Exact Combinatorial Bounds on the Probability of Overfitting for Empirical Risk Minimization

K. V. Vorontsov

Dorodnicyn Computing Centre, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia
e-mail: vokov@forecsys.ru

Abstract—Three general methods for obtaining exact bounds on the probability of overfitting are proposed within statistical learning theory: a method of generating and destroying sets, a recurrent method, and a blockwise method. Six particular cases are considered to illustrate the application of these methods. These are the following model sets of predictors: a pair of predictors, a layer of a Boolean cube, an interval of a Boolean cube, a monotonic chain, a unimodal chain, and a unit neighborhood of the best predictor. For the interval and the unimodal chain, the results of numerical experiments are presented that demonstrate the effects of splitting and similarity on the probability of overfitting.

Key words: statistical learning theory, generalizing bounds, probability of overfitting, empirical risk minimization, splitting and similarity profile, layer of boolean cube, interval of boolean cube, monotonic chain of predictors, unimodal chain of predictors.

DOI: 10.1134/S105466181003003X

1. INTRODUCTION

Obtaining exact generalization bounds remains an open problem in statistical learning theory. The first bounds obtained in the Vapnik–Chervonenkis (VC) theory were highly overestimated [18, 16] and were subjected to many improvements later [19, 10, 7, 14]. However, the cases of small samples and complex function sets, which are of most practical interest, still remain beyond the scope of the theory because bounds are trivial in these cases. Overestimated bounds provide only a qualitative insight into the relation between overfitting and the complexity of the function set and do not always admit exact quantitative control of the learning process. The question of whether or not overfitting is related to some finer and not-yet-studied phenomena remains open.

The aim of the present study is to obtain exact bounds for the *overfitting probability*, i.e., for the probability that, for a given $\varepsilon \in (0, 1)$, the predictor with the least error rate on the training sample will have an error rate greater by ε on an independent test sample. Note that, to date, the problem of obtaining exact (rather than asymptotic or upper) bounds has not even been posed in statistical learning theory and was likely to be considered hopeless. Usually, one's goal is to find "tight bounds."

Experiments [20, 22] have shown that the overfitting probability depends not only on the complexity of the set of predictors (the number of different predictors in the set), but also on the diversity of these predictors. To obtain exact bounds, one should simultaneously take into account two fine effects: the splitting of the set into error levels and the similarity of predictors in the set. Experiments on real classification tasks [20] have shown that neglect of the splitting may increase the upper bound on the probability of overfitting by a factor of 10^2–10^5, while neglect of the similarity increases it by a factor of 10^3–10^4.

The *effect of splitting* is due to the fact that the number of predictors with a low error rate that are suitable for solving a given task is usually much smaller than the total number of predictors in the set. This is a consequence of the universality of the sets used in practice, which contain predictors suitable for solving a wide range of tasks. For each task, only a small "localized" subset of predictors is relevant, where a large part of the set actually remains idle. Taking into account idle predictors while defining the complexity measure of the set substantially weakens the bounds. However, it is rather difficult to describe this effect quantitatively because the localization of relevant predictors depends on the specific task and the specific learning algorithm.

The *effect of similarity* is due to the fact that, for any predictor, many similar predictors can exist in the set. Two predictors that differ in their error only on a single object manifest themselves as nearly a single predictor from the viewpoint of overfitting; hence, when evaluating the complexity of the set, one should also consider them as nearly a single predictor. Most of the classifiers used in practice have a separating surface that is continuous in the parameters; hence, the set of these classifiers is connected. In [15], this property was defined as the connectedness of a graph vertices of

Received April 14, 2010

which are various predictors and the edges of which connect predictors that differ in error only on a single object. In this paper, we will show that the existence of a path between any two predictors is not an essential property of a set. For estimating the probability of overfitting of a predictor, it is much more important to know the average number of other predictors in the set that differ from the given one in error only on a single object.

Experiments [20, 22] have shown that the neglect of one of these effects makes it impossible to obtain an exact bound. Known attempts to take into account these effects separately [7, 10, 5, 15] have not radically improved the accuracy of bounds and have not allowed one to approach the overfitting probabilities observed in the experiments. The author is unaware of any attempts to consider both effects simultaneously.

Experiments with a monotonic chain of predictors [22]—a model of a set of predictors with a single continuous parameter—confirm that learning algorithms that are effective in practice should necessarily possess both splitting and similarity properties. Otherwise the probability of overfitting would be close to one even for a set with a few tens of predictors.

Most of the known complexity bounds, except for Rademacher complexity bounds [9, 8] and the PAC-Bayes [13, 12] bounds, are derived from the union bound, which is the main cause of the overestimation. In the present paper, we develop a combinatorial approach that does not use the union bound and is based on weak probabilistic assumptions [20, 22, 23]. We propose three general methods for obtaining exact bounds: a method of generating and destroying sets (Section 3), a recurrent method (Section 4), and a blockwise method (Section 5). Then, we apply the above general methods to obtain exact bounds for the probability of overfitting in six particular cases. Most of these cases are constructed to demonstrate the possibility of simultaneous consideration of the effects of splitting and similarity.

This paper is an extended version of paper [23].

2. THE PROBABILITY OF OVERFITTING

Let there be given a finite set of objects $\mathbb{X} = \{x_1, ..., x_L\}$, called a *full*, or *general, sample*, and a finite set $A = \{a_1, ..., a_D\}$, whose elements are called *predictors*. There exists a binary function I: $A \times \mathbb{X} \longrightarrow \{0, 1\}$, called an *error indicator*. If $I(a, x) = 1$, it is said that the predictor a makes an error on the object x. The L-dimensional binary vector $\vec{a} = (I(a, x_i))_{i=1}^{L}$ is called an *error vector*. An *error matrix* of the set A is an $L \times D$ matrix composed of the column vectors $\vec{a}_1, ..., \vec{a}_D$. We assume that all column vectors are pairwise distinct. Therefore, $D = |A| \le 2^L$.

The number of errors of a predictor a on a sample $X \subseteq \mathbb{X}$ is defined as

$$n(a, X) = \sum_{x \in X} I(a, x),$$

and the *error rate*, or the *empirical risk*, of a predictor a on the sample X is defined as $\nu(a, X) = \frac{1}{|X|} n(a, X)$.

Fix a natural number $l < L$. Denote by $[\mathbb{X}]^l$ the set of all l-element subsets of the general sample $\mathbb{X}$. The cardinality of this set is C_L^l.

A *learning algorithm* is a mapping that assigns a certain predictor μX from A to an arbitrary *training sample* $X \in [\mathbb{X}]^l$.

The difference $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ is called the *deviation of the error rates* of the predictor a on the samples X and $\bar{X} = \mathbb{X} \backslash X$.

The deviation of the error rates of the predictor $a = \mu X$ is called the *overfitting* of the algorithm μ on the sample X.

$$\delta_\mu(X) = \delta(\mu X, X) = \nu(\mu X, \bar{X}) - \nu(\mu X, X).$$

An algorithm μ is said to be *overfitted* on a sample X if $\delta_\mu(X) \ge \varepsilon$ for a given $\varepsilon \in (0, 1)$.

A learning algorithm μ is called an *empirical risk minimization* (ERM) algorithm if

$$\mu X \in A(X) = \operatorname*{Argmin}_{a \in A} n(a, X). \qquad (1)$$

An ERM algorithm μ is said to be *optimistic* if

$$\mu X = \arg\min_{a \in A(X)} n(a, \bar{X}),$$

and *pessimistic* if

$$\mu X = \arg\max_{a \in A(X)} n(a, \bar{X}).$$

We assume that a sample $\bar{X}$ plays the role of a test sample and cannot be known at the time when a learning algorithm is applied to a training sample X. Therefore, optimistic and pessimistic ERMs cannot be implemented in practice. However, they are of significant theoretical interest because they provide sharp lower and upper bounds for the probability of overfitting.

Under the weak probabilistic axiom [20], we will assume that all C_L^l partitions of the set of objects $\mathbb{X}$ into an observed training sample X of length l and a hidden test sample $\bar{X}$ of length $k = L - l$ can be realized with equal probability.

The goal of this paper is to obtain exact bounds for the *probability of overfitting* for an ERM algorithm μ:

$$Q_\varepsilon \equiv \mathrm{P}[\delta_\mu(X) \ge \varepsilon] = \frac{1}{C_L^l} \sum_{X \in [\mathbb{X}]^l} [\delta_\mu(X) \ge \varepsilon]. \qquad (2)$$

Here and in what follows, a logical expression in square brackets means [true] = 1 and [false] = 0.

Consider a particular case when $A = \{a\}$ is a one-element set. For a fixed predictor a that makes $m = n(a, \mathbb{X})$, $m \in \{0, ..., L\}$, errors on the general sample, the probability to obtain exactly s errors on a subsample of X is described by the *hypergeometric probability function*

$$\mathrm{P}[n(a, X) = s] = C_m^s C_{L-m}^{l-s} / C_L^l \equiv h_L^{l,m}(s),$$

where the argument s takes integer values from $s_0 = \max\{0, m - k\}$ to $s_1 = \min\{m, l\}$. For any other integers m and s, we agree that the binomial coefficients C_m^s and the function $h_L^{l,m}(s)$ are zero.

Lemma 1. *Suppose that a predictor a makes $m = n(a, \mathbb{X})$ errors on the general sample. Then the probability of a large deviation of the error rates of the predictor a is described by a hypergeometric distribution function: for any $\varepsilon \in [0, 1)$,*

$$Q_\varepsilon = \mathrm{P}[\delta(a, X) \geq \varepsilon] = \mathrm{P}[n(a, X) \leq s_m(\varepsilon)] = \sum_{s' = s_0}^{\lfloor s_m(\varepsilon) \rfloor} h_L^{l,m}(s') \equiv H_L^{l,m}(s_m(\varepsilon)), \quad (3)$$

where $s_m(\varepsilon) = \left\lfloor \frac{l}{L}(m - \varepsilon k) \right\rfloor$ is the maximal number of errors $n(a, X)$ on the training sample, such that $\delta(a, X) = \frac{m-s}{k} - \frac{s}{l} \geq \varepsilon$.

Remark 1. When $l, k \longrightarrow \infty$, the right-hand side of (3) tends to zero and provides an exact bound for the convergence of error rates in the two samples.

Further, in the general case when $|A| > 1$, a hypergeometric distribution will also play an important role.

3. GENERATING AND DESTROYING SETS

In this section, we give exact bounds for the probability of overfitting that are based on the assumption that, for each predictor $a \in A$, one can write explicit conditions under which $\mu X = a$. We assume that A is a finite set and all predictors have pairwise distinct error vectors.

Conjecture 1. *Suppose that a set A, a sample $\mathbb{X}$, and an algorithm μ are such that, for each predictor $a \in A$, there exists a pair of non-intersecting subsets $X_a \subset \mathbb{X}$ and $X_a' \subset \mathbb{X}$ such that*

$$[\mu X = a] = [X_a \subseteq X][X_a' \subseteq \bar{X}]$$

for any $X \in [\mathbb{X}]^l$, where $\bar{X} = \mathbb{X}/X$.

For a predictor a, the objects from X_a are said to be *generating*; the objects from X_a', *destructive*; and the remaining objects, *neutral*. In other words, the learning algorithm returns the predictor if and only if its generating objects belong to the training sample and its destroying objects are outside the training sample. For each $a \in A$, introduce the following notation:

$L_a = L - |X_a| - |X_a'|$ is the number of neutral objects,

$l_a = l - |X_a|$ is the number of neutral training objects,

$m_a = n(a, \mathbb{X}) - n(a, X_a) - n(a, X_a')$ is the number of errors on neutral objects, and

$s_a(\varepsilon) = \frac{l}{L}(n(a, \mathbb{X}) - \varepsilon k) - n(a, X_a)$ is the maximum number of errors on neutral training objects such that $\delta(a, X) \geq \varepsilon$.

Lemma 2. *If Conjecture 1 is valid, then the probability to obtain a predictor a as a result of learning is*

$$P_a = \mathrm{P}[\mu X = a] = \frac{C_{L_a}^{l_a}}{C_L^l}.$$

Proof. According to Conjecture 1,

$$\mathrm{P}[\mu X = a] = \mathrm{P}[X_a \subseteq X][X_a' \subseteq \bar{X}].$$

This is a fraction of partitions of the general sample $\mathbb{X} = X \sqcup \bar{X}$ such that the set of objects X_a lies completely in X and the set of objects X_a' lies completely in $\bar{X}$. The number of such partitions is equal to the number of ways to choose l_a objects from among L_a neutral objects for a training subsample $X \backslash X_a$, which is obviously equal to $C_{L_a}^{l_a}$. The total number of partitions is C_L^l, and their ratio is exactly P_a.

Theorem 1. *If Conjecture 1 is valid, then the probability of overfitting is given by the formula*

$$Q_\varepsilon = \sum_{a \in A} P_a H_{L_a}^{l_a, m_a}(s_a(\varepsilon)).$$

Proof. The probability of overfitting Q_ε is expressed by the formula of total probability if, for each predictor a from A, the probability P_a to obtain it as a result of learning and the conditional probability $\mathrm{P}(\delta(a, X) \geq \varepsilon | a)$ of a large deviation of error rates under the condition that a predictor a is obtained are known:

$$Q_\varepsilon = \sum_{a \in A} P_a \mathrm{P}(\delta(a, X) \geq \varepsilon | a).$$

The conditional probability is given by Lemma 1 with regard to the fact that, for a fixed predictor a, the subsets X_a and X_a' do not take part in the partitions. Considering L_a neutral objects and all possible partitions of these objects into l_a training and $L_a - l_a$ test ones, we obtain

$$\mathrm{P}(\delta(a, X) \geq \varepsilon | a) = H_{L_a}^{l_a, m_a}(s_a(\varepsilon)),$$

which completes the proof.

Conjecture 1 imposes constraints on the sample $\mathbb{X}$, the set A, and the algorithm μ that are too restrictive. Therefore, Theorem 1 can be applied only in special cases. Consider a natural generalization of Conjecture 1. Suppose that, for each predictor, there exist many pairs of generating and destroying sets.

Conjecture 2. *Suppose that a set A, a sample $\mathbb{X}$, and an algorithm μ are such that, for each predictor $a \in A$, there exists a finite set of indices V_a and, for each index, there exist subsets of objects $X_{av} \subset \mathbb{X}$, $X'_{av} \subset \mathbb{X}$ and coefficients $c_{av} \in \mathbb{R}$, such that, for any $X \in [\mathbb{X}]^l$,*

$$[\mu X = a] = \sum_{v \in V_a} c_{av}[X_{av} \subseteq X][X'_{av} \subseteq \bar{X}]. \qquad (4)$$

Introduce the following notations for each $a \in A$ and $v \in V_a$:

$$L_{av} = L - |X_{av}| - |X'_{av}|,$$

$$l_{av} = l - |X_{av}|,$$

$$m_{av} = n(a, \mathbb{X}) - n(a, X_{av}) - n(a, X'_{av}),$$

$$s_{av}(\varepsilon) = \frac{l}{L}(n(a, \mathbb{X}) - \varepsilon k) - n(a, X_{av}).$$

Theorem 2. *If Conjecture 2 is valid, then the probability to obtain a predictor a as a result of learning is*

$$P_a = \mathrm{P}[\mu X = a] = \sum_{v \in V_a} c_{av} P_{av}, \qquad (5)$$

$$P_{av} = \mathrm{P}[X_{av} \subseteq X][X'_{av} \subseteq \bar{X}] = \frac{C^{l_{av}}_{L_{av}}}{C^l_L}, \qquad (6)$$

and the probability of overfitting is

$$Q_\varepsilon = \sum_{a \in A} \sum_{v \in V_a} c_{av} P_{av} H^{l_{av}, m_{av}}_{L_{av}}(s_{av}(\varepsilon)). \qquad (7)$$

The proof is largely similar to the proof of Lemma 2 and Theorem 1. The following theorem states that Conjecture 2 is not restrictive at all since it is always valid.

Theorem 3. *For any $\mathbb{X}$, A, and μ, there exist sets V_a, X_{av}, and X'_{av} such that representation* (4) *holds, where $c_{av} = 1$ for any $a \in A$ and $v \in V_a$.*

Proof. Fix an arbitrary predictor $a \in A$. Take, as the index set V_a, the set of all subsamples $v \in [\mathbb{X}]^l$ such that $\mu v = a$. For each $v \in V_a$, set $X_{av} = v$, $X'_{av} = \mathbb{X} \setminus v$, and $c_{av} = 1$. Then, for any $X \in [\mathbb{X}]^l$, the following representation of type (4) holds:

$$[\mu X = a] = \sum_{v \in V_a} [v = X]$$

$$= \sum_{v \in V_a} [v = X][\mathbb{X} \setminus v = \mathbb{X} \setminus X]$$

$$= \sum_{v \in V_a} [v \subseteq X][\mathbb{X} \setminus v \subseteq \bar{X}];$$

here, if $\mu X = a$, then exactly one term in this sum is equal to unity and other terms vanish, whereas, if $\mu X \neq a$, then all the terms vanish.

Remark 2. Theorem 2 is a typical existence theorem. The method of constructing index sets V_a used in the proof of this theorem requires explicit enumeration of all partitions of a sample, thus leading to computationally ineffective bounds on the probability of overfitting. However, representation (4) is not generally unique. A search for the representation for which the cardinalities of the sets $|V_a|$, $|X_{av}|$, $|X'_{av}|$ are as small as possible remains a key problem. Below, we will show that such representations can be obtained on the basis of the splitting and similarity properties in the sets of predictors.

A predictor a_0 that does not make errors on a sample $U \subseteq \mathbb{X}$ is said to be *correct on the sample U*. Formula (7) is strongly simplified if the set A contains a predictor that is correct on the whole general sample.

Theorem 4. *Suppose that Conjecture 2 is valid, the algorithm μ is an ERM algorithm, and the set A contains a predictor a_0 such that $n(a_0, \mathbb{X}) = 0$. Then the probability of overfitting reduces to*

$$Q_\varepsilon = \sum_{a \in A} [n(a, \mathbb{X}) \geq \varepsilon k] P_a. \qquad (8)$$

Proof. Consider an arbitrary predictor $a \in A$ and an arbitrary index $v \in V_a$. If an object on which a makes an error is contained in the training sample X, then the algorithm μ cannot choose this predictor because there exists a correct predictor a_0 that does not make errors on X. Hence, the set of objects on which the predictor a makes an error is completely contained in X'_{av}. Thus, the predictor a makes no errors on neutral objects and $m_{av} = 0$. In this case, the hypergeometric function $H^{l_{av}, 0}_{L_{av}}(s_{av}(\varepsilon))$ degenerates: for $s_{av}(\varepsilon) \geq 0$, it represents the sum of a single term equal to 1; when $s_{av}(\varepsilon) < 0$, the number of summands is zero, and the whole sum vanishes:

$$H^{l_{av}, 0}_{L_{av}}(s_{av}(\varepsilon)) = [s_{av}(\varepsilon) \geq 0] = [n(a, \mathbb{X}) \geq \varepsilon k].$$

Substituting this expression into (7), we obtain (8).

4. RECURRENT METHOD

Suppose that the error vectors of all predictors from the set A are known and pairwise distinct, and there is a predictor in $\mathbb{X}$ that is correct on X. Suppose that μ is a pessimistic ERM algorithm. Let us solve the following problem: for each predictor $a \in A$, find all the information necessary for calculating the overfitting probability Q_ε by Theorem 2:

$$\Im(a) = \langle X_{av}, X'_{av}, c_{av}\rangle_{v \in V_a},$$

where V_a is the index set, X_{av} is the set of generating objects, X'_{av} is the set of destroying objects, and $c_{av} \in \mathbb{R}$.

Let us renumber the predictors in the order of nondecreasing $n(a, \mathbb{X})$—the number of errors on the general sample: $A = \{a_0, \ldots, a_D\}$. It is obvious that $n(a_0, \mathbb{X}) = 0$. Denote by μ_d a pessimistic ERM algorithm that chooses predictors only from the subset $A_d = \{a_0, \ldots, a_d\}$. Consider a procedure of successive addition of predictors that upgrades from algorithm μ_{d-1} to algorithm μ_d at every step. Suppose that, for any predictor a_t, $t < d$, information $\Im(a_t)$ with respect to the algorithm μ_{d-1} is already calculated. Let us calculate information $\Im(a_d)$ and update the information $\Im(a_t)$, $t < d$ with respect to the algorithm μ_d. Note that such an update is necessary because the predictor a_d can "take away some partitions" from each of the preceding predictors a_t.

Lemma 3. *The algorithm μ_d chooses the predictor a_d if and only if all the objects on which a_d makes an error fall into the test sample*:

$$[\mu_d X = a_d] = [X'_d \subseteq \bar{X}],$$

$$X'_d = \{x_i \in \mathbb{X}\colon I(a_d, x_i) = 1\}.$$

Proof. If at least one object on which a_d makes an error belongs to the training sample X, then the algorithm μ_d chooses a predictor with a smaller number of errors on X. Such a predictor indeed exists; for example, this is a_0. Thus, the condition $X'_d \subseteq \bar{X}$ is necessary for the algorithm μ_d to choose the predictor a_d. Let us show that it is also a sufficient condition. To this end, it suffices to show that if there are several predictors in A_d that do not make errors on the trainings sample X, then the algorithm chooses precisely a_d among these predictors. Since the set A_d is ordered, the predictor a_d makes the maximum number of errors on $\mathbb{X}$, and, among the predictors with the same number of errors on $\mathbb{X}$, it has the maximal number. Therefore, according to the definition of a pessimistic ERM algorithm, the predictor a_d will be chosen by the algorithm μ_d from A_d whenever $X'_d \subseteq \bar{X}$.

Suppose that, immediately before the addition of the predictor a_d, the selection conditions for each preceding predictor a_t were expressed in the form (4):

$$[\mu_{d-1}X = a_t] = \sum_{v \in V_t} c_{tv} \underbrace{[X_{tv} \subseteq X][X'_{tv} \subseteq \bar{X}]}_{J_{tv}(d-1)}, \quad t < d.$$

After the addition of the predictor a_d, these conditions are changed. They are supplemented with the requirement that the set X'_d should not lie completely in the test sample $\bar{X}$; otherwise, the algorithm μ_d will choose the predictor a_d instead of a_t:

$$\begin{aligned}[\mu_d X = a_t] &= [\mu_{d-1}X = a_t][X'_d \not\subseteq \bar{X}] \\ &= \sum_{v \in V_t} c_{tv} \underbrace{[X_{tv} \subseteq X][X'_{tv} \subseteq \bar{X}][X'_d \not\subseteq \bar{X}]}_{J_{tv}(d)}, \quad t < d.\end{aligned} \tag{9}$$

To obtain an update rule for the information $\Im(a_t)$, it suffices to reduce expression (9) to the form (4). This will be done in the following lemma.

Lemma 4. *The update of the information $\Im(a_t)$, $t < d$, after the addition of the predictor a_d reduces to the verification of the following three conditions for every $v \in V_t$ such that $X_{tv} \cap X'_d = \varnothing$*:

(i) *if $X'_d \setminus X'_{tv} = \{x_i\}$ is a one-element set, then x_i is added to X_{tv}*;

(ii) *if $|X'_d \setminus X'_{tv}| > 1$, then the index set V_t is incremented with a new element (denote it by w), taking $c_{tw} = -c_{tv}$, $X_{tw} = X_{tv}$, $X'_{tw} = X'_{tv} \cup X'_d$*;

(iii) *if $|X'_d \setminus X'_{tv}| = 0$, then the index v is removed from the index set V_t; accordingly, the whole triple $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$ is removed from $\Im(a_t)$.*

Proof. If $X_{tv} \cap X'_d \neq \varnothing$, then it follows from $X_{tv} \subseteq X$ that the set X'_d does not completely belong to the test sample $\bar{X}$. Hence, the triple $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$ does not need any update:

$$J_{tv}(d) = [X_{tv} \subseteq X][X'_{tv} \subseteq \bar{X}] = J_{tv}(d-1).$$

If $X_{tv} \cap X'_d = \varnothing$, then three cases are possible depending on the cardinality of the set $X'_d \setminus X'_{tv}$.

The first case: $X'_d \setminus X'_{tv} = \{x_i\}$ is a one-element set. Then the following chain of equalities holds: $[X'_d \not\subseteq \bar{X}] = [x_i \notin \bar{X}] = [x_i \in X]$. Substituting this into (9), we obtain

$$J_{tv}(d) = [X_{tv} \sqcup \{x_i\} \subseteq X][X'_{tv} \subseteq \bar{X}].$$

The second case: $|X'_d \setminus X'_{tv}| > 1$. Then

$$\begin{aligned}J_{tv}(d) &= [X_{tv} \subseteq X][X'_{tv} \subseteq \bar{X}](1 - [X'_d \subseteq \bar{X}]) \\ &= J_{tv}(d-1) - [X_{tv} \subseteq X][X'_{tv} \cup X'_d \subseteq \bar{X}].\end{aligned} \tag{10}$$

Thus, one more term appears in the expression for $[\mu_d X = a_t]$; this is equivalent to adding one more index (denote it by w) to the set V_t, such that $c_{tw} = -c_{tv}$, $X_{tw} = X_{tv}$, and $X'_{tw} = X'_{tv} \cup X'_d$.

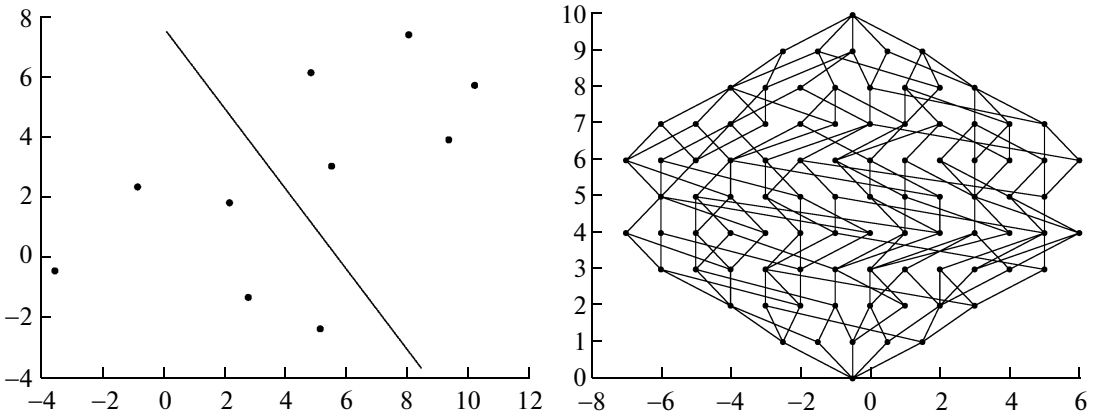

Fig. 1. Linearly separable general sample and the connectivity graph of a set of linear classifiers.

Finally, the third case: $|X'_d \setminus X'_{tv}| = 0$. Then representation (10) remains valid; however, the difference $J_{tv}(d)$ turns out to be zero because $X'_{tv} \cup X'_d = X'_{tv}$. The vanishing of $J_{tv}(d)$ is equivalent to removing the index v from the index set V_t together with the removal of the corresponding triple $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$ from the information $\Im(a_t)$.

Lemmas 3 and 4 and Theorem 4 allow one to recurrently calculate the probability of overfitting Q_ε. At every dth step, $d = 0, \ldots, D$, a predictor a_d is added, and the information $\Im(a_d)$ is calculated; then, for every t, $t = 0, \ldots, d-1$, the information $\Im(a_t)$ and the probabilities P_{tv} are updated. Based on the updated information, the current bound for Q_ε is recalculated. After the last Dth step, the current bound for Q_ε should coincide with the exact value of the probability of overfitting.

This procedure may be computationally inefficient if condition (ii) of Lemma 4 is fulfilled too often. Every time this condition is fulfilled, one more term is added to the sum (7). Hence, the number of terms may increase exponentially with respect to the number of predictors D. The following theorem allows one to trade off between the computing time and the accuracy of the upper bound of Q_ε.

Theorem 5. *Let, in Lemma* 4, *condition* (ii) *hold and* $c_{tv} = 1$. *If the index set* V_t *is not incremented, then the calculated value of* Q_ε *will be an upper bound of the probability of overfitting.*

Proof. The increment of the index set V_t when $c_{tv} = 1$ and $c_{tw} = -1$ leads to a decrease by $P_{tw} \geq 0$ in the current calculated value of Q_ε. The neglect of the increment leads to the elimination, from the sum (8), of the negative term $-P_{tw}$ and, possibly, of a few other positive and negative terms that appear in this sum as a result of further updates of the triple $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$ under condition (iii). Each such update arises as a result of addition of a certain predictor a_d, $d > t$, that takes away a part of the partitions from the predictor a_t, thus reducing the term P_{tw} to the value $\tilde{P}_{tw}$:

$$\tilde{P}_{tw} = \mathrm{P}[X_{tw} \subseteq X][X'_{tw} \subseteq \overline{X}][X'_d \not\subseteq \overline{X}] \leq P_{tw}.$$

The elimination from the sum (8) of the negative term $-P_{tw}$ together with all the subsequent terms that update the triple $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$ can only increase the resulting value of Q_ε. The theorem is proved.

Remark 3. One can similarly prove that if the index set V_t is not incremented under condition (ii) when $c_{tv} = -1$, then the calculated value of Q_ε provides a lower bound for the probability of overfitting.

If one never increments the index set under condition (ii) of Lemma 4, then one obtains a *simplified recurrent procedure* for calculating the probability of overfitting. In this case, triples $\langle X_{tw}, X'_{tw}, c_{tw}\rangle$ with negative values of c_{tw} will never appear, each predictor a_d will correspond to a single triple, and all the index sets V_d, $d = 0, \ldots, D$, will consist of a single element. According to Theorem 5, the calculated value of Q_ε will be an upper bound of the probability of overfitting. This bound can be expressed in explicit form in terms of a splitting and similarity profile of the set A.

A subset of predictors $A_m = \{a \in A: n(a, \mathbb{X}) = m\}$ is called the *m-th layer* of the set A. The partition of $A = A_0 \sqcup \ldots \sqcup A_L$ is called a *splitting* of the set of predictors A.

The *connectivity* $q(a)$ of a predictor $a \in A$ is the number of predictors in the next layer that make errors on the same objects as a:

$$q(a) = \#\{a' \in A_{n(a,\mathbb{X})+1}: I(a, x) \leq I(a', x), x \in \mathbb{X}\}.$$

Thus, the connectivity $q(a)$ is the number of error vectors in A that are worse than a on some object.

For every predictor $a \in A$, denote by E_a the set of objects of the general sample $\mathbb{X}$ on which the predictor makes an error: $E_a = \{x_i \in \mathbb{X}: I(a, x_i) = 1\}$. It is obvious that $n(a, \mathbb{X}) = |E_a|$.

A *graph of connectivity*, or simply a graph of the set of predictors A, is a directed graph whose vertices correspond to predictors and the edges (a, a') connect pairs of predictors such that $E_{a'} \backslash E_a = 1$. Then, the connectivity $q(a)$ of the predictor a is the number of edges of the graph that emanate from the vertex a.

A *splitting and similarity profile* of a set A is an $L \times L$ matrix $(\Delta_{mq})_{m=0}^{L}{}_{q=0}^{L}$, where Δ_{mq} is the number of predictors in the mth layer with connectivity q.

Example 1. Figure 1 (left) shows a two-dimensional linearly separable sample of length $L = 10$ that consists of objects of two classes, with five objects in each class. The figure on the right shows the connectivity graph of the set of linear classifiers for a given sample. The vertical axis enumerates layers m. The only vertex of the graph at $m = 0$ corresponds to the classifier that separates two classes without errors. The next layer $m = 1$ contains five classifiers that separate the sample with one error. The layer $m = 2$ contains eight classifiers with two errors, etc.

Theorem 6. *Suppose that the error vectors of all the predictors of the set A are pairwise distinct, A contains a predictor that is correct on* $\mathbb{X}$, *and* Δ_{mq} *is the number of predictors in the m-th layer with connectivity q. Then the following upper bound is valid*:

$$Q_\varepsilon \le \sum_{m=[\varepsilon k]}^{L} \sum_{q=0}^{L} \Delta_{mq} \frac{C_{L-m-q}^{l-q}}{C_L^l}. \qquad (11)$$

Proof. Consider a simplified recurrent procedure that gives an upper bound on the probability of overfitting. For each predictor $a \in A$, a unique triple $\langle X_a, X_a', 1\rangle$ is constructed in which the destroying set X_a' coincides with E_a and the generating set consists of all objects that are added when satisfying condition (i) of Lemma 4. These are those and only those objects x_i for which there exists a predictor $a' \in A$ that makes one error more than a. Obviously, $X_{a'}' \backslash X_a' = E_{a'} \backslash E_a = \{x_i\}$ is a one-element set. The number of such objects x_i coincides with the value of the connectivity $q(a)$. Thus, $|X_a'| = n(a, \mathbb{X})$ and $|X_a| = q(a)$ for an arbitrary predictor $a \in A$. Hence, bound (8) is rewritten as

$$Q_\varepsilon \le \sum_{a \in A} [n(a, \mathbb{X}) \ge \varepsilon k] \frac{C_{L_a}^{l_a}}{C_L^l}$$

$$= \sum_{a \in A} [n(a, X) \ge \varepsilon k] \frac{C_{L-n(a,\mathbb{X})-q(a)}^{l-q(a)}}{C_L^l}$$

$$= \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{L} \underbrace{\sum_{a \in A} [n(a, \mathbb{X}) = m][q(a) = q]}_{\Delta_{mq}} \frac{C_{L-m-q}^{l-q}}{C_L^l}.$$

According to bound (11), the maximal contribution to the probability of overfitting is made by predictors with a smaller number of errors, starting with $m = \lceil \varepsilon k \rceil$. As m increases, the combinatorial multiplier $\frac{C_{L-m-q}^{l-q}}{C_L^l}$ decreases exponentially.

The increase in the connectivity q improves the bound. In the experiments with linear classifiers, the mean value of connectivity q was proportional to the dimension of the space (the number of features) with the proportionality factor close to unity [3]. Generally, an increase in the dimension of the space gives rise to two opposite phenomena: on the one hand, the number of predictors in each layer increases, which leads to the increase in Q_ε; on the other hand, the connectivity q increases, which decreases the growth rate of Q_ε.

Preliminary experiments have shown that the *splitting and similarity profiles* Δ_{mq} for a certain set of predictors are separable to a high degree of accuracy: $\Delta_{mq} \approx \Delta_m \lambda_q$, where Δ_m is the number of different predictors in the mth layer and λ_q is the fraction of predictors of the mth layer that have connectivity q. It is reasonable to call the vector $(\Delta_m)_{m=0}^{L}$ a *splitting profile*, and the vector $(\lambda_q)_{q=0}^{L}$, a *similarity profile* of the set of predictors A. The similarity profile satisfies the normalization condition $\lambda_0 + \ldots + \lambda_L = 1$.

In terms of the splitting and similarity profiles, the bound (11) is rewritten as

$$Q_\varepsilon \lessapprox \underbrace{\sum_{m=\lceil \varepsilon k \rceil}^{k} \Delta_m \frac{C_{L-m}^{l}}{C_L^l}}_{\text{VC bound}} \underbrace{\sum_{q=0}^{L} \lambda_q \left(\frac{l}{L-m}\right)^q}_{\text{correction for connectivity}}. \qquad (12)$$

The first part of this bound is the VC bound expressed in terms of the splitting profile [17, 2], which is valid when the ERM algorithm always finds a predictor that makes no error on the training sample. In the present situation, this is the case, because the set A contains a correct predictor, $n(a_0, \mathbb{X}) = 0$. The second part of the bound is a correction for connectivity. It decreases exponentially as q increases, thus making the bound much more accurate than the classical VC-type bounds.

5. BLOCKWISE BOUND

Suppose that the error vectors of all predictors from the set $A = \{a_1, \ldots, a_D\}$ are known and pairwise distinct. Assume that μ is a pessimistic ERM algorithm: when $n(a, X)$ attains its minimum on several predictors, μ chooses a predictor with larger $n(a, \bar{X})$, and if there are several such predictors, then it chooses a predictor with a larger ordinal number.

The values of $I(a_d, x_i)$ form a binary $L \times D$ error matrix the columns of which are error vectors of the predictors and the rows of which correspond to objects. Denote by $b = (b_1, \ldots, b_D)$ an arbitrary binary

vector of dimension D. The sample $\mathbb{X}$ is partitioned into disjoint *blocks* so that all the objects in a block correspond to the same row $b = (b_1, \ldots, b_D)$ in the error matrix:

$$U_b = \{x_i \in \mathbb{X} \mid I(a_d, x_i) = b_d, d = 1, \ldots, D\}.$$

Denote by B the set of binary vectors b that correspond to nonempty blocks U_b. It is obvious that $|B| \le \min\{L, 2^D\}$.

Denote $m_b = |U_b|$.

To each training sample $X \in [\mathbb{X}]^l$, we assign an integer-valued vector $(s_b)_{b \in B}$ such that $s_b = |X \cap U_b|$ is the number of objects in the block U_b that fall into the training sample. Denote the set of all such vectors corresponding to all possible training samples by S. Obviously, S can also be defined in a different way:

$$S = \left\{ s = (s_b)_{b \in B} \mid s_b = 0, \ldots, m_b, \sum_{b \in B} s_b = l \right\}.$$

Let us write the numbers of errors of the predictor a_d made on the training sample X and on the test sample $\bar{X}$ as sums over blocks:

$$n(a_d, X) = \sum_{b \in B} b_d |X \cap U_b| = \sum_{b \in B} b_d s_b,$$

$$n(a_d, \bar{X}) = \sum_{b \in B} b_d |\bar{X} \cap U_b| = \sum_{b \in B} b_d (m_b - s_b).$$

Thus, the choice of a predictor by an algorithm μ depends only on how many objects s_b from each block fall into the training sample and does not depend on what these objects are. Define the function $d^*: S \longrightarrow \{1, \ldots, D\}$ as the number of a predictor chosen by algorithm μ from the training sample. If μ is a pessimistic ERM algorithm, we set

$$\begin{aligned} A(s) &= \operatorname*{Argmin}_{d = 1, \ldots, D} \sum_{b \in B} b_d s_b, \\ A'(s) &= \operatorname*{Argmax}_{d \in A(s)} \sum_{b \in B} b_d (m_b - s_b), \\ d^*(s) &= \max\{d: d \in A'(s)\}, \end{aligned} \tag{13}$$

where $\operatorname*{Argmin}_{d = 1, \ldots, D} f(d)$ denotes the set of values of d such that the function $f(d)$ attains its minimum.

Theorem 7. *Suppose that μ is a pessimistic ERM and the error vectors of all predictors $a \in A$ are pairwise distinct. Then the probability to obtain a predictor a_d as a result of learning is*

$$\mathrm{P}[\mu X = a_d] = \frac{1}{C_L^l} \sum_{s \in S} \left(\prod_{b \in B} C_{m_b}^{s_b} \right) [d^*(s) = d], \tag{14}$$

while the probability of overfitting is

$$Q_\varepsilon = \frac{1}{C_L^l} \sum_{s \in S} \left(\prod_{b \in B} C_{m_b}^{s_b} \right) \left[\sum_{b \in B} b_{d^*(s)} (m_b l - s_b L) \ge \varepsilon k l \right]. \tag{15}$$

Proof. To an arbitrary set of values $(s_b)_{b \in B}$ from S, there corresponds a set of samples $X \in [\mathbb{X}]^l$ such that $|X \cap U_b| = s_b$. The number of such samples is given by the product $\prod_{b \in B} C_{m_b}^{s_b}$ because, for every block U_b, there exist $C_{m_b}^{s_b}$ ways to select s_b objects into the subsample $X \cap U_b$.

Since the conditions $\mu X = a_d$ and $d^*(s) = d$ are equivalent, the probability to obtain a predictor a_d as a result of learning is expressed as

$$\begin{aligned} \mathrm{P}[\mu X = a_d] &= \frac{1}{C_L^l} \sum_{X \in [\mathbb{X}]^l} [d^*(s) = d] \\ &= \frac{1}{C_L^l} \sum_{s \in S} \left(\prod_{b \in B} C_{m_b}^{s_b} \right) [d^*(s) = d]. \end{aligned}$$

Now, let us write the probability of overfitting:

$$Q_\varepsilon = \mathrm{P}[\delta_\mu(X) \ge \varepsilon] = \mathrm{P} \sum_{d = 1}^{D} [\mu X = a_d] [\delta(a_d, X) \ge \varepsilon].$$

The deviation of error rates of the predictor a_d can be expressed as a sum over blocks:

$$\begin{aligned} \delta(a_d, X) &= \frac{1}{k} \sum_{b \in B} b_d (m_b - s_b) - \frac{1}{l} \sum_{b \in B} b_d s_b \\ &= \frac{1}{lk} \sum_{b \in B} b_d (m_b l - s_b L). \end{aligned}$$

Then, the expression for the probability of overfitting is rewritten as

$$\begin{aligned} Q_\varepsilon = \frac{1}{C_L^l} \sum_{s \in S} \left(\prod_{b \in B} C_{m_b}^{s_b} \right) \sum_{d = 1}^{D} [d^*(s) = d] \\ \times \left[\sum_{b \in B} b_d (m_b l - s_b L) \ge \varepsilon l k \right]. \end{aligned}$$

This implies the required formula (15).

Remark 4. If the set B contains vectors b that correspond to empty blocks U_b, then formulas (14) and (15) remain valid because then $m_b = s_b = 0$.

Remark 5. Direct calculations by formulas (14) and (15) may require considerable time exponential in the sample length L. In the worst case, when all U_b are one-element blocks, the set S consists of all possible Boolean vectors of length L that contain exactly l units. In this case, the number of terms in (14) and

(15) is C_L^l. Calculations by Theorem 7 are only effective when the number of blocks $|B|$ is small, in particular, when the number of predictors is small.

6. A TWO-ELEMENT SET OF PREDICTORS

Consider a particular case of a two-element set $A = \{a_1, a_2\}$. Even this simple case illustrates both the overfitting phenomenon itself and the effects of splitting and similarity, which reduce the probability of overfitting. An exact estimate for the probability of overfitting in this special case was obtained in [21]. Consider a shorter proof based on the blockwise method. Set $B =$ (1.1), (1.0), (0.1), (0.0).

Suppose that, in a sample $\mathbb{X}$, there are m_{11} objects on which both predictors make an error, m_{10} objects on which only a_1 makes an error, m_{01} objects on which only a_2 makes an error, and $m_{00} = L - m_{11} - m_{10} - m_{01}$ objects on which both predictors give a correct answer:

$$\vec{a}_1 = (1, \ldots, 1, 1, \ldots, 1, 0, \ldots, 0, 0, \ldots, 0),$$

$$\vec{a}_2 = (\underbrace{1, \ldots, 1}_{m_{11}}, \underbrace{0, \ldots, 0}_{m_{10}}, \underbrace{1, \ldots, 1}_{m_{01}}, \underbrace{0, \ldots, 0}_{m_{00}}).$$

Theorem 8. *Suppose that μ is a pessimistic ERM algorithm and the set consists of two predictors, $A = \{a_1, a_2\}$. Then the following exact bound is valid for any $\varepsilon \in [0, 1)$:*

$$Q_\varepsilon = \sum_{s_{11}=0}^{m_{11}} \sum_{s_{10}=0}^{m_{10}} \sum_{s_{01}=0}^{m_{01}} \frac{C_{m_{11}}^{s_{11}} C_{m_{10}}^{s_{10}} C_{m_{01}}^{s_{01}} C_{L-m_{11}-m_{10}-m_{01}}^{l-s_{11}-s_{10}-s_{01}}}{C_L^l}$$

$$\times \left([s_{10} < s_{01}] \left[s_{11} + s_{10} \le \frac{l}{L}(m_{11} + m_{10} - \varepsilon k) \right] \right. \quad (16)$$

$$\left. + [s_{10} \ge s_{01}] \left[s_{11} + s_{01} \le \frac{l}{L}(m_{11} + m_{01} - \varepsilon k) \right] \right).$$

Proof. Let us apply Theorem 7.

The set S consists of integer-valued vectors $s = (s_{11}, s_{10}, s_{01}, s_{00})$ such that $s_{11} + s_{10} + s_{01} + s_{00} = l$. Therefore, the sum $\sum_{s \in S}$ is transformed into a triple sum $\sum_{s_{11}=0}^{m_{11}} \sum_{s_{10}=0}^{m_{10}} \sum_{s_{01}=0}^{m_{01}}$, and s_{00} is expressed in terms of other components of the vector s.

The number $d^*(s)$ of a predictor chosen by algorithm μ from the training sample is 1 when $s_{10} < s_{01}$ and 2 when $s_{10} \ge s_{01}$.

Now, we substitute the values of m_b, s_b, and $d^*(s)$ into (15):

$$\left[\sum_{b \in B} b_{d^*(s)}(m_b l - s_b L) \ge \varepsilon l k \right]$$

$$= [d^*(s) = 1][(m_{10} + m_{11})l - (s_{10} + s_{11})L \ge \varepsilon l k]$$

$$+ [d^*(s) = 2][(m_{01} + m_{11})l - (s_{01} + s_{11})L \ge \varepsilon l k].$$

This implies the required equality (16).

In the case of $m_{10} = m_{01} = L/2$, when the predictors are maximally different and equally bad, the value of Q_ε is maximal and is twice the value of Q_ε of an individual predictor (3). Hence, we can conclude that overfitting arises whenever a choice out of several alternatives is made by incomplete information, even if there are only two alternatives. If the two predictors are very similar, or if one of them is much better than the other, then the overfitting vanishes. Thus, the effects of splitting and similarity reduce the probability of overfitting even in the case of two predictors [21].

7. A LAYER OF A BOOLEAN CUBE

Consider a set A consisting of all C_L^m predictors that make exactly m errors on the general sample $\mathbb{X}$ and have pairwise distinct error vectors. Since all possible error vectors form a Boolean cube of size L, the error vectors of the set A form the mth layer of the Boolean cube.

Theorem 9. *Suppose that μ is a pessimistic ERM algorithm and A is the m-th layer of a Boolean cube. Then*

$$Q_\varepsilon = [\varepsilon k \le m \le L - \varepsilon l]$$

for any $\varepsilon \in [0, 1]$.

Proof. If $m \le k$, then the empirical risk attains its minimum on an $a \in A$ such that all m errors fall into the test sample and no error falls into the training sample. Then $\nu(a, \bar{X}) = \frac{m}{k}$, $\nu(a, X) = 0$, and

$$Q_\varepsilon = \mathrm{P}\left[\frac{m}{k} - 0 \ge \varepsilon \right] = [\varepsilon k \le m \le k].$$

If $m > k$, then the empirical risk attains its minimum on the $a \in A$ that makes errors on all test objects. Then

$$Q_\varepsilon = \mathrm{P}\left[1 - \frac{m-k}{l} \ge \varepsilon \right] = [k < m \le L - \varepsilon l].$$

Combining two mutually exclusive cases $m \le k$ and $m > k$, we complete the proof.

Thus, the probability of overfitting takes values of either 0 or 1. Although this result is trivial and, in a sense, negative, it allows one to draw a few important conclusions. First, the predictors of the lowest layers, $m < \lceil \varepsilon k \rceil$, do not contribute to overfitting. Second, the lowest layer of the set of predictors, which contains predictors with the number of errors not less than $\lceil \varepsilon k \rceil$, should not contain *all* such predictors. The ERM

algorithm within a too rich set of predictors leads to overfitting.

8. AN INTERVAL OF A BOOLEAN CUBE

Suppose that the error vectors of all predictors from A are pairwise distinct and form an interval of rank m in an L-dimensional Boolean cube. This means that the objects are divided into three groups: m_0 "internal" objects, on which none of the predictors makes errors; m_1 "noise" objects, on which all predictors make errors; and m "boundary" objects, on which all 2^m variants of making an error are realized. There are no other objects: $m_0 + m_1 + m = L$.

An interval of a Boolean cube possesses the properties of splitting and similarity and can be considered as a model of practically used sets of predictors. The number of predictors in this interval is 2^m. The predictors make from m_1 to $m_1 + m$ errors. None of the layers of the Boolean cube is completely contained in A, except for a particular case of no interest where $m = L$ and A coincides with the Boolean cube. The parameter m characterizes the complexity, or the "dimension," of this set.

For the sake of greater generality, consider a set of predictors A_t formed by t lower layers of a Boolean cube interval. The number of different error vectors in A_t is $C_m^0 + C_m^1 + \ldots + C_m^t$. The predictors make from m_1 to $m_1 + t$ errors. The parameter t can take values of $0, \ldots, m$. This model set is of interest in that it allows one to analyze the effect of splitting on the probability of overfitting by considering how Q_ε depends on the number of lower layers t.

Theorem 10. *Suppose that μ is a pessimistic ERM algorithm and A is the set of t lower layers of a Boolean cube interval with m boundary and m_1 noise objects. Then, for any $\varepsilon \in [0, 1]$, the probability of overfitting is given by*

$$Q_\varepsilon = \sum_{s=0}^{m}\sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{l-s-s_1}}{C_L^l} \times \left[s_1 \le \frac{l}{L}(m_1 + \min\{t, m-s\} - \varepsilon k)\right].$$

Proof. Denote by X_0, X_1, and S the sets of all internal, noise, and boundary objects, respectively, and by s_0, s_1, and s, the numbers of internal, noise, and boundary objects, respectively, that fall into the training sample X.

Since the algorithm μ is pessimistic, it always chooses a predictor from A that makes no errors on all training boundary objects but makes errors on all test boundary objects. Therefore,

$$\nu(\mu X, X) = \frac{s_1}{l};$$

$$\nu(\mu X, \bar{X}) = \frac{(m_1 - s_1) + \min\{t, m-s\}}{k}.$$

The number of partitions $X \sqcup \bar{X}$ such that $|X_0 \cap X| = s_0$, $|X_1 \cap X| = s_1$, and $|S \cap X| = s$ is $C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s$. Hence, the probability of overfitting can be represented as

$$Q_\varepsilon = \sum_{\substack{s_0=0 \\ s_0+s_1+s=l}}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s=0}^{m} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_m^s}{C_L^l} \times \left[\frac{(m_1 - s_1) + \min\{t, m-s\}}{k} - \frac{s_1}{l} \ge \varepsilon\right].$$

To obtain the assertion of the theorem, it suffices to apply the relations $m_0 + m_1 + m = L$ and $s_0 + s_1 + s = l$ and transform the inequality in square brackets to $s_1 \le \frac{l}{L}(m_1 + \min\{t, m-s\} - \varepsilon k)$.

Figure 2 shows the probability of overfitting Q_ε as a function of the error level $t = m_1, \ldots, m_1 + m$. The three experiments differ by the length of the general sample (200, 400, 1000), while the proportions of $\frac{m}{L} = 0.2$ and $\frac{m_1}{L} = 0.05$ are preserved; in other words, the general sample contains 20% of boundary and 5% of noise objects in all three cases. The diagrams also demonstrate the contributions of layers to the value of the functional Q_ε. Only the lower layers make nonzero contributions. The ruggedness of the graphs is attributed to the fact that, in view of the relation $\frac{l}{L} = \frac{1}{2}$, every second layer makes no contribution to Q_ε.

In this experiment it turned out that 20% of boundary objects represents such a powerful interval that the probability of overfitting reaches a value of 1 too rapidly. The probability of overfitting is close to zero only if one takes the lowest layers of the interval, which amount to at most 2% of the sample length.

This implies two conclusions. First, a good generalization is hardly possible if the sample contains a considerable number of boundary objects on which predictors can make errors in all possible ways. The fraction of such objects is actually added to the value of overfitting. Second, an interval of a Boolean cube is not a quite adequate model of real sets. The hypothesis on the existence of boundary objects seems reasonable. However, it is likely that, in real problems, the predictors of the set by no means realize all variants of making errors on boundary objects. Perhaps a model

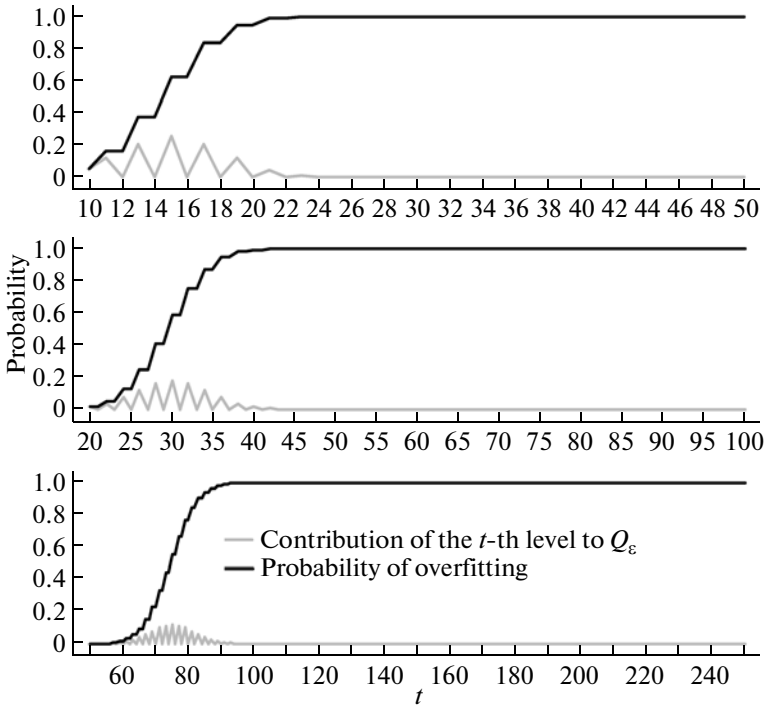

Fig. 2. Probability of overfitting Q_ε as a function of the number of errors for $\varepsilon = 0.05$. Top: $l = k = 100$, $m_1 = 10$, and $m = 40$; middle: $l = k = 200$, $m_1 = 20$, and $m = 80$; and bottom: $l = k = 500$, $m_1 = 50$, and $m = 200$.

in which one somehow introduces a characteristic of the "degree of boundariness" of objects and estimates the distribution of this characteristic over the sample would be more adequate.

9. A MONOTONIC CHAIN OF PREDICTORS

A monotonic chain of predictors seems to be the simplest model set that possesses the properties of splitting and similarity. A monotonic chain is generated by a one-parameter connected set of predictors under the assumption that continuous variation of the parameter away from its optimal value can only increase the number of errors made on the general sample.

Introduce a Hamming distance between the error vectors of predictors:

$$\rho(a, a') = \sum_{i=1}^{L} \left| I(a, x_i) - I(a', x_i) \right|, \quad \forall a, a' \in A.$$

A set of predictors $a_0, a_1, \ldots, a_D$ is called a *chain* if $\rho(a_{d-1}, a_d) = 1$, $d = 1, \ldots, D$. A chain of predictors is said to be *monotonic* if $n(a_d, \mathbb{X}) = m + d$ for some $m \geq 0$. The predictor a_0 is said to be *the best in the chain*.

Example 2. Let $\mathbb{X}$ be a set of points in $\mathbb{R}^n$ and A be a set of linear classifiers—parametric mappings from $\mathbb{X}$ into $\{-1, +1\}$ of the form

$$a(x, w) = \operatorname{sgn}(x_1 w_1 + \ldots + x_n w_n),$$
$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

with parameter $w \in \mathbb{R}^n$. Suppose that a loss function is given by $I(a, x) = [a(x, w) \neq y(x)]$, where $y(x)$ is the true classification of the object x and the set of objects in $\mathbb{X}$ is linearly separable; i.e., there exists a $w^* \in \mathbb{R}^n$ such that a classifier $a(x, w^*)$ makes no errors on $\mathbb{X}$. Then, under some additional technical assumptions, the set of classifiers $\{a(x, w^* + t\delta)\colon t \in [0, +\infty)\}$ obtained by shifting or rotating the direction vector w of the separating hyperplane forms a monotonic chain for any given vector $\delta \in \mathbb{R}^n$ except for some finite set of vectors. In this case, $m = 0$.

Theorem 11. *Let $A = \{a_0, a_1, \ldots, a_D\}$ be a monotonic chain and $L \geq m + D$. Then*

$$Q_\varepsilon = \sum_{d=0}^{k} P_d H^{l-1,m}_{L-d-1}(s_d(\varepsilon)),$$
$$P_d = \frac{C^{l-1}_{L-d-1}}{C^l_L}, \quad d = 0, \ldots, k$$

when $D \geq k$ and

$$Q_\varepsilon = \sum_{d=0}^{D-1} P_d H^{l-1,m}_{L-d-1}(s_d(\varepsilon)) + P_D H^{l,m}_{L-D}(s_D(\varepsilon)),$$

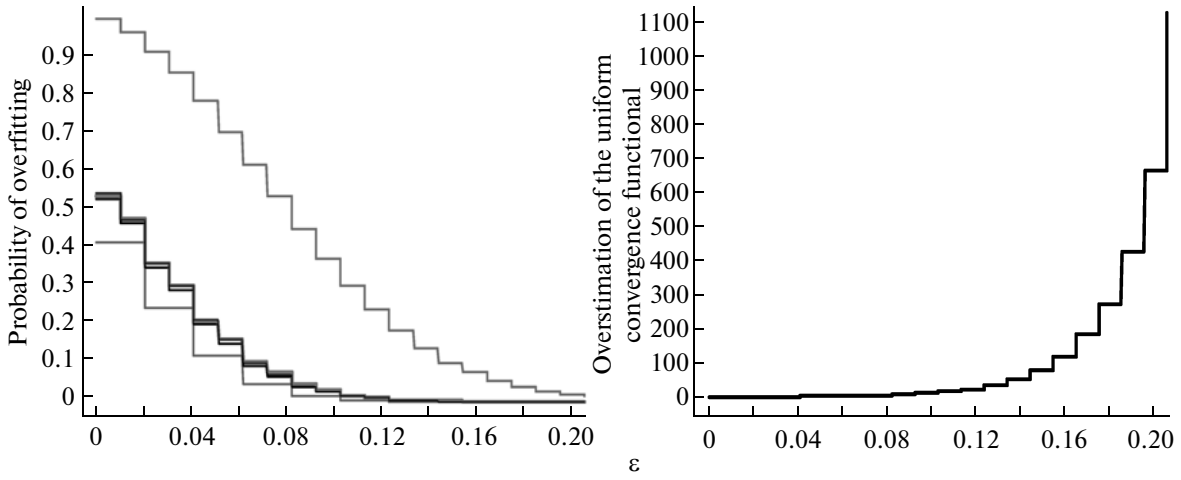

Fig. 3. (left) Bounds for the probability of overfitting Q_ε as a function of ε: exact bound from Theorem 9.1 and four bounds calculated by the Monte Carlo method using 1000 random partitions: for optimistic (lower curve), pessimistic, and randomized ERM. The upper curve corresponds to the bound for a uniform functional R_ε. (right) The factor of overestimation of the uniform functional $R_\varepsilon/Q_\varepsilon$. All the charts are plotted for $l = k = 100$ and $m = 20$.

$$P_d = \frac{C_{L-d-1}^{l-1}}{C_L^l}, \quad d = 0, \ldots, D-1, \quad P_D = \frac{C_{L-D}^{l}}{C_L^l},$$

when $D < k$, *where* P_d *is the probability to obtain a predictor* a_d *by algorithm* μ *and* $s_d(\varepsilon) = \frac{l}{L}(m + d - \varepsilon k)$.

Proof. Let us renumber the objects so that each predictor a_d, $d = 1, \ldots, D$, makes an error on the objects $x_1, \ldots, x_d$. Obviously, the best predictor a_0 makes no error on any of these objects. The numbering of other objects does not matter because the predictors are indistinguishable on these objects.

For the sake of clarity, let us partition the sample $\mathbb{X}$ into three blocks:

$$\begin{aligned}
&\qquad x_1\ x_2\ x_3 \qquad x_D \qquad\qquad \overbrace{\qquad}^{m}\\
\mathbf{a}_0 &= (\ 0,\ 0,\ 0,\ \ldots\ 0,\ 0, \ldots, 0,\ 1, \ldots, 1\ );\\
\mathbf{a}_1 &= (\ 1,\ 0,\ 0,\ \ldots\ 0,\ 0, \ldots, 0,\ 1, \ldots, 1\ );\\
\mathbf{a}_2 &= (\ 1,\ 1,\ 0,\ \ldots\ 0,\ 0, \ldots, 0,\ 1, \ldots, 1\ );\\
\mathbf{a}_3 &= (\ 1,\ 1,\ 1,\ \ldots\ 0,\ 0, \ldots, 0,\ 1, \ldots, 1\ );\\
&\ \ldots \qquad\quad \ldots \qquad \ldots \qquad \ldots\\
\mathbf{a}_D &= (\ 1,\ 1,\ 1,\ \ldots\ 1,\ 0, \ldots, 0,\ 1, \ldots, 1\ ).
\end{aligned}$$

There are three possible cases for the predictor a_d.

1. If $k < d$, then the number of errors made by a_d on the objects $\{x_1, \ldots, x_d\}$ is greater than the length of the test sample. A part of errors will certainly fall into the training subsample X, and the algorithm μ chooses another predictor. In this case,

$$[\mu X = a_d] = 0.$$

2. If $d = D < k$, then the algorithm μ chooses the worst predictor in the chain a_D if and only if all the objects $\{x_1, \ldots, x_D\}$ are contained in the test subsample $\bar{X}$. In this case,

$$[\mu X = a_d] = [x_1, \ldots, x_D \in \bar{X}].$$

3. In all the other cases, the algorithm μ chooses the predictor a_d only if all the objects $\{x_1, \ldots, x_d\}$ are contained in the test subsample $\bar{X}$, while the object x_{d+1} is contained in the training subsample X. In this case,

$$[\mu X = a_d] = [x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}].$$

Now we can apply Theorem 1.

If $D \geq k$, then the predictor a_d corresponds to the following set of parameters (to simplify the notation, we will use single subscripts (L_d) instead of double ones (L_{a_d})): $L_d = L - d - 1$, $l_d = l - 1$, $m_d = m + d - d = m$, and $s_d(\varepsilon) = \frac{l}{L}(m + d - \varepsilon k)$. Hence we obtain the assertion of the theorem in the case of $D \geq k$.

If $D < k$, then the predictors $a_0, \ldots a_{D-1}$ have the same values of the parameters as for $D \geq k$. For the worst predictor a_D, only the parameter $l_D = l$ is different. Hence, we obtain the assertion of the theorem in the case of $D < k$.

Remark 6. During the proof of the theorem, it is useful to check if the probabilities P_d are calculated correctly and their sum is one. In the cases of $D \geq k$ and $D < k$, this verification is made somewhat differently using well-known combinatorial identities.

We compared the exact bound from Theorem 11 with the result of empirical measurement of Q_ε by the Monte Carlo method using $N = 1000$ random partitions. The experimental results shown in Fig. 3 are obtained for $l = k = 100$ and $m = 20$, i.e., in the case

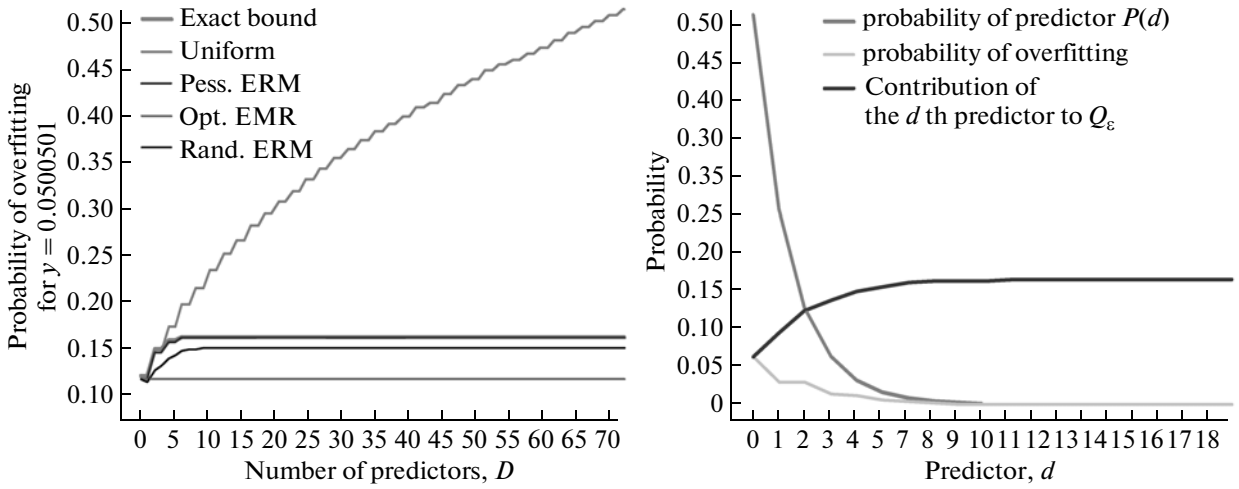

Fig. 4. (left) Bounds for the probability of overfitting Q_ε as a function of the number D of predictors in the chain. (right) Probability of each predictor $P_d = \mathrm{P}[\mu X = a_d]$, the contribution of each predictor to the probability of overfitting Q_ε, and the value of Q_ε for the pessimistic ERM over the set of predictors $\{a_0, \ldots, a_d\}$ as a function of the number d of predictors. All the charts are plotted for $l = k = 100$, $m = 20$, and $\varepsilon = 0.05$.

when the best predictor in the chain makes 10% of errors on the general sample. The optimistic ERM algorithm yields a noticeably underestimated bound for Q_ε, whereas the pessimistic and randomized bounds are very close (see the left-hand diagram in Fig. 3). This means that, for the given set, the pessimistic bound is very tight and is "more reasonable" than the optimistic one.

In this experiment, we also estimated the probability of large uniform deviation of error rates, which underlies many generalization bounds, e.g., VC bounds [6]:

$$R_\varepsilon = \mathrm{P}[\max_{a \in A} \delta(a, X, \bar{X}) \geq \varepsilon].$$

It is obvious that this functional provides an overestimated upper bound for the probability of overfitting, $Q_\varepsilon \leq R_\varepsilon$. The right-hand chart in Fig. 3 shows that the uniform functional R_ε may give a bound overestimated hundreds of times for the probability of overfitting.

The left-hand chart in Fig. 4 shows that the functional R_ε continues to grow with the number of predictors in a monotonic chain, whereas the probability of overfitting Q_ε reaches a horizontal asymptote after 5–8 predictors. Thus, the uniform convergence principle, which was originally introduced in the VC theory [18, 17] and is widely used in statistical learning theory, may give highly overestimated bounds for split sets of predictors.

A term in the sum over all $a \in A$ in formula (7) is called a *contribution $Q_\varepsilon(a)$ of a predictor a* to the probability of overfitting Q_ε. The right-hand chart in Fig. 4 shows that only predictors of 5–8 lower layers make a significant contribution to the probability of overfitting. Apparently, a similar result is characteristic not only of monotonic chains but also of any sets of predictors when the *splitting effect* takes place.

The main conclusion is that a monotonic chain of predictors is hardly overfitted. This fact can serve as a basis for the procedures of one-dimensional optimization, which are frequently used in machine learning for choosing a certain critical parameter in hold-out model selection, for example, a regularization constant or the width of a smoothing window.

10. A UNIMODAL CHAIN OF PREDICTORS

A unimodal chain of predictors is a more realistic model of a one-parameter connected set of predictors compared with a monotonic chain. Here it is assumed that the deviation of a real parameter to either greater or smaller values from its optimal value, corresponding to the best predictor a_0, leads to an increase in the number of errors.

A set of predictors $a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_D$ is called a *unimodal chain* if the *left branch* $a_0, a_1, \ldots, a_D$ and the *right branch* $a_0, a'_1, \ldots, a'_{D'}$ are monotonic chains. The predictor a_0 is said to be *the best in the unimodal chain*. Denote by $m = n(a_0, \mathbb{X})$ the number of errors of the best predictor.

Example 3 (continuation of Example 2). Suppose that a set of objects $\mathbb{X} \subset \mathbb{R}^n$ is linearly separable; i.e., there exists a linear classifier $a(x, w^*)$ with parameter $w^* \in \mathbb{R}^n$ that makes no errors on $\mathbb{X}$. Then, the set of classifiers $\{a(x, w^* + t\delta)\colon t \in \mathbb{R}\}$ forms a unimodal chain for almost any direction vector $\delta \in \mathbb{R}^n$.

Consider a unimodal chain with branches of equal length, $D = D'$. Renumber the objects so that each pre-

dictor a_d, $d = 1, \ldots, D$ makes an error on the objects $x_1, \ldots, x_d$ and each predictor a'_d, $d = 1, \ldots, D$ makes an error on the objects $x'_1, \ldots, x'_d$. Assume that the sets of objects $\{x_1, \ldots, x_D\}$ and $\{x'_1, \ldots, x'_D\}$ are disjoint. It is obvious that the best predictor a_0 makes no error on any of these objects. The numbering of other objects does not matter because the predictors are indistinguishable on these objects.

We will assume that if the empirical risk attains its minimum on several predictors with the same number of errors on both the training and general samples, then the algorithm μ chooses a predictor from the left branch.

Theorem 12. *Let $A = \{a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_D\}$ be a unimodal chain, $k \le D$, and $2D + m \le L$. Then the probability to obtain each predictor of the chain as a result of training is*

$$P_0 = \mathrm{P}[\mu X = a_0] = \frac{C^{l-2}_{L-2}}{C^l_L},$$

$$P_d = \mathrm{P}[\mu X = a_d] = \frac{C^{l-1}_{L-d-1} - C^{l-1}_{L-2d-2}}{C^l_L},$$

$$P'_d = \mathrm{P}[\mu X = a'_d] = \frac{C^{l-1}_{L-d-1} - C^{l-1}_{L-2d-1}}{C^l_L};$$

the probability of overfitting for $s_d(\varepsilon) = \frac{l}{L}(m + d - \varepsilon k)$ is expressed as

$$Q_\varepsilon = \frac{C^{l-2}_{L-2}}{C^l_L} H^{l-2,m}_{L-2}(s_0(\varepsilon)) + \sum_{d=1}^{k}\left(2\frac{C^{l-1}_{L-d-1}}{C^l_L} H^{l-1,m}_{L-d-1}(s_d(\varepsilon)) - \frac{C^{l-1}_{L-2d-2}}{C^l_L} H^{l-1,m}_{L-2d-2}(s_d(\varepsilon)) - \frac{C^{l-1}_{L-2d-1}}{C^l_L} H^{l-1,m}_{L-2d-1}(s_d(\varepsilon))\right).$$

Proof. Introduce auxiliary variables:

$$\beta_d = [x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}], \quad d = 1, \ldots, D-1,$$

$$\beta_D = [x_1, \ldots, x_D \in \bar{X}],$$

$$\beta'_d = [x'_{d+1} \in X][x'_1, \ldots, x'_d \in \bar{X}], \quad d = 1, \ldots, D-1,$$

$$\beta'_D = [x'_1, \ldots, x'_D \in \bar{X}].$$

The conditions $\beta_1, \ldots, \beta_D$ are mutually exclusive; moreover, one of them is valid if and only if $x_1 \in \bar{X}$. Hence,

$$[x_1 \in X] + \beta_1 + \ldots + \beta_D = 1.$$

Analogously,

$$[x'_1 \in X] + \beta'_1 + \ldots + \beta'_D = 1.$$

If the left and right branches were considered as separate monotonic chains, then one could assert that $[\mu X = a_d] = \beta_d$ and $[\mu X = a'_d] = \beta'_d$. However, in the case of a unimodal chain, the conditions for obtaining the predictors a_d and a'_d have a more complicated form. If the condition β_d and simultaneously one of the conditions $\beta'_{d+1} \ldots, \beta'_D$ are satisfied, then the algorithm μ chooses one of the predictors $a'_{d+1}, \ldots, a'_D$ from the right branch according to the convention that the algorithm should choose the worst predictor among all those that make the minimal number of errors on X. Similarly, if the condition β'_d and simultaneously one of the conditions $\beta_d, \ldots, \beta_D$ are satisfied, then the algorithm μ chooses one of the predictors $a_d \ldots, a_D$ from the left branch. Notice that the predictors of the left branch have priority. Thus, the conditions for obtaining all the predictors of the unimodal chain are expressed in terms of auxiliary variables as follows:

$$[\mu X = a_0] = [x_1, x'_1 \in X] = (1 - \beta_1 - \ldots - \beta_D)(1 - \beta'_1 - \ldots - \beta'_D),$$

$$[\mu X = a_d] = \beta_d(1 - \beta'_{d+1} - \ldots - \beta'_D), \quad d = 1, \ldots, D-1,$$

$$[\mu X = a'_d] = \beta'_d(1 - \beta_d - \ldots - \beta_D), \quad d = 1, \ldots, D-1,$$

$$[\mu X = a_D] = \beta_D,$$

$$[\mu X = a'_D] = \beta'_D(1 - \beta_D).$$

Let us determine the probabilities of all the predictors of the chain by applying Theorem 2.

$$P_0 = \mathrm{P}[\mu X = a_0] = \mathrm{P}[x_1, x'_1 \in X] = \frac{C^{l-2}_{L-2}}{C^l_L},$$

$$P_d = \mathrm{P}[\mu X = a_d] = \mathrm{P}[x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}] - \sum_{t=d+1}^{k-d} \mathrm{P}[x_{d+1}, x'_{t+1} \in X][x_1, \ldots, x_d, x'_1, \ldots, x'_t \in \bar{X}]$$

$$= \frac{1}{C^l_L}\left(C^{l-1}_{L-d-1} - \sum_{t=d+1}^{k-d} C^{l-2}_{L-d-t-2}\right) = \frac{C^{l-1}_{L-d-1} - C^{l-1}_{L-2d-2}}{C^l_L},$$

$$P'_d = \mathrm{P}[\mu X = a'_d] = \mathrm{P}[x'_{d+1} \in X][x'_1, \ldots, x'_d \in \bar{X}] - \sum_{t=d}^{k-d} \mathrm{P}[x'_{d+1}, x_{t+1} \in X][x'_1, \ldots, x'_d, x_1, \ldots, x_t \in \bar{X}]$$

$$= \frac{1}{C_L^l}\left(C_{L-d-1}^{l-1} - \sum_{t=d}^{k-d} C_{L-d-t-2}^{l-2}\right) = \frac{C_{L-d-1}^{l-1} - C_{L-2d-1}^{l-1}}{C_L^l}.$$

Now, let us write the probability of overfitting using Theorem 2.

$$Q_\varepsilon = \frac{C_{L-2}^{l-2}}{C_L^l} H_{L-2}^{l-2,m}(s_0(\varepsilon)) + \sum_{d=1}^{k}\left(\frac{C_{L-d-1}^{l-1}}{C_L^l} H_{L-d-1}^{l-1,m}(s_d(\varepsilon)) - \sum_{t=d+1}^{k-d} \frac{C_{L-d-t-2}^{l-2}}{C_L^l} H_{L-d-t-2}^{l-2,m}(s_d(\varepsilon))\right)$$

$$+ \sum_{d=1}^{k}\left(\frac{C_{L-d-1}^{l-1}}{C_L^l} H_{L-d-1}^{l-1,m}(s_d(\varepsilon)) - \sum_{t=d}^{k-d} \frac{C_{L-d-t-2}^{l-2}}{C_L^l} H_{L-d-t-2}^{l-2,m}(s_d(\varepsilon))\right).$$

This expression can be simplified if we notice that

$$\sum_{t=d+1}^{k-d} \frac{C_{L-d-t-2}^{l-2}}{C_L^l} H_{L-d-t-2}^{l-2,m}(s_d(\varepsilon)) = \frac{C_{L-2d-2}^{l-1}}{C_L^l} H_{L-2d-2}^{l-1,m}(s_d(\varepsilon)),$$

$$\sum_{t=d}^{k-d} \frac{C_{L-d-t-2}^{l-2}}{C_L^l} H_{L-d-t-2}^{l-2,m}(s_d(\varepsilon)) = \frac{C_{L-2d-1}^{l-1}}{C_L^l} H_{L-2d-1}^{l-1,m}(s_d(\varepsilon)).$$

Substituting these expressions into the formula for Q_ε, we obtain the required bound.

A natural generalization of monotonic and unimodal chains of predictors are multidimensional monotonic and unimodal grids of predictors. They model multidimensional parametric sets of predictors with splitting and similarity. Note that exact bounds for the probability of overfitting for h-dimensional monotonic and unimodal grids were obtained by Botov in [1]. Another multidimensional generalization—pencils of h monotonic chains—was considered by Frey in [4].

11. UNIT NEIGHBORHOOD OF THE BEST PREDICTOR

Another example of a connected set is given by a unit neighborhood of the best predictor. This is an extreme particular case when predictors are maximally close to each other, and the classical bounds based on counting the number of different predictors are highly overestimated. Moreover, this is a set of predictors that form *two lower layers* in an arbitrary connected set with a single best predictor.

A set of predictors $A = \{a_0, a_1, ..., a_D\}$ is called a *unit neighborhood* of the predictor a_0 if all error vectors a_d are pairwise distinct, $n(a_d, \mathbb{X}) = n(a_0, \mathbb{X}) + 1$, and $\rho(a_0, a_d) = 1$ for any $d = 1, ..., D$. The predictor a_0 is said to be *the best in the neighborhood*, or the *center of the neighborhood*.

Assume that if the empirical risk attains its minimum on several predictors with the same number of errors on both the training and general samples, then the algorithm μ chooses a predictor with a smaller number.

Theorem 13. *Let $A = \{a_0, a_1, ..., a_D\}$ be a unit neighborhood of the predictor a_0, $m = n(a_0, \mathbb{X})$, and $L \geq m + D$. Then*

$$Q_\varepsilon = P_0 H_{L-D}^{l-D,m}\left(\frac{l}{L}(m - \varepsilon k)\right) + \sum_{d=1}^{D} P_d H_{L-d}^{l-d+1,m}\left(\frac{l}{L}(m + 1 - \varepsilon k)\right),$$

$$P_0 = \frac{C_{L-D}^k}{C_L^k}, \quad P_d = \frac{C_{L-d}^{k-1}}{C_L^k}, \quad d = 1, \ldots, D,$$

where P_d is the probability to obtain a predictor a_d as a result of learning.

Proof. Let us renumber the objects so that each predictor a_d, $d = 1, ..., D$ makes an error on the object x_d. Obviously, the best predictor a_0 makes no error on any of these objects. The numbering of other objects does not matter because the predictors are indistinguishable on these objects.

For the sake of clarity, we partition the sample $\mathbb{X}$ into three blocks:

$$\begin{array}{lccccccccccc} & x_1 & x_2 & x_3 & & x_D & & & & \overbrace{\quad}^{m} & & \\ \mathbf{a}_0 = (& 0, & 0, & 0, & \ldots & 0, & 0, & \ldots, & 0, & 1, & \ldots, & 1); \\ \mathbf{a}_1 = (& 1, & 0, & 0, & \ldots & 0, & 0, & \ldots, & 0, & 1, & \ldots, & 1); \\ \mathbf{a}_2 = (& 0, & 1, & 0, & \ldots & 0, & 0, & \ldots, & 0, & 1, & \ldots, & 1); \\ \mathbf{a}_3 = (& 0, & 0, & 1, & \ldots & 0, & 0, & \ldots, & 0, & 1, & \ldots, & 1); \\ \ldots & & & & \ldots & & & \ldots & & \ldots & & \\ \mathbf{a}_D = (& 0, & 0, & 0, & \ldots & 1, & 0, & \ldots, & 0, & 1, & \ldots, & 1). \end{array}$$

It is easily seen that the set of partitions for which the algorithm μ chooses a predictor a_d is represented as

$$[\mu X = a_0] = [x_1, \ldots, x_D \in X],$$

$$[\mu X = a_d] = [x_1, \ldots, x_{d-1} \in X][x_d \in \overline{X}], \quad d = 1, \ldots, D.$$

The parameters that should be substituted into the formula of Theorem 1 are as follows:

$$L_0 = L - D, \quad l_0 = l - D, \quad m_0 = m,$$

$$s_0(\varepsilon) = \frac{l}{L}(m - \varepsilon k),$$

$$L_d = L - d, \quad l_d = l - d + 1, \quad m_d = m,$$

$$s_d(\varepsilon) = \frac{l}{L}(m + 1 - \varepsilon k), \quad d = 1, \ldots, D.$$

Substituting these parameters into the formula of Theorem 1, we obtain the bound required.

Remark 7. We can easily verify that the probabilities P_d are determined correctly.

$$\sum_{d=0}^{D} P_d = \frac{1}{C_L^k}(C_{L-D}^k + \underbrace{C_{L-D}^{k-1} + \ldots + C_{L-1}^{k-1}}_{C_L^k - C_{L-D}^k}) = 1.$$

CONCLUSIONS

In this paper, we have proposed three general methods for obtaining exact bounds for the probability of overfitting. To illustrate the application of these methods, we considered six model sets of predictors: a pair of predictors, a layer and an interval of a Boolean cube, a monotonic and a unimodal chain, and a unit neighborhood. For the interval and the monotonic chain, we presented the results of numerical experiments that illustrate the effects of splitting and similarity on the probability of overfitting.

A penalty for the accuracy of bounds has two drawbacks, the elimination of which still remains an open problem.

First, to date, exact bounds have been obtained only for a number of artificial cases. The model sets of predictors are defined directly by their error matrices, regardless of any applied problem or any practical set of predictors. It seems reasonable to assume that a gradual generalization of these models will make it possible to analyze the probability of overfitting as a function of the dimensional characteristics of the lowest layers in predictor sets and then adapt these results to practical situations. This approach to the development of combinatorial learning theory seems to be the most realistic.

Second, the bounds obtained are unobserved ones; i.e., they depend on the hidden testing subsample of the general sample. Examples of transition from unobserved bounds to observed ones (which are calculated only using the training sample) can be found in [11, 10]. It is reasonable to assume that a similar approach can also be applied to combinatorial bounds. We did not consider this problem in the present study.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 08-07-00422) and by the Program "Algebraic and Combinatorial Methods in Mathematical Cybernetics" of the Department of Mathematics, Russian Academy of Sciences.

REFERENCES

1. P. V. Botov, "Exact Bounds for the Probability of Overfitting for Monotone and Unimodal Sets of Predictors," in *Proceedings of the 14th Russian Conference on Mathematical Methods of Pattern Recognition* (MAKS Press, Moscow, 2009), pp. 7–10.
2. K. V. Vorontsov, "Combinatorial Approach to Estimating the Quality of Learning Algorithms," in *Mathematics Problems of Cybernetics* Ed. by O.B. Lupanov (Fizmatlit, Moscow, 2004), Vol. 13, pp. 5–36.
3. D. A. Kochedykov, "Similarity Structures in Sets of Classifiers and Generalization Bounds," in *Proceedings of the 14th Russian Conference on Mathematical Methods of Pattern Recognition* (MAKS Press, Moscow, 2009), pp. 45–48.
4. A. I. Frey, "Exact Bounds for the Probability of Overfitting for Symmetric Sets of Predictors," in *Proceedings of the 14th Russian Conference on Mathematical Methods of Pattern Recognition* (MAKS Press, Moscow, 2009), pp. 66–69.
5. E. T. Bax, "Similar Predictors and VC Error Bounds," Tech. Rep. CalTech-CS-TR97-14: 6 1997.
6. S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of Classification: A Survey of Some Recent Advances," *ESIAM: Probab. Stat.*, No. 9, 323–375 (2005).
7. R. Herbrich and R. Williamson, "Algorithmic Luckiness," *J. Machine Learning Res.,* No. 3, 175–212 (2002).
8. V. Koltchinskii, "Rademacher Penalties and Structural Risk Minimization," *IEEE Trans. Inf. Theory* **47** (5) 1902–1914 (2001).
9. V. Koltchinskii and D. Panchenko, "Rademacher Processes and Bounding the Risk of Function Learning," in *High Dimensional Probability, II,* Ed. by D.E. Gine and J Wellner (Birkhauser, 1999) pp. 443–457.
10. J. Langford, "Quantitatively Tight Sample Complexity Bounds," Ph.D. Thesis (Carnegie Mellon Thesis, 2002).
11. J. Langford and D. McAllester, "Computable Shell Decomposition Bounds," in *Proceedings of the 13th Annual Conference on Computer Learning Theory* (Morgan Kaufmann, San Francisco, CA, 2000), pp. 25–34.
12. J. Langford and J. Shawe-Taylor, "PAC-Bayes and Margins," in *Advances in Neural Information Processing Systems 15* (MIT Press, 2002), pp. 439–446.
13. D. McAllester, "PAC-Bayesian Model Averaging," in *COLT: Proceedings of the Workshop on Computational Learning Theory* (Morgan Kaufmann, San Francisco, CA, 1999).
14. P. Philips, "Data-Dependent Analysis of Learning Systems," Ph.D. Thesis (The Australian National University, Canberra, 2005).

15. J. Sill, "Monotonicity and Connectedness in Learning Systems," Ph.D. Thesis (California Inst. Technol., 1998).

16. V. Vapnik, *Estimation of Dependencies Based on Empirical Data* (Springer, New York, 1982).

17. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

18. V. Vapnik and A. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory Probab. Its. Appl.* **16** (2), 264–280 (1971).

19. N. Vayatis and R. Azencott, "Distribution-dependent Vapnik–Chervonenkis Bounds," *Lecture Notes in Computer Science* **1572** 230–240 (1999).

20. K. V. Vorontsov, "Combinatorial Probability and the Tightness of Generalization Bounds," *Pattern Recognit. Image Anal.* **18** (2), 243–259 (2008).

21. K. V. Vorontsov, "On the Influence of Similarity of Classifiers on the Probability of Overfitting," in *Pattern Recognition and Image Analysis: New Information Technologies (PRIA-9)* (Nizhni Novgorod, 2008), Vol. 2, pp. 303–306.

22. K. V. Vorontsov, "Splitting and Similarity Phenomena in the Sets of Classifiers and Their Effect on the Probability of Overfitting," *Pattern Recognit. Image Anal.* **19** (3), 412–420 (2009).

23. K. V. Vorontsov, "Tight Bounds for the Probability of Overfitting," *Dokl. Math.* **80** (3) 793–796 (2009).

Konstantin Vorontsov. Born 1971. Graduated from the Faculty of Applied Mathematics and Control, Moscow Institute of Physics and Technology, in 1994. Received candidate's degree in 1999 and doctoral degree in 2010. Currently is with the Dorodnicyn Computing Centre, Russian Academy of Sciences. Scientific interests: statistical learning theory, machine learning, data mining, probability theory, and combinatorics. Author of 75 papers. Homepage: www.ccas.ru/voron.