2 Empirical Risk Minimization

For a class of classification rules, $\mathcal{F}$, consider the approach of finding a rule f_n^* that minimizes the empirical risk $R_n(f)$. As the empirical error is a discrete non-convex function of f, actual computation necessary to get the best rule over $\mathcal{F}$ could be complex. However, we set aside this computational issue for a moment and focus on the relationship between the empirical error and the true probability of error.

In this process of determining a classification rule based on the data, three different kinds of *errors* come up. For a rule f_n in general (including the empirically best rule f_n^*),

(i) $R_n(f_n) - R(f_n)$ is due to the estimation of the risk R from the data by R_n.

(ii) $R(f_n) - \inf_{f\in\mathcal{F}} R(f)$: **estimation error** is the excess error of f_n relative to the best rule within the class $\mathcal{F}$. It measures how close f_n is to the best possible rule in $\mathcal{F}$ measured in terms of the theoretical risk R.

(iii) $R(f_n) - R^*$: **Bayes regret** is the excess error of f_n relative to the Bayes decision rule.

The Bayes regret is decomposed as follows:

$$R(f_n) - R^* = [R(f_n) - \inf_{f\in\mathcal{F}} R(f)] + [\inf_{f\in\mathcal{F}} R(f) - R^*].$$

The second term, $\inf_{f\in\mathcal{F}} R(f) - R^*$, is called the **approximation error**, which measures how well the functions or classifiers in $\mathcal{F}$ approximate f^*, the Bayes rule. Note that the approximation error does not depend on the data and it is a property of $\mathcal{F}$. There is a trade-off between the estimation error and the approximation error. When the class $\mathcal{F}$ is large, $\inf_{f\in\mathcal{F}} R(f)$ may be close to R^*, but the estimation error may be large as well. On the other hand, if $\mathcal{F}$ is too small, then there would be an irreducible gap between $\inf_{f\in\mathcal{F}} R(f)$ and R^*, so the approximation error can be substantial.

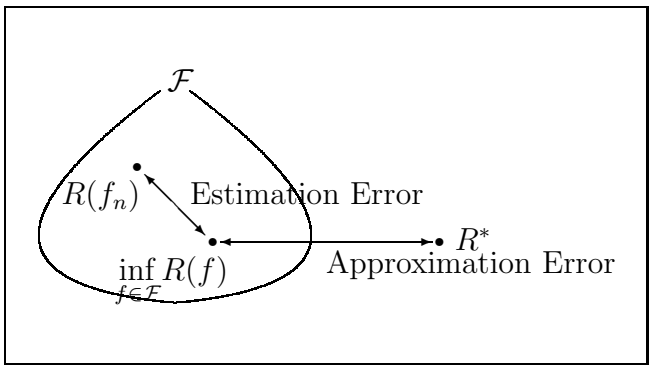

Figure 1: Schematic diagram of risks

Differently from the approximation error, the first two types of error (or the difference of the risk) depend on data and can be reduced by increasing the sample size. How much difference do we expect between the empirical error and the true error or how much estimation error do we expect given a sample size n and the class $\mathcal{F}$? As both of the errors involve random quantities ($R_n(f_n)$ and $R(f_n)$), statistical learning theory mainly aims at their probabilistic bounds. That is, how to bound the tail probabilities of the differences in risk. The following inequalities suggest that the two types of error for the best rule f_n^* can be dealt with as one by examining $\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$.

Theorem 3 (**Fundamental Inequalities**)**.**

(i) $|R_n(f_n^*) - R(f_n^*)| \leq \sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$.

(ii) $R(f_n^*) - \inf_{f\in\mathcal{F}} R(f) \leq 2 \sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$.

Proof. (i) is trivially true because $f_n^* \in \mathcal{F}$ by our choice. For (ii) we prove as follows.

$$
\begin{aligned}
R(f_n^*) - \inf_{f\in\mathcal{F}} R(f) &= R(f_n^*) - R_n(f_n^*) + R_n(f_n^*) - \inf_{f\in\mathcal{F}} R(f) \\
&\leq \sup_{f\in\mathcal{F}} |R_n(f) - R(f)| + R_n(f_n^*) - \inf_{f\in\mathcal{F}} R(f) \quad \text{[by (i)]} \\
&= \sup_{f\in\mathcal{F}} |R_n(f) - R(f)| + \sup_{f\in\mathcal{F}} (R_n(f_n^*) - R(f)) \\
&\leq \sup_{f\in\mathcal{F}} |R_n(f) - R(f)| + \sup_{f\in\mathcal{F}} (R_n(f) - R(f)) \quad \text{[by the definition of } f_n^*\text{]} \\
&\leq \sup_{f\in\mathcal{F}} |R_n(f) - R(f)| + \sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \\
&= 2 \sup_{f\in\mathcal{F}} |R_n(f) - R(f)|.
\end{aligned}
$$

□

Hence, upper bounds for $\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$ provide upper bounds for

(i) $R(f_n^*) - \inf_{f\in\mathcal{F}} R(f)$, the suboptimality of f_n^* in $\mathcal{F}$ and

(ii) $|R_n(f_n^*) - R(f_n^*)|$, the error made in estimating the true risk $R(f_n^*)$ by the empirical risk $R_n(f_n^*)$.

So, an important question is how to find bounds for $P\{\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \geq \epsilon\}$.

For a fixed f, $R_n(f) - R(f) = \frac{1}{n}\sum L(f(X_i), Y_i) - E\, L(f(X), Y)$. The convergence $R_n(f)$ to $R(f)$ is guaranteed by the law of large numbers. For finite sample analysis, some probability inequalities associated with the law of large numbers will be discussed. They are called *concentration inequalities*. They tell us how rapidly $R_n(f)$ converges to $R(f)$. Note that in our case we are interested in the supremum over all $f \in \mathcal{F}$. Therefore, such

probability bounds regard uniform convergence of empirical risk to the true risk. We will learn some standard techniques useful for proving uniform convergence as well. By studying this uniform deviation $\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$, we can establish a probabilistic bound for the estimation error of the form

$$P\left(R(f_n^*) - \inf_{f\in\mathcal{F}} R(f) > \epsilon\right) \leq \delta.$$

Here δ is a distribution free upper bound of the probability of the tail event. Of course, δ depends on ϵ, n and the size (richness) of $\mathcal{F}$. From the inequality, we have

$$P\left(R(f_n^*) - \inf_{f\in\mathcal{F}} R(f) \leq \epsilon\right) \geq 1 - \delta;$$

that is, with confidence of at least $1 - \delta$,

$$R(f_n^*) \leq \inf_{f\in\mathcal{F}} R(f) + \epsilon.$$

Thereby we can determine how much data we need to guarantee that the risk of the empirically optimal rule is within ϵ-bound of the minimum risk of decision rules in $\mathcal{F}$ with confidence $1 - \delta$ without any distributional assumption on (X, Y). This formalism of learning is called *Probably Approximately Correct* (PAC) learning (Valiant, 1984) in the literature.