A Counterexample Concerning Uniform Ergodic Theorems for a Class of Functions

Andrew Nobel [*]

June 13, 1994

Abstract

Vapnik and Cervonenkis, and Talagrand, have characterized the Glivenko-Cantelli property for independent random variables. We show that these characterizations fail to hold for general stationary ergodic processes.

Appears in *Statistics and Probability Letters*, 24 165-168, 1995.

[*] Andrew Nobel is a Beckman Institute Fellow at the University of Illinois Urbana-Champaign. *Correspondence to*: Andrew B. Nobel, Beckman Institute, 405 N. Mathews, Urbana, IL 61801. Email: nobel@assistant.beckman.uiuc.edu

Key Words: Uniform strong law, Glivenko-Cantelli class, Vapnik-Cervonenkis class, stationary ergodic processes.

1 Introduction

Uniform strong laws extend the classical strong law of large numbers from a single real-valued function to a collection of such functions. They provide a useful methodology with which one can establish the consistency of simple inductive procedures for a variety of statistical problems, including pattern recognition (Vapnik (1982), Devroye (1988)), the training of neural networks (Baum and Haussler (1989), Hausser (1992), Faragó and Lugosi (1992)), and machine learning (Blumer *et al.* (1989)).

Consider a sequence $X_0, X_1, X_2, \ldots$ of independent, identically distributed (i.i.d.) random variables taking values in a measurable space $(\mathcal{X}, \mathcal{S})$, and having distribution P. If C is a measurable subset of $\mathcal{X}$, the strong law of large numbers guarantees that the relative frequency $\frac{1}{n}\sum_{i=0}^{n-1} I_C(X_i) \to P(C)$ with probability one as $n \to \infty$.

Let $\mathcal{C} \subseteq \mathcal{S}$ be a countable class of sets. A *uniform strong law* establishes the simultaneous convergence of relative frequencies for every set $C \in \mathcal{C}$. Specifically, $\mathcal{C}$ satisfies a uniform strong law with respect to $\{X_i\}$ if

$$\sup_{\mathcal{C}} \left| \frac{1}{n} \sum_{i=0}^{n-1} I_C(X_i) - P(C) \right| \to 0 \quad \text{w.p.1.} \tag{1}$$

In this case $\mathcal{C}$ is said to be a *Glivenko-Cantelli* class. Consideration of uncountable classes is straightforward when measurability issues are taken into account. Here we will consider only countable classes of sets.

Vapnik and Cervonenkis (1971) and later Talagrand (1987) found necessary and sufficient conditions under which (1) holds when $\{X_i\}$ is i.i.d.. Their results are presented in Theorems A and B of the next section. However, many applications of uniform strong laws occur in statistical problems where the sequence $\{X_i\}$ may exhibit short-term or long-term dependencies, e.g. samples of a speech waveform or contiguous blocks of pixels from a medical image. The application of Theorems A and B (or suitable variations) to such situations is clearly of interest. Nobel and Dembo (1993) showed that if the convergence in (1) holds for an i.i.d. sequence $\{X_i\}$, then it also holds for every stationary, absolutely regular process having the same one-dimensional marginal distribution as $\{X_i\}$. Peškir and Weber (1992) investigated necessary and sufficient conditions for uniform ergodic theorems. Yu (1993) showed that random entropy conditions are sufficient to guarantee a uniform ergodic theorem for suitable weakly Bernoulli processes. Peškir and Yukich (1994) extended these results and considered the more general case of absolutely regular dynamical systems. It is shown below that the combinatorial conditions of Vapnik, Cervonenkis, and Talagrand do not apply to stationary ergodic proceses. Standard random entropy conditions are not sufficient to insure uniform ergodic theorems in the case of strongly dependent random variables.

2 Preliminaries

The necessary and sufficient conditions of Vapnik and Cervonenkis and Talagrand can be expressed in terms of a simple combinatorial quantity that relates the class $\mathcal{C}$ and sample sequences of the process $\{X_i\}$. Let S be a finite subset of the set $\mathcal{X}$, and let $\mathcal{C}$ be any collection of subsets of $\mathcal{X}$. Define the *index* of $\mathcal{C}$ with respect to S

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

to be the number of distinct subsets of S induced by sets $C \in \mathcal{C}$. Clearly, $\Delta^{\mathcal{C}}(S) \leq 2^{|S|}$. If $\Delta^{\mathcal{C}}(S) = 2^{|S|}$, then $\mathcal{C}$ is said to *shatter* the set S: the class $\mathcal{C}$ shatters S precisely when $\mathcal{C}$ induces every subset of S. We say that $\mathcal{C}$ shatters an *infinite* set $S \subseteq \mathcal{X}$ if $\mathcal{C}$ shatters every finite subset of S.

Let $\mathcal{C} \subset \mathcal{S}$ be a countable class of measurable subsets of $\mathcal{X}$, and let the i.i.d. sequence $\{X_i\}$ be defined as above. In what follows $\Delta^{\mathcal{C}}(X_0, \ldots, X_{n-1})$ will denote the quantity $\Delta^{\mathcal{C}}(\{X_0, \ldots, X_{n-1}\})$. The theorem of Vapnik and Cervonenkis (1971) relates uniform strong laws and the asymptotic growth rate of the index $\Delta^{\mathcal{C}}(X_0, \ldots, X_{n-1})$.

Theorem A (Vapnik and Cervonenkis) *The uniform strong law (1) holds if and only if*

$$\frac{1}{n} \log \Delta^{\mathcal{C}}(X_0, \ldots, X_{n-1}) \to 0$$

in probability. □

A *VC class* $\mathcal{C}$ is one for which there exists an integer k such that $\Delta^{\mathcal{C}}(x_1, \ldots, x_k) < 2^k$ for every sequence $x_1, \ldots x_k$ in $\mathcal{X}$. In this case, it can be shown (*cf.* Sauer (1972)) that $\Delta^{\mathcal{C}}(x_1, \ldots, x_n) \leq n^k$ for every $n \geq k$, so the conditions of Theorem A are satisfied.

Assume now that the probability space $(\mathcal{X}, \mathcal{S}, P)$ is non-atomic. The theorem of Talagrand (1987) equates the failure of uniform strong laws with local instability of the class $\mathcal{C}$.

Theorem B (Talagrand) *The uniform strong law (1) fails to hold if and only if there is a set $A \in \mathcal{S}$ with $P(A) > 0$ having the property that, for almost every realization of the process $\{X_i\}$, $\mathcal{C}$ shatters the set $\{X_{n_1}, X_{n_2}, \ldots\}$ consisting of those X_i that lie in A.*

3 A Counterexample

Theorem 1 *For a suitable measurable space* $(\mathcal{X}, \mathcal{S})$ *and a suitable distribution* P*, there exists a countable collection* $\mathcal{C} \subseteq \mathcal{S}$*, and a stationary ergodic process* $\{Y_i\}$ *defined on* $(\mathcal{X}, \mathcal{S})$ *with marginal distribution* P *such that*

a. $\sup_{\mathcal{C}} \left| \frac{1}{n} \sum_{i=0}^{n-1} I_C(Y_i) - P(C) \right| \not\to 0$ *in probability*

b. $\frac{1}{n} \log \Delta^{\mathcal{C}}(Y_0, \ldots, Y_{n-1}) \to 0$ *in probability*

c. There is no set $A \in \mathcal{S}$ *with* $P(A) > 0$ *having the property that, for almost every realization of the process* $\{Y_i\}$*,* $\mathcal{C}$ *shatters the set* $\{Y_{n_1}, Y_{n_2}, \ldots\}$ *consisting of those* Y_i *that lie in* A*.*

Parts a and b of the theorem indicate that subexponential growth of the index $\Delta^{\mathcal{C}}$ along sample sequences does not guarantee the uniform behavior of sample averages. Parts a and c indicate that the Glivenko-Cantelli property may fail to hold even when $\mathcal{C}$ is “stable”.

The proof of the theorem makes use of the Rohklin-Kakutani Lemma. In Györfi *et al.* (1989 p.60), Shields used this lemma to show that density estimation with a general histogram method is not always consistent for stationary ergodic observations. Let $(\mathcal{X}, \mathcal{S}, P)$ be a non-atomic probability space and let the transformation $T : \mathcal{X} \to \mathcal{X}$ be measure-preserving, invertible, and ergodic. The following lemma is well known (*cf.* Petersen (1983)).

Lemma A (Rohklin-Kakutani) *For every* $\epsilon > 0$ *and every positive integer* n *there exists a measurable set* $A \subseteq \mathcal{X}$ *such that* $A, TA, \ldots, T^{n-1}A$ *are pairwise disjoint and satisfy* $P(\mathcal{X} \setminus \bigcup_{i=0}^{n-1} T^i A) < \epsilon$*.* □

The counterexample is based on repeated application application of this lemma. For $r = 2, 3, 4, \ldots$ let $A_r \in \mathcal{S}$ be such that

a. $A_r, TA_r, \ldots, T^{r-1}A_r$ are pairwise disjoint

b. $P(\mathcal{X} \setminus \bigcup_{i=0}^{r-1} T^i A_r) < 1/r$.

For each $r \geq 2$ define $C_r = \bigcup_{i=0}^{\lceil r/2 \rceil - 1} T^i A_r$, and note that $P(C_r) \leq 1/2$. Define the collection $\mathcal{C} = \{C_r : r = 2, 3, 4, \ldots\}$.

The process $\{Y_i\}$ is defined on $(\mathcal{X}, \mathcal{S}, P)$ in terms of the transformation T in the usual way:

$$Y_i(x) = T^i x \quad \text{for} \quad i = 0, 1, 2, \ldots.$$

It follows that $\{Y_i\}$ is stationary and ergodic, and that each random variable Y_k is distributed according to P. By definition, a sample sequence $Y_0(x), Y_1(x), Y_2(x), \ldots$ of $\{Y_i\}$ corresponds to a trajectory $x, Tx, T^2x, \ldots$, obtained by repeatedly applying the transformation T to a point $x \in \mathcal{X}$.

Proposition 1 *For each $n \geq 1$ and every $x \in \mathcal{X}$, $\Delta^{\mathcal{C}}(x, Tx, \ldots, T^{n-1}x) \leq n^2 + 2n$.*

Proof: Fix $n \geq 1$ and a point $x \in \mathcal{X}$. Consider the *trace* of the class $\mathcal{C}$ on an orbit of length n beginning at x:

$$\mathcal{T}_{\mathcal{C}}^n(x) = \{(I_C(x), I_C(Tx), \ldots, I_C(T^{n-1}x)) : C \in \mathcal{C}\}.$$

Note that $\mathcal{T}_{\mathcal{C}}^n(x)$ is a subset of $\{0,1\}^n$ and that $\Delta^{\mathcal{C}}(x, Tx, \ldots, T^{n-1}x) = |\mathcal{T}_{\mathcal{C}}^n(x)|$. Therefore, it is enough to show that $|\mathcal{T}_{\mathcal{C}}^n(x)|$ is bounded by a polynomial in n.

Every vector in $\{0,1\}^n$ consists of alternating blocks of 0's and 1's. Consider a vector $\mathbf{b}_r = (I_{C_r}(x), I_{C_r}(Tx), \ldots, I_{C_r}(T^{n-1}x))$ in $\mathcal{T}_{\mathcal{C}}^n(x)$. By design, every block of ones in this vector, with the possible exception of the last, is followed by a block of at least $r/2$ zeros. In particular, if $r \geq 2n$ the vector $\mathbf{b}_r$ can have at most one block of 1's; there are at most n^2 binary n-vectors of this form. Including those vectors $\mathbf{b}_r$ for $r = 2, 3, \ldots, 2n-1$, we find that $|\mathcal{T}_{\mathcal{C}}^n(x)| \leq n^2 + 2n$. As this bound is independent of x, the proof is complete. □

Proposition 2 *The sequence of random variables*

$$\sup_{\mathcal{C}} \left| \frac{1}{n} \sum_{i=0}^{n-1} I_C(T^i x) - P(C) \right| \qquad n = 0, 1, 2, \ldots$$

does not converge to zero in probability as $n \to \infty$.

Proof: Fix an integer n and consider the set A_{4n}. If $x \in \bigcup_{j=0}^{n-1} T^j A_{4n}$ then each of $x, Tx, \ldots, T^{n-1}x \in C_{4n}$ and consequently the average $\frac{1}{n}\sum_{i=0}^{n-1} I_{C_{4n}}(T^i x) = 1$. As $P(C_{4n}) \leq 1/2$, it follows that

$$\sup_{\mathcal{C}} \left| \frac{1}{n} \sum_{i=0}^{n-1} I_C(T^i x) - P(C) \right| \geq 1/2$$

for every $x \in \bigcup_{j=0}^{n-1} T^j A_{4n}$. By an easy calculation $P(A_{4n}) \geq (4n-1)/(4n)^2$, and consequently

$$P(\bigcup_{j=0}^{n-1} T^j A_{4n}) = \sum_{j=0}^{n-1} P(T^j A_{4n}) \geq \frac{4n-1}{16n},$$

which is greater than 1/8. This establishes our claim. □

Proof of Theorem 1: Parts a and b of the theorem follow immediately from Propositions 1 and 2. To establish part c, assume to the contrary that there exists a set $A \in \mathcal{S}$ with $P(A) > 0$ such that for almost every $x \in \mathcal{X}$ the class $\mathcal{C}$ shatters those points of the trajectory $x, Tx, T^2x, \ldots$ that lie in A. Let $\alpha = P(A)/2 > 0$, and for $n \geq 1$ define the event $A_n = \{x : \frac{1}{n}\sum_{i=0}^{n-1} I_A(T^i x) \geq \alpha\}$.

The ergodic theorem insures that $P(A_n) \to 1$ as $n \to \infty$. Moreover, by virtue of our assumption above,

$$\Delta^{\mathcal{C}}(x, Tx, \ldots, T^{n-1}x) \geq \exp_2\left(\sum_{i=0}^{n-1} I_A(T^i x)\right) \geq \exp_2(\alpha n)$$

for almost every x in A_n (here $\exp_2(a) = 2^a$). Thus, for n large enough

$$P\{x : \Delta^{\mathcal{C}}(x, Tx, \ldots, T^{n-1}x) \geq n^2 + 2n\} > 0.$$

This contradicts Proposition 1 and completes the proof. □

References

[1] E. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1, 1989.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Computing Machinery*, 36:929–965, 1989.

[3] L. Devroye. Automatic pattern recognition: a study of the probability of error. *IEEE Trans. on Pattern Anal. and Mach. Intelligence*, 10(4):530–543, July 1988.

[4] A. Faragó and G. Lugosi. Strong universal consistency of neural network classifiers. 1992. Preprint.

[5] L. Györfi, W. Härdle, P. Sarda, and P. Vieu. *Nonparametric Curve Estimation from Time Series.* Springer-Verlag, Berlin, 198.

[6] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[7] A.B. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Stat. and Prob. Letters*, 17:169–172, 1993.

[8] K. Petersen. *Ergodic Theory.* Cambridge University Press, 1983.

[9] G. Peškir and M. Weber. Necessary and sufficient conditions for the uniform law of large numbers in the stationary case. Technical Report 27, Institute of Mathematics, University of Aarhus, Denmark, 1992.

[10] G. Peškir and J.E. Yukich. Uniform ergodic theorems for dynamical systems under VC entropy conditions. In *Proceedings of the Ninth International Conference on Probability in Banach Spaces*, pages 104–127. Birkhauser, 1994.

[11] N. Sauer. On the density of families of sets. *J. Comb. Theory (A)*, 13:145–147, 1972.

[12] M. Talagrand. The Glivenko-Cantelli problem. *Ann. Probab.*, 15:837–870, 1987.

[13] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data.* Springer, New York, 1982.

[14] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

[15] B. Yu. Rates of convergence and central limit theorems for empirical processes of stationary mixing sequences. Ann. Probab., to appear, 1993.