IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 1–4

DOI: 10.1214/12-IMSCOLL901

A counterexample concerning the extension of uniform strong laws to ergodic processes

Terrence M. Adams[1] **and Andrew B. Nobel**[*,2]

Department of Defense and University of North Carolina, Chapel Hill

Abstract: We present a construction showing that a class of sets $\mathcal{C}$ that is Glivenko-Cantelli for an i.i.d. process need not be Glivenko-Cantelli for every stationary ergodic process with the same one dimensional marginal distribution. This result provides a counterpoint to recent work extending uniform strong laws to ergodic processes, and a recent characterization of universal Glivenko Cantelli classes.

1. Introduction and result

Let $\mathbf{X} = X_1, X_2, \ldots$ be an independent, identically distributed sequence of random variables defined on an underlying probability space $(\Omega, \mathcal{F}, P)$ and taking values in a measurable space $(\mathcal{X}, \mathcal{S})$. The strong law of large numbers ensures that, for every set $C \in \mathcal{S}$, the sample averages $n^{-1} \sum_{i=1}^n I_C(X_i)$ converge almost surely to $P(X \in C)$. A countable family $\mathcal{C} \subseteq \mathcal{S}$ is said to be a Glivenko-Cantelli class for $\mathbf{X}$ if the discrepancy

$$\Delta_n(\mathcal{C} : \mathbf{X}) \;=\; \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n I_C(X_i) - P(X \in C) \right|$$

tends to zero almost surely as n tends to infinity. In other words, $\mathcal{C}$ is a Glivenko-Cantelli class if the relative frequencies of sets in $\mathcal{C}$ converge uniformly to their limiting probabilities. The notion of a Glivenko-Cantelli class extends in an obvious way to uncountable families $\mathcal{C}$ under appropriate measurability conditions. For simplicity, we restrict our attention to countable classes $\mathcal{C}$ in what follows.

The discrepancy $\Delta_n(\mathcal{C} : \mathbf{X})$ plays an important role in the theory of machine learning and empirical processes (*cf.* [4, 5, 9, 10]). Necessary and sufficient conditions under which a family of sets $\mathcal{C}$ is a Glivenko-Cantelli class for an i.i.d. process $\mathbf{X}$ were first established by Vapnik and Chervonenkis [11], and later strengthened by Talagrand [8]. In both cases the conditions are combinatorial, and place limits on the ability of the family $\mathcal{C}$ to separate points in the trajectory of $\mathbf{X}$.

The ergodic theorem extends the classical strong law of large numbers to the larger family of ergodic processes, and it is natural to consider uniform laws of

*Work supported in part by NSF grant DMS-0907177.

[1]Department of Defense, 9800 Savage Rd. Suite 6513, Ft. Meade, MD 20755

[2]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260, e-mail: nobel@email.unc.edu, url: http://www.unc.edu/∼nobel/

AMS 2000 subject classifications: Primary 60F15; Secondary 60G10

Keywords and phrases: Glivenko-Cantelli, Uniform laws of large numbers, Ergodic process, Cutting and stacking

large numbers in the ergodic setting as well. A stationary process $\mathbf{X} = X_1, X_2, \ldots$ with values in $(\mathcal{X}, \mathcal{S})$ is ergodic if for each $k \geq 1$ and every $A, B \in \mathcal{S}^k$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P\big(X_1^k \in A, X_{i+1}^{i+k} \in B\big) = P\big(X_1^k \in A\big)\, P\big(X_1^k \in B\big),$$

where X_i^j denotes the tuple $(X_i, \ldots, X_j)$ when $i \leq j$. Let $\mathcal{C} \subseteq \mathcal{S}$ be a countable family of sets. Extending the definition above, we will say that $\mathcal{C}$ is Glivenko-Cantelli for $\mathbf{X}$ if $\Delta_n(\mathcal{C} : \mathbf{X}) \to 0$ with probability one as n tends to infinity.

Although necessary and sufficient conditions analogous to those of [8, 11] are not known in the general ergodic case, there has been some recent progress in regards to sufficiency and universality. Adams and Nobel [1] showed that if $\mathcal{C}$ has finite Vapnik-Chervonenkis (VC) dimension, then $\mathcal{C}$ is Glivenko-Cantelli for every ergodic process $\mathbf{X}$. Extensions to families of real-valued functions with finite VC-major, VC-graph, and fat-shattering dimensions can be found in [1, 2]. In a subsequent paper, Adams and Nobel [3] showed that classes of sets with finite VC dimension have finite bracketing numbers. In recent work, von Handel [12] has obtained generalizations of these results, and has established connections between universal Glivenko-Cantelli classes, covering numbers, and bracketing numbers. His principal result has the following immediate corollary: if $\mathcal{C}$ is Glivenko-Cantelli for every i.i.d. process, then $\mathcal{C}$ is Glivenko-Cantelli for every stationary ergodic process. In other words, the Glivenko-Cantelli property extends from the family of i.i.d. processes to the family of stationary ergodic processes.

In light of these results, it is natural to ask if the Glivenko-Cantelli property can be extended from a *single* i.i.d. process $\mathbf{X}$ to a related family of dependent processes. On the positive side, Nobel and Dembo [7] showed that if $\mathcal{C}$ is Glivenko-Cantelli for an i.i.d. process $\mathbf{X}$, then $\mathcal{C}$ is Glivenko-Cantelli for every beta-mixing (weakly Bernoulli) process $\mathbf{Y}$ with the same one dimensional marginal distribution, regardless of the mixing rate. In contrast with this result, we show below that a class $\mathcal{C}$ that is Glivenko-Cantelli for an i.i.d. process $\mathbf{X}$ need *not* be Glivenko-Cantelli for every stationary ergodic process with the same one dimensional marginal distribution as $\mathbf{X}$. In general then, extension of the Glivenko-Cantelli property from the i.i.d. setting to the ergodic one requires consideration of multiple marginal distributions.

In what follows, let $[0, 1)$ be the half-open unit interval, equipped with its Borel subsets $\mathcal{B}$ and Lebesgue measure λ.

Theorem 1.1. *There exist stationary processes* $\mathbf{X}$ *and* $\mathbf{Y}$ *with values in* $[0, 1)$ *and a countable family* $\mathcal{D}$ *of Borel subsets of* $[0, 1)$ *such that*

(a) $\mathbf{X}$ *is independent with* $X_i \sim \lambda$

(b) $\mathbf{Y}$ *is ergodic with* $Y_j \sim \lambda$

(c) $\Delta_n(\mathcal{D} : \mathbf{X}) \to 0$ *with probability* 1 *but* $\Delta_n(\mathcal{D} : \mathbf{Y}) \geq 1/2$ *with probability* 1 *for each* $n \geq 1$.

The ergodic process $\mathbf{Y}$ in the theorem is defined by the repeated application of a fixed, Lebesgue measure preserving transformation $T : [0, 1) \to [0, 1)$ known as the von Neumann-Kakutani adding machine. An iterative construction of T via the method of cutting and stacking is outlined below; a more detailed presentation may be found in Friedman [6]. The sets in $\mathcal{D}$ are unions of intervals used to construct T, and are chosen in such a way that they are mutually independent under Lebesgue measure. Arguments of Dudley [5] show that $\Delta_n(\mathcal{D} : \mathbf{X}) \to 0$ with probability one, while the construction of the sets in $\mathcal{D}$ ensures that $\Delta_n(\mathcal{D} : \mathbf{Y}) \geq 1/2$.

Proof. We define a transformation $T : [0, 1) \to [0, 1)$ in an iterative fashion using a sequence of ordered intervals, known as columns. The intervals in each column can be viewed as a stack, with each interval lying directly below the interval to its right. Let the initial column $C_0 = \{[0, 1)\}$. For $k \geq 0$ define a new column C_{k+1} as follows. First, split each interval in C_k in half, creating two stacks with the same height, but half the width, of the stack defined by C_k. Then, place the right stack on top of the left. Thus, ordering intervals in the column from top to bottom, $C_1 = \{[0, 1/2), [1/2, 1)\}$ and $C_2 = \{[0, 1/4), [1/2, 3/4), [1/4, 1/2), [3/4, 1)\}$. In general C_k contains 2^k dyadic intervals of length 2^{-k}. The transformation T maps each point $x \in [0, 1)$ into the point directly above it in one of the columns C_k. The definition of the columns ensures that this assignment is consistent across columns, and that T is defined for every point in $[0, 1)$. It is easy to see that T is measurable, and that T preserves the (Lebesgue) measure of dyadic intervals. Thus T is measure preserving, and one may show in addition that T is ergodic (see Friedman [6] for more details).

The sets $D_1, D_2, \ldots$ in $\mathcal{D}$ are constructed inductively from the intervals used to define T. Let $D_1 = [0, 1)$ and $k_1 = 1$. Suppose that for some $m \geq 2$ the set D_{m-1} has been defined as a union of intervals in the column $C_{k_{m-1}}$. Choose integers $l_m, r_m \geq 1$ such that $(2m)^{-1} \leq 2l_m\, 2^{-r_m} \leq m^{-1}$. Let $k_m = k_{m-1} + r_m$, and define D_m to be the top $2l_m\, 2^{k_{m-1}}$ intervals of C_{k_m}.

The definition of D_m ensures that $\lambda(D_m) = 2l_m \cdot 2^{-r_m}$ so that $(2m)^{-1} \leq \lambda(D_m) \leq m^{-1}$. The construction of the columns C_k ensures that the intervals defining D_{m-1} appear in a regular fashion among the intervals in the column C_{k_m}. (In particular, the intervals in D_{m-1} appear $2l_m$ times among the intervals defining D_m.) One may readily show that D_m is independent of D_{m-1}, and of $D_1, D_2, \ldots, D_{m-2}$ as well. Define $\mathcal{D} = \{D_m : m \geq 1\}$.

Let $\mathbf{X} = X_1, X_2, \ldots \in [0, 1)$ be any i.i.d. sequence with $X_i \sim \lambda$. Using the bounds on $\lambda(D_m)$ above, arguments like those in Proposition 7.1.6 of Dudley [5] show that, for every $0 < \epsilon < 1$,

$$\sum_{n \geq 2/\epsilon} \sum_{m \geq 1} P\left(\left| n^{-1} \sum_{i=1}^{n} I_{D_m}(X_i) - \lambda(D_m) \right| > \epsilon \right) < \infty.$$

It follows from the first Borel-Cantelli lemma that $\Delta_n(\mathcal{D} : \mathbf{X}) \to 0$ with probability one. Using independence of the sets D_m one may also show that the ϵ bracketing numbers of $\mathcal{D}$ are infinite if $\epsilon < 1/2$. See Dudley [5] for more details.

Define a (deterministic) process $\mathbf{Y} = Y_0, Y_1, \ldots$ on $([0, 1), \mathcal{B}, \lambda)$ by letting $Y_i(x) = T^i x$, where T^i denotes the i-fold composition of the transformation T with itself. As T is measure preserving and ergodic, the process $\mathbf{Y}$ is stationary and ergodic, and moreover $Y_i \sim \lambda$. For each $m \geq 2$, let D'_m contain the "bottom" $l_m\, 2^{k_{m-1}}$ intervals comprising D_m. The sets $D'_1, D'_2, \ldots$ are independent, and the lower bound on $\lambda(D_m)$ ensures that $\lambda(D'_m) \geq (4m)^{-1}$. Thus $\sum_{m \geq 2} \lambda(D'_m) = \infty$, and the second Borel-Cantelli lemma implies that $\lambda(\{D'_m \text{ i.o.}\}) = 1$. Fix $n \geq 1$ and let $x \in \{D'_m \text{ i.o.}\}$. Then there exists $m \geq 3$ such that $x \in D'_m$ and $l_m\, 2^{k_{m-1}} > n$. The definition of D'_m and T ensure that $T^j x \in D_m$ for $j = 1, \ldots, n$, and as $\lambda(D_m) < 1/2$ we find that $\Delta_n(\mathcal{D} : \mathbf{Y}) \geq 1/2$ at x. It follows that $\Delta_n(\mathcal{D} : \mathbf{Y}) \geq 1/2$ with probability one for each $n \geq 1$. □

Acknowledgements

The authors would like to acknowledge helpful conversations with Ramon van Handel.

References

[1] Adams, T. M. and Nobel, A. B. (2010). Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Ann. Probab.* **38** 1345–1367.

[2] Adams, T. M. and Nobel, A. B. (2010). The gap dimension and uniform laws of large numbers for ergodic processes. Preprint. arXiv:1007.2964v1

[3] Adams, T. M. and Nobel, A. B. (2012). Uniform approximation and bracketing properties of VC classes. *Bernoulli.* To appear.

[4] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York. MR1383093 (97d:68196)

[5] Dudley, R. M. (1999). *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge. MR1720712 (2000k:60050)

[6] Friedman, N. A. (1992). Replication and stacking in ergodic theory. *Amer. Math. Monthly* **99** 31–41.

[7] Nobel, A. B. and Dembo, A. (1993). A note on uniform laws of averages for dependent processes. *Statist. Probab. Lett.* **17** 169–172. MR1229933 (94e:60031)

[8] Talagrand, M. (1987). The Glivenko-Cantelli problem. *Ann. Probab.* **15** 837–870. MR0893902 (88h:60012)

[9] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York. MR1385671 (97g:60035)

[10] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory.* Second edition. Springer-Verlag, New York. MR1719582 (2001c:68110)

[11] Vapnik, V. N. and Chervonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280. MR0627861 (83d:60031)

[12] von Handel, R. (2012). The universal Glivenko-Cantelli property. *Probab. Theory Related Fields.* To appear.