we would be very unlucky for this to occur *precisely* for the function f chosen by empirical risk minimization.

At first sight, it seems that empirical risk minimization should work — in contradiction to our lengthy explanation in the last section, arguing that we have to do more than that. What is the catch?

5.3 When Does Learning Work: the Question of Consistency

It turns out that in the last section, we were too sloppy. When we find a function f by choosing it to minimize the training error, we are no longer looking at independent Bernoulli trials. We are actually choosing f such that the mean of the ξ_i is as small as possible. In this sense, we are actively looking for the worst case, for a function which is very atypical, with respect to the average loss (i.e., the empirical risk) that it will produce.

We should thus state more clearly what it is that we actually need for empirical risk minimization to work. This is best expressed in terms of a notion that statisticians call *consistency*. It amounts to saying that as the number of examples m tends to infinity, we want the function f^m that minimizes $R_{\text{emp}}[f]$ (note that f^m need not be unique), to lead to a test error which converges to the lowest achievable value. In other words, f^m is asymptotically as good as whatever we could have done if we were able to directly minimize $R[f]$ (which we cannot, as we do not even know it). In addition, consistency requires that asymptotically, the training and the test error of f^m be identical.[3]

Consistency

It turns out that *without restricting the set of admissible functions*, empirical risk minimization is not consistent. The main insight of VC (Vapnik-Chervonenkis) theory is that actually, the *worst case* over all functions that the learning machine can implement determines the consistency of empirical risk minimization. In other words, we need a version of the law of large numbers which is *uniform* over all functions that the learning machine can implement.

5.4 Uniform Convergence and Consistency

The present section will explain how consistency can be characterized by a uniform convergence condition on the set of functions $\mathcal{F}$ that the learning machine can implement. Figure 5.2 gives a simplified depiction of the question of consistency. Both the empirical risk and the actual risk are drawn as functions of f. For

3. We refrain from giving a more formal definition of consistency, the reason being that there are some caveats to this classical definition of consistency; these would necessitate a discussion leading us away from the main thread of the argument. For the precise definition of the required notion of "nontrivial consistency," see [561].

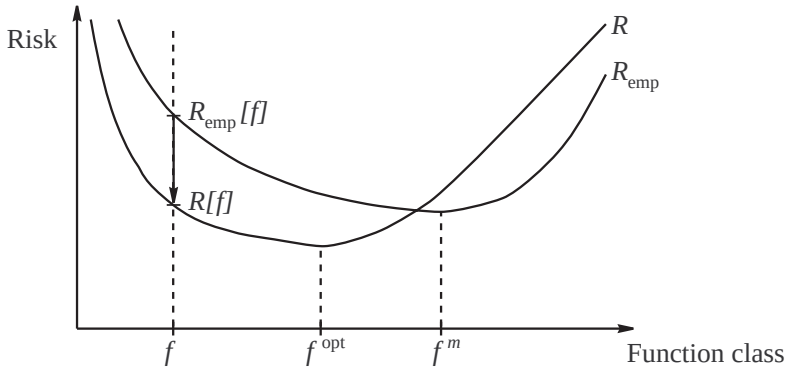

Figure 5.2 Simplified depiction of the convergence of empirical risk to actual risk. The x-axis gives a one-dimensional representation of the function class; the y axis denotes the risk (error). For each *fixed* function f, the law of large numbers tells us that as the sample size goes to infinity, the empirical risk $R_{\text{emp}}[f]$ converges towards the true risk $R[f]$ (indicated by the downward arrow). This does not imply, however, that in the limit of infinite sample sizes, the minimizer of the empirical risk, f^m, will lead to a value of the risk that is as good as the best attainable risk, $R[f^{\text{opt}}]$ *(consistency)*. For the latter to be true, we require convergence of $R_{\text{emp}}[f]$ towards $R[f]$ to be uniform over all functions that the learning machines can implement (see text).

simplicity, we have summarized all possible functions f by a single axis of the plot. Empirical risk minimization consists in picking the f that yields the minimal value of R_{emp}. If it is consistent, then the minimum of R_{emp} converges to that of R in probability. Let us denote the minimizer of R by f^{opt}, satisfying

$$R[f] - R[f^{\text{opt}}] \geq 0 \tag{5.12}$$

for all $f \in \mathcal{F}$. This is the optimal choice that we could make, given complete knowledge of the distribution P.[4] Similarly, since f^m minimizes the empirical risk, we have

$$R_{\text{emp}}[f] - R_{\text{emp}}[f^m] \geq 0, \tag{5.13}$$

for all $f \in \mathcal{F}$. Being true for all $f \in \mathcal{F}$, (5.12) and (5.13) hold in particular for f^m and f^{opt}. If we substitute the former into (5.12) and the latter into (5.13), we obtain

$$R[f^m] - R[f^{\text{opt}}] \geq 0, \tag{5.14}$$

and

$$R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \geq 0. \tag{5.15}$$

4. As with f^m, f^{opt} need not be unique.

The sum of these two inequalities satisfies

$$\begin{aligned} 0 &\leq R[f^m] - R[f^{\text{opt}}] + R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \\ &= R[f^m] - R_{\text{emp}}[f^m] + R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}] \\ &\leq \sup_{f \in \mathcal{F}} \big(R[f] - R_{\text{emp}}[f]\big) + R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}]. \end{aligned} \tag{5.16}$$

Let us first consider the second half of the right hand side. Due to the law of large numbers, we have convergence in probability, i.e., for all $\epsilon > 0$,

$$|R_{\text{emp}}[f^{\text{opt}}] - R[f^{\text{opt}}]| \xrightarrow{\text{P}} 0 \text{ as } m \to \infty. \tag{5.17}$$

This holds true since f^{opt} is a fixed function, which is independent of the training sample (see (5.11)).

The important conclusion is that if the empirical risk converges to the actual risk one-sided *uniformly*, over all functions that the learning machine can implement,

Uniform Convergence of Risk

$$\sup_{f \in \mathcal{F}} (R[f] - R_{\text{emp}}[f]) \xrightarrow{\text{P}} 0 \;\text{ as }\; m \to \infty, \tag{5.18}$$

then the left hand sides of (5.14) and (5.15) will likewise converge to 0;

$$R[f^m] - R[f^{\text{opt}}] \xrightarrow{\text{P}} 0, \tag{5.19}$$

$$R_{\text{emp}}[f^{\text{opt}}] - R_{\text{emp}}[f^m] \xrightarrow{\text{P}} 0. \tag{5.20}$$

As argued above, (5.17) is not always true for f^m, since f^m is chosen to minimize R_{emp}, and thus depends on the sample. Assuming that (5.18) holds true, however, then (5.19) and (5.20) imply that in the limit, $R[f^m]$ cannot be larger than $R_{\text{emp}}[f^m]$. One-sided uniform convergence on $\mathcal{F}$ is thus a sufficient condition for consistency of the empirical risk minimization over $\mathcal{F}$.[5]

What about the other way round? Is one-sided uniform convergence also a *necessary* condition? Part of the mathematical beauty of VC theory lies in the fact that this is the case. We cannot go into the necessary details to prove this [571, 561, 562], and only state the main result. Note that this theorem uses the notion of nontrivial consistency that we already mentioned briefly in footnote 3. In a nutshell, this concept requires that the induction principle be consistent even after the "best" functions have been removed. Nontrivial consistency thus rules out, for instance, the case in which the problem is trivial, due to the existence of a function which uniformly does better than all other functions. To understand this, assume that there exists such a function. Since this function is uniformly better than all others, we can already select this function (using ERM) from *one* (arbitrary) data point. Hence the method would be trivially consistent, no matter what the

5. Note that the onesidedness of the convergence comes from the fact that we only require consistency of empirical risk *minimization*. If we required the same for empirical risk *maximization*, then we would end up with standard uniform convergence, and the parentheses in (5.18) would be replaced with modulus signs.

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