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].