Strong minimax lower bounds for learning *

András Antos   Gábor Lugosi

December, 1996

Abstract

Minimax lower bounds for concept learning state, for example, 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 not tell anything about the rate of decrease of the error for a *fixed* distribution-concept pair.

In this paper we investigate minimax lower bounds in such a—stronger—sense. We show that for several natural k-parameter concept classes, including the class of linear halfspaces, the class of balls, the class of polyhedra with a certain number of faces, and a class of neural networks, 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 k/n for *infinitely many* n. We also obtain such strong minimax lower bounds for the tail distribution of the probability of error, which extend the corresponding minimax lower bounds.

* A. Antos is with the 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). G. Lugosi is with the Department of Economics, Pompeu Fabra University, Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain. (email: lugosi@upf.es).

1 Introduction

Let X be a random variable on a domain $\mathcal{X}$ with distribution μ, that is, 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 (or target) $C \in \mathcal{C}$ 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. All random variables are defined on a common probability space $(\Omega, \mathcal{A}, \mathbf{P})$, and $\mathbf{E}$ denotes expectation with respect to $\mathbf{P}$. 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

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

Thus, $L(g_n)$ is the probability that, trained on the data sequence D_n, the classifier g_n makes a mistake. Its value depends on the actual value of the data sequence D_n. The *expected probability of error*

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

is the expected value of $L(g_n)$. The joint distribution of the pair $(X, I_{\{X \in C\}})$ is determined by the pair (μ, C), which will be referred to as a *distribution-target pair.*

The minimax behavior of the expected probability of error has been thoroughly studied. Here the question is the size of the minimax error

$$\inf_{g_n} \sup_{(\mu,C)} \mathbf{E}L(g_n),$$

where the infimum is taken over all (measurable) learning rules, while the supremum is taken over all possible distribution-target pairs with $C \in \mathcal{C}$. The minimax error expresses the minimal achievable worst-case error for a given sample size n, and concept class $\mathcal{C}$.

It is a beautiful fact that for a given n, the minimax expected error $\inf_{g_n} \sup_{(\mu,C)} \mathbf{E}L(g_n)$ is basically determined by V, the VC *dimension* V of the class $\mathcal{C}$, and it is insensitive to other properties of $\mathcal{C}$. V is defined as the largest integer $k \geq 1$ with $s(k) = 2^k$, where the k-th *shatter coefficient* $s(k)$ of the class $\mathcal{C}$ is defined as the maximal number of different sets in

$$\{\{x_1, \ldots, x_k\} \cap C; C \in \mathcal{C}\},$$

where the maximum is taken over all $x_1, \ldots, x_k \in \mathcal{X}$. If $s(k) = 2^k$ for all k, then, by definition, $V = \infty$.

Haussler, Littlestone, and Warmuth [7] showed that there exists a learning rule such that for all distribution-target pairs,

$$\mathbf{E}L(g_n) \leq \frac{V}{n}. \tag{1}$$

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 [12] showed that for every $n \geq V-1$, and every classifier g_n, there exists a distribution-target pair such that

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

(1) and (2) together essentially solve the minimax problem for the expected probability of error, since they state that for each $n \geq V-1$, the minimax expected error is sandwiched between constant multiples of V/n, that is,

$$\frac{V-1}{2en}\left(1-\frac{1}{n}\right) \leq \inf_{g_n} \sup_{(\mu,C)} \mathbf{E}L(g_n) \leq \frac{V}{n}.$$

The expected probability of error is a useful quantity in describing the behavior of $L(g_n)$. However, it is rather the tail probabilities

$$\mathbf{P}\{L(g_n) \geq \epsilon\}$$

(where $\epsilon \in [0,1]$) that completely describe the distribution of the probability of error. The minimax problem for the tail probabilities is thus a more interesting (and harder) problem. Here one is interested in the quantity

$$\inf_{g_n} \sup_{(\mu,C)} \mathbf{P}\{L(g_n) \geq \epsilon\},$$

if n, ϵ, and $\mathcal{C}$ are given. The VC dimension also features minimax upper and lower bounds for the tail probabilities. For example, a classical result of Vapnik and Chervonenkis [12] (see also [3]) states that if g_n is any classifier such that $g_n(X_i) = I_{\{X_i \in C\}}$ for all $i = 1, \ldots, n$, and $\{x : g_n(x) = 1\} \in \mathcal{C}$, then for $n \geq V$,

$$\mathbf{P}\{L(g_n) \geq \epsilon\} \leq 2\left(\frac{2ne}{V}\right)^V e^{-n\epsilon \log 2/2}. \tag{3}$$

An improved version of this inequality can be found in [11]. Corresponding minimax lower bounds were first proved by Blumer, Ehrenfeucht, Haussler, and Warmuth [3], and Ehrenfeucht, Haussler, Kearns, and Valiant [6]. Devroye and Lugosi [5] (see also [4]) show that for any classifier g_n, there exists a distribution-target pair such that

$$\mathbf{P}\{L(g_n) \geq \epsilon\} \geq \frac{1}{2e\sqrt{\pi(V-1)}}\left(\frac{2ne\epsilon}{V-1}\right)^{(V-1)/2} e^{-4n\epsilon/(1-4\epsilon)}, \tag{4}$$

whenever $V \geq 2$, $n \geq V-1$ and $\epsilon < 1/4$.

The combination of (3) and (4) yield that for any concept class $\mathcal{C}$, with $V \geq 2$, $n \geq V$, and $\epsilon < 1/4$,

$$\frac{1}{2e\sqrt{\pi(V-1)}}\left(\frac{2ne\epsilon}{V-1}\right)^{(V-1)/2} e^{-4n\epsilon/(1-4\epsilon)} \leq \inf_{g_n} \sup_{(\mu,C)} \mathbf{P}\{L(g_n) \geq \epsilon\} \leq 2\left(\frac{2ne}{V}\right)^V e^{-n\epsilon \log 2/2}.$$

Note that in terms of n and ϵ, there is an order-of-magnitude gap between the upper and lower bounds: the pre-exponent of the upper bound is roughly of the order of n^V, while that of the lower bound is $(n\epsilon)^V$. The "interesting" values of ϵ are clearly around $1/n$, therefore the difference may be quite significant. (The improved version of (3) in [11] also leaves the gap open.) For some concept classes—for example for the class of unions of V initial segments discussed below—it is possible to prove an upper bound which, apart from constant factors, coincides with the lower bound (4). It is an interesting open question if the lower bound is tight for all concept classes.

In some sense, lower bounds of the form of (2) and (4) 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-target pair as the sample size n increases. In other words, the "bad" distribution-target pair, causing the largest error for a learning rule, may be different for each n. For example, the lower bound (2) does not exclude the possibility that there exists a sequence of classifiers $\{g_n\}$ such that for *every* μ and C the expected error $\mathbf{E}L(g_n)$ decreases at an exponential rate in n. Indeed, it is easy to see that such classes exist with arbitrarily large, and even with infinite, VC dimension (see Propositions 1 and 2 below). Schuurmans [10] studied the question when such exponential decrease occurs, and characterized it among certain "one-dimensional" problems. We are interested in "strong" minimax lower bounds that describe the behavior of the error for a fixed distribution-target pair (μ, C) as the sample size n grows. For example, the sequence $\{a_n\}$ of positive numbers is a *strong minimax lower bound* for the expected error for $\mathcal{C}$ if

$$\inf_{\{g_n\}} \sup_{(\mu,C)} \limsup_{n\to\infty} \frac{\mathbf{E}L(g_n)}{a_n} \geq 1,$$

where the infimum is taken over all *sequences* $\{g_n\}$ of classifiers and the supremum is taken over all distribution-target pairs with $C \in \mathcal{C}$. A slightly different, but essentially equivalent, definition requires that for all sequences $\{g_n\}$, there exists a fixed pair (μ, C) such that $\mathbf{E}L(g_n) \geq a_n$ for *infinitely many* n. The notion of strong minimax lower bounds can be defined similarly for tail probabilities by replacing $\mathbf{E}L(g_n)$ by $\mathbf{P}\{L(g_n) \geq \epsilon_n\}$, where $\{\epsilon_n\}$ is a fixed sequence of positive numbers.

The purpose of this paper is to establish minimax lower bounds in the described strong sense. The main results extend the lower bounds of (2) and (4). Because of the reason mentioned above, this is clearly not possible for all VC classes. However, the extension is possible for many important geometric concept classes, and the role of the VC dimension is played by the number of parameters of the class, which, in all of our examples, is closely related to the VC dimension of the class. Thus, the situation here significantly differs from that of the usual minimax theory, where a single combinatorial parameter—the VC dimension—completely determines the behavior of the concept class.

We close this introduction by illustrating through a simple example why it is impossible to give a "strong" extension of the lower bound of (2) for all VC classes. (See Schuurmans [10] for much more on this.) It can be seen similarly that no strong extension of (4) can be given for all VC classes either.

As the simplest example, let $\mathcal{C}$ be any class containing finitely many concepts. Then consider a learning rule that selects a concept C_n from $\mathcal{C}$ which is consistent with the data D_n, that is, $g_n(x) = I_{\{x \in C_n\}}$ for some $C_n \in \mathcal{C}$, and

$$g_n(X_i) = I_{\{X_i \in C\}} \quad \text{for all } i = 1, \ldots, n,$$

where $C \in \mathcal{C}$ is the true concept. Then (3) implies that

$$\mathbf{E}L(g_n) \leq \frac{2V \log(2n) + 4}{n \log 2},$$

where V is the VC dimension of $\mathcal{C}$. The beauty of this bound is that it is independent of the distribution-target pair, and that it is essentially the best such bound (see [7, Theorem 4.2]). However, for *all* distribution-target pairs, the error decreases at a much faster rate. This can be seen from the simple fact that g_n can only make an error if there is at least one concept $C' \in \mathcal{C}$ with $\mu(C' \triangle C) > 0$ 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

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

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

Proposition 1 *Let V be an arbitrary positive integer. There exists a class $\mathcal{C}$ with* VC *dimension V and a corresponding sequence of learning rules $\{g_n\}$ such that for all distribution-target pairs (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 depend on the distribution-target pair.

If a concept class $\mathcal{C}$ is finite, its n-th shatter coefficient $s(n)$ is bounded above by $|\mathcal{C}|$ 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 distribution-target pairs. It is natural to ask if the growth of $s(n)$ determines the rate of convergence of the error. This conjecture is false, and in fact, we may have an exponential rate of convergence for all distribution-target pairs even for classes with infinite VC dimension (for which $s(n) = 2^n$ for all n):

Proposition 2 *There exists a class $\mathcal{C}$ whith $V = \infty$ and a corresponding sequence of learning rules $\{g_n\}$, such that for all distribution-target pairs and for all n,*

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

where $b < 1$. The positive constants a and b depend on the distribution-target pair.

Proof. Let $\mathcal{X} = \mathcal{R}$ and let $\mathcal{C}$ contain all finite subsets of $\mathcal{R}$. Let $g_n(x) = 1$ if and only if there exists an X_i such that $x = X_i$ and $I_{\{X_i \in C\}} = 1$. $\Box$

The rest of the paper is organized as follows. In Section 2 we introduce a general tool, an application of the "probabilistic method," for obtaining strong minimax lower bounds. In Section 3 we provide strong minimax lower bounds for the expected probability of error $\mathbf{E}L(g_n)$ if the concept class $\mathcal{C}_k$ is the class of unions of k initial segments. (This class was introduced in [7].) In particular, we show that for every sequence of learning rules $\{g_n\}$, there exist a distribution-target pair (μ, C) such that if C is the "true" concept, then

$$\mathbf{E}L(g_n) > (1-\epsilon)\frac{k}{2n} \quad \text{for infinitely many } n,$$

where ϵ is an arbitrarily small fixed number; see Theorem 1 for the precise statement. Since the VC dimension of this class is k, this result states that there always exists a distribution-target pair such that the error is essentially within a factor of two of the upper bound of (1) infinitely many times. We extend this result to more general classes of concepts showing that the above lower bound remains true for many other important classes of "dimension" k. (Here by dimension we mean the number of parameters of the class, which, in most of our cases, essentially coincides with the VC dimension of the class.) These examples include the class of halfspaces, the class of d-dimensional intervals, the class of euclidean balls, the class of all ellipsoids, certain classes of neural networks, etc.

Section 4 extends the results of Section 3 for the expected *cumulative* error.

Section 5 contains the main results of the paper. Here we present analogous lower bounds for the tail probabilities $\mathbf{P}\{L(g_n) \geq \epsilon\}$, which extend (4). Clearly, these bounds are much more informative than bounds for the expected value of $L(g_n)$, however, their proof is much more technical. Parts of the proofs are given in Appendix 2.

2 The probabilistic method

In this technical section we present a simple lemma that equips us with a general tool for proving strong minimax lower bounds. Let $R_n(z)$ be a sequence of nonnegative numbers parametrized by an abstract parameter z from a set $\mathcal{Z}$. Assume that we wish to prove that for some fixed n, there exists a $z \in \mathcal{Z}$ such that $R_n(z) \geq a_n$, where $a_n > 0$. Then it suffices to find a random variable Z on $\mathcal{Z}$ such that $\mathbf{P}\{R_n(Z) \geq a_n\} > 0$. This simple trick is the basic idea of the powerful "probabilistic method." Another, equally simple, way for obtaining a lower bound by the probabilistic method is using the trivial fact that for any random variable Z,

$$\sup_{z \in \mathcal{Z}} R_n(z) \geq \mathbf{E}R_n(Z). \tag{5}$$

For example, we may take $R_n(z) = \mathbf{E}L(g_n)$, where the parameter space $\mathcal{Z}$ is obtained by a suitable parametrization of the concept class $\mathcal{C}$, and the parameter z corresponds to a concept $C \in \mathcal{C}$. Thus, to show that there exists a concept such that the error $\mathbf{E}L(g_n)$ is greater than a_n, one might use one of the ideas above. In fact, (5) is at the heart of the proof of essentially all minimax lower bounds we are aware of.

In this paper we wish to prove something of a stronger form: there exists a fixed $z \in \mathcal{Z}$ such that $R_n(z) \geq a_n$ for *infinitely many* n, where $a_1, a_2, \ldots$ is a sequence of positive numbers. The first difficulty to overcome is that the randomizing distribution cannot depend on n any longer. However, it is not sufficient to find a fixed random variable Z such that $\mathbf{E}R_n(Z) \geq a_n$ for all n. An additional stability property is needed to obtain a bound of the desired form.

The following lemma provides a simple way of proving lower bounds of the desired form. A somewhat weaker version is implicitly used by Schuurmans [10]. A significantly stronger form and more discussion is added in Appendix 1.

Lemma 1 *Let $R_n(z)$ be a sequence of nonnegative numbers parametrized by an abstract parameter z from a set $\mathcal{Z}$. If there exists a random variable Z taking its values from $\mathcal{Z}$ such that*

$$\frac{\operatorname{ess\,sup} R_n(Z)}{\mathbf{E}R_n(Z)} \not\to \infty \quad \text{as } n \to \infty, \tag{6}$$

then there exists a $z \in \mathcal{Z}$ such that for every $0 < \epsilon < 1$,

$$R_n(z) > (1-\epsilon)\mathbf{E}\{R_n(Z)\} \quad \text{for infinitely many } n.$$

(Recall that for a random variable X, $\operatorname{ess\,sup} X = \inf\{x : \mathbf{P}\{X > x\} = 0\}$.)

Proof. It suffices to prove that

$$\mathbf{P}\left\{\limsup_{n\to\infty} \frac{R_n(Z)}{\mathbf{E}R_n(Z)} \geq 1\right\} > 0.$$

Condition (6) means that there exists a sequence $\{n_i\}$ of indices along which the subsequence $\operatorname{ess\,sup} R_{n_i}(Z)/\mathbf{E}R_{n_i}(Z)$ is bounded. Thus, Fatou's lemma may be applied to the sequence of random variables $R_{n_i}(Z)/\mathbf{E}R_{n_i}(Z)$, $i = 1, 2, \ldots$:

$$\mathbf{E}\left\{\limsup_{n\to\infty} \frac{R_n(Z)}{\mathbf{E}R_n(Z)}\right\} \geq \mathbf{E}\left\{\limsup_{i\to\infty} \frac{R_{n_i}(Z)}{\mathbf{E}R_{n_i}(Z)}\right\} \geq \limsup_{i\to\infty} \mathbf{E}\left\{\frac{R_{n_i}(Z)}{\mathbf{E}R_{n_i}(Z)}\right\} = 1,$$

and the statement follows. □

Remark. Lemma 1 states that minimax lower bounds obtained by using the simplest form (5) of the probabilistic method can be extended to their strong form if the randomization Z does not depend on n, and, in addition, the stability property

$$\frac{\operatorname{ess\,sup} R_n(Z)}{\mathbf{E}R_n(Z)} \not\to \infty$$

is verified. The following example demonstrates that this additional condition cannot be dropped, and some kind of stability condition is necessary. Let $\mathcal{C} = \{\{x\} : x \in \mathcal{X}\}$ be the class of one-point concepts on the domain $\mathcal{X}$ of positive integers. Let $\{g_n\}$ be an arbitrary sequence of learning rules, and for $z \in \mathcal{X}$, define $R_n(z) = \mathbf{P}\{g_n(X, D_n(z)) \neq Y(z)\}$, where $Y(z) = I_{\{X=z\}}$, and $D_n(z) = ((X_1, I_{\{X_1=z\}}), \ldots, (X_n, I_{\{X_n=z\}}))$. Let $\mathbf{P}\{X = i\} = c/(i \log^2 i)$

for an appropriate normalizing constant c, and introduce the random variable Z distributed as X, and independent of $X, X_1, \ldots, X_n$. Using a similar argument as in the proof of Theorem 1, one sees that for every n, $\mathbf{E}R_n(Z) \geq \text{const.}/(n+2)^{5/2}$. However, similarly to the proof of Proposition 2, one can show easily that there exists a sequence $\{g_n\}$ such that for every z, $R_n(z)$ converges to zero exponentially rapidly. This demonstrates the fact that a lower bound for $\mathbf{E}R_n(Z)$ cannot necessarily be converted into a strong minimax lower bound, even if the randomization Z is independent of n. An additional condition, such as (6) needs to be satisfied.

Remark. Finding a fixed random variable Z such that $\mathbf{E}R_n(Z) \geq a_n$ for all n, is useful in a different situation, even if the additional stability property (6) cannot be verified. It allows us to derive lower bounds for the cumulative error. In particular, in such a case we have, for every n, that

$$\sup_{z \in \mathcal{Z}} \left(\sum_{i=0}^{n} R_i(z) \right) \geq \sum_{i=0}^{n} \mathbf{E}R_i(Z). \tag{7}$$

We discuss lower bounds for the cumulative error in Section 4.

3 Bounds for the expected probability of error

In this section we provide examples of concept classes for which the minimax lower bound (2) for the expected probability of error $\mathbf{E}L(g_n) = \mathbf{P}\{g_n(X) \neq I_{\{X \in C\}}\}$ can be extended to its strong version. All examples shown here are based on lower bounds obtained for a very simple concept class. The class $\mathcal{C}_k$ of unions of k initial segments is defined as follows: let $\mathcal{X} = [0,1] \times \{1, 2, \ldots, k\}$, and

$$\mathcal{C}_k = \left\{ \bigcup_{j=1}^{k} ([0, z_j] \times \{j\}) : z \in [0,1]^k \right\}. \tag{8}$$

The class $\mathcal{C}_k$ is therefore parametrized by a vector of k parameters: $z = (z_1, \ldots, z_k) \in [0,1]^k$. Clearly, the VC dimension of $\mathcal{C}_k$ is also k. For this class, we have the following result:

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

$$\mathbf{E}L(g_n) > (1-\epsilon)\frac{k}{2n} \quad \textit{for infinitely many } n.$$

Remark. Haussler, Littlestone, and Warmuth [7, Theorem 3.2] showed for the class $\mathcal{C}_k$ of unions of k initial segments that for every learning rule, and for every n, there exists a $C \in \mathcal{C}_k$ such that

$$\mathbf{E}L(g_n) \geq \frac{k}{2n} - O\left(n^{-2}\right).$$

Furthermore, in their proof of this lower bound, the randomization Z is independent of n. To make the proof of Theorem 1 short, we use many elements of the proof of the above inequality. A different proof (for a slightly weaker version) of Theorem 1 may be found in [1].

Remark. Note that the lower bound $k/(2n) - O(n^{-2})$ for the minimax expected error is better in the constant factor than the bound of (2). However, it is less general, since it does not apply to any VC class.

Remark. It is clear from the proof of the theorem that the uniform distribution may be replaced by any nonatomic distribution on $\mathcal{X}$.

Proof. Let $z \in [0,1]^k$ be the parameter that determines $C \in \mathcal{C}_k$. First we introduce some notation. Let $Y(z) = I_{\{X \in C\}}$, $Y_i(z) = I_{\{X_i \in C\}}$, and $D_n(z) = ((X_1, Y_1(z)), \ldots, (X_n, Y_n(z)))$. Denote $X = \langle U, M \rangle$ so that U is uniformly distributed on $[0,1]$, M is uniform on $\{1, \ldots, k\}$, and U and M are independent. Introduce

$$l = \max\{u \in [0,1] : u \leq U \text{ and } \langle u, M\rangle \in \{X_1, \ldots, X_n\} \cup \langle 0, M\rangle\}$$

and

$$r = \min\{u \in [0,1] : u \geq U \text{ and } \langle u, M\rangle \in \{X_1, \ldots, X_n\} \cup \langle 1, M\rangle\},$$

that is, l and r are the left and right neighbors of U among the data points falling on the M-th segment. Finally, define the following (random) sets of parameters:

$$L_n = \left\{z \in [0,1]^k : z_M \in [l, U)\right\} \quad \text{and} \quad R_n = \left\{z \in [0,1]^k : z_M \in [U, r)\right\}.$$

Clearly,

$$\mathbf{E}L(g_n) \geq R_n(z) \stackrel{\text{def}}{=} \mathbf{P}\{g_n(X, D_n(z)) \neq Y(z), z \in L_n \cup R_n\}.$$

We apply Lemma 1 for $R_n(z)$. (The reason why we do not define $R_n(z)$ as the expected probability of error $\mathbf{E}L(g_n)$ itself is that this is the only way we can ensure that the additional stability property (6) required by Lemma 1 holds.) We will show that if the random vector $Z = (Z_1, \ldots, Z_k)$ is uniformly distributed on $[0,1]^k$ and independent of $X, X_1, \ldots, X_n$, then

$$\mathbf{E}\{R_n(Z)\} \geq \frac{k}{2(n+1)} - \frac{k^2}{2(n+1)(n+2)}\left(1 - \left(1 - \frac{1}{k}\right)^{n+2}\right), \tag{9}$$

and $\sup_z R_n(z)/\mathbf{E}\{R_n(Z)\}$ does not tend to infinity, from which the theorem follows.

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

$$R_n(Z) = \mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z), Z \in L_n \cup R_n | Z\}, \tag{10}$$

and

$$\begin{aligned}
\mathbf{E}\{R_n(Z)\} &= \mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z), Z \in L_n \cup R_n\} \\
&= \mathbf{E}\left\{\mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z), Z \in L_n \cup R_n | X, X_1, \ldots, X_n\}\right\} \\
&= \mathbf{E}\Big\{\mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z) | Z \in L_n, X, X_1, \ldots, X_n\}\mathbf{P}\{Z \in L_n | X, X_1, \ldots, X_n\} \\
&\quad + \mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z) | Z \in R_n, X, X_1, \ldots, X_n\}\mathbf{P}\{Z \in R_n | X, X_1, \ldots, X_n\}\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbf{E}\Big\{\min(\mathbf{P}\{Z \in L_n | X, X_1, \ldots, X_n\}, \mathbf{P}\{Z \in R_n | X, X_1, \ldots, X_n\})\Big\} \\
&= \mathbf{E}\{\min(U - l, r - U)\} \\
&= \frac{k}{2(n+1)} - \frac{k^2}{2(n+1)(n+2)}\left(1 - \left(1 - \frac{1}{k}\right)^{n+2}\right),
\end{aligned}
$$

where the last equality follows from direct calculation, which is detailed in the proof of Theorem 3.2 in [7].

On the other hand, we observe that for each fixed $z \in [0,1]^k$,

$$
R_n(z) \leq \mathbf{P}\{z \in L_n \cup R_n\} \leq \frac{2k}{n+1}. \tag{11}
$$

This may be seen by conditioning on the *set* $\{X, X_1, \ldots, X_n\}$, and observing that since $X, X_1, \ldots, X_n$ are i.i.d., the probability remains the same by permuting them. Of the $(n+1)!$ permutations, there are at most $2kn!$ such that X is a neighbor of one of the z_i's. Therefore,

$$
\frac{\sup_z R_n(z)}{\mathbf{E}R_n(Z)} \leq 4 + o(1),
$$

so the condition of Lemma 1 is satisfied, and the proof of the theorem is complete. □

We may extend Theorem 1 to other important classes of geometric concepts by embedding. For example, we have the following straightforward corollary of Theorem 1.

Corollary 1 *Let $\mathcal{X} = \mathcal{R}^d$, and let $\mathcal{C}$ be a class of concepts defined on $\mathcal{X}$. If there exist invertible measurable mappings $f_1, \ldots, f_k : [0,1] \to \mathcal{R}^d$ such that the sets $f_i([0,1])$ are disjoint and for all $z = (z_1, \ldots, z_k) \in [0,1]^k$ there exists a $C \in \mathcal{C}$ with*

$$
C \cap (f_1([0,1]) \cup \cdots \cup f_k([0,1])) = f_1([0,z_1]) \cup \cdots \cup f_k([0,z_k]),
$$

then for any sequence of classifiers $\{g_n\}$, there exists a distribution-target pair (μ, C) with μ concentrated on $f_1([0,1]) \cup \cdots \cup f_k([0,1])$ and $C \in \mathcal{C}$ such that for all $0 < \epsilon < 1$,

$$
\mathbf{E}L(g_n) > \frac{(1-\epsilon)k}{2n} \quad \textit{for infinitely many } n.
$$

The above corollary may be applied to many important geometric concept classes. Below we give a short list of examples. The proofs are quite straightforward, most of them can be found in [7, p.279].

1. If $\mathcal{C}$ is the class of subsets of $\mathcal{R}$ that can be written as a union of m intervals, then $k = 2m$.
2. $k = d$ for the class of d-dimensional octants:

$$
\Big\{x \in \mathcal{R}^d : x_i \leq a_i, i = 1, \ldots, d\Big\}, \quad a_1, \ldots, a_d \in \mathcal{R},
$$

where $x_1, \ldots, x_d$ are the components of the vector x.

3. $k = 2d$ for the class of d-dimensional intervals:

$$\left\{x \in \mathcal{R}^d : a_i \le x_i \le b_i, i = 1, \ldots, d\right\},$$

where $a_1, b_1, \ldots, a_d, b_d \in \mathcal{R}$.

4. $k = d$ if $\mathcal{C}$ is the class of halfspaces of $\mathcal{R}^d$, that is, sets of the form

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

5. $k = d$ if

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

6. $k = d + 1$ for the class of balls in $\mathcal{R}^d$:

$$\left\{x \in \mathcal{R}^d : \sum_{i=1}^{d}(x_i - a_i)^2 \le r\right\},$$

where $a_1, \ldots, a_d, r \in \mathcal{R}, r \ge 0$.

7. $k = 2d$ for the class of all d-dimensional ellipsoids:

$$\left\{x \in \mathcal{R}^d : \sum_{i=1}^{d} \frac{(x_i - a_i)^2}{b_i} \le 1\right\},$$

where $a_1, b_1, \ldots, a_d, b_d \in \mathcal{R}$.

8. $k = md$ for the class of convex polyhedra of m faces in $\mathcal{R}^d$.

9. $k = md$ for the class $\mathcal{C}$ of all neural network classifiers on $\mathcal{R}^d$ with m hidden nodes in their single hidden layer, that is, each $C \in \mathcal{C}$ is of the form

$$\left\{x : \sum_{i=1}^{m} a_i \sigma(b_i x^T + c_i) + a_0 \ge 0\right\},$$

where $a_0, \ldots, a_m, c_1, \ldots, c_m \in \mathcal{R}$, $b_1, \ldots, b_m \in \mathcal{R}^d$, and σ is the threshold sigmoid $\sigma(x) = I_{\{x>0\}}$. (x^T denotes the transpose of a vector x.)

Remark. Based on Corollary 1, we may define a new "dimension" Δ for a concept class $\mathcal{C}$ as follows: let Δ be the largest integer k such that there exist k invertible measurable mappings $f_1, \ldots, f_k : [0,1] \to \mathcal{R}^d$ such that the sets $f_i([0,1])$ are disjoint and for all $z = (z_1, \ldots, z_k) \in [0,1]^k$ there exists a $C \in \mathcal{C}$ with

$$C \cap (f_1([0,1]) \cup \cdots \cup f_k([0,1])) = f_1([0,z_1]) \cup \cdots \cup f_k([0,z_k]).$$

If no such mapping exists then $\Delta = 0$, and if for each k there are k mappings with the above property, then $\Delta = \infty$.

Corollary 1 shows the relation of Δ to strong minimax lower bounds for the expected error. Lower bounds for Δ may be obtained in specific cases by construction. For upper bounds, note that it is easy to see that $\Delta \leq V$, since a set $\{x_1, \ldots, x_\Delta\}$ is shattered by $\mathcal{C}$ if for each $i \leq \Delta$, $x_i \in f_i([0,1])$. Further, it is easy to see that for each n,

$$s(n) \geq \left\lfloor \frac{n}{\Delta} \right\rfloor^{\Delta}$$

(just put at least $\lfloor n/\Delta \rfloor$ of the n points on each image $f_i([0,1])$ of the segment $[0,1]$, $i = 1, \ldots, \Delta$), which means that also $\Delta \leq D$, where D is the *Assouad density* of $\mathcal{C}$, defined as

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

see Assouad [2]. (It is well-known that $D \leq V$, $D < \infty$ if and only if $V < \infty$, and that for each k there exists a class $\mathcal{C}$ with $V = k$ and $D = 0$.) Thus, we have

$$\Delta \leq D \leq V.$$

On the other hand, it follows from Proposition 2 and Corollary 1 that there exists a class $\mathcal{C}$ such that $D = V = \infty$, but $\Delta = 0$.

4 Cumulative error bounds

Let $\{g_n\}$ be a sequence of learning rules. The *cumulative error* is defined as

$$\sum_{i=0}^{n} I_{\{g_i(X_{i+1}, D_i) \neq I_{\{X_{i+1} \in C\}}\}},$$

that is, the number of errors committed by the sequence in the first n steps, if the first i labelled examples are always used to predict the label of the $i+1$-th example. Based on the results of the previous section, it is easy to obtain strong minimax lower bounds for the expected value of the cumulative error.

It is a direct consequence of (1) that there exists a sequence of learning rules such that for all distribution-target pairs

$$\mathbf{E} \left\{ \sum_{i=0}^{n} I_{\{g_i(X_{i+1}, D_i) \neq I_{\{X_{i+1} \in C\}}\}} \right\} \leq V \log(n+1) + 1.$$

(see [7]).

Haussler, Littlestone, and Warmuth [7] considered minimax lower bounds for the expected cumulative error. The observation (7) is at the basis of the proof of their Corollary 3.1 where it is proved [1] for the class $\mathcal{C}_k$ introduced in Section 3 that for every n, and for every sequence

[1]Corollary 3.1 of [7] states more than what is actually proved there. It states that for every sequence of learning rules, there exists a distribution-target pair such that for *every* n, the expected cumulative error is lower bounded as in (12). The proof of Corollary 3.1 of [7] has the quantifiers reversed, so it in fact does not show that there is a fixed C such that the lower bound holds for all n.

of learning rules, there exists a distribution-target pair such that the expected cumulative error satisfies

$$\mathbf{E}\left\{\sum_{i=0}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}}\right\} \geq \frac{k}{2}\left(\log\frac{n+1}{k}-1\right). \tag{12}$$

We have the following "strong" extension of the above minimax lower bound:

Theorem 2 *Let μ be the uniform distribution on $[0,1]\times\{1,2,\ldots,k\}$. For every sequence of learning rules $\{g_n\}$, there exist a $C\in\mathcal{C}_k$ such that for all $0<\epsilon<1$,*

$$\mathbf{E}\left\{\sum_{i=0}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}}\right\} > (1-\epsilon)\frac{k}{2}\log n \quad \textit{for infinitely many } n.$$

Proof. We apply Lemma 1 with

$$R_n(z) \stackrel{\text{def}}{=} \sum_{i=0}^{n} \mathbf{P}\{g_i(X_{i+1},D_i(z))\neq Y_{i+1}(z), z\in L_i\cup R_i\}.$$

(Recall the definition of $Y_i(z), D_i(z), L_i$ and R_i from Section 3.) Clearly,

$$\mathbf{E}\left\{\sum_{i=0}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}}\right\} = \sum_{i=0}^{n}\mathbf{P}\{g_i(X_{i+1},D_i(z))\neq Y_{i+1}(z)\} \geq R_n(z).$$

Then it follows from (9) that if Z is uniform on $[0,1]\times\{1,2,\ldots,k\}$, and independent of $X, X_1, X_2\ldots$, then

$$\mathbf{E}R_n(Z) \geq \frac{k}{2}\left(\log\frac{n+1}{k}-1\right).$$

(For the details see [7, p.278].) On the other hand, (11) implies that for each z,

$$R_n(z) \leq \sum_{i=0}^{n}\frac{2k}{i+1} \leq 2k(\log(n+1)+1),$$

so condition (6) is satisfied, which completes the proof. □

5 Bounds for the tail probabilities

The purpose of this section is to give strong lower bounds of the following type: let $\mathcal{C}$ be a class of concepts, and let $\{\epsilon_n\}$ be a sequence of positive numbers. Then for any sequence of learning rules $\{g_n\}$, there exists a distribution-target pair (μ, C) with $C\in\mathcal{C}$ such that

$$\mathbf{P}\{L(g_n)\geq\epsilon_n\}\geq a_n \quad \text{for infinitely many } n.$$

Here we would like to have

$$a_n \approx (c_0 n\epsilon_n)^{kc_1} e^{-n\epsilon_n c_2}$$

for some constants c_0, c_1, c_2, where k is the "dimension" of $\mathcal{C}$ so that the result is indeed an extension of (4). For some sequences of ϵ_n's (which we believe to be the most interesting ones) we will be able to prove such results if $\mathcal{C}$ is one of the geometric concept classes discussed in Section 3.

Clearly, the most interesting values of ϵ_n are constant multiples of $1/n$, since this is the range where the probability of error $L(g_n)$ of a good learning rule g_n is expected to be with high probability. Our main result extends (4) to such values of ϵ_n:

Theorem 3 *Let $\mathcal{C}_k$ be the class of unions of k initial segments as defined in (8), and let μ be the uniform distribution on $\mathcal{X} = [0,1] \times \{1, \ldots, k\}$. Let $\gamma \geq 0$ be fixed, and define $\epsilon_n = \gamma/n$. For any sequence $\{g_n\}$ there exists a $C \in \mathcal{C}_k$ such that for each $\delta \in (0,1)$,*

$$\mathbf{P}\{L(g_n) \geq \epsilon_n\} \geq (1-\delta)\frac{1}{2}\sum_{i=0}^{k-1}\frac{(c\gamma)^i}{i!}e^{-c\gamma} \quad \textit{for infinitely many } n, \tag{13}$$

where $c = \log 256 \approx 5.545$.

Remark. At the price of more complicated arguments, the value of the constant c may be improved to something slightly larger than 2. However, it is not our aim here to search for the sharpest constants, so we stay with suboptimal constants and simpler arguments.

Note that since

$$\sum_{i=0}^{k-1}\frac{(c\gamma)^i}{i!}e^{-c\gamma} \geq \left(\frac{c\gamma}{k-1}\right)^{k-1} e^{-c\gamma},$$

apart from constants, the lower bound of Theorem 3 has the same form as that of (4). By the same embedding argument as the one used in Corollary 1, Theorem 3 can be extended to the concept classes listed in Section 3. The intuitive idea behind the proof of Theorem 3 is that in each of the k initial segments, inside the interval between the rightmost data point labelled by 1 and the lefttmost data point labelled by 0, no learning rule can do better than mere guessing. Thus, the sum of the lengths of these intervals determines the size of the minimal probability of error. In the proof we exploit the fact that the length of these intervals have approximately exponential distribution, and they are almost independent, therefore we may approximate the minimax tail distribution of $L(g_n)$ by the tail of an appropriate gamma distribution.

Proof of Theorem 3. First we introduce some notation:

$$U_{nj}^- = \max\{u \in [0,1] : \langle u, j\rangle \in \{X_i : Y_i = 1\} \cup \langle 0, j\rangle\},$$

$$U_{nj}^+ = \min\{u \in [0,1] : \langle u, j\rangle \in \{X_i : Y_i = 0\} \cup \langle 1, j\rangle\},$$

$$A'_{nj} = (U_{nj}^-, U_{nj}^+),$$

$$A_{nj} = A'_{nj} \times \{j\},$$

$$A_n = \bigcup_{j=1}^{k} A_{nj}.$$

Step 1. We apply Lemma 1 for $R_n(z) = R_{n,\epsilon_n}(z)$, where for each $\epsilon > 0$,

$$R_{n,\epsilon}(z) \stackrel{\text{def}}{=} \mathbf{P}\{L(g_n) > \epsilon\} = \mathbf{P}\left\{ \int_{\mathcal{X}} I_{\{g_n(x,D_n(z)) \neq Y(x,z)\}}\, d\mu(x) > \epsilon \right\}.$$

(Here $Y(x,z) = I_{\{x \in C_z\}}$, where C_z is the concept associated with the parameter $z \in [0,1]^k$.) Just like in the proof of Theorem 1, let $Z = (Z_1, \ldots, Z_k)$ be uniformly distributed on $[0,1]^k$, and independent of $X, X_1, \ldots, X_n$. Since $R_{n,\epsilon_n}(z)$ is always bounded above by 1, and since the desired lower bound of (13) is independent of n, it suffices to prove a suitable lower bound for $\mathbf{E}R_{n,\epsilon}(Z)$, as the stability property (6) is automatically satisfied. We will show that for each n and ϵ,

$$\mathbf{E}R_{n,\epsilon}(Z) \geq \frac{1}{2}\sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} - \frac{1}{2}\left[\left(\sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \right) k e^{-n/k} + k e^{-\frac{n}{k}\left(1 - \frac{\sqrt{\epsilon}}{2}\right)} \right] \tag{14}$$

(where $\gamma = n\epsilon$), which proves the theorem, since the term inside the brackets converges to zero rapidly as $n \to \infty$.

Clearly, by the independence of Z and $X, X_1, \ldots, X_n$,

$$\begin{aligned} R_{n,\epsilon}(Z) &= \mathbf{P}\left\{ \int_{\mathcal{X}} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}}\, d\mu(x) > \epsilon \,\middle|\, Z \right\} \\ &\geq \mathbf{P}\left\{ \int_{A_n} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}}\, d\mu(x) > \epsilon \,\middle|\, Z \right\}. \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathbf{E}R_{n,\epsilon}(Z) &\geq \mathbf{P}\left\{ \int_{A_n} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}}\, d\mu(x) > \epsilon \right\} \\ &= \mathbf{E}\left\{ \mathbf{P}\left\{ \int_{A_n} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}}\, d\mu(x) > \epsilon \,\middle|\, D_n(Z) \right\} \right\} \\ &= \mathbf{E}\left\{ \mathbf{P}\left\{ \sum_{j=1}^{k} \int_{A_{nj}} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}}\, d\mu(x) > \epsilon \,\middle|\, D_n(Z) \right\} \right\}. \end{aligned}$$

Step 2. In this step we obtain a lower bound for $\mathbf{P}\{L(g_n) > \epsilon\}$ in terms of the spacings containing the Z_i's. Let $\xi_{nj} = U_{nj}^+ - U_{nj}^-$. For all n and $\epsilon > 0$,

$$\mathbf{E}R_{n,\epsilon}(Z) \geq \frac{1}{2}\mathbf{P}\left\{ \frac{1}{k}\sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon \right\}.$$

Proof. Clearly,

$$\int_{A_{nj}} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\, d\mu(x) = \frac{1}{k}\lambda(A_{nj}\cap(B_{nj}\triangle C)) = \frac{1}{k}\lambda(B_{nj}\triangle C_{nj}),$$

where λ is the one-dimensional Lebesgue measure, and

$$B_{nj} = \{x \in A_{nj} : g_n(x, D_n(Z)) = 1\},$$

and

$$C_{nj} = C \cap A_{nj}.$$

Then it follows by Lemma 3 in Appendix 2 that

$$\int_{A_{nj}} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\, d\mu(x) \geq \frac{1}{k}\left|U_{nj}^- + \lambda(B_{nj}) - Z_j\right|,$$

and therefore

$$\begin{aligned}
&\mathbf{P}\left\{\sum_{j=1}^{k}\int_{A_{nj}} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\, d\mu(x) > \epsilon \,\middle|\, D_n(Z)\right\}\\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}|U_{nj}^- + \lambda(B_{nj}) - Z_j| > \epsilon \,\middle|\, D_n(Z)\right\}\\
&\geq \frac{1}{2} I_{\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj}\geq 4\epsilon\right\}},
\end{aligned}$$

where the last inequality is proved in Lemma 4 (see Appendix 2). Taking expected values of both sides, we obtain

$$\mathbf{E}R_{n,\epsilon}(Z) \geq \frac{1}{2}\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon\right\}$$

as desired. □

Step 3. Obviously,

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon\right\} \geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon, \forall N_j > 0\right\},$$

where N_j denotes the number of X_i's falling on the j'th initial segment. If $N_j > 0$, given N_j, the conditional distribution of ξ_{nj} is the same as the distribution of the sum of two spacings defined by $N_j + 1$ i.i.d. uniform random variables on $[0,1]$, that is, for all $\epsilon \in [0,1]$,

$$\mathbf{P}\{\xi_{nj} \geq \epsilon | N_1, \ldots, N_k\} = \mathbf{P}\{\xi_{nj} \geq \epsilon | N_j\} = (1-\epsilon)^{N_j}(1 + N_j\epsilon)$$

(see, e.g., Reiss [9]). A crucial step of the proof is approximating the conditional distributions of the ξ_{nj}'s by appropriate exponential distributions. For $N_j > 0$, define $\lambda_j = N_j \log 4 - 2\log(1 + N_j/2)$, and define the random variables $\xi'_{n1}, \ldots, \xi'_{nk}$ such that given $N_1, \ldots, N_k$, they are conditionally independent, and the conditional distribution of ξ'_{nj} is exponential with parameter λ_j, that is,

$$\mathbf{P}\{\xi'_{nj} \geq \epsilon | N_1, \ldots, N_k\} = \mathbf{P}\{\xi'_{nj} \geq \epsilon | N_j\} = e^{-\lambda_j \epsilon}.$$

(If $N_j = 0$ for some j, then $\mathbf{P}\{\xi'_{nj} \geq \epsilon | N_1, \ldots, N_k\}$ is defined arbitrarily.) Then, by Lemma 5 in Appendix 2, for all ϵ,

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon, \forall N_j > 0\right\} \geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon, \forall N_j > 0\right\} - ke^{-\frac{n}{k}\left(1-\frac{\sqrt{\epsilon}}{2}\right)}.$$

Step 4. To finish the proof of (14), it remains to show that

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon, \forall N_j > 0\right\} \geq \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma}\left(1 - ke^{-n/k}\right).$$

To do this, we may proceed as follows:

$$\begin{aligned}
\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon, \forall N_j > 0\right\} &= \mathbf{E}\left\{I_{\{\forall N_j > 0\}} \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\geq \mathbf{E}\left\{I_{\{\forall N_j > 0\}} \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi''_{nj} \geq 4\epsilon \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(where the } \xi''_{nj} \text{ are defined exactly as the} \\
&\quad\ \xi'_{nj} \text{ but with } \lambda_j \text{ replaced by } \lambda'_j = N_j \log 4) \\
&\geq \mathbf{E}\left\{I_{\{\forall N_j > 0\}} \mathbf{P}\left\{\frac{\sum_{j=1}^{k} \lambda'_j \xi''_{nj}}{\sum_{i=1}^{k} \lambda'_j} \geq 4\epsilon \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(by Lemma 6 in Appendix 2)} \\
&= \mathbf{E}\left\{I_{\{\forall N_j > 0\}} \mathbf{P}\left\{\Phi_k \geq 4\epsilon \sum_{i=1}^{k} \lambda'_j \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(where given } N_1, \ldots, N_k \text{, the random variable } \Phi_k \text{ has} \\
&\quad\ k\text{-th order gamma distribution with parameter 1,} \\
&\quad\ \text{since given } N_j \text{, each } \lambda'_j \xi''_{nj} \text{ has exponential} \\
&\quad\ \text{distribution with parameter 1.)} \\
&= \mathbf{E}\left\{I_{\{\forall N_j > 0\}} \mathbf{P}\left\{\Phi_k \geq 4n\epsilon \log 4 \,|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(since } \textstyle\sum_{i=1}^{k} \lambda'_j = n \log 4)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\left\{ I_{\{\forall N_j > 0\}} \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \right\} \\
&= \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \mathbf{P}\{\forall N_j > 0\} \\
&\geq \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \left(1 - k e^{-n/k}\right),
\end{aligned}
$$

since

$$
\mathbf{P}\{\forall N_j > 0\} \geq 1 - k\mathbf{P}\{N_1 = 0\} = 1 - k\left(1 - \frac{1}{k}\right)^n \geq 1 - k e^{-n/k},
$$

and the proof of (14) is finished, so the proof of the theorem is complete. □

The reason why we can prove strong tail lower bounds only for certain sequences of ϵ_n's is that the stability condition (6) is difficult to check in more general cases than $\epsilon_n = \gamma/n$ for some fixed γ. It is possible, however, to generalize Theorem 3 for other sequences. The proof of the following generalization is identical to that of Theorem 3:

Theorem 4 *Let $\mathcal{C}_k$ be the class of unions of k initial segments, and let μ be the uniform distribution on $\mathcal{X} = [0,1] \times \{1, \ldots, k\}$. Let $n_j \to \infty$ ($j = 1, 2, \ldots$) be a sequence of positive integers, and let ϵ_j be nonnegative numbers such that $\gamma_j = n_j \epsilon_j$ does not tend to ∞ as $j \to \infty$. Then for any sequence $\{g_n\}$ there exists a $C \in \mathcal{C}_k$ such that for each $\delta \in (0,1)$,*

$$
\mathbf{P}\{L(g_{n_j}) \geq \epsilon_j\} \geq (1-\delta)\frac{1}{2}\sum_{i=0}^{k-1} \frac{(c\gamma_j)^i}{i!} e^{-c\gamma_j} \quad \textit{for infinitely many } j,
$$

where $c = \log 256$.

Appendix 1. Sharpening and remarks to Section 2

The following is a significantly stronger form of Lemma 1.

Lemma 2 *Let $R_n(z)$ be a sequence of nonnegative numbers parametrized by an abstract parameter z from a set $\mathcal{Z}$, and let Z be a random variable taking its values from $\mathcal{Z}$. If for some $p > 1$, $\mathbf{E}\{R_n(Z)^p\}/\mathbf{E}^p\{R_n(Z)\}$ does not tend to ∞ then either*

$$
\mathbf{P}\left\{ \limsup_{n\to\infty} \frac{R_n(Z)}{\mathbf{E}R_n(Z)} > 1 \right\} > 0
$$

or

$$
\limsup_{n\to\infty} \frac{R_n(Z)}{\mathbf{E}R_n(Z)} = 1 \quad \textit{with probability one.}
$$

Remark. In any case,

$$\mathbf{P}\left\{\limsup_{n\to\infty}\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\geq 1\right\}>0,$$

which implies that there exists a $z\in\mathcal{Z}$ such that for every $\epsilon\in(0,1)$ $R_n(z)>(1-\epsilon)\mathbf{E}R_n(Z)$ for infinitely many n.

Proof. Let $S=\limsup_{n\to\infty}\frac{R_n(Z)}{\mathbf{E}R_n(Z)}$. Assume, on the contrary, that $S\leq 1$ with probability one and $\mathbf{P}\{A\}>0$, where $A=\{S<1\}$. For $\delta>0$, define

$$X=\max\left(\frac{S+1}{2},0\right)I_A+(1+\delta)(1-I_A).$$

Since

$$\begin{aligned}\mathbf{E}X &= \mathbf{E}\left\{\max\left(\frac{S+1}{2},0\right)\Big|\,A\right\}\mathbf{P}\{A\}+(1+\delta)(1-\mathbf{P}\{A\})\\ &= 1-\left(1-\mathbf{E}\left\{\max\left(\frac{S+1}{2},0\right)\Big|\,A\right\}\right)\mathbf{P}\{A\}+\delta(1-\mathbf{P}\{A\}),\end{aligned}$$

and $\left(1-\mathbf{E}\left\{\max\left(\frac{S+1}{2},0\right)\Big|\,A\right\}\right)\mathbf{P}\{A\}>0$, we may choose δ so small that $\mathbf{E}X<1$. Also, $0\leq X\leq 1+\delta$ with probability one. Introduce the events $A_n=\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\leq X\right\}$, and $B_n=\bigcap_{m\geq n}A_m$. Since $S<X$ with probability one, with probability one A_n occurs for sufficiently large n, that is,

$$1=\mathbf{P}\{\bigcup_{n=1}^{\infty}\bigcap_{m\geq n}A_m\}=\mathbf{P}\{\bigcup_{n=1}^{\infty}B_n\}=\lim_{n\to\infty}\mathbf{P}\{B_n\}.$$

Thus, for every k we have $\mathbf{P}\{B_n\}>1-1/k$ if n is sufficiently large. Since $B_n\subseteq A_n$, we have

$$\mathbf{E}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}I_{B_n}\right\}\leq\mathbf{E}\left\{XI_{B_n}\right\},$$

and therefore

$$\begin{aligned}\mathbf{E}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}(1-I_{B_n})\right\} &= \mathbf{E}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\right\}-\mathbf{E}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}I_{B_n}\right\}\\ &\geq \mathbf{E}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\right\}-\mathbf{E}\left\{XI_{B_n}\right\}\\ &\geq 1-\mathbf{E}\{X\}>0.\end{aligned}$$

Let $1/q=1-1/p$, and apply Hölder's inequality for the random variables $\frac{R_n(Z)}{\mathbf{E}R_n(Z)}(1-I_{B_n})$ and $(1-I_{B_n})$:

$$\mathbf{E}^{1/p}\left\{\left(\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\right)^p(1-I_{B_n})\right\}\mathbf{E}^{1/q}\{(1-I_{B_n})\}\geq\mathbf{E}\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}(1-I_{B_n})\right\}.$$

Thus,

$$\begin{aligned}
\frac{\mathbf{E}\{R_n(Z)^p\}}{\mathbf{E}^p R_n(Z)} &= \mathbf{E}\left\{\left(\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\right)^p\right\} \geq \mathbf{E}\left\{\left(\frac{R_n(Z)}{\mathbf{E}R_n(Z)}\right)^p (1 - I_{B_n})\right\} \\
&\geq \mathbf{E}^p\left\{\frac{R_n(Z)}{\mathbf{E}R_n(Z)}(1 - I_{B_n})\right\} \frac{1}{(1 - \mathbf{P}\{B_n\})^{p-1}} > (1 - \mathbf{E}\{X\})^p k^{p-1}
\end{aligned}$$

for sufficiently large n. Therefore, $\mathbf{E}\{R_n(Z)^p\}/\mathbf{E}^p R_n(Z)$ tends to infinity, a contradiction. □

Remark. Note that Lemma 2 is indeed stronger than Lemma 1, since condition (6) implies that $\mathbf{E}\{R_n(Z)^p\}/\mathbf{E}^p\{R_n(Z)\} \not\to \infty$, but not vice versa. However, the inequality

$$\mathbf{P}\left\{\limsup_{n\to\infty} \frac{R_n(Z)}{\mathbf{E}R_n(Z)} \geq 1\right\} > 0.$$

cannot be strengthened even if (6) is assumed in the sense that even under (6) we may have

$$\begin{aligned}
\left\{\limsup_{n\to\infty} \frac{R_n(Z)}{\mathbf{E}R_n(Z)} > 1\right\} &\subseteq \{R_n(Z) > \mathbf{E}R_n(Z) \text{ for infinitely many } n\} \\
&\subseteq \{R_n(Z) \geq \mathbf{E}R_n(Z) \text{ for infinitely many } n\} \\
&= \emptyset.
\end{aligned}$$

To see this, let $Z \sim \text{Uniform}[0,1]$, and consider $R_n(Z) = nI_{\{Z\in(0,1/n)\}} + n - 2$. Then $\mathbf{E}R_n(Z) = n - 1$, $\operatorname{ess\,sup} R_n(Z) = 2n - 2$, but

$$\begin{aligned}
&\{R_n(Z) \geq \mathbf{E}R_n(Z) \text{ for infinitely many } n\} \\
&= \{nI_{\{Z\in(0,1/n)\}} + n - 2 \geq n - 1 \text{ for infinitely many } n\} \\
&= \{nI_{\{Z\in(0,1/n)\}} \geq 1 \text{ for infinitely many } n\} \\
&= \{Z \in (0, 1/n) \text{ for infinitely many } n\} = \emptyset.
\end{aligned}$$

□

Appendix 2. Lemmas for the proof of Theorem 3

The following lemmas are used in the proof of Theorem 3. We use the notation introduced in the text.

Lemma 3 *Let*

$$E_{nj} = (U_{nj}^-, U_{nj}^- + \lambda(B_{nj})) \times \{j\}.$$

For all n, $j \in \{1, \ldots, k\}$, $z \in [0,1]^k$, and data points $X_1, \ldots, X_n$,

$$\lambda(B_{nj} \triangle C_{nj}) \geq \lambda(E_{nj} \triangle C_{nj}).$$

Proof. Clearly, $\lambda(E_{nj}) = \lambda(B_{nj})$. Assume, on the contrary, that

$$\lambda(E_{nj} \bigtriangleup C_{nj}) > \lambda(B_{nj} \bigtriangleup C_{nj}).$$

Then

$$\text{either} \quad \lambda(E_{nj} \cap \overline{C_{nj}}) > \lambda(B_{nj} \cap \overline{C_{nj}}) \quad \text{or} \quad \lambda(\overline{E_{nj}} \cap C_{nj}) > \lambda(\overline{B_{nj}} \cap C_{nj}),$$

where $\overline{A} = [0,1] \times \{j\} - A$ is the complement of a set $A \subset [0,1] \times \{j\}$. In the first case $C_{nj} \subseteq E_{nj}$, so we have

$$\lambda(E_{nj}) = \lambda(E_{nj} \cap \overline{C_{nj}}) + \lambda(E_{nj} \cap C_{nj}) > \lambda(B_{nj} \cap \overline{C_{nj}}) + \lambda(C_{nj}) \geq \lambda(B_{nj}),$$

a contradiction. In the second case $E_{nj} \subseteq C_{nj}$. Then similarly to the first case,

$$\lambda(\overline{E_{nj}}) = \lambda(\overline{E_{nj}} \cap \overline{C_{nj}}) + \lambda(\overline{E_{nj}} \cap C_{nj}) > \lambda(\overline{C_{nj}}) + \lambda(\overline{B_{nj}} \cap C_{nj}) \geq \lambda(\overline{B_{nj}}),$$

again a contradiction. $\square$

Lemma 4

$$\mathbf{P}\left\{ \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - Z_j| > \epsilon \,\middle|\, D_n(Z) \right\} \geq \frac{1}{2} I_{\{\frac{1}{k}\sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon\}}.$$

Proof. Since given D_n, the Z_j's are independent and uniform on the sets A'_{nj},

$$\begin{aligned}
&\mathbf{P}\left\{ \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - Z_j| > \epsilon \,\middle|\, D_n(Z) \right\} \\
&= \frac{\lambda_k(\{z \in \bigotimes_{j=1}^{k} A'_{nj} : \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - z_j| > \epsilon\})}{\lambda_k(\bigotimes_{j=1}^{k} A'_{nj})}
\end{aligned}$$

where λ is the one-dimensional, and λ_k is the k-dimensional Lebesgue measure. Define $M_{nj} = \frac{1}{2}(U_{nj}^{-} + U_{nj}^{+})$, and

$$T_n \stackrel{\text{def}}{=} \bigotimes_{j=1}^{k} (M_{nj}, U_{nj}^{+}).$$

Then

$$\begin{aligned}
&\frac{\lambda_k(\{z \in \bigotimes_{j=1}^{k} A'_{nj} : \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - z_j| > \epsilon\})}{\lambda_k(\bigotimes_{j=1}^{k} A'_{nj})} \geq \\
&\geq \frac{\lambda_k(\{z \in \bigotimes_{j=1}^{k} A'_{nj} : \frac{1}{k}\sum_{j=1}^{k} |z_j - M_{nj}| > \epsilon\})}{\lambda_k(\bigotimes_{j=1}^{k} A'_{nj})} \\
&= \frac{2^k \lambda_k(\{z \in T_n : \frac{1}{k}\sum_{j=1}^{k} (z_j - M_{nj}) > \epsilon\})}{2^k \lambda_k(T_n)} \\
&\geq \frac{\lambda_k(\{z \in T_n : \frac{1}{k}\sum_{j=1}^{k} (z_j - M_{nj}) > \frac{1}{k}\sum_{j=1}^{k} \frac{\xi_{nj}}{4}\})}{\lambda_k(T_n)} \\
&= \frac{\lambda_k(T_n \cap \{z : \sum_{j=1}^{k} (z_j - (M_{nj} + \frac{\xi_{nj}}{4})) > 0\})}{\lambda_k(T_n)},
\end{aligned}$$

whenever $\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon$. Observe that the last expression equals $1/2$, since the numerator is the volume of the intersection of the rectangle T_n with a halfspace defined by a hyperplane containing the center of the rectangle. The proof is complete. □

Lemma 5 *Let the random variables ξ_{nj} and ξ'_{nj} be as defined in the proof of Theorem 3. Then for all ϵ,*

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq \epsilon, \forall N_j > 0\right\} \geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon, \forall N_j > 0\right\} - ke^{-\frac{n}{k}\left(1-\frac{\sqrt{\epsilon}}{2}\right)}.$$

Proof. It is easy to see that for all $N_j > 0$,

$$\mathbf{P}\{\xi_{nj} \geq \epsilon | N_j\} \geq \mathbf{P}\{\xi'_{nj} \geq \epsilon | N_j\},$$

whenever $\epsilon \leq 1/2$, and we have equality for $\epsilon = 1/2$. Thus, if $\forall N_j > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq \epsilon \,\middle|\, \max_{j\leq k}\xi_{nj} < \frac{1}{2}, N_1,\ldots,N_k\right\}\mathbf{P}\left\{\max_{j\leq k}\xi_{nj} < \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, \max_{j\leq k}\xi'_{nj} < \frac{1}{2}, N_1,\ldots,N_k\right\}\mathbf{P}\left\{\max_{j\leq k}\xi'_{nj} < \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&= \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon, \max_{j\leq k}\xi'_{nj} < \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&= \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} \\
&\quad -\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon, \max_{j\leq k}\xi'_{nj} \geq \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} - \mathbf{P}\left\{\max_{j\leq k}\xi'_{nj} \geq \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} - \sum_{j=1}^{k}\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2} \,\middle|\, N_j\right\}.
\end{aligned}$$

Clearly, for each j with $N_j > 0$,

$$\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2} \,\middle|\, N_j\right\} = e^{-\lambda_j/2} \leq e^{-N_j \log 2 + N_j/2}.$$

Using the fact that N_j is a binomial random variable with parameters n and $1/k$, we get, by straightforward calculation, that

$$\begin{aligned}
\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2}, N_j > 0\right\} &= \mathbf{E}\left\{I_{\{N_j>0\}}\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2}\middle| N_j\right\}\right\} \\
&\leq \mathbf{E}\left\{e^{-N_j(\log 2 - 1/2)}\right\} \\
&= \left[1 - \frac{1}{k}\left(1 - \frac{\sqrt{e}}{2}\right)\right]^n \\
&\leq e^{-\frac{n}{k}\left(1-\frac{\sqrt{e}}{2}\right)}.
\end{aligned}$$

Taking expected values, we get the desired inequality. □

Lemma 6 *Let $\xi_1, \ldots, \xi_k$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_k > 0$, respectively. Then for each ϵ,*

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_j > \epsilon\right\} \geq \mathbf{P}\left\{\frac{\sum_{j=1}^{k}\lambda_j\xi_j}{\sum_{j=1}^{k}\lambda_j} > \epsilon\right\} = \mathbf{P}\{\Phi_k > k\epsilon\lambda\},$$

where $\lambda = \frac{1}{k}\sum_{j=1}^{k}\lambda_j$, and the random variable Φ_k has k-th order gamma distribution with parameter 1.

Proof. We prove the lemma by induction for k. We will use two simple facts:

FACT 1. Let η, ξ, ξ' be real-valued random variables such that η is independent of (ξ, ξ') and $\mathbf{P}\{\xi > x\} \geq \mathbf{P}\{\xi' > x\}$ for all $x \in \mathcal{R}$. Then $\mathbf{P}\{\xi + \eta > x\} \geq \mathbf{P}\{\xi' + \eta > x\}$ for all x.

FACT 2. Let ξ_1 and ξ_2 be independent, exponential random variables with parameters λ_1 and λ_2, respectively. Assume $0 < \lambda_1 \leq \lambda_2$, and let $\delta = (\lambda_2 - \lambda_1)/2$. Then for all $\epsilon > 0$, the probability $\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\}$ is monotone increasing in δ (while holding $\lambda_1 + \lambda_2$ fixed). In particular, $\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} \geq \mathbf{P}\{\Phi_2 > \lambda'\epsilon\}$ for $\lambda' = (\lambda_2 + \lambda_1)/2$.

PROOF. Straightforward calculation shows that for $\delta > 0$,

$$\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} = \frac{\lambda_2 e^{-\lambda_1\epsilon} - \lambda_1 e^{-\lambda_2\epsilon}}{\lambda_2 - \lambda_1} = e^{-\lambda'\epsilon}\left(\lambda'\epsilon\frac{\sinh(\delta\epsilon)}{\delta\epsilon} + \cosh(\delta\epsilon)\right),$$

and for $\delta = 0$,

$$\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} = \mathbf{P}\{\Phi_2 > \lambda'\epsilon\} = (1 + \lambda'\epsilon)e^{-\lambda'\epsilon}.$$

Since $\sinh(x)/x$ and $\cosh(x)$ are monotone increasing on $[0, \infty)$, Fact 2 follows.

Now we are ready to prove the lemma. The statement is trivially true for $k = 1$, and by Fact 2 for $k = 2$. Let $k \geq 3$, and assume that the statement is true for $k - 1$. There exist two indices $j, j' \leq k$ such that $\lambda_j \leq \lambda$ and $\lambda_{j'} \geq \lambda$. Without loss of generality, we assume that $\lambda_1 \leq \lambda$ and $\lambda_2 \geq \lambda$. Let ξ_1' and ξ_2' be independent exponential random variables with parameter λ and $\lambda_1 + \lambda_2 - \lambda$, also independent of all ξ_j. Since

$$\frac{|\lambda_1 - \lambda_2|}{2} \geq \left| \lambda - \frac{\lambda_1 + \lambda_2}{2} \right|,$$

Fact 2 implies that

$$\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} \geq \mathbf{P}\{\xi_1' + \xi_2' > \epsilon\}.$$

Also, by the inductive assumption,

$$\mathbf{P}\left\{\xi_2' + \sum_{j=3}^{k} \xi_j > \epsilon\right\} \geq \mathbf{P}\{\Phi_{k-1}/\lambda > \epsilon\}.$$

Using these and Fact 1 twice, we obtain

$$\begin{aligned} \mathbf{P}\left\{\sum_{j=1}^{k} \xi_j > \epsilon\right\} &\geq \mathbf{P}\left\{\xi_1' + \xi_2' + \sum_{j=3}^{k} \xi_j > \epsilon\right\} \\ &\geq \mathbf{P}\{\xi_1' + \Phi_{k-1}/\lambda > \epsilon\} \\ &= \mathbf{P}\{\Phi_k > \lambda\epsilon\}. \end{aligned}$$

$\square$

Acknowledgements. We thank Peter Bartlett for pointing out reference [10].

References

[1] A. Antos and G. Lugosi. Strong minimax lower bounds for learning. In *Proceedings of the Nineth Annual ACM Conference on Computational Learning Theory*, pages 303–309. Association for Computing Machinery, New York, 1996.

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

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

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

[5] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018, 1995.

[6] 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.

[7] D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994.

[8] G. Lugosi. Improved upper bounds for probabilities of uniform deviations. *Statistics and Probability Letters*, 25:71–77, 1995.

[9] R.-D. Reiss. *Approximate distributions of order statistics.* Springer-Verlag, New York, 1989.

[10] D. Schuurmans. Characterizing rational versus exponential learning curves. In *Computational Learning Theory: Second European Conference. EuroCOLT'95*, pages 272–286. Springer Verlag, 1995.

[11] J. Shawe-Taylor, M. Anthony, and N.L. Biggs. Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 42:65–73, 1993.

[12] 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.