rest of the function class looks like. Having one function which gets picked as soon as we have seen one data point would essentially void the inherently *asymptotic* notion of consistency.

Theorem 5.3 (Vapnik & Chervonenkis (e.g., [562])) *One-sided uniform convergence in probability,*

$$\lim_{m\to\infty} \mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} = 0, \tag{5.21}$$

for all $\epsilon > 0$*, is a* necessary and sufficient *condition for nontrivial consistency of empirical risk minimization.*

As explained above, consistency, and thus learning, crucially depends on the set of functions. In Section 5.1, we gave an example where we considered the set of all possible functions, and showed that learning was impossible. The dependence of learning on the set of functions has now returned in a different guise: the condition of uniform convergence will crucially depend on the set of functions for which it must hold.

The abstract characterization in Theorem 5.3 of consistency as a uniform convergence property, whilst theoretically intriguing, is not all that useful in practice. We do not want to check some fairly abstract convergence property every time we want to use a learning machine. Therefore, we next address whether there are properties of learning machines, i.e., of sets of functions, which *ensure* uniform convergence of risks.

5.5 How to Derive a VC Bound

We now take a closer look at the subject of Theorem 5.3; the probability

$$\mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f] - R_{\mathrm{emp}}[f]) > \epsilon\}. \tag{5.22}$$

We give a simplified account, drawing from the expositions of [561, 562, 415, 238]. We do not aim to describe or even develop the theory to the extent that would be necessary to give precise bounds for SVMs, say. Instead, our goal will be to convey central insights rather than technical details. For more complete treatments geared specifically towards SVMs, cf. [562, 491, 24]. We focus on the case of pattern recognition; that is, on functions taking values in $\{\pm 1\}$.

Two tricks are needed along the way: the *union bound* and the method of *symmetrization by a ghost sample*.

5.5.1 The Union Bound

Suppose the set $\mathcal{F}$ consists of two functions, f_1 and f_2. In this case, uniform convergence of risk trivially follows from the law of large numbers, which holds

for each of the two. To see this, let

$$C_\epsilon^i := \{(x_1, y_1), \ldots, (x_m, y_m) | (R[f_i] - R_{\text{emp}}[f_i]) > \epsilon\} \tag{5.23}$$

denote the set of samples for which the risks of f_i differ by more than ϵ. Then, by definition, we have

$$\mathrm{P}\{\sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon\} = \mathrm{P}(C_\epsilon^1 \cup C_\epsilon^2). \tag{5.24}$$

The latter, however, can be rewritten as

$$\mathrm{P}(C_\epsilon^1 \cup C_\epsilon^2) = \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2) - \mathrm{P}(C_\epsilon^1 \cap C_\epsilon^2) \leq \mathrm{P}(C_\epsilon^1) + \mathrm{P}(C_\epsilon^2), \tag{5.25}$$

where the last inequality follows from the fact that P is nonnegative. Similarly, if $\mathcal{F} = \{f_1, \ldots, f_n\}$, we have

$$\mathrm{P}\{\sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon\} = \mathrm{P}(C_\epsilon^1 \cup \ldots \cup C_\epsilon^n) \leq \sum_{i=1}^{n} \mathrm{P}(C_\epsilon^i). \tag{5.26}$$

Union Bound

This inequality is called the *union bound*. As it is a crucial step in the derivation of risk bounds, it is worthwhile to emphasize that it becomes an equality if and only if all the events involved are *disjoint*. In practice, this is rarely the case, and we therefore lose a lot when applying (5.26). It is a step with a large "slack."

Nevertheless, when $\mathcal{F}$ is finite, we may simply apply the law of large numbers (5.11) for each individual $\mathrm{P}(C_\epsilon^i)$, and the sum in (5.26) then leads to a constant factor n on the right hand side of the bound — it does not change the exponentially fast convergence of the empirical risk towards the actual risk. In the next section, we describe an ingenious trick used by Vapnik and Chervonenkis, to reduce the infinite case to the finite one. It consists of introducing what is sometimes called a *ghost sample*.

5.5.2 Symmetrization

The central observation in this section is that we can bound (5.22) in terms of a probability of an event referring to a *finite* function class. Note first that the empirical risk term in (5.22) effectively refers only to a finite function class: for any given training sample of m points $x_1, \ldots, x_m$, the functions of $\mathcal{F}$ can take at most 2^m different values $y_1, \ldots, y_m$ (recall that the y_i take values only in $\{\pm 1\}$). In addition, the probability that the empirical risk differs from the actual risk by more than ϵ, can be bounded by the twice the probability that it differs from the empirical risk on a *second* sample of size m by more than $\epsilon/2$.

Lemma 5.4 (Symmetrization (Vapnik & Chervonenkis) (e.g. [559])) *For $m\epsilon^2 \geq 2$, we have*

Symmetrization

$$\mathrm{P}\{\sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon\} \leq 2\mathrm{P}\{\sup_{f \in \mathcal{F}}(R_{\text{emp}}[f] - R'_{\text{emp}}[f]) > \epsilon/2\}. \tag{5.27}$$

Here, the first P *refers to the distribution of iid samples of size m, while the second one*

refers to iid samples of size $2m$*. In the latter case,* R_{emp} *measures the loss on the first half of the sample, and* R'_{emp} *on the second half.*

Although we do not prove this result, it should be fairly plausible: if the empirical error rates on two independent m-samples are close to each other, then they should also be close to the true error rate.

5.5.3 The Shattering Coefficient

The main result of Lemma 5.4 is that it implies, for the purpose of bounding (5.22), that the function class $\mathcal{F}$ is effectively finite: restricted to the $2m$ points appearing on the right hand side of (5.27), it has *at most* 2^{2m} elements. This is because only the outputs of the functions on the patterns of the sample count, and there are $2m$ patterns with two possible outputs, ± 1. The number of effectively different functions can be smaller than 2^{2m}, however; and for our purposes, this is the case that will turn out to be interesting.

Let $Z_{2m} := \big((x_1, y_1), \ldots, (x_{2m}, y_{2m})\big)$ be the given $2m$-sample. Denote by $\mathcal{N}(\mathcal{F}, Z_{2m})$ the cardinality of $\mathcal{F}$ when restricted to $\{x_1, \ldots, x_{2m}\}$, that is, the number of functions from $\mathcal{F}$ that can be distinguished from their values on $\{x_1, \ldots, x_{2m}\}$. Let us, moreover, denote the maximum (over all possible choices of a $2m$-sample) number of functions that can be distinguished in this way as $\mathcal{N}(\mathcal{F}, 2m)$.

Shattering Coefficient

The function $\mathcal{N}(\mathcal{F}, m)$ is referred to as the *shattering coefficient*, or in the more general case of regression estimation, the *covering number* of $\mathcal{F}$.[6] In the case of pattern recognition, which is what we are currently looking at, $\mathcal{N}(\mathcal{F}, m)$ has a particularly simple interpretation: it is the number of different outputs $(y_1, \ldots, y_m)$ that the functions in $\mathcal{F}$ can achieve on samples of a given size.[7] In other words, it simply measures the *number of ways that the function class can separate the patterns into two classes*. Whenever $\mathcal{N}(\mathcal{F}, m) = 2^m$, all possible separations can be implemented by functions of the class. In this case, the function class is said to *shatter* m points. Note that this means that there *exists* a set of m patterns which can be separated in all possible ways — it does not mean that this applies to *all* sets of m patterns.

Shattering

5.5.4 Uniform Convergence Bounds

Let us now take a closer look at the probability that for a $2m$-sample Z_{2m} drawn iid from P, we get a one-sided uniform deviation larger than $\epsilon/2$ (cf. (5.27)),

$$\mathrm{P}\{\sup_{f \in \mathcal{F}} (R_{\text{emp}}[f] - R'_{\text{emp}}[f]) > \epsilon/2\}. \tag{5.28}$$

6. In regression estimation, the covering number also depends on the accuracy within which we are approximating the function class, and on the loss function used; see Section 12.4 for more details.

7. Using the zero-one loss $c(x, y, f(x)) = 1/2|f(x) - y| \in \{0, 1\}$, it also equals the number of different loss vectors $(c(x_1, y_1, f(x_1)), \ldots, c(x_m, y_m, f(x_m)))$.

The basic idea now is to pick a maximal set of functions $\{f_1, \ldots, f_{\mathcal{N}(\mathcal{F}, Z_{2m})}\}$ that can be distinguished based on their values on Z_{2m}, then use the union bound, and finally bound each term using the Chernoff inequality. However, the fact that the f_i depend on the sample Z_{2m} will make things somewhat more complicated. To deal with this, we have to introduce an auxiliary step of randomization, using a uniform distibution over permutations σ of the $2m$-sample Z_{2m}.

Let us denote the empirical risks on the two halves of the sample after the permutation σ by $R^{\sigma}_{\mathrm{emp}}[f]$ and $R'^{\sigma}_{\mathrm{emp}}[f]$, respectively. Since the $2m$-sample is iid, the permutation does not affect (5.28). We may thus instead consider

$$\mathrm{P}_{Z_{2m},\sigma}\{\sup_{f\in\mathcal{F}}(R^{\sigma}_{\mathrm{emp}}[f] - R'^{\sigma}_{\mathrm{emp}}[f]) > \epsilon/2\}, \tag{5.29}$$

where the subscripts of P were added to clarify what the distribution refers to. We next rewrite this as

$$\int_{(\mathcal{X}\times\{\pm1\})^{2m}} \mathrm{P}_{\sigma|Z_{2m}}\{\sup_{f\in\mathcal{F}|_{Z_{2m}}} (R^{\sigma}_{\mathrm{emp}}[f] - R'^{\sigma}_{\mathrm{emp}}[f]) > \epsilon/2 \mid Z_{2m}\}\, d\mathrm{P}(Z_{2m}). \tag{5.30}$$

We can now express the event $C_\epsilon := \{\sigma | \sup_{f\in\mathcal{F}|_{Z_{2m}}} (R^{\sigma}_{\mathrm{emp}}[f] - R'^{\sigma}_{\mathrm{emp}}[f]) > \epsilon/2\}$ as

$$C_\epsilon = \bigcup_{n=1}^{\mathcal{N}(\mathcal{F},Z_{2m})} C_\epsilon(f_n), \tag{5.31}$$

where the events $C_\epsilon(f_n) := \{\sigma | (R^{\sigma}_{\mathrm{emp}}[f_n] - R'^{\sigma}_{\mathrm{emp}}[f_n]) > \epsilon/2\}$ refer to individual functions f_n chosen such that $\left(\bigcup_n \{f_n\}\right)|_{Z_{2m}} = \mathcal{F}|_{Z_{2m}}$. Note that the functions f_n may be considered as fixed, since we have conditioned on Z_{2m}.

We are now in a position to appeal to the classical law of large numbers. Our random experiment consists of drawing σ from the uniform distribution over all permutations of $2m$-samples. This turns our sequence of losses $\xi^{\sigma}_i = \frac{1}{2}|f(x^{\sigma}_i) - y^{\sigma}_i|$ $(i = 1, \ldots, 2m)$ into an iid sequence of independent Bernoulli trials. We then apply a modified Chernoff inequality to bound the probability of each event $C_\epsilon(f_n)$. It states that given a $2m$-sample of Bernoulli trials, we have (see Problem 5.4)

$$\mathrm{P}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi_i - \frac{1}{m}\sum_{i=m+1}^{2m}\xi_i \geq \epsilon\right\} \leq 2\,\exp\left(-\frac{m\epsilon^2}{2}\right). \tag{5.32}$$

For our present problem, we thus obtain

$$\mathrm{P}_{\sigma|Z_{2m}}(C_\epsilon(f_n)) \leq 2\,\exp\left(-\frac{m\epsilon^2}{8}\right), \tag{5.33}$$

independent of f_n. We next use the union bound to get a bound on the probability of the event C_ϵ defined in (5.31). We obtain a sum over $\mathcal{N}(\mathcal{F}, Z_{2m})$ identical terms of the form (5.33). Hence (5.30) (and (5.29)) can be bounded from above by

$$\begin{aligned}&\int_{(\mathcal{X}\times\{\pm1\})^{2m}} \mathcal{N}(\mathcal{F}, Z_{2m})\, 2\,\exp\left(-\frac{m\epsilon^2}{8}\right) d\mathrm{P}(Z_{2m})\\ &= 2\,\mathbf{E}\left[\mathcal{N}(\mathcal{F}, Z_{2m})\right]\exp\left(-\frac{m\epsilon^2}{8}\right),\end{aligned} \tag{5.34}$$

where the expectation is taken over the random drawing of Z_{2m}. The last step is to combine this with Lemma 5.4, to obtain

$$\begin{aligned} \mathrm{P}\{\sup_{f\in\mathcal{F}}(R[f]-R_{\mathrm{emp}}[f])>\epsilon\} &\leq 4\,\mathbf{E}\left[\mathcal{N}(\mathcal{F},Z_{2m})\right]\,\exp\left(-\frac{m\epsilon^2}{8}\right) \\ &= 4\,\exp\left(\ln\mathbf{E}\left[\mathcal{N}(\mathcal{F},Z_{2m})\right]-\frac{m\epsilon^2}{8}\right). \end{aligned} \tag{5.35}$$

Inequality of Vapnik-Chervonenkis Type

We conclude that provided $\mathbf{E}\left[\mathcal{N}(\mathcal{F},Z_{2m})\right]$ does not grow exponentially in m (i.e., $\ln\mathbf{E}\left[\mathcal{N}(\mathcal{F},Z_{2m})\right]$ grows sublinearly), it is actually possible to make nontrivial statements about the *test* error of learning machines.

The above reasoning is essentially the VC style analysis. Similar bounds can be obtained using a strategy which is more common in the field of empirical processes, first proving that $\sup_f(R[f]-R_{\mathrm{emp}}[f])$ is concentrated around its mean [554, 14].

5.5.5 Confidence Intervals

It is sometimes useful to rewrite (5.35) such that we specify the probability with which we want the bound to hold, and then get the confidence interval, which tells us how close the risk should be to the empirical risk. This can be achieved by setting the right hand side of (5.35) equal to some $\delta>0$, and then solving for ϵ. As a result, we get the statement that with a probability at least $1-\delta$,

Risk Bound

$$R[f]\leq R_{\mathrm{emp}}[f]+\sqrt{\frac{8}{m}\left(\ln\mathbf{E}\left[\mathcal{N}(\mathcal{F},Z_{2m})\right]+\ln\frac{4}{\delta}\right)}. \tag{5.36}$$

Note that this bound holds independent of f; in particular, it holds for the function f^m minimizing the empirical risk. This is not only a strength, but also a weakness in the bound. It is a strength since many learning machines do not truly minimize the empirical risk, and the bound thus holds for them, too. It is a weakness since by taking into account more information on which function we are interested in, one could hope to get more accurate bounds. We will return to this issue in Section 12.1.

Bounds like (5.36) can be used to justify induction principles different from the empirical risk minimization principle. Vapnik and Chervonenkis [569, 559] proposed minimizing the right hand side of these bounds, rather than just the empirical risk. The confidence term, in the present case, $\sqrt{\frac{8}{m}\left(\ln\mathbf{E}\left[\mathcal{N}(\mathcal{F},Z_{2m})\right]+\ln\frac{4}{\delta}\right)}$, then ensures that the chosen function, denoted f_*, not only leads to a small risk, but also comes from a function class with small capacity.

The capacity term is a property of the function class $\mathcal{F}$, and not of any individual function f. Thus, the bound cannot simply be minimized over choices of f. Instead, we introduce a so-called *structure* on $\mathcal{F}$, and minimize over the choice of the structure. This leads to an induction principle called *structural risk minimization*. We leave out the technicalities involved [559, 136, 562]. The main idea is depicted in Figure 5.3.

Structural Risk Minimization

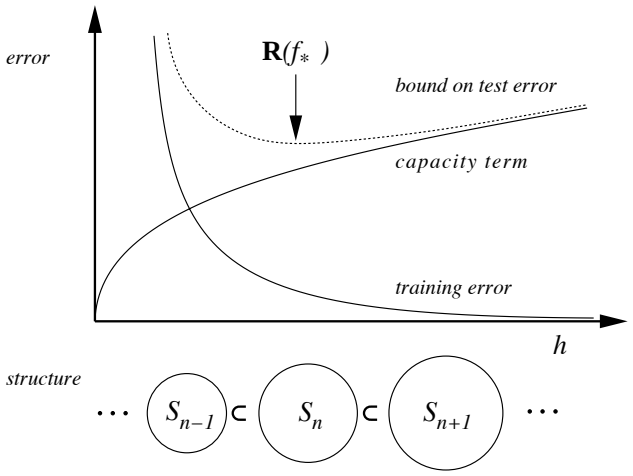

Figure 5.3 Graphical depiction of the structural risk minimization (SRM) induction principle. The function class is decomposed into a nested sequence of subsets of increasing size (and thus, of increasing capacity). The SRM principle picks a function f_* which has small training error, and comes from an element of the structure that has low capacity h, thus minimizing a risk bound of type (5.36).

For practical purposes, we usually employ bounds of the type (5.36) as a guideline for coming up with risk functionals (see Section 4.1). Often, the risk functionals form a compromise between quantities that *should* be minimized from a statistical point of view, and quantities that *can* be minimized efficiently (cf. Problem 5.7).

There exists a large number of bounds similar to (5.35) and its alternative form (5.36). Differences occur in the constants, both in front of the exponential and in its exponent. The bounds also differ in the exponent of ϵ — in some cases, by a factor greater than 2. For instance, if a training error of zero is achievable, we can use Bernstein's inequality instead of Chernoff's result, which leads to ϵ rather than ϵ^2. For further details, cf. [136, 562, 492, 238]. Finally, the bounds differ in the way they measure capacity. So far, we have used covering numbers, but this is not the only method.

5.5.6 The VC Dimension and Other Capacity Concepts

So far, we have formulated the bounds in terms of the so-called *annealed entropy* $\ln \mathbf{E}[\mathcal{N}(\mathcal{F}, Z_{2m})]$. This led to statements that depend on the distribution and thus can take into account characteristics of the problem at hand. The downside is that they are usually difficult to evaluate; moreover, in most problems, we do not have knowledge of the underlying distribution. However, a number of different capacity concepts, with different properties, can take the role of the term $\ln(\mathbf{E}[\mathcal{N}(\mathcal{F}, Z_{2m})])$ in (5.36).

- Given an example (x, y), $f \in \mathcal{F}$ causes a loss that we denote by $c(x, y, f(x)) := \frac{1}{2}|f(x) - y| \in \{0, 1\}$. For a larger sample $(x_1, y_1) \ldots, (x_m, y_m)$, the different functions

VC Entropy

$f \in \mathcal{F}$ lead to a *set* of loss vectors $\boldsymbol{\xi}_f = (c(x_1, y_1, f(x_1)), \ldots, c(x_m, y_m, f(x_m)))$, whose cardinality we denote by $\mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)$. The *VC entropy* is defined as

$$H_{\mathcal{F}}(m) = \mathbf{E}\left[\ln \mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)\right], \tag{5.37}$$

where the expectation is taken over the random generation of the m-sample $(x_1, y_1) \ldots, (x_m, y_m)$ from P.
One can show [562] that the convergence

$$\lim_{m\to\infty} H_{\mathcal{F}}(m)/m = 0, \tag{5.38}$$

is equivalent to uniform (two-sided) convergence of risk,

$$\lim_{m\to\infty} \mathrm{P}\{\sup_{f\in\mathcal{F}} |R[f] - R_{\mathrm{emp}}[f]| > \epsilon\} = 0, \tag{5.39}$$

for all $\epsilon > 0$. By Theorem 5.3, (5.39) thus implies consistency of empirical risk minimization.

Annealed Entropy

- If we exchange the expectation **E** and the logarithm in (5.37), we obtain the annealed entropy used above,

$$H^{\mathrm{ann}}_{\mathcal{F}}(m) = \ln \mathbf{E}\left[\mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right)\right]. \tag{5.40}$$

Since the logarithm is a concave function, the annealed entropy is an upper bound on the VC entropy. Therefore, whenever the annealed entropy satisfies a condition of the form (5.38), the same automatically holds for the VC entropy.
One can show that the convergence

$$\lim_{m\to\infty} H^{\mathrm{ann}}_{\mathcal{F}}(m)/m = 0, \tag{5.41}$$

implies exponentially fast convergence [561],

$$P\{\sup_{f\in\mathcal{F}} |R[f] - R_{\mathrm{emp}}[f]| > \epsilon\} \leq 4\, \exp(((H^{\mathrm{ann}}_{\mathcal{F}}(2m)/m) - \epsilon^2) \cdot m). \tag{5.42}$$

It has recently been proven that in fact (5.41) is not only sufficient, but also necessary for this [66].

Growth Function

- We can obtain an upper bound on both entropies introduced so far, by taking a supremum over all possible samples, instead of the expectation. This leads to the *growth function*,

$$G_{\mathcal{F}}(m) = \max_{(x_1,y_1),\ldots,(x_m,y_m)\in\mathcal{X}\times\{\pm 1\}} \ln \mathcal{N}\left(\mathcal{F}, (x_1, y_1) \ldots, (x_m, y_m)\right). \tag{5.43}$$

Note that by definition, the growth function is the logarithm of the shattering coefficient, $G_{\mathcal{F}}(m) = \ln \mathcal{N}(\mathcal{F}, m)$.
The convergence

$$\lim_{m\to\infty} G_{\mathcal{F}}(m)/m = 0, \tag{5.44}$$

is necessary and sufficient for exponentially fast convergence of risk *for all underlying distributions* P.

■ The next step will be to summarize the main behavior of the growth function with a single number. If $\mathcal{F}$ is as rich as possible, so that for any sample of size m, the points can be chosen such that by using functions of the learning machine, they can be separated in all 2^m possible ways (i.e., they can be shattered), then

$$G_{\mathcal{F}}(m) = m \cdot \ln(2). \tag{5.45}$$

In this case, the convergence (5.44) does not take place, and learning will not generally be successful. What about the other case? Vapnik and Chervonenkis [567, 568] showed that either (5.45) holds true for all m, or there exists some *maximal* m for which (5.45) is satisfied. This number is called the *VC dimension* and is denoted by h. If the maximum does not exist, the VC dimension is said to be infinite.

VC Dimension

By construction, the VC dimension is thus the maximal number of points which can be shattered by functions in $\mathcal{F}$. It is possible to prove that for $m > h$ [568],

$$G_{\mathcal{F}}(m) \leq h \left(\ln \frac{m}{h} + 1 \right). \tag{5.46}$$

This means that up to $m = h$, the growth function increases linearly with the sample size. Thereafter, it only increases logarithmically, i.e., *much* more slowly. This is the regime where learning can succeed.

Although we do not make use of it in the present chapter, it is worthwhile to also introduce the *VC dimension of a class of real-valued functions* $\{f_{\mathbf{w}} | \mathbf{w} \in \Lambda\}$ at this stage. It is defined to equal the VC dimension of the class of indicator functions

$$\left\{ \operatorname{sgn}(f_{\mathbf{w}} - \beta) | \mathbf{w} \in \Lambda, \beta \in \left(\inf_x f_{\mathbf{w}}(x), \sup_x f_{\mathbf{w}}(x) \right) \right\}. \tag{5.47}$$

VC Dimension for Real-Valued Functions

In summary, we get a succession of capacity concepts,

$$H_{\mathcal{F}}(m) \leq H_{\mathcal{F}}^{\text{ann}}(m) \leq G_{\mathcal{F}}(m) \leq h \left(\ln \frac{m}{h} + 1 \right). \tag{5.48}$$

From left to right, these become less precise. The entropies on the left are distribution-dependent, but rather difficult to evaluate (see, e.g., [430, 391]). The growth function and VC dimension are distribution-independent. This is less accurate, and does not always capture the essence of a given problem, which might have a much more benign distribution than the worst case; on the other hand, we want the learning machine to work for all distributions. If we knew the distribution beforehand, then we would not need a learning machine anymore.

VC Dimension Example

Let us look at a simple example of the VC dimension. As a function class, we consider hyperplanes in $\mathbb{R}^2$, i.e.,

$$f(x) = \operatorname{sgn}(a + b[x]_1 + c[x]_2), \quad \text{with parameters } a, b, c \in \mathbb{R}. \tag{5.49}$$

Suppose we are given three points x_1, x_2, x_3 which are not collinear. No matter how they are labelled (that is, independent of our choice of $y_1, y_2, y_3 \in \{\pm 1\}$), we can always find parameters $a, b, c \in \mathbb{R}$ such that $f(x_i) = y_i$ for all i (see Figure 1.4 in the introduction). In other words, there exist three points that we can shatter. This

shows that the VC dimension of the set of hyperplanes in $\mathbb{R}^2$ satisfies $h \geq 3$. On the other hand, we can never shatter *four* points. It follows from simple geometry that given any four points, there is always a set of labels such that we cannot realize the corresponding classification. Therefore, the VC dimension is $h = 3$. More generally, for hyperplanes in $\mathbb{R}^N$, the VC dimension can be shown to be $h = N + 1$. For a formal derivation of this result, as well as of other examples, see [523].

VC Dimension of Hyperplanes

How does this fit together with the fact that SVMs can be shown to correspond to hyperplanes in feature spaces of possibly infinite dimension? The crucial point is that SVMs correspond to *large margin* hyperplanes. Once the margin enters, the capacity can be much smaller than the above general VC dimension of hyperplanes. For simplicity, we consider the case of hyperplanes containing the origin.

VC Dimension of Margin Hyperplanes

Theorem 5.5 (Vapnik [559]) *Consider hyperplanes* $\langle \mathbf{w}, \mathbf{x} \rangle = 0$, *where* $\mathbf{w}$ *is normalized such that they are in canonical form w.r.t. a set of points* $X^* = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$; *i.e.,*

$$\min_{i=1,\ldots,r} | \langle \mathbf{w}, \mathbf{x}_i \rangle | = 1. \tag{5.50}$$

The set of decision functions $f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn} \langle \mathbf{x}, \mathbf{w} \rangle$ *defined on* X^*, *and satisfying the constraint* $\|\mathbf{w}\| \leq \Lambda$, *has a VC dimension satisfying*

$$h \leq R^2 \Lambda^2. \tag{5.51}$$

Here, R is the radius of the smallest sphere centered at the origin and containing X^*.

Before we give a proof, several remarks are in order.

- The theorem states that we can control the VC dimension *irrespective of the dimension of the space* by controlling the length of the weight vector $\|\mathbf{w}\|$. Note, however, that this needs to be done a priori, by choosing a value for Λ. It therefore does not strictly motivate what we will later see in SVMs, where $\|\mathbf{w}\|$ is minimized in order to control the capacity. Detailed treatments can be found in the work of Shawe-Taylor et al. [491, 24, 125].
- There exists a similar result for the case where R is the radius of the smallest sphere (not necessarily centered at the origin) enclosing the data, and where we allow for the possibility that the hyperplanes have a nonzero offset b [562]. In this case, we give a simple visualization in figure Figure 5.4, which shows it is plausible that enforcing a large margin amounts to reducing the VC dimension.
- Note that the theorem talks about functions defined on X^*. To extend it to the case where the functions are defined on all of the input domain $\mathcal{X}$, it is best to state it for the *fat shattering dimension*. For details, see [24].

The proof [24, 222, 559] is somewhat technical, and can be skipped if desired.

Proof Let us assume that $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are shattered by canonical hyperplanes with $\|\mathbf{w}\| \leq \Lambda$. Consequently, for all $y_1, \ldots, y_r \in \{\pm 1\}$, there exists a $\mathbf{w}$ with $\|\mathbf{w}\| \leq \Lambda$, such that

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \quad \text{for all } i = 1, \ldots, r. \tag{5.52}$$

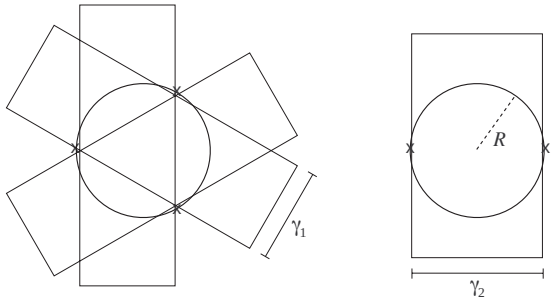

Figure 5.4 Simple visualization of the fact that enforcing a large margin of separation amounts to limiting the VC dimension. Assume that the data points are contained in a ball of radius R (cf. Theorem 5.5). Using hyperplanes with margin γ_1, it is possible to separate three points in all possible ways. Using hyperplanes with the larger margin γ_2, this is only possible for *two* points, hence the VC dimension in that case is two rather than three.

The proof proceeds in two steps. In the first part, we prove that the more points we want to shatter (5.52), the larger $\| \sum_{i=1}^r y_i \mathbf{x}_i \|$ must be. In the second part, we prove that we can upper bound the size of $\| \sum_{i=1}^r y_i \mathbf{x}_i \|$ in terms of R. Combining the two gives the desired condition, which tells us the maximum number of points we can shatter.

Summing (5.52) over $i = 1, \ldots, r$ yields

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^r y_i \mathbf{x}_i \right) \right\rangle \geq r. \tag{5.53}$$

By the Cauchy-Schwarz inequality, on the other hand, we have

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^r y_i \mathbf{x}_i \right) \right\rangle \leq \|\mathbf{w}\| \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\| \leq \Lambda \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\|. \tag{5.54}$$

Here, the second inequality follows from $\|\mathbf{w}\| \leq \Lambda$.

Combining (5.53) and (5.54), we get the desired lower bound,

$$\frac{r}{\Lambda} \leq \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\|. \tag{5.55}$$

We now move on to the second part. Let us consider independent random labels $y_i \in \{\pm 1\}$ which are uniformly distributed, sometimes called *Rademacher variables*. Let $\mathbf{E}$ denote the expectation over the choice of the labels. Exploiting the linearity of $\mathbf{E}$, we have

$$\begin{aligned} \mathbf{E} \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\|^2 &= \sum_{i=1}^r \mathbf{E} \left\langle y_i \mathbf{x}_i, \sum_{j=1}^r y_j \mathbf{x}_j \right\rangle \\ &= \sum_{i=1}^r \mathbf{E} \left\langle y_i \mathbf{x}_i, \left(\left(\sum_{j \neq i} y_j \mathbf{x}_j \right) + y_i \mathbf{x}_i \right) \right\rangle \end{aligned}$$

$$= \sum_{i=1}^{r} \left(\left(\sum_{j \neq i} \mathbf{E} \left\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \right\rangle \right) + \mathbf{E} \left\langle y_i \mathbf{x}_i, y_i \mathbf{x}_i \right\rangle \right)$$

$$= \sum_{i=1}^{r} \mathbf{E} \|y_i \mathbf{x}_i\|^2, \tag{5.56}$$

where the last equality follows from the fact that the Rademacher variables have zero mean and are independent. Exploiting the fact that $\|y_i \mathbf{x}_i\| = \|\mathbf{x}_i\| \leq R$, we get

$$\mathbf{E} \left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\|^2 \leq rR^2. \tag{5.57}$$

Since this is true for the expectation over the random choice of the labels, there must be at least one set of labels for which it also holds true. We have so far made no restrictions on the labels, hence we may now use this specific set of labels. This leads to the desired upper bound,

$$\left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\|^2 \leq rR^2. \tag{5.58}$$

Combining the upper bound with the lower bound (5.55), we get

$$\frac{r^2}{\Lambda^2} \leq rR^2; \tag{5.59}$$

hence,

$$r \leq R^2 \Lambda^2. \tag{5.60}$$

In other words, if the r points are shattered by a canonical hyperplane satisfying the assumptions we have made, then r is constrained by (5.60). The VC dimension h also satisfies (5.60), since it corresponds to the maximum number of points that can be shattered. ■

In the next section, we give an application of this theorem. Readers only interested in the theoretical background of learning theory may want to skip this section.

5.6 A Model Selection Example

In the following example, taken from [470], we use a bound of the form (5.36) to predict which kernel would perform best on a character recognition problem (USPS set, see Section A.1). Since the problem is essentially separable, we disregard the empirical risk term in the bound, and choose the parameters of a polynomial kernel by minimizing the second term. Note that the second term is a monotonic function of the capacity. As a capacity measure, we use the upper bound on the VC dimension described in Theorem 5.5, which in turn is an upper bound on the logarithm of the covering number that appears in (5.36) (by the arguments put forward in Section 5.5.6).