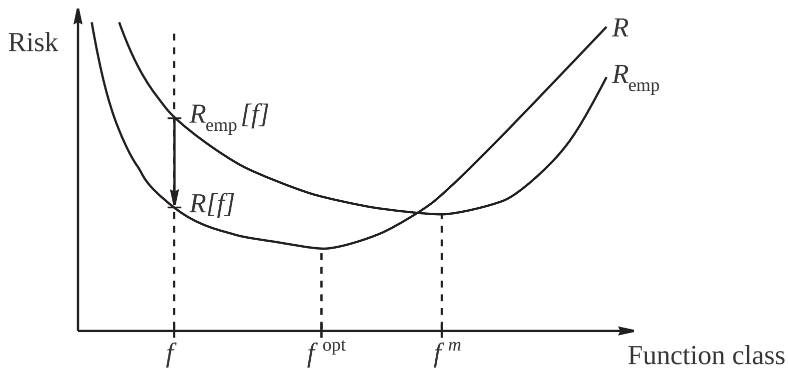

Figure 5.2 Simplified depiction of the convergence of empirical risk to actual risk. The x-axis gives a one-dimensional representation of the function class; the y axis denotes the risk (error). For each *fixed* function f, the law of large numbers tells us that as the sample size goes to infinity, the 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 the 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.