Chapter 2
On the Uniform Convergence of the Frequencies of Occurrence of Events to Their Probabilities

Vladimir N. Vapnik and Alexey Ya. Chervonenkis

Abstract This chapter is a translation of Vapnik and Chervonenkis's pathbreaking note

> В. Н. Вапник, А. Я. Червоненкис, О равномерной сходимости частот появления событий к их вероятностям, Доклады Академии Наук СССР 181(4), 781–783 (1968)

essentially following the excellent translation

> V. N. Vapnik, A. Ya. Červonenkis, Uniform Convergence of Frequencies of Occurrence of Events to Their Probabilities, Soviet Mathematics Doklady 9(4), 915–918 (1968)

by Lisa Rosenblatt (the editors only corrected a few minor mistakes and in some places made the translation follow more closely the Russian original).

(Presented by Academician V. A. Trapeznikov, 6 October 1967)

2.1 Introduction

According to the classical theorem of Bernoulli, the frequency of occurrence of an event A converges (in probability, in a sequence of independent trials to the probability of this event). In some applications, however, it is necessary to draw conclusions about the probabilities of the events of an entire class S from one and the same sample. (In particular, this is necessary in the construction of learning algorithms.) Here it is important to find out whether the frequencies converge to the probabilities uniformly over the entire class of events S. More precisely, it is important to find out whether the probability that the maximal deviation of frequency from the corresponding probability over the class S exceeds a given small number approaches zero in an unbounded number of trials. It turns out that even in the simplest examples such uniform convergence may not take place. Therefore we would like to have a criterion by which we can decide whether there is such

B. Schölkopf et al. (eds.), *Empirical Inference*, DOI 10.1007/978-3-642-41136-6_2,

convergence or not. In this note we consider sufficient conditions for such uniform convergence which do not depend on the properties of the distribution but are related only to the internal properties of the class S, and we give a bound on the rate of convergence also not depending on the distribution, and finally we point out necessary and sufficient conditions for the uniform convergence of the frequencies to the probabilities over the class of events S.

2.2 Statement of the Problem

Let X be a set of elementary events on which a probability measure μ is defined. Let S be a collection of random events, i.e., of subsets of the space X measurable relative to the measure μ (the system S belongs to a Borel system but does not necessarily coincide with it).

Let $X^{(l)}$ denote the space of samples from X of length l. On the space $X^{(l)}$ we define the probability product measure by the condition $P(Y_1 \cdot Y_2 \cdot \ldots \cdot Y_l) = P(Y_1) \cdot P(Y_2) \cdot \ldots \cdot P(Y_l)$, where Y_i are measurable subsets of X. This formalises the fact that sampling is repeated, i.e., the elements are chosen independently with a fixed distribution.

For every sample $x_1, \ldots, x_l$ and an event A we can define the frequency $\nu_A^l = \nu_A(x_1, \ldots, x_l)$ of occurrence of the event A as equal to the ratio of the number n_A of those elements of the sample which belong to A to the overall length l of the sample:

$$\nu_A(x_1, \ldots, x_l) = n_A / l.$$

Bernoulli's theorem asserts that

$$\lim_{l \to \infty} P(|\nu_A^l - P_A| > \varepsilon) = 0.$$

We, however, will be interested in the maximal deviation of the frequency from the probability

$$\pi^{(l)} = \sup_{A \in S} \left| \nu_A^l - P_A \right|$$

over the class. The quantity $\pi^{(l)}$ is a point function on the space $X^{(l)}$.

We will assume that this function is measurable relative to the measure on $X^{(l)}$, i.e., that $\pi^{(l)}$ is a random variable. If $\pi^{(l)}$ approaches 0 in probability with unbounded increase of the sample size l, then we will say that the frequencies of the events $A \in S$ converge in probability to the probabilities of these events uniformly over the class S.

The theorems below are concerned with estimating the probability of the event

$$\pi^{(l)} \underset{l\to\infty}{\longrightarrow} 0$$

and finding out conditions when

$$P\left(\pi^{(l)} \underset{l\to\infty}{\longrightarrow} 0\right) = 1.$$

2.3 Some Additional Definitions

Let $X_r = x_1, \ldots, x_r$ be a finite sample of elements from X. Every set A from S determines a subsample $X_r^A = x_{i_1}, \ldots, x_{i_k}$ on this sample consisting of those elements of the sample X_r which are in A. We will say that the set A induces the subsample X_r^A on the sample X_r.

Denote the set of all distinct subsamples induced by the sets from S on the sample X_r by $S(x_1, \ldots, x_r)$. The number of distinct subsamples of the sample X_r induced by the sets from S (the number of elements of the set $S(x_1, \ldots, x_r)$) will be called the index of the system S relative to the sample X_r and denoted by $\Delta^S(x_1, \ldots, x_r)$.

Obviously it is always true that

$$\Delta^S(x_1, \ldots, x_r) \leq 2^r.$$

The function $m^S(r) = \max_{x_1,\ldots,x_r} \Delta^S(x_1, \ldots, x_r)$, where the maximum is taken over all samples of length r, is called the growth function of the class S.

Example 2.1. Let X be a straight line and S the set of all rays of the form $x < a$; $m^S(r) = r + 1$.

Example 2.2. X is the segment $[0, 1]$; S consists of all open sets; $m^S(r) = 2^r$.

Example 2.3. Let X be n-dimensional Euclidean space. The set of events S consists of all half-spaces of the form $(x\phi) > c$, where ϕ is a vector and c a constant; $m^S(r) < r^n$ $(r > n)$.

Along with the growth function $m^S(r)$ consider the function

$$M^S(r) = \int_{X^{(r)}} \ln \Delta^S(x_1, \ldots, x_r) d\mu(X^r);$$

$M^S(r)$ is the mathematical expectation of the logarithm of the index $\Delta^S(x_1, \ldots, x_r)$ of the system S.

2.4 A Property of the Growth Function

The main property of the growth function of the class S is established by the following theorem.

Theorem 2.1. *The growth function $m^S(r)$ is either identically equal to 2^r or majorized by the function r^n, where n is the first value of r for which $m^S(n) \neq 2^n$.*

2.5 Sufficient Conditions for Uniform Convergence Not Depending on Properties of the Distribution

Sufficient conditions for the uniform convergence (with probability 1) of the frequencies to the probabilities are established by the following theorem.

Theorem 2.2. *If $m^S(r) \leq r^n$, then*

$$P\left(\pi^{(l)} \underset{l\to\infty}{\longrightarrow} 0\right) = 1.$$

To prove this theorem, we establish the following lemma.

Take a sample $x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}$ of length $2l$ and compute the frequencies of occurrence of an event A on the first half-sample $x_1, \ldots, x_l$ and the second half-sample $x_{l+1}, \ldots, x_{2l}$. Denote the corresponding frequencies by ν'_A and ν''_A and consider $\rho_A^{(l)} = \left|\nu'_A - \nu''_A\right|$. We will be interested in the maximal deviation of $\rho^{(l)}{}_A$ over all events of S, i.e., $\rho^{(l)} = \sup_{A\in S} \rho_A^{(l)}$.

Lemma 2.1. *For each ε with $l > 2/\varepsilon^2$ we have the inequality*

$$P\left(\pi^{(l)} > \varepsilon\right) \leq 2P\left(\rho^{(l)} > \varepsilon/2\right).$$

We further establish for the proof of Theorem 2.2 that

$$P\left(\rho^{(l)} > \varepsilon/2\right) < 2m^S(2l)e^{-\varepsilon^2 l/16},$$

whence

$$P\left(\pi^{(l)} > \varepsilon\right) < 4m^S(2l)e^{-\varepsilon^2 l/16}. \qquad (*)$$

In the case where $m^S(r) < r^n$, the inequality (*) implies uniform convergence in probability. By a well-known lemma [1] from probability theory, we also establish convergence with probability 1 under the conditions of the theorem.

According to Theorem 2.2 there is uniform convergence in Examples 2.1 and 2.3 considered in Sect. 2.3. The fact that there is uniform convergence in Example 2.1 coincides with the assertion of Glivenko's theorem.

In many applications it is necessary to know the required sample size in order to assert with probability at least $1-\eta$ that the maximal deviation of the frequency from the probability over the class of events S does not exceed ε.

In the case where the growth function $m^S(l) \le l^n$ for the class S, the inequality (*) easily yields

$$l \ge \frac{32n}{\varepsilon^2}\left(\ln\frac{32n}{\varepsilon^2} - \ln\frac{\eta}{4}\right).$$

2.6 Necessary and Sufficient Conditions for the Uniform Convergence of Frequencies to Probabilities

Theorem 2.3. *For the uniform convergence (with probability 1) of the frequencies to the probabilities over the class of events S the condition*

$$\lim_{l\to\infty}\frac{M^S(l)}{l} = 0; \quad (M^S(l) = E(\ln \Delta^S(x_1,\ldots,x_l)))$$

is necessary and sufficient (here we assume the measurability of the function $\Delta^S(x_1,\ldots,x_l)$).

For the proof of Theorem 2.3 we consider a lemma.

Lemma 2.2. *The sequence $M^S(l)/l$ has a limit as $l\to\infty$.*

In the case where this limit is equal to 0, the sufficiency of the condition is proved analogously to Theorem 2.2. For the proof of necessity we first establish that

$$P(\pi^{(l)} > \varepsilon) > \frac{1}{2}P(\rho^{(l)} > 2\varepsilon).$$

We further establish that if $\lim_{l\to\infty} M^S(l)/l = t \ne 0$ then there is a δ such that

$$\lim_{l\to\infty} P\left(\rho^{(l)} > 2\delta\right) = 1,$$

whence $\lim_{l\to\infty} P\left(\pi^{(l)} > \delta\right) \ne 0$.

The theorem is proved.

V.N. Vapnik and A. Ya. Chervonenkis
Institute of Automation and Remote Control
(Technical Cybernetics)

Received
6 October 1967

Reference

1. Гнеденко, Б.В.: Курс теории вероятностей, 3rd edn. Fizmatgiz, Moscow (1961). English translation: Gnedenko, B.V.: A Course in Probability Theory, p. 212. Chelsea, New York, 1962. MR 25 #2622

http://www.springer.com/978-3-642-41135-9

Empirical Inference
Festschrift in Honor of Vladimir N. Vapnik
Schoelkopf, B.; Luo, Z.; Vovk, V. (Eds.)
2013, XIX, 287 p. 33 illus., 26 illus. in color., Hardcover
ISBN: 978-3-642-41135-9