# Chapter 1
# Chervonenkis's Recollections

**Alexey Chervonenkis**

**Abstract** These recollections about the origins of VC theory were written by Alexey Chervonenkis in 2004 for several colleagues and not intended for publication. They are now published for the first time. (Eds.) Translated by Vladimir Vovk.

The original setting of the problem (Autumn 1962) of pattern recognition learning was as follows. There are $N$ decision rules (ways of dividing objects into classes). The teacher is using one of them. A training sequence $x_1, \ldots, x_l$ is given, and the teacher classifies it naming for each point its class $\omega_1, \ldots, \omega_l$ using one of the $N$ known rules. The learning machine excludes from the list those rules that make an error, i.e., work *differently* from the teacher. There remain $N_1$ rules, and they are bound to contain the true one. (We called such algorithms *algorithms with complete memory*, as opposed to recurrent ones.)

The idea was to prove that there exists a training sequence such that $N_1$ becomes equal to 1, i.e., *only the true decision rule remains*, and, moreover, $l \sim \log N$. This scheme is reminiscent of searching for a counterfeit coin using a series of weighings (in which case $l \sim \log N$ is indeed sufficient).

An almost inverse statement is easy to show for recognition of binary vectors. If we want each of the $N$ decision rules to be chosen given some training sequence, the number $N$ must be at most the number of all variants of the training sequence of length $l$. For binary vectors of dimension $n$ the number of such variants is equal to $2^{(n+1)l}$. From this we get

A. Chervonenkis
Institute of Control Sciences, Laboratory 38, Profsoyuznaya Ulitsa 65,
Moscow 117997, Russia

A. Chervonenkis
Department of Computer Science, Royal Holloway, University of London,
Egham, Surrey, UK

A. Chervonenkis
Yandex, Moscow, Russia



V. Vovk et al. (eds.), *Measures of Complexity*,
DOI 10.1007/978-3-319-21852-6_1

$$N \leq 2^{(n+1)l},$$
$$l \geq \frac{\log_2 N}{n+1}.$$

Therefore the length of the sample must be at least[1] $\frac{\log_2 N}{n+1}$.

It turned out, however, that the opposite inequality

$$l \lesssim \ln N$$

in this setting is, in general, not correct.

For example, let there be $N-1$ objects, $N-1$ decision rules each of which assigns one of the objects to class I and the rest to class II, and the $N$th decision rule assigning all objects to class II. If the teacher is using the last rule, all the given objects will be assigned to class II and only at most $l$ decision rules will be discarded, and to discard all of them (except for the right one) everything has to be shown. That is, $l = N - 1$ rather than $\log N$.

Up to this point the problem did not involve *probability*. The indicated failure, and also other considerations, forced us to change the setting in March 1963.

The training sequence is not *given* but *generated* by some source $\Gamma$ independently with a constant, but unknown, distribution $P(x)$ (the i.i.d. hypothesis). On the other hand, we do not require that only one decision rule remains in reserve, but allow arbitrarily many provided they are *close* to the true one, i.e., make an error with probability $< \varkappa$ (under the same distribution as for training). Then it is easy to get a logarithmic estimate.

The probability that a rule that is different from the true one by more than $\varkappa$ will not be eliminated on a sample of length $l$ is less than

$$p = (1 - \varkappa)^l.$$

The probability that at least one such rule will not be eliminated is less than

$$N(1 - \varkappa)^l.$$

[1] In fact, in the setting of the problem as described here it is also true that

$$N \leq 2^l$$
$$l \geq \log_2 N$$

($\log_2$ standing for base 2 logarithm). Alexey's weaker (but sufficient for his purpose) bound $(\log_2 N)/(n+1)$ also holds in a situation that is easier for the learner: he knows the true decision rule, and his goal is to choose a training sequence $x_1, \ldots, x_l$ proving that the known decision rule is indeed the true one (in the sense that the observed $\omega_1, \ldots, \omega_l$ is compatible with only one rule). (Eds.)

Setting

$$N(1-\varkappa)^l = \eta$$

($\eta$ is a given small number $> 0$), we obtain

$$l \ln(1-\varkappa) + \ln N = \ln \eta,$$
$$l = \frac{\ln N - \ln \eta}{-\ln(1-\varkappa)} \approx \frac{\ln N - \ln \eta}{\varkappa}.$$

At that time we already knew that the number of ways to divide $K$ points by a hyperplane in an $n$-dimensional space is

$$\sim N = \frac{K^n}{n!}.$$

Since at that time people worked mainly with binary vectors, and there are only $2^n$ of them in $n$-dimensional space, then $K = 2^n$,

$$N \leq (2^n)^n = 2^{n^2},$$

which implies

$$\ln N \leq n^2,$$

and this can be regarded as acceptable.

We were very glad that for the first time one could justify theoretically a learning method of the same type as algorithms with complete memory. But in Autumn 1963 Aizerman talked about Novikoff's result that if a training sequence is rotated cyclically on a perceptron (1-layer, giving rise to a linear decision rule), then there can be at most $D^2/\rho^2$ errors overall, where $D$ is the diameter of the point set and $\rho$ is the distance between the convex hulls. From this they (the Aizermans[2]), after some tricks, managed to show that for a good performance on a test (exam) it is sufficient that

$$l \sim D^2/\rho^2.$$

Comparing with our result we can see that this *does not involve the dimension* and does not require discreteness. On the other hand, we do not require that classes should be separable by a wide band, i.e., $\rho$ can be arbitrarily small.

[2]Here Alexey jokingly refers to Aizerman, Braverman, and Rozonoer (members of Aizerman's laboratory at the Institute of Control Sciences) as the Aizermans. (Eds.)

Soon afterwards it became clear that the dimension $n$ and the value $D^2/\rho^2$ are in some respect interchangeable. Without additional assumptions it is impossible to get a good estimate without bounding either *dimension n* or $D^2/\rho^2$.

However, we also required the discreteness of the space (otherwise it is impossible to get a finite $N$) unlike Novikoff–Aizerman, and this appeared redundant, although in 1964–65 we did not manage to do anything about it. This was the setting of the problem: to obtain an estimate depending only on the dimension, but without the requirement of discreteness.

At that time new competitors appeared. Tsypkin started saying that all learning methods could be easily justified with the help of the method of "stochastic approximation" (1964–65), for which asymptotic convergence was proved (but without any rates). The Aizermans also were concerned only with convergence, not rates.

Tsypkin did not even want to listen to us when we said that something was proved in the discrete case. He used to say, "Spare me your talk of some finite set of decision rules, everything is proved a long time ago in the continuous case, for a continuum of decision rules."

As early as in 1962 *Highleyman's* work [1] appeared, where for the first time he considered the learning problem as minimization of *empirical risk*. But the justification of convergence was rather "wild." He wrote that since by the *Bernoulli* theorem for any decision rule the empirical risk converges to the true one (both are considered as a function of the decision rule, in his case of the coefficients of the hyperplane), the minimization of the empirical risk is asymptotically equivalent to (will lead to the same result as) the minimization of the true risk.

Ya.I. Khurgin and his Ph.D. student Loginov, following Highleyman's idea, went even further. Using Chebyshev's inequality (although using the binomial distribution would be more precise) they obtained absolute figures: in order to approach the true minimum with accuracy 10 % approximately 300 observations are sufficient; for 1 %, it appears that approximately 10,000 are sufficient. And this was without any restrictions whatsoever.

Since we were asked to review their paper, and Khurgin was Lerner's friend, there was a heated discussion between us in Summer 1965. We gave explicit examples where getting an acceptable result required an arbitrarily long training sequence (even for linear rules). It is here that *uniform convergence* was mentioned for the first time. Khurgin was saying, "You are playing on the non-compactness of the Hilbert ball," or "You are demanding uniform convergence." I was replying, "Yes, we are!"

Loginov was saying, "Don't you believe in the consistency of the method of empirical risk minimization for linear rules?" We were replying, "We believe but cannot prove it (without assuming discreteness)."

Khurgin was saying, "For a fixed decision rule the Bernoulli theorem (and the binomial distribution) is true. One can get a good estimate which is true for any decision rule. Therefore, it is true for all rules, and we are right." I was objecting, "The probability to meet randomly a syphilitic in Moscow is, say, $10^{-5}$. But if you went to a venereal clinic, it is significantly greater, even though it is also in Moscow. Looking for the best decision rule is like a trip to a venereal clinic."

In September 1965 at the All-Union Conference on Automatic Control[3] (it took place on the Black Sea aboard the ship "Admiral Nakhimov," which later sank) there was a flood of talks from the Aizermans, Tsypkin, Khurgin, et al., about learning algorithms that always converge without any restrictions (the dimension can be arbitrarily large, the distance between the classes arbitrarily small, etc.).

Prof. Kovalevskii from Kiev (later he was an examiner for my candidate thesis) said, "Why don't you stop them?" He was doing practical recognition: building a reading automaton.

Nevertheless we could not offer any alternatives.

Only in June 1966 I realized a thing very close to what you are doing now.[4] Given a sample $x_1, \ldots, x_l$, if we add $x_{l+1}$ to it, build a generalized portrait from $x_1, \ldots, x_l, x_{l+1}$, and then remove $x_{l+1}$, then the GP will change only in the case when $x_{l+1}$ was a *support* vector. But the probability that the last vector in the sequence will be a support vector is $k/(l+1)$, where $k$ is the number of support vectors. And in the general case this number does not exceed the dimension. If, on the other hand, $x_{l+1}$ is not a support vector, no error will be made on it when learning only on $x_1, \ldots, x_l$.

Therefore, the mathematical expectation of the number of errors made based on a sample of length $l$ (averaged over all samples of this length) will not exceed

$$\frac{n}{l+1}.$$

This is how, at last, an estimate appeared (although only for the case of GP) that depends only on the dimension and is not connected with discreteness.

Vapnik then suggested that we should not publish this result, because it is too simple, and it is embarrassing that we had "overlooked" it earlier. It was first published only in the book [4] in 1974.

After that, events developed quickly. It became clear that instead of the general population one can use an *exam*. In the simple case mentioned above the exam consisted of only one point. But the exam sample is usually sufficiently long, for example, as long as the training one. The following two experiments are equivalent:

1. We take a training sample of length $l$, learn on it, and are examined on a random sample of the same length.
2. We take straight away a random sample of length $2l$ and *randomly* divide it into two halves. Learn on the first and are examined on the second.

In the second case one can forget about the general population and assume that the world has narrowed down to this double sample. In the first case, on the contrary, we can use the *usual* Bernoulli theorem as applied to the exam (or the *usual* binomial distribution) and assume that the exam result is close to the true risk.

---

[3] The Third All-Union Conference on the Theory of Automatic Control, Odessa, September 20–26, 1965. (Eds.)

[4] Alexey means the method of conformal prediction; the first monograph [5] on the subject was being prepared by his colleagues at that time. (Eds.)

If we prove that in the second case *all is well*, then because of the equivalence all is well also in the first.

But the second case is discrete by definition. The sample itself gives a discrete set of points. But then one could use the *old* idea for finitely many decision rules. This way the *growth function* was born (on the same day) and an estimate was obtained for an arbitrary system of decision rules with a polynomial growth function. (Its being $2^l$ or a polynomial (VC dimension) was proven later in the course of proving necessary and sufficient conditions for uniform convergence, but that is another story.)

At the same time in July 1966 (published in 1968) we wrote and submitted for publication (in *Automation and Remote Control*) the paper "Algorithms with complete memory and recurrent algorithms in pattern recognition learning" [2], where we gave these results and compared them with what can be obtained for the perceptron.

And after that I remembered that Khurgin had been saying, "You demand uniform convergence." It became clear that the result for algorithms with complete memory could be turned into a proof of sufficient conditions of uniform convergence of frequencies to probabilities.

In a draft form this was done during the same month. But our inferiority complex did not allow us to stop here. We thought: what if they say, "You invented some growth function, and for the case when it grows as a polynomial proved that uniform convergence is present. But maybe it is always present, or present in a much wider range of cases?" It was necessary to show that it is present in this and only this case. And we also managed to do it in a draft form in July 1966.

At the time we did not know anything about the Glivenko(–Cantelli) theorem, and met it only when preparing the final version of [3] (towards the end of 1966 and in 1967).

## References

1. Highleyman, W.H.: Linear decision functions, with application to pattern recognition. Proc. IRE **50**, 1501–1514 (1962)
2. Vapnik, V.N., Chervonenkis, A.Y.: Algorithms with complete memory and recurrent algorithms in pattern recognition learning. Autom. Remote Control **29**, 606–616 (1968)
3. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of the frequencies of occurrence of events to their probabilities. Dokl. Akad. Nauk SSSR **181**, 781–783 (1968) (Sov. Math. Dokl. **9**, 915–918)
4. Vapnik, V.N., Chervonenkis, A.Y.: Теория распознавания образов: Статистические проблемы обучения (Theory of Pattern Recognition: Statistical Problems of Learning: in Russian). Nauka, Moscow (1974). German translation: Theorie der Zeichenerkennung, transl. K.G. Stöckel and B. Schneider, ed. S. Unger and B. Fritzsch, Akademie Verlag, Berlin (1979)
5. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)