1

Pattern classification and learning theory

Gábor Lugosi

1.1 A binary classification problem

Pattern recognition (or *classification* or *discrimination*) is about guessing or predicting the unknown class of an observation. An *observation* is a collection of numerical measurements, represented by a d-dimensional vector x. 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 $g(x) : \mathcal{R}^d \to \{0, 1\}$ which represents one's guess of y given x. The mapping g is called a *classifier*. A 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 $\mathcal{R}^d \times \{0, 1\}$-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 $A \subseteq \mathcal{R}^d$,

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

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

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

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 $g : \mathcal{R}^d \to \{0, 1\}$ 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 1.1. *For any classifier* $g : \mathcal{R}^d \to \{0,1\}$,

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

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

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

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

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

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 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,

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

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

In our model, we have access to a data base 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. Even though averaging over the data as well is unnatural, since in a given application, one has to live with the data at hand, the number $\mathbb{E}L_n = \mathbb{P}\{g_n(X) \neq Y\}$ which indicates the quality on an average data sequence, provides useful information, especially if the random variable L_n is concentrated around its mean with high probability.

1.2 Empirical risk minimization

Assume that a class $\mathcal{C}$ of classifiers $g : \mathcal{R}^d \to \{0, 1\}$ 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

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

of the selected rule we have:

Lemma 1.1.

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

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

Proof.

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

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.

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.

1.3 Concentration inequalities

1.3.1 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 $t > 0$,

$$\mathbb{P}\{X \ge t\} \le \frac{\mathbb{E}X}{t}.$$

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

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

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

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

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 (1963):

Lemma 1.2. *Let X be a random variable with $\mathbb{E}X = 0$, $a \le X \le b$. Then for $s > 0$,*

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

Proof. Note that by convexity of the exponential function

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

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

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

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,

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

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

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

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

$$\begin{aligned} &\mathbb{P}\{S_n - \mathbb{E}S_n \ge \epsilon\} \\ &\le e^{-s\epsilon}\prod_{i=1}^n \mathbb{E}\left\{e^{s(X_i - \mathbb{E}X_i)}\right\} \\ &\le e^{-s\epsilon}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} \quad \text{(by Lemma 1.2)} \\ &= e^{-s\epsilon}e^{s^2\sum_{i=1}^n(b_i-a_i)^2/8} \\ &= e^{-2\epsilon^2/\sum_{i=1}^n(b_i-a_i)^2} \quad \text{(by choosing } s = 4\epsilon/\textstyle\sum_{i=1}^n(b_i-a_i)^2\text{)}. \end{aligned}$$

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 1.2. (Hoeffding's inequality). *Let $X_1, \dots, X_n$ 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 1.1 to bound the performance of empirical risk minimization in the special case when the class $\mathcal{C}$ contains finitely many classifiers:

Theorem 1.3. *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 1.2 with the following simple result:

Lemma 1.3. *Let $\sigma > 0$, $n \geq 2$, and let $Y_1, \dots, Y_n$ 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, $\mathbb{E}\{e^{s(-Y_i)}\} \le e^{s^2\sigma^2/2}$ *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 $s > 0$,

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

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

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

1.3.2 Other inequalities for sums

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

Theorem 1.4. 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* $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}\{X_i\}$*. 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

$$\mathbb{E}\{e^{sX_i}\} = 1 + s\mathbb{E}\{X_i\} + \sum_{r=2}^{\infty}\frac{s^r\mathbb{E}\{X_i^r\}}{r!} = 1 + s^2\,\mathbf{Var}\{X_i\}F_i \le e^{s^2\,\mathbf{Var}\{X_i\}F_i}$$

with $F_i = \sum_{r=2}^{\infty} s^{r-2}\mathbb{E}\{X_i^r\}/\left(r!\,\mathbf{Var}\{X_i\}\right)$. We may use the boundedness of the X_i's to show that $\mathbb{E}\{X_i^r\} \le c^{r-2}\,\mathbf{Var}\{X_i\}$, 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 1.5. Bernstein's inequality. *Under the conditions of the previous exercise, for any $t > 0$,*

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

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

Theorem 1.6. *Let $X_1, \ldots, X_n$ be independent random variables, taking their values from $[0,1]$. If $m = \mathbb{E}S_n$, then for any $m \leq t \leq n$,*

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

Also,

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

and for all $\epsilon > 0$,

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

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

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

1.3.3 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 $g : A^n \to \mathcal{R}$ 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 $\mathbb{E}\{XY\} = \mathbb{E}\{\mathbb{E}\{XY|Y\}\} = \mathbb{E}\{Y\mathbb{E}\{X|Y\}\}$.

Te first result of this section is an improvement of an inequality of Efron and Stein (1981) proved by Steele (1986). We have learnt the short proof given here from Stéphane Boucheron.

Theorem 1.7. Efron-Stein inequality. *If $X_1', \ldots, X_n'$ form an independent copy of $X_1, \ldots, X_n$, then*

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

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

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

since, for any $i > j$,

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

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

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

and therefore

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

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

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

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

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

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

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

Theorem 1.8. THE BOUNDED DIFFERENCE INEQUALITY. *Under the bounded difference assumption above, for all $t > 0$,*

$$\mathbb{P}\left\{g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n) \geq t\right\} \leq e^{-2t^2 / \sum_{i=1}^n c_i^2} ,$$

and

$$\mathbb{P}\left\{\mathbb{E}g(X_1,\ldots,X_n) - g(X_1,\ldots,X_n) \geq t\right\} \leq e^{-2t^2 / \sum_{i=1}^n c_i^2} .$$

McDiarmid (1989) proved this inequality using martingale techniques, which we reproduce here. The proof of Theorem 1.8 uses the following straightforward extension of Lemma 1.2:

Lemma 1.4. *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 1.8. Just like in the proof of Theorem 1.7, 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

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

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, $Z_i \le V_i \le W_i$ with probability one, and also

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

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

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

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

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

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 $g : \mathcal{R}^d \to \{0,1\}$, then

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

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 1.8 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.

1.4 Vapnik-Chervonenkis theory

1.4.1 The Vapnik-Chervonenkis inequality

Recall from Section 1.3.1 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 $X_1,\ldots,X_n$ be i.i.d. random variables taking values in $\mathcal{R}^d$ with common distribution

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

Define the empirical distribution

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

Consider a class $\mathcal{A}$ of subsets of $\mathcal{R}^d$. Our main concern here is the behavior of the random variable $\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|$. 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

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

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 1.9. 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

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

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, $|\widehat{\mathcal{A}}| \le \mathbb{S}_{\mathcal{A}}(2n)$, 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 1.2 that for any $s > 0$,

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

Since the distribution of $\sigma_i\left(\mathbb{I}_{[x_i\in A]} - \mathbb{I}_{[x_i'\in A]}\right)$ is symmetric, Lemma 1.3 immediately implies that

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

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

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

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

A combination of Theorem 1.9 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 1.4.3 is devoted to some key combinatorial results related to shatter coefficients.

We close this section by a refinement of Theorem 1.9 due to Massart (2000). The bound below substantially improves the bound of Theorem 1.9 whenever $\sup_{A\in\mathcal{A}} \mu(A)(1-\mu(A))$ is very small.

Theorem 1.10. *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 1.9, we have

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

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 1.3 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

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

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. □

1.4.2 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 1.11. *For every $\epsilon > 0$,*

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

and

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

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

1. Symmetrization.

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

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

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

and

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

Sketch of proof. Take $\alpha > 0$. Considering the cases $\mu(A) < (\alpha+1)^2\epsilon^2\alpha^{-2}$ and $\mu(A) \ge (\alpha+1)^2\epsilon^2\alpha^{-2}$ separately, it is easy to show that $\mu(A)-\mu_n(A) \le \epsilon\sqrt{\mu(A)}$ implies that $\mu(A) \le (1+\alpha)\mu_n(A)+\epsilon^2(1+\alpha)/\alpha$. Then choosing $\alpha = 2t/(1-t)$ and $\epsilon^2 = st^2/(1-t^2)$ we easily prove that the first inequality in Theorem 1.11 implies the first inequality. The second inequality follows similarly from the second inequality of Theorem 1.11. □

Finally, we point out another corollary of Theorem 1.11 which has interesting applications in statistical learning theory:

Corollary 1.2.

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

In particular, setting $t = 1$,

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

1.4.3 Shatter coefficients

Consider a class $\mathcal{A}$ of subsets of $\mathcal{R}^d$, and let $x_1, \ldots, x_n \in \mathcal{R}^d$ 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 $\mathcal{A}(x_1^n)$

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

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

Theorem 1.12. *Let $\mathcal{A}$ and $\mathcal{B}$ be classes of subsets of $\mathcal{R}^d$, and let $n, m \ge 1$ 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 $x_1, \ldots, x_n$, let $N = |\mathcal{A}(x_1^n)| \le \mathbb{S}_{\mathcal{A}}(n)$, and denote by $A_1, A_2, \ldots, A_N$ the different sets of the form $\{x_1, \ldots, x_n\} \cap A$ 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 $x_1, \ldots, x_n$, then we say that $\mathcal{A}$ *shatters* the set $x_1^n = \{x_1, \ldots, x_n\}$. As the next basic result shows, the VC dimension provides a useful bound for the shatter coefficient of a class.

Theorem 1.13. SAUER'S LEMMA. *Let $\mathcal{A}$ be a class of sets with* VC *dimension $V < \infty$. Then for all n,*

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

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 $1 \le i \le n$, 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 1.3. *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 1.5. *If $\mathcal{A}$ is the class of all rectangles in $\mathcal{R}^d$, 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 1.6. *Let $\mathcal{G}$ be an m-dimensional vector space of real-valued functions defined on $\mathcal{R}^d$. 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 $L : \mathcal{G} \to \mathcal{R}^{m+1}$ 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 $\mathcal{R}^{m+1}$ of dimension not exceeding m. This implies the existence of a nonzero vector $\gamma = (\gamma_1, \ldots, \gamma_{m+1}) \in \mathcal{R}^{m+1}$ 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. □

Generalizing a result of Schläffli (1950), Cover (1965) showed that if $\mathcal{G}$ is defined as the linear space of functions spanned by functions $\psi_1, \ldots, \psi_m : \mathcal{R}^d \to \mathcal{R}$, 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 1.4. *(1) If $\mathcal{A}$ is the class of all linear halfspaces, that is, subsets of $\mathcal{R}^d$ of the form $\{x : a^T x \geq b\}$, where $a \in \mathcal{R}^d, b \in \mathcal{R}$ take all possible values, then $V \leq d+1$.*

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

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

then $V \leq d+2$.

(3) If $\mathcal{A}$ is the class of all ellipsoids in $\mathcal{R}^d$, 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).

1.4.4 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 $g : \mathcal{R}^d \to \{0,1\}$. 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,

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

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 1.1 shows that

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

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, the random variable $\sup_{g\in C}\left|\widehat{L}_n(g) - L(g)\right|$ is 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

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

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

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

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 1.7. *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 $\mathcal{R}^d \times \{0,1\}$, 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 $\mathcal{R}^d$, 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 $\mathbb{S}_{\bar{\mathcal{A}}}(n) \leq \mathbb{S}_{\mathcal{A}}(n)$. 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 1.5.

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

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 1.2 implies that

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

and therefore

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

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

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 1.5 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 1.5:

Theorem 1.14.

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

Also, for every $\epsilon > 0$,

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

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 1.11 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. □

1.4.5 Convex combinations of classifiers

Several important classification methods form a classifier as a convex combination of simple functions. To describe such a situation, consider a class $\mathcal{C}$ of classifiers $g : \mathcal{R}^d \to \{0, 1\}$. Think of $\mathcal{C}$ as a small class of "base" classifiers such as the class of all linear splits of $\mathcal{R}^d$. In general we assume that the VC dimension $V_{\mathcal{C}}$ of $\mathcal{C}$ is finite. Define the class $\mathcal{F}$ as the class of functions $f : \mathcal{R}^d \to [0, 1]$ of the form

$$f(x) = \sum_{j=1}^N w_j g_j(x)$$

where N is any positive integer, $w_1, \ldots, w_N$ are nonnegative weights with $\sum_{j=1}^N w_j = 1$, and $g_1, \ldots, g_N \in \mathcal{C}$. Thus, $\mathcal{F}$ may be considered as the convex hull of $\mathcal{C}$. Each function $f \in \mathcal{F}$ defines a classifier g_f, in a natural way, by

$$g_f(x) = \begin{cases} 1 & \text{if } f(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

A large variety of "boosting" and "bagging" methods, based mostly on the work of Schapire (1990), Freund (1995) and Breiman (1996), construct classifiers as convex combinations of very simple functions. Typically the class of classifiers defined this way is too large in the sense that it is impossible to obtain meaningful distribution-free upper bounds for $\sup_{f\in\mathcal{F}}\Big(L(g_f)-\widehat{L}_n(g_f)\Big)$. Indeed, even in the simple case when $d=1$ and $\mathcal{C}$ is the class of all linear splits of the real line, the class of all g_f is easily seen to have an infinite VC dimension.

Surprisingly, however, meaningful bounds may be obtained if we replace the empirical probability of error $\widehat{L}_n(g_f)$ by a slightly larger quantity. To this end, let $\gamma>0$ be a fixed parameter, and define the *margin error* by

$$L_n^\gamma(g_f)=\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{[f(X_i)(1-2Y_i)<\gamma]}.$$

Notice that for all $\gamma>0$, $L_n^\gamma(g_f)\ge\widehat{L}_n(g_f)$ and the $L_n^\gamma(g_f)$ is increasing in γ. An interpretation of the margin error $L_n^\gamma(g_f)$ is that it counts, apart from the number of misclassified pairs (X_i,Y_i), also those which are well classified but only with a small "confidence" (or "margin") by g_f.

The purpose of this section is to present a result of Freund, Schapire, Bartlett, and Lee (1998) which states that the margin error is always a good approximate upper bound for the probability of error, at least if γ is not too small. The elegant proof shown here is due to Koltchinskii and Panchenko (2002).

Theorem 1.15. *For every $\epsilon>0$,*

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left(L(g_f)-L_n^\gamma(g_f)\right)>\frac{2\sqrt{2}}{\gamma}\sqrt{\frac{V_{\mathcal{C}}\log(n+1)}{n}}+\epsilon\right\}\le e^{-2n\epsilon^2}.$$

Thus, with very high probability, the probability of error of any classifier g_f, $f\in\mathcal{F}$, may be simultaneously upper bounded by the sum

$$L_n^\gamma(g_f)+\frac{2\sqrt{2}}{\gamma}\sqrt{\frac{V_{\mathcal{C}}\log(n+1)}{n}}$$

plus a term of the order $n^{-1/2}$. Notice that, as γ grows, the first term of the sum increases, while the second decreases. The bound can be very useful whenever a classifier has a small margin error for a relatively large γ (i.e., if the classifier classifies the training data well with high "confidence") since the second term only depends on the VC dimension of the small base class $\mathcal{C}$. As shown in the next section, the second term in the above sum may be replaced by $(c/\gamma)\sqrt{V_{\mathcal{C}}/n}$ for some universal constant c.

The proof of the theorem crucially uses the following simple lemma, called the "contraction principle". Here we cite a version tailored for our needs. For the proof, see Ledoux and Talagrand (1991), pages 112–113.

Lemma 1.8. *Let $Z_1(f), \ldots, Z_n(f)$ be arbitrary real-valued bounded random variables indexed by an abstract parameter f and let $\sigma_1, \ldots, \sigma_n$ be independent symmetric sign variables, independent of the $Z_i(f)$'s (i.e., $\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = 1\} = 1/2$). If $\phi : \mathcal{R} \to \mathcal{R}$ is a Lipschitz function such that $|\phi(x) - \phi(y)| \le |x - y|$ with $\phi(0) = 0$, then*

$$\mathbb{E}\sup_f \sum_{i=1}^n \sigma_i \phi(Z_i(f)) \le \mathbb{E}\sup_f \sum_{i=1}^n \sigma_i Z_i(f).$$

Proof of Theorem 1.15. For any $\gamma > 0$, introduce the function

$$\phi_\gamma(x) = \begin{cases} 1 & \text{if } x \le 0 \\ 0 & \text{if } x \ge \gamma \\ 1 - x/\gamma & \text{if } x \in (0, \gamma) \end{cases}$$

Observe that $\mathbb{I}_{[x \le 0]} \le \phi_\gamma(x) \le \mathbb{I}_{[x \le \gamma]}$. Thus,

$$\sup_{f \in \mathcal{F}} (L(g_f) - L_n^\gamma(g_f)) \le \sup_{f \in \mathcal{F}} \left(\mathbb{E}\phi_\gamma((1-2Y)f(X)) - \frac{1}{n}\sum_{i=1}^n \phi_\gamma((1-2Y_i)f(X)) \right).$$

Introduce the notation $Z(f) = (1-2Y)f(X)$ and $Z_i(f) = (1-2Y_i)f(X_i)$. Clearly, by the bounded difference inequality,

$$\begin{aligned} \mathbb{P}\Bigg\{ \sup_{f \in \mathcal{F}} & \left(\mathbb{E}\phi_\gamma(Z(f)) - \frac{1}{n}\sum_{i=1}^n \phi_\gamma(Z_i(f)) \right) \\ & > \mathbb{E}\sup_{f \in \mathcal{F}} \left(\mathbb{E}\phi_\gamma(Z(f)) - \frac{1}{n}\sum_{i=1}^n \phi_\gamma(Z_i(f)) \right) + \epsilon \Bigg\} \le e^{-2n\epsilon^2} \end{aligned}$$

and therefore it suffices to prove that the expected value of the supremum is bounded by $\frac{2\sqrt{2}}{\gamma}\sqrt{\frac{V_{\mathcal{C}} \log(n+1)}{n}}$. As a first step, we proceed by a symmetrization argument just like in the proof of Theorem 1.9 to obtain

$$\begin{aligned} \mathbb{E}\sup_{f \in \mathcal{F}} \left(\mathbb{E}\phi_\gamma(Z(f)) - \frac{1}{n}\sum_{i=1}^n \phi_\gamma(Z_i(f)) \right) & \le \mathbb{E}\sup_{f \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n \sigma_i \left(\phi_\gamma(Z_i'(f)) - \phi_\gamma(Z_i(f))\right) \right) \\ & \le 2\mathbb{E}\sup_{f \in \mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n \sigma_i \left(\phi_\gamma(Z_i(f)) - \phi_\gamma(0)\right) \right) \end{aligned}$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. symmetric sign variables and $Z_i'(f) = (1 - 2Y_i')f(X_i')$ where the (X_i', Y_i') are independent of the (X_i, Y_i) and have the same distribution as that of the pairs (X_i, Y_i).

Observe that the function $\phi(x) = \gamma(\phi_\gamma(x) - \phi_\gamma(0))$ is Lipschitz and $\phi(0) = 0$, therefore, by the contraction principle (Lemma 1.8),

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\left(\phi_\gamma(Z_i(f)) - \phi_\gamma(0)\right) \le \frac{1}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i Z_i(f) = \frac{1}{\gamma}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$$

where at the last step we used the fact that $\sigma_i(1-2Y_i)$ is a symmetric sign variable, independent of the X_i and therefore $\sigma_i(1-2Y_i)f(X_i)$ has the same distribution as that of $\sigma_i f(X_i)$. The last expectation may be rewritten as

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) = \frac{1}{n}\mathbb{E}\sup_{N\ge 1}\sup_{g_1,\dots,g_N\in\mathcal{C}}\sup_{w_1,\dots,w_N}\sum_{i=1}^n\sum_{j=1}^N w_j\sigma_i g_j(X_i).$$

The key observation is that for any N and base classifiers $g_i,\dots,g_N$, the supremum in

$$\sup_{w_1,\dots,w_N}\sum_{i=1}^n\sum_{j=1}^N w_j\sigma_i g_j(X_i)$$

is achieved for a weight vector which puts all the mess in one index, that is, when $w_j = 1$ for some j. (This may be seen by observing that a linear function over a convex polygon achieves its maximum at one of the vertices of the polygon.) Thus,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) = \frac{1}{n}\mathbb{E}\sup_{g\in\mathcal{C}}\sum_{i=1}^n \sigma_i g(X_i).$$

However, repeating the argument in the proof of Theorem 1.9 with the necessary adjustments, we obtain

$$\frac{1}{n}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\sum_{i=1}^n \sigma_i g(X_i)\right| \le \sqrt{\frac{2\log \mathbb{S}_{\mathcal{C}}(n)}{n}} \le \sqrt{\frac{2V_{\mathcal{C}}\log(n+1)}{n}}$$

which completes the proof of the desired inequality. □

1.4.6 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,

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

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

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

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 $[0,1]^n \subset \mathcal{R}^n$, 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 $\mathcal{R}^d$ and a set of n points $x_1^n = \{x_1, \ldots, x_n\} \subset \mathcal{R}^d$ 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 1.16.

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

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 1.9 is always of larger order of magnitude, trivial cases excepted. The main additional idea is Dudley's *chaining* trick.