About the origins of the VC lemma

Léon Bottou

Goals

Preliminary

It is often said that the fundamental combinatorial lemma of the Vapnik-Chervonenkis theory was independently established by Vapnik and Chervonenkis (1971), Sauer (1972), Shelah (1972), and sometimes Perles and Shelah (to my knowledge, without reference).

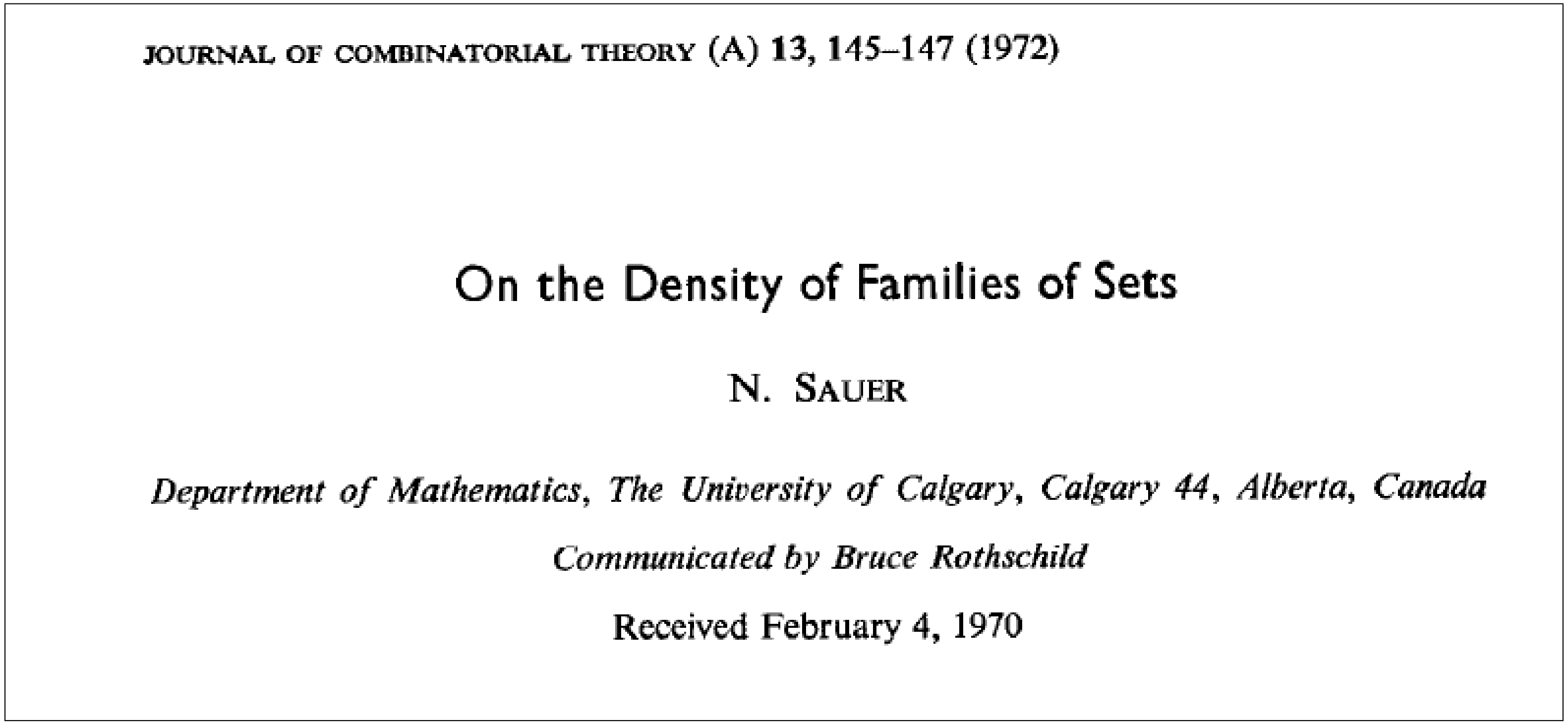
Questions

– Simultaneous discoveries sometimes occur.
– This happens when many teams work on the same problems.
– Learning theory was not then a common object of study.
– What can we find out about the origin of the lemma?

I. The Papers

Let's first focus on the documents.

Sauer 1972



JOURNAL OF COMBINATORIAL THEORY (A) **13**, 145–147 (1972)

On the Density of Families of Sets

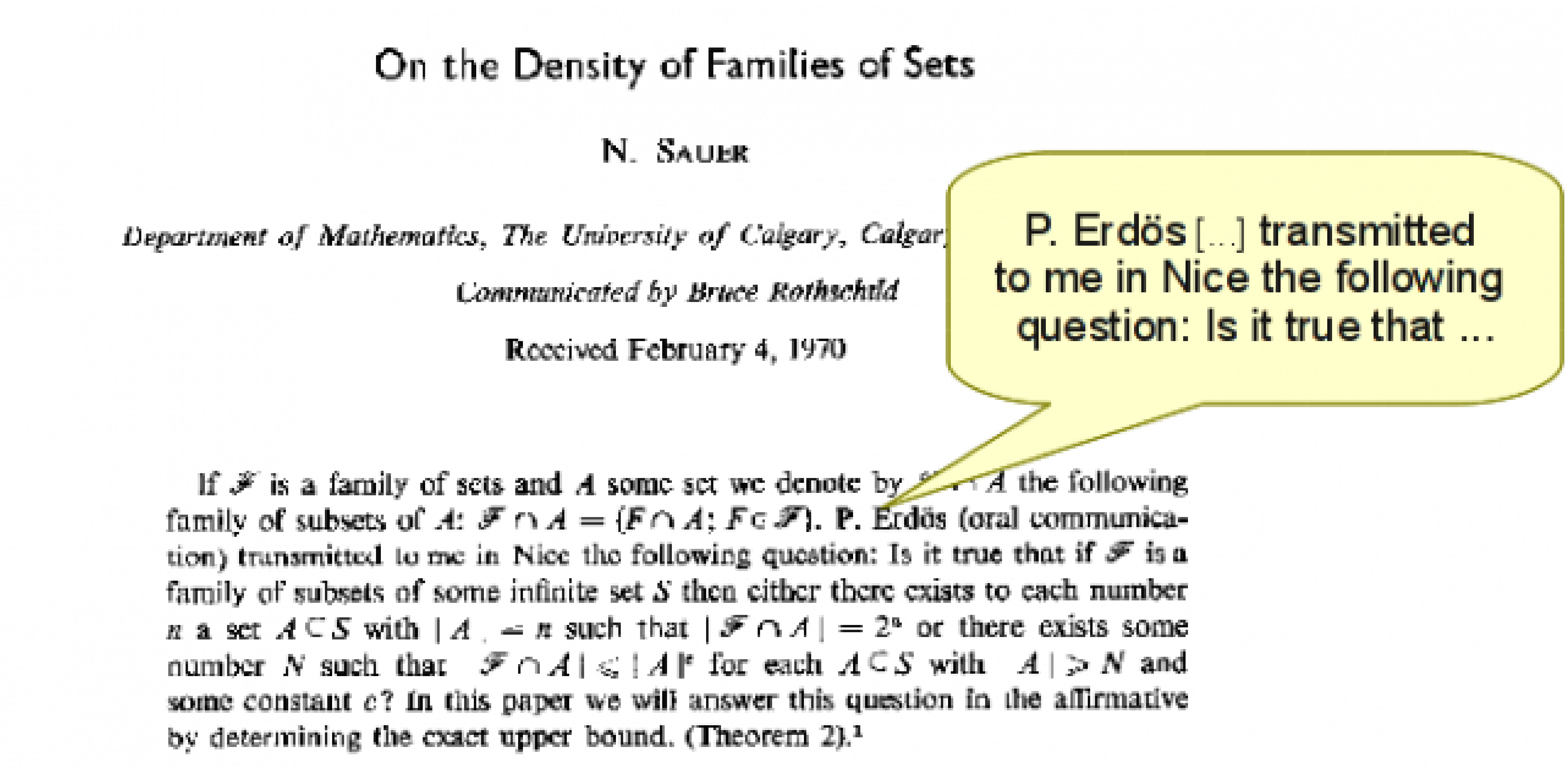
N. SAUER

Department of Mathematics, The University of Calgary, Calgary 44, Alberta, Canada

Communicated by Bruce Rothschild

Received February 4, 1970

Sauer 1972



On the Density of Families of Sets

N. Sauer

Department of Mathematics, The University of Calgary, Calgar

Communicated by Bruce Rothschild

Received February 4, 1970

If $\mathcal{F}$ is a family of sets and A some set we denote by $\mathcal{F} \cap A$ the following family of subsets of A: $\mathcal{F} \cap A = \{F \cap A; F \in \mathcal{F}\}$. P. Erdös (oral communication) transmitted to me in Nice the following question: Is it true that if $\mathcal{F}$ is a family of subsets of some infinite set S then either there exists to each number n a set $A \subset S$ with $|A| = n$ such that $|\mathcal{F} \cap A| = 2^n$ or there exists some number N such that $|\mathcal{F} \cap A| \leqslant |A|^c$ for each $A \subset S$ with $|A| \geqslant N$ and some constant c? In this paper we will answer this question in the affirmative by determining the exact upper bound. (Theorem 2).[1]

P. Erdös [...] transmitted to me in Nice the following question: Is it true that ...

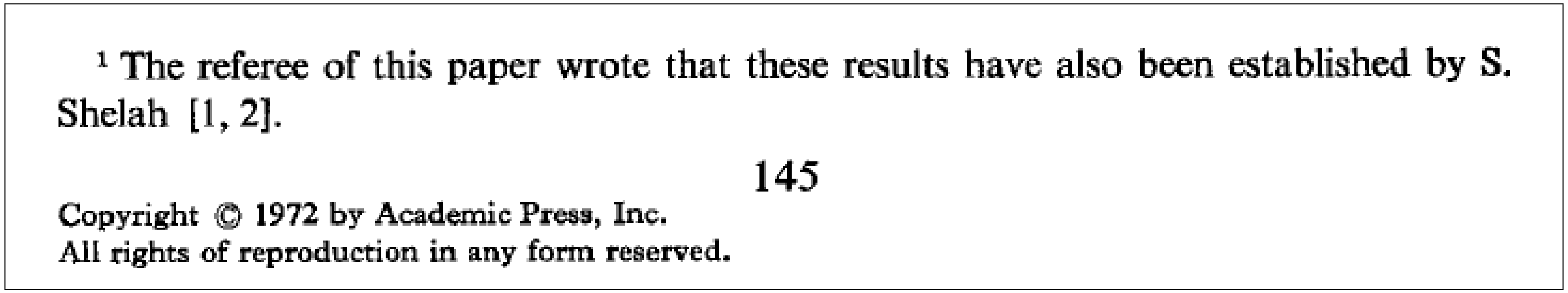
This is followed by a proof by induction.

Sauer 1972

"P. Erdös transmitted to me in Nice the following question. . . "

Remarks

- Sauer's proof is entirely motivated by Erdös question.
- Sauer does not attribute the conjecture to Erdös.
- Sauer did not know about Shelah's work either.



[1] The referee of this paper wrote that these results have also been established by S. Shelah [1, 2].

145

Sauer 1972

“P. Erdös transmitted to me in Nice the following question. . . ”

What about Nice?

– Every reader is expected to know what Nice represents.

Sauer 1972

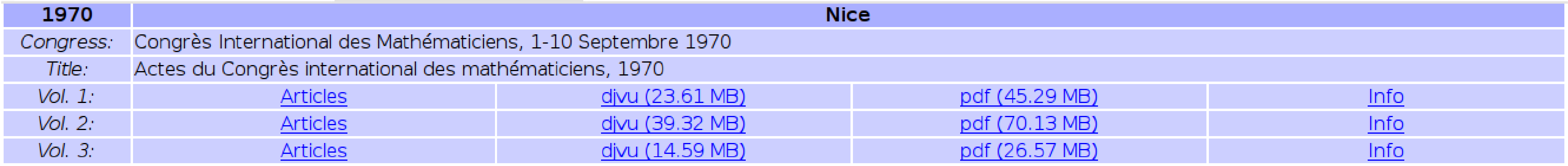
"P. Erdös transmitted to me in Nice the following question. . . "

What about Nice?

– Every reader is expected to know what Nice represents.

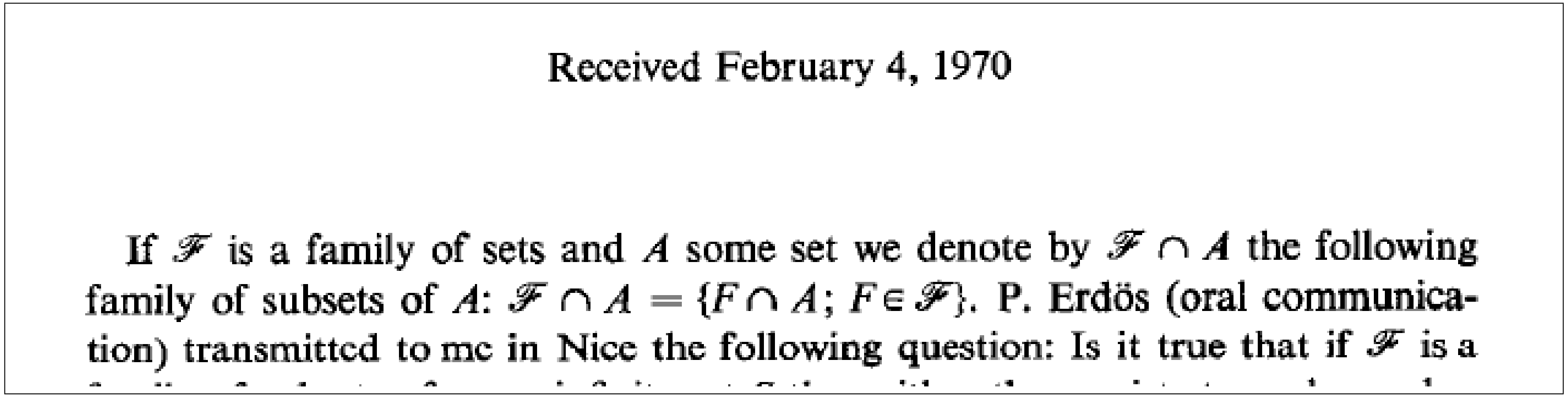
The International Congress Of Mathematicians
September 1-10, 1970, Nice.

Proceedings available on `http://www.mathunion.org/ICM/#1970`:



1970	**Nice**			
Congress:	Congrès International des Mathématiciens, 1-10 Septembre 1970			
Title:	Actes du Congrès international des mathématiciens, 1970			
Vol. 1:	Articles	djvu (23.61 MB)	pdf (45.29 MB)	Info
Vol. 2:	Articles	djvu (39.32 MB)	pdf (70.13 MB)	Info
Vol. 3:	Articles	djvu (14.59 MB)	pdf (26.57 MB)	Info

Sauer 1972



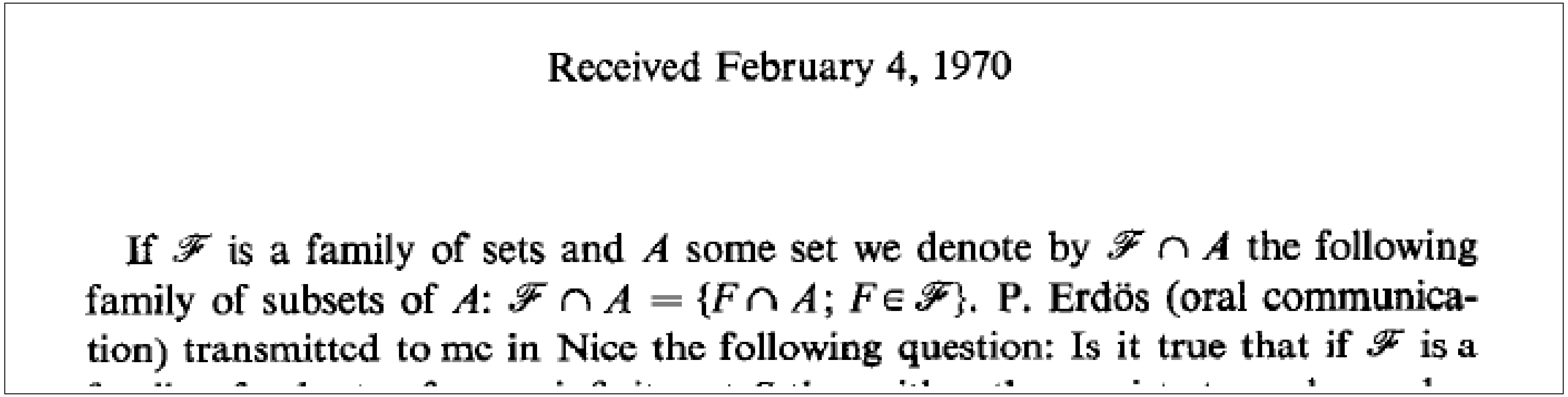
Received February 4, 1970

If $\mathscr{F}$ is a family of sets and A some set we denote by $\mathscr{F} \cap A$ the following family of subsets of A: $\mathscr{F} \cap A = \{F \cap A;\ F \in \mathscr{F}\}$. P. Erdös (oral communication) transmitted to me in Nice the following question: Is it true that if $\mathscr{F}$ is a

The International Congress Of Mathematicians
September 1-10, 1970, Nice.

Search the inconsistency

Sauer 1972



Received February 4, 1970

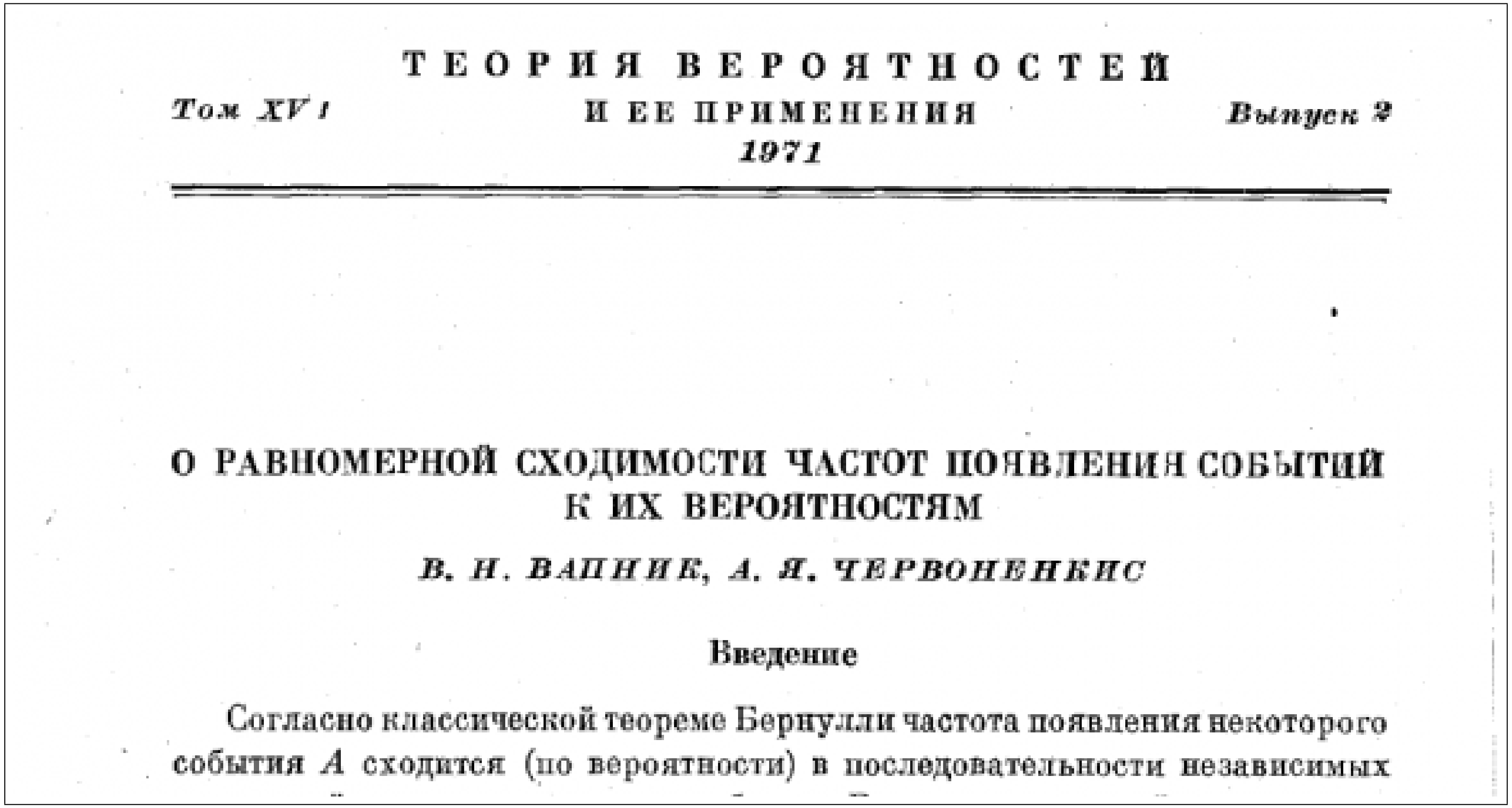
If $\mathscr{F}$ is a family of sets and A some set we denote by $\mathscr{F} \cap A$ the following family of subsets of A: $\mathscr{F} \cap A = \{F \cap A; F \in \mathscr{F}\}$. P. Erdös (oral communication) transmittcd to mc in Nice the following question: Is it true that if $\mathscr{F}$ is a

The International Congress Of Mathematicians
September 1-10, 1970, Nice.

Search the inconsistency

– Did Sauer and Erdös meet in Nice before the Nice congress?
– The motivation sentence was probably edited for the final version.

Vapnik and Chervonenkis 1971



ТЕОРИЯ ВЕРОЯТНОСТЕЙ
И ЕЕ ПРИМЕНЕНИЯ
1971

Том XVI *Выпуск 2*

О РАВНОМЕРНОЙ СХОДИМОСТИ ЧАСТОТ ПОЯВЛЕНИЯ СОБЫТИЙ К ИХ ВЕРОЯТНОСТЯМ

В. Н. ВАПНИК, А. Я. ЧЕРВОНЕНКИС

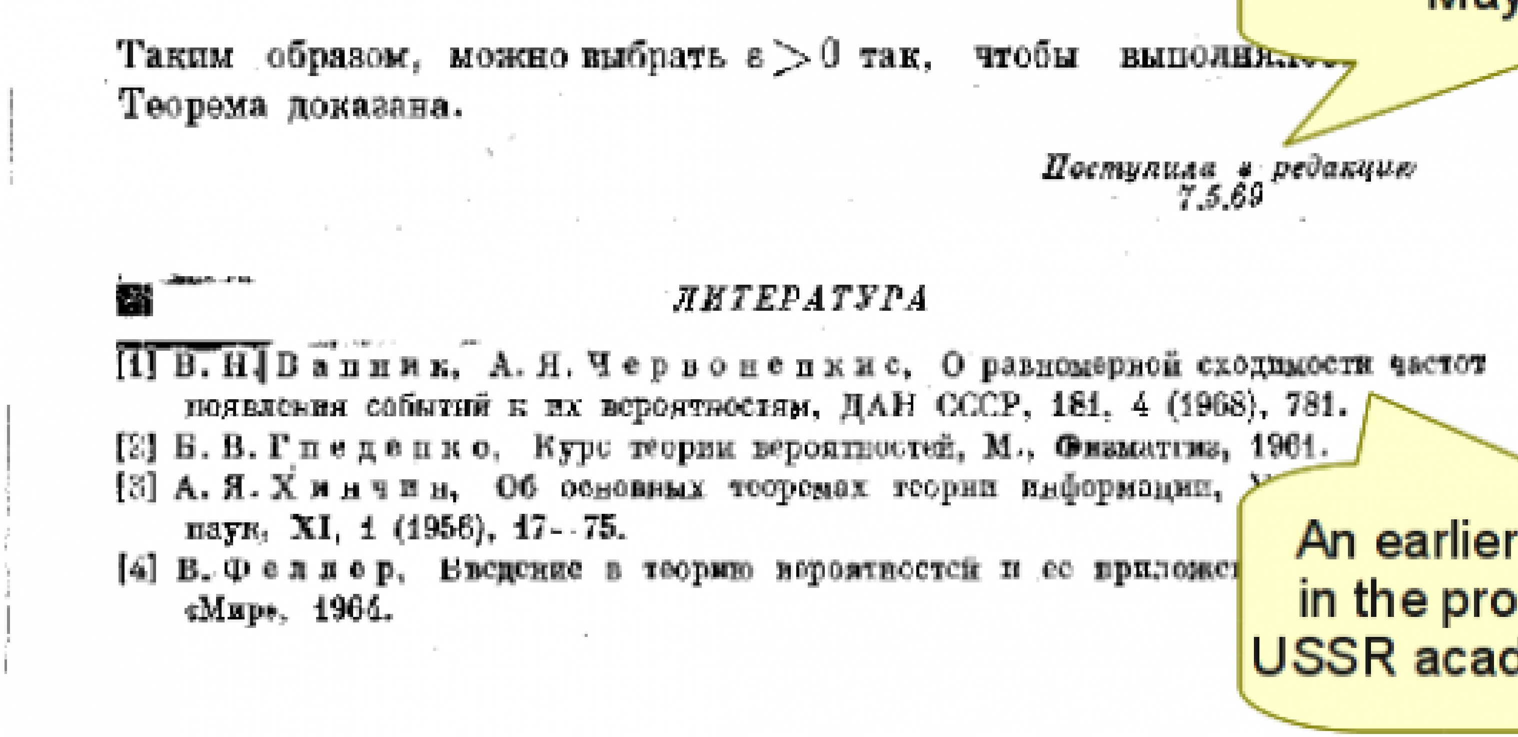
Введение

Согласно классической теореме Бернулли частота появления некоторого события A сходится (по вероятности) в последовательности независимых

Most of us know the english translation. . .

Vapnik and Chervonenkis 1971

but the russian version contains interesting details.



Таким образом, можно выбрать $\varepsilon > 0$ так, чтобы выполня…
Теорема доказана.

Поступила в редакцию
7.5.69

ЛИТЕРАТУРА

[1] В. Н. Вапник, А. Я. Червоненкис, О равномерной сходимости частот появления событий к их вероятностям, ДАН СССР, 181, 4 (1968), 781.

[2] Б. В. Гнеденко, Курс теории вероятностей, М., Физматгиз, 1961.

[3] А. Я. Хинчин, Об основных теоремах теории информации, … наук, XI, 1 (1956), 17--75.

[4] В. Феллер, Введение в теорию вероятностей и ее приложе… «Мир», 1964.

Submitted for review
May 7th, 1969

An earlier V-C publication
in the proceedings of the
USSR academy of sciences?

Vapnik and Chervonenkis 1968

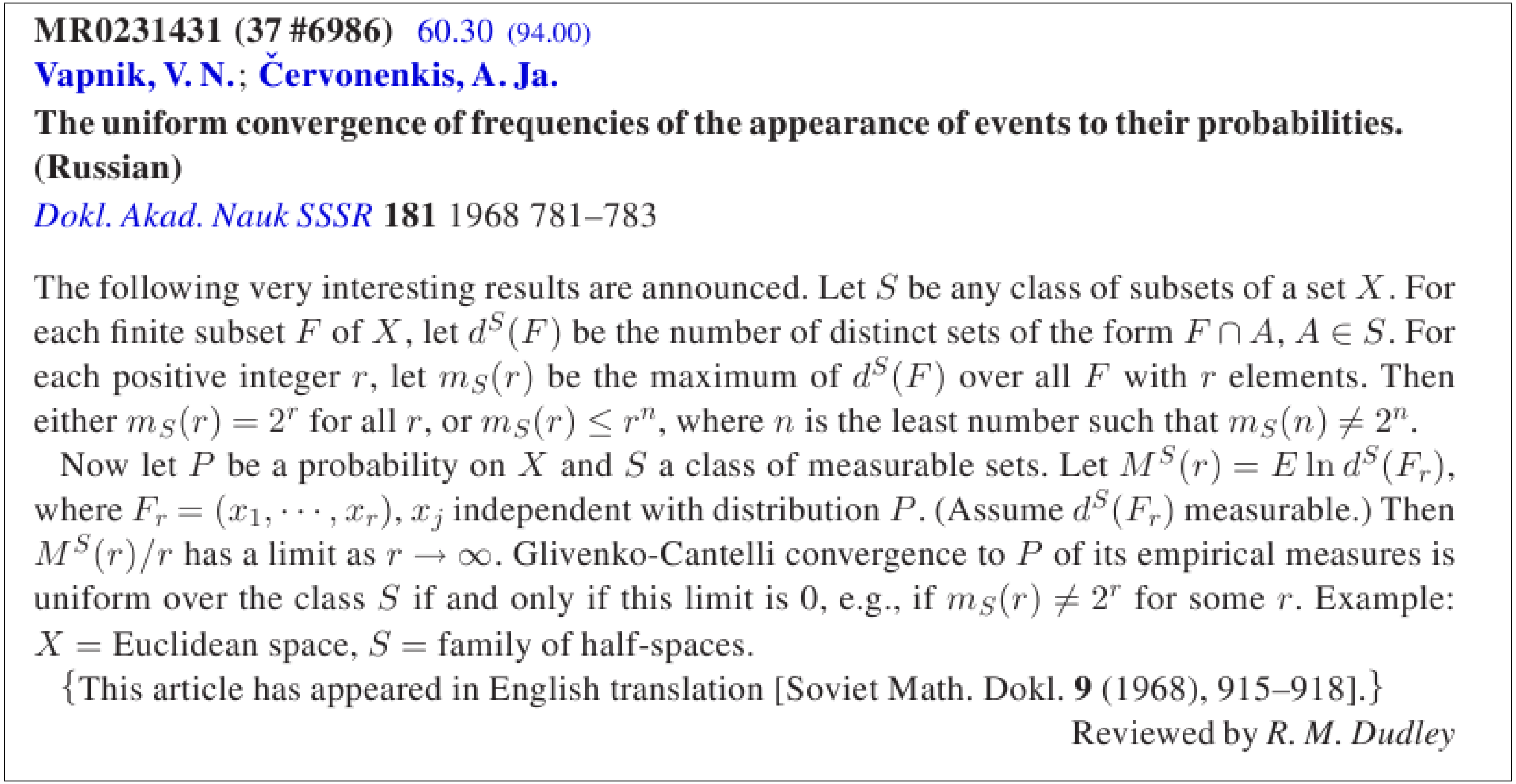
V. N. Vapnik and A. Ya. Chervonenkis:
Uniform convegence of the frequencies of occurence
of events to their probabilities.
Proceedings of the Academy of Sciences of the USSR, 181, 4(1968).

The American Mathematical Society used to publish translations of the Proceedings of the USSR Academy of Sciences.

This particular paper was translated in 1969 and was immediately noticed.

Vapnik and Chervonenkis 1968

Mathematical Reviews, 1969



MR0231431 (37 #6986) 60.30 (94.00)

Vapnik, V. N.; **Červonenkis, A. Ja.**

The uniform convergence of frequencies of the appearance of events to their probabilities. (Russian)

Dokl. Akad. Nauk SSSR **181** 1968 781–783

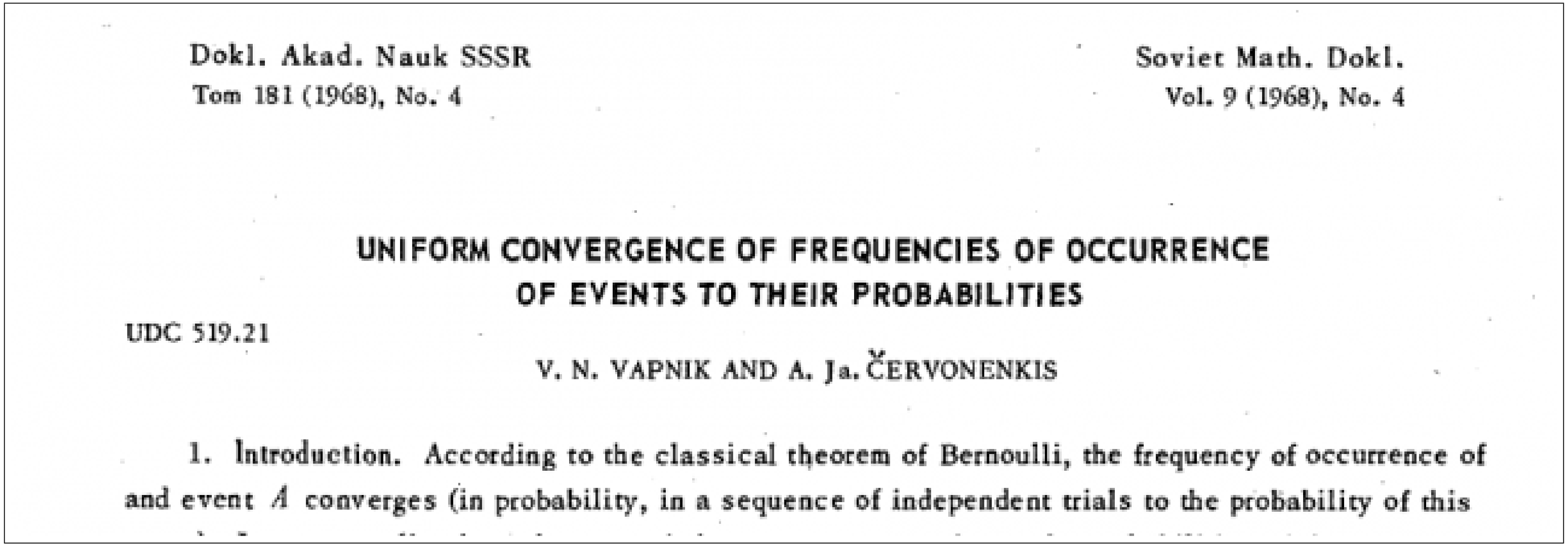
The following very interesting results are announced. Let S be any class of subsets of a set X. For each finite subset F of X, let $d^S(F)$ be the number of distinct sets of the form $F \cap A$, $A \in S$. For each positive integer r, let $m_S(r)$ be the maximum of $d^S(F)$ over all F with r elements. Then either $m_S(r) = 2^r$ for all r, or $m_S(r) \leq r^n$, where n is the least number such that $m_S(n) \neq 2^n$.

Now let P be a probability on X and S a class of measurable sets. Let $M^S(r) = E \ln d^S(F_r)$, where $F_r = (x_1, \cdots, x_r)$, x_j independent with distribution P. (Assume $d^S(F_r)$ measurable.) Then $M^S(r)/r$ has a limit as $r \to \infty$. Glivenko-Cantelli convergence to P of its empirical measures is uniform over the class S if and only if this limit is 0, e.g., if $m_S(r) \neq 2^r$ for some r. Example: X = Euclidean space, S = family of half-spaces.

{This article has appeared in English translation [Soviet Math. Dokl. **9** (1968), 915–918].}

Reviewed by *R. M. Dudley*

Vapnik and Chervonenkis 1968



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

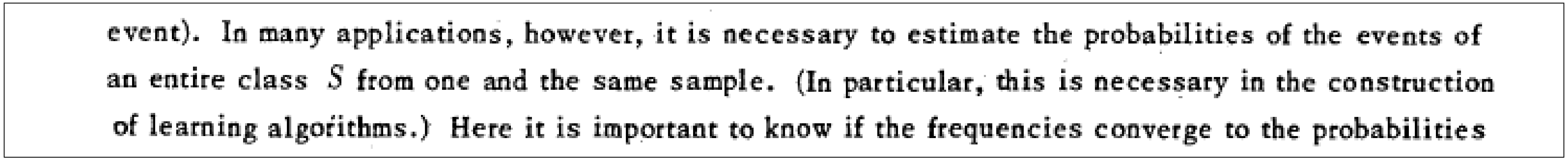
1. Introduction. According to the classical theorem of Bernoulli, the frequency of occurrence of and event A converges (in probability, in a sequence of independent trials to the probability of this

The paper is only four pages long.

This the clearest introduction to the VC theory I have read so far.

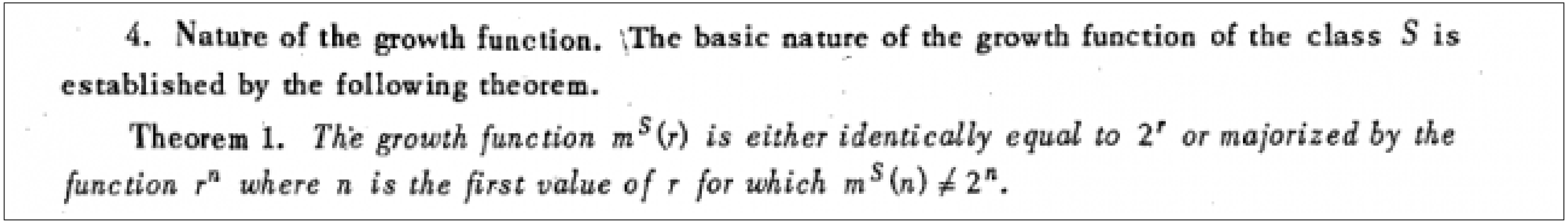
Vapnik and Chervonenkis 1968

Motivation



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

Theorem 1



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* $m^S(n) \neq 2^n$.

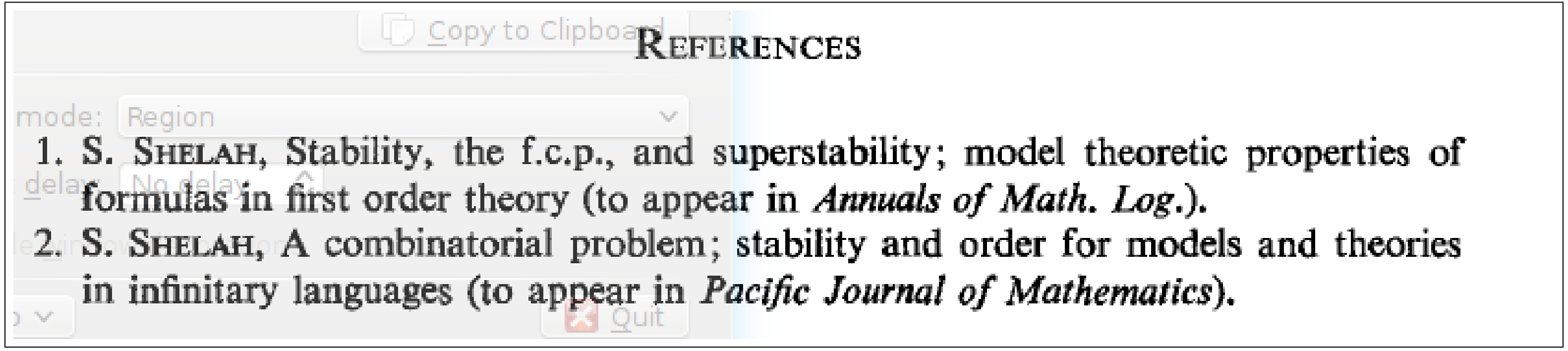
unfortunately given without proof.

Theorems 2 and 3

Give the distribution independent sufficient conditions
and the distribution dependent necessary and sufficient conditions
for uniform convergence. With short proofs.

Shelah's 1972 papers

Shelah's papers are cited by Sauer (as preprints).



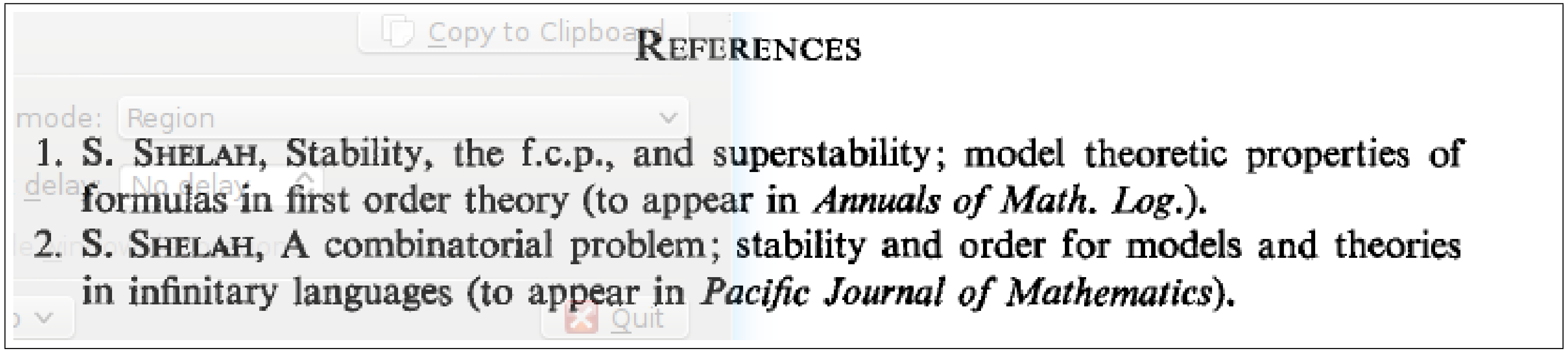
REFERENCES

1. S. SHELAH, Stability, the f.c.p., and superstability; model theoretic properties of formulas in first order theory (to appear in *Annuals of Math. Log.*).
2. S. SHELAH, A combinatorial problem; stability and order for models and theories in infinitary languages (to appear in *Pacific Journal of Mathematics*).

I found both papers but could not locate the lemma.

Shelah's 1972 papers

Shelah's papers are cited by Sauer (as preprints).



REFERENCES

1. S. SHELAH, Stability, the f.c.p., and superstability; model theoretic properties of formulas in first order theory (to appear in *Annuals of Math. Log.*).
2. S. SHELAH, A combinatorial problem; stability and order for models and theories in infinitary languages (to appear in *Pacific Journal of Mathematics*).

- "*. . . whereas it is easy to find the result in the paper of Vapnik-Chernovenkis, I would be hard put to give a precise location in Shelah's paper where he actually states this dichotomy.*" E Kowalsky[1]
- "*. . . could not see my way to it through the thicket of mathematical logic.*" M. J. Steele[2]

[1] http://blogs.ethz.ch/kowalski/2008/05/23/a-combinatorial-dichotomy

[2] http://www-stat.wharton.upenn.edu/~steele/Rants/ShatteredSets.html

Unresolved issues

The motivation

– The motivation of Vapnik and Chervonenkis was clear.
What were the motivations of Erdös, Shelah, Sauer ?

Erdös' question

– Erdös poses the question in september 1970.
How different was Sauer's paper in february 1970?
Did Erdös read the VC 1968 paper?
Why didn't he prove it himself?

II. Testimonies

I exchanged emails with a couple people. . .

Michael J. Steele (UPenn)

M. J. Steele coined the expression “shattered sets”.

- *“I learned the VC lemma from their 1971 paper. I mentioned this to Erdös in 1973 or 1974 and he told me about Sauer and Shelah. [...] Erdös definitely thought at that time that Sauer and Shelah were the first to answer his question [. . .] Incidentally, I think Erdos spoke more affectionately about Shelah than any other mathematician he ever mentioned to me.”*

- *“Lovasz is probably to ”blame” for the VC lemma becoming known as Sauer's Lemma — e.g. his Problems and Exercises in Combinatorics book.”*

Richard M. Dudley (MIT)

Richard M. Dudley wrote the review of the 1968 paper.

- *"I reviewed the 1968 announcement for Mathematical Reviews [. . .] I also reviewed for MR the 1971 paper with proofs, and their book on pattern recognition I think in 1974."*

- *"In my reading Vapnik and Cervonenkis did not have Sauer's Lemma but a weaker lemma of the same form where instead of the VC dimension S, they had $S + 1$. Even that, I saw in the 1971 paper and had not noticed it in the announcement. [...] By 1974 VC had the exact statement but that was after Sauer."*

Richard M. Dudley (MIT)

R. M. Dudley points out the following difference:

> When $n \geq h$, where h denotes the VC dimension of S,
>
> – Vapnik and Chervonenkis (1971) prove that $m_S(n) < \displaystyle\sum_{k=0}^{h} \binom{n}{k}$,
>
> – Sauer (1972) proves that $m_S(n) \leq \displaystyle\sum_{k=0}^{h-1} \binom{n}{k}$.

This difference only appears as a bound in the proof. Since both bounds imply a polynomial growth, this difference does not show in the statement of the 1968 paper, and it does not change the VC theory.

Richard M. Dudley (MIT)

The book *Theory of Pattern Recognition* (VC1974) contains many improvements to the 1971 proofs. In particular, the lemma is proven using the same bound as Sauer's paper.

- *"I don't know a reason for Vapnik or Chervonenkis to have been reading papers such as Sauer's, so I could believe the work was independent. In fact their 1974 reference list includes works not in Russian only from 1967 or earlier."*

- *"Sauer did not make other contributions to VC theory that I know of, but he did make this one. By the way Shelah is also mentioned for this lemma, but I could not find it in Shelah's paper, which deals with shattering for possibly infinite sets."*

Richard M. Dudley (MIT)

About the expression “VC-dimension”.

- *“I coined only the abbreviation ”VC”. I believe the first use of ”dimension” in relation to VC classes was in the title of a paper by P. Assouad, ”Densité et dimension” in 1983. Moreover he was concerned not only with the size of the largest shattered set, but with the behavior of the collection of sets more globally. I have never actually used ”VC dimension” in the current widely accepted sense, as in learning theory. Rather I talk about ”VC classes of sets,” ”VC index”, ”VC subgraph classes of functions,” etc. I don’t think Assouad used ”dimension” to mean the VC index either. So I don’t know where the usage began.”*

Norbert W. Sauer (U. Calgary)

- *“When I proved that Lemma, I was very young and have since moved my interest more towards model theoretic type questions. As far as I can remember, Erdos visited Calgary and told me at that occasion that this question has 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. I found the proof quite soon, a few weeks at most, after the visit by Erdös.*

Norbert W. Sauer (U. Calgary)

- *“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. There are many generalizations of that Lemma in many different directions.”*

My interpretation is that Shelah and Perles probably knew about such a result but did not considere it important enough to deserve a publication. Things were different for Sauer, who was certainly was very happy to have solved one of Erdös puzzles. . .

I found no trace of Perles proofs. What did he prove? When?
We could also ask when Vapnik and Chervonenkis first found the lemma.

The Eigenlemma

Vapnik and Chervonenkis offer a clear motivation for the lemma.
What was the motivation for Erdös, Shelah, Sauer, Perles ?

> Consider an hypothetical set S whose cardinality is smaller than the cardinality of $\mathbb{R}$. Let a *predicate* be a logic formula with a free variable. A collection of predicates partitions set S into classes of equivalence. Since such predicates can be numbered we can consider how the partition size grows with the number of predicates.
>
> – The partition size cannot grows like 2^n forever:
> otherwise we could build an injection from S to $\mathbb{R}$.
>
> – Therefore (lemma) the partition size grows polynomially.
> This means that a countable number of predicates
> cannot test whether set S is countable or larger.

The next step is to generalize this result to all formal statements one can express about the set in our logic system.
This quickly gets very complicated. . .

III. Conclusions

Conclusions

The publications

– Earliest publications of the lemma: Vapnik and Chervonenkis, 1968.

– Earliest proof: Vapnik and Chervonenkis, 1971.

– Improved proof: Sauer, 1972.

The motivations

– Vapnik and Chervonenkis motivation was learning theory.

– Shelah, Perles, and Erdös were probably seeking insights in the foundation of mathematics. Shelah's stability theory does not rely on the lemma in its simplest form.

– They probably had results of comparable nature (but when?) and did not consider them important enough be published.