - \theta_0||) + n^{-1}v_n f(\cdot, \theta_0) + o_p(n^{-1}). \]  \hspace{1cm} (4.5)

First deduce from the nonsingularity of \( D \) that \( ||\hat{\theta}_n - \theta_0|| = O_p(n^{-1}) \):
\[ |n^{-1}v_n f(\cdot, \theta_0) + o_p(n^{-1})| = |D(\hat{\theta}_n - \theta_0) + o(||\hat{\theta}_n - \theta_0||)| \text{ from (4.5)} \]
\[ \Rightarrow (\kappa - o_p(1)) ||\hat{\theta}_n - \theta_0|| \]

for some positive constant \( \kappa \). The left-hand side is of order \( O_p(n^{-1}) \) by (ii). Next consolidate the error terms in (4.5) to get
\[ n^{-1}||\hat{\theta}_n - \theta_0|| = D^{-1}v_n f(\cdot, \theta_0) + o_p(1). \]

The random vector on the right-hand side has an asymptotic normal distribution.

5 Overview and Extensions

The main lesson that we learned from our efforts to develop empirical central limit theorems for dependent variables was: everything depends on the existence of good probabilistic bounds for the increments of the empirical process. Let us explain.

There have been two major lines of development in the literature on abstract empirical central limit theorems. One line, which starts from a symmetrization argument, has evolved from the method of Vapnik & Červonenkis (1971). It depends on the following two requirements.

(i) In the sense of either tail probability bounds or moment bounds, the quantity \( \sup_{\mathcal{X}} |v_n f| \) is less variable than the ‘symmetrized’ process
\[ \sup_{\mathcal{X}} n^{-1} \left| \sum_{i \in n} (f(\xi_{n,i}) - f(\xi_{n,i}')) \right|. \]

The new variables \( \{\xi_{n,i}'\} \) are typically an independent copy of the \( \{\xi_{n,i}\} \).

(ii) Conditional on certain information symmetric in both \( \{\xi_{n,i}\} \) and \( \{\xi_{n,i}'\} \), the symmetrized process has a tractable distribution.

Requirement (ii) is analogous to the property that justifies the calculations with permutation distributions for experimental designs with randomization, as in Box & Andersen (1955). In practice it has required independence of \( \{\xi_{n,i}\} \) from \( \{\xi_{n,i}'\} \). For then the symmetrized process has the same distribution as
\[ \sup_{\mathcal{X}} n^{-1} \left| \sum_{i \in n} \sigma_i (f(\xi_{n,i}) - f(\xi_{n,i}')) \right|. \]  \hspace{1cm} (5.1)

where the \( \{\sigma_i\} \) are independent Rademacher variables (+1 and -1, each with probability \( \frac{1}{2} \)). The conditional distribution, given \( \{\xi_{n,i}\} \) and \( \{\xi_{n,i}'\} \), of this expression is amenable to various types of chaining argument (cf. Section 3), because good bounds exist for the increments of the underlying process: the Hoeffding inequalities give exponential bounds.
on tail probabilities and moments up to exponential order. It is largely a matter of taste whether one applies the chaining argument to moment quantities (as in Pollard 1990) or to tail probabilities (as in Pollard 1984). The moment bounds require slightly less machinery (one fewer sequence of constants to adjust correctly), at the slight cost of results not quite as refined as those for tail probabilities (Alexander 1984, Massart 1986). For other related applications, however, such as the $U$-processes of Nolan & Pollard (1987, 1988), only moment bounds seem to work.

Dependence between the $\{\xi_n\}$ variables usually greatly complicates the requirement (ii). Leventhal (1988) was able to extend the method to regenerative processes, in which excursions between renewal times are independent, by replacing the $\{\xi_n\}$ by the whole excursion; he symmetrized over whole blocks of variables. Yu (1994) coupled $\phi$-mixing and $\beta$-mixing processes with processes of independent blocks, leading back to the theory for the independence case. Further progress with symmetrization applied to dependent variables seems unlikely, except in special cases that allow similar reduction to independence.

The second major line of development of abstract empirical central limit theorems has involved the use of bracketing arguments. These appear promising for dependent variables because they work directly with the empirical process. Again one needs some sort of probabilistic bounds for the increments of the process in order to invoke a chaining argument. For the classical one-dimensional empirical central limit theorem in the independent case, a Tchebychev bound based on fourth moments of the binomial distribution suffices—for example, see pages 262–266 of Parthasarathy (1967).

For his abstract empirical central limit theorem (for independent summands) under a bracketing condition, Dudley (1978, 1981) applied the Bernstein exponential tail bound to the increments of the empirical process indexed by classes of sets and uniformly bounded classes of functions. The Bernstein inequality requires bounded summands. The bound loses its power for increments with small variance—compare with the $n^{-1}$ that appears on the right-hand side of (3.5). A separate argument is needed to handle the contributions from such increments, as in our proof of (3.2).

Ossiander (1987) combined the Bernstein bound with a delicate truncation argument to remove the boundedness assumption from the class of functions. The chaining can continue forever; the bound on the truncated summands decreases with the variance, keeping the empirical process within the range where the Bernstein inequality has power. Andrews & Pollard (1991) have shown that a first moment version of Ossiander's method slightly simplifies the argument.

Analogues of the Bernstein bound exist for dependent summands, such as martingale difference arrays. Leventhal (1989) invoked such a bound, but was left with an unpalatable uniformity assumption involving the small increments. For $\phi$-mixing sequences Collomb's (1984) Lemma 4.1 can support a chaining argument. Yukich (1986) applied it with a bracketing argument to derive rates of convergence for uniform laws of large numbers. Massart (1987, Theorem 5) applied it to prove a functional central limit theorem for uniformly bounded classes of functions under bracketing conditions only slightly stronger than those needed for independent variables. Andrews & Pollard (1991) modified Collomb's method to establish an inequality that supports an analog of Ossiander's truncation argument; they extended Massart's theorem to classes of unbounded functions.

For strong mixing sequences with mixing coefficients decreasing at a geometric rate, Massart (1987, Theorem 5) was able to handle exponential rates of growth in bracketing numbers by strengthening the maximal inequality corresponding to our Lemma 3.1. He cited an unpublished paper of Doukhan & Portal (which we have not seen) to determine
the dependence of the constant $C$ on the moment $Q$; for constants $\theta < 1$ and $K$ that do not depend on $Q$ or our $\gamma$, it appears that one can take

$$
P \left| \sum_{i=1}^{n} Z(i) \right|^{Q} \leq (KQ^2/(1 - \theta^\gamma) n\sigma^{1-\delta})^{Q} \text{ for } n\sigma^2 \geq 1,$n

where $\sigma^2$ is the upper bound on the variances of the summands (that is, our $\tau^2 + \gamma$), and $\delta = \gamma/(2 + \gamma)$. We are surprised that the constraint is not $n\tau^2 \geq 1$. As a strengthening of our inequality (3.5) one would get

$$
P \max_{i \in N} |\nu_n h_i| \leq \max_{i \in N} |\nu_n h_i|_Q \leq C^Q Q^2 N^{1/2} \max \left( n^{-1}, \max_{i \in N} \tau(h_i) \right),$$

where $C^Q$ does not depend on $Q$. Choosing $Q$ to increase like $\log N$, one would end up with a factor $(\log N)^2$ instead of the $N^{1/2}$. The integral condition in Theorem 2.2 could then be weakened to

$$\int_{0}^{1} x^{-\gamma(2 + \gamma)} (\log N(x))^2 \, dx < \infty,$$

with $L^Q$ norm replaced by an $L^1$ norm in the stochastic equicontinuity assertion. For example, if $\log N(x, F) = O(x^{-\beta})$ for some $\beta < 1$ one could choose $\gamma$ small enough to make the integral converge. We hesitate to proclaim such a result as a true theorem before we have verified the form of the constant asserted by Massart. In principle one might be able to control the constants in the Appendix more carefully. We defer that exercise to a later paper.

We have encountered some extra difficulties in trying to extend this technique for strong mixing arrays to classes of unbounded functions. Modified forms of Carbon’s (1983) inequality (1) seem to offer some promise. As it stands, his inequality is not completely suitable, because it does not take advantage of small variances of the summands—the factor $c(k)$ in the coefficient of $\sigma^2$ does not decrease with $D$. By modifying the argument leading to the second moment bound at the top of Collomb’s page 451, we were able to replace Carbon’s $c(k)$ by a factor involving the square root of an $L^1$ norm. Unfortunately, we were not able to find an appropriate substitute for the last term in his exponential bound; we still have not transformed it into a bound that could support a chaining argument with truncation. See Andrews & Pollard (1991) for more about the type of inequality needed for truncation.

For related approaches, which might extend the range of empirical central limit theorems for dependent variables, we refer the reader to Andrews (1991), and to the work of Goldie & Greenwood (1986a, b) on set-indexed partial-sum processes. There is also a substantial literature on dependent sequences in general Banach spaces, with particular emphasis on invariance principles that translate the limit theorems into approximation results—papers of Philipp (1982) and Dehling (1983), for example. The paper of Dudley & Philipp (1983) treats the corresponding theorems for the independence case. Our stochastic equicontinuity property corresponds to approximation of the Banach-valued processes by processes concentrated in finite dimensional subspaces. Probabilistic bounds for the error of approximation again depend on suitable moment or exponential inequalities. Massart (1987) stated several of his results as rates of convergence for such invariance principles. We do not have sufficient familiarity with that branch of the Banach space literature to make pronouncements of any authority, but we surmise that the effort of translating from Banach-valued to empirical process terminology would yield a rich supply of limit theorems for dependent processes.
Appendix: Proof of Lemma 3.1

We will make repeated use of the following standard strong mixing inequality for random variables (Hall & Heyde 1980, Corollary A.2). For fixed \( n, m, \) and \( d \), suppose \( X \) is \( \mathcal{A}_n(m) \)-measurable, \( Y \) is \( \mathcal{B}_n(m + d) \)-measurable. Let \( s, p, \) and \( q \) be positive numbers whose reciprocals sum to 1. Then

\[
|\mathbb{P}XY - \mathbb{P}X \mathbb{P}Y| \leq 8\alpha(d)^{1/2} \|X\|_p \|Y\|_q. \tag{A.1}
\]

If \( X \) happens to be a product \( X_1 \ldots X_m \), Hölder’s inequality bounds the factor \( \|X\|_p \) by

\[
\left( \prod_i \mathbb{P} |X_i|^{mp} \right)^{1/mp}.
\]

If \( X_i \leq 1 \) and \( \mathbb{P}X_i^2 \leq \tau^{2+\gamma} \leq 1 \) for every \( i \), and if \( mp \geq 2 \), the product is less than \( \tau^{(2+\gamma)/p} \).

If \( Y \) has a similar decomposition into a product of \((k - m)\) factors, \( \|Y\|_q \) is similarly bounded by \( \tau^{(2+\gamma)/q} \). Choosing \( s = (Q + \gamma)/\gamma \) and \( mp = (k - m)q = k/(1 - 1/s) \), then decreasing the resulting exponent of \( \tau \) from \((2 + \gamma)Q/(Q + \gamma)\) to 2, we arrive at our working inequality,

\[
|\mathbb{P}X_1 \ldots X_m Y \ldots Y_{k-m}| \leq |\mathbb{P}X_1 \ldots X_m \mathbb{P}Y_1 \ldots Y_{k-m}| + 8\alpha(d)^{\gamma Q(2+\gamma)} \tau^2. \tag{A.2}
\]

Here the choice of \( p \) and \( q \) is not critical; we need only ensure that \( mp \geq 2 \) and \((k - m)q \geq 2 \).

For positive integers \( k \) and \( n \), with \( \tau \) fixed, define a bounding function

\[
B_n(k) = nt^2 + (nt^2)^2 + \ldots + (nt^2)^{k/2},
\]

where \( \lfloor k/2 \rfloor \) stands for the integer part of \( k/2 \). We will establish the existence of constants \( C_k \), for \( k = 1, \ldots, Q \), such that

\[
\sum_1^k |\mathbb{P}Z(i_1) \ldots Z(i_k)| \leq C_k B_n(k) \quad \text{for all } n, \tag{A.3}
\]

where the sum runs over all choices of \( i = (i_1, \ldots, i_k) \) such that \( 1 \leq i_1 \leq \ldots \leq i_k \leq n \). With \( k = Q \) the left-hand side of (A.3) is greater than \( 1/Q! \) times the \( Q \)th moment quantity that we are seeking to bound.

Inequality (A.3) holds for \( k = 1 \), since the \( Z(i) \) have zero expected values. We will argue inductively to establish it for a \( k > 1 \), assuming that it holds for all values less than \( k \).

For a given \( i = (i_1, \ldots, i_k) \), let \( G(i) \) denote the largest of the differences \( i_j - i_{j-1} \), and let \( m(i) \) be the smallest \( j \) for which the difference equals \( G(i) \). To simplify the notation, write \( \beta(d) \) for \( a(d)^{\gamma Q(2+\gamma)} \). Apply the inequality (A.2) to each term on the left-hand side of (A.3) to bound the sum by

\[
\sum_{m=1}^{k-1} \sum_i \{m(i) = m|(|\mathbb{P}Z(i_1) \ldots Z(i_m)\mathbb{P}Z(i_{m+1}) \ldots Z(i_k)| + 8\tau^2 \beta(G(i))). \tag{A.4}
\]

Consider first the contribution from the product of expectations. If \( m = 1 \) or \( m = k - 1 \), one of the expectations is zero. For other values of \( m \) we invoke the inductive hypothesis. Fixing \( m \) and \( i_1, \ldots, i_m \) and letting \( i_{m+1}, \ldots, i_k \) range from 1 to \( n \) instead of just from \( i_m \) to \( n \), we bound the contribution by

\[
\sum_{m=2}^{k-2} \sum_{i_1, \ldots, i_m} |\mathbb{P}Z(i_1) \ldots Z(i_m)|C_{k-m}B_n(k - m),
\]

which, by a second appeal to the inductive hypothesis, is less than

\[
\sum_{m=2}^{k-2} C_mC_{k-m}B_n(m)B_n(k - m).
\]
The product $B_n(m)B_n(k - m)$ is a polynomial in $n\tau^2$ of degree
\[ \left\lfloor \frac{m}{2} \right\rfloor + \left\lfloor \frac{k - m}{2} \right\rfloor \leq \left\lfloor \frac{k}{2} \right\rfloor. \]

Thus the product of expectations contributes to (A.4) at most a constant multiple of $B_n(k)$.

For the contribution to (A.4) from the mixing coefficients we further decompose the sum over $i$ according to the location and size of the largest gap $G(i)$. The contribution equals
\[ \sum_{m=1}^{k-1} \sum_{i=1}^{n} \sum_{g=1}^{n} [m(i) = m, i_m = l, G(i) = g]8\tau^2\beta(g). \]

Given $m(i) = m$ and $i_m = l$ and $G(i) = g$, the indices $i_1, \ldots, i_{m-1}$ are subject to the constraints
\[ 1 \leq i_1 \leq \ldots \leq i_{m-1} \leq i_m = l, \]
\[ i_{j+1} - i_j \leq g \quad \text{for } j = 1, \ldots, m - 1. \]

Summing first over $i_1$, for fixed $i_2, \ldots, i_m$, and then over $i_2$, and so on, we constrain each index to a range of $1 + g$ or fewer integers. There are at most $(1 + g)^{m-1}$ choices for the first $m - 1$ indices. Similarly, because the two equalities $i_m = l$ and $G(i) = g$ fix $i_{m+1}$, at the value $l + g$, there are at most $(1 + g)^{k-1-1}$ choices for $i_{m+1}, \ldots, i_k$. Thus the mixing coefficients contribute to (A.4) at most
\[ 8\tau^2 \sum_{m=1}^{k-1} \sum_{i=1}^{n} \sum_{g=1}^{n} (1 + g)^{m-1}(1 + g)^{k-1-m-1}\beta(g) \leq 8\tau^2n \sum_{g=1}^{\infty} (1 + g)^{k-2}\beta(g). \]

Assumption (ii) ensures finiteness of the sum over $g$; the whole contribution to (A.4) is less than a constant multiple of $n\tau^2$, which can be absorbed into $B_n(k)$.

\[ \Box \]

Acknowledgement

We thank both the referees for pointing out the relevance of work by Massart, Philipp, and Yu.

Remark added in proof

Subsequent to the acceptance of this paper we received from Paul Doukhan a copy of a book manuscript “Mixing: Properties and Examples”. The manuscript renders superfluous some of our hedging comments in Section 5 regarding the results of Massart (1987) based on the unpublished work of Doukhan & Portal. We are loath to further delay publication of our paper by undertaking a fourth major revision. Not only would we want to modify our remarks concerning our ignorance of some of the French work, but we would also feel obliged to comment on recent developments. The subject has progressed rapidly in the last two years. We instead recommend readers to seek out the papers “Invariance principles for absolutely regular empirical process” by Doukhan, Massart, and Rio which will appear in Annales de l’Institut Henri Poincaré, and “Central limit theorems of empirical and U-Proceses of stationary mixing sequences”, by Arcones and Yu, which will appear in Journal of Theoretical Probability.
References


**Résumé**

Cet article montre comment l'outillage moderne pour générer les théorèmes empiriques abstraits de limite centrale peuvent s'appliquer aux vecteurs de variables dépendantes. Nous développons une approximation de crochets (similaire aux résultats de Philipp et Massart) fondée sur une inégalité de moments de sommes pour des vecteurs satisfaisant des conditions de mélange fort, afin d'illustrer le genre de difficultés qui doivent être surmontées lors de l'adaptation de la théorie des processus empiriques de vecteurs indépendants. Quelques suggestions de développements futurs sont proposées. L'article est en grande partie autonome.

*[Received May 1990, accepted January 1992]*