PROOF. As in the proof of Theorem 1.9, we see that

$$\begin{aligned}
&\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\right\}\\
&\quad\le\ \frac{1}{n}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)\right|\right\}\\
&\quad\le\ \frac{1}{n}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\mathbb{I}_{[X_i\in A]}\right|\right\}+\frac{1}{n}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\mathbb{I}_{[X_i'\in A]}\right|\right\}\\
&\quad=\ \frac{2}{n}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\mathbb{I}_{[X_i\in A]}\right|\right\}\\
&\quad=\ \frac{2}{n}\mathbb{E}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\mathbb{I}_{[X_i\in A]}\right|\,\middle|\,X_1,\ldots,X_n\right\}.
\end{aligned}$$

Just as in the proof of theorem 1, we fix the values $X_1=x_1,\ldots,X_n=x_n$ and study

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\mathbb{I}_{[x_i\in A]}\right|\right\}=\mathbb{E}\left\{\max_{b\in\mathcal{A}(x_1^n)}\left|\sum_{i=1}^n\sigma_i b_i\right|\right\}\ .$$

Now let $B_0\stackrel{\text{def}}{=}\{b^{(0)}\}$ be the singleton set containing the all-zero vector $b^{(0)}=(0,\ldots,0)$, and let $B_1,B_2,\ldots,B_M$ be subsets of $\{0,1\}^n$ such that each B_k is a minimal cover of $\mathcal{A}(x_1^n)$ of radius 2^{-k}, and $M=\lfloor\log_2\sqrt{n}\rfloor+1$. Note that B_0 is also a cover of radius 2^0, and that $B_M=\mathcal{A}(x_1^n)$. Now denote the (random) vector reaching the maximum by $b^*=(b_1^*,\ldots,b_n^*)\in\mathcal{A}(x_1^n)$, that is,

$$\left|\sum_{i=1}^n\sigma_i b_i^*\right|=\max_{b\in\mathcal{A}(x_1^n)}\left|\sum_{i=1}^n\sigma_i b_i\right|\ ,$$

and, for each $k\le M$, let $b^{(k)}\in B_k$ be a nearest neighbor of b^* in the k-th cover, that is,

$$\rho(b^{(k)},b^*)\le\rho(b,b^*)\quad\text{for all } b\in B_k.$$

Note that $\rho(b^{(k)},b^*)\le 2^{-k}$, and therefore

$$\rho(b^{(k)},b^{(k-1)})\le\rho(b^{(k)},b^*)+\rho(b^{(k-1)},b^*)\le 3\cdot 2^{-k}\ .$$

Now clearly,

$$\begin{aligned}
\sum_{i=1}^n\sigma_i b_i^* &= \sum_{i=1}^n\sigma_i b_i^{(0)}+\sum_{k=1}^M\sum_{i=1}^n\sigma_i\left(b_i^{(k)}-b_i^{(k-1)}\right)\\
&= \sum_{k=1}^M\sum_{i=1}^n\sigma_i\left(b_i^{(k)}-b_i^{(k-1)}\right)\ ,
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}\left\{\max_{b\in\mathcal{A}(x_1^n)}\left|\sum_{i=1}^n \sigma_i b_i\right|\right\} &= \mathbb{E}\left|\sum_{k=1}^M\sum_{i=1}^n \sigma_i\left(b_i^{(k)}-b_i^{(k-1)}\right)\right| \\
&\le \sum_{k=1}^M \mathbb{E}\left|\sum_{i=1}^n \sigma_i\left(b_i^{(k)}-b_i^{(k-1)}\right)\right| \\
&\le \sum_{k=1}^M \mathbb{E}\max_{b\in B_k, c\in B_{k-1}:\rho(b,c)\le 3\cdot 2^{-k}}\left|\sum_{i=1}^n \sigma_i\left(b_i-c_i\right)\right| .
\end{aligned}$$

Now it follows from Lemma 1.2 that for each pair $b\in B_k, c\in B_{k-1}$ with $\rho(b,c)\le 3\cdot 2^{-k}$, and for all $s>0$,

$$e^{s\sum_{i=1}^n \sigma_i(b_i-c_i)} \le e^{s^2 n\left(3\cdot 2^{-k}\right)^2/2}.$$

On the other hand, the number of such pairs is bounded by $|B_k|\cdot|B_{k-1}| \le |B_k|^2 = N(2^{-k},\mathcal{A}(x_1^n))^2$. Then Lemma 1.3 implies that for each $1\le k\le M$,

$$\mathbb{E}\max_{b\in B_k, c\in B_{k-1}:\rho(b,c)\le 3\cdot 2^{-k}}\left|\sum_{i=1}^n \sigma_i\left(b_i-c_i\right)\right| \le 3\sqrt{n}2^{-k}\sqrt{2\log 2N(2^{-k},\mathcal{A}(x_1^n))^2} .$$

Summarizing, we obtain

$$\begin{aligned}
\mathbb{E}\left\{\max_{b\in\mathcal{A}(x_1^n)}\left|\sum_{i=1}^n \sigma_i b_i\right|\right\} &\le 3\sqrt{n}\sum_{k=1}^M 2^{-k}\sqrt{2\log 2N(2^{-k},\mathcal{A}(x_1^n))^2} \\
&\le 12\sqrt{n}\sum_{k=1}^\infty 2^{-(k+1)}\sqrt{\log 2N(2^{-k},\mathcal{A}(x_1^n))} \\
&\le 12\sqrt{n}\int_0^1 \sqrt{\log 2N(r,\mathcal{A}(x_1^n))}\,dr ,
\end{aligned}$$

where at the last step we used the fact that $N(r,\mathcal{A}(x_1^n))$ is a monotonically decreasing function of r. The proof is finished. □

To complete our argument, we need to relate the VC dimension of a class of sets $\mathcal{A}$ to the covering numbers $N(r,\mathcal{A}(x_1^n))$ appearing in Theorem 3.10.

Theorem 1.17. *Let $\mathcal{A}$ be a class of sets with* VC *dimension $V<\infty$. For every $x_1,\dots,x_n\in\mathcal{R}^d$ and $0\le r\le 1$,*

$$N(r,\mathcal{A}(x_1^n)) \le \left(\frac{4e}{r^2}\right)^{V/(1-1/e)} .$$

Theorem 1.17 is due to Dudley (1978). Haussler (1995) refined Dudley's probabilistic argument and showed that the stronger bound

$$N(r, \mathcal{A}(x_1^n)) \le e(V+1)\left(\frac{2e}{r^2}\right)^V .$$

also holds.

PROOF. Fix $x_1, \ldots, x_n$, and consider the set $B_0 = \mathcal{A}(x_1^n) \in \{0,1\}^n$. Fix $r \in (0,1)$, and let $B_r \subset \{0,1\}^n$ be a minimal cover of B_0 of radius r with respect to the metric

$$\rho(b,c) = \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[b_i \neq c_i]}}.$$

We need to show that $|B_r| \le \left(4e/r^2\right)^{V/(1-1/e)}$.

First note that there exists a "packing set" $C_r \subset B_0$ such that $|B_r| \le |C_r|$ and any two elements $b, c \in C_r$ are *r-separated*, that is, $\rho(b,c) > r$. To see this, suppose that C_r is such an r-separated set of maximal cardinality. Then for any $b \in B_0$, there exists a $c \in C_r$ with $\rho(b,c) \le r$, since otherwise adding b to the set C_r would increase its cardinality, and it would still be r-separated. Thus, C_r is a cover of radius r, which implies that $|B_r| \le |C_r|$. Denote the elements of C_r by $c^{(1)}, \ldots, c^{(M)}$, where $M = |C_r|$. For any $i, j \le M$, define $A_{i,j}$ as the set of indices where the binary vectors $c^{(i)}$ and $c^{(j)}$ disagree:

$$A_{i,j} = \left\{1 \le m \le n : c_m^{(i)} \neq c_m^{(j)}\right\} .$$

Note that any two elements of C_r differ in at least nr^2 components. Next define K independent random variables $Y_1, \ldots, Y_K$, distributed uniformly over the set $\{1, 2, \ldots, n\}$, where K will be specified later. Then for any $i, j \le M$, $i \neq j$, and $k \le K$,

$$\mathbb{P}\{Y_k \in A_{i,j}\} \ge r^2 ,$$

and therefore the probability that no one of $Y_1, \ldots, Y_K$ falls in the set $A_{i,j}$ is less than $\left(1 - r^2\right)^K$. Observing that there are less than M^2 sets $A_{i,j}$, and applying the union bound, we obtain that

$$\begin{aligned}\mathbb{P}\{&\text{for all } i \neq j, i, j \le M, \text{at least one } Y_k \text{ falls in } A_{i,j}\} \\ &\ge 1 - M^2(1-r^2)^K \ge 1 - M^2 e^{-Kr^2} .\end{aligned}$$

If we choose $K = \lceil 2\log M/r^2 \rceil + 1$, then the above probability is strictly positive. This implies that there exist $K = \lceil 2\log M/r^2 \rceil + 1$ indices $y_1, \ldots, y_K \in \{1, 2, \ldots, n\}$ such that at least one y_k falls in each set $A_{i,j}$. Therefore, restricted to the K components $y_1, \ldots, y_K$,

the elements of C_r are all different, and since $C_r \subset B_0$, C_r does not shatter any set of size larger than V. Therefore, by Sauer's lemma we obtain

$$|C_r| = M \le \left(\frac{eK}{V}\right)^V$$

for $K \le V$. Thus, if $\log M \ge V$, then

$$\begin{aligned}
\log M &\le V \log \frac{e\left(\lceil 2\log M/r^2 \rceil + 1\right)}{V} \\
&\le V\left(\log \frac{4e}{r^2} + \log \frac{\log M}{V}\right) \\
&\le V \log \frac{4e}{r^2} + \frac{1}{e}\log M \quad (\text{since } \log x \le x/e \text{ for } x > 0) \ .
\end{aligned}$$

Therefore,

$$\log M \le \frac{V}{1-1/e} \log \frac{4e}{r^2} \ .$$

If $\log M < V$, then the above inequality holds trivially. This concludes the proof. □

Combining this result with Theorem 3.10 we obtain that for any class $\mathcal{A}$ with VC dimension V,

$$\mathbb{E}\left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right\} \le c\sqrt{\frac{V}{n}} \ ,$$

where c is a universal constant.

1.5 Minimax lower bounds

The purpose of this section is to investigate how good the bounds obtained in the previous chapter for empirical risk minimization are. We have seen that for any class $\mathcal{C}$ of classifiers with VC dimension V, a classifier g_n^* minimizing the empirical risk satisfies

$$\mathbb{E}L(g_n^*) - L_{\mathcal{C}} \le O\left(\sqrt{\frac{L_{\mathcal{C}} V_{\mathcal{C}} \log n}{n}} + \frac{V_{\mathcal{C}} \log n}{n}\right),$$

and also

$$\mathbb{E}L(g_n^*) - L_{\mathcal{C}} \le O\left(\sqrt{\frac{V_{\mathcal{C}}}{n}}\right).$$

In this section we seek answers for the following questions: Are these upper bounds (at least up to the order of magnitude) tight? Is there a much better way of selecting a classifier than minimizing the empirical error?

Let us formulate exactly what we are interested in. Let $\mathcal{C}$ be a class of decision functions $g : \mathcal{R}^d \to \{0,1\}$. The training sequence $D_n = ((X_1,Y_1),\ldots,(X_n,Y_n))$ is used to select the classifier $g_n(X) = g_n(X,D_n)$ from $\mathcal{C}$, where the selection is based on the data D_n. We emphasize here that g_n can be an arbitrary function of the data, we do not restrict our attention to empirical error minimization, where g_n is a classifier in $\mathcal{C}$ that minimizes the number errors committed on the data D_n.

As before, we measure the performance of the selected classifier by the difference between the error probability $L(g_n) = \mathbb{P}\{g_n(X) \neq Y|D_n\}$ of the selected classifier and that of the best in the class, $L_{\mathcal{C}}$. In particular, we seek lower bounds for

$$\sup \mathbb{E}L(g_n) - L_{\mathcal{C}},$$

where the supremum is taken over all possible distributions of the pair (X,Y). A lower bound for this quantities means that no matter what our method of picking a rule from $\mathcal{C}$ is, we may face a distribution such that our method performs worse than the bound.

Actually, we investigate a stronger problem, in that the supremum is taken over all distributions with $L_{\mathcal{C}}$ kept at a fixed value between zero and $1/2$. We will see that the bounds depend on n, $V_{\mathcal{C}}$, and $L_{\mathcal{C}}$ jointly. As it turns out, the situations for $L_{\mathcal{C}} > 0$ and $L_{\mathcal{C}} = 0$ are quite different. Because of its simplicity, we first treat the case $L_{\mathcal{C}} = 0$. All the proofs are based on a technique called "the probabilistic method." The basic idea here is that the existence of a "bad" distribution is proved by considering a large class of distributions, and bounding the average behavior over the class.

1.5.1 The zero-error case

Here we obtain lower bounds under the assumption that the best classifier in the class has zero error probability. Recall that by Corollary 1.2 the expected probability of error of an empirical risk minimizer is bounded by $O(V_{\mathcal{C}} \log n/n)$. Next we obtain minimax lower bounds close to the upper bounds.

Theorem 1.18. *Let $\mathcal{C}$ be a class of discrimination functions with* VC *dimension V. Let $\mathcal{X}$ be the set of all random variables (X,Y) for which $L_{\mathcal{C}} = 0$. Then, for every discrimination rule g_n based upon $X_1,Y_1,\ldots,X_n,Y_n$, and $n \geq V-1$,*

$$\sup_{(X,Y)\in\mathcal{X}} \mathbb{E}L(g_n) \geq \frac{V-1}{2en}\left(1-\frac{1}{n}\right) .$$

PROOF. The idea is to construct a family $\mathcal{F}$ of 2^{V-1} distributions within the distributions with $L_{\mathcal{C}} = 0$ as follows: first find points $x_1,\ldots,x_V$ that are shattered by $\mathcal{C}$. Each distribution in $\mathcal{F}$ is concentrated on the set of these points. A member in $\mathcal{F}$ is described by $V-1$ bits,

$b_1, \ldots, b_{V-1}$. For convenience, this is represented as a bit vector b. Assume $V-1 \le n$. For a particular bit vector, we let $X = x_i$ $(i < V)$ with probability $1/n$ each, while $X = x_V$ with probability $1-(V-1)/n$. Then set $Y = f_b(X)$, where f_b is defined as follows:

$$f_b(x) = \begin{cases} b_i & \text{if } x = x_i, i < V \\ 0 & \text{if } x = x_V. \end{cases}$$

Note that since Y is a function of X, we must have $L^* = 0$. Also, $L_{\mathcal{C}} = 0$, as the set $\{x_1, \ldots, x_V\}$ is shattered by $\mathcal{C}$, i.e., there is a $g \in \mathcal{C}$ with $g(x_i) = f_b(x_i)$ for $1 \le i \le V$. Clearly,

$$\begin{aligned} &\sup_{(X,Y):L_{\mathcal{C}}=0} \mathbb{E}\{L(g_n) - L_{\mathcal{C}}\} \\ &\quad \ge \sup_{(X,Y)\in\mathcal{F}} \mathbb{E}\{L(g_n) - L_{\mathcal{C}}\} \\ &\quad = \sup_b \mathbb{E}\{L(g_n) - L_{\mathcal{C}}\} \\ &\quad \ge \mathbb{E}\{L(g_n) - L_{\mathcal{C}}\} \\ &\qquad \text{(where } b \text{ is replaced by } B\text{, uniformly distributed over } \{0,1\}^{V-1}) \\ &\quad = \mathbb{E}\{L(g_n)\}\ , \\ &\quad = \mathbb{P}\{g_n(X, X_1, Y_1, \ldots, X_n, Y_n) \ne f_B(X)\}\ . \end{aligned}$$

The last probability may be viewed as the error probability of the decision function $g_n : \mathcal{R}^d \times (\mathcal{R}^d \times \{0,1\})^n \to \{0,1\}$ in predicting the value of the random variable $f_B(X)$ based on the observation $Z_n = (X, X_1, Y_1, \ldots, X_n, Y_n)$. Naturally, this probability is bounded from below by the Bayes probability of error

$$L^*(Z_n, f_B(X)) = \inf_{g_n} \mathbb{P}\{g_n(Z_n) \ne f_B(X)\}$$

corresponding to the decision problem $(Z_n, f_B(X))$. By the results of Chapter 1,

$$L^*(Z_n, f_B(X)) = \mathbb{E}\{\min(\eta^*(Z_n), 1-\eta^*(Z_n))\}\ ,$$

where $\eta^*(Z_n) = \mathbb{P}\{f_B(X) = 1 | Z_n\}$. Observe that

$$\eta^*(Z_n) = \begin{cases} 1/2 & \text{if } X \ne X_1, \ldots, X \ne X_n, X \ne x_V \\ 0 \text{ or } 1 & \text{otherwise.} \end{cases}$$

Thus, we see that

$$\begin{aligned}
\sup_{(X,Y):L_{\mathcal{C}}=0} \mathbb{E}\{L(g_n) - L_{\mathcal{C}}\} &\geq L^*(Z_n, f_B(X)) \\
&= \frac{1}{2}\mathbb{P}\{X \neq X_1, \ldots, X \neq X_n, X \neq x_V\} \\
&= \frac{1}{2}\sum_{i=1}^{V-1} \mathbb{P}\{X = x_i\}(1 - \mathbb{P}\{X = x_i\})^n \\
&= \frac{V-1}{2n}(1 - 1/n)^n \\
&\geq \frac{V-1}{2en}\left(1 - \frac{1}{n}\right) \quad (\text{since } (1-1/n)^{n-1} \downarrow 1/e).
\end{aligned}$$

This concludes the proof. □

1.5.2 The general case

In the more general case, when the best decision in the class $\mathcal{C}$ has positive error probability, the upper bounds derived in Chapter 2 for the expected error probability of the classifier obtained by minimizing the empirical risk are much larger than when $L_{\mathcal{C}} = 0$. Theorem 1.19 below gives a lower bound for $\sup_{(X,Y):L_{\mathcal{C}} \text{ fixed}} \mathbb{E}L(g_n) - L_{\mathcal{C}}$. As a function of n and $V_{\mathcal{C}}$, the bound decreases basically as in the upper bound obtained from Theorem 1.11. Interestingly, the lower bound becomes smaller as $L_{\mathcal{C}}$ decreases, as should be expected. The bound is largest when $L_{\mathcal{C}}$ is close to $1/2$.

Theorem 1.19. *Let $\mathcal{C}$ be a class of discrimination functions with* VC *dimension $V \geq 2$. Let $\mathcal{X}$ be the set of all random variables (X, Y) for which for fixed $L \in (0, 1/2)$,*

$$L = \inf_{g \in \mathcal{C}} \mathbb{P}\{g(X) \neq Y\} .$$

Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$,

$$\sup_{(X,Y)\in\mathcal{X}} \mathbb{E}(L(g_n) - L) \geq \sqrt{\frac{L(V-1)}{24n}} e^{-8} \quad \text{if } n \geq \tfrac{V-1}{2L} \max(9, 1/(1-2L)^2).$$

Proof. Again we consider the finite family $\mathcal{F}$ from the previous section. The notation b and B is also as above. X now puts mass p at x_i, $i < V$, and mass $1 - (V-1)p$ at x_V. This imposes the condition $(V-1)p \leq 1$, which will be satisfied. Next introduce the constant $c \in (0, 1/2)$. We no longer have Y as a function of X. Instead, we have a uniform $[0, 1]$

random variable U independent of X and define

$$Y = \begin{cases} 1 & \text{if } U \le \frac{1}{2} - c + 2cb_i,\ X = x_i,\ i < V \\ 0 & \text{otherwise.} \end{cases}$$

Thus, when $X = x_i$, $i < V$, Y is 1 with probability $1/2 - c$ or $1/2 + c$. A simple argument shows that the best rule for b is the one which sets

$$f_b(x) = \begin{cases} 1 & \text{if } x = x_i\ ,\ i < V,\ b_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Also, observe that

$$L = (V-1)p(1/2 - c)\ .$$

Noting that $|2\eta(x_i) - 1| = c$ for $i < V$, for fixed b, we may write

$$L(g_n) - L \ge \sum_{i=1}^{V-1} 2pcI_{\{g_n(x_i, X_1, Y_1, \ldots, X_n, Y_n) = 1 - f_b(x_i)\}}\ .$$

It is sometimes convenient to make the dependence of g_n upon b explicit by considering $g_n(x_i)$ as a function of $x_i, X_1, \ldots, X_n, U_1, \ldots, U_n$ (an i.i.d. sequence of uniform $[0,1]$ random variables), and b_i. We replace b by a uniformly distributed random B over $\{0,1\}^{V-1}$. After this randomization, denote $Z_n = (X, X_1, Y_1, \ldots, X_n, Y_n)$. Thus,

$$\begin{aligned} \sup_{(X,Y)\in\mathcal{F}} \mathbb{E}\{L(g_n) - L\} &= \sup_b \mathbb{E}\{L(g_n) - L\} \\ &\ge \mathbb{E}\{L(g_n) - L\} \quad \text{(with random } B\text{)} \\ &\ge \sum_{i=1}^{V-1} 2pc\mathbb{E}I_{\{g_n(x_i, X_1, \ldots, Y_n) = 1 - f_B(x_i)\}} \\ &= 2c\mathbb{P}\{g_n(Z_n) \ne f_B(X)\} \\ &\ge 2cL^*(Z_n, f_B(X)), \end{aligned}$$

where, as before, $L^*(Z_n, f_B(X))$ denotes the Bayes probability of error of predicting the value of $f_B(X)$ based on observing Z_n. All we have to do is to find a suitable lower bound for

$$L^*(Z_n, f_B(X)) = \mathbb{E}\left\{\min(\eta^*(Z_n), 1 - \eta^*(Z_n))\right\},$$

where $\eta^*(Z_n) = \mathbb{P}\{f_B(X) = 1 | Z_n\}$. Observe that

$$\eta^*(Z_n) = \begin{cases} 1/2 & \text{if } X \ne X_1, \ldots, X \ne X_n \text{ and } X \ne x_V \\ \mathbb{P}\{B_i = 1 | Y_{i_1}, \ldots, Y_{i_k}\} & \text{if } X = X_{i_1} = \cdots = X_{i_k} = x_i,\ i < V. \end{cases}$$

Next we compute $\mathbb{P}\{B_i = 1 | Y_{i_1} = y_1, \ldots, Y_{i_k} = y_k\}$ for $y_1, \ldots, y_k \in \{0,1\}$. Denoting the numbers of zeros and ones by $k_0 = |\{j \le k : y_j = 0\}|$ and $k_1 = |\{j \le k : y_j = 1\}|$, we see that

$$\begin{aligned}
&\mathbb{P}\{B_i = 1 | Y_{i_1} = y_1, \ldots, Y_{i_k} = y_k\} \\
&\quad = \frac{(1-2c)^{k_1}(1+2c)^{k_0}}{(1-2c)^{k_1}(1+2c)^{k_0} + (1+2c)^{k_1}(1-2c)^{k_0}}.
\end{aligned}$$

Therefore, if $X = X_{i_1} = \cdots = X_{i_k} = x_i$, $i < V$, then

$$\begin{aligned}
&\min(\eta^*(Z_n), 1 - \eta^*(Z_n)) \\
&\quad = \frac{\min\left((1-2c)^{k_1}(1+2c)^{k_0}, (1+2c)^{k_1}(1-2c)^{k_0}\right)}{(1-2c)^{k_1}(1+2c)^{k_0} + (1+2c)^{k_1}(1-2c)^{k_0}} \\
&\quad = \frac{\min\left(1, \left(\frac{1+2c}{1-2c}\right)^{k_1-k_0}\right)}{1 + \left(\frac{1+2c}{1-2c}\right)^{k_1-k_0}} \\
&\quad = \frac{1}{1 + \left(\frac{1+2c}{1-2c}\right)^{|k_1-k_0|}}.
\end{aligned}$$

In summary, denoting $a = (1+2c)/(1-2c)$, we have

$$\begin{aligned}
L^*(Z_n, f_B(X)) &= \mathbb{E}\left\{ \frac{1}{1 + a^{\left|\sum_{j:X_j = X}(2Y_j - 1)\right|}} \right\} \\
&\ge \mathbb{E}\left\{ \frac{1}{2a^{\left|\sum_{j:X_j = X}(2Y_j - 1)\right|}} \right\} \\
&\ge \frac{1}{2}\sum_{i=1}^{V-1} \mathbb{P}\{X = x_i\} \mathbb{E}\left\{ a^{-\left|\sum_{j:X_j = x_i}(2Y_j - 1)\right|} \right\} \\
&\ge \frac{1}{2}(V-1)pa^{-\mathbb{E}\left\{\left|\sum_{j:X_j = x_i}(2Y_j - 1)\right|\right\}} \\
&\quad \text{(by Jensen's inequality).}
\end{aligned}$$

Next we bound $\mathbb{E}\left\{\left|\sum_{j:X_j = x_i}(2Y_j - 1)\right|\right\}$. Clearly, if $B(k,q)$ denotes a binomial random variable with parameters k and q,

$$\mathbb{E}\left\{\left|\sum_{j:X_j = x_i}(2Y_j - 1)\right|\right\} = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\mathbb{E}\{|2B(k, 1/2 - c) - k|\}.$$

However, by straightforward calculation we see that

$$\begin{aligned}\mathbb{E}\{|2B(k,1/2-c)-k|\} &\le \sqrt{\mathbb{E}\{(2B(k,1/2-c)-k)^2\}} \\ &= \sqrt{k(1-4c^2)+4k^2c^2} \\ &\le 2kc+\sqrt{k}.\end{aligned}$$

Therefore, applying Jensen's inequality once again, we get

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\mathbb{E}\{|2B(k,1/2-c)-k|\} \le 2npc+\sqrt{np}.$$

Summarizing what we have obtained so far, we have

$$\begin{aligned}\sup_b \mathbb{E}\{L(g_n)-L\} &\ge 2cL^*(Z_n,f_B(X)) \\ &\ge 2c\frac{1}{2}(V-1)pa^{-2npc-\sqrt{np}} \\ &\ge c(V-1)pe^{-2npc(a-1)-(a-1)\sqrt{np}} \\ &\quad \text{(by the inequality } 1+x\le e^x\text{)} \\ &= c(V-1)pe^{-8npc^2/(1-2c)-4c\sqrt{np}/(1-2c)}.\end{aligned}$$

A rough asymptotic analysis shows that the best asymptotic choice for c is given by

$$c=\frac{1}{\sqrt{4np}}\ .$$

Then the constraint $L=(V-1)p(1/2-c)$ leaves us with a quadratic equation in c. Instead of solving this equation, it is more convenient to take $c=\sqrt{(V-1)/(8nL)}$. If $2nL/(V-1)\ge 9$, then $c\le 1/6$. With this choice for c, using $L=(V-1)p(1/2-c)$, straightforward calculation provides

$$\sup_{(X,Y)\in\mathcal{F}}\mathbb{E}(L(g_n)-L)\ge\sqrt{\frac{(V-1)L}{24n}}e^{-8}.$$

The condition $p(V-1)\le 1$ implies that we need to ask that $n\ge(V-1)/(2L(1-2L)^2)$. This concludes the proof of Theorem 1.19. □

1.6 Complexity regularization

This section deals with the problem of automatic model selection. Our goal is to develop some data-based methods to find the class $\mathcal{C}$ of classifiers in a way that approximately minimizes the probability of error of the empirical risk minimizer.

1.6.1 Model selection by penalization

In empirical risk minimization one selects a classifier from a given class $\mathcal{C}$ by minimizing the error estimate $\widehat{L}_n(g)$ over all $g \in \mathcal{C}$. This provides an estimate whose loss is close to the optimal loss L^* if the class $\mathcal{C}$ is (i) sufficiently large so that the loss of the best function in $\mathcal{C}$ is close to L^* and (ii) is sufficiently small so that finding the best candidate in $\mathcal{C}$ based on the data is still possible. These two requirements are clearly in conflict. The trade-off is best understood by writing

$$\mathbb{E}L(g_n^*) - L^* = \left(\mathbb{E}L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)\right) + \left(\inf_{g\in\mathcal{C}} L(g) - L^*\right).$$

The first term is often called *estimation error*, while the second is the *approximation error*.

It is common to fix in advance a sequence of model classes $\mathcal{C}_1, \mathcal{C}_2, \ldots$, which, typically, become richer for larger indices. Given the data D_n, one wishes to select a good model from *one* of these classes. This is the problem of model selection.

Denote by $\widehat{g}_k$ a function in $\mathcal{C}_k$ having minimal empirical risk. One hopes to select a model class $\mathcal{C}_K$ such that the excess error $\mathbb{E}L(\widehat{g}_K) - L^*$ is close to

$$\min_k \mathbb{E}L(\widehat{g}_k) - L^* = \min_k \left[\left(\mathbb{E}L(\widehat{g}_k) - \inf_{g\in\mathcal{C}_k} L(g)\right) + \left(\inf_{g\in\mathcal{C}_k} L(g) - L^*\right)\right].$$

The idea of *structural risk minimization*, (also known as *complexity regularization*, is to add a complexity penalty to each of the $\widehat{L}_n(\widehat{g}_k)$'s to compensate for the overfitting effect. This penalty is usually closely related to a distribution-free upper bound for $\sup_{g\in\mathcal{C}_k} |\widehat{L}_n(g) - L(g)|$ so that the penalty eliminates the effect of overfitting.

The first general result shows that any approximate upper bound on error can be used to define a (possibly data-dependent) complexity penalty $C_n(k)$ and a model selection algorithm for which the excess error is close to

$$\min_k \left[\mathbb{E}C_n(k) + \left(\inf_{g\in\mathcal{C}_k} L(g) - L^*\right)\right].$$

Our goal is to select, among the classifiers $\widehat{g}_k$ one which has approximately minimal loss. The key assumption for our analysis is that the true loss of $\widehat{g}_k$ can be estimated for all k.

Assumption 1 *There are positive numbers c and m such that for each k an estimate $R_{n,k}$ on $L(\widehat{g}_k)$ is available which satisfies*

$$\mathbb{P}\left[L(\widehat{g}_k) > R_{n,k} + \epsilon\right] \le ce^{-2m\epsilon^2}$$

for all $\epsilon > 0$.

Now define the complexity penalty by

$$C_n(k) = R_{n,k} - \widehat{L}_n(\widehat{g}_k) + \sqrt{\frac{\log k}{m}}.$$

The last term is required because of technical reasons that will become apparent shortly. It is typically small. The difference $R_{n,k} - \widehat{L}_n(\widehat{g}_k)$ is simply an estimate of the 'right' amount of penalization $L(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k)$. Finally, define the prediction rule:

$$g_n^* = \arg\min_{k=1,2,\ldots} \tilde{L}_n(\widehat{g}_k),$$

where

$$\tilde{L}_n(\widehat{g}_k) = \widehat{L}_n(\widehat{g}_k) + C_n(k) = R_{n,k} + \sqrt{\frac{\log k}{m}}.$$

The following theorem summarizes the main performance bound for g_n^*.

Theorem 1.20. *Assume that the error estimates $R_{n,k}$ satisfy Assumption 1 for some positive constants c and m. Then*

$$\mathbb{E}L(g_n^*) - L^* \le \min_k \left[\mathbb{E}C_n(k) + \left(\inf_{g \in \mathcal{C}_k} L(g) - L^* \right) \right] + \sqrt{\frac{\log(ce)}{2m}}.$$

Theorem 1.20 shows that the prediction rule minimizing the penalized empirical loss achieves an almost optimal trade-off between the approximation error and the expected complexity, provided that the estimate $R_{n,k}$ on which the complexity is based is an approximate upper bound on the loss. In particular, if we knew in advance which of the classes $\mathcal{C}_k$ contained the optimal prediction rule, we could use the error estimates $R_{n,k}$ to obtain an upper bound on $\mathbb{E}L(\widehat{g}_k) - L^*$, and this upper bound would not improve on the bound of Theorem 1.20 by more than $O\left(\sqrt{\log k/m}\right)$.

PROOF. For brevity, introduce the notation

$$L_k^* = \inf_{g \in \mathcal{C}_k} L(g).$$

Then for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left[L(g_n^*) - \tilde{L}_n(g_n^*) > \epsilon\right] &\leq \mathbb{P}\left[\sup_{j=1,2,\ldots}\left(L(\widehat{g}_j) - \tilde{L}_n(\widehat{g}_j)\right) > \epsilon\right] \\
&\leq \sum_{j=1}^{\infty} \mathbb{P}\left[L(\widehat{g}_j) - \tilde{L}_n(\widehat{g}_j) > \epsilon\right] \\
&\quad \text{(by the union bound)} \\
&= \sum_{j=1}^{\infty} \mathbb{P}\left[L(\widehat{g}_j) - R_{n,j} > \epsilon + \sqrt{\frac{\log j}{m}}\right] \\
&\quad \text{(by definition)} \\
&\leq \sum_{j=1}^{\infty} ce^{-2m\left(\epsilon + \sqrt{\frac{\log j}{m}}\right)^2} \quad \text{(by Assumption 1)} \\
&\leq \sum_{j=1}^{\infty} ce^{-2m\left(\epsilon^2 + \frac{\log j}{m}\right)} \\
&< 2ce^{-2m\epsilon^2} \quad (\text{since } \textstyle\sum_{j=1}^{\infty} j^{-2} < 2).
\end{aligned}$$

To prove the theorem, for each k, we decompose $L(g_n^*) - L_k^*$ as

$$L(g_n^*) - L_k^* = \left(L(g_n^*) - \inf_j \tilde{L}_n(\widehat{g}_j)\right) + \left(\inf_j \tilde{L}_n(\widehat{g}_j) - L_k^*\right).$$

The first term may be bounded, by standard integration of the tail inequality shown above, as $\mathbb{E}\left[L(g_n^*) - \inf_j \tilde{L}_n(\widehat{g}_j)\right] \leq \sqrt{\log(ce)/(2m)}$. Choosing g_k^* such that $L(g_k^*) = L_k^*$, the second term may be bounded directly by

$$\begin{aligned}
\mathbb{E}\inf_j \tilde{L}_n(\widehat{g}_j) - L_k^* &\leq \mathbb{E}\tilde{L}_n(\widehat{g}_k) - L_k^* \\
&= \mathbb{E}\widehat{L}_n(\widehat{g}_k) - L_k^* + \mathbb{E}C_n(k) \\
&\quad \text{(by the definition of } \tilde{L}_n(\widehat{g}_k)) \\
&\leq \mathbb{E}\widehat{L}_n(g_k^*) - L(g_k^*) + \mathbb{E}C_n(k) \\
&\quad \text{(since } \widehat{g}_k \text{ minimizes the empirical loss on } \mathcal{C}_k) \\
&= \mathbb{E}C_n(k),
\end{aligned}$$

where the last step follows from the fact that $\mathbb{E}\widehat{L}_n(g_k^*) = L(g_k^*)$. Summing the obtained bounds for both terms yields that for each k,

$$\mathbb{E}L(g_n^*) \leq \mathbb{E}C_n(k) + L_k^* + \sqrt{\log(ce)/(2m)},$$

which implies the second statement of the theorem. □

1.6.2 Selection based on a test sample

In our first application of Theorem 1.20, we assume that m independent sample pairs

$$(X_1', Y_1'), \ldots, (X_m', Y_m')$$

are available. This may always be achieved by simply removing m samples from the training data. Of course, this is not very attractive, but m may be small relative to n. In this case we can estimate $L(\widehat{g}_k)$ by the hold-out error estimate

$$R_{n,k} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}_{\widehat{g}_k(X_i') \neq Y_i'}.$$

We apply Hoeffding's inequality to show that Assumption 1 is satisfied with $c = 1$, notice that $\mathbb{E}[R_{n,k}|D_n] = L(\widehat{g}_k)$, and apply Theorem 1.20 to give the following result.

COROLLARY 1.6. *Assume that the model selection algorithm is performed with the hold-out error estimate. Then*

$$\begin{aligned} &\mathbb{E}L(g_n^*) - L^* \\ &\quad \leq \min_k \left[\mathbb{E}\left[L(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k) \right] + \left(\inf_{g \in \mathcal{C}_k} L(g) - L^* \right) + \sqrt{\frac{\log k}{m}} \right] + \frac{1}{\sqrt{2m}}. \end{aligned}$$

In other words, the estimate achieves a nearly optimal balance between the approximation error, and the quantity

$$\mathbb{E}\left[L(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k) \right],$$

which may be regarded as the amount of overfitting.

1.6.3 Penalization by the VC dimension

In the remaining examples we consider error estimates $R_{n,k}$ which avoid splitting the data. First recall that by the Vapnik-Chervonenkis inequality, $2\sqrt{(V_{\mathcal{C}_k} \log(n+1) + \log 2)/n}$ is an upper bound for the expected maximal deviation, within class $\mathcal{C}_k$, between $L(g)$ and its empirical counterpart, $\widehat{L}_n(g)$. This suggests that penalizing the empirical error by this complexity term should compensate the overfitting within class $\mathcal{C}_k$. Thus, we introduce the error estimate

$$R_{n,k} = \widehat{L}_n(\widehat{g}_k) + 2\sqrt{\frac{V_{\mathcal{C}} \log(n+1) + \log 2}{n}}$$

Indeed, it is easy to show that this estimate satisfies Assumption 1. Indeed,

$$\begin{aligned}
&\mathbb{P}\left[L(\widehat{g}_k) > R_{n,k} + \epsilon\right] \\
&= \mathbb{P}\left[L(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k) > 2\sqrt{\frac{V_{\mathcal{C}}\log(n+1)+\log 2}{n}} + \epsilon\right] \\
&\leq \mathbb{P}\left[\sup_{g\in\mathcal{C}_k}\left|L(g) - \widehat{L}_n(g)\right| > 2\sqrt{\frac{V_{\mathcal{C}}\log(n+1)+\log 2}{n}} + \epsilon\right] \\
&\leq \mathbb{P}\left[\sup_{g\in\mathcal{C}_k}\left|L(g) - \widehat{L}_n(g)\right| > \mathbb{E}\sup_{g\in\mathcal{C}_k}\left|L(g) - \widehat{L}_n(g)\right| + \epsilon\right] \\
&\qquad \text{(by the Vapnik-Chervonenkis inequality)} \\
&\leq e^{-2n\epsilon^2} \quad \text{(by the bounded difference inequality).}
\end{aligned}$$

Therefore, satisfies Assumption 1 with $m = n$. Substituting this into Theorem 1.20 gives

$$\begin{aligned}
&\mathbb{E}L(g_n^*) - L^* \\
&\leq \min_k\left[2\sqrt{\frac{V_{\mathcal{C}_k}\log(n+1)+\log 2}{n}} + \left(\inf_{g\in\mathcal{C}_k} L(g) - L^*\right) + \sqrt{\frac{\log k}{n}}\right] + \sqrt{\frac{1}{2n}}.
\end{aligned}$$

Thus, structural risk minimization finds the best trade-off between the approximation error and a distribution-free upper bound on the estimation error.

1.6.4 Penalization by maximum discrepancy

In this section we propose a data-dependent way of computing the penalties with improved performance guarantees. Assume, for simplicity, that n is even, divide the data into two equal halves, and define, for each predictor f, the empirical loss on the two parts by

$$\widehat{L}_n^{(1)}(g) = \frac{2}{n}\sum_{i=1}^{n/2}\mathbb{I}_{g(X_i)\neq Y_i}$$

and

$$\widehat{L}_n^{(2)}(g) = \frac{2}{n}\sum_{i=n/2+1}^{n}\mathbb{I}_{g(X_i)\neq Y_i}.$$

Define the error estimate $R_{n,k}$ by

$$R_{n,k} = \widehat{L}_n(\widehat{g}_k) + \max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right).$$

Observe that the maximum discrepancy $\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g)-\widehat{L}_n^{(2)}(g)\right)$ may be computed using the following simple trick: first flip the labels of the first half of the data, thus obtaining the modified data set $D_n'=(X_1',Y_1'),\ldots,(X_n',Y_n')$ with $(X_i',Y_i')=(X_i,1-Y_i)$ for $i\le n/2$ and $(X_i',Y_i')=(X_i,Y_i)$ for $i>n/2$. Next find $f_k^-\in\mathcal{C}_k$ which minimizes the empirical loss based on D_n',

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{g(X_i')\neq Y_i'} &= \frac{1}{2}-\frac{1}{n}\sum_{i=1}^{n/2}\mathbb{I}_{g(X_i)\neq Y_i}+\frac{1}{n}\sum_{i=n/2+1}^{n}\mathbb{I}_{g(X_i)\neq Y_i}\\ &= \frac{1-\widehat{L}_n^{(1)}(g)+\widehat{L}_n^{(2)}(g)}{2}.\end{aligned}$$

Clearly, the function f_k^- maximizes the discrepancy. Therefore, the same algorithm that is used to compute the empirical loss minimizer $\widehat{g}_k$ may be used to find f_k^- and compute the penalty based on maximum discrepancy. This is appealing: although empirical loss minimization is often computationally difficult, the same approximate optimization algorithm can be used for both finding prediction rules and estimating appropriate penalties. In particular, if the algorithm only approximately minimizes empirical loss over the class $\mathcal{C}_k$ because it minimizes over some proper subset of $\mathcal{C}_k$, the theorem is still applicable.

Theorem 1.21. *If the penalties are defined using the maximum-discrepancy error estimates, and $m=n/21$, then*

$$\begin{aligned}\mathbb{E}L(g_n^*)-L^* \le \min_k\Bigg[&\mathbb{E}\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g)-\widehat{L}_n^{(2)}(g)\right)\\ &+\left(\inf_{g\in\mathcal{C}_k}L(g)-L^*\right)+4.59\sqrt{\frac{\log k}{n}}\Bigg]+\frac{4.70}{\sqrt{n}}.\end{aligned}$$

Proof. Once again, we check Assumption 1 and apply Theorem 1.20. Introduce the ghost sample $(X_1',Y_1'),\ldots,(X_n',Y_n')$, which is independent of the data and has the same distribution. Denote the empirical loss based on this sample by $L_n'(g)=\frac{1}{n}\sum_{i=1}^n\mathbb{I}_{g(X_i')\neq Y_i'}$. The

proof is based on the simple observation that for each k,

$$\begin{aligned}
\mathbb{E}\max_{g\in\mathcal{F}_k}\left(L'_n(g)-\hat{L}_n(g)\right) &= \frac{1}{n}\mathbb{E}\max_{g\in\mathcal{F}_k}\sum_{i=1}^{n}\left(\mathbb{I}_{g(X'_i)\neq Y'_i}-\mathbb{I}_{g(X_i)\neq Y_i}\right)\\
&\leq \frac{1}{n}\mathbb{E}\left(\max_{g\in\mathcal{F}_k}\sum_{i=1}^{n/2}\left(\mathbb{I}_{g(X'_i)\neq Y'_i}-\mathbb{I}_{g(X_i)\neq Y_i}\right)\right.\\
&\qquad\left.+\max_{g\in\mathcal{F}_k}\sum_{i=n/2+1}^{n}\left(\mathbb{I}_{g(X'_i)\neq Y'_i}-\mathbb{I}_{g(X_i)\neq Y_i}\right)\right)\\
&= \frac{2}{n}\mathbb{E}\max_{g\in\mathcal{F}_k}\sum_{i=1}^{n/2}\left(\mathbb{I}_{g(X'_i)\neq Y'_i}-\mathbb{I}_{g(X_i)\neq Y_i}\right)\\
&= \mathbb{E}\max_{g\in\mathcal{F}_k}\left(\hat{L}_n^{(1)}(g)-\hat{L}_n^{(2)}(g)\right). \qquad (1.1)
\end{aligned}$$

The bounded difference inequality inequality (Theorem 1.8) implies

$$\mathbb{P}\left[\max_{g\in\mathcal{C}_k}\left(L'_n(g)-\widehat{L}_n(g)\right)>\mathbb{E}\max_{g\in\mathcal{C}_k}\left(L'_n(g)-\widehat{L}_n(g)\right)+\epsilon\right]\leq e^{-n\epsilon^2}, \qquad (1.2)$$

$$\mathbb{P}\left[\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g)-\widehat{L}_n^{(2)}(g)\right)<\mathbb{E}\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g)-\widehat{L}_n^{(2)}(g)\right)-\epsilon\right]\leq e^{-n\epsilon^2/2} \qquad (1.3)$$

and so for each k,

$$
\begin{aligned}
&\mathbb{P}\left[L(\widehat{g}_k) > R_{n,k} + \epsilon\right] \\
&\quad = \mathbb{P}\left[L(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k) > \max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right) + \epsilon\right] \\
&\quad \le \mathbb{P}\left[L'_n(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k) > \max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right) + \frac{7\epsilon}{9}\right] \\
&\qquad + \mathbb{P}\left[L(\widehat{g}_k) - L'_n(\widehat{g}_k) > \frac{2\epsilon}{9}\right] \\
&\quad \le \mathbb{P}\left[L'_n(\widehat{g}_k) - \widehat{L}_n(\widehat{g}_k) > \max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right) + \frac{7\epsilon}{9}\right] \\
&\qquad + e^{-8n\epsilon^2/81} \quad \text{(by Hoeffding)} \\
&\quad \le \mathbb{P}\left[\max_{g\in\mathcal{C}_k}\left(L'_n(g) - \widehat{L}_n(g)\right) > \max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right) + \frac{7\epsilon}{9}\right] \\
&\qquad + e^{-8n\epsilon^2/81} \\
&\quad \le \mathbb{P}\left[\max_{g\in\mathcal{C}_k}\left(L'_n(g) - \widehat{L}_n(g)\right) > \mathbb{E}\max_{g\in\mathcal{C}_k}\left(L'_n(g) - \widehat{L}_n(g)\right) + \frac{\epsilon}{3}\right] \\
&\qquad + \mathbb{P}\left[\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right) < \mathbb{E}\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g) - \widehat{L}_n^{(2)}(g)\right) - \frac{4\epsilon}{9}\right] \\
&\qquad + e^{-8n\epsilon^2/81} \quad \text{(where we used (1.1))} \\
&\quad \le e^{-n\epsilon^2/9} + e^{-8n\epsilon^2/81} + e^{-8n\epsilon^2/81} \quad \text{(by (1.2) and (1.3))} \\
&\quad < 3e^{-8n\epsilon^2/81}.
\end{aligned}
$$

Thus, Assumption 1 is satisfied with $m = n/21$ and $c = 3$ and the proof is finished. □

Bibliography

General [1]

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, 1999.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

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

[4] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[5] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[6] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[7] 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.

Concentration for sums of independent random variables

[8] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[9] S.N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

[10] H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[1] The list of references given below contains, apart from the literature cited in the text, some of the key references in each covered topics. The list is far from being complete. Its purpose it to suggest some starting points for further reading.

[11] T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33:305–308, 1990.

[12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[13] R.M. Karp. *Probabilistic Analysis of Algorithms*. Class Notes, University of California, Berkeley, 1988.

[14] M. Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10:29–35, 1958.

Concentration

[15] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[16] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications in random combinatorics and learning. *Random Structures and Algorithms*, 16:277-292, 2000.

[17] L. Devroye. Exponential inequalities in nonparametric estimation. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 31–44. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

[18] J. H. Kim. The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$. *Random Structures and Algorithms*, 7:173–207, 1995.

[19] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, **1**, 63–87, `http://www.emath.fr/ps/`, (1996).

[20] K. Marton. A simple proof of the blowing-up lemma. *IEEE Transactions on Information Theory*, 32:445–446, 1986.

[21] K. Marton. Bounding $\bar{d}$-distance by informational divergence: a way to prove measure concentration. *Annals of Probability*, to appear:0–0, 1996.

[22] K. Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6:556–571, 1996. Erratum: 7:609–613, 1997.

[23] P. Massart. About the constant in Talagrand's concentration inequalities from empirical processes. *Annals of Probability*, 28:863–884, 2000.

[24] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[25] W. Rhee and M. Talagrand. Martingales, inequalities, and NP-complete problems. *Mathematics of Operations Research*, 12:177–181, 1987.

[26] J.M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[27] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. I.H.E.S. Publications Mathématiques, 81:73–205, 1996.

[28] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.* 126:505-563, 1996.

[29] M. Talagrand. A new look at independence. *Annals of Probability*, 24:0–0, 1996. special invited paper.

VC theory

[30] K. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*, 4:1041–1067, 1984.

[31] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

[32] P. Bartlett and G. Lugosi. An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, 44:55–62, 1999.

[33] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[34] Devroye, L. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

[35] 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.

[36] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

[37] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.

[38] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[39] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers, *Annals of Statistics*, 30, 2002.

[40] M. Ledoux and M. Talagrand. *Probability in Banach Space*, Springer-Verlag, New York, 1991.

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

[42] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[43] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods, *Annals of Statistics*, 26:1651–1686, 1998.

[44] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[45] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

[46] S. Van de Geer. Estimating a regression function. *Annals of Statistics*, 18:907–924, 1990.

[47] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[48] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[49] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[50] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[51] 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.

[52] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*, Springer-Verlag, New York, 1996.

Shatter coefficients, VC dimension

[53] P. Assouad, Sur les classes de Vapnik-Chervonenkis, *C.R. Acad. Sci. Paris*, vol. 292, Sér.I, pp. 921–924, 1981.

[54] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Transactions on Electronic Computers*, vol. 14, pp. 326–334, 1965.

[55] R. M. Dudley, Central limit theorems for empirical measures, *Annals of Probability*, vol. 6, pp. 899–929, 1978.

[56] R. M. Dudley, Balls in R^k do not cut all subsets of $k+2$ points, *Advances in Mathematics*, vol. 31 (3), pp. 306–308, 1979.

[57] P. Frankl, On the trace of finite sets, *Journal of Combinatorial Theory, Series A*, vol. 34, pp. 41–45, 1983.

[58] D. Haussler, Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension, *Journal of Combinatorial Theory, Series A*, vol. 69, pp. 217–232, 1995.

[59] N. Sauer, On the density of families of sets, *Journal of Combinatorial Theory Series A*, vol. 13, pp. 145–147, 1972.

[60] L. Schläffli, *Gesammelte Mathematische Abhandlungen*, Birkhäuser-Verlag, Basel, 1950.

[61] S. Shelah, A combinatorial problem: stability and order for models and theories in infinity languages, *Pacific Journal of Mathematics*, vol. 41, pp. 247–261, 1972.

[62] J. M. Steele, Combinatorial entropy and uniform limit laws, Ph.D. dissertation, Stanford University, Stanford, CA, 1975.

[63] J. M. Steele, Existence of submatrices with all possible columns, *Journal of Combinatorial Theory, Series A*, vol. 28, pp. 84–88, 1978.

[64] R. S. Wenocur and R. M. Dudley, Some special Vapnik-Chervonenkis classes, *Discrete Mathematics*, vol. 33, pp. 313–318, 1981.

Lower bounds.

[65] A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, vol.30, 31–56, 1998.

[66] P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus de l'Académie des Sciences de Paris*, 296:1021–1024,1983.

[67] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65:181–237, 1983.

[68] L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71:271–291,1986.

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

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

[71] 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.

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

[73] E. Mammen, A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27:1808–1829, 1999.

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

[75] 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.

Complexity regularization

[76] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[77] A.R. Barron. Logically smooth density estimation. Technical Report TR 56, Department of Statistics, Stanford University, 1985.

[78] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

[79] A.R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related fields*, 113:301–413, 1999.

[80] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[81] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, March 1998.

[82] P. Bartlett, S. Boucheron, and G. Lugosi, Model selection and error estimation. *Proceedings of the 13th Annual Conference on Computational Learning Theory*, ACM Press, pp.286–297, 2000.

[83] L. Birgé and P. Massart. From model selection to adaptive estimation. In E. Torgersen D. Pollard and G. Yang, editors, *Festschrift for Lucien Le Cam: Research papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

[84] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[85] Y. Freund. Self bounding learning algorithms. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 247–258, 1998.

[86] A.R. Gallant. *Nonlinear Statistical Models*. John Wiley, New York, 1987.

[87] S. Geman and C.R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:401–414, 1982.

[88] M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth Annual ACM Workshop on Computational Learning Theory*, pages 21–30. Association for Computing Machinery, New York, 1995.

[89] A. Krzyzak and T. Linder. Radial basis function networks and complexity regularization in function learning. *IEEE Transactions on Neural Networks*, 9:247–256, 1998.

[90] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *Annals of Statistics*, vol. 27, no.6, 1999.

[91] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41:677–678, 1995.

[92] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[93] C.L. Mallows. Some comments on c_p. *IEEE Technometrics*, 15:661–675, 1997.

[94] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la faculté des sciences de l'université de Toulouse, Mathématiques*, série 6, **IX**:245–303, 2000.

[95] R. Meir. Performance bounds for nonlinear time series prediction. In *Proceedings of the Tenth Annual ACM Workshop on Computational Learning Theory*, page 122–129. Association for Computing Machinery, New York, 1997.

[96] D.S. Modha and E. Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42:2133–2145, 1996.

[97] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[98] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[99] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[100] X. Shen and W.H. Wong. Convergence rate of sieve estimates. *Annals of Statistics*, 22:580–615, 1994.

[101] Y. Yang and A.R. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, to appear, 1997.

[102] Y. Yang and A.R. Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44:to appear, 1998.