ON UNIFORM CONVERGENCE
FOR DEPENDENT PROCESSES

by
ANDREW NOBEL
and
AMIR DEMBO
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 74
JUNE 1990

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT NCR-89-14538

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON UNIFORM CONVERGENCE
FOR DEPENDENT PROCESSES

by
ANDREW NOBEL

and
AMIR DEMBO
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 74
JUNE 1990

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT NCR-89-14538

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
On Uniform Convergence for Dependent Processes

Andrew Nobel* and Amir Dembo †

June 23, 1990

Abstract

Uniform convergence with respect to an i.i.d process yields uniform convergence with respect to all $\beta$-mixing processes with the same marginal distribution. Several interesting closure properties of uniformly convergent classes of functions are derived. The performance of empirical risk minimization when the class of loss functions is not uniformly convergent is asymptotically bounded by a simple distribution-free empirical overfit estimator.

---

*Andrew Nobel is with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

†Amir Dembo is with the Statistics and Mathematics Departments, Stanford University, Stanford, CA 94305.

Key Words: uniform convergence, empirical risk minimization, V-C class, absolutely regular processes, symmetrization, subadditive ergodic theorem.

AMS (1980) subject classification: 60F20, 60F15, 60B12

This work was partially supported by the National Science Foundation under Grant NCR-89-14538.
1 Introduction

The Vapnik-Cervonenkis theory of uniform convergence extends the classical strong law of large numbers (SLLN) from a single real-valued observation to an ensemble of such observations. If \( X_0, X_1, \ldots \) is an independent, identically distributed (i.i.d.) sequence of random variables taking values in some set \( \mathcal{X} \), and \( f \) is some real-valued function on \( \mathcal{X} \), then the SLLN asserts that the sample averages

\[
\frac{1}{n} \sum_{i=0}^{n-1} f(X_i)
\]

converge almost surely to the expected value of \( f \), \( Ef \). In many applications of interest, it is necessary to consider, simultaneously, the sample averages of an ensemble \( \mathcal{F} \) of such functions. Although the SLLN applies individually to each function \( f \in \mathcal{F} \) we often need to know whether or not the sample averages of functions in \( \mathcal{F} \) converge \textit{uniformly} to their expectations; that is, whether or not

\[
\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \to 0,
\]

almost surely as \( n \) tends to infinity.

If the convergence in (1) obtains, the class \( \mathcal{F} \) is said to be \textit{uniformly convergent} with respect to the sequence \( X_0, X_1, \ldots \). Note that the term 'uniform' here refers to the class \( \mathcal{F} \) rather than the underlying sample space. Perhaps the best known result concerning uniform convergence is the classical Glivenko-Cantelli theorem which asserts that the class \( \mathcal{F} = \{I_{(-\infty, t]} : t \in \mathbb{R}\} \) of indicator functions of half-infinite intervals is uniformly convergent with respect to any real-valued i.i.d. process \( \{X_t\} \).

In statistical terminology, this is just the assertion that the empirical c.d.f. \( \hat{F}_n(t) \) converges uniformly in \( t \) to the true c.d.f. \( F(t) \).

The following pattern recognition problem illustrates a typical application in which uniform convergence plays an important role. Suppose that we wish to assign each vector, or pattern, \( t \) in \( \mathbb{R}^k \) to one of \( N \) classes \( \{\theta_1, \ldots, \theta_N\} \) by means of a classification rule \( \phi : \mathbb{R}^k \to \{\theta_1, \ldots, \theta_N\} \). In particular, \( t \) might be a vector of symptoms such as blood pressure, temperature, heart rate and so on, and each class \( \theta_i \) might denote the presence of a particular disease. Each observation \( X_i \) is a pair of random variables
\((T_i, \Theta_i)\), where \(T_i\) takes values in \(\mathbb{R}^k\), and \(\Theta_i\) takes values in \(\{\theta_1, \ldots, \theta_N\}\). We assume that each pair \((T_i, \Theta_i)\) is drawn according to some fixed, but unknown, distribution \(P\).

On the basis of observations \(X_0, \ldots, X_{n-1}\), we wish to select an appropriate classification rule \(\phi\) from a fixed family \(\Phi\) of such rules. For instance, \(\Phi\) might be the set of all classification rules \(\phi\) satisfying some geometrical or computational constraint. Note that the distribution \(P\) need not preserve a direct functional relationship between patterns and classes. Nevertheless, we may think of the observations as representing noisy measurements of some true, underlying rule \(\phi^*\).

Let \(D : \{\theta_1, \ldots, \theta_N\} \times \{\theta_1, \ldots, \theta_N\} \to \mathbb{R}\) be a nonnegative error function. We can measure the loss, \(L_\phi(X_i)\), of a rule \(\phi \in \Phi\) on an observation \(X_i\) in terms of the error function \(D\):

\[
L_\phi(X_i) = D(\phi(T_i), \Theta_i)
\]  

(2)

Given the observations \(X_0, \ldots, X_{n-1}\), we wish to find a rule \(\phi \in \Phi\) with small risk (expected loss) \(E L_\phi\). On the average, such a rule will agree well with future observations. More specifically, we would like to find a natural procedure which will produce from the data \(X_0 = x_0, X_1 = x_1, \ldots\) a sequence of rules \(\phi_n = \phi_n(x_0, \ldots, x_{n-1})\) from \(\Phi\) such that

\[
\lim_{n \to \infty} E L_{\phi_n} = \inf_{\phi \in \Phi} E L_\phi.
\]  

(3)

If the class \(\mathcal{F}_\Phi = \{L_\phi : \phi \in \Phi\}\) of loss functions is uniformly convergent with respect to the observations \(\{X_i\}\), then it can be shown that (3) holds if, at time \(n\), we select a rule \(\phi_n\) with minimal empirical risk \(\frac{1}{n} \sum_{i=0}^{n-1} L_\phi(X_i)\).

Thus, the uniform convergence of \(\mathcal{F}_\Phi\) insures the consistency of a natural inductive procedure that Vapnik calls empirical risk minimization. Empirical risk minimization is based on the principle that one can make accurate inferences about the expectations of functions in \(\mathcal{F}_\Phi\) by means of a simultaneous evaluation of their finite-sample averages. This and other applications of uniform convergence to empirical risk minimization are discussed in Vapnik (1989). An application of uniform convergence to a general model of learning can be found in Haussler (1989).

3
When \( \Phi \) is a class of rules involving say pairs of patterns \( T_i \) and \( T_{i-1} \), then \( X_i = (T_i, T_{i-1}, \Theta_i) \) and the observations are no longer independent. Such a scenario arise for example in nonparametric time series prediction where the classes \( \theta_1, \ldots, \theta_N \) represent possible estimates of future patterns. Typically, while the observations are now dependent, this dependence has a "finite range" so the observation process still satisfies certain mixing conditions.

In the pioneering work of Vapnik and Cervonenkis (1981), they found necessary and sufficient conditions under which a class of functions \( \mathcal{F} \) is uniformly convergent with respect to an i.i.d. process \( \{X_i\} \). Their result is briefly described in the next section where its extension to sequences of dependent random variables is motivated by considering direct methods for establishing uniform convergence for any stationary, ergodic process, but for much restricted classes of functions.

In Section 3 we observe that for any stationary and ergodic process the sequence \( \sup_{\mathcal{F}} |\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0)| \) converges almost surely to a constant as \( n \) tends to infinity (see Lemma 1). Our main result (Theorem 1) is that uniform convergence with respect to a sequence of i.i.d. random variables \( X_0, X_1, \ldots \) with common distribution \( P \) is sufficient for uniform convergence with respect to any absolutely regular process \( \{X_i\} \) with marginal distribution \( P \). For the exact definition of absolutely regular processes see Section 3 (absolute regularity is a strong mixing condition also known as \( \beta \)-mixing). An immediate consequence of Theorem 1 is that the empirical c.d.f converges uniformly in \( t \) to the true c.d.f for any absolutely regular sequence of samples (this is the implied extension of the classical Gilvenko-Cantelli theorem).

Unlike the recent works of Bin Yu (1990) and Massart (1988) we do not aim at a uniform CLT for empirical processes and thus obtain the uniform convergence with no conditions imposed on the mixing rates. We also use a different indirect technique, based on Kingman's subadditive ergodic theorem and a regular spacing argument.

Section 4 is devoted to closure properties of uniformly convergent classes of functions. The main result is Theorem 2 which allows for continuous operations on pairs of uniformly convergent classes. Possible applications of this theorem include products, minimum and maximum, logarithms, and \( L_q \) norms. Pollard (1989) contains a
similar exposition of closure (stability) properties in the context of uniform CLT-s.

In the last section we consider the method of empirical risk minimization which motivated this work. In particular we show in Theorem 3 that this method is applicable even when the convergence in (1) is to a nonzero constant \( \eta(\mathcal{F}) \) provided the latter is less than the tolerated error. Lemma 3 and Theorem 4 show that when \( \eta(\mathcal{F}) \) is nonzero and the underlying distribution is unknown, one can still estimate \( \eta(\mathcal{F}) \) within a factor of 2 using differences of sample averages. This result holds for any absolutely regular process under a mild regularity condition on \( \mathcal{F} \).

2 The Theorem of Vapnik and Cervonenkis and the Method of Direct Approximation

Let \( \mathcal{F} \subseteq \mathbb{R}^\mathcal{X} \) be any class of real-valued functions on \( \mathcal{X} \). A function \( F : \mathcal{X} \rightarrow \mathbb{R} \) is said to be an envelope for \( \mathcal{F} \) if for every \( f \in \mathcal{F} \) \( |f| \leq F \). The class \( \mathcal{F} \) is said to be uniformly bounded if for some \( 0 \leq K < \infty \), \( F = K \) is an envelope for \( \mathcal{F} \). Now let \( x_0, \ldots, x_{n-1} \) be a sequence of points in \( \mathcal{X} \), and let \( \epsilon > 0 \). A finite subset \( \mathcal{H} \subseteq \mathbb{R}^\mathcal{X} \) is said to be an \( \epsilon \)-cover of \( \mathcal{F} \) on \( x_0^{n-1} \) if for each \( f \in \mathcal{F} \) there is some \( h \in \mathcal{H} \) for which \( \frac{1}{n} \sum_{i=0}^{n-1} |f(x_i) - h(x_i)| < \epsilon \). The covering number \( N(x_0^{n-1}, \epsilon, \mathcal{F}) \) denotes the cardinality of a smallest \( \epsilon \)-cover of \( \mathcal{F} \) on \( x_0^{n-1} \); if no such \( \epsilon \)-cover exists, then \( N(x_0^{n-1}, \epsilon, \mathcal{F}) \) is infinite.

Remark 1: To every minimal \( \epsilon \)-cover \( \mathcal{H} \) of \( \mathcal{F} \) there corresponds a \( 2\epsilon \)-cover \( \mathcal{H}' \) of \( \mathcal{F} \) with \( \mathcal{H}' \subseteq \mathcal{F} \): simply replace each \( h \in \mathcal{H} \) by a function \( h' \in \mathcal{F} \) such that \( \frac{1}{n} \sum_{i=0}^{n-1} |h'(x_i) - h(x_i)| < \epsilon \). Thus, by doubling \( \epsilon \), we may assume that any minimal \( \epsilon \)-covering of \( \mathcal{F} \) is contained in \( \mathcal{F} \).

If the class \( \mathcal{F} \) is uncountable, it must satisfy regularity conditions in order to insure that no problems of measurability will arise. Following Pollard (1984), we will use the term 'permissible' to indicate that a class \( \mathcal{F} \) satisfies such conditions. For more details, we refer the interested reader to Dudley (1978) or Pollard (1984). The following is a slightly stronger version of the result presented in Vapnik and Cervonenkis (1981). The improvement, due to Steele (1978), will be discussed in the
next section.

Theorem A (Vapnik and Cervonenkis) Let \( \{ X_i \} \) be an i.i.d. sequence of random variables and let \( F \) be a uniformly bounded, permissible class of functions. Then

\[
\sup_{F} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \to 0 \text{ almost surely as } n \to \infty \text{ if and only if for every } \epsilon > 0
\]

\[
\frac{1}{n} \log N(X_0^{n-1}, \epsilon, F) \to 0 \quad (4)
\]

in probability as \( n \to \infty \). □

The condition (4) may be characterized as being local in so far as it concerns the behavior of the class \( F \) along particular realizations \( X_0(\omega), X_1(\omega), \ldots \) of the process \( \{ X_i \} \). By contrast, one might consider instead the global behavior of \( F \) in terms of the marginal distribution of the process. This leads to what Pollard (1984) calls the 'method of direct approximation', which allows us to establish uniform convergence results for stationary, ergodic processes. To illustrate, we state without proof an elementary result of Pollard (1984) and give some of its consequences. Let \( \{ X_i \}_{i=0}^{\infty} \) be a stationary, ergodic process with marginal distribution \( P \), and assume that the random variables \( X_i \) take values in a set \( \mathcal{X} \).

Theorem B Let \( F \subseteq \mathbb{R}^\mathcal{X} \) be permissible. Suppose that for each \( \epsilon > 0 \) there exists a finite class \( F_\epsilon \) of upper and lower approximations to \( F \); that is, for each \( f \) in \( F \) there exist \( f_1, f_2 \) in \( F_\epsilon \) for which

\[
f_1 \leq f \leq f_2 \quad \text{and} \quad \int (f_2 - f_1) dP \leq \epsilon.
\]

Then \( \sup_{F} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \to 0 \) almost surely as \( n \to \infty \). □

Corollary 1 If a permissible class \( F \) is totally bounded as a subset of \( L_\infty(\mathcal{X}) \) (i.e. the set \( \mathbb{R}^\mathcal{X} \) equipped with the supremum norm), then \( \sup_{F} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \to 0 \) almost surely as \( n \to \infty \).
Proof: Let \( \varepsilon > 0 \) be given, and let \( \mathcal{H} \) be a finite \( \varepsilon \)-net for \( \mathcal{F} \). Then the class \( \mathcal{F}_{2\varepsilon} = \{ h + \varepsilon : h \in \mathcal{H} \} \cup \{ h - \varepsilon : h \in \mathcal{H} \} \) contains lower and upper approximations to every \( f \in \mathcal{F} \).

Corollary 2 If the random variables \( X_i \) take values in a finite set \( \mathcal{X} \) and \( \mathcal{F} \) is uniformly bounded, then \( \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \to 0 \) almost surely as \( n \to \infty \).

Proof: Any uniformly bounded class \( \mathcal{F} \) is a totally bounded subset of \( L_\infty(\mathcal{X}) \), as \( \mathcal{X} \) is finite.

These results indicate that a thorough study of uniform convergence need not be limited to independent processes. However, the method of direct approximation lacks both the scope and the applicability of the method of Vapnik and Cervonenkis. In the next section we show that the sufficiency part of Theorem A holds, with only minor changes, for absolutely regular processes.

3 Uniform Convergence for Absolutely Regular Processes

Throughout the remainder of this paper, we assume that \( \{X_i\}_{\infty=-\infty} \) is a two-sided, strictly stationary sequence of random variables taking values in the measurable space \((\mathcal{X}, \mathcal{S})\). Without loss of generality, we assume the sequence \( \{X_i\} \) is defined on the product space \((\mathcal{X}_{\infty}^{\infty}, \mathcal{S}_{\infty}^{\infty}, \mathbb{P})\) in the usual way via the left shift operator \( T \) and the coordinate projection \( X_0 \); that is, \( X_i(\omega) = X_0(T^i\omega) \). If \( \mathcal{G} \) is a sub \( \sigma \)-field of \( \mathcal{S}_{\infty}^{\infty} \) and \( h \) is a real-valued function on \( \mathcal{X}_{\infty}^{\infty} \), then we write the conditional expectation of \( h \) with respect to \( \mathcal{G} \) as \( E\mathcal{G}h \). The \( \sigma \)-field \( \sigma(X_{\infty}^{\infty}) \) will be denoted by \( \mathcal{G}_{\infty} \).

Given a process \( \{X_i\}_{\infty=-\infty} \) as above, we may define dependence coefficients \( \beta(k) \) as follows (c.f. Bradley (1986) and Volkonskii and Rozanov (1959, 1961)): \[
\beta(k) = \frac{1}{2} \sup \sum_{i,j} |P(A_i \cap B_j) - P(A_i)P(B_j)|,\]
where the supremum is over all finite partitions \( \{A_i\} \) measurable \( \sigma(X_{\infty}^{\infty}) \) and all finite partitions \( \{B_j\} \) measurable \( \sigma(X_{k}^{\infty}) \). As the coefficients \( \beta(k) \) are bounded by 1,
nonnegative and non-increasing, \( \lim_{k \to \infty} \beta(k) \) always exists. If \( \lim_{k \to \infty} \beta(k) = 0 \) then the process \( \{X_i\}_{i=-\infty}^{\infty} \) is said to be absolutely regular. In particular, if the process \( \{X_i\}_{i=-\infty}^{\infty} \) is absolutely regular, then it is strongly mixing in the ergodic theory sense (c.f. Bradley (1986)). Additional properties of the \( \beta \) coefficients are discussed in Bradley (1983).

The following alternative characterizations of the coefficients \( \beta(k) \) (c.f. Bradley (1986) and Volkonskii and Rozanov (1959, 1961)) will be essential in what follows.

**Lemma A** Let \( \mathbb{P}^0_{-\infty} \) and \( \mathbb{P}^\infty_1 \) be the half-infinite marginals of the distribution \( \mathbb{P} \), and \( \mathbb{P}^0_{-\infty} \times \mathbb{P}^\infty_1 \) the product distribution under which the two collections \( \{\ldots, X_{-1}, X_0\} \) and \( \{X_1, X_2, \ldots\} \) are independent. Then

\[
\beta(k) = \sup\{\|\mathbb{P}(A) - \mathbb{P}^0_{-\infty} \times \mathbb{P}^\infty_1(A)\| : A \in \sigma(X^0_{-\infty}, X^\infty_1)\}.
\]

**Lemma B** \( \beta(k) = E\left[\sup_{A \in \sigma(X^\infty_{-\infty})} \|\mathbb{P}_0(A) - \mathbb{P}(A)\|\right] \).

**Remark:** Additional conditions on the process \( \{X_i\}_{i=-\infty}^{\infty} \) are necessary to insure that the expression in Lemma B is meaningful. It may be enough, for instance, to assume the existence of a regular conditional distribution with respect to the \( \sigma \)-field \( \mathcal{G}_0 \). In any case, \( \beta(k) \) is certainly an upper bound if the supremum is taken over a countable number of sets.

The following observation is due in part to Steele (1978).

**Lemma 1** Let \( \{X_i\}_{i=-\infty}^{\infty} \) be a stationary, ergodic process and let \( \mathcal{F} \) be a permissible class of functions with integrable envelope \( F \). Then \( \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \) converges almost surely to a constant \( \eta \) as \( n \) tends to infinity. Moreover,

\[
\eta = \inf_{\mathcal{F}} E \left[ \sup_{n} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef \right| \right].
\]

**Proof:** Define a sequence of functions \( \{g_n\}_{n=1}^{\infty} \) via

\[
g_n(\omega) = \sup_{\mathcal{F}} \left| \sum_{i=0}^{n-1} (f(X_i(\omega)) - Ef) \right| = \sup_{\mathcal{F}} \left| \sum_{i=0}^{n-1} (f(X_0(T_i(\omega))) - Ef) \right|.
\]
Then it is easy to see that for every $n, m \geq 1$, $g_{n+m}(\omega) \leq g_n(\omega) + g_m(T^n\omega)$. Moreover, each function $g_n$ is integrable as $0 \leq E g_n \leq 2nEF < \infty$. It follows from Kingman’s subadditive ergodic theorem that
\[
\lim_{n \to \infty} \frac{g_n}{n} = \lim_{n \to \infty} \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| = \eta \quad \text{a.s.}
\]
where $0 \leq \eta \leq 2EF$ is the constant given by (5). □

So we see that $\sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right|$ will converge almost surely to a constant for any permissible class $\mathcal{F}$ with integrable envelope. This was first noted by Steele (1978) who strengthened the conclusion of Theorem A from convergence in probability to almost sure convergence. Here we will also make use of equation (5) which identifies the constant with a limit of expectations. Our primary concern in this section is the case $\eta = 0$, which characterizes uniform convergence. In the last section we will consider the interpretation and estimation of $\eta$ in the case $\eta > 0$.

Let the stationary process $\{X_i\}_{i=-\infty}^{\infty}$ have distribution $\mathbb{P}$ with one-dimensional marginal $P$ and let $\mathbb{P}_0$ denote the product distribution generated by $P$ (this is the independent approximation to $\mathbb{P}$). We are now in position to state our main result.

**Theorem 1** Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary absolutely regular stochastic process with distribution $\mathbb{P}$ having independent approximation $\mathbb{P}_0$. Let $\mathcal{F}$ be a uniformly bounded, permissible class of functions. Then if $\frac{1}{n} \log N(X_0^{n-1}, \varepsilon, \mathcal{F}) \to 0$ in probability $\mathbb{P}_0$,
\[
\sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \to 0
\]
almost surely $\mathbb{P}$.

The theorem tells us that it is sufficient to evaluate the asymptotic behavior of the covering numbers under the assumption that the sequence $\{X_i\}$ is i.i.d.. One need not know the precise nature of the dependence between the $X_i$ in order to verify the conditions of the theorem.

**Proof:** Without loss of generality we may assume that $Ef = 0$ for every $f$ in $\mathcal{F}$. Then we wish to show $\sup_{\mathcal{F}} \left| \frac{1}{n} \sum f(X_i) \right| \to 0$ a.s. $\mathbb{P}$. Let $K$ be an envelope for $\mathcal{F}$. If $n = km$ then
\[
\sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(x_i) \right| \leq \frac{1}{k} \sup_{\mathcal{F}} \left| \frac{1}{k} \sum_{j=0}^{k-1} f(x_{jk+j}) \right| \leq \frac{1}{m} \sup_{\mathcal{F}} \left| \frac{1}{m} \sum_{l=0}^{m-1} f(x_{lk+l}) \right|
\]
(6)
Let $\delta > 0$. Taking expectations of both sides in (6) gives
\[
E \left[ \sup_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \right] \leq \frac{1}{k} \sum_{j=0}^{k-1} E \left[ \sup_{\mathcal{F}} \frac{1}{m} \sum_{l=0}^{m-1} f(X_{l+k}) \right] \\
= E \left[ \sup_{\mathcal{F}} \frac{1}{m} \sum_{l=0}^{m-1} f(X_{lk}) \right] \\
\leq \delta + K \mathbb{P} \left\{ \sup_{\mathcal{F}} \frac{1}{m} \sum_{l=0}^{m-1} f(X_{lk}) > \delta \right\},
\]
where the equality in (7) follows by stationarity. Lemma 2 allows us to express the second term of (8) in terms of the probability $\mathbb{P}_0$ plus an additional error term involving $\beta(k)$. We make use of the following

**Lemma 2** If $A \in \sigma(X_0, X_k, \ldots, X_{(m-1)k})$ then $|\mathbb{P}(A) - \mathbb{P}_0(A)| \leq m\beta(k)$.

**Proof of Lemma 2:** The event $A$ is determined by a regularly spaced sequence of random variables. We may apply Lemma A repeatedly, first to $A$ itself and then, by integrating, to sections of $A$ determined by progressively longer initial segments of $(x_0, x_k, \ldots, x_{(m-1)k})$. At each stage we add an additional $\beta(k)$ to our estimate of the difference. \(\square\)

Returning now to the proof of the theorem, we apply Lemma 2 to (8) to get
\[
E \left[ \sup_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \right] \leq \delta + Km\beta(k) + K \mathbb{P}_0 \left\{ \sup_{\mathcal{F}} \frac{1}{m} \sum_{l=0}^{m-1} f(X_{lk}) > \delta \right\} \\
= \delta + Km\beta(k) + K \mathbb{P}_0 \left\{ \sup_{\mathcal{F}} \frac{1}{m} \sum_{l=0}^{m-1} f(X_{lk}) > \delta \right\}
\]

As $\frac{1}{n} \log N(X_0^{n-1}, \epsilon, \mathcal{F}) \to 0$ in probability $\mathbb{P}_0$, the last term above tends to zero as $m \to \infty$ by virtue of Theorem 1. Since $\beta(k) \to 0$ as $k \to \infty$ we can select increasing sequences of integers $\{m_i\}$ and $\{k_i\}$ such that $m_i \to \infty$, $k_i \to \infty$ and $m_i\beta(k_i) \to 0$.

Let $n_i = m_i k_i$. It follows from (9) and the choice of $m_i$, $k_i$ that
\[
\inf_n E \left[ \sup_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \right] \leq \inf_i E \left[ \sup_{\mathcal{F}} \frac{1}{n_i} \sum_{j=0}^{n_i-1} f(X_j) \right] \\
\leq \delta .
\]

As $\delta > 0$ is arbitrary, we have
\[
\inf_n E \left[ \sup_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \right] = 0 ,
\]
and the result follows from Lemma 1. □

**Corollary 3** Theorem 1 continues to hold if the class $\mathcal{F}$ has an integrable envelope $F$.

**Proof:** Fix $\epsilon > 0$. As $F$ is integrable there is a real number $K > 0$ such that

$$\int_{\{F>K\}} F(x) dP(x) < \epsilon .$$

Then we have

$$\sup_{\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef \right\} \leq \sup_{\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} fI_{\{F\leq K\}}(X_i) - EfI_{\{F\leq K\}} \right\}$$

$$+ \sup_{\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} fI_{\{F>K\}}(X_i) \right\} + \sup_{\mathcal{F}} EfI_{\{F>K\}} \quad (11)$$

As $|f| \leq F$ for each $f$ in $\mathcal{F}$ the last two terms in (11) are bounded by

$$\frac{1}{n} \sum_{i=0}^{n-1} F I_{\{F>K\}}(X_i) + E F I_{\{F>K\}} \quad (12)$$

As $n$ tends to infinity, the ergodic theorem and our choice of $K$ guarantee that the sum of the terms in (12) is almost surely less than $2\epsilon$. Thus, it is enough to show that

$$\sup_{\mathcal{F}_K} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef \right\} \to 0 \text{ a.s.} \quad (13)$$

where the class $\mathcal{F}_K = \{fI_{\{F\leq K\}} : f \in \mathcal{F}\}$ is uniformly bounded. But this follows immediately from Theorem 1 and the fact that $N(x_0^{n-1}, \epsilon, \mathcal{F}_K) \leq N(x_0^{n-1}, \epsilon, \mathcal{F})$. □

### 4 Closure Properties

In order to apply the results of Theorem A or Theorem 1 to problems of interest, it is necessary to develop techniques which enable us to demonstrate that a particular class of functions $\mathcal{F}$ is uniformly convergent. Typically, one begins with some small collection of known uniformly convergent classes and then generates the class $\mathcal{F}$ by applying various closure properties to these fundamental classes. In this section we present several methods for combining or modifying uniformly convergent classes of functions to get new uniformly convergent classes. For the existence of some basic
uniformly convergent classes, we refer the reader to Dudley (1978), Pollard (1984) and Steele (1975).

Let \( \{X_i\}_{i=0}^{\infty} \) be a stationary, ergodic process defined on \((\mathcal{X}_0^\infty, S_0^\infty, \mathbb{P})\) and let \( \mathbb{P} \) be the one-dimensional marginal of \( \mathbb{P} \). For \( 1 \leq q \leq \infty \) we define

\[
VC_q(\mathbb{P}) = \left\{ \mathcal{F} \subseteq \mathbb{R}^\mathcal{X} : \mathcal{F} \text{ is permissible with envelope } F \in L_q(\mathbb{P}) \text{ and } \lim_{n \to \infty} \sup_{\mathcal{F}} |\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0)| = 0 \text{ a.s. } \mathbb{P} \right\}
\]

When the subscript \( q \) is irrelevant to our discussion, we shall omit it. If the process is i.i.d. under \( \mathbb{P} \), we write \( VC_q(\mathbb{P}) \) rather than \( VC_q(\mathbb{P}) \).

**Proposition 1** The following relations and closure properties are valid:

a. If \( \mathcal{F}_1, \mathcal{F}_2 \in VC(\mathbb{P}) \) then \( \mathcal{F}_1 + \mathcal{F}_2 = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\} \in VC(\mathbb{P}) \)

b. If \( \mathcal{F} \in VC(\mathbb{P}), \alpha \in \mathbb{R} \) then \( \alpha \mathcal{F} = \{\alpha f : f \in \mathcal{F}\} \in VC(\mathbb{P}) \)

c. If \( \mathcal{F}_1, \mathcal{F}_2 \in VC(\mathbb{P}) \) then \( \mathcal{F}_1 \cup \mathcal{F}_2 \in VC(\mathbb{P}) \)

d. If \( \mathcal{F}_1 \in VC(\mathbb{P}), \mathcal{F}_2 \subseteq \mathcal{F}_1 \), then \( \mathcal{F}_2 \in VC(\mathbb{P}) \)

e. If \( \mathbb{Q} \ll \mathbb{P} \) then \( VC(\mathbb{P}) \subseteq VC(\mathbb{Q}) \)

f. If \( \mathcal{F} \in VC(\mathbb{Q}), VC(\mathbb{P}) \) then \( \mathcal{F} \in VC(\lambda \mathbb{Q} + (1 - \lambda)\mathbb{P}) \) \( \forall \lambda \in [0, 1] \)

g. If \( \mathbb{P} \) is the distribution of an absolutely regular process and \( \mathcal{F} \in VC(\mathbb{P}) \), then \( \mathcal{F} \in VC(\mathbb{P}) \)

**Proof:** Each of a through f follows directly from the definition of uniform convergence, while g follows from Theorem 1 and Theorem A. □

In particular, \( VC(\mathbb{P}) \) is closed under addition, scaling and union. Note that the scalar multiplication defined in b is associative, and that the addition operation defined in a is commutative and associative. Together these operations obey the usual distributive laws.

The next proposition indicates that any class \( \mathcal{F} \) which can be approximated by a uniformly convergent class is itself uniformly convergent. Its proof is virtually identical to the proof of Theorem B, which can be found in Pollard (1984).
Proposition 2 Let $\mathcal{F} \subseteq \mathbb{R}^H \subseteq \mathbb{R}^H$ be permissible, and let $\mathcal{F}' \in VC(\mathbb{P})$. Suppose that for each $\epsilon > 0$ and for each $f$ in $\mathcal{F}$ there exist $f_1, f_2$ in $\mathcal{F}'$ for which

$$f_1 \leq f \leq f_2 \quad \text{and} \quad \int (f_2 - f_1) d\mathbb{P} \leq \epsilon.$$ 

Then $\mathcal{F} \in VC(\mathbb{P})$.

The following theorem relies on the relationship established in Theorem A between uniform convergence and the asymptotic behavior of the covering numbers. In what follows, we restrict our attention to i.i.d processes.

Theorem 2 Let $\mathcal{F}_1, \mathcal{F}_2 \in VC(\mathbb{P})$ and let $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be continuous. If $g(\mathcal{F}_1, \mathcal{F}_2) = \{g(f_1, f_2) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ has envelope $G \in L_1(\mathbb{P})$, then $g(\mathcal{F}_1, \mathcal{F}_2) \in VC(\mathbb{P})$.

Proof: By Theorem A it is enough to show that for every $\delta > 0$

$$\frac{1}{n} \log N(x_0^{n-1}, \delta, g(\mathcal{F}_1, \mathcal{F}_2)) \rightarrow 0 \quad \text{in probability}.$$ 

Let $\delta > 0$ be fixed. Let $F_1, F_2 \in L_1(\mathbb{P})$ be envelopes for $\mathcal{F}_1, \mathcal{F}_2$, respectively. As $G \in L_1(\mathbb{P})$ we may choose $K$ so large that

$$\int_{(\max(F_1, F_2) > K)} G d\mathbb{P} < \frac{\delta}{8}. \quad (14)$$

Now $g$ is uniformly continuous on $[-K, K] \times [-K, K] = [-K, K]^2$, so there exists a number $\gamma > 0$ such that for all $(y, z), (y', z') \in [-K, K]^2$ if $|y - y'| \leq \gamma$ and $|z - z'| \leq \gamma$ then $|g(y, z) - g(y', z')| < \frac{\delta}{8}$.

Fix an arbitrary sequence $x_0, \ldots, x_{n-1}$ of elements of $\mathcal{X}$, and let $\mathcal{H}_1, \mathcal{H}_2$ be minimal $\gamma$-coverings of $\mathcal{F}_1, \mathcal{F}_2$ on $x_0^{n-1}$ such that $\mathcal{H}_1 \subseteq \mathcal{F}_1$ and $\mathcal{H}_2 \subseteq \mathcal{F}_2$. By virtue of Remark 1,

$$|\mathcal{H}_1| \leq N(x_0^{n-1}, \gamma/2, \mathcal{F}_1), \quad |\mathcal{H}_2| \leq N(x_0^{n-1}, \gamma/2, \mathcal{F}_2)$$

Our goal is to show that for most sequences $x_0^{n-1}$ the class $g(\mathcal{H}_1, \mathcal{H}_2)$ is a $\delta$-covering of $g(\mathcal{F}_1, \mathcal{F}_2)$ on $x_0^{n-1}$. 13
Let \( f_1, f_2 \) be arbitrary functions in \( \mathcal{F}_1, \mathcal{F}_2 \), respectively. Select functions \( h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2 \) for which
\[
\frac{1}{n} \sum_{i=0}^{n-1} |f_1(x_i) - h_1(x_i)| < \gamma \quad \text{and} \quad \frac{1}{n} \sum_{i=0}^{n-1} |f_2(x_i) - h_2(x_i)| < \gamma.
\]
(15)
If \( x_i \in \{\max(F_1, F_2) \leq K\} \), then \((f_1(x_i), f_2(x_i)), (h_1(x_i), h_2(x_i)) \in [-K, K]^2\), so that
\[
|g(f_1(x_i), f_2(x_i)) - g(h_1(x_i), h_2(x_i))| \\
\leq |g(f_1(x_i), f_2(x_i)) - g(h_1(x_i), f_2(x_i))| + |g(h_1(x_i), h_2(x_i)) - g(h_1(x_i), f_2(x_i))| \\
\leq \frac{\delta}{8} \left[ 1 + \frac{|f_1(x_i) - h_1(x_i)|}{\gamma} \right] + \frac{\delta}{8} \left[ 1 + \frac{|f_2(x_i) - h_2(x_i)|}{\gamma} \right],
\]
(16)
where the last inequality follows from the uniform continuity of \( g \) by dividing the intervals between \( f_1(x_i) \) and \( h_1(x_i) \), and between \( f_2(x_i) \) and \( h_2(x_i) \) into segments of length \( \gamma \). If \( x_i \not\in \{\max(F_1, F_2) \leq K\} \) then
\[
|g(f_1(x_i), f_2(x_i)) - g(h_1(x_i), h_2(x_i))| \leq 2G(x_i)
\]
(17)
From (15), (16) and (17) we have
\[
\frac{1}{n} \sum_{i=0}^{n-1} |g(f_1(x_i), f_2(x_i)) - g(h_1(x_i), h_2(x_i))| \leq \frac{\delta}{2} + \frac{1}{n} \sum_{i=0}^{n-1} 2G(x_i)\mathbb{1}_{\{\max(F_1, F_2) > K\}}(x_i)
\]
(18)
Now let \( A_n \) be the set of all sequences \( x_0^{n-1} \) for which the second term in (18) is large:
\[
A_n = \left\{ x_0^{n-1} \in \mathcal{X}^n : \frac{1}{n} \sum_{i=0}^{n-1} 2G(x_i)\mathbb{1}_{\{\max(F_1, F_2) > K\}}(x_i) > \frac{\delta}{2} \right\}.
\]
From (18) it follows that for any sequence \( x_0^{n-1} \in A_n^c \)
\[
N(x_0^{n-1}, \delta, g(\mathcal{F}_1, \mathcal{F}_2)) \leq |g(\mathcal{H}_1, \mathcal{H}_2)| \\
\leq |\mathcal{H}_1||\mathcal{H}_2| \\
\leq N(x_0^{n-1}, \gamma/2, \mathcal{F}_1)N(x_0^{n-1}, \gamma/2, \mathcal{F}_2),
\]
from which we conclude that
\[
\frac{1}{n} \log N(x_0^{n-1}, \delta, g(\mathcal{F}_1, \mathcal{F}_2)) \leq \frac{1}{n} \log N(x_0^{n-1}, \gamma/2, \mathcal{F}_1) + \frac{1}{n} \log N(x_0^{n-1}, \gamma/2, \mathcal{F}_2).
\]

14
Thus, for any $\epsilon > 0$

$$\Pr\left\{ \frac{1}{n} \log N(x_0^{n-1}, \delta, g(\mathcal{F}_1, \mathcal{F}_2)) \geq \epsilon \right\} \leq \Pr(A_n) + \Pr\left\{ \frac{1}{n} \log N(x_0^{n-1}, \delta, g(\mathcal{F}_1, \mathcal{F}_2)) \geq \epsilon, A^c_n \right\}$$

$$\leq \Pr(A_n) + \Pr\left\{ \frac{1}{n} \log N(x_0^{n-1}, \gamma/2, \mathcal{F}_1) \geq \frac{\epsilon}{2} \right\} + \Pr\left\{ \frac{1}{n} \log N(x_0^{n-1}, \gamma/2, \mathcal{F}_2) \geq \frac{\epsilon}{2} \right\}.$$  \hspace{1cm} (19)

Now, from (14) and our assumption that $\mathcal{F}_1, \mathcal{F}_2 \in VC(P)$, we see that the right hand side of (19) tends to 0 as $n$ tends to infinity.

**Remarks:**

a. When Theorem 2 is combined with part $g$ of Proposition 1 it implies that $g(\mathcal{F}_1, \mathcal{F}_2) \in VC(\Pr)$ for any absolutely regular process with one-dimensional marginal $P$.

b. The function $g$ need not be continuous on all of $\mathbb{R}^2$. Careful inspection of the proof indicates that it is enough for $g$ to be continuous on some simply connected, closed subset of $\mathbb{R}^2$ containing the set $\{(f_1(x), f_2(x)) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2, x \in \mathcal{X}\}$.

The following corollaries indicate some possible applications of theorem 2.

**Corollary 4** If $\mathcal{F}_1, \mathcal{F}_2 \in VC_1(P)$, then $\max(\mathcal{F}_1, \mathcal{F}_2) \in VC_1(P)$.

**Proof:** It suffices to establish the existence of an integrable envelope for the class $\max(\mathcal{F}_1, \mathcal{F}_2)$. But if $F_1, F_2 \in L_1(P)$ are envelopes for $\mathcal{F}_1$ and $\mathcal{F}_2$, then $F_1 + F_2 \in L_1(P)$ is an envelope for $\max(\mathcal{F}_1, \mathcal{F}_2)$. \hspace{1cm} $\square$

**Corollary 5** If $\mathcal{F}_1, \mathcal{F}_2 \in VC_2(P)$, then $\mathcal{F}_1 \mathcal{F}_2 \in VC_1(P)$

**Proof:** Let $F_1 \in L_2(P), F_2 \in L_2(P)$ be envelopes for $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively. Then $F_1 F_2$ is an envelope for $\mathcal{F}_1 \mathcal{F}_2$, and by the Cauchy-Schwartz inequality $F_1 F_2 \in L_1(P)$. \hspace{1cm} $\square$

**Corollary 6** Let $\mathcal{F} \in VC_1(P)$ and suppose that there exists a number $a > 0$ such that $f \geq a$ for every $f$ in $\mathcal{F}$. Then $\log \mathcal{F} \in VC_1(P)$

15
Proof: Let $F \in L_1(P)$ be an envelope for $\mathcal{F}$. Then $\log F$ is an envelope for $\log \mathcal{F}$, and by Jensen's inequality, $\log F \in L_1(P)$. □

Corollary 7 If $\mathcal{F} \in VC_q(P)$ with $1 \leq q < \infty$, then

$$\sup_{\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} |f(X_i)|^q \right\}^{1/q} - \|f\|_q \to 0$$

almost surely as $n$ tends to infinity.

Proof: Let $g(y) = |y|^q$, and let $F \in L_1(P)$ be an envelope for $\mathcal{F}$. Then $F^q \in L_1(P)$ is an envelope for $g(\mathcal{F})$, so by Theorem 2 $g(\mathcal{F}) \in VC_1(P)$. Writing this in terms of $\mathcal{F}$ we have

$$\sup_{\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=0}^{n-1} |f(X_i)|^q - E|f|^q \right\} \to 0 \text{ a.s. } P_0$$

The result now follows by applying the uniform continuity of the function $h = (\cdot)^{1/q}$ to each sample sequence $X_0(\omega), X_1(\omega), \ldots$ for which the convergence above holds. □

5 Empirical Risk Minimization and a Measure of Asymptotic Overfit

Consider again the application of uniform convergence given in the introduction. There we saw that a specific pattern recognition problem reduced to that of selecting a function $f$ with minimum expectation from a class of functions $\mathcal{F}$. One can analyze this latter problem in terms of uniform convergence and the method of empirical risk minimization. Vapnik (1989, 1982) has shown that one may carry out a similar reduction and analysis for many other problems involving, either implicitly or explicitly, risk minimization. Density estimation and composite hypothesis testing are examples of such problems. The reduction typically depends on the particular problem at hand, while the subsequent analysis is quite general. Accordingly, in this section we will consider inductive inference in the more general framework suggested by the example. We begin by restating the general problem and then giving a precise definition of empirical risk minimization.
Let $\mathcal{F}$ be a class of real-valued functions on a set $\mathcal{X}$, and let $\{X_i\}_{i=0}^{\infty}$ be a stationary, ergodic sequence of random variables taking values in the set $\mathcal{X}$. We wish to select a function $f \in \mathcal{F}$ with minimal expectation $Ef$. In practice, often the common distribution $P$ of the random variables $X_i$ is unknown (or the evaluation of $Ef$ is not feasible). Then we have to resort to empirical selection rules based on the observations $X_0, X_1, \ldots$. More precisely, we wish to find a procedure which associates with almost every realization $X_0(\omega), X_1(\omega), \ldots$ of the process $\{X_i\}$, a sequence of functions $\{f_n\}_{n=1}^{\infty}$ from $\mathcal{F}$ such that

$$\lim_{n \to \infty} Ef_n = \inf_{\mathcal{F}} Ef.$$ 

Here $Ef_n = \int f_n dP$ is the expectation of $f_n$ with respect to a new independent observation and not with respect to $X_0(\omega), X_1(\omega), \ldots$. Such a procedure is said to be consistent. A procedure is called inductive if $f_n$ is fully determined by the first $n$ observations $X_0(\omega), \ldots, X_{n-1}(\omega)$. Thus, we are led to the following

**Definition:** Let $\{\epsilon_n\}_{n=1}^{\infty}$ be a fixed sequence of positive numbers tending to zero. An empirical risk minimization rule (with respect to $\mathcal{F}$ and $\{\epsilon_n\}$) is a rule that assigns to every realization $X_0(\omega), X_1(\omega), \ldots$ of the process $\{X_i\}_{i=0}^{\infty}$, a sequence of functions $\{f_n\}_{n=1}^{\infty}$ from $\mathcal{F}$ with the property that for $n = 1, 2, \ldots$,

a. $f_n$ is determined by the first $n$ observations $X_0(\omega), X_1(\omega), \ldots, X_{n-1}(\omega)$, and

b. $\frac{1}{n} \sum_{i=0}^{n-1} f_n(X_i(\omega)) \leq \inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i(\omega)) + \epsilon_n$.

**Proposition 3** *Any empirical risk minimization rule with respect to a uniformly convergent class $\mathcal{F}$ is consistent.*

**Proof:**

$$|Ef_n - \inf_{\mathcal{F}} Ef| \leq |Ef_n - \frac{1}{n} \sum_{i=0}^{n-1} f_n(X_i)| + |\frac{1}{n} \sum_{i=0}^{n-1} f_n(X_i) - \inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i)|$$

$$+ |\inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \inf_{\mathcal{F}}Ef|$$

$$\leq \epsilon_n + 2 \sup_{\mathcal{F}} |\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef|.$$  

(20)
By definition $\epsilon_n \to 0$. The uniform convergence of $\mathcal{F}$ means that $\sup_{\mathcal{F}} |\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef| \to 0$ and so by (20)

$$\lim_{n \to \infty} Ef_n = \inf_{\mathcal{F}} Ef,$$

almost surely in $\mathcal{F}$. □

The closure properties given in the previous section enable us to extend this result. Indeed, if the class $\mathcal{F}$ is uniformly convergent, we can establish the consistency of simple inductive procedures for solving a wide variety of problems involving $\mathcal{F}$. As an example, suppose that the process $\{X_i\}_{i=-\infty}^{\infty}$ is absolutely regular, and consider the problem of finding a function $f$ in $\mathcal{F}$ with minimum $L_q$ norm, $\|f\|_q$. If $\mathcal{F} \in VC_q(P)$, it follows from Corollary 7 and the argument above that the obvious inductive procedure, which selects $f_n$ to minimize $(\frac{1}{n} \sum_{i=0}^{n-1} |f(X_i)|^q)^{1/q}$, is consistent.

What can be said about the consistency of empirical risk minimization rules when the class $\mathcal{F}$ is not uniformly convergent? In this case, the constant

$$\eta(\mathcal{F}) = \inf_n E \left[ \sup_{\mathcal{F}} |\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0)| \right]$$

given by Lemma 1 is greater than zero. This indicates that the class $\mathcal{F}$ overfits the data: sample averages of functions $f \in \mathcal{F}$ do not give uniformly good estimates of their expectations. The following result extends the applicability of empirical risk minimization to families $\mathcal{F}$ where $\eta(\mathcal{F}) > 0$.

**Theorem 3** Empirical risk minimization rules produce, with probability one, a sequence of functions $\{f_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ for which

$$\limsup_{n \to \infty} Ef_n \leq \inf_{\mathcal{F}} Ef + \eta(\mathcal{F}).$$

**Proof:** Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of functions corresponding to a realization $X_0(\omega), X_1(\omega), \ldots$ of the process for which $\sup_{\mathcal{F}} |\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0)|$ converges to $\eta(\mathcal{F})$. Henceforth we will omit any reference to the sample point $\omega$. From the definition of $\eta(\mathcal{F})$ we see that for every $\epsilon > 0$ there is an integer $N$ such that for every $n \geq N$, $Ef_n \leq \frac{1}{n} \sum_{i=0}^{n-1} f_n(X_i) + \eta(\mathcal{F}) + \epsilon$. Hence,

$$\limsup_{n \to \infty} Ef_n \leq \limsup_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f_n(X_i) + \eta(\mathcal{F}) + \epsilon.$$
\[
\leq \limsup_{n \to \infty} \inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) + \eta(\mathcal{F}) + \epsilon,
\]
where the last inequality follows from the definition of empirical risk minimization. Letting \( \epsilon \to 0 \) gives
\[
\limsup_{n \to \infty} Ef_n \leq \limsup_{n \to \infty} \inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) + \eta(\mathcal{F}),
\]
so it is enough to show that
\[
\limsup_{n \to \infty} \inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \leq \inf_{\mathcal{F}} Ef.
\]
To see this, fix \( \delta > 0 \) and select a function \( f' \in \mathcal{F} \) such that \( Ef' \leq \inf_{\mathcal{F}} Ef + \delta \). By the ergodic theorem, there is an integer \( N \) such that for every \( n \geq N \), \( \frac{1}{n} \sum_{i=0}^{n-1} f'(X_i) \leq Ef' + \delta \). So, for \( n \geq N \),
\[
\inf_{\mathcal{F}} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \leq \frac{1}{n} \sum_{i=0}^{n-1} f'(X_i) \leq Ef' + \delta \leq \inf_{\mathcal{F}} Ef + 2\delta.
\]
Taking the lim sup of both sides of (21) and letting \( \delta \to 0 \) completes the proof. \( \square \)

Consider an application where we wish to apply an inductive procedure to find a function with small expectation in a large class of functions \( \mathcal{F} \). Suppose that the specific nature of the application allows us a certain tolerance, or margin of error, \( \epsilon \), in the asymptotic performance of the inductive procedure. If \( \eta(\mathcal{F}) \leq \epsilon \), then Theorem 3 indicates that the asymptotic performance of empirical risk minimization rules is within the prescribed margin of error. In these situations we may say that the method of empirical risk minimization is \( \epsilon \)-consistent. In order to apply Theorem 3 to a particular problem, we need to know the value of \( \eta(\mathcal{F}) \), or at least a good approximation to it. A direct calculation involving the definition is impossible, since we have assumed that we do not know the marginal distribution \( P \) of the observations. Instead, we must try to use the observations themselves to determine \( \eta(\mathcal{F}) \). In view of the fact that \( \eta(\mathcal{F}) \) is a measure of what the observations do not tell us about the class \( \mathcal{F} \), such an endeavor seems unlikely to succeed.

Indeed, it appears that a precise empirical determination of \( \eta(\mathcal{F}) \) is not possible in general. However, we can exhibit a procedure which yields good upper and lower
bounds on $\eta(F)$ when the class $F$ is well-behaved, and the sequence of observations is stationary and absolutely regular. The method we propose is based on a resampling idea from statistics known as cross-validation (c.f. Stone (1974)). Our application of the idea is simple: replace the expectations appearing in the definition of $\eta(F)$ by sample averages over additional observations. So, at time $n$ we collect $2n$ samples, $X_0, \ldots, X_{2n-1}$, and replace each term

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0)$$

appearing in the definition of $\eta(F)$ by a difference of averages

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j).$$

So we are interested in

$$\sup_{F} \left| \frac{1}{n} \sum_{i=-n}^{-1} f(X_i) - \frac{1}{n} \sum_{j=0}^{n-1} f(X_j) \right|.$$

Consider for the moment a more symmetric version of the quantity in (22), namely

$$\sup_{F} \frac{1}{n} \sum_{i=-n}^{-1} f(X_i) - \frac{1}{n} \sum_{j=0}^{n-1} f(X_j) = \sup_{F} \frac{1}{n} \left| \sum_{i=-n}^{-1} f(X_i) - \sum_{j=0}^{n-1} f(X_j) \right|.$$

Lemma 3 Let $F \subseteq \mathbb{R}^X$ be a permissible class of functions with integrable envelope, and let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary absolutely regular process with distribution $P$. Then

$$\sup_{F} \frac{1}{n} \left| \sum_{i=-n}^{-1} f(X_i) - \sum_{j=0}^{n-1} f(X_j) \right|$$

converges almost surely to a constant $\hat{\eta}(F)$ as $n$ tends to infinity.

Proof: Define the set

$$E = \left\{ \lim_{n \to \infty} \sup_{F} \frac{1}{n} \left| \sum_{i=-n}^{-1} f(X_i) - \sum_{j=0}^{n-1} f(X_j) \right| \text{ exists} \right\}.$$

Then $E \in \sigma(X_{-\infty}^0, X_k^\infty)$ for every $k \geq 1$. It follows by Lemma A that

$$|P(E) - P_{-\infty}^0 \times P_1^\infty(E)| \leq \beta(k)$$

for every $k \geq 1$. As the process $\{X_i\}_{i=-\infty}^{\infty}$ is absolutely regular, $P(E) = P_{-\infty}^0 \times P_1^\infty(E)$.
Consider now the product of two identical copies of the underlying probability space \((X_{\infty,-\infty}, S_{\infty,-\infty}, \mathbb{P})\). Let the process \(\{\hat{X}_i\}\) correspond to the action of the left shift \(T\) on a sample point \(\hat{\omega}\) in the first copy of \(X_{\infty,-\infty}\), and let the process \(\{\hat{X}_i\}\) correspond to the action of \(T\) on a point \(\hat{\omega}\) in the second copy of \(X_{\infty,-\infty}\). Then the processes \(\{\hat{X}_i\}_{i=-\infty}^{\infty}\) and \(\{\hat{X}_i\}_{i=-\infty}^{\infty}\) are independent, identical versions of \(\{X_i\}_{i=-\infty}^{\infty}\). Define the set \(\tilde{E} \subseteq X_{\infty,-\infty} \times X_{\infty,-\infty}\) as follows:

\[
\tilde{E} = \left\{ (\hat{\omega}, \hat{\omega}) : \limsup_{n \to \infty} \frac{1}{n} \sum_{i=-n}^{-1} f(\hat{X}_i(\hat{\omega})) - \sum_{j=0}^{n-1} f(\hat{X}_j(\hat{\omega})) \right\}
\]

Then it is clear that \(\mathbb{P}^0_{-\infty} \times \mathbb{P}^\infty_1(E) = \mathbb{P} \times \mathbb{P}(\tilde{E})\).

Define on the product space the stationary transformation \(S = T^{-1} \times T\). Since the process \(\{X_i\}_{i=-\infty}^{\infty}\) is absolutely regular, \(T\) is weakly mixing, and it follows that \(S\) is ergodic (c.f. Petersen (1983)). Let \(\nu = (\hat{\omega}, \hat{\omega})\) denote a point in the product space. With every \(f \in F\) we may associate a function \(h_f(\nu) = f(\hat{X}_{-1}(\hat{\omega})) - f(\hat{X}_0(\hat{\omega}))\). Let \(\mathcal{H} = \{h_f : f \in F\}\) be the collection of all such functions. Then

\[
\tilde{E} = \left\{ \nu : \limsup_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} h(S^i \nu) \right\}
\]

and by Lemma 1, \(\mathbb{P} \times \mathbb{P}(\tilde{E}) = 1\). \(\Box\)

Remark: The same argument shows that for any uniformly bounded, permissible class of functions \(\mathcal{G}\),

\[
\sup_{\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} g(X_{-i}, X_i)
\]

converges almost surely to the constant

\[
\inf_n E \left[ \sup_{\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} g(\hat{X}_{-i}, \hat{X}_i) \right]
\]

as \(n\) tends to infinity.

**Corollary 8** The random variables

\[
\sup_{F} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j)
\]

converge in probability to the constant \(\hat{\eta}(F)\) as \(n\) tends to infinity.
Proof: By stationarity, the distribution of the random variable
\[ \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j) \right| \]
is the same as that of
\[ \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=-n}^{-1} f(X_i) - \sum_{j=0}^{n-1} f(X_j) \right| , \]
and the result follows from the previous lemma. □

Thus, if we can establish a relationship between \( \hat{\eta}(\mathcal{F}) \) and \( \eta(\mathcal{F}) \), we can estimate \( \eta(\mathcal{F}) \) by calculating \( \hat{\eta}(\mathcal{F}) \), which depends solely on the observations.

Consider a permissible class of functions \( \mathcal{F} \subseteq \mathbb{R}^X \) with integrable envelope. We say that \( \mathcal{F} \) is asymptotically separable if there is a countable subclass \( \mathcal{F}' \subseteq \mathcal{F} \) for which \( \eta(\mathcal{F}) = \eta(\mathcal{F}') \). Asymptotic separability insures that for large \( n \),
\[ \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \]
is determined with high probability by a fixed, countable class of functions \( \mathcal{F}' \). Asymptotic separability is related to the notion of universal separability, mentioned by Dudley (1978) in the context of regularity conditions for classes of functions. In particular, universal separability implies asymptotic separability.

Theorem 4 Let \( \{X_i\}_{i=-\infty}^{\infty} \) be a stationary, absolutely regular process with distribution \( \mathbb{P} \). If \( \mathcal{F} \in \text{VC}_1(\mathbb{P}) \) is asymptotically separable, then \( \eta(\mathcal{F}) \leq \hat{\eta}(\mathcal{F}) \leq 2\eta(\mathcal{F}) \).

Proof: The upper bound follows easily from the definition of \( \hat{\eta}(\mathcal{F}) \) by subtracting \( Ef \) from each of the two sample averages whose difference is considered in (23), and then applying the triangle inequality.

It remains to establish the lower bound. Let \( E \in L_1(\mathbb{P}) \) be an envelope for \( \mathcal{F} \). For every \( K > 0 \) the class \( \mathcal{F}_K = \{fI_{\{f \leq K\}} : f \in \mathcal{F}\} \) is uniformly bounded, with envelope \( K \). An argument similar to that given in the Corollary 3 shows that for every \( \epsilon > 0 \) there exists a number \( K' \) so large that \( |\eta(\mathcal{F}) - \eta(\mathcal{F}_{K'})| < \epsilon \) and \( |\hat{\eta}(\mathcal{F}) - \hat{\eta}(\mathcal{F}_{K'})| < \epsilon \). Thus, it is enough to show that \( \eta(\mathcal{F}) \leq \hat{\eta}(\mathcal{F}) \) whenever \( \mathcal{F} \) is an asymptotically separable class with envelope \( K' \).

Our approach is based in part on a symmetrization technique from the theory of probability in Banach spaces. The technique allows us to bound \( \eta(\mathcal{F}) \) by the sum

22
of $\eta(\mathcal{F})$ and an error term which depends on the mixing coefficients $\beta(j)$ and the constant $K'$. The absolute regularity of the process will insure that the error term tends to zero as $j \to \infty$.

We begin by splitting into two parts the term $\left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right|$ that appears in the definition of $\eta(\mathcal{F})$:

$$\left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \leq \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} \mathbb{E} \beta_n f(X_j) \right|$$

$$+ \left| \frac{1}{n} \sum_{j=n}^{2n-1} \mathbb{E} \beta_n f(X_j) - Ef(X_0) \right|. \tag{24}$$

Now let $\mathcal{F}'$ be a countable subset of $\mathcal{F}$ with $\eta(\mathcal{F}') = \eta(\mathcal{F})$. Then we see that

$$\eta(\mathcal{F}) = \lim_{n \to \infty} \mathbb{E} \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - Ef(X_0) \right| \right]$$

$$\leq \limsup_{n \to \infty} \mathbb{E} \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} \mathbb{E} \beta_n f(X_j) \right| \right]$$

$$+ \limsup_{n \to \infty} \mathbb{E} \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{j=n}^{2n-1} \mathbb{E} \beta_n f(X_j) - Ef(X_0) \right| \right]. \tag{25}$$

We now claim that the first term on the right hand side of equation (25) is bounded above by $\eta(\mathcal{F})$. To show this, we begin by noting that the basic properties of conditional expectations and the convexity of the absolute value allow us to write

$$\sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} \mathbb{E} \beta_n f(X_j) \right| = \sup_{\mathcal{F}'} \left| \mathbb{E} \beta_n \left( \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j) \right) \right|$$

$$\leq \sup_{\mathcal{F}'} \mathbb{E} \beta_n \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j) \right|$$

$$\leq \mathbb{E} \beta_n \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j) \right| \right]$$

Taking expectations of the first and last terms above gives

$$\mathbb{E} \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} \mathbb{E} \beta_n f(X_j) \right| \right] \leq \mathbb{E} \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) - \frac{1}{n} \sum_{j=n}^{2n-1} f(X_j) \right| \right]$$

$$= \mathbb{E} \left[ \sup_{\mathcal{F}'} \left| \frac{1}{n} \sum_{i=-n}^{1} f(X_i) - \frac{1}{n} \sum_{j=0}^{n-1} f(X_j) \right| \right],$$

23
where the last equality follows by stationarity. Applying Lemma 3 and the dominated convergence theorem to this last inequality establishes the claim. Therefore, the proof of Theorem 4 is complete if we can establish that the second term on the right side of equation (25) is zero.

We begin with a crude upper bound:

\[
E \left[ \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{j=n}^{2n-1} E_{\mathcal{G}_n} f(X_j) - Ef(X_0) \right| \right] \leq \frac{1}{n} \sum_{j=n}^{2n-1} E \left[ \sup_{\mathcal{F}} \left| E_{\mathcal{G}_n} f(X_j) - Ef(X_0) \right| \right] \\
\leq \frac{1}{n} \sum_{j=0}^{n-1} E \left[ \sup_{\mathcal{F}} \left| E_{\mathcal{G}_n} f(X_j) - Ef(X_0) \right| \right]
\]  

(26)

where in the last step we have used the fact that for any measure-preserving transformation \(T, E_{\mathcal{G}}(h) \circ T^r = E_{T^r \circ \mathcal{G}}(h \circ T^r)\). Therefore, if

\[
\lim_{j \to \infty} E \left[ \sup_{\mathcal{F}} \left| E_{\mathcal{G}_n} f(X_j) - Ef(X_0) \right| \right] = 0,
\]  

(27)

then the Cesaro averages in (26) converge to zero and we have

\[
\limsup_{n \to \infty} E \left[ \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{j=n}^{2n-1} E_{\mathcal{G}_n} f(X_j) - Ef(X_0) \right| \right] = 0,
\]

as desired. Equation (27) follows from the absolute regularity of the process and the following consequence of Lemma B.

**Lemma 4** Let \(\mathcal{F}\) be a countable class of functions with envelope \(K\), and let \(\{X_i\}_{i=-\infty}^{\infty}\) be a stationary stochastic process. Then for every \(j \geq 1\),

\[
E \left[ \sup_{\mathcal{F}} \left| E_{\mathcal{G}_n} f(X_j) - Ef(X_0) \right| \right] \leq 2K\beta(j).
\]

**Proof:** Define the class \(\mathcal{H} = \{f(X_j) : f \in \mathcal{F}\}\). Then \(\mathcal{H}\) is a countable collection of real valued functions on the underlying sample space \(\Omega\). Moreover, \(\mathcal{H}\) has envelope \(K\), and every function \(h \in \mathcal{H}\) is measurable \(\sigma(X_j^{\infty})\). We wish to show that

\[
E \left[ \sup_{\mathcal{H}} \left| E_{\mathcal{G}_n} h - h \right| \right] \leq 2K\beta(j)
\]

The right hand side of this equation is unchanged if we add \(K\) to every function \(h\). Hence we may assume that \(0 \leq h \leq 2K\) for any \(h \in \mathcal{H}\).
Given $\epsilon > 0$, choose a number $\Delta < \epsilon$ such that $L = 2K/\Delta$ is an integer. For each $h \in \mathcal{H}$ define a new function

$$h_\Delta = \sum_{k=1}^{L} \Delta I_{\{h \leq k\Delta\}}.$$

Then $|h_\Delta - h| \leq \Delta < \epsilon$ for every $h \in \mathcal{H}$, so

$$E \left[ \sup_{\mathcal{H}} |E_{\mathcal{G}_0} h - Eh| \right] \leq E \left[ \sup_{\mathcal{H}} |E_{\mathcal{G}_0} h_\Delta - h_\Delta| \right] + 2\epsilon. \quad (28)$$

Now consider the quantity $|E_{\mathcal{G}_0} h_\Delta - h_\Delta|$ appearing on the right hand side of (28):

$$|E_{\mathcal{G}_0} h_\Delta - h_\Delta| = \left| \sum_{k=1}^{L} \Delta \mathbb{P}_{\mathcal{G}_0} \{h \leq k\Delta\} - \sum_{k=1}^{L} \Delta \mathbb{P} \{h \leq k\Delta\} \right|$$

$$\leq \sum_{k=1}^{L} \Delta |\mathbb{P}_{\mathcal{G}_0} \{h \leq k\Delta\} - \mathbb{P} \{h \leq k\Delta\}|. \quad (29)$$

Taking expectations gives

$$E \left[ \sup_{\mathcal{H}} |E_{\mathcal{G}_0} h_\Delta - h_\Delta| \right] \leq \sum_{k=1}^{L} \Delta E \left[ \sup_{\mathcal{H}} |\mathbb{P}_{\mathcal{G}_0} \{h \leq k\Delta\} - \mathbb{P} \{h \leq k\Delta\}| \right]. \quad (30)$$

Note that for every $1 \leq k \leq L$ and every $h \in \mathcal{H}$ the set $\{h \leq k\Delta\}$ is contained in $\sigma(X_j^{\infty})$. Moreover, for each $k$ the supremum in (30) is taken over a countable collection of sets, so by Lemma B

$$E \left[ \sup_{\mathcal{H}} |E_{\mathcal{G}_0} h_\Delta - h_\Delta| \right] \leq \sum_{k=1}^{L} \Delta \beta(j) = 2K \beta(j)$$

Combine this inequality with (28) to get

$$E \left[ \sup_{\mathcal{H}} |E_{\mathcal{G}_0} h - h| \right] \leq 2K \beta(j) + 2\epsilon,$$

and letting $\epsilon \to 0$ completes the proof. □

One final remark is in order. In considering the asymptotic consistency of empirical risk minimization, with or without a margin of error, we are interested only in those functions $f \in \mathcal{F}$ whose expectations are greater than their sample averages. Following Vapnik and Cervonenkis (1988), we may refine our analysis by considering the behavior of the random variables

$$\sup_{\mathcal{F}} |Ef - \frac{1}{n} \sum_{i=0}^{n-1} f(X_i)|^+ \quad n = 1, 2, \ldots,$$  

(31)
where $|\cdot|^+$ is defined by

$$
|a|^+ = \begin{cases} 
a & \text{if } a > 0 \\
0 & \text{otherwise.}
\end{cases}
$$

Since $|a+b|^+ \leq |a|^++|b|^+$, we can apply the subadditive ergodic theorem as in Lemma 1 to show that for any stationary ergodic process $\{X_i\}$ the random variables in (31) will converge almost surely to a constant $\eta^+(\mathcal{F})$ as $n$ tends to infinity. Theorem 1, Lemma 3 and Theorem 4 all hold with $\eta^+(\mathcal{F})$ in place of $\eta(\mathcal{F})$. The new proofs are virtually identical to the old ones; they require only the convexity of $|\cdot|$, the inequality mentioned above and the fact that $|a|^+ \leq |a|$. 

Acknowlegements

One of us (A. Nobel) thanks Prof. Y. Katznelson, Prof. B. Efron, Prof. T. Cover and J. Gisladottir for their time and their advice.

References


