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

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

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

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

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 elements 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