ON THE UNIFORM CONVERGENCE OF FREQUENCIES OF THE APPEARANCE OF EVENTS TO THEIR PROBABILITIES

MATHEMATICS

1968

SovietRxiv

Abstract

Full Text

UDC 519.21

MATHEMATICS

V. N. VAPNIK, A. Ya. CHERVONENKIS

ON THE UNIFORM CONVERGENCE OF FREQUENCIES OF THE APPEARANCE OF EVENTS TO THEIR PROBABILITIES

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

§ 1. **Introduction.** According to Bernoulli' s classical theorem, the frequency of occurrence of a certain event A converges (in probability, in a sequence of independent trials, to the probability of this event). In a number of applications, however, it becomes necessary to judge the probabilities of events of an entire class S from one and the same sample. (In particular, this is necessary in constructing learning algorithms.) It is then important to ascertain whether the frequencies converge to the probabilities uniformly over the whole class of events S. More precisely, it is important to find out whether, as the number of trials grows without bound, the probability that the maximum over the class S of the deviation of the frequency from the corresponding probability exceeds a prescribed small quantity tends to zero. It turns out that even in the simplest examples such uniform convergence may fail to occur. Therefore one would like to have a criterion by which it would be possible to judge whether such convergence exists or not.

In the present note we consider sufficient conditions for such uniform convergence, conditions that do not depend on the properties of the distribution but are connected only with the internal properties of the class S; an estimate of the rate of convergence, also independent of the distribution, is given; and, finally, necessary and sufficient conditions for the uniform convergence of frequencies to probabilities over the class of events S are indicated.

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

Denote by $X^{(l)}$ the space of samples from X of length l. On the space $X^{(l)}$ a probability product measure is defined 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 formalizes the fact that the sample is repeated, i.e. the elements are chosen independently with an unchanged distribution.

For each sample $x_1, \ldots, x_l$ and event A there is defined the frequency $\nu_A^l = \nu_A(x_1, \ldots, x_l)$ of occurrence of the event A, equal to the ratio of the number n_A of those elements of the sample that belong to A, to the total length l of the sample:

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

Bernoulli' s theorem asserts that

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

We, however, shall be interested in the maximum, over the class, of the deviation of the frequency from the probability

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

The quantity $\pi^{(l)}$ is a function of a point in the space $X^{(l)}$.

We shall assume that this function is measurable with respect to the measure in

$X^{(l)}$, i.e., $\pi^{(l)}$ is a random variable. If $\pi^{(l)}$ tends in probability to zero as the sample size l increases without bound, then we shall say that the frequencies of the events $A_i \in S$ converge in probability to the probabilities of these events uniformly over the class S.

The subsequent theorems are devoted to estimating the probability of the event

$$\pi^{(l)} \xrightarrow[l\to\infty]{} 0$$

and to clarifying the conditions under which

$$P\left(\pi^{(l)} \xrightarrow[l\to\infty]{} 0\right) = 1.$$

§ 3. **Some additional definitions.** Let $X_r = x_1, \ldots, x_r$ be some finite sample of elements from X. Each set A from S determines on this sample a subsample

$X_r^A = x_{i_1}, \dots, \dots, x_{i_k}$ consisting of those terms of the sample X_r that belong to A. We shall 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 sets from S on the sample X_r by $S(x_1, \dots, x_r)$. The number of distinct subsamples of the sample X_r induced by sets from S (the number of elements of the set $S(x_1, \dots, x_r)$) will be called the index of the system S relative to the sample X_r and denoted by $\Delta^S(x_1, \dots, x_r)$.

Obviously, always

$$\Delta^S(x_1, \dots, x_r) \leqslant 2^r.$$

The function

$$m^S(r) = \max_{x_1, \dots, x_r} \Delta^S(x_1, \dots, x_r),$$

where the maximum is taken over all samples of length r, will be called the growth function of the class S.

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

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

Example 3. Let X be Euclidean space of dimension n. The set of events S consists of all half-spaces of the form $(x\varphi) > c$, where φ is a vector and c a constant; $m^S(r) < r^n$ $(r > n)$.

In addition to the growth function $m^S(r)$, consider the function

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

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

§ 4. **A property of the growth function.** The principal property of the growth function of the class S is established by the following theorem.

Theorem 1. *The growth function $m^S(r)$ is either identically equal to 2^r, or is majorized by the function r^n, where n is the first value of r for which $m^S(n) \neq 2^n$.*

§ 5. **Sufficient conditions for uniform convergence independent of properties of the distribution.** Sufficient conditions for the uniform convergence (with probability one) of frequencies to probabilities are established by the following theorem.

Theorem 2. *If $m^S(r) \leqslant r^n$, then*

$$P\left(\pi^{(l)} \xrightarrow[l\to\infty]{} 0\right) = 1.$$

To prove this theorem, the validity of the following lemma is established.

Let a sample of length $2l$ be taken: $x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}$, and let the frequencies of occurrence of the event A be counted on the first half-sample $x_1, \ldots, x_l$ and the second

subsample $x_{l+1}, \ldots, x_{2l}$. Denote the corresponding frequencies by ν'_A and ν''_A and consider $\rho_A^{(l)} = |\nu'_A - \nu''_A|$. We shall be interested in the maximal deviation $\rho_A^{(l)}$ over all events S, i.e. $\rho^{(l)} = \sup_{A\in S} \rho_A^{(l)}$.

Lemma 1. For any ε, when $l > 2/\varepsilon^2$, the inequality

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

holds.

Next, for the proof of Theorem 2 it is established that

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

whence

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

In the case where $m^S(r) < r^n$, the inequality (*) implies uniform convergence in probability. With the help of a well-known lemma from probability theory [(1)], it is established that under the hypotheses of the theorem convergence with probability one also holds.

According to Theorem 2, in examples 1 and 3 considered in § 3 there is uniform convergence. The fact that in example 1 uniform convergence exists coincides with the assertion of Glivenko' s theorem.

In many applications it is necessary to know what the sample size must be so that, with probability not less than $1 - \eta$, one may assert that the maximal deviation of the frequency from the probabilities over the class of events S will not exceed ε.

In the case where, for the class S, the growth function $m^S(l) \leq l^n$, from inequality (*) one easily obtains

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

§ 6. Necessary and sufficient conditions for the uniform convergence of frequencies to probabilities.

Theorem 3. For the uniform convergence (with probability one) of frequencies to probabilities over a class of events S, it is necessary and sufficient that the condition

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

be fulfilled

(where the measurability of the function $\Delta^S(x_1,\dots,x_l)$ is assumed).

For the proof of Theorem 3 a lemma is considered.

Lemma 2. The sequence $M^S(l)/l$ as $l\to\infty$ has a limit.

In the case where this limit is equal to zero, the sufficiency of the condition is proved analogously to Theorem 2. To prove necessity, it is first established that

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

It is then established that if

$$\lim_{l\to\infty}M^S(l)/l=t\neq 0,$$

then there is a δ such that

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

whence

$$\lim_{l\to\infty}P(\pi^{(l)}>\delta)\neq 0.$$

The theorem is proved.

Institute of Automatics and Telemechanics
(Technical Cybernetics)

Received
6 X 1967

References

[1] B. V. Gnedenko, *A Course in Probability Theory*, 3rd ed., Moscow, 1961, p. 212.

Note: Figure translations are in progress. See original paper for figures.