Journal of Machine Learning Research 16 (2015) 2051-2066 Published 9/15

Alexey Chervonenkis's Bibliography: Introductory Comments

Alex Gammerman ALEX@CS.RHUL.AC.UK
Vladimir Vovk V.VOVK@RHUL.AC.UK
Computer Learning Research Centre, Department of Computer Science
Royal Holloway, University of London

This introduction to Alexey Chervonenkis's bibliography, which is published next in this issue, mainly consists of historical notes. The bibliography is doubtless incomplete, and it is just a first step in compiling more comprehensive ones. *En route* we also give some basic information about Alexey as a researcher and person; for further details, see, e.g., the short biography (Editors, 2015) in the Chervonenkis Festschrift. In this introduction, the numbers in square brackets refer to Chervonenkis's bibliography, and the author/year citations refer to the list of references at the end of this introduction.

Alexey Chervonenkis was born in Moscow in 1938. In 1955 he became a student at the MIPT, Moscow Institute of Physics and Technology (Faculty 1, Radio Engineering, nowadays Radio Engineering and Cybernetics). As part of his course of studies at the MIPT, he was attached to a laboratory at the ICS (the Institute of Control Sciences, called the Institute of Automation and Remote Control at the time), an institution in a huge system known as the Soviet Academy of Sciences.

In 1961 Alexey graduated from the MIPT and started his work for the ICS, where he stayed for the rest of his life. His first project at the ICS was very applied and devoted to designing a light organ for an exhibition in London (Russian Trade Fair, Earls Court, 1961). After completion of this project, Alexey was given an opportunity to concentrate on problems of cybernetics, namely pattern recognition; at that time cybernetics became extremely popular in the USSR, perhaps as a reaction to its earlier perception as a pseudo-science invented by the capitalist society and a "whore of imperialism" (Novoseltsev, 2015, p. 43).

In 1962 the joint work of Vapnik and Chervonenkis began. At that time they were members of the laboratory headed by Aleksandr Lerner, a leading cyberneticist. Lerner's laboratory was allowed to work on pattern recognition as a counterbalance to another laboratory, led by Mark Aizerman, which was the first to start work on this topic at the ICS: it was part of the strategy of Vadim Trapeznikov, the Institute director, to foster rivalry between different laboratories. Vapnik, a newly admitted PhD student, and Chervonenkis, hired a few months earlier as an engineer, were supposed to work as a pair. In hindsight, it appears that it was a perfect match; as Novoseltsev (2015) writes in his reminiscences, it is said that Vapnik was often inventing new things while Chervonenkis was proving them.

1. Foundations of Statistical Learning

Now Alexey Chervonenkis and Vladimir Vapnik are known, first of all, as the creators of the statistical theory of machine learning. However, their earliest joint work was devoted to non-statistical approaches to learning, as Alexey describes in [66]; it appears that this

work is not reflected at all in their joint publications. It was only in March 1963 that they brought statistics into their research.

When they started their joint work in Autumn 1962, they were interested in a problem that had more to do with the power of the teacher than the power of the learner. Suppose there are N decision rules, and one of them, say F, is used for classifying each point in a sequence $x_1, \dots, x_l$. The question is how small l can be so that there is only one decision rule (namely, F itself) compatible with the observed sequence $x_1, \dots, x_l$ and the classes $F(x_1), \dots, F(x_l)$. By choosing such a sequence the teacher can teach the learner to classify new points perfectly.

This problem is somewhat reminiscent of the problem of finding the counterfeit coin in a pile of N coins all but one of which are genuine. In the latter problem, we can take l of the order $\log N$, and the hope was that this remains true for the former. However, this is not the case. Consider N decision rules $F_1, \dots, F_N$ and $N-1$ points $x_1, \dots, x_{N-1}$ such that

$$F_i(x) := \begin{cases} 1 & \text{if } x = x_i \\ 0 & \text{if not,} \end{cases} \qquad i = 1, \dots, N-1,$$

and $F_N(x) = 0$ for all x. If F_N is used for classifying the points, seeing a new labelled point will allow the learner to discard at most one decision rule, and in order to discard all apart from the true one, we need $N-1$ observations. Therefore, the required value of l can be as large as $N-1$, which the two young researchers perceived as a failure of their non-statistical setting.

The statistical approach was first successfully used in [2] and [3]; the latter was prepared for a conference of young specialists in Spring 1963 (Editors, 2015). It was applicable to *algorithms with full memory*, introduced by Vapnik, Lyudmila Dronfort, and Chervonenkis in 1963, i.e., to algorithms that make no errors on the training set (the term "with full memory" was coined by Lyudmila). Suppose we are given two parameters, κ (the desired upper bound on the risk of the chosen decision rule) and η (the desired upper bound on the probability that the risk will in fact exceed κ). Let the true decision rule be F_i and the decision rule F_j chosen after l observations have a risk exceeding κ. The probability of getting all l labels right for F_j is at most $(1-\kappa)^l$, and the probability that at least one decision rule with risk exceeding κ will get all l labels right is at most $N(1-\kappa)^l$. To ensure that $N(1-\kappa)^l \le \eta$ it suffices to require

$$l \ge \frac{\log \eta - \log N}{\log(1-\kappa)}. \tag{1}$$

(This is Theorem 1 in [2].) As $-\log(1-\kappa) \ge \kappa$, the simpler inequality

$$l \ge \frac{\log N - \log \eta}{\kappa}$$

is also sufficient (and approximately equivalent to (1) for a small κ).

In might seem strange nowadays, but the presence of N in (1) was a real breakthrough. In [66] Alexey vividly describes discussions about the necessity of such an adjustment. The common reasoning among their colleagues (in related contexts) was that, since the probability of getting all labels right is at most $(1-\kappa)^l$ for **any** F_j, this is also true for the F_j

actually chosen by the algorithm, and so the fact that F_j is chosen *a posteriori* is irrelevant. Their colleagues, some of them very distinguished, could not be impressed by results like (1) believing that better results could be proved for a continuum of decision rules.

Alexey remembered a heated discussion with Yakov I. Khurgin (Lerner's friend and Professor of the Russian State University of Oil and Gas at the time) in Summer 1965. Khurgin's argument was, as usual, that a probabilistic statement that is true for all decision rules must be true for the one that was chosen by the algorithm. Alexey's counterargument was "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 the context of infinitely many decision rules (e.g., linear), Khurgin argued that Vapnik and Chervonenkis were playing on the non-compactness of the Hilbert ball and, crucially, that they were demanding uniform convergence. Alexey agreed. This was the first time that the words "uniform convergence" were mentioned in this context. Later they became part of the titles of the fundamental papers [10,11,28,57].

Paper [2] applied the general performance guarantee (1) to the problem of classification of binary vectors of dimension n using perceptrons. For simplicity, in this introduction we will only discuss binary decision rules, in which case a decision rule can be identified with a set in the input space, and perceptrons then become half-spaces. The authors bounded above the number of ways in which an $n-1$-dimensional hyperplane can split the n-dimensional Boolean cube by $2^{n(n+1)}/(n+1)!$, which gives the sample size

$$l \geq \frac{\log \eta - n(n+1) + \log((n+1)!)}{\log(1-\kappa)} \approx \frac{\log \eta - n^2}{\log(1-\kappa)}$$

for the Boolean input space with n attributes, where log now stands for binary logarithm.

In [5] Vapnik and Chervonenkis introduced a new learning framework, which they called "extremal imitation teaching". Suppose we observe a sequence of random pairs $(X(k), Y(k))$, $k = 1, 2, \ldots$, generated independently from the same distribution P and also observe, at each step k, the value $g(X(k), Y(k))$ of a "reward function" g. For each k, we are allowed to replace $Y(k)$ by our chosen value $Y^*(k)$, in which case we observe $g(X(k), Y^*(k))$ instead of $g(X(k), Y(k))$. This can be a model of, e.g., a chemical process at a chemical plant: $X(k)$ describes the kth batch of raw materials, $Y(k)$ describes the parameters of the process (such as temperature, pressure, or reagents) chosen by an experienced plant operator, and $g(X(k), Y(k))$ is the quality of the choice (assumed observable and determined by $X(k)$ and $Y(k)$ alone). It is supposed that g belongs to a known finite set Q of functions, say of size N, and that it takes values in a known interval $[a, b]$. The learning problem involves three positive parameters, ϵ, κ, and η, and is to find $l = l(N, \epsilon, \kappa, \eta)$ (as small as possible) such that after l steps we can come up with a strategy of choosing Y^* as a function of X that satisfies, with a probability at least $1 - \eta$ over the training set,

$$P\left(g(X, Y) > g(X, Y^*) + \epsilon\right) < \kappa,$$

where P is the probability over the random choice of (X, Y). Vapnik and Chervonenkis's proposed strategy is, in their terminology, sequential: first it observes $(X(k), Y(k))$, $k = 1, 2, \ldots$, and then it "trains", replacing $Y(k)$ by its own $Y^*(k)$ from some k on. The strategy

requires

$$l = O\left(\frac{\log N}{\epsilon\kappa}\left(\log\log N - \log\epsilon - \log\eta\right)\right),$$

where the O notation refers to $N \to \infty$, $\eta \to 0$, $\kappa \to 0$, and $\epsilon \to 0$; a and b are regarded as fixed constants. Namely, the strategy first makes

$$l_1 = O\left(\frac{\log N - \log\eta}{\epsilon\kappa}\right) \tag{2}$$

passive observations, and then it starts $d = O((\log N)/\epsilon)$ active training periods of length

$$l_0 = O\left(\frac{\log\log N - \log\epsilon - \log\eta}{\kappa}\right)$$

each; in each training period it tests a new strategy of choosing Y^* until it fails, in some sense (if it never fails during the l_0 steps, training is complete and the learning procedure is stopped).

They derive a very interesting corollary in the spirit of prediction with expert advice (Cesa-Bianchi and Lugosi, 2006). Suppose, in the language of our example, we can observe n experienced plant operators instead of just one. Observing each of the operators for l_1 steps (see (2)) and then training as before, our resulting strategy is likely to be competitive with the best operator at each step: with a probability at least $1 - \eta$ over the training set,

$$P\left(\max_{i=1,\dots,n} g(X, Y^i) > g(X, Y^*) + \epsilon\right) < \kappa,$$

where Y^i is the ith operator's output (being competitive at each step is unusual from the point of view of prediction with expert advice, where the goal is to be competitive in the sense of cumulative rewards or losses). Now passively observing takes nl_1 steps, whereas active training still takes $l_2 := dl_0$ steps.

In [6] Vapnik and Chervonenkis extended the methods of [5] to limited infinite classes of reward functions, and in [7] they applied them to the problem of playing an unknown zero-sum game.

Until Summer 1966 Vapnik and Chervonenkis could prove performance guarantees only for a finite number of decision rules. At the level of mathematical rigour that they set for themselves, they could not accept their colleagues' argument (see above) although they shared their optimism about, say, the learnability of the linear decision rules in Euclidean space. The first breakthrough came in July 1966 [66], when they extended their learnability results to classes of decision rules with a slow growth function (see below), and the second in September 1966, when they characterized the classes with slow growth functions in terms of what is now known as VC dimension. The main definitions (very well known by now) that we will need to talk about these developments are:

- Given a class S of decision rules on (i.e., subsets of) an input space X, let $\Delta^S(x_1, \dots, x_l)$ be the number of different restrictions of those decision rules to the finite set $\{x_1, \dots, x_l\}$ in X; Vapnik and Chervonenkis called $\Delta(x_1, \dots, x_l)$ the *index* of S with respect to $x_1, \dots, x_l$.

- They called the maximum

$$m^S(l) := \max_{x_1,\dots,x_l} \Delta^S(x_1,\dots,x_l)$$

  (as function of l) the *growth function* of S.

- The value

$$\mathrm{VC}(S) := \max\{l \mid m^S(l) = 2^l\}$$

  is now known as the *VC-dimension* of the class S. See Dudley (2015b, Section 4.6) for a discussion of the origin of the expression "VC-dimension"; the earliest mention of it seems to be in the article by Blumer et al. (1986) (in the form "Vapnik–Chervonenkis dimension", which was abbreviated to "VC dimension" in the journal version), whereas the abbreviation VC was coined by Dudley himself (Bottou, 2013).

The key fact about the growth function is now known as *Sauer's lemma*: if $\mathrm{VC}(S) = \infty$, $m^S(l) = 2^l$ for all l; otherwise,

$$m^S(l) \le \sum_{j=0}^{\mathrm{VC}(S)} \binom{l}{j} = O(l^{\mathrm{VC}(S)}) \tag{3}$$

for all l (the binomial coefficient is defined to be 0 when $j > l$). Therefore, we have the *Vapnik–Chervonenkis dichotomy*: the rate of growth of m^S is either exponential or at most polynomial. Sauer's lemma is more precise and also gives the degree of the polynomial ($\mathrm{VC}(S)$, which Sauer referred to as the density of S).

The main papers in which the 1966 breakthrough was announced and described were published in 1968 and 1971:

- In the first 1968 paper [9] Vapnik and Chervonenkis showed that the class S is learnable, in the now standard sense, if the rate of growth of m^S is polynomial (or slower).

- The second 1968 paper [10] is the famous announcement of their main results obtained during this period; it was "the true beginnings of Statistical Learning Theory" according to Bottou (2013).

- In [11] (1971), they gave detailed proofs.

- In their other 1971 paper [12] they explained in detail how the results of [11] can be applied to machine learning.

The results of [9] were obtained in July 1966, as Alexey describes in [66]. At that time Vapnik and Chervonenkis started to suspect that there are only two kinds of growth functions: exponential and (at most) polynomial. Vapnik said that even if this were true, it would be very difficult to prove it, but Chervonenkis presented a proof two months later (Novoseltsev, 2015).

A footnote in [9] says that after the paper had been submitted, the authors discovered that either $m^S(l) = 2^l$ for all l or $m^S(l) \le l^{\mathrm{VC}(S)+1}$ for all l. (This should have said "for all $l > 1$".) The date of submission is given as 20 September 1966.

Announcement [10] was published in *Doklady AN SSSR* (usually translated as *Proceedings of the USSR Academy of Sciences*). This journal only publishes papers presented by full and corresponding members of the Academy. The announcement stated the Vapnik–Chervonenkis dichotomy, in the form of the footnote in [9], as Theorem 1. It was first submitted for publication in 1966. The authors wanted Academician Andrei N. Kolmogorov to present their note, but submitted it directly to the journal, which forwarded it to Kolmogorov, who gave it to Boris V. Gnedenko to read. The authors did not hear from the journal for a long time, and a chain of enquiries led them to Gnedenko. Gnedenko explained to the young authors that what they were doing was not statistics; statistics was what Kolmogorov, Gnedenko himself, and their students were doing, and there was no chance that Gnedenko or his students would work in this new area. In the end the note was presented by the ICS Director Trapeznikov and submitted for publication on the same date, 6 October 1967; as compared to the original 1966 submission the authors only changed 2 lines in their manuscript: "Presented by Academician A. N. Kolmogorov" became "Presented by Academician V. A. Trapeznikov". The topic to which the note was assigned in the journal changed from "Probability theory" to "Cybernetics".

In [11], the authors still have Sauer's lemma with $\mathrm{VC}(S)+1$ instead of $\mathrm{VC}(S)$ (for the first time they will give the optimal exponent $\mathrm{VC}(S)$ in (3) in their book [18]). That paper was written in 1966, at the same time as their *Doklady* announcement [10], as it was customary for such announcements to be submitted together with a full paper, so that proofs of their statements could be checked. The key results of [11] were the VC dichotomy (Theorem 1), the uniform convergence of frequencies to probabilities over classes with polynomial growth functions (Theorem 3 and its small-sample counterpart Theorem 2), and an elegant necessary and sufficient condition for the uniform convergence of frequencies to probabilities in terms of the *entropy* $H^S(l) := \mathbb{E} \log \Delta^S(x_1, \dots, x_l)$ (Theorem 4).

In the four papers [9–12], Vapnik and Chervonenkis made a great leap forward in mathematical rigour. However, the required assumptions of measurability were very subtle, and even Vapnik and Chervonenkis did not get them quite right. After a modest description of his own mistakes in related measurability conditions, Dudley (2015a) points out that their requirement (in the penultimate paragraph of the Introduction to [11]) of

$$(x_1, \dots, x_l) \mapsto \sup_{A \in S} \left| \nu_A^{(l)}(x_1, \dots, x_l) - P(A) \right| \tag{4}$$

(where $\nu_A^{(l)}(x_1, \dots, x_l) := n_A/l$ and n_A is the frequency of A in the sample $x_1, \dots, x_l$) being measurable is not sufficient, as shown in the introduction to Dudley (1999, Chap. 5). The condition that is actually needed in the proof is that

$$(x_1, \dots, x_{2l}) \mapsto \sup_{A \in S} \left| \nu_A^{(l)}(x_1, \dots, x_l) - \nu_A^{(l)}(x_{l+1}, \dots, x_{2l}) \right|$$

be measurable.

The notion of growth function introduced in [9,10] was innovative but had had several interesting precursors, as described by Dudley (2015b). Already by 1852 Schläfli (1814–1895) found the growth function for the class S of all half-spaces in $\mathbb{R}^d$ containing 0 on their

boundary,

$$m^S(l) = 2\sum_{j=0}^{d-1} \binom{l-1}{j} = O(l^{d-1}) \qquad (l \to \infty). \tag{5}$$

Schläfli's memoir containing this result was published only in 1901 despite being written in 1850–1852. Among other fundamental achievements of this memoir were the introduction of d-dimensional Euclidean geometry (mathematicians had only treated the case $d \le 3$ before) and the extension of the ancient Greeks' result that there are only five platonic solids, i.e., convex regular polytopes in $\mathbb{R}^3$, to the case of $\mathbb{R}^d$ with $d > 3$ (it turned out that for $d > 4$ there are only three trivial platonic solids, the generalizations of the tetrahedron, cube, and octahedron, whereas for $d = 4$ there are six). Cover (1965) pointed out that, using Schläfli's method, one can obtain

$$m^S(l) = 2\sum_{j=0}^{d} \binom{l-1}{j} = O(l^d) \qquad (l \to \infty) \tag{6}$$

for the class S of all half-spaces in $\mathbb{R}^d$; he also obtained similar results for other classes, such as the parts of $\mathbb{R}^d$ bounded by hyperspheres or hypercones.

Richard Dudley wrote enthusiastic reviews of both [10] and [11] for *Mathematical Reviews*; interestingly, his review of [10] was instrumental in obtaining the permission to publish [11] (Bottou, 2012, with a reference to Vapnik). These reviews attracted attention of some leading mathematicians, and it seems likely that they were the means through which the VC dichotomy, in the form of a conjecture, reached the attention of Sauer and another independent discoverer, Shelah (together with his PhD student Perles).

The first statement of convergence of frequencies of events to their probabilities was James Bernoulli's (1713) celebrated law of large numbers, stating that, for all $\epsilon > 0$, events A, and probability measures P,

$$P^l\left(\left|\nu_A^{(l)} - P(A)\right| > \epsilon\right) \to 0 \tag{7}$$

as $l \to \infty$ (using the notation $\nu_A^{(l)}$ introduced in (4) and under unnecessary but mild restrictions on $P(A)$ and ϵ). Now we know that the convergence (7) is uniform in P, but Bernoulli did not know that, which might have been one of his reasons for not completing his manuscript (published in 1713 posthumously by his nephew): if the convergence is uniform, we can easily invert (7) to obtain a confidence interval for $P(A)$ given the observed frequency $\nu_A^{(l)}$. (This is one of the two reasons put forward by Hald 2003, p. 263; other authors have come up with more.)

Uspensky (1937) gives a "modernized" version of James Bernoulli's proof that does give a uniform convergence in (7) (cf. Hald 2003, p. 268). Nowadays, the most standard proof is based on Chebyshev's inequality and immediately gives uniform convergence:

$$P^l\left(\left|\nu_A^{(l)} - P(A)\right| > \epsilon\right) \le \frac{P(A)(1-P(A))}{l\epsilon^2} \le \frac{1}{4l\epsilon^2} \to 0.$$

(Although large-deviation inequalities, such as Hoeffding's, often give better results.)

Vapnik and Chervonenkis came up with a much deeper, and entirely different, statement of uniformity: for a fixed P,

$$P^l\left(\sup_{A\in S}\left|\nu_A^{(l)}-P(A)\right|>\epsilon\right)\to 0$$

for many interesting classes S of events; the requirement of uniformity was again motivated by statistical applications. If we require uniformity in both A and P,

$$\sup_P P^l\left(\sup_{A\in S}\left|\nu_A^{(l)}-P(A)\right|>\epsilon\right)\to 0$$

(i.e., that S be a *uniformly Glivenko–Cantelli class*), the condition $\mathrm{VC}(S)<\infty$ becomes both necessary and sufficient, for any $\epsilon\in(0,1)$; Vapnik and Chervonenkis understood this well already in 1966 (Editors, 2015).

In 1974 Vapnik and Chervonenkis published their book [18] in which they gave a survey of their work so far on the foundations of statistical learning (in Part II) and the method of generalized portrait (see the next section). They introduced a name for $\mathrm{VC}(S)$ (Chapter V, Section 7), namely the *capacity* of S (емкость S). The authors sent a copy of the book to *Mathematical Reviews*, requesting that it be sent to Richard Dudley to review. Whereas the papers [10] and [11] were reviewed quickly, in 1969 and 1972, respectively, reviewing [18] took five years, and Dudley's review appeared only in 1979. Dudley (2015b) explains this by the Peter principle: as reviewer, he was promoted to reviewing more and more difficult publications by Vapnik and Chervonenkis until his knowledge of the Russian language and pattern recognition became insufficient for the task.

In their book Vapnik and Chervonenkis gave Sauer's (1972) form of their dichotomy, which is obviously sharp in general: it suffices to take as S the class of all sets of cardinality $\mathrm{VC}(S)$ in an infinite input space X. For specific classes, however, even very important ones, the bound can be far from being sharp: e.g., for Schläfli's and Cover's cases Sauer's lemma only gives

$$m^S(l)\le\sum_{j=0}^{d}\binom{l}{j}=\Omega(l^d)\quad\text{and}\quad m^S(l)\le\sum_{j=0}^{d+1}\binom{l}{j}=\Omega(l^{d+1})$$

in place of (5) and (6), respectively.

An important contribution of the book [18], alongside with the papers [16,17], was the introduction of the method of Structural Risk Minimization (in Chapter VI) and its application to various specific problems. An appendix to Chapter VI (Section 14) gives lower bounds for the performance guarantees of learning algorithms in terms of the VC dimension.

In [28] Vapnik and Chervonenkis extended their necessary and sufficient condition of uniform convergence for classes of events (Theorem 4 of [11]) to classes of functions, defining the functional analogue of the entropy function H^S using Kolmogorov and Tikhomirov's ϵ-entropy. In [38] they found necessary and sufficient conditions for one-sided uniform convergence (Theorem Б of [38], where Б is the second letter of the Russian alphabet), which is particularly important from the viewpoint of machine learning because of its equivalence to the consistency of the method of empirical risk minimization (Theorem A of [38], where A is the first letter of the Russian alphabet).

Alexey's last great mathematical achievement [57,58] was the definitive quantitative form of Michel Talagrand's result about the existence of "bad sets" in machine learning (Talagrand, 1987, Theorem 5, and Talagrand, 1996, Theorem 2.1), which he first discovered just a few year's after Talagrand discovered his (Talagrand, 2014) but could prove rigorously only in the last years of his life (see [67], Theorem 13.1). Typically, he did this without any knowledge of Talagrand's work (Novoseltsev, 2015). Alexey's result says that if $H^S(l)/l \to c$ as $l \to \infty$ (the limit always exists), there exists a set $E \subseteq X$ of probability c such that for any n almost all sequences in E^n are shattered by S (a sequence $x_1, \ldots, x_n$ is *shattered* by S if $\Delta^S(x_1, \ldots, x_n) = 2^n$). Talagrand's result only asserts, for $c > 0$, the existence of E of positive probability satisfying the last condition. A precursor of this result was stated in [38] as Theorem B (B being the third letter of the Russian alphabet), and the result found its way into Vapnik (1998, Theorem 3.6).

Chervonenkis and Talagrand met in Paris in May 2011 and discussed the former's quantitative form of the latter's result (which Talagrand was really proud of but which, as he says, would not have been even conceivable without Chervonenkis's previous contributions). Chervonenkis asked Talagrand whether the quantitative form should be published. Talagrand replied that the quantitative form did not seem to have much use and so discouraged Chervonenkis from its publication (Talagrand, 2014).

2. Generalized Portrait and Optimal Separating Hyperplane

Vapnik and Chervonenkis's first joint paper [1] introduced the method of generalized portrait, which is a linear precursor of support vector machines, in the case of supervised learning. The idea of the method itself was first published by Vapnik and Lerner (1963) a year earlier, and Vapnik, Lerner, and Chervonenkis started discussing the method already in 1962 (see [61], which is an excellent source for the early history of support vector machines).

Vapnik and Lerner (1963) work in the context of unsupervised learning. The starting point of the early versions of the method of generalized portrait was that patterns were represented by points on the unit sphere in a Hilbert space. (Vapnik and Lerner consider a family of mappings from the patterns to the unit sphere, but let us, for simplicity, fix such a mapping, assume that it is a bijection, and identify patterns with the corresponding points of the unit sphere.) A set F of patterns *divides into n images* $F_1, \ldots, F_n$ (these are disjoint subsets of F) if for each F_i there is a point ϕ_i on the sphere such that, for all i, $j \neq i$, $f_i \in F_i$, and $f_j \in F_j$, it is true that $(\phi_i, f_i) > (\phi_i, f_j)$. Under a further restriction (the images should be "definite"), ϕ_i is called a *generalized portrait* for F_i. In their definition, Vapnik and Lerner do not specify a precise optimization problem with a unique solution that generalized portraits are required to solve. Later in the paper they do give two ideas for such optimization problems:

- In Section 4, they say that, for a given F_i, ϕ_i can be defined to maximize the *recognition threshold* $\min_{f \in F_i}(\phi_i, f)$. (This is the optimization problem that Alexey describes in the section devoted to Vapnik and Lerner's 1963 paper in his historical contribution to the Vapnik Festschrift: see [61], Section 3.1.1.) The overall optimization problem (to be solved before dividing F into images), however, remains unspecified.

- In the concluding Section 5 of their paper, Vapnik and Lerner make the problem of "self-learning" (unsupervised learning in this context) more precise by requiring that the generalized portraits $\phi_1, \dots, \phi_n$ maximize the *order of distinguishability* $1 - \max_{i,j}(\phi_i, \phi_j)$. This optimization problem will rarely determine generalized portraits completely: e.g., in the case of two images, $n = 2$, this condition only restricts ϕ_1 and ϕ_2 to being anti-collinear. And only rarely will any of its solutions maximize the recognition thresholds.

In [1] Vapnik and Chervonenkis made several important steps in the development of the method of generalized portrait; in particular, they defined it in the case of supervised learning and expressed it as a precise optimization problem. Suppose we are interested in a class K_1 of patterns and K_2 is the union of the other classes; these are assumed to be subsets of the unit sphere in a Hilbert space. The generalized portrait of K_1 is defined in this paper as the unit vector ϕ solving the optimization problem

$$\begin{aligned} (\phi, X) &\geq c, \quad \forall X \in K_1, \\ (\phi, Y) &\leq c, \quad \forall Y \in K_2, \\ c &\to \max . \end{aligned} \tag{8}$$

When the solution $(\phi, c) = (\phi_0, C(\phi_0))$ exists (i.e., when the class K_1 is linearly separable from the rest of data), it is unique, and the vectors $X \in K_1$ and $Y \in K_2$ satisfying $(\phi_0, X) = C(\phi_0)$ or $(\phi_0, Y) = C(\phi_0)$, respectively, were called the *marginal vectors*; these are precursors of support vectors. It was shown that the generalized portrait is a linear combination of marginal vectors (with nonnegative coefficients if they belong to K_1 and nonpositive if not).

Another contribution of [1] was that the method was rewritten in terms of scalar products between input vectors, which was an important step towards support vector machines. As it often happens, necessity was the mother of invention ([61], Section 3.3; Editors, 2015). At that time the ICS only had analogue computers, and inputting data was difficult. The easiest way was to calculate the scalar products by hand or using calculators, and then input them into the analogue computers by adjusting corresponding resistors. In 1964 the first digital computers arrived, and the dual form of the method lost much of its appeal for a few dozen years.

Vapnik and Chervonenkis kept the name "method of generalized portrait" in [1]. This might have been the first application of their decision (Novoseltsev, 2015) not to coin a new name for each new modification of their main recognition method; Vladimir proposed to use the same name for all modifications, the method of generalized portrait, and Alexey agreed. (There might have been one exception: it appears that in print the method of optimal separating hyperplane has not been explicitly referred to as that of "generalized portrait". In particular, the methods of generalized portrait and optimal separating hyperplane are treated as different ones in [61].)

As Alexey discusses in his historical paper [61] (Sections 3.1–3.2), already in 1962 he and Vladimir considered a more general version of the method, with $(\phi, Y) \leq kc$ in place of

$(\phi, Y) \le c$ in (8), for a given constant $k < 1$:

$$(\phi, X) \ge c, \quad \forall X \in K_1, \tag{9}$$

$$(\phi, Y) \le kc, \quad \forall Y \in K_2, \tag{10}$$

$$c \to \max. \tag{11}$$

In the same year they obtained the possibility of decomposition of the generalized portrait via marginal vectors directly, without the use of the Kuhn–Tucker theorem.

The generalization (9)–(11) was first published in [9], where the assumption that the training patterns should belong to a unit hypersphere is no longer mentioned. The authors retained the name "generalized portrait" for this more general setting. Using the Kuhn–Tucker theorem, they showed that the generalized portrait can be found by minimizing a quadratic function over the positive quadrant and developed several algorithms for solving such problems.

Further important developments were made in the 1973 papers [14,15] published in the same book edited by Vapnik and describing a library of computer programs written by Zhuravel' and Glazkova and implementing the method of generalized portrait (improved versions are described in [18], Chapter XV). In [14], Vapnik, Chervonenkis, and their co-authors consider the method of generalized portrait (9)–(11), whereas in [15] they consider a new method, that of optimal separating hyperplane. Given two linearly separable sets of vectors, X and $\bar{X}$ (the notation they use for K_1 and K_2 in this paper), they define the optimal separating hyperplane as the hyperplane that separates the two sets and is as far as possible from their convex closures. They notice that the optimal separating hyperplane can be represented by the equation $(\psi, x) = (c_1 + c_2)/2$, where ψ is the shortest vector satisfying $(\psi, z) \ge 1$ for all z of the form $z = x - \bar{x}$, $x \in X$ and $\bar{x} \in \bar{X}$, and

$$c_1 = \min_X (\psi, x), \quad c_2 = \max_{\bar{X}} (\psi, \bar{x}).$$

Together with the fact that ψ can be represented as a linear combination of margin vectors, this serves as the basis of their algorithm GP-4 for finding the optimal separating hyperplane.

The fundamental 1974 book [18] consists of three parts, one of which, Part III, is devoted to the methods of generalized portrait and optimal separating hyperplane (Part I is introductory and Part II is called "Statistical foundations of the theory"). In this part (Chapter XIV, Section 12) the authors derive another kind of performance guarantees for the two methods, which, as they say, are much closer to the lower bounds of Section VI.14 (already mentioned in Section 1 above) and so demonstrate special statistical properties of the method. A simple performance guarantee of this kind is that the (unconditional) probability of error does not exceed $m/(l+1)$, where l is the length of the training sequence and m is the expectation of the number of essential support vectors (which they called informative marginal vectors at the time). Since, in their context, m does not exceed the dimension n (assumed finite) of the input space, the probability of error is also bounded by $n/(l+1)$. This result was obtained by Alexey in June 1966 [66], but Vladimir was reluctant to publish it as it was embarrassingly simple. Let us call this type of error bounds *VC74 bounds* and the type of bounds discussed in the previous section *VC68 bounds* (following Vovk 2015). There were hints of VC74 bounds in [9], Section 5.3, and [14], pp. 91–92; however, the first

precise statements were first published only in the 1974 book [18]. It is interesting that, as Alexey says at the end of Section 3.6 of [61], VC74 bounds led to the notions of the growth function and VC dimension and to conditions for uniform convergence; it can be concluded that VC74 bounds led to VC68 bounds.

At the beginning of Chapter XIV the authors emphasize that in many cases the optimal separating hyperplane should be constructed not in the original input space but in a feature space (спрямляемое пространство). They only discuss finite-dimensional feature spaces, but since they already have the dual form of the optimization problem, there is only one step to support vector machines: to combine their algorithms with the idea of kernels that was already used by their competitors in Aizerman's laboratory (Aizerman et al., 1964); but this step had to wait for another 20 years.

The book [18] treats the methods of generalized portrait and optimal separating hyperplane more or less on equal footing, and studies relations between them, such as the latter being a special case of the former corresponding to a certain value of k. In the historical paper [61] Alexey mentions that in his and Vladimir's experience the number of support vectors for the optimal separating hyperplane often turned out to be larger than that for the generalized portrait for other values of k. His suggestion is to return to the method of generalized portrait (surely in combination with kernel methods—Eds.) looking for k providing the fewest number of support vectors. His intuition was that in the case of two approximately equal classes the method of optimal separating hyperplane is preferable. However, in the case where a small class is being separated from a much larger one (such as separating the letter "a" from the other letters of the English alphabet) the method of generalized portrait with a constant k close to 1 is preferable.

3. Other Publications

Approximately one half of Alexey's publications are devoted to applications of machine learning in various fields, such as natural language systems, geology, and medicine. This work was mainly done in collaboration with colleagues at the Institute of Control Sciences, the University of London, and Yandex.

In 1975–1983 Alexey and his colleagues at the ICS published a series of papers [19,21–27,29] describing their interactive data-retrieval system using a subset of the Russian language to control a large sea port. Alexey's main co-author was Leonid Mikulich, who also worked in Lerner's laboratory starting from 1961. In the course of numerous conversations between them Alexey proposed a formal logical calculus for describing non-trivial linguistic structures [19]. They also often discussed modelling evolution, and much later they were surprised to discover that it had become popular under the name of evolutionary and genetic programming.

Alexey's next significant area of applied research was geology [30–33,35,37,39–41,43,44]. This work included designing mathematical models for geological processes and non-parametric alternatives to the popular method of Kriging for restoring conditional distributions from empirical data. On the practical side, Alexey created a system for optimal automatic delineation of ore bodies that has been in operation at the world's largest gold deposit Murun-Tau since 1986 (Novoseltsev, 2015). For the creation of this system he was awarded the State Prize of the USSR (formerly Stalin Prize) in 1987.

Alexey's first work in medicine was done in 1971 [13] jointly with Vapnik and Lerner, but most of his papers in this area [51,53,56,59,60,68] were written together with his colleagues at Royal Holloway, University of London (whose Professor he formally became in 2000). A closely related application area in which Alexey was active is bioinformatics: see [45–47]. In the course of his work on bioinformatics he independently (albeit significantly later) rediscovered Watkins's (2000) and Haussler's (1999) string kernels. In general, independent rediscoveries were a typical feature of his research, arising naturally when a creative mind does not follow current literature preferring instead to invent new directions for itself at the risk of "discovering" well-known results and concepts. (A good example of this is Werner Heisenberg's rediscovery of matrix algebra in developing his approach to quantum mechanics.) Another independent rediscovery was his combination of Bayes and maximum likelihood methods for regression [42], which he later found in the work of David MacKay and finally [42,52] traced to a 1970 paper (Turchin et al., 1971).

Among Alexey's other applied papers were those devoted to energy load forecasting [49] and aircraft engineering [55]. One of Alexey's last applied research areas was the problem of optimal placement of advertisements among the results of a web search [63–65], which is of great interest to Yandex, the Russian analogue of Google, with which he was affiliated (alongside the ICS and Royal Holloway, University of London) since 2011.

From 2007 Alexey lectured at the School of Data Analysis founded by Yandex, and it is due to this activity that we owe his excellent textbook [52]. We are also lucky to have historical papers and notes published or prepared for publication during the last years of his life, namely the two sets of reminiscences in Vladimir Vapnik's and his own Festschriften [61,66] and his review [67] (preceded by the abstract [62]).

A lot remains unwritten or unfinished. Alexey was active, physically and mentally, and full of ideas until the very moment of his tragic death in the early hours of 22 September 2014 in Elk Island just outside Moscow.

Acknowledgments

The editors were greatly helped in their work by the late Alexey Chervonenkis, who generously shared his recollections with them. Many thanks to Vladimir Vapnik, Leonid Mikulich, and Michel Talagrand, who were invaluable sources of further information. Anatolii Mikhalsky's help is also gratefully appreciated.

References

Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

Jacob Bernoulli. *Ars Conjectandi*. Thurnisius, Basel, 1713.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred Warmuth. Classifying learnable geometric concepts with the Vapnik–Chervonenkis dimension. In *Proceedings of the 18th ACM Symposium on Theory of Computing*, pages 273–282, New York,

1986. ACM. Extended abstract. The full journal paper appeared as "Learnability and the Vapnik–Chervonenkis dimension" in *Journal of the Association for Computing Machinery*, 36:929–965, 1989.

Léon Bottou. On the Vapnik–Chervonenkis–Sauer lemma, 2012. URL `http://leon.bottou.org/news/vapnik-chervonenkis_sauer`. Accessed in September 2015.

Léon Bottou. In hindsight: Doklady Akademii Nauk SSSR, 181(4), 1968. In *Empirical Inference: A Festschrift in Honor of Vladimir N. Vapnik*, pages 13–20. Springer, Berlin, 2013.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14:326–334, 1965.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, 1999. Second edition: 2014.

R. M. Dudley. A paper that created three new fields: Teoriya veroyatnosteĭ i ee primeneniya 16(2), 1971, pp. 264–279. In *Measures of Complexity: Festschrift for Alexey Chervonenkis*, chapter 2, pages 9–10. Springer, Berlin, 2015a.

R. M. Dudley. Sketched history: VC combinatorics, 1826 up to 1975. In *Measures of Complexity: Festschrift for Alexey Chervonenkis*, chapter 4, pages 31–42. Springer, Berlin, 2015b.

Editors. Short biography of Alexey Chervonenkis. In Vladimir Vovk, Harris Papadopoulos, and Alex Gammerman, editors, *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pages ix–xvii. Springer, Berlin, 2015. This biography draws heavily on Chervonenkis's unpublished reminiscences.

Anders Hald. *History of Probability and Statistics and Their Applications before 1750*. Wiley, Hoboken, NJ, 2003.

David Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California at Santa Cruz, Computer Science Department, July 1999.

Vasily N. Novoseltsev. Institute of Control Sciences through the lens of VC dimension. In *Measures of Complexity: Festschrift for Alexey Chervonenkis*, chapter 5, pages 43–53. Springer, Berlin, 2015.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972. Submitted on 4 February 1970.

Ludwig Schläfli. *Theorie der vielfachen Kontinuität*, volume 38 (1st half) of *Denkschriften der Schweizerischen Naturforschenden Gesellschaft*. Zürcher & Furrer, Bern, 1901. Written in 1850–1852.

Michel Talagrand. The Glivenko–Cantelli problem. *Annals of Probability*, 15:837–870, 1987.

Michel Talagrand. The Glivenko–Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9:371–384, 1996.

Michel Talagrand. An email to Alex Gammerman of 12 October, 2014.

V. F. Turchin, V. P. Kozlov, and M. S. Malkevich. The use of mathematical-statistics methods in the solution of incorrectly posed problems. *Soviet Physics Uspekhi*, 13: 681–703, 1971. Russian original: В. Ф. Турчин, В. П. Козлов, М. С. Малкевич. Использование методов математической статистики для решения некорректных задач. *Успехи физических наук*, 102(3):345–386, 1970.

J. V. Uspensky. *Introduction to Mathematical Probability.* McGraw-Hill, New York, 1937.

Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Vladimir N. Vapnik and Aleksandr Ya. Lerner. Pattern recognition using generalized portraits. *Automation and Remote Control*, 24:709–715, 1963. Russian original: В. Н. Вапник, А. Я. Лернер. Узнавание образов при помощи обобщенных портретов. *Автоматика и телемеханика*, 24(6):774–780, 1964. The original article submitted on 26 December 1962.

Vladimir Vovk. Comment: The two styles of VC bounds. In *Measures of Complexity: Festschrift for Alexey Chervonenkis*, chapter 11, pages 161–164. Springer, Berlin, 2015.

Chris Watkins. Dynamic alignment kernels. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 39–50. MIT Press, Cambridge, MA, 2000.

Figure 1: Alexey Chervonenkis (1938–2014)