Lectures on statistical learning theory

Presented at
Garchy Seminar on Mathematical Statistics and Applications
August 27–September 1, 2000

Gábor Lugosi
Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain,
lugosi@upf.es
http://www.econ.upf.es/~lugosi

Contents

1. A binary classification problem

Pattern recognition (or *classification* or *discrimination*) is about guessing or predicting the unknown class of an observation, determined by a discrete quantity. An observation is a collection of numerical measurements, represented by a d-dimensional vector x, called the *observation*. The unknown nature of the observation is called a *class*. It is denoted by y and takes values in the set $\{0, 1\}$. (For simplicity, we restrict our attention to binary classification.) In pattern recognition, one creates a function $g(x) : \mathbb{R}^d \to \{0, 1\}$ which represents one's guess of y given x. The mapping g is called a *classifier*. Our classifier errs on x if $g(x) \neq y$.

To model the learning problem, we introduce a probabilistic setting, and let (X, Y) be an $\mathbb{R}^d \times \{0, 1\}$-valued random pair.

The random pair (X, Y) may be described in a variety of ways: for example, it is defined by the pair (μ, η), where μ is the probability measure for X and η is the regression of Y on X. More precisely, for a Borel-measurable set $A \subseteq \mathbb{R}^d$,

$$\mu(A) = \mathbb{P}\{X \in A\},$$

and for any $x \in \mathbb{R}^d$,

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\} = \mathbb{E}\{Y | X = x\}.$$

Thus, $\eta(x)$ is the conditional probability that Y is 1 given $X = x$. The distribution of (X, Y) is determined by (μ, η). The function η is called the *a posteriori probability*.

Any function $g : \mathbb{R}^d \to \{0, 1\}$ defines a *classifier*. An *error* occurs if $g(X) \neq Y$, and the *probability of error* for a classifier g is

$$L(g) = \mathbb{P}\{g(X) \neq Y\} .$$

The Bayes classifier given by

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

minimizes the probability of error:

Theorem 1 *For any classifier* $g : \mathbb{R}^d \to \{0, 1\}$,

$$\mathbb{P}\{g^*(X) \neq Y\} \leq \mathbb{P}\{g(X) \neq Y\}.$$

Proof. Given $X = x$, the conditional probability of error of any decision g may be expressed as

$$\begin{aligned}
&\mathbb{P}\{g(X) \neq Y | X = x\} \\
&= 1 - \mathbb{P}\{Y = g(X) | X = x\} \\
&= 1 - (\mathbb{P}\{Y = 1, g(X) = 1 | X = x\} + \mathbb{P}\{Y = 0, g(X) = 0 | X = x\}) \\
&= 1 - \left(\mathbb{I}_{\{g(x)=1\}} \mathbb{P}\{Y = 1 | X = x\} + \mathbb{I}_{\{g(x)=0\}} \mathbb{P}\{Y = 0 | X = x\}\right) \\
&= 1 - \left(\mathbb{I}_{\{g(x)=1\}} \eta(x) + \mathbb{I}_{\{g(x)=0\}} (1 - \eta(x))\right),
\end{aligned}$$

where $\mathbb{I}_A$ denotes the indicator of the set A. Thus, for every $x \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{P}\{g(X) \neq Y | X = x\} - \mathbb{P}\{g^*(X) \neq Y | X = x\} \\
&= \eta(x) \left(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}\right) + (1 - \eta(x)) \left(\mathbb{I}_{\{g^*(x)=0\}} - \mathbb{I}_{\{g(x)=0\}}\right) \\
&= (2\eta(x) - 1) \left(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}\right) \\
&\geq 0
\end{aligned}$$

by the definition of g^*. The statement now follows by integrating both sides with respect to $\mu(dx)$. ∎

L^* is called the Bayes probability of error, Bayes error, or Bayes risk. The proof given above reveals that

$$L(g) = 1 - \mathbb{E}\left\{\mathbb{I}_{\{g(X)=1\}} \eta(X) + \mathbb{I}_{\{g(X)=0\}} (1 - \eta(X))\right\},$$

and in particular,

$$L^* = 1 - \mathbb{E}\left\{\mathbb{I}_{\{\eta(X)>1/2\}} \eta(X) + \mathbb{I}_{\{\eta(X)\leq 1/2\}} (1 - \eta(X))\right\}.$$

Note that g^* depends upon the distribution of (X, Y). If this distribution is known, g^* may be computed. Mostly, the distribution of (X, Y) is unknown, so that g^* is unknown too.

In our model, we have access to a good database of pairs (X_i, Y_i), $1 \leq i \leq n$, observed in the past. We assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$, the *data*, is a sequence of independent identically distributed (*i.i.d.*) random pairs with the same distribution as that of (X, Y).

A classifier is constructed on the basis of $X_1, Y_1, \ldots, X_n, Y_n$ and is denoted by g_n; Y is guessed by $g_n(X; X_1, Y_1, \ldots, X_n, Y_n)$. The process of constructing g_n is called *learning*, *supervised learning*, or *learning with a teacher*. The performance of g_n is measured by the conditional *probability of error*

$$L_n = L(g_n) = \mathbb{P}\{g_n(X; X_1, Y_1, \ldots, X_n, Y_n) \neq Y | X_1, Y_1, \ldots, X_n, Y_n)\} .$$

This is a random variable because it depends upon the data. So, L_n averages over the distribution of (X, Y), but the data is held fixed. Averaging over the data as well would be unnatural, because in a given application, one has to live with the data at hand. It would be marginally useful to know the number $\mathbb{E}L_n = \mathbb{P}\{g_n(X) \neq Y\}$ as this number would indicate the quality of an average data sequence.

2. Empirical risk minimization

Assume that a class $\mathcal{C}$ of classifiers $g : \mathbb{R}^d \to \{0,1\}$ is given and our task is to find one with a small probability of error. In the lack of the knowledge of the underlying distribution, one has to resort to using the data to estimate the probabilities of error for the classifiers in $\mathcal{C}$. It is tempting to pick a classifier from $\mathcal{C}$ that minimizes an estimate of the probability of error over the class. The most natural choice to estimate the probability of error $L(g) = \mathbb{P}\{g(X) \neq Y\}$ is the error count

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_{\{g(X_j) \neq Y_j\}}.$$

$\widehat{L}_l(g)$ is called the *empirical error* of the classifier g.

A good method should pick a classifier with a probability of error that is close to the minimal probability of error in the class. Intuitively, if we can estimate the error probability for the classifiers in $\mathcal{C}$ *uniformly* well, then the classification function that minimizes the estimated probability of error is likely to have a probability of error that is close to the best in the class.

Denote by g_n^* the classifier that minimizes the estimated probability of error over the class:

$$\widehat{L}_n(g_n^*) \leq \widehat{L}_n(g) \qquad \text{for all } g \in \mathcal{C}.$$

Then for the probability of error

$$L(g_n^*) = \mathbb{P}\{\, g_n^*(X) \neq Y | D_n\}$$

of the selected rule we have:

Lemma 1

$$L(g_n^*) - \inf_{g \in \mathcal{C}} L(g) \leq 2 \sup_{g \in \mathcal{C}} |\widehat{L}_n(g) - L(g)|,$$

$$|\widehat{L}_n(g_n^*) - L(g_n^*)| \leq \sup_{g \in \mathcal{C}} |\widehat{L}_n(g) - L(g)|.$$

PROOF.

$$\begin{aligned} L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) &= L(g_n^*) - \widehat{L}_n(g_n^*) + \widehat{L}_n(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \\ &\leq L(g_n^*) - \widehat{L}_n(g_n^*) + \sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| \\ &\leq 2\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|. \end{aligned}$$

The second inequality is trivially true. ∎

We see that upper bounds for $\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$ provide us with upper bounds for two things simultaneously:

(1) An upper bound for the suboptimality of g_n^* within $\mathcal{C}$, that is, a bound for $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)$.

(2) An upper bound for the error $|\widehat{L}_n(g_n^*) - L(g_n^*)|$ committed when $\widehat{L}_n(g_n^*)$ is used to estimate the probability of error $L(g_n^*)$ of the selected rule.

It is particularly useful to know that even though $\widehat{L}_n(g_n^*)$ is usually optimistically biased, it is within given bounds of the unknown probability of error with g_n^*, and that no other test sample is needed to estimate this probability of error. Whenever our bounds indicate that we are close to the optimum in $\mathcal{C}$, we must at the same time have a good estimate of the probability of error, and vice versa.

Clearly, the random variable $n\widehat{L}_n(g)$ is binomially distributed with parameters n and $L(g)$. Thus, to obtain bounds for the success of empirical error minimization, we need to study uniform deviations of binomial random variables from their means. In the next two sections we summarize the basics of the underlying theory.

3. Concentration inequalities

3.1 Hoeffding's inequality

The simplest inequality to bound the difference between a random variable and its expected value is *Markov's inequality*: for any nonnegative random variable X, and $t > 0$,

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}.$$

From this, we deduce *Chebyshev's inequality*: if X is an arbitrary random variable and $t > 0$, then

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathbb{P}\left\{|X - \mathbb{E}X|^2 \geq t^2\right\} \leq \frac{\mathbb{E}\left\{|X - \mathbb{E}X|^2\right\}}{t^2} = \frac{\mathbb{V}\{X\}}{t^2}.$$

As an example, we derive inequalities for $\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\}$ with $S_n = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent real-valued random variables. Chebyshev's inequality and independence immediately gives us

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq t\} \leq \frac{\mathbb{V}\{S_n\}}{t^2} = \frac{\sum_{i=1}^n \mathbb{V}\{X_i\}}{t^2}.$$

The meaning of this is perhaps better seen if we assume that the X_i's are i.i.d. Bernoulli(p) random variables (i.e., $\mathbb{P}\{X_i = 1\} = 1 - \mathbb{P}\{X_i = 0\} = p$), and normalize:

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i - p\right| \geq \epsilon\right\} \leq \frac{p(1-p)}{n\epsilon^2}.$$

To illustrate the weakness of this bound, let $\Phi(y) = \int_{-\infty}^{y} e^{-t^2/2}/\sqrt{2\pi}\,dt$ be the normal distribution function. The central limit theorem states that

$$\mathbb{P}\left\{\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{i=1}^n X_i - p\right) \geq y\right\} \to 1 - \Phi(y) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-y^2/2}}{y},$$

from which we would expect something like

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i - p \geq \epsilon\right\} \approx e^{-n\epsilon^2/(2p(1-p))}.$$

Clearly, Chebyshev's inequality is off mark. An improvement may be obtained by *Chernoff's bounding method.* By Markov's inequality, if s is an arbitrary positive number, then for any random variable X, and any $t > 0$,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}e^{sX}}{e^{st}}.$$

In Chernoff's method, we find an $s > 0$ that minimizes the upper bound or makes the upper bound small. In the case of a sum of independent random variables,

$$\begin{aligned}
\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st}\mathbb{E}\left\{\exp\left(s\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right)\right\} \\
&= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left\{e^{s(X_i - \mathbb{E}X_i)}\right\} \quad \text{(by independence).}
\end{aligned}$$

Now the problem of finding tight bounds comes down to finding a good upper bound for the moment generating function of the random variables $X_i - \mathbb{E}X_i$. There are many ways of doing this. For bounded random variables perhaps the most elegant version is due to Hoeffding:

Lemma 2 *Let* X *be a random variable with* $\mathbb{E}X = 0$, $a \leq X \leq b$. *Then for* $s > 0$,

$$\mathbb{E}\left\{e^{sX}\right\} \leq e^{s^2(b-a)^2/8}.$$

Proof. Note that by convexity of the exponential function

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{for } a \leq x \leq b.$$

Exploiting $\mathbb{E}X = 0$, and introducing the notation $p = -a/(b-a)$ we get

$$\begin{aligned}
\mathbb{E}e^{sX} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\
&= \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)} \\
&\stackrel{\text{def}}{=} e^{\phi(u)},
\end{aligned}$$

where $u = s(b-a)$, and $\phi(u) = -pu + \log(1-p+pe^u)$. But by straightforward calculation it is easy to see that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p+(1-p)e^{-u}},$$

therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq \frac{1}{4}.$$

Thus, by Taylor series expansion with remainder, for some $\theta \in [0, u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

Now we may directly plug this lemma into the bound obtained by Chernoff's method:

$$\begin{aligned}
&\mathbb{P}\{S_n - \mathbb{E}S_n \geq \epsilon\} \\
&\leq \; e^{-s\epsilon} \prod_{i=1}^{n} \mathbb{E}\left\{e^{s(X_i - \mathbb{E}X_i)}\right\} \\
&\leq \; e^{-s\epsilon} \prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} \quad \text{(by Lemma 2)} \\
&= \; e^{-s\epsilon} e^{s^2 \sum_{i=1}^{n}(b_i-a_i)^2/8} \\
&= \; e^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2} \quad \text{(by choosing } s = 4\epsilon/\textstyle\sum_{i=1}^{n}(b_i-a_i)^2\text{)}.
\end{aligned}$$

The result we have just derived is generally known as *Hoeffding's inequality.* For binomial random variables it was proved by Chernoff (1952) and Okamoto (1952). Summarizing, we have:

Theorem 2 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Denote their sum by $S_n = \sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$ we have*

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq \epsilon\} \leq e^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}$$

and

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -\epsilon\} \leq e^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}.$$

If we specialize this to the binomial distribution, that is, when the X_i's are i.i.d. Bernoulli(p), we get

$$\mathbb{P}\{S_n/n - p \geq \epsilon\} \leq e^{-2n\epsilon^2},$$

which is just the kind of inequality we hoped for.

We may combine this inequality with that of Lemma 1 to bound the performance of empirical risk minimization in the special case when the class $\mathcal{C}$ contains finitely many classifiers:

Theorem 3 *Assume that the cardinality of $\mathcal{C}$ is bounded by N. Then we have for all $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| > \epsilon\right\} \leq 2Ne^{-2n\epsilon^2}.$$

An important feature of the result above is that it is completely distribution free. The actual distribution of the data does not play a role at all in the upper bound.

To have an idea about the size of the error, one may be interested in the expected maximal deviation

$$\mathbb{E}\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|.$$

The inequality above may be used to derive such an upper bound by observing that for any nonnegative random variable X,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\{X \geq t\}dt.$$

Sharper bounds result by combining Lemma 2 with the following simple result:

Lemma 3 *Let $\sigma > 0$, $n \geq 2$, and let $Y_1, \ldots, Y_n$ be real-valued random variables such that for all $s > 0$ and $1 \leq i \leq n$, $\mathbb{E}\left\{e^{sY_i}\right\} \leq e^{s^2\sigma^2/2}$. Then*

$$\mathbb{E}\left\{\max_{i\leq n} Y_i\right\} \leq \sigma\sqrt{2\ln n}\ .$$

If, in addition, $\mathbb{E}\left\{e^{s(-Y_i)}\right\} \le e^{s^2\sigma^2/2}$ *for every* $s > 0$ *and* $1 \le i \le n$*, then for any* $n \ge 1$*,*

$$\mathbb{E}\left\{\max_{i\le n}|Y_i|\right\} \le \sigma\sqrt{2\ln(2n)}\ .$$

Proof. By Jensen's inequality, for all $s > 0$,

$$e^{s\mathbb{E}\{\max_{i\le n}Y_i\}} \le \mathbb{E}\left\{e^{s\max_{i\le n}Y_i}\right\} = \mathbb{E}\left\{\max_{i\le n}e^{sY_i}\right\} \le \sum_{i=1}^{n}\mathbb{E}\left\{e^{sY_i}\right\} \le ne^{s^2\sigma^2/2}\ .$$

Thus,

$$\mathbb{E}\left\{\max_{i\le n}Y_i\right\} \le \frac{\ln n}{s} + \frac{s\sigma^2}{2}\ ,$$

and taking $s = \sqrt{2\ln n/\sigma^2}$ yields the first inequality. Finally, note that $\max_{i\le n}|Y_i| = \max(Y_1, -Y_1, \ldots, Y_n, -Y_n)$ and apply the first inequality to prove the second. ▮

Now we obtain

$$\mathbb{E}\sup_{g\in\mathcal{C}}|\widehat{L}_n(g) - L(g)| \le \sqrt{\frac{\ln(2N)}{2n}}.$$

3.2 Other inequalities for sums

Here we summarize some other useful inequalities for the deviations of sums of independent random variables from their means.

Theorem 4 BENNETT'S INEQUALITY. *Let* $X_1, \ldots, X_n$ *be independent real-valued random variables with zero mean, and assume that* $|X_i| \le c$ *with probability one. Let* $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}\{X_i\}$*. Then that for any* $t > 0$*,*

$$\mathbb{P}\{S_n > t\} \le \exp\left(-\frac{n\sigma^2}{c^2}h\left(\frac{ct}{n\sigma^2}\right)\right),$$

where the function h *is defined by* $h(u) = (1+u)\log(1+u) - u$ *for* $u \ge 0$.

SKETCH OF PROOF. We use Chernoff's method as in the proof of Hoeffding's inequality. Write

$$\mathbb{E}\left\{e^{sX_i}\right\} = 1 + s\mathbb{E}\{X_i\} + \sum_{r=2}^{\infty} \frac{s^r \mathbb{E}\{X_i^r\}}{r!} = 1 + s^2 \mathbb{V}\{X_i\} F_i \le e^{s^2 \mathbb{V}\{X_i\} F_i}$$

with $F_i = \sum_{r=2}^{\infty} s^{r-2} \mathbb{E}\{X_i^r\} / (r! \mathbb{V}\{X_i\})$. We may use the boundedness of the X_i's to show that $\mathbb{E}\{X_i^r\} \le c^{r-2} \mathbb{V}\{X_i\}$, which implies $F_i \le (e^{sc} - 1 - sc)/(sc)^2$. Choose the s which minimizes the obtained upper bound for the tail probability. ∎

Theorem 5 BERNSTEIN'S INEQUALITY. *Under the conditions of the previous exercise, for any $t > 0$,*

$$\mathbb{P}\{S_n > t\} \le \exp\left(-\frac{t^2}{2n\sigma^2 + 2ct/3}\right).$$

PROOF. The result follows from Bennett's inequality and the inequality $h(u) \ge u^2/(2 + 2u/3)$, $u \ge 0$. ∎

Theorem 6 *Let $X_1, \dots, X_n$ be independent random variables, taking their values from $[0,1]$. If $m = \mathbb{E}S_n$, then for any $m \le t \le n$,*

$$\mathbb{P}\{S_n \ge t\} \le \left(\frac{m}{t}\right)^t \left(\frac{n-m}{n-t}\right)^{n-t}.$$

Also,

$$\mathbb{P}\{S_n \ge t\} \le \left(\frac{m}{t}\right)^t e^{t-m},$$

and for all $\epsilon > 0$,

$$\mathbb{P}\{S_n \ge m(1+\epsilon)\} \le e^{-mh(\epsilon)},$$

where h is the function defined in the previous theorem. Finally,

$$\mathbb{P}\{S_n \le m(1-\epsilon)\} \le e^{-m\epsilon^2/2}.$$

3.3 The bounded difference inequality

In this section we give some powerful extensions of concentration inequalities for sums to to general functions of independent random variables.

Let A be some set, and let $g : A^n \to \mathbb{R}$ be some measurable function of n variables. We derive inequalities for the difference between $g(X_1, \ldots, X_n)$ and its expected value when $X_1, \ldots, X_n$ are arbitrary independent random variables taking values in A. Sometimes we will write g instead of $g(X_1, \ldots, X_n)$ whenever it does not cause any confusion.

We recall the elementary fact that if X and Y are arbitrary bounded random variables, then $\mathbb{E}\{XY\} = \mathbb{E}\{\mathbb{E}\{XY|Y\}\} = \mathbb{E}\{Y\mathbb{E}\{X|Y\}\}$.

Theorem 7 EFRON-STEIN INEQUALITY. *If $X_1', \ldots, X_n'$ form an independent copy of $X_1, \ldots, X_n$, then*

$$\mathbb{V}(g(X_1, \ldots, X_n)) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left\{ (g(X_1, \ldots, X_n) - g(X_1, \ldots, X_i', \ldots, X_n))^2 \right\}$$

PROOF. Introduce the notation $V = g - \mathbb{E}g$, and define

$$V_i = \mathbb{E}\{g|X_1, \ldots, X_i\} - \mathbb{E}\{g|X_1, \ldots, X_{i-1}\}, \qquad i = 1, \ldots, n.$$

Clearly, $V = \sum_{i=1}^{n} V_i$. Then

$$\begin{aligned} \mathbb{V}(g) &= \mathbb{E}\left\{ \left(\sum_{i=1}^{n} V_i \right)^2 \right\} \\ &= \mathbb{E} \sum_{i=1}^{n} V_i^2 + 2\mathbb{E} \sum_{i>j} V_i V_j \\ &= \mathbb{E} \sum_{i=1}^{n} V_i^2 , \end{aligned}$$

since, for any $i > j$,

$$\mathbb{E}V_iV_j = \mathbb{E}\mathbb{E}\{V_iV_j|X_1, \ldots, X_j\} = \mathbb{E}\{V_j\mathbb{E}\{V_i|X_1, \ldots, X_j\}\} = 0 .$$

To bound $\mathbb{E}V_i^2$, note that, by Jensen's inequality,

$$\begin{aligned}
V_i^2 &= (\mathbb{E}\{g|X_1,\dots,X_i\} - \mathbb{E}\{g|X_1,\dots,X_{i-1}\})^2 \\
&= \left(\mathbb{E}\left[\mathbb{E}\{g|X_1,\dots,X_n\} - \mathbb{E}\{g|X_1,\dots,X_{i-1},X_{i+1},\dots,X_n\}\middle|X_1,\dots,X_i\right]\right)^2 \\
&\le \mathbb{E}\left[(\mathbb{E}\{g|X_1,\dots,X_n\} - \mathbb{E}\{g|X_1,\dots,X_{i-1},X_{i+1},\dots,X_n\})^2\middle|X_1,\dots,X_i\right] ,
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathbb{E}V_i^2 &\le \mathbb{E}\left[(g - \mathbb{E}\{g|X_1,\dots,X_{i-1},X_{i+1},\dots,X_n\})^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[(g(X_1,\dots,X_n) - g(X_1,\dots,X_i',\dots,X_n\})^2\right] ,
\end{aligned}$$

where at the last step we used (conditionally) the elementary fact that if X and Y are independent and identically distributed random variables, then $\mathbb{V}(X) = (1/2)\mathbb{E}\{(X-Y)^2\}$. ∎

Assume that a function $g : A^n \to \mathbb{R}$ satisfies the *bounded difference assumption*

$$\sup_{\substack{x_1,\dots,x_n,\\ x_i'\in A}} |g(x_1,\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)| \le c_i \ , \ \ 1 \le i \le n \ .$$

In other words, we assume that if we change the i-th variable of g while keeping all the others fixed, then the value of the function does not change by more than c_i. Then the Efron-Stein inequality implies that

$$\mathbb{V}(g) \le \frac{1}{2}\sum_{i=1}^n c_i^2 \ .$$

For such functions is is possible to prove the following exponential tail inequality, a powerful extension of Hoeffding's inequality.

Theorem 8 THE BOUNDED DIFFERENCE INEQUALITY. *Under the bounded difference assumption above, for all* $t > 0$*,*

$$\mathbb{P}\{g(X_1,\dots,X_n) - \mathbb{E}g(X_1,\dots,X_n) \ge t\} \le e^{-2t^2 \big/ \sum_{i=1}^n c_i^2} \ ,$$

and

$$\mathbb{P}\{\mathbb{E}g(X_1,\dots,X_n) - g(X_1,\dots,X_n) \ge t\} \le e^{-2t^2 \big/ \sum_{i=1}^n c_i^2} \ .$$

McDiarmid (1989) proved this inequality using martingale techniques, which we reproduce here. The proof of Theorem 8 uses the following straightforward extension of Lemma 2:

Lemma 4 *Let* V *and* Z *be random variables such that* $\mathbb{E}\{V|Z\} = 0$ *with probability one, and for some function* h *and constant* $c \geq 0$

$$h(Z) \leq V \leq h(Z) + c.$$

Then for all $s > 0$

$$\mathbb{E}\left\{e^{sV}|Z\right\} \leq e^{s^2c^2/8}.$$

Proof of Theorem 8. Just like in the proof of Theorem 7, introduce the notation $V = g - \mathbb{E}g$, and define

$$V_i = \mathbb{E}\{g|X_1, \ldots, X_i\} - \mathbb{E}\{g|X_1, \ldots, X_{i-1}\}, \qquad i = 1, \ldots, n.$$

Then $V = \sum_{i=1}^{n} V_i$. Also introduce the random variables

$$H_i(X_1, \ldots, X_i) = \mathbb{E}\{g(X_1, \ldots, X_n)|X_1, \ldots, X_i\}.$$

Then, denoting the distribution of X_i by F_i for $i = 1, \ldots, n$,

$$V_i = H_i(X_1, \ldots, X_i) - \int H_i(X_1, \ldots, X_{i-1}, x)F_i(dx).$$

Define the random variables

$$W_i = \sup_u \left(H_i(X_1, \ldots, X_{i-1}, u) - \int H_i(X_1, \ldots, X_{i-1}, x)F_i(dx) \right),$$

and

$$Z_i = \inf_v \left(H_i(X_1, \ldots, X_{i-1}, v) - \int H_i(X_1, \ldots, X_{i-1}, x)F_i(dx) \right).$$

Clearly, $Z_i \leq V_i \leq W_i$ with probability one, and also

$$W_i - Z_i = \sup_u \sup_v \left(H_i(X_1, \ldots, X_{i-1}, u) - H_i(X_1, \ldots, X_{i-1}, v)\right) \leq c_i ,$$

by the bounded difference assumption. Therefore, we may apply the lemma above to obtain, for all $i = 1, \ldots, n$,

$$\mathbb{E}\left\{e^{sV_i}|X_1, \ldots, X_{i-1}\right\} \leq e^{s^2 c_i^2/8}.$$

Finally, by Chernoff's bound, for any $s > 0$,

$$\begin{aligned}
&\mathbb{P}\{g - \mathbb{E}g \geq t\} \\
&\leq \frac{\mathbb{E}\left\{e^{s\sum_{i=1}^n V_i}\right\}}{e^{st}} = \frac{\mathbb{E}\left\{e^{s\sum_{i=1}^{n-1} V_i}\mathbb{E}\left\{e^{sV_n}|X_1, \ldots, X_{n-1}\right\}\right\}}{e^{st}} \\
&\leq e^{s^2 c_n^2/8}\frac{\mathbb{E}\left\{e^{s\sum_{i=1}^{n-1} V_i}\right\}}{e^{st}} \\
&\leq e^{-st}e^{s^2\sum_{i=1}^n c_i^2/8} \quad \text{(by repeating the same argument } n \text{ times)}.
\end{aligned}$$

Choosing $s = 4t \Big/ \sum_{i=1}^n c_i^2$ proves the first inequality. The proof of the second inequality is similar. ∎

An important application of the bounded difference inequality shows that if $\mathcal{C}$ is any class of classifiers of form $g : \mathbb{R}^d \to \{0, 1\}$, then

$$\mathbb{P}\left\{\left|\sup_{g\in\mathcal{C}}|\widehat{L}_n(g) - L(g)| - \mathbb{E}\sup_{g\in\mathcal{C}}|\widehat{L}_n(g) - L(g)|\right| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

Indeed, if we view $\sup_{g\in\mathcal{C}}|\widehat{L}_n(g) - L(g)|$ as a function of the n independent random pairs (X_i, Y_i), $i = 1, \ldots, n$, then we immediately see that the bounded difference assumption is satisfied with $c_i = 1/n$, and Theorem 8 immediately implies the statement.

The interesting fact is that regardless of the size of its expected value, the random variable $\sup_{g\in\mathcal{C}}|\widehat{L}_n(g) - L(g)|$ is sharply concentrated around its mean with very large probability. In the next section we study the expected value.

4. Vapnik-Chervonenkis theory

4.1 The Vapnik-Chervonenkis inequality

Recall from Section 3.1 that for any finite class $\mathcal{C}$ of classifiers, and for all $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| > \epsilon\right\} \le 2Ne^{-2n\epsilon^2},$$

and

$$\mathbb{E}\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| \le \sqrt{\frac{\ln(2N)}{2n}}.$$

These simple bounds may be useless if the cardinality N of the class is very large, or infinite. The purpose of this section is to introduce a theory to handle such cases.

Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}^d$ with common distribution

$$\mu(A) = \mathbb{P}\{X_1 \in A\} \quad (A \subset \mathbb{R}^d).$$

Define the empirical distribution

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[X_i\in A]} \quad (A \subset \mathbb{R}^d).$$

Consider a class $\mathcal{A}$ of subsets of $\mathbb{R}^d$. Our main concern here is the behavior of $\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|$. We saw in the previous chapter that a simple consequence of the bounded difference inequality is that

$$\mathbb{P}\left\{\left|\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| - \mathbb{E}\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right| > t\right\} \le 2e^{-2nt^2}$$

for any n and $t > 0$. This shows that for any class $\mathcal{A}$, the maximal deviation is sharply concentrated around its mean. In the rest of this chapter we derive inequalities for the expected value, in terms of certain combinatorial quantities related to $\mathcal{A}$. The first such quantity is the VC *shatter coefficient*, defined by

$$\mathbb{S}_{\mathcal{A}}(n) = \max_{x_1,\dots,x_n\in\mathbb{R}^d} |\{\{x_1, \dots, x_n\} \cap A; A \in \mathcal{A}\}|.$$

Thus, $\mathbb{S}_{\mathcal{A}}(n)$ is the maximal number of different subsets of a set of n points which can be obtained by intersecting it with elements of $\mathcal{A}$. The main theorem is the following version of a classical result of Vapnik and Chervonenkis:

Theorem 9 VAPNIK-CHERVONENKIS INEQUALITY.

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\right\} \le 2\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}}(n)}{n}}.$$

PROOF. Introduce $X_1',\ldots,X_n'$, an independent copy of $X_1,\ldots,X_n$. Also, define n i.i.d. sign variables $\sigma_1,\ldots,\sigma_n$ such that $\mathbb{P}\{\sigma_1=-1\}=\mathbb{P}\{\sigma_1=1\}=1/2$, independent of $X_1,X_1',\ldots,X_n,X_n'$. Then, denoting $\mu_n'(A)=(1/n)\sum_{i=1}^n \mathbb{I}_{[X_i'\in A]}$, we may write

$$\begin{aligned}
&\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\right\}\\
&= \mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mathbb{E}\{\mu_n(A)-\mu_n'(A)|X_1,\ldots,X_n\}|\right\}\\
&\le \mathbb{E}\left\{\sup_{A\in\mathcal{A}}\mathbb{E}\left\{|\mu_n(A)-\mu_n'(A)|\,\middle|\,X_1,\ldots,X_n\right\}\right\}\\
&\qquad\text{(by Jensen's inequality)}\\
&\le \mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_n'(A)|\right\}\\
&\qquad\text{(since }\sup\mathbb{E}(\cdot)\le\mathbb{E}\sup(\cdot))\\
&= \frac{1}{n}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)\right|\right\}\\
&\qquad\text{(because }X_1,X_1',\ldots,X_n,X_n'\text{ are i.i.d.)}\\
&= \frac{1}{n}\mathbb{E}\left\{\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)\right|\,\middle|\,X_1,X_1',\ldots,X_n,X_n'\right\}\right\}.
\end{aligned}$$

Now because of the independence of the σ_i's of the rest of the variables, we may fix the values of $X_1=x_1,X_1'=x_1',\ldots,X_n=x_n,X_n'=x_n'$, and investigate

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

Denote by $\widehat{\mathcal{A}} \subset \mathcal{A}$ a collection of sets such that any two sets in $\widehat{\mathcal{A}}$ have different intersections with the set $\{x_1, x_1', \ldots, x_n, x_n'\}$, and every possible intersection is represented once. Thus, $|\widehat{\mathcal{A}}| \leq \mathbb{S}_{\mathcal{A}}(2n)$, and

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

Observing that each $\sigma_i\left(\mathbb{I}_{[x_i\in A]} - \mathbb{I}_{[x_i'\in A]}\right)$ has zero mean and takes values in $[-1, 1]$, we obtain from Lemma 2 that for any $s > 0$,

$$\mathbb{E}e^{s\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]} - \mathbb{I}_{[x_i'\in A]}\right)} = \prod_{i=1}^{n}\mathbb{E}e^{s\sigma_i\left(\mathbb{I}_{[x_i\in A]} - \mathbb{I}_{[x_i'\in A]}\right)} \leq e^{ns^2/2} .$$

Since the distribution of $\sigma_i\left(\mathbb{I}_{[x_i\in A]} - \mathbb{I}_{[x_i'\in A]}\right)$ is symmetric, Lemma 3 immediately implies that

$$\mathbb{E}\left\{\max_{A\in\widehat{\mathcal{A}}}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]} - \mathbb{I}_{[x_i'\in A]}\right)\right|\right\} \leq \sqrt{2n\log 2\mathbb{S}_{\mathcal{A}}(2n)} .$$

Conclude by observing that $\mathbb{S}_{\mathcal{A}}(2n) \leq \mathbb{S}_{\mathcal{A}}(n)^2$. ∎

Remark. The original form of the Vapnik-Chervonenkis inequality is

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu(A)| > t\right\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-nt^2/8}.$$

A combination of Theorem 9 with the concentration inequality for the supremum quickly yields an inequality of a similar form.

The main virtue of the Vapnik-Chervonenkis inequality is that it converts the problem of uniform deviations of empirical averages into a combinatorial problem. Investigating the behavior of $\mathbb{S}_{\mathcal{A}}(n)$ is the key to the understanding of the behavior of the maximal deviations. Classes for which $\mathbb{S}_{\mathcal{A}}(n)$ grows at a subexponential rate with n are managable in the sense that $\mathbb{E}\{\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu(A)|\}$ converges to zero. More importantly, explicit upper bounds for $\mathbb{S}_{\mathcal{A}}(n)$ provide nonasymptotic distribution-free bounds for the expected maximal deviation (and also for the tail probabilities). Section 4.3 is devoted to some key combinatorial results related to shatter coefficients.

We close this section by a refinement of Theorem 9 due to Massart. The bound below substantially improves the bound of Theorem 9 whenever $\sup_{A\in\mathcal{A}} \mu(A)(1-\mu(A))$ is very small.

Theorem 10 *Let* $\Sigma = \sup_{A\in\mathcal{A}} \sqrt{\mu(A)(1-\mu(A))}$. *Then*

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A)-\mu(A)|\right\} \le \frac{16\log 2\mathbb{S}_{\mathcal{A}}(2n)}{n} + +\sqrt{\frac{32\Sigma^2 \log 2\mathbb{S}_{\mathcal{A}}(2n)}{n}}.$$

PROOF. From the proof of Theorem 9, we have

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

By Hoeffding's inequality, for each set A,

$$\mathbb{E}\left\{e^{s\sum_{i=1}^n \sigma_i\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)}\,\middle|\, X_1,X_1',\dots,X_n,X_n'\right\} \le e^{s^2\sum_{i=1}^n\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)^2/2}\ ,$$

so by Lemma 3 we obtain

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A)-\mu(A)|\right\} \le \frac{1}{n}\mathbb{E}\sup_{A\in\mathcal{A}}\sqrt{\sum_{i=1}^n\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)^2}\sqrt{2\log 2\mathbb{S}_{\mathcal{A}}(2n)}\ .$$

To bound the right-hand side, note that

$$
\begin{aligned}
& \mathbb{E} \sup_{A\in\mathcal{A}} \sqrt{\sum_{i=1}^{n} \left(\mathbb{I}_{[X_i\in A]} - \mathbb{I}_{[X_i'\in A]}\right)^2} \\
\le & \sqrt{\mathbb{E} \sup_{A\in\mathcal{A}} \sum_{i=1}^{n} \left(\mathbb{I}_{[X_i\in A]} - \mathbb{I}_{[X_i'\in A]}\right)^2} \\
\le & \sqrt{\mathbb{E} \sup_{A\in\mathcal{A}} \sum_{i=1}^{n} \left(\left(\mathbb{I}_{[X_i\in A]} - \mu(A)\right) + \left(\mu(A) - \mathbb{I}_{[X_i'\in A]}\right)\right)^2} \\
\le & \sqrt{4\mathbb{E} \sup_{A\in\mathcal{A}} \sum_{i=1}^{n} \left(\mathbb{I}_{[X_i\in A]} - \mu(A)\right)^2} \\
= & 2\sqrt{\mathbb{E} \sup_{A\in\mathcal{A}} \sum_{i=1}^{n} \left[\left(\mathbb{I}_{[X_i\in A]} - \mu(A)\right)(1-\mu(A)) + \mu(A)\left(\mu(A) - \mathbb{I}_{[X_i\in A]}\right) + \mu(A)(1-\mu(A))\right]} \\
\le & 2\sqrt{n\Sigma^2} + 2\sqrt{\mathbb{E} \sup_{A\in\mathcal{A}} \sum_{i=1}^{n} \left(\mathbb{I}_{[X_i\in A]} - \mu(A)\right)} \\
\le & 2\sqrt{n\Sigma^2} + 2\sqrt{n\mathbb{E} \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|} \ .
\end{aligned}
$$

Summarizing, if we denote $\mathbb{E}\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| = M$, we have obtained

$$
M \le \sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}}(2n)}{2n}} \left(\Sigma + \sqrt{M}\right) \ .
$$

This is a quadratic inequality for $\sqrt{M}$, whose solution is just the statement of the theorem. ∎

4.2 Inequalities for relative deviations

In this section we summarize some important improvements of the basic Vapnik-Chervonenkis inequality. The basic result is the following pair of inequalities, due to Vapnik and Chervonenkis (1974). The proof sketched here is due to Anthony and Shawe-Taylor (1993).

Theorem 11 *For every* $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} > \epsilon\right\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-n\epsilon^2/4}$$

and

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\sqrt{\mu_n(A)}} > \epsilon\right\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-n\epsilon^2/4} \;.$$

Sketch of proof. The main steps of the proof are as follows:

1. Symmetrization.

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\sqrt{\mu(A)}} > \epsilon\right\} \leq 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{\mu_n'(A) - \mu_n(A)}{\sqrt{(1/2)(\mu_n'(A) + \mu_n(A))}} > \epsilon\right\}.$$

2. Randomization, conditioning.

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

3. Tail bound. Use the union bound and Hoeffding's inequality to bound the conditional probability inside. ∎

Using the bounds above, we may derive other interesting inequalities. The first inequalities are due to Pollard (1995) and Haussler (1992).

Corollary 1 *For all* $t \in (0,1)$ *and* $s > 0$,

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{\mu(A) - \mu_n(A)}{\mu(A) + \mu_n(A) + s/2} > t\right\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-nst^2/4}$$

and

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{\mu_n(A) - \mu(A)}{\mu(A) + \mu_n(A) + s/2} > t\right\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-nst^2/4}.$$

SKETCH OF PROOF. Take $\alpha > 0$. Considering the cases $\mu(A) < (\alpha + 1)^2\epsilon^2\alpha^{-2}$ and $\mu(A) \geq (\alpha + 1)^2\epsilon^2\alpha^{-2}$ separately, it is easy to show that $\mu(A) - \mu_n(A) \leq \epsilon\sqrt{\mu(A)}$ implies that $\mu(A) \leq (1+\alpha)\mu_n(A) + \epsilon^2(1+\alpha)/\alpha$. Then choosing $\alpha = 2t/(1-t)$ and $\epsilon^2 = st^2/(1-t^2)$ we easily prove that the first inequality in Theorem 11 implies the first inequality. The second inequality follows similarly from the second inequality of Theorem 11. ∎

Finally, we point out another corollary of Theorem 11 which has interesting applications in statistical learning theory:

Corollary 2

$$\mathbb{P}\{\exists A \in \mathcal{A} : \mu(A) > \epsilon \textit{ and } \mu_n(A) \leq (1-t)\mu(A)\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-n\epsilon t^2/4} .$$

In particular, setting $t = 1$*,*

$$\mathbb{P}\{\exists A \in \mathcal{A} : \mu(A) > \epsilon \textit{ and } \mu_n(A) = 0\} \leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-n\epsilon/4} .$$

4.3 Shatter coefficients

Consider a class $\mathcal{A}$ of subsets of $\mathbb{R}^d$, and let $x_1, \ldots, x_n \in \mathbb{R}^d$ be arbitrary points. Recall from the previous section that properties of the finite set $\mathcal{A}(x_1^n) \subset \{0,1\}^n$ defined by

$$\mathcal{A}(x_1^n) = \big\{b = (b_1, \ldots, b_n) \in \{0,1\}^n : \\ b_i = \mathbb{I}_{[x_i \in A]},\ i = 1, \ldots, n \text{ for some } A \in \mathcal{A}\big\}$$

play an essential role in bounding uniform deviations of the empirical measure. In particular, the maximal cardinality of $\mathcal{A}(x_1^n)$

$$\mathbb{S}_{\mathcal{A}}(n) = \max_{x_1,\ldots,x_n \in \mathbb{R}^d} |\mathcal{A}(x_1^n)|$$

(i.e., the shatter coefficient) provides simple bounds via the Vapnik-Chervonenkis inequality. We begin with some elementary properties of the shatter coefficient.

Theorem 12 *Let $\mathcal{A}$ and $\mathcal{B}$ be classes of subsets of $\mathbb{R}^d$, and let $n, m \geq 1$ be integers. Then*

(1) $\mathbb{S}_{\mathcal{A}}(n+m) \le \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{A}}(m)$*;*

(2) If $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$*, then* $\mathbb{S}_{\mathcal{C}}(n) \le \mathbb{S}_{\mathcal{A}}(n) + \mathbb{S}_{\mathcal{B}}(n)$*;*

(3) If $\mathcal{C} = \{C = A^c : A \in \mathcal{A}\}$*, then* $\mathbb{S}_{\mathcal{C}}(n) = \mathbb{S}_{\mathcal{A}}(n)$*;*

(4) If $\mathcal{C} = \{C = A \cap B : A \in \mathcal{A} \text{ and } B \in \mathcal{B}\}$*, then* $\mathbb{S}_{\mathcal{C}}(n) \le \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$*;*

(5) If $\mathcal{C} = \{C = A \cup B : A \in \mathcal{A} \text{ and } B \in \mathcal{B}\}$*, then* $\mathbb{S}_{\mathcal{C}}(n) \le \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$*;*

(6) If $\mathcal{C} = \{C = A \times B : A \in \mathcal{A} \text{ and } B \in \mathcal{B}\}$*, then* $\mathbb{S}_{\mathcal{C}}(n) \le \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$*.*

PROOF. Parts (1), (2), (3), and (6) are immediate from the definition. To show (4), fix $x_1, \dots, x_n$, let $N = |\mathcal{A}(x_1^n)| \le \mathbb{S}_{\mathcal{A}}(n)$, and denote by $A_1, A_2, \dots, A_N$ the different sets of the form $\{x_1, \dots, x_n\} \cap A$ for some $A \in \mathcal{A}$. For all $1 \le i \le N$, sets in $\mathcal{B}$ pick at most $\mathbb{S}_{\mathcal{B}}(|A_i|) \le \mathbb{S}_{\mathcal{B}}(n)$ different subsets of A_i. Thus,

$$|\mathcal{A}(x_1^n)| \le \sum_{i=1}^{N} \mathbb{S}_{\mathcal{B}}(|A_i|) \le \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n).$$

(5) follows from (4) and (3). ∎

The VC *dimension* V of a class $\mathcal{A}$ of sets is defined as the largest integer n such that

$$\mathbb{S}_{\mathcal{A}}(n) = 2^n.$$

If $\mathbb{S}_{\mathcal{A}}(n) = 2^n$ for all n, then we say that $V = \infty$. Clearly, if $\mathbb{S}_{\mathcal{A}}(n) < 2^n$ for some n, then for all $m > n$, $\mathbb{S}_{\mathcal{A}}(m) < 2^m$, and therefore the VC dimension is always well-defined. If $|\mathcal{A}(x_1^n)| = 2^n$ for some points $x_1, \dots, x_n$, then we say that $\mathcal{A}$ *shatters* the set $x_1^n = \{x_1, \dots, x_n\}$. As the next basic result shows, the VC dimension provides a useful bound for the shatter coefficient of a class.

Theorem 13 SAUER'S LEMMA. *Let* $\mathcal{A}$ *be a class of sets with* VC *dimension* $V < \infty$*. Then for all* n*,*

$$\mathbb{S}_{\mathcal{A}}(n) \le \sum_{i=0}^{V} \binom{n}{i}.$$

PROOF. Fix $x_1, \ldots, x_n$, such that $|\mathcal{A}(x_1^n)| = \mathbb{S}_{\mathcal{A}}(n)$. Denote $B_0 = \mathcal{A}(x_1^n) \in \{0,1\}^n$. We say that a set $B \subset \{0,1\}^n$ *shatters* a set $S = \{s_1, \ldots, s_m\} \subset \{1, 2, \ldots, n\}$ if the restriction of B to the components $s_1, \ldots, s_m$ is the full m-dimensional binary hypercube, that is,

$$\{(b_{s_1}, \ldots, b_{s_m}) : b = (b_1, \ldots, b_n) \in B\} = \{0,1\}^m.$$

It suffices to show that the cardinality of any set $B_0 \subset \{0,1\}^n$ that cannot shatter any set of size $m > V$, is at most $\sum_{i=0}^{V} \binom{n}{i}$. This is done by transforming B_0 into a set B_n with $|B_n| = |B_0|$ such that any set shattered by B_n is also shattered by B_0. Moreover, it will be easy to see that $|B_n| \leq \sum_{i=0}^{V} \binom{n}{i}$.

For every vector $b = (b_1, \ldots, b_n) \in B_0$, if $b_1 = 1$, then flip the first component of b to zero unless $(0, b_2, \ldots, b_n) \in B_0$. If $b_1 = 0$, then keep the vector unchanged. The set of vectors B_1 obtained this way obviously has the same cardinality as that of B_0. Moreover, if B_1 shatters a set $S = \{s_1, s_2, \ldots, s_m\} \subset \{1, \ldots, n\}$, then B_0 also shatters S. This is trivial if $1 \notin S$. If $1 \in S$, then we may assume without loss of generality that $s_1 = 1$. The fact that B_1 shatters S implies that for any $v \in \{0,1\}^{m-1}$ there exists a $b \in B_1$ such that $b_1 = 1$ and $(b_{s_2}, \ldots, b_{s_m}) = v$. By the construction of B_1 this is only possible if for any $u \in \{0,1\}^m$ there exists a $b' \in B_0$ such that $(b'_{s_1}, \ldots, b'_{s_m}) = u$. This means that B_0 also shatters S.

Now starting from B_1, execute the same transformation, but now by flipping the second component of each vector, if necessary. Again, the cardinality of the obtained set B_2 remains unchanged, and any set shattered by B_2 is also shattered by B_1 (and therefore also by B_0). Repeat the transformation for all components, arriving at the set B_n. Clearly, B_n cannot shatter sets of cardinality larger than V, since otherwise B_0 would shatter sets of the same size. On the other hand, it is easy to see that B_n is such that for every $b \in B_n$, all vectors of form $c = (c_1, \ldots, c_n)$ with $c_i \in \{b_i, 0\}$ for $1 \leq i \leq n$, are also in B_n. Then B_n is a subset of a set of form

$$T = \{b \in \{0,1\}^n : b_i = 0 \text{ if } v_i = 0\},$$

where $v = (v_1, \ldots, v_n)$ is a fixed vector containing at most V 1's. This implies that

$$\mathbb{S}_{\mathcal{A}}(n) = |B_0| = |B_n| \le |T| = \sum_{i=0}^{V} \binom{n}{i} ,$$

concluding the proof. ∎

The following corollary makes the meaning of Sauer's lemma more transparent:

Corollary 3 *Let $\mathcal{A}$ be a class of sets with* VC *dimension $V < \infty$. Then for all n,*

$$\mathbb{S}_{\mathcal{A}}(n) \le (n+1)^V,$$

and for all $n \ge V$,

$$\mathbb{S}_{\mathcal{A}}(n) \le \left(\frac{ne}{V}\right)^V .$$

PROOF. By the binomial theorem,

$$(n+1)^V = \sum_{i=0}^{V} n^i \binom{V}{i} = \sum_{i=0}^{V} \frac{n^i V!}{i!(V-i)!} \ge \sum_{i=0}^{V} \frac{n^i}{i!} \ge \sum_{i=0}^{V} \binom{n}{i} .$$

On the other hand, if $V/n \le 1$, then

$$\left(\frac{V}{n}\right)^V \sum_{i=0}^{V} \binom{n}{i} \le \sum_{i=0}^{V} \left(\frac{V}{n}\right)^i \binom{n}{i} \le \sum_{i=0}^{n} \left(\frac{V}{n}\right)^i \binom{n}{i} = \left(1 + \frac{V}{n}\right)^n \le e^V ,$$

where again we used the binomial theorem. ∎

Recalling the Vapnik-Chervonenkis inequality, we see that if $\mathcal{A}$ is any class of sets with VC dimension V, then

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right\} \le 2\sqrt{\frac{V\log(n+1) + \log 2}{n}} ,$$

that is, whenever $\mathcal{A}$ has a finite VC dimension, the expected largest deviation over $\mathcal{A}$ converges to zero at a rate $O(\sqrt{\log n/n})$.

Next we calculate the VC dimension of some simple classes.

Lemma 5 *If $\mathcal{A}$ is the class of all rectangles in $\mathbb{R}^d$, then $V = 2d$.*

PROOF. To see that there are $2d$ points that can be shattered by $\mathcal{A}$, just consider the $2d$ vectors with $d-1$ zero components, and one non-zero component which is either 1 or -1. On the other hand, for any given set of $2d+1$ points we can choose a subset of at most $2d$ points with the property that it contains a point with largest first coordinate, a point with smallest first coordinate, a point with largest second coordinate, and so forth. Clearly, there is no set in $\mathcal{A}$ which contains these points, but not the rest. ∎

Lemma 6 *Let $\mathcal{G}$ be an m-dimensional vector space of real-valued functions defined on $\mathbb{R}^d$. The class of sets*

$$\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}$$

has VC *dimension $V \leq m$.*

PROOF. It suffices to show that no set of size $m+1$ can be shattered by sets of the form $\{x : g(x) \geq 0\}$. Fix $m+1$ arbitrary points $x_1, \ldots, x_{m+1}$, and define the linear mapping $L : \mathcal{G} \to \mathbb{R}^{m+1}$ as

$$L(g) = (g(x_1), \ldots, g(x_{m+1})) \ .$$

Then the image of $\mathcal{G}$, $L(\mathcal{G})$, is a linear subspace of $\mathbb{R}^{m+1}$ of dimension not exceeding m. This implies the existence of a nonzero vector $\gamma = (\gamma_1, \ldots, \gamma_{m+1}) \in \mathbb{R}^{m+1}$ orthogonal to $L(\mathcal{G})$, that is, for every $g \in \mathcal{G}$,

$$\gamma_1 g(x_1) + \ldots + \gamma_{m+1} g(x_{m+1}) = 0 \ .$$

We may assume that at least one of the γ_i's is negative. Rearranging this equality so that all terms with nonnegative γ_i stay on the left-hand side, we get

$$\sum_{i:\gamma_i \geq 0} \gamma_i g(x_i) = \sum_{i:\gamma_i < 0} -\gamma_i g(x_i) \ .$$

Now suppose that there exists a $g \in \mathcal{G}$ such that the set $\{x : g(x) \geq 0\}$ picks exactly the x_i's on the left-hand side. Then all terms on the left-hand side are nonnegative, while the terms on the right-hand side must be negative, which is a contradiction, so $x_1, \ldots, x_{m+1}$ cannot be shattered, which implies the statement. $\square$

Generalizing a result of Schläffli (1950), Cover (1965) showed that if $\mathcal{G}$ is defined as the linear space of functions spanned by functions $\psi_1, \ldots, \psi_m : \mathbb{R}^d \to \mathbb{R}$, and the vectors $\Psi(x_i) = (\psi_1(x_i), \ldots, \psi_m(x_i))$, $i = 1, 2, \ldots, n$ are linearly independent, then for the class of sets $\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}$ we have

$$|\mathcal{A}(x_1^n)| = 2\sum_{i=0}^{m-1}\binom{n-1}{i},$$

which often gives a slightly sharper estimate than Sauer's lemma. The proof is left as an exercise. Now we may immediately deduce the following:

Corollary 4 *(1) If $\mathcal{A}$ is the class of all linear halfspaces, that is, subsets of $\mathbb{R}^d$ of the form $\{x : a^T x \geq b\}$, where $a \in \mathbb{R}^d, b \in \mathbb{R}$ take all possible values, then $V \leq d + 1$.*

(2) If $\mathcal{A}$ is the class of all closed balls in $\mathbb{R}^d$, that is, sets of the form

$$\left\{x = (x^{(1)}, \ldots, x^{(d)}) : \sum_{i=1}^{d} |x^{(i)} - a_i|^2 \leq b\right\}, \quad a_1, \ldots, a_d, b \in \mathbb{R},$$

then $V \leq d + 2$.

(3) If $\mathcal{A}$ is the class of all ellipsoids in $\mathbb{R}^d$, that is, sets of form $\{x : x^T \Sigma^{-1} x \leq 1\}$, where Σ is a positive definite symmetric matrix, then $V \leq d(d+1)/2 + 1$.

Note that the above-mentioned result implies that the VC dimension of the class of all linear halfspaces actually equals $d+1$. Dudley (1979) proved that in the case of the class of all closed balls the above inequality is not tight, and the VC dimension equals $d + 1$ (see exercise 5).

4.4 Applications to empirical risk minimization

In this section we apply the main results of the previous sections to obtain upper bounds for the performance of empirical risk minimization.

Recall the scenario set up in Chapter 2: $\mathcal{C}$ is a class of classifiers containing decision functions of the form $g : \mathbb{R}^d \to \{0,1\}$. The data $(X_1,Y_1),\ldots,(X_n,Y_n)$ may be used to calculate the empirical error $\widehat{L}_n(g)$ for any $g \in \mathcal{C}$. g_n^* denotes a classifier minimizing $\widehat{L}_n(g)$ over the class, that is,

$$\widehat{L}_n(g_n^*) \le \widehat{L}_n(g) \qquad \text{for all } g \in \mathcal{C}.$$

Denote the probability of error of the optimal classifier in the class by $L_{\mathcal{C}}$, that is,

$$L_{\mathcal{C}} = \inf_{g\in\mathcal{C}} L(g).$$

(Here we implicitely assume that the infimum is achieved. This assumption is motivated by convenience in the notation, it is not essential.)

The basic Lemma 1 shows that

$$L(g_n^*) - L_{\mathcal{C}} \le 2\sup_{g\in\mathcal{C}} \left|\widehat{L}_n(g) - L(g)\right|.$$

Thus, the quantity of interest is the maximal deviation between empirical probabilities of error and their expectation over the class. Such quantities are estimated by the Vapnik-Chervonenkis inequality. Indeed, $\sup_{g\in\mathcal{C}}\left|\widehat{L}_n(g) - L(g)\right|$ is exactly of the form of $\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu(A)|$, where the role of the class of sets $\mathcal{A}$ is now played by the class of error sets

$$\left\{(x,y)\in\mathbb{R}^d\times\{0,1\} : g(x)\neq y\right\}; \quad g\in\mathcal{C}.$$

Denote the class of these error sets by $\bar{\mathcal{A}}$. Thus, the Vapnik-Chervonenkis inequality immediately bounds the expected maximal deviation in terms of the xshatter coefficients (or VC dimension) of the class of error sets.

Instead of error sets, it is more convenient to work with classes of sets of the form

$$\left\{x\in\mathbb{R}^d : g(x) = 1\right\}; \quad g\in\mathcal{C}.$$

We denote the class of sets above by $\mathcal{A}$. The next simple fact shows that the classes $\bar{\mathcal{A}}$ and $\mathcal{A}$ are equivalent from a combinatorial point of view:

Lemma 7 *For every n we have $\mathbb{S}_{\bar{\mathcal{A}}}(n) = \mathbb{S}_{\mathcal{A}}(n)$, and therefore the corresponding* VC *dimensions are also equal: $V_{\bar{\mathcal{A}}} = V_{\mathcal{A}}$.*

PROOF. Let N be a positive integer. We show that for any n pairs from $\mathbb{R}^d \times \{0,1\}$, if N sets from $\bar{\mathcal{A}}$ pick N different subsets of the n pairs, then there are N corresponding sets in $\mathcal{A}$ that pick N different subsets of n points in $\mathbb{R}^d$, and vice versa. Fix n pairs $(x_1, 0), \ldots, (x_m, 0), (x_{m+1}, 1), \ldots, (x_n, 1)$. Note that since ordering does not matter, we may arrange any n pairs in this manner. Assume that for a certain set $A \in \mathcal{A}$, the corresponding set $\bar{A} = A \times \{0\} \bigcup A^c \times \{1\} \in \bar{\mathcal{A}}$ picks out the pairs $(x_1, 0), \ldots, (x_k, 0), (x_{m+1}, 1), \ldots, (x_{m+l}, 1)$, that is, the set of these pairs is the intersection of $\bar{A}$ and the n pairs. Again, we can assume without loss of generality that the pairs are ordered in this way. This means that A picks from the set $\{x_1, \ldots, x_n\}$ the subset $\{x_1, \ldots, x_k, x_{m+l+1}, \ldots, x_n\}$, and the two subsets uniquely determine each other. This proves $\mathbb{S}_{\bar{\mathcal{A}}}(n) \le \mathbb{S}_{\mathcal{A}}(n)$. To prove the other direction, notice that if A picks a subset of k points $x_1, \ldots, x_k$, then the corresponding set $\bar{A} \in \bar{\mathcal{A}}$ picks the pairs with the same indices from $\{(x_1, 0), \ldots, (x_k, 0)\}$. Equality of the VC dimensions follows from the equality of the shatter coefficients. ∎

From this point on, we will denote the common value of $\mathbb{S}_{\bar{\mathcal{A}}}(n)$ and $\mathbb{S}_{\mathcal{A}}(n)$ by $\mathbb{S}_{\mathcal{C}}(n)$, and refer to is as the n-th shatter coefficient of the class $\mathcal{C}$. It is simply the maximum number of different ways n points can be classified by classifiers in the class $\mathcal{C}$. Similarly, $V_{\bar{\mathcal{A}}} = V_{\mathcal{A}}$ will be referred to as the VC dimension of the class $\mathcal{C}$, and will be denoted by $V_{\mathcal{C}}$.

Now we are prepared to summarize our main performance bound for empirical risk minimization:

Corollary 5

$$\mathbb{E}L(g_n^*) - L_{\mathcal{C}} \le 4\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{C}}(n)}{n}} \le 4\sqrt{\frac{V_{\mathcal{C}} \log(n+1) + \log 2}{n}}$$

Bounds for $\mathbb{P}\{L(g_n^*) - L_{\mathcal{C}} > \epsilon\}$ may now be easily obtained by combining the corollary above with the bounded difference inequality.

The inequality above may be improved in various different ways. In the appendix of this chapter we show that the factor of $\log n$ in the upper bound is unnecessary, it may be replaced by a suitable constant. In practice, however, often the sample size is so small that the inequality above provides smaller numerical values.

On the other hand, the main performance may be improved in another direction. To understand the reason, consider first an extreme situation when $L_{\mathcal{C}} = 0$, that is, there exists a classifier in $\mathcal{C}$ which classifies without error. (This also means that for som $g' \in \mathcal{C}$, $Y = g'(X)$ with probability one, a very restrictive assumption. Nevertheless, the assumption that $L_{\mathcal{C}} = 0$ is common in computational learning theory, see Blumer, Ehrenfeucht, Haussler, and Warmuth (1989). In such a case, clearly $\widehat{L}_n(g^*) = 0$, and the second statement of Corollary 2 implies that

$$\mathbb{P}\{L(g_n^*) - L_{\mathcal{C}} > \epsilon\} = \mathbb{P}\{L(g_n^*) > \epsilon\} \le 4\mathbb{S}_{\mathcal{C}}(2n)e^{-n\epsilon/4} \ ,$$

and therefore

$$\mathbb{E}L(g_n^*) - L_{\mathcal{C}} = \mathbb{E}L(g_n^*) \le \frac{4\ln\left(4\mathbb{S}_{\mathcal{C}}(2n)\right)}{n}.$$

(The bound on the expected value may be obtained by the following simple bounding argument: assume that for some random variable Z, for all $\epsilon > 0$, $\mathbb{P}\{Z > \epsilon\} \le Ce^{-K\epsilon}$ for some positive constants. Then $\mathbb{E}Z = \int_0^\infty \mathbb{P}\{Z > \epsilon\}d\epsilon \le u + \int_u^\infty Ce^{-K\epsilon}$ for any $u > 0$. Integrating, and choosing u to minimize the upper bound, we obtain $\mathbb{E}Z \le \ln C/K$.)

The main point here is that the upper bound obtained in this special case is of smaller order of magnitude than in the general case ($O(V_{\mathcal{C}} \ln n/n)$ as opposed to $O\left(\sqrt{V_{\mathcal{C}} \ln n/n}\right)$.) Intuition suggests that if $L_{\mathcal{C}}$ is nonzero but very small, the general bound of Corollary 5 should be improvable. In fact, the argument below shows that it is possible interpolate between the special case $L_{\mathcal{C}} = 0$ and the fully distribution-free bound of Corollary 5:

Theorem 14

$$\mathbb{E}L(g_n^*) - L_{\mathcal{C}} \le \sqrt{\frac{8L_{\mathcal{C}}\ln(5\mathbb{S}_{\mathcal{C}}(2n)) + 2}{n}} + \frac{8\ln(10\mathbb{S}_{\mathcal{C}}(2n)) + 4}{n}.$$

Also, for every $\epsilon > 0$,

$$\mathbb{P}\{L(g_n^*) - L_{\mathcal{C}} > \epsilon\} \le 5\mathbb{S}_{\mathcal{C}}(2n)e^{-n\epsilon^2/16(L_{\mathcal{C}}+\epsilon)}.$$

Proof. For any $\epsilon > 0$, if

$$\sup_{g\in\mathcal{C}} \frac{L(g)-\widehat{L}_n(g)}{\sqrt{L(g)}} \le \frac{\epsilon}{\sqrt{L_{\mathcal{C}}+2\epsilon}},$$

then for each $g \in \mathcal{C}$

$$\widehat{L}_n(g) \ge L(g) - \epsilon\sqrt{\frac{L(g)}{L_{\mathcal{C}}+2\epsilon}}.$$

If, in addition, g is such that $L(g) > L_{\mathcal{C}} + 2\epsilon$, then by the monotonicity of the function $x - c\sqrt{x}$ (for $c > 0$ and $x > c^2/4$),

$$\widehat{L}_n(g) \ge L_{\mathcal{C}} + 2\epsilon - \epsilon\sqrt{\frac{L_{\mathcal{C}}+2\epsilon}{L_{\mathcal{C}}+2\epsilon}} = L_{\mathcal{C}} + \epsilon.$$

Therefore,

$$\mathbb{P}\left\{\inf_{g:L(g)>L_{\mathcal{C}}+2\epsilon} \widehat{L}_n(g) < L_{\mathcal{C}} + \epsilon\right\} \le \mathbb{P}\left\{\sup_{g\in\mathcal{C}} \frac{L(g)-\widehat{L}_n(g)}{\sqrt{L(g)}} > \frac{\epsilon}{\sqrt{L_{\mathcal{C}}+2\epsilon}}\right\}.$$

But if $L(g_n^*) - L_{\mathcal{C}} > 2\epsilon$, then, denoting by g' a classifier in $\mathcal{C}$ such that $L(g') = L_{\mathcal{C}}$, there exists an $g \in \mathcal{C}$ such that $L(g) > L_{\mathcal{C}} + 2\epsilon$ and $\widehat{L}_n(g) \le \widehat{L}_n(g')$. Thus,

$$\begin{aligned}
&\mathbb{P}\{L(g_n^*) - L_{\mathcal{C}} > 2\epsilon\} \\
&\quad \le \mathbb{P}\left\{\inf_{g:L(g)>L_{\mathcal{C}}+2\epsilon} \widehat{L}_n(g) < \widehat{L}_n(g')\right\} \\
&\quad \le \mathbb{P}\left\{\inf_{g:L(g)>L_{\mathcal{C}}+2\epsilon} \widehat{L}_n(g) < L_{\mathcal{C}} + \epsilon\right\} + \mathbb{P}\{\widehat{L}_n(g') > L_{\mathcal{C}} + \epsilon\} \\
&\quad \le \mathbb{P}\left\{\sup_{g\in\mathcal{C}} \frac{L(g)-\widehat{L}_n(g)}{\sqrt{L(g)}} > \frac{\epsilon}{\sqrt{L_{\mathcal{C}}+2\epsilon}}\right\} + \mathbb{P}\{\widehat{L}_n(g') - L_{\mathcal{C}} > \epsilon\}.
\end{aligned}$$

Bounding the last two probabilities by Theorem 11 and Bernstein's inequality, respectively, we obtain the probability bound of the statement.

The upper bound for the expected value may now be derived by some straightforward calculations which we sketch here: let $u \le L_{\mathcal{C}}$ be a positive number. Then, using the tail inequality obtained above,

$$
\begin{aligned}
& \mathbb{E}L(g_n^*) - L_{\mathcal{C}} \\
& = \int_0^\infty \mathbb{P}\{L(g_n^*) - L_{\mathcal{C}} > \epsilon\} d\epsilon \\
& \leq u + \int_u^\infty 5\mathbb{S}_{\mathcal{C}}(2n) \max\left(e^{-n\epsilon^2/8L_{\mathcal{C}}}, e^{-n\epsilon/8}\right) d\epsilon \\
& \leq \left(u/2 + \int_u^\infty 5\mathbb{S}_{\mathcal{C}}(2n) e^{-n\epsilon^2/8L_{\mathcal{C}}} d\epsilon\right) \\
& \quad + \left(u/2 + \int_u^\infty 5\mathbb{S}_{\mathcal{C}}(2n) e^{-n\epsilon/8} d\epsilon\right).
\end{aligned}
$$

The second term may be bounded as in the argument given fot the case $L_{\mathcal{C}} = 0$, while the first term may be calculated similarly, using the additional observation that

$$
\begin{aligned}
\int_u^\infty e^{-n\epsilon^2} d\epsilon & \leq \frac{1}{2} \int_u^\infty \left(2 + \frac{1}{n\epsilon^2}\right) e^{-n\epsilon^2} d\epsilon \\
& = \frac{1}{2} \left[\frac{1}{n\epsilon} e^{-n\epsilon^2}\right]_u^\infty .
\end{aligned}
$$

The details are omitted. ▮

4.5 Appendix: sharper bounds via chaining

In this section we present an improvement of the Vapnik-Chervonenkis inequality stating that for any class $\mathcal{A}$ of sets of VC dimension V,

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

where c is a universal constant. This in turn implies for empirical risk minimization that

$$\mathbb{E}L(g_n^*) - L_{\mathcal{C}} \leq 2c\sqrt{\frac{V_{\mathcal{C}}}{n}}.$$

The new bound involves some geometric and combinatorial quantities related to the class $\mathcal{A}$. Consider a pair of bit vectors $b = (b_1, \ldots, b_n)$ and

$c = (c_1, \ldots, c_n)$ from $\{0,1\}^n$, and define their distance by

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

Thus, $\rho(b,c)$ is just the square root of the normalized Hamming distance between b and c. Observe that ρ may also be considered as the normalized euclidean distance between the corners of the hypercube $[0,1]^n \subset \mathbb{R}^n$, and therefore it is indeed a distance.

Now let $B \subset \{0,1\}^n$ be any set of bit vectors, and define a *cover* of radius $r > 0$ as a set $B_r \subset \{0,1\}^n$ such that for any $b \in B$ there exists a $c \in B_r$ such that $\rho(b,c) \leq r$. The *covering number* $N(r,B)$ is the cardinality of the smallest cover of radius r.

A class $\mathcal{A}$ of subsets of $\mathbb{R}^d$ and a set of n points $x_1^n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ define a set of bit vectors by

$$\mathcal{A}(x_1^n) = \left\{ b = (b_1, \ldots, b_n) \in \{0,1\}^n : b_i = \mathbb{I}_{[x_i \in A]},\ i = 1, \ldots, n \ \text{ for some } A \in \mathcal{A} \right\}.$$

That is, every bit vector $b \in \mathcal{A}(x_1^n)$ describes the intersection of $\{x_1, \ldots, x_n\}$ with a set A in $\mathcal{A}$. We have the following:

Theorem 15

$$\mathbb{E}\left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right\} \leq \frac{24}{\sqrt{n}} \max_{x_1,\ldots,x_n \in \mathbb{R}^d} \int_0^1 \sqrt{\log 2N(r, \mathcal{A}(x_1^n))}\, dr \ .$$

The theorem implies that $\mathbb{E}\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\} = O(1/\sqrt{n})$ whenever the integral in the bound is uniformly bounded over all $x_1, \ldots, x_n$ and all n. Note that the bound of Theorem 9 is always of larger order of magnitude, trivial cases excepted. The main additional idea is Dudley's *chaining* trick.

PROOF. As in the proof of Theorem 9, we see that

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

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

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

Now it follows from Lemma 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 3 implies that for each $1 \le k \le M$,

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

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 15.

Theorem 16 *Let $\mathcal{A}$ be a class of sets with* VC *dimension $V < \infty$. For every $x_1, \dots, x_n \in \mathbb{R}^d$ 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 16 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 $x_1, \dots, x_n$, 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 $c^{(1)}, \dots, c^{(M)}$, 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:

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

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 $i, j \le M$, $i \ne j$, 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

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

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 15 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.

5. Minimax lower bounds

The purpose of this chapter 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

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

and also

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

In this chapter 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 $g : \mathbb{R}^d \to \{0, 1\}$. 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.

5.1 The zero-error case

In this section we obtain lower bounds under the assumption that the best classifier in the class has zero error probability. Recall that by Corollary 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 17 *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 $n \geq V - 1$,*

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

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 $V - 1 \leq n$. 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,

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

The last probability may be viewed as the error probability of the decision function $g_n : \mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\})^n \to \{0,1\}$ 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

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

This concludes the proof. ▮

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 18 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 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 18 *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 $X_1, Y_1, \ldots, X_n, Y_n$,

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

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

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

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

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

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,

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

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

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

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

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

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 $k_0 = |\{j \leq k : y_j = 0\}|$ and $k_1 = |\{j \leq k : y_j = 1\}|$, we see that

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

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

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

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

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

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

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

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

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

Summarizing what we have obtained so far, we have

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

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 $2nL/(V-1) \geq 9$, then $c \leq 1/6$. With this choice for c, using $L = (V-1)p(1/2-c)$, straightforward calculation provides

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

The condition $p(V-1) \leq 1$ implies that we need to ask that $n \geq (V-1)/(2L(1-2L)^2)$. This concludes the proof of Theorem 18. ∎

6. Complexity regularization

This chapter 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.

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*

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

for all ϵ.

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:

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

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 19 *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 19 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 19 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 $\epsilon > 0$,

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

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 f_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. ∎

6.2 Selection based on a test sample

In our first application of Theorem 19, 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 19 to give the following result.

Corollary 6 *Assume that the model selection algorithm is performed with the hold-out error estimate. Then*

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

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.

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,

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

Therefore, satisfies Assumption 1 with $m = n$. Substituting this into Theorem 19 gives

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

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

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

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

Proof. Once again, we check Assumption 1 and apply Theorem 19. 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,

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

The bounded difference inequality inequality (Theorem 8) implies

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

$$\mathbb{P}\left[\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g)-\widehat{L}_n^{(2)}(g)\right)<\mathbb{E}\max_{g\in\mathcal{C}_k}\left(\widehat{L}_n^{(1)}(g)-\widehat{L}_n^{(2)}(g)\right)-\epsilon\right]\le e^{-n\epsilon^2/2} \quad (3)$$

and so for each k,

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

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

References

General

[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 r.v.'s

[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.

[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*, to appear.

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

[31] Anthony, M. and Shawe-Taylor, J. 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] Devroye, L. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

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

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

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

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

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

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

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

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

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

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

[44] 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.

[45] 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.

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

Shatter coefficients, VC dimension

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

[48] 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.

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

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

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

[52] 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.

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

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

[55] 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.

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

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

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

Lower bounds.

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

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

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

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

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

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

[65] 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.

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

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

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

[69] 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

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

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

[72] 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.

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

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

[75] 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.

[76] 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.

[77] 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.

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

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

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

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

[82] 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.

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

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

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

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

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

[88] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la faculté des sciences de l'université de Toulouse, Mathématiques*, série 6, **IX**, to appear.

[89] 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.

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

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

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

[93] 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.

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

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

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