As we noted in Section 12.1, such bounds yield performance bounds for any classifier selected by minimizing the empirical error. To make the material more digestible, we first present the main ideas in a simple one-dimensional setting, and then prove the general theorem in the next section.

We drop the pattern recognition setting momentarily, and return to probability theory. The following theorem is sometimes referred to as the fundamental theorem of mathematical statistics, stating uniform almost sure convergence of the empirical distribution function to the true one:

Theorem 12.4. (GLIVENKO-CANTELLI THEOREM). *Let $Z_1, \ldots, Z_n$ be i.i.d. real-valued random variables with distribution function $F(z) = \mathbf{P}\{Z_1 \leq z\}$. Denote the standard empirical distribution function by*

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i \leq z\}}.$$

Then

$$\mathbf{P}\left\{ \sup_{z \in \mathcal{R}} |F(z) - F_n(z)| > \epsilon \right\} \leq 8(n+1)e^{-n\epsilon^2/32},$$

and, in particular, by the Borel-Cantelli lemma,

$$\lim_{n \to \infty} \sup_{z \in \mathcal{R}} |F(z) - F_n(z)| = 0 \quad \textit{with probability one.}$$

PROOF. The proof presented here is not the simplest possible, but it contains the main ideas leading to a powerful generalization. Introduce the notation $\nu(A) = \mathbf{P}\{Z_1 \in A\}$ and $\nu_n(A) = (1/n) \sum_{j=1}^{n} I_{\{Z_j \in A\}}$ for all measurable sets $A \subset \mathcal{R}$. Let $\mathcal{A}$ denote the class of sets of form $(-\infty, z]$ for $z \in \mathcal{R}$. With these notations,

$$\sup_{z \in \mathcal{R}} |F(z) - F_n(z)| = \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|.$$

We prove the theorem in several steps, following symmetrization ideas of Dudley (1978), and Pollard (1984). We assume that $n\epsilon^2 \geq 2$, since otherwise the bound is trivial. In the first step we introduce a symmetrization.

STEP 1. FIRST SYMMETRIZATION BY A GHOST SAMPLE. Define the random variables $Z_1', \ldots, Z_n' \in \mathcal{R}$ such that $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$ are all independent and identically distributed. Denote by ν_n' the empirical measure corresponding to the new sample:

$$\nu_n'(A) = \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i' \in A\}}.$$

Then for $n\epsilon^2 \geq 2$ we have

$$\mathbf{P}\left\{ \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon \right\} \leq 2\mathbf{P}\left\{ \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu_n'(A)| > \frac{\epsilon}{2} \right\}.$$

To see this, let $A^* \in \mathcal{A}$ be a set for which $|\nu_n(A^*) - \nu(A^*)| > \epsilon$ if such a set exists, and let A^* be a fixed set in $\mathcal{A}$ otherwise. Then

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu_n'(A)| > \epsilon/2\right\} \\
&\quad \geq \mathbf{P}\left\{|\nu_n(A^*) - \nu_n'(A^*)| > \epsilon/2\right\} \\
&\quad \geq \mathbf{P}\left\{|\nu_n(A^*) - \nu(A^*)| > \epsilon, |\nu_n'(A^*) - \nu(A^*)| < \frac{\epsilon}{2}\right\} \\
&\quad = \mathbf{E}\left\{I_{\{|\nu_n(A^*)-\nu(A^*)|>\epsilon\}}\mathbf{P}\left\{|\nu_n'(A^*) - \nu(A^*)| < \frac{\epsilon}{2}\middle| Z_1, \ldots, Z_n\right\}\right\}.
\end{aligned}$$

The conditional probability inside may be bounded by Chebyshev's inequality as follows:

$$\begin{aligned}
\mathbf{P}\left\{|\nu_n'(A^*) - \nu(A^*)| < \frac{\epsilon}{2}\middle| Z_1, \ldots, Z_n\right\} &\geq 1 - \frac{\nu(A^*)(1-\nu(A^*))}{n\epsilon^2/4} \\
&\geq 1 - \frac{1}{n\epsilon^2} \geq \frac{1}{2}
\end{aligned}$$

whenever $n\epsilon^2 \geq 2$. In summary,

$$\begin{aligned}
\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu_n'(A)| > \epsilon/2\right\} &\geq \frac{1}{2}\mathbf{P}\{|\nu_n(A^*) - \nu(A^*)| > \epsilon\} \\
&\geq \frac{1}{2}\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\}.
\end{aligned}$$

STEP 2. SECOND SYMMETRIZATION BY RANDOM SIGNS. Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. sign variables, independent of $Z_1, \ldots, Z_n$ and $Z_1', \ldots, Z_n'$, with $\mathbf{P}\{\sigma_i = -1\} = \mathbf{P}\{\sigma_i = 1\} = 1/2$. Clearly, because $Z_1, Z_1', \ldots, Z_n, Z_n'$ are all independent and identically distributed, the distribution of

$$\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^n (I_A(Z_i) - I_A(Z_i'))\right|$$

is the same as the distribution of

$$\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i(I_A(Z_i) - I_A(Z_i'))\right|.$$

Thus, by Step 1,

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\} \\
&\quad \leq 2\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^n (I_A(Z_i) - I_A(Z_i'))\right| > \frac{\epsilon}{2}\right\} \\
&\quad = 2\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i(I_A(Z_i) - I_A(Z_i'))\right| > \frac{\epsilon}{2}\right\}.
\end{aligned}$$

Simply applying the union bound, we can remove the auxiliary random variables $Z_1', \ldots, Z_n'$:

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i (I_A(Z_i) - I_A(Z_i'))\right| > \frac{\epsilon}{2}\right\} \\
&\quad \leq \mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4}\right\} + \mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i')\right| > \frac{\epsilon}{4}\right\} \\
&\quad = 2\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4}\right\}.
\end{aligned}$$

STEP 3. CONDITIONING. To bound the probability

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4}\right\} = \mathbf{P}\left\{\sup_{z\in\mathcal{R}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_{\{Z_i\leq z\}}\right| > \frac{\epsilon}{4}\right\},$$

we condition on $Z_1, \ldots, Z_n$. Fix $z_1, \ldots, z_n \in \mathcal{R}^d$, and note that as z ranges over $\mathcal{R}$, the number of different vectors $\left(I_{\{z_1\leq z\}}, \ldots, I_{\{z_n\leq z\}}\right)$ is at most $n+1$. Thus, conditional on $Z_1, \ldots, Z_n$, the supremum in the probability above is just a maximum taken over at most $n+1$ random variables. Thus, applying the union bound gives

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \,\middle|\, Z_1, \ldots, Z_n\right\} \\
&\quad \leq (n+1) \sup_{A\in\mathcal{A}} \mathbf{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \,\middle|\, Z_1, \ldots, Z_n\right\}.
\end{aligned}$$

With the supremum now outside the probability, it suffices to find an exponential bound on the conditional probability

$$\mathbf{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \,\middle|\, Z_1, \ldots, Z_n\right\}.$$

STEP 4. HOEFFDING'S INEQUALITY. With $z_1, \ldots, z_n$ fixed, $\sum_{i=1}^{n} \sigma_i I_A(z_i)$ is the sum of n independent zero mean random variables bounded between -1 and 1. Therefore, Theorem 8.1 applies in a straightforward manner:

$$\mathbf{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \,\middle|\, Z_1, \ldots, Z_n\right\} \leq 2e^{-n\epsilon^2/32}.$$

Thus,

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \,\middle|\, Z_1, \ldots, Z_n\right\} \leq 2(n+1)e^{-n\epsilon^2/32}.$$

Taking the expected value on both sides we have

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4}\right\} \le 2(n+1)e^{-n\epsilon^2/32}.$$

In summary,

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)| > \epsilon\right\} \le 8(n+1)e^{-n\epsilon^2/32}. \quad \square$$

12.4 Uniform Deviations of Relative Frequencies from Probabilities

In this section we prove the Vapnik-Chervonenkis inequality, a mighty generalization of Theorem 12.4. In the proof we need only a slight adjustment of the proof above. In the general setting, let the independent identically distributed random variables $Z_1, \ldots, Z_n$ take their values from $\mathcal{R}^d$. Again, we use the notation $\nu(A) = \mathbf{P}\{Z_1 \in A\}$ and $\nu_n(A) = (1/n)\sum_{j=1}^{n} I_{\{Z_j\in A\}}$ for all measurable sets $A \subset \mathcal{R}^d$. The Vapnik-Chervonenkis theory begins with the concepts of *shatter coefficient* and *Vapnik-Chervonenkis (or* VC*) dimension*:

DEFINITION 12.1. *Let $\mathcal{A}$ be a collection of measurable sets. For $(z_1, \ldots, z_n) \in \{\mathcal{R}^d\}^n$, let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ be the number of different sets in*

$$\{\{z_1, \ldots, z_n\} \cap A; A \in \mathcal{A}\}.$$

The n-th shatter coefficient of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(z_1,\ldots,z_n)\in\{\mathcal{R}^d\}^n} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

That is, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$.

The shatter coefficients measure the richness of the class $\mathcal{A}$. Clearly, $s(\mathcal{A}, n) \le 2^n$, as there are 2^n subsets of a set with n elements. If $N_{\mathcal{A}}(z_1, \ldots, z_n) = 2^n$ for some $(z_1, \ldots, z_n)$, then we say that $\mathcal{A}$ *shatters* $\{z_1, \ldots, z_n\}$. If $s(\mathcal{A}, n) < 2^n$, then any set of n points has a subset such that there is no set in $\mathcal{A}$ that contains exactly that subset of the n points. Clearly, if $s(\mathcal{A}, k) < 2^k$ for some integer k, then $s(\mathcal{A}, n) < 2^n$ for all $n > k$. The first time when this happens is important:

DEFINITION 12.2. *Let $\mathcal{A}$ be a collection of sets with $|\mathcal{A}| \ge 2$. The largest integer $k \ge 1$ for which $s(\mathcal{A}, k) = 2^k$ is denoted by $V_{\mathcal{A}}$, and it is called the Vapnik-Chervonenkis dimension (or* VC *dimension) of the class $\mathcal{A}$. If $s(\mathcal{A}, n) = 2^n$ for all n, then by definition, $V_{\mathcal{A}} = \infty$.*

For example, if $\mathcal{A}$ contains all halflines of form $(-\infty, x]$, $x \in \mathcal{R}$, then $s(\mathcal{A}, 2) = 3 < 2^2$, and $V_{\mathcal{A}} = 1$. This is easily seen by observing that for any two different points $z_1 < z_2$ there is no set of the form $(-\infty, x]$ that contains z_2, but not z_1. A class of sets $\mathcal{A}$ for which $V_{\mathcal{A}} < \infty$ is called a Vapnik-Chervonenkis (or VC) class. In a sense, $V_{\mathcal{A}}$ may be considered as the complexity, or size, of $\mathcal{A}$. Several properties of the shatter coefficients and the VC dimension will be shown in Chapter 13. The main purpose of this section is to prove the following important result by Vapnik and Chervonenkis (1971):

Theorem 12.5. (VAPNIK AND CHERVONENKIS (1971)). *For any probability measure ν and class of sets $\mathcal{A}$, and for any n and $\epsilon > 0$,*

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\} \le 8s(\mathcal{A}, n)e^{-n\epsilon^2/32}.$$

PROOF. The proof parallels that of Theorem 12.4. We may again assume that $n\epsilon^2 \ge 2$. In the first two steps we prove that

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\} \le 4\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4}\right\}.$$

This may be done exactly the same way as in Theorem 12.4; we do not repeat the argument. The only difference appears in Step 3:

STEP 3. CONDITIONING. To bound the probability

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4}\right\},$$

again we condition on $Z_1, \ldots, Z_n$. Fix $z_1, \ldots, z_n \in \mathcal{R}^d$, and observe that as A ranges over $\mathcal{A}$, the number of different vectors $(I_A(z_1), \ldots, I_A(z_n))$ is just the number of different subsets of $\{z_1, \ldots, z_n\}$ produced by intersecting it with sets in $\mathcal{A}$, which, by definition, cannot exceed $s(\mathcal{A}, n)$. Therefore, with $Z_1, \ldots, Z_n$ fixed, the supremum in the above probability is a maximum of at most $N_{\mathcal{A}}(Z_1, \ldots, Z_n)$ random variables. This number, by definition, is bounded from above by $s(\mathcal{A}, n)$. By the union bound we get

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \middle| Z_1, \ldots, Z_n\right\} \\
&\quad \le s(\mathcal{A}, n) \sup_{A\in\mathcal{A}} \mathbf{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \middle| Z_1, \ldots, Z_n\right\}.
\end{aligned}$$

Therefore, as before, it suffices to bound the conditional probability

$$\mathbf{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_A(Z_i)\right| > \frac{\epsilon}{4} \middle| Z_1, \ldots, Z_n\right\}.$$

This may be done by Hoeffding's inequality exactly as in Step 4 of the proof of Theorem 12.4. Finally, we obtain

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)|>\epsilon\right\}\le 8s(\mathcal{A},n)e^{-n\epsilon^2/32}. \quad \square$$

The bound of Theorem 12.5 is useful when the shatter coefficients do not increase too quickly with n. For example, if $\mathcal{A}$ contains all Borel sets of $\mathcal{R}^d$, then we can shatter any collection of n different points at will, and obtain $s(\mathcal{A},n)=2^n$. This would be useless, of course. The smaller $\mathcal{A}$, the smaller the shatter coefficient is. To apply the VC bound, it suffices to compute shatter coefficients for certain families of sets. Examples may be found in Cover (1965), Vapnik and Chervonenkis (1971), Devroye and Wagner (1979a), Feinholz (1979), Devroye (1982a), Massart (1983), Dudley (1984), Simon (1991), and Stengle and Yukich (1989). This list of references is far from exhaustive. More information about shatter coefficients is given in Chapter 13.

REMARK. MEASURABILITY. The supremum in Theorem 12.5 is not always measurable. Measurability must be verified for every family $\mathcal{A}$. For all our examples, the quantities are indeed measurable. For more on the measurability question, see Dudley (1978; 1984), Massart (1983), and Gaenssler (1983). Giné and Zinn (1984) and Yukich (1985) provide further work on suprema of the type shown in Theorem 12.5. $\square$

REMARK. OPTIMAL EXPONENT. For the sake of readability we followed the line of Pollard's proof (1984) instead of the original by Vapnik and Chervonenkis. In particular, the exponent $-n\epsilon^2/32$ in Theorem 12.5 is worse than the $-n\epsilon^2/8$ established in the original paper. The best known exponents together with some other related results are mentioned in Section 12.8. The basic ideas of the original proof by Vapnik and Chervonenkis appear in the proof of Theorem 12.7 below. $\square$

REMARK. NECESSARY AND SUFFICIENT CONDITIONS. It is clear from the proof of the theorem that it can be strengthened to

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)|>\epsilon\right\}\le 8\mathbf{E}\left\{N_{\mathcal{A}}(Z_1,\ldots,Z_n)\right\}e^{-n\epsilon^2/32},$$

where $Z_1,\ldots,Z_n$ are i.i.d. random variables with probability measure ν. Although this upper bound is tighter than that in the stated inequality, it is usually more difficult to handle, since the coefficient in front of the exponential term depends on the distribution of Z_1, while $s(\mathcal{A},n)$ is purely combinatorial in nature. However, this form is important in a different setting: we say that the *uniform law of large numbers* holds if

$$\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)|\to 0 \quad \text{in probability.}$$