Strong minimax lower bounds for learning *

András Antos Gábor Lugosi

December, 1995

*** *Extended abstract* ***

Abstract

Known minimax lower bounds for learning state that for each sample size n, and learning rule g_n, there exists a distribution of the observation X and a concept C to be learnt such that the expected error of g_n is at least a constant times V/n, where V is the VC dimension of the concept class. However, these bounds do no tell anything about the rate of decrease of the error for a *fixed* distribution and fixed concept.

In this paper we investigate minimax lower bounds in such—much stronger—sense. We show that for several natural d-parameter concept classes, for any *sequence* of learning rules $\{g_n\}$, there exists a fixed distribution of X and a fixed concept C such that the expected error is larger than a constant times d/n for *infinitely many* n.

We also show that it is neither the VC dimension, nor the rate of increase of the shatter coefficients of the class that determine the asymptotic behavior of the concept class.

*Department of Mathematics and Computer Science, Faculty of Electrical Engineering, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary. (email: `antos@inf.bme.hu` and `lugosi@vma.bme.edu`). The work of the second author was supported by ...

1 Introduction

Let X be a random variable on a domain $\mathcal{X}$, with distribution μ, i.e., for each measurable subset A of $\mathcal{X}$, $\mu(A) = \mathbf{P}\{X \in A\}$. Let $\mathcal{C}$ be a class of subsets of $\mathcal{X}$. Members of $\mathcal{C}$ are called concepts, and $\mathcal{C}$ is a concept class. A fixed, but unknown concept $C \in \mathcal{C}$ is to be learnt based on the data

$$D_n = ((X_1, I_{\{X_1 \in C\}}), \ldots, (X_n, I_{\{X_n \in C\}})),$$

where $X_1, \ldots, X_n$ are independent, identically distributed copies of X, and I_A denotes the indicator of an event A. A learning rule—or classifier—intends to decide, based on the data D_n and X, if $X \in C$. Formally, it is a function $g_n : \mathcal{X} \times (\mathcal{X} \times \{0,1\})^n \to \{0,1\}$, whose probability of error is the random variable

$$L(g_n) = \mathbf{P}\{g_n(X, D_n) \neq I_{\{X \in C\}} | D_n\}.$$

The joint distribution of the pair $(X, I_{\{X \in C\}})$ is determined by the pair (μ, C), which, therefore, will shortly be referred to as *distribution*.

Haussler, Littlestone, and Warmuth [6] showed that there exists a learning rule such that for all distributions,

$$\mathbf{E}\{L(g_n)\} \leq \frac{2V}{n},$$

where V is the VC dimension of the class $\mathcal{C}$. Minimax lower bounds show that, in a sense, this is the smallest possible distribution-free upper bound obtainable for any learning function. For example, Vapnik and Chervonenkis [8] showed that for every n, and every classifier g_n, there exists a distribution such that

$$\mathbf{E}\{L(g_n)\} \geq \frac{V-1}{2en}\left(1 - \frac{1}{n}\right).$$

For related minimax lower bounds, see Haussler, Littlestone, and Warmuth [6], Ehrenfeucht, Haussler, Kearns, and Valiant [5], Simon [7], and Devroye and Lugosi [4].

However, in some sense, these lower bounds are not satisfactory. They do not tell us anything about the way the error decreases as the sample size is increased for a given classification problem. These bounds, for each n, give information about the maximal error within the class, but not about the behavior of the error for a single fixed distribution as the sample size n increases. For example, the above-mentioned lower bound does not exclude the possibility that there exists a classifier g_n such that for every μ and C the expected error decreases exponentially. Indeed, we will see that such classes exist with arbitrarily large VC dimension, and also with arbitrarily large Assouad density.

The purpose of this paper is to establish minimax lower bounds (matching the corresponding upper bounds) in the described strong sense. This is not possible for all VC classes. We introduce a general method to obtain such "strong" minimax lower bounds. As applications of the general principle, we give some general sufficient conditions for classes that have such strong minimax lower bounds. Two typical examples of our results along this line are the following:

Proposition 1 *Let V be an arbitrary positive integer. Then there exists a class $\mathcal{C}$ with VC dimension V and a corresponding sequence of learning rules g_n such that for all distributions (μ, C) with $C \in \mathcal{C}$ and for all n,*

$$\mathbf{E}\{L(g_n)\} \leq a \cdot b^n,$$

where $b < 1$. The positive constants a and b may depend on the distribution.

Thus, the error decreases at an exponential rate for all distributions, and it is not possible to have a "strong" version of the above-mentioned lower bound for all VC classes. However, for many "natural" classes, the error indeed behaves the way it is expected, i.e., as $\Omega(d/n)$ for some distribution, where d is the "dimension" of the class. For example, we will see the following:

Corollary 1 *Let $\mathcal{X} = \mathcal{R}^d$, and let $\mathcal{C}$ contain all closed halfspaces. Then for any sequence of classifiers $\{g_n\}$, there exists a distribution (μ, C) such that*

$$\mathbf{E}\{L(g_n)\} \geq \frac{(1-\epsilon)d}{2n}$$

for infinitely many n, where ϵ is any fixed small positive number.

2 Negative examples

Since this is a paper on lower bounds, paradoxically, by a negative example we mean that there is no large lower bound, i.e., that there exists a learning rule with rapidly decreasing error. This section is devoted to such examples of concept classes.

As the simplest example, consider any class $\mathcal{C}$ containing finitely many concepts. Then a learning rule that selects any concept from $\mathcal{C}$ which is consistent with the data D_n can only make an error if there is at least one concept $C' \in \mathcal{C}$ with $\mu(C' \triangle C) > 0$ (where $C \in \mathcal{C}$ is the true concept) such that no one of $X_1, \ldots, X_n$ falls in the symmetric difference of C' and C. The probability of this event is at most

$$\sum_{i=1}^{|\mathcal{C}|} (1 - \mu(C_i \triangle C))^n \, I_{\{\mu(C_i \triangle C) > 0\}} \leq |\mathcal{C}| \max_{C' \in \mathcal{C}: \mu(C' \triangle C) > 0} (1 - \mu(C' \triangle C))^n,$$

which converges to zero exponentially rapidly. Since a finite concept class can have an arbitrary VC dimension, this proves Proposition 1.

If a concept class $\mathcal{C}$ is finite, its n-th shatter coefficient $s(n)$ (i.e., the Vapnik-Chervonenkis growth function at n) is bounded above by $2^{|\mathcal{C}|}$ for all n, that is, the shatter coefficients do not increase with n for large n. In such cases it is not surprising that the error can decrease at an exponential rate for all distributions. It is natural to ask if the growth of $s(n)$ determines the rate of convergence of the error. For example, one might conjecture, that if d is the Assouad density of $\mathcal{C}$, then the error of all learning rules behave as $\Omega(d/n)$ for some fixed distribution. Recall that the Assouad density of $\mathcal{C}$ is defined by

$$\inf\left\{r > 0 : \sup_n \frac{s(n)}{n^r} < \infty\right\}$$

(Assouad [1]). This conjecture is false, as the following simple result shows:

Proposition 2 *Let d be an arbitrary positive integer. Then there exists a class $\mathcal{C}$ whose n-th shatter coefficient equals $s(n) = \sum_{i=0}^{d} \binom{n}{i}$ and a corresponding sequence of learning rules g_n such that for all distributions and for all n,*

$$\mathbf{E}\{L(g_n)\} \leq a \cdot b^n,$$

where $b < 1$. The positive constants a and b may depend on the distribution.

Proof. Let $\mathcal{X} = \mathcal{R}$ and let $\mathcal{C}$ contain all d-point subsets of $\mathcal{R}$. A similar argument to the proof of Proposition 1 proves the claim.

3 A general lemma

We prove strong minimax lower bounds using the probabilistic method. The main tool of our analysis is the following lemma:

Lemma 1 *Let $R_n(z)$ be a sequence of real numbers parametrized by an abstract parameter z taking its value from a set $\mathcal{Z}$. If there exists a random variable Z taking its values from $\mathcal{Z}$ such that $\mathbf{E}^2\{R_n(Z)\}/\mathbf{E}\{R_n^2(Z)\}$ does not tend to zero, then for every $0 < \lambda < 1$ there exists a $z \in \mathcal{Z}$ such that*

$$R_n(z) > \lambda \mathbf{E}\{R_n(Z)\}$$

for infinitely many n. □

Proof. Clearly, it suffices to prove that for every λ,

$$\mathbf{P}\left\{\limsup_{n\to\infty}\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\}>0.$$

Now,

$$\begin{aligned}
\mathbf{P}\left\{\limsup_{n\to\infty}\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\} &= \mathbf{E}\left\{I_{\left\{\limsup_{n\to\infty}\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\}}\right\}\\
&\geq \mathbf{E}\left\{\limsup_{n\to\infty} I_{\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\}}\right\}\\
&\geq \limsup_{n\to\infty}\mathbf{E}\left\{I_{\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\}}\right\}\\
&\quad\text{(by Fatou's lemma)}\\
&= \limsup_{n\to\infty}\mathbf{P}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\}.
\end{aligned}$$

Thus, we need to prove that $\mathbf{P}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\}$ does not tend to zero.

$$\begin{aligned}
\mathbf{P}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq\lambda\right\} &= \mathbf{P}\left\{\frac{R_n(Z)-\mathbf{E}R_n(Z)}{\mathbf{E}R_n(Z)}\geq-(1-\lambda)\right\}\\
&= 1-\mathbf{P}\{|R_n(Z)-\mathbf{E}R_n(Z)|>(1-\lambda)|\mathbf{E}R_n(Z)|\}\\
&\geq 1-\frac{\mathbf{Var}\{R_n(Z)\}}{\mathbf{Var}\{R_n(Z)\}+(1-\lambda)^2\mathbf{E}^2\{R_n(Z)\}}\\
&\quad\text{(by the Chebyshev-Cantelli inequality, see, e.g., [2])}\\
&= 1-\frac{1}{1+(1-\lambda)^2\mathbf{E}^2\{R_n(Z)\}/\mathbf{Var}\{R_n\}(Z)}.
\end{aligned}$$

However, the last expression converges to zero if and only if $\mathbf{E}^2\{R_n(Z)\}/\mathbf{E}\{R_n^2(Z)\}$ converges to zero, which case is excluded by assumption. $\square$

4 Strong minimax lower bounds

In this section we provide examples of concept classes for which the known minimax lower bounds can be extended to strong ones. All examples shown here are based on lower bounds obtained for a very simple concept class: the class of the union of a certain number of one-dimensional intervals. More precisely for each positive integer k, define the class $\mathcal{C}_k$ as follows: For each $\mathbf{z}=(z_1,\ldots,z_k)\in\mathcal{R}^k$ let $z_{(1)}\leq\cdots\leq z_{(k)})$ be the components of $\mathbf{z}$ in increasing order. Define

$$\mathcal{C}_k=\begin{cases}\left\{\bigcup_{j=1}^{k/2}(z_{(2j-1)},z_{(2j)}];z\in\mathcal{R}^k\right\} & \text{if } k \text{ is even}\\ \left\{\bigcup_{j=1}^{(k-1)/2}(z_{(2j-1)},z_{(2j)}]\cup(z_{(k)},\infty);z\in\mathcal{R}^k\right\} & \text{if } k \text{ is odd.}\end{cases}$$

Let $C \in \mathcal{C}_k$ denote the "true" concept and recall that $L(g_n) = \mathbf{P}\{g_n(X) \neq I_{\{X \in C\}} | D_n\}$ denotes the error of the learning rule g_n.

Theorem 1 *Let μ be the uniform distribution on $[0,1]$. For every sequence of learning rules $\{g_n\}$ and $0 < \lambda < 1$, there exist a $C \in \mathcal{C}_k$ such that if C is the "true" concept, then*

$$\mathbf{E}\{L(g_n)\} > \lambda \frac{k}{2n}$$

for infinitely many n.

Remark. It is clear from the proof of the theorem that the uniform distribution may be replaced by any nonatomic distribution on the real line.

Sketch of proof. Let $\mathbf{z} \in \mathcal{R}^k$ be the parameter that determines $C \in \mathcal{C}_k$. Define $A_n(\mathbf{z})$ as the event that for some $j \in \{1, \ldots, k\}$, no X_i falls in $(z_j, X]$ or $(X, z_j]$. Let

$$R_n(\mathbf{z}) = \mathbf{P}\{g_n(X, D_n(\mathbf{z})) \neq Y(\mathbf{z}), A_n(\mathbf{z})\},$$

where $Y(\mathbf{z}) = I_{\{X \in C\}}$, $Y_i(\mathbf{z}) = I_{\{X_i \in C\}}$, and $D_n(\mathbf{z}) = ((X_1, Y_1(\mathbf{z})), \ldots, (X_n, Y_n(\mathbf{z})))$. Clearly,

$$\mathbf{E}\{L(g_n)\} \geq R_n(\mathbf{z}).$$

We apply Lemma 1 for $R_n(\mathbf{z})$. We will show that if the random vector $\mathbf{Z} = (Z_1, \ldots, Z_k)$ is uniformly distributed on $[0,1]^k$ and independent of $X, X_1, \ldots, X_n$, then

$$\mathbf{E}\{R_n(\mathbf{Z})\} \geq \frac{k}{2(n+k+1)},$$

and $\mathbf{E}^2\{R_n(\mathbf{Z})\}/\mathbf{E}\{R_n^2(\mathbf{Z})\}$ does not tend to zero, from which the theorem follows.

First we prove the lower bound on the expected value of $R_n(\mathbf{Z})$. By the independence of $\mathbf{Z}$ and $X, X_1, \ldots, X_n$,

$$R_n(\mathbf{Z}) = \mathbf{P}\{g_n(X, D_n(\mathbf{Z})) \neq Y(\mathbf{Z}), A_n(\mathbf{Z}) | \mathbf{Z}\}, \tag{1}$$

and

$$\begin{aligned}
\mathbf{E}\{R_n(\mathbf{Z})\} &= \mathbf{P}\{g_n(X, D_n(\mathbf{Z})) \neq Y(\mathbf{Z}), A_n(\mathbf{Z})\} \\
&= \mathbf{P}\{g_n(X, D_n(\mathbf{Z})) \neq Y(\mathbf{Z}) | A_n(\mathbf{Z})\}\mathbf{P}\{A_n(\mathbf{Z})\}.
\end{aligned}$$

The first term on the right-hand side may be thought of as the conditional error probability of the decision function g_n classifying $Y(\mathbf{Z})$ based on observing $(X, X_1, Y_1(\mathbf{Z}), \ldots, X_n, Y_n(\mathbf{Z}))$.

This error probability may be bounded below by the conditional Bayes-error of this discrimination problem, i.e.,

$$\mathbf{E}\{\min(\eta^*(X, D_n(\mathbf{Z})), 1-\eta^*(X, D_n(\mathbf{Z})))|A_n(\mathbf{Z})\} \tag{2}$$

where

$$\eta^*(X, D_n(\mathbf{Z})) = \mathbf{P}\{Y(\mathbf{Z}) = 1|X, X_1, Y_1(\mathbf{Z}), \ldots, X_n, Y_n(\mathbf{Z})\}$$

(see, e.g., [3]). Let $X_{(-)}$ be the largest X_i not larger than X, and let $X_{(+)}$ be the smallest X_i larger than X. Define the event $B_{j,n}(\mathbf{Z})$ as

$$B_{j,n}(\mathbf{Z}) = \left\{[X_{(-)}, X_{(+)}) \text{ contains exactly } j \text{ of the } Z_i\text{'s}\right\}.$$

Write $\eta^*(X, D_n(\mathbf{Z})) = \sum_{j=1}^k \mathbf{P}\{Y(\mathbf{Z}) = 1|X, D_n(\mathbf{Z}), B_{j,n}(\mathbf{Z})\}\mathbf{P}\{B_{j,n}(\mathbf{Z})|X, D_n(\mathbf{Z})\}$. Then

$$\begin{aligned}
&\min(\eta^*(X, D_n(\mathbf{Z})), 1-\eta^*(X, D_n(\mathbf{Z}))) \\
&\geq \sum_{j=1}^k \mathbf{P}\{B_{j,n}(\mathbf{Z})|X, D_n(\mathbf{Z})\} \\
&\qquad \min\left(\mathbf{P}\{Y(\mathbf{Z}) = 1|X, D_n(\mathbf{Z}), B_{j,n}(\mathbf{Z})\}, 1-\mathbf{P}\{Y(\mathbf{Z}) = 1|X, D_n(\mathbf{Z}), B_{j,n}(\mathbf{Z})\}\right) \\
&= \sum_{j=1}^k \mathbf{P}\{B_{j,n}(\mathbf{Z})|X, D_n(\mathbf{Z})\} \min\left(O_j\left(\frac{\mu([X_{(-)}, X))}{\mu([X_{(-)}, X_{(+)}))}\right), E_j\left(\frac{\mu([X_{(-)}, X))}{\mu([X_{(-)}, X_{(+)}))}\right)\right),
\end{aligned}$$

where

$$\begin{aligned}
O_j(p) &= \sum_{m=0}^{\lfloor (j-1)/2\rfloor} \binom{j}{2m+1} p^{2m+1}(1-p)^{j-2m-1} \\
E_j(p) &= \sum_{m=0}^{\lfloor j/2\rfloor} \binom{j}{2m} p^{2m}(1-p)^{j-2m}
\end{aligned}$$

are the probabilities that the value of a binomial (j, p) random variable is odd, and even, respectively. The above identity follows by observing that

$$Y(\mathbf{Z}) = I_{\{\text{there is an odd number of } Z_j\text{'s smaller than } X\}}$$

with probability one and a similar expression is valid for $Y_i(\mathbf{Z})$ for $i = 1, \ldots, n$.

By simple calculation, $O_j(p) + E_j(p) = 1$ and $E_j(p) - O_j(p) = (2p-1)^j$, therefore $\min(O_j(p), E_j(p)) = \left(1 - |2p-1|^j\right)/2$. Thus, using the abbreviation $p = \frac{\mu([X_{(-)}, X))}{\mu([X_{(-)}, X_{(+)}))}$, (2) may be bounded from below by

$$\begin{aligned}
&\mathbf{E}\{\min(\eta^*(X, D_n(\mathbf{Z})), 1-\eta^*(X, D_n(\mathbf{Z})))|A_n(\mathbf{Z})\} \\
&\geq \sum_{j=1}^k \mathbf{E}\left\{\mathbf{P}\{B_{j,n}(\mathbf{Z})|X, D_n(\mathbf{Z})\}\frac{1-|2p-1|^j}{2}\middle| A_n(\mathbf{Z})\right\} \\
&= \sum_{j=1}^k \mathbf{P}\{B_{j,n}(\mathbf{Z})|A_n(\mathbf{Z})\}\mathbf{E}\left\{\frac{1-|2p-1|^j}{2}\middle| A_n(\mathbf{Z})\right\},
\end{aligned}$$

by the conditional independence of $B_{j,n}(\mathbf{Z})$ and p (which can be verified by straightforward calculation). Also by elementary calculations, $\mathbf{E}\left\{(1-|2p-1|^j)/2|\, A_n(\mathbf{Z})\right\} = j/(2j+2)$. Therefore, we have that

$$\begin{aligned}
\mathbf{E}\{R_n(\mathbf{Z})\} &\geq \mathbf{E}\{\min(\eta^*(X,D_n(\mathbf{Z})),1-\eta^*(X,D_n(\mathbf{Z})))|A_n(\mathbf{Z})\}\mathbf{P}\{A_n(\mathbf{Z})\} \\
&\geq \sum_{j=1}^{k} \frac{j}{2j+2}\mathbf{P}\{B_{j,n}(\mathbf{Z})\} \\
&\quad \text{(since } B_{j,n}(\mathbf{Z}) \subset A_n(\mathbf{Z})\text{)} \\
&= \sum_{j=1}^{k} \frac{j}{2j+2} \cdot \frac{\binom{k}{j} n(n+1)}{\binom{n+k}{j+1}(n+k+1)} \\
&= \frac{k}{2(n+k+1)},
\end{aligned}$$

which completes the proof of the desired lower bound on $\mathbf{E}\{R_n(\mathbf{Z})\}$. To finish the proof of the theorem, we need to prove that $\mathbf{E}^2\{R_n(\mathbf{Z})\}/\mathbf{E}\{R_n^2(\mathbf{Z})\}$ does not tend to zero. To see this, simply observe, that for each fixed $\mathbf{z} \in \mathcal{R}^k$, $\mathbf{P}\{A_n(\mathbf{z})\} \leq 2k/(n+1)$. Thus,

$$\mathbf{E}\{R_n^2(\mathbf{Z})\} \leq \mathbf{E}\left\{\mathbf{P}^2\{A_n(\mathbf{Z})|\mathbf{Z}\}\right\} = \int_{\mathcal{R}^k} \mathbf{P}^2\{A_n(\mathbf{z})\}\mu(d\mathbf{z}) \leq \frac{4k^2}{(n+1)^2},$$

which, combined with the obtained lower bound on $\mathbf{E}\{R_n(\mathbf{Z})\}$, implies the desired stability property. □

Now, we are prepared to prove Corollary 1:

Proof of Corollary 1. The result follows by Theorem 1. We define a curve in $\mathcal{R}^d$ such that the class of sets obtained by intersecting the curve with all halfspaces of $\mathcal{R}^d$ is the union of at most $d/2$ subintervals of the curve if d is even, and for d odd, it is the union of $(d-1)/2$ intervals and one halfline. Thus, if the distribution μ of X is concentrated on this curve, then the behavior of the error is identical to that for the class $\mathcal{C}_k$ of Theorem 1, which implies the corollary.

The class of halfspaces of $\mathcal{R}^d$ contains all sets of the form

$$H = \left\{x : \sum_{i=1}^{d} a_i x_i + a_0 \geq 0\right\}; \quad a_0, a_1, \ldots, a_d \in \mathcal{R},$$

where $x_1, \ldots, x_d$ are the components of the vector x. The above-mentioned curve (i.e., subset of $\mathcal{R}^d$) is given parametrically by

$$S = \{x = x(t) = (x_1(t), x_2(t), \ldots, x_d(t)),\ t \in [0,1]\},$$

where $x_i(t) = t^i$. The intersection of the curve and a halfspace is the set of those $x(t) \in S$ for which the parameter t satisfies $a_d t^d + a_{d-1} t^{d-1} + \cdots + a_1 t + a_0 \geq 0$. Clearly, given d numbers in $[0,1]$, there exists a polynomial of degree d (i.e., coefficients $a_0, \ldots, a_d$) such that the roots of the polynomial are exactly these numbers. Thus, the class of subsets of $[0,1]$ defined by the sets of parameters corresponding to sets of the form $S \cap H$ is a subset of the class $\mathcal{C}_k$ of Theorem 1. $\square$

The method of the proof of Corollary 1 may be used to obtain similar lower bounds for other concept classes as well. One such example is the class of subsets of $\mathcal{R}^d$ of the form

$$\left\{ x : \prod_{i=1}^{d} (x_i - a_i) \geq 0 \right\}, \quad a_1, \ldots, a_d \in \mathcal{R}.$$

In this case the distribution of X might be concentrated on the straight line segment

$$\{ x = x(t) = (t, t, \ldots, t); t \in [0,1] \}.$$

A similar argument to the proof of Corollary 1 shows that $\mathcal{C}$ is this class of concepts, then for any sequence of classifiers $\{g_n\}$, there exists a distribution (μ, C) (with μ being the uniform distribution on the above line segment) such that

$$\mathbf{E}\{L(g_n)\} \geq \frac{(1-\epsilon)d}{2n}$$

for infinitely many n, where ϵ is any fixed small positive number.

References

[1] P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33:233–282, 1983.

[2] Y.S. Chow and H. Teicher. *Probability Theory, Independence, Interchangeability, Martingales.* Springer-Verlag, New York, 1978.

[3] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[4] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 1995. To appear.

[5] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

[6] D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0, 1\}$ functions from randomly drawn points. In *Proceedings of the 29th IEEE Symposium on the Foundations of Computer Science*, pages 100–109. IEEE Computer Society Press, Los Alamitos, CA, 1988.

[7] H.U. Simon. General lower bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, pages 402–412. Association for Computing Machinery, New York, 1993.

[8] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.