In hindsight:
Doklady Akademii Nauk SSSR, 181(4), 1968

Léon Bottou

This short contribution presents the first paper in which Vapnik and Chervonenkis describe the foundations of the Statistical Learning Theory [10]. The original paper was published in the *Doklady*, the Proceedings of the USSR Academy of Sciences, in 1968. An English translation was published the same year in *Soviet Mathematics*, a journal from the American Mathematical Society publishing translations of the mathematical section of the Doklady.[1] The importance of the work of Vapnik and Chervonenkis was noticed immediately. Dudley begins his 1969 review for *Mathematical Reviews* [3] with the simple sentence "*the following very interesting results are announced.*"

This concise paper is historically more interesting than the celebrated 1971 paper [11] because its three page limit forced its authors to reveal what they consider essential. Every word in this paper counts. In particular, the introduction explains that a uniform law of large numbers "*is necessary in the construction of learning algorithms.*" The mention of learning algorithms in 1968 seems to be an anachronism. In fact, learning machines were a popular subject in the sixties at the Institute of Automation and Remote Control in Moscow. The trend possibly started with the works of Aizerman and collaborators on pattern recognition [1] and the work of Fel'dbaum on dual control [4]. Tsypkin then wrote two monographs [7, 8] that clearly define machine learning as a topic for both research and engineering.

These early works on machine learning are supported by diverse mathematical arguments suggesting that learning takes place. The uniform convergence results introduced in the 1968 paper provide powerful tools to construct such arguments. In fact, in their following works [12, 9], Vapnik and Chervonenkis show how the uniform convergence concept splits such arguments in three clearly defined parts: the approximation properties of the model, the estimation properties of the induction principle, and the computational properties of the learning algorithm. Instead

Léon Bottou
Microsoft Research, Redmond, WA. e-mail: leon@bottou.org

[1] A reproduction of this English translation of the 1968 paper follows this brief introduction.

of simply establishing proofs for specific cases, the work of Vapnik and Chervonenkis reveals the structure of the space of all learning algorithms. This is a higher achievement in mathematics.

Whereas the law of large numbers tells how to estimate the probability of a single event, the uniform law of large numbers explains how to simultaneously estimate the probabilities of an infinite family of events. The passage from the simple law to the uniform law relies on a remarkable combinatorial result (theorem 1 in the 1968 paper). This result was given without proof, most likely because the paper would have exceeded the three page limit. The independent discovery of this combinatorial result is usually attributed to Vapnik and Chervonenkis [11], Sauer [5], or Shelah [6]. Although this cannot be established with certainty, several details suggest that the 1968 paper and its review by Dudley attracted the attention of eminent mathematicians and diffused into the work of their collaborators.[2] However, Sauer gives a better bound in his 1972 paper than Vapnik and Chervonenkis in their 1971 paper.[3]

The combinatorial result of the first theorem directly leads to the best know Vapnik-Chervonenkis theorem, namely, the distribution–independent sufficient condition for uniform convergence. A detailled sketch of the proof supports this second theorem. Although the paper mentions the connection with the Glivenko-Cantelli theorem [2], the paper does not spell out the notion of capacity, now known as the *Vapnik-Chervonenkis dimension*. However, the definition of the growth function is followed by its expression for three simple families of events, including the family of half-spaces associated with linear classifiers.

The third and final theorem states the distribution–dependent necessary and sufficient condition for uniform convergence. The paper provides a minimal proof sketch. The proof takes in fact seven pages in [11] and twenty-three pages in [9].

In conclusion, this concise paper deserves recognition because it contains the true beginnings of Statistical Learning Theory. The work is clearly motivated by the design of learning algorithms and its results have provided a new foundation for statistics in the computer age.

[2] Sauer motivates his work with a single sentence, “*P. Erdös transmitted to me in Nice the following question: is it true that (insert result here)*”, without attributing the conjecture to anyone. Sauer kindly replied to my questions with interesting details: “*When I proved that Lemma, I was very young, and have since moved my interest more towards model theoretic type questions. Erdös visited Calgary and told me at that occasion that this question had come up. But I do not remember the context in which he claimed that it did come up. I then produced a proof and submitted it as a paper. I did not know about that question before the visit by Erdös.*” and “*the only thing I can contribute is that, I believe Weiss in Israel, told me that Shelah had asked Perles to prove such a Lemma, which he did, and subsequently both forgot about it and Shelah then asked Perles again to prove that Lemma.*”.

[3] In fact, Sauer gives the optimal bound (Dudley, personal communication.)

References

1. Aizerman, M.A., Braverman, É..M., Rozonoér, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control **25**, 821–837 (1964)
2. Cantelli, F.P.: Sulla determinazione empirica della legi di probabilita. Giornale dell'InstitutoItaliano dgli Attuari **4** (1933)
3. Dudley, R.M.: Mathematical Reviews MR0231431 (37#6986) (1969)
4. Fel'dbaum, A.A.: Optimal Control Systems. Nauka, Moscow (1963). English translation: Academic Press, New York, 1965
5. Sauer, N.: On the density of families of sets. J. Combinatorial Theory, Ser. A **13**, 145–147 (1972)
6. Shelah, S.: A combinatorial problem; stability and order for models and theories in innitary languages. Pacic J. Math. **41**, 247–261 (1972)
7. Tsypkin, Y.: Adaptation and Learning in Automatic Systems. Nauka, Moscow (1969). English translation: Academic Press, New York, 1971
8. Tsypkin, Y.: Foundations of the Theory of Learning Systems. Nauka, Moscow (1970). English translation: Academic Press, New York, 1973
9. Vapnik, V.N.: Estimation of dependences based on empirical data. Springer Series in Statistics. Springer Verlag, Berlin, New York (1982)
10. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Proceedings of the USSR Academy of Sciences **181**(4), 781–783 (1968). English translation: Soviet Math. Dokl., 9:915-918, 1968
11. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16**(2), 264–281 (1971)
12. Vapnik, V.N., Chervonenkis, A.Y.: Theory of pattern recognition. Nauka (1974). German translation: Akademie–Verlag, Berlin, 1979

Dokl. Akad. Nauk SSSR
Tom 181 (1968), No. 4

Soviet Math. Dokl.
Vol. 9 (1968), No. 4

UNIFORM CONVERGENCE OF FREQUENCIES OF OCCURRENCE OF EVENTS TO THEIR PROBABILITIES

UDC 519.21

V. N. VAPNIK AND A. Ja. ČERVONENKIS

1. Introduction. According to the classical theorem of Bernoulli, the frequency of occurrence of and event A converges (in probability, in a sequence of independent trials to the probability of this event). In many applications, however, it is necessary to estimate 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 know if the frequencies converge to the probabilities *uniformly on the entire class of events* S. More precisely, *it is important to know if the probability* that the maximal deviation of frequency over the class S from the corresponding probability 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 criteria by which we can decide if there is such 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 bounds for the rate of convergence also not depending on the distribution, and finally we find necessary and sufficient conditions for uniform convergence of the frequencies to the probabilities over the class of events S.

2. Statement of the problem. Let X be a set of elementary events on which the 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 sequences of length l of elements of X. On the space $X^{(l)}$ we define the probabilistic product measure from the condition $P(Y_1 \cdot Y_2 \cdot \cdots \cdot Y_l) = P(Y_1) \cdot P(Y_2) \cdot \cdots \cdots P(Y_l)$, 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 a fixed distribution.

For every sample $x_1, \cdots, x_l$ and an event A we define the frequency $\nu_A^l = \nu_A(x_1, \cdots, x_l)$ of occurrence of the event A, equal to the ratio of the number n_A of those elements of the sample which belong to A to the overall size 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 are interested in the maximal deviation of the frequency from the probability in the class

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

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

We 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 zero in probability with unbounded increase of the sample size l, then we say that the frequencies of the events $A_i \in S$ approach the probabilities of these events uniformly over the class S in probability.

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

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

and finding conditions when

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

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

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

Obviously we always have

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

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

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

Example 3. Let X be n-dimensional Euclidean space. The set of events S consists of those 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 $m^S(r)$ we consider the function

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

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

4. Nature of the growth function. \The basic nature 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 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 not depending on properties of the distribution. Sufficient conditions for uniform convergence (with probability one) of frequencies to probabilities are established by the following theorem.

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

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

To prove this theorem, we establish the following lemma.

Suppose we have taken a sample of size $2l$: $x_1, \cdots, x_l, x_{l+1}, \cdots, x_{2l}$ and computed the frequencies of occurrence of the event A on the first half-sample $x_1, \cdots, x_l$ and the second half-sample $x_{l+1}, \cdots, x_{2l}$. We denote the corresponding frequencies by ν'_A and ν''_A and consider $\rho_A^{(l)} = |\nu'_A - \nu''_A|$. We are interested in the maximal deviation of $\rho_A^{(l)}$ over all events of S, i.e., $\rho^{(l)} = \sup_{A \in S} \rho_A^{(l)}$.

Lemma 1. *For every ϵ with $l > 2/\epsilon^2$ we have the inequality*

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

We further establish for the proof of Theorem 2 that

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

whence

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

In the case in which $m^S(r) < r^n$, the inequality (*) implies uniform convergence in probability. Using a well-known lemma [1] in probability theory, we establish that under the hypotheses of the theorem we also have convergence with probability one.

According to Theorem 2 there is uniform convergence in Examples 1 and 3 of §3. The fact that there is uniform convergence in Example 1 coincides with the assertion of Glivenko's theorem.

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

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

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

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

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

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

(here we assume measurability of the function $\Delta^S(x_1, \cdots, x_l)$).

For the proof of Theorem 3 we consider a lemma.

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

In the case where this limit is equal to zero, sufficiency of the condition is proved analogously to Theorem 2. For the proof of necessity we first establish that

$$P(\pi^{(l)} > \varepsilon) > {}^1\!/_2 P(\rho^{(l)} > 2\varepsilon).$$

We further establish 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 Automation and Remote Control

Received 6/OCT/67

BIBLIOGRAPHY

[1] B. V. Gnedenko, *A course in probability theory*, 3rd ed., Fizmatgiz, Moscow, 1961, p. 212; English transl., Chelsea, New York, 1962. MR 25 #2622.

Translated by:
Lisa Rosenblatt