■ 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,

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

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 unknown 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