4 Model Selection

A key problem in the design of learning algorithms is the choice of the hypothesis set $\mathcal{H}$. This is known as the *model selection* problem. How should the hypothesis set $\mathcal{H}$ be chosen? A rich or complex enough hypothesis set could contain the ideal Bayes classifier. On the other hand, learning with such a complex family becomes a very difficult task. More generally, the choice of $\mathcal{H}$ is subject to a trade-off that can be analyzed in terms of the *estimation* and *approximation errors*.

Our discussion will focus on the particular case of binary classification but much of what is discussed can be straightforwardly extended to different tasks and loss functions.

4.1 Estimation and approximation errors

Let $\mathcal{H}$ be a family of functions mapping $\mathcal{X}$ to $\{-1, +1\}$. The *excess error* of a hypothesis h chosen from $\mathcal{H}$, that is the difference between its error $R(h)$ and the Bayes error R^*, can be decomposed as follows:

$$R(h) - R^* = \underbrace{\left(R(h) - \inf_{h \in \mathcal{H}} R(h)\right)}_{\text{estimation}} + \underbrace{\left(\inf_{h \in \mathcal{H}} R(h) - R^*\right)}_{\text{approximation}}. \tag{4.1}$$

The first term is called the *estimation error*, the second term the *approximation error*. The estimation error depends on the hypothesis h selected. It measures the error of h with respect to the infimum of the errors achieved by hypotheses in $\mathcal{H}$, or that of the best-in-class hypothesis h^* when that infimum is reached. Note that the definition of agnostic PAC-learning is precisely based on the estimation error.

The approximation error measures how well the Bayes error can be approximated using $\mathcal{H}$. It is a property of the hypothesis set $\mathcal{H}$, a measure of its richness. For a more complex or richer hypothesis $\mathcal{H}$, the approximation error tends to be smaller at the price of a larger estimation error. This is illustrated by Figure 4.1.

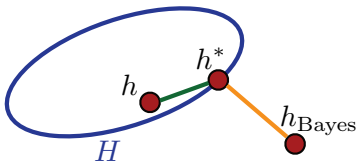

Figure 4.1
Illustration of the estimation error (in green) and approximation error (in orange). Here, it is assumed that there exists a best-in-class hypothesis, that is h^* such that $R(h^*) = \inf_{h \in \mathcal{H}} R(h)$.

Model selection consists of choosing $\mathcal{H}$ with a favorable trade-off between the approximation and estimation errors. Note, however, that the approximation error is not accessible, since in general the underlying distribution $\mathcal{D}$ needed to determine R^* is not known. Even with various noise assumptions, estimating the approximation error is difficult. In contrast, the *estimation error of an algorithm* $\mathcal{A}$, that is, the estimation error of the hypothesis h_S returned after training on a sample S, can sometimes be bounded using generalization bounds as shown in the next section.

4.2 Empirical risk minimization (ERM)

A standard algorithm for which the estimation error can be bounded is *Empirical Risk Minimization* (ERM). ERM seeks to minimize the error on the training sample:[4]

$$h_S^{\text{ERM}} = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}_S(h). \tag{4.2}$$

Proposition 4.1 *For any sample S, the following inequality holds for the hypothesis returned by ERM:*

$$\mathbb{P}\left[R(h_S^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) > \epsilon\right] \leq \mathbb{P}\left[\sup_{h \in \mathcal{H}} |R(h) - \widehat{R}_S(h)| > \frac{\epsilon}{2}\right]. \tag{4.3}$$

Proof: By definition of $\inf_{h \in \mathcal{H}} R(h)$, for any $\epsilon > 0$, there exists h_ϵ such that $R(h_\epsilon) \leq \inf_{h \in \mathcal{H}} R(h) + \epsilon$. Thus, using $\widehat{R}_S(h_S^{\text{ERM}}) \leq \widehat{R}_S(h_\epsilon)$, which holds by the

[4] Note that, if there exists multiple hypotheses with minimal error on the training sample, then ERM returns an arbitrary one.

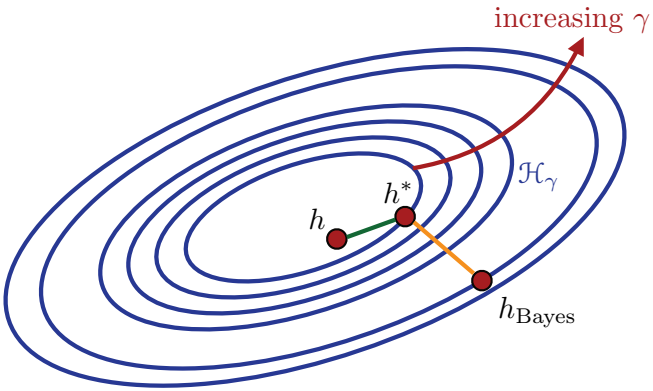

Figure 4.2
Illustration of the decomposition of a rich family $\mathcal{H} = \bigcup_{\gamma \in \Gamma} \mathcal{H}_\gamma$.

definition of the algorithm, we can write

$$\begin{aligned} R(h_S^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) &= R(h_S^{\text{ERM}}) - R(h_\epsilon) + R(h_\epsilon) - \inf_{h \in \mathcal{H}} R(h) \\ &\leq R(h_S^{\text{ERM}}) - R(h_\epsilon) + \epsilon \\ &= R(h_S^{\text{ERM}}) - \widehat{R}_S(h_S^{\text{ERM}}) + \widehat{R}_S(h_S^{\text{ERM}}) - R(h_\epsilon) + \epsilon \\ &\leq R(h_S^{\text{ERM}}) - \widehat{R}_S(h_S^{\text{ERM}}) + \widehat{R}_S(h_\epsilon) - R(h_\epsilon) + \epsilon \\ &\leq 2 \sup_{h \in \mathcal{H}} |R(h) - \widehat{R}_S(h)| + \epsilon. \end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, it implies the following:

$$R(h_S^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) \leq 2 \sup_{h \in \mathcal{H}} |R(h) - \widehat{R}_S(h)|,$$

which concludes the proof. □

The right-hand side of (4.3) can be upper-bounded using the generalization bounds presented in the previous chapter in terms of the Rademacher complexity, the growth function, or the VC-dimension of $\mathcal{H}$. In particular, it can be bounded by $2e^{-2m[\epsilon - \mathfrak{R}_m(\mathcal{H})]^2}$. Thus, when $\mathcal{H}$ admits a favorable Rademacher complexity, for example a finite VC-dimension, for a sufficiently large sample, with high probability, the estimation error is guaranteed to be small. Nevertheless, the performance of ERM is typically very poor. This is because the algorithm disregards the complexity of the hypothesis set $\mathcal{H}$: in practice, either $\mathcal{H}$ is not complex enough, in which case the approximation error can be very large, or $\mathcal{H}$ is very rich, in which case the bound on the estimation error becomes very loose. Additionally, in many cases, determining the ERM solution is computationally intractable. For example, finding

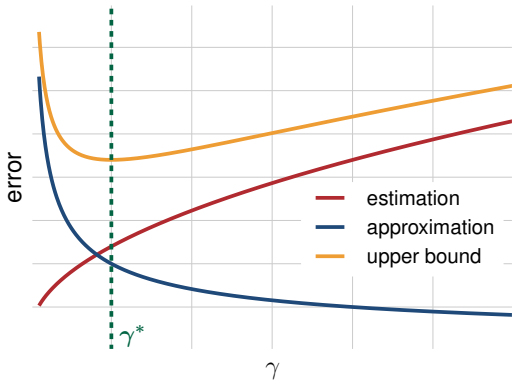

Figure 4.3
Choice of γ^* with the most favorable trade-off between estimation and approximation errors.

a linear hypothesis with the smallest error on the training sample is NP-hard, as a function of the dimension of the space.

4.3 Structural risk minimization (SRM)

In the previous section, we showed that the estimation error can be sometimes bounded or estimated. But, since the approximation error cannot be estimated, how should we choose $\mathcal{H}$? One way to proceed is to choose a very complex family $\mathcal{H}$ with no approximation error or a very small one. $\mathcal{H}$ may be too rich for generalization bounds to hold for $\mathcal{H}$, but suppose we can decompose $\mathcal{H}$ as a union of increasingly complex hypothesis sets $\mathcal{H}_\gamma$, that is $\mathcal{H} = \bigcup_{\gamma \in \Gamma} \mathcal{H}_\gamma$, with the complexity of $\mathcal{H}_\gamma$ increasing with γ, for some set Γ. Figure 4.2 illustrates this decomposition. The problem then consists of selecting the parameter $\gamma^* \in \Gamma$ and thus the hypothesis set $\mathcal{H}_{\gamma^*}$ with the most favorable trade-off between estimation and approximation errors. Since these quantities are not known, instead, as illustrated by Figure 4.3, a uniform upper bound on their sum, the excess error (also called excess risk), can be used.

This is precisely the idea behind the *Structural Risk Minimization* (SRM) method. For SRM, $\mathcal{H}$ is assumed to be decomposable into a countable set, thus, we will write its decomposition as $\mathcal{H} = \bigcup_{k \geq 1} \mathcal{H}_k$. Additionally, the hypothesis sets $\mathcal{H}_k$ are assumed to be nested: $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ for all $k \geq 1$. However, many of the results presented in this section also hold for non-nested hypothesis sets. Thus, we will not make use of that assumption, unless explicitly specified. SRM consists of choosing the index $k^* \geq 1$ and the ERM hypothesis h in $\mathcal{H}_{k^*}$ that minimize an upper bound on the excess error.

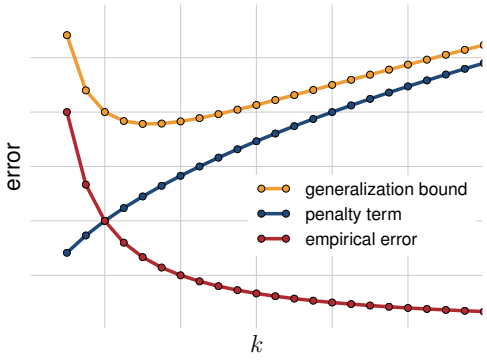

Figure 4.4
Illustration of structural risk minimization. The plots of three errors are shown as a function of the index k. Clearly, as k, or equivalently the complexity the hypothesis set $\mathcal{H}_k$, increases, the training error decreases, while the penalty term increases. SRM selects the hypothesis minimizing a bound on the generalization error, which is a sum of the empirical error and the penalty term.

As we shall see, the following learning bound holds for all $h \in \mathcal{H}$: for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of a sample S of size m from $\mathcal{D}^m$, for all $h \in \mathcal{H}_k$ and $k \geq 1$,

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Thus, to minimize the resulting bound on the excess error $(R(h) - R^*)$, the index k and the hypothesis $h \in \mathcal{H}_k$ should be chosen to minimize the following objective function:

$$F_k(h) = \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}.$$

This is precisely the definition of the SRM solution h_S^{SRM}:

$$h_S^{\mathrm{SRM}} = \operatorname*{argmin}_{k \geq 1, h \in \mathcal{H}_k} F_k(h) = \operatorname*{argmin}_{k \geq 1, h \in \mathcal{H}_k} \widehat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}. \tag{4.4}$$

Thus, SRM identifies an optimal index k^* and therefore hypothesis set $\mathcal{H}_{k^*}$, and returns the ERM solution based on that hypothesis set. Figure 4.4 further illustrates the selection of the index k^* and hypothesis set $\mathcal{H}_{k^*}$ by SRM by minimizing an upper bound on the sum of the training error and the penalty term $R_m(\mathcal{H}_k) + \sqrt{\log k/m}$. The following theorem shows that the SRM solution benefits from a strong learning guarantee. For any $h \in \mathcal{H}$, we will denote by $\mathcal{H}_{k(h)}$ the least complex hypothesis set among the $\mathcal{H}_k$s that contain h.

Theorem 4.2 (SRM Learning guarantee) *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{D}^m$, the generalization error of the hypothesis h_S^{SRM} returned by the SRM method is bounded as follows:*

$$R(h_S^{\mathrm{SRM}}) \leq \inf_{h \in \mathcal{H}} \left(R(h) + 2\mathfrak{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log k(h)}{m}} \right) + \sqrt{\frac{2 \log \frac{3}{\delta}}{m}}.$$

Proof: Observe first that, by the union bound, the following general inequality holds:

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h \in \mathcal{H}} R(h) - F_{k(h)}(h) > \epsilon \right] \\
&\qquad = \mathbb{P}\left[\sup_{k \geq 1} \sup_{h \in \mathcal{H}_k} R(h) - F_k(h) > \epsilon \right] \\
&\qquad \leq \sum_{k=1}^{\infty} \mathbb{P}\left[\sup_{h \in \mathcal{H}_k} R(h) - F_k(h) > \epsilon \right] \\
&\qquad = \sum_{k=1}^{\infty} \mathbb{P}\left[\sup_{h \in \mathcal{H}_k} R(h) - \widehat{R}_S(h) - \mathfrak{R}_m(\mathcal{H}_k) > \epsilon + \sqrt{\frac{\log k}{m}} \right] \\
&\qquad \leq \sum_{k=1}^{\infty} \exp\left(-2m \left[\epsilon + \sqrt{\tfrac{\log k}{m}} \right]^2 \right) \\
&\qquad \leq \sum_{k=1}^{\infty} e^{-2m\epsilon^2} e^{-2 \log k} \\
&\qquad = e^{-2m\epsilon^2} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} e^{-2m\epsilon^2} \leq 2e^{-2m\epsilon^2}.
\end{aligned} \tag{4.5}$$

Next, for any two random variables X_1 and X_2, if $X_1 + X_2 > \epsilon$, then either X_1 or X_2 must be larger than $\epsilon/2$. In view of that, by the union bound, $\mathbb{P}[X_1 + X_2 > \epsilon] \leq \mathbb{P}[X_1 > \frac{\epsilon}{2}] + \mathbb{P}[X_2 > \frac{\epsilon}{2}]$. Using this inequality, inequality (4.5), and the inequality $F_{k(h_S^{\mathrm{SRM}})}(h_S^{\mathrm{SRM}}) \leq F_{k(h)}(h)$, which holds for all $h \in \mathcal{H}$, by definition of h_S^{SRM}, we

can write, for any $h \in \mathcal{H}$,

$$\begin{aligned}
&\mathbb{P}\left[R(h_S^{\mathrm{SRM}}) - R(h) - 2\mathfrak{R}_m(\mathcal{H}_{k(h)}) - \sqrt{\tfrac{\log k(h)}{m}} > \epsilon\right] \\
&\quad \leq \mathbb{P}\left[R(h_S^{\mathrm{SRM}}) - F_{k(h_S^{\mathrm{SRM}})}(h_S^{\mathrm{SRM}}) > \frac{\epsilon}{2}\right] \\
&\qquad + \mathbb{P}\left[F_{k(h_S^{\mathrm{SRM}})}(h_S^{\mathrm{SRM}}) - R(h) - 2\mathfrak{R}_m(\mathcal{H}_{k(h)}) - \sqrt{\tfrac{\log k(h)}{m}} > \frac{\epsilon}{2}\right] \\
&\quad \leq 2e^{-\frac{m\epsilon^2}{2}} + \mathbb{P}\left[F_{k(h)}(h) - R(h) - 2\mathfrak{R}_m(\mathcal{H}_{k(h)}) - \sqrt{\tfrac{\log k(h)}{m}} > \frac{\epsilon}{2}\right] \\
&\quad = 2e^{-\frac{m\epsilon^2}{2}} + \mathbb{P}\left[\widehat{R}_S(h) - R(h) - \mathfrak{R}_m(\mathcal{H}_{k(h)}) > \frac{\epsilon}{2}\right] \\
&\quad = 2e^{-\frac{m\epsilon^2}{2}} + e^{-\frac{m\epsilon^2}{2}} = 3e^{-\frac{m\epsilon^2}{2}}.
\end{aligned}$$

Setting the right-hand side to be equal to δ completes the proof. □

The learning guarantee just proven for SRM is remarkable. To simplify its discussion, let us assume that there exists h^* such that $R(h^*) = \inf_{h \in \mathcal{H}} R(h)$, that is, that there exists a best-in-class classifier $h^* \in \mathcal{H}$. Then, the theorem implies in particular that, with probability at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$:

$$R(h_S^{\mathrm{SRM}}) \leq R(h^*) + 2\mathfrak{R}_m(\mathcal{H}_{k(h^*)}) + \sqrt{\frac{\log k(h^*)}{m}} + \sqrt{\frac{2\log\frac{3}{\delta}}{m}}. \tag{4.6}$$

Observe that, remarkably, this bound is similar to the estimation error bound for $\mathcal{H}_{k(h^*)}$: it differs from it only by the term $\sqrt{\log k(h^*)/m}$. Thus, modulo that term, the guarantee for SRM is as favorable as the one we would have obtained, had an oracle informed us of the index $k(h^*)$ of the best-in-class classifier's hypothesis set.

Furthermore, observe that when $\mathcal{H}$ is rich enough that $R(h^*)$ is close to the Bayes error, the learning bound (4.6) is approximately a bound on the excess error of the SRM solution. Note that, if for some k_0, the empirical error of the ERM solution for $\mathcal{H}_{k_0}$ is zero, which holds in particular if $\mathcal{H}_{k_0}$ contains the Bayes error, then, we have $\min_{h \in \mathcal{H}_k} F_{k_0}(h) \leq \min_{h \in \mathcal{H}_k} F_k(h)$ for all $k > k_0$ and only finitely many indices need to be considered in SRM.

Assume more generally that if $\min_{h \in \mathcal{H}_k} F_k(h) \leq \min_{h \in \mathcal{H}_{k+1}} F_k(h)$ for some k, then indices beyond $k+1$ need not be inspected. This property may hold for example if the empirical error cannot be further improved after some index k. In that case, the minimizing index k^* can be determined via a binary search in the interval $[1, k_{\max}]$, given some maximum value $k_{\max}$. $k_{\max}$ itself can be found by inspecting $\min_{h \in \mathcal{H}_{2^n}} F_k(h)$ for exponentially growing indices 2^n, $n \geq 1$, and setting $k_{\max} = 2^n$ for n such that $\min_{h \in \mathcal{H}_{2^n}} F_k(h) \leq \min_{h \in \mathcal{H}_{2^{n+1}}} F_k(h)$. The number of ERM computations needed to find $k_{\max}$ is in $O(n) = O(\log k_{\max})$ and similarly the

number of ERM computations due to the binary search is in $O(\log k_{\max})$. Thus, if n is the smallest integer such that $k^* < 2^n$, the overall number of ERM computations is in $O(\log k^*)$.

While it benefits from a very favorable guarantee, SRM admits several drawbacks. First, the decomposability of $\mathcal{H}$ into countably many hypothesis sets, each with a converging Rademacher complexity, remains a strong assumption. As an example, the family of all measurable functions cannot be written as a union of countably many hypothesis sets with finite VC-dimension. Thus, the choice of $\mathcal{H}$ or that of the hypothesis sets $\mathcal{H}_k$ is a key component of SRM. Second, and this is the main disadvantage of SRM, the method is typically computationally intractable: for most hypothesis sets, finding the solution of ERM is NP-hard and in general SRM requires determining that solution for a large number of indices k.

4.4 Cross-validation

An alternative method for model selection, *cross-validation*, consists of using some fraction of the training sample as a *validation set* to select a hypothesis set $\mathcal{H}_k$. This is in contrast with the SRM model which relies on a theoretical learning bound assigning a penalty to each hypothesis set. In this section, we analyze the cross-validation method and compare its performance to that of SRM.

As in the previous section, let $(\mathcal{H}_k)_{k\geq 1}$ be a countable sequence of hypothesis sets with increasing complexities. The cross-validation (CV) solution is obtained as follows. Let S be an i.i.d. labeled sample of size m. S is divided into a sample S_1 of size $(1-\alpha)m$ and a sample S_2 of size αm, with $\alpha \in (0,1)$ typically chosen to be relatively small. S_1 is reserved for training, S_2 for validation. For any $k \in \mathbb{N}$, let $h^{\mathrm{ERM}}_{S_1,k}$ denote the solution of ERM run on S_1 using the hypothesis set $\mathcal{H}_k$. The hypothesis h^{CV}_S returned by cross-validation is the ERM solution $h^{\mathrm{ERM}}_{S_1,k}$ with the best performance on S_2:

$$h^{\mathrm{CV}}_S = \operatorname*{argmin}_{h \in \left\{h^{\mathrm{ERM}}_{S_1,k} \colon k \geq 1\right\}} \widehat{R}_{S_2}(h). \tag{4.7}$$

The following general result will help us derive learning guarantees for cross-validation.

Proposition 4.3 *For any $\alpha > 0$ and any sample size $m \geq 1$, the following general inequality holds:*

$$\mathbb{P}\left[\sup_{k\geq 1}\left|R(h^{\mathrm{ERM}}_{S_1,k}) - \widehat{R}_{S_2}(h^{\mathrm{ERM}}_{S_1,k})\right| > \epsilon + \sqrt{\frac{\log k}{\alpha m}}\right] \leq 4e^{-2\alpha m\epsilon^2}.$$

Proof: By the union bound, we can write

$$\begin{aligned}\mathbb{P}\left[\sup_{k\geq 1}\left|R(h_{S_1,k}^{\mathrm{ERM}})-\widehat{R}_{S_2}(h_{S_1,k}^{\mathrm{ERM}})\right|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right]&\\ \leq\sum_{k=1}^{\infty}\mathbb{P}\left[\left|R(h_{S_1,k}^{\mathrm{ERM}})-\widehat{R}_{S_2}(h_{S_1,k}^{\mathrm{ERM}})\right|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right]&\\ =\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{P}\left[\left|R(h_{S_1,k}^{\mathrm{ERM}})-\widehat{R}_{S_2}(h_{S_1,k}^{\mathrm{ERM}})\right|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\;\middle|\;S_1\right]\right].&\end{aligned}\tag{4.8}$$

The hypothesis $h_{S_1,k}^{\mathrm{ERM}}$ is fixed conditioned on S_1. Furthermore, the sample S_2 is independent from S_1. Therefore, by Hoeffding's inequality, we can bound the conditional probability as follows:

$$\begin{aligned}\mathbb{P}\left[\left|R(h_{S_1,k}^{\mathrm{ERM}})-\widehat{R}_{S_2}(h_{S_1,k}^{\mathrm{ERM}})\right|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\;\middle|\;S_1\right]&\leq 2e^{-2\alpha m\left(\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right)^2}.\\ &\leq 2e^{-2\alpha m\epsilon^2-2\log k}\\ &=\frac{2}{k^2}e^{-2\alpha m\epsilon^2}.\end{aligned}$$

Plugging in the right-hand side of this bound in (4.8) and summing over k yields

$$\mathbb{P}\left[\sup_{k\geq 1}\left|R(h_{S_1,k}^{\mathrm{ERM}})-\widehat{R}_{S_2}(h_{S_1,k}^{\mathrm{ERM}})\right|>\epsilon+\sqrt{\frac{\log k}{\alpha m}}\right]\leq\frac{\pi^2}{3}e^{-2\alpha m\epsilon^2}<4e^{-2\alpha m\epsilon^2},$$

which completes the proof. □

Let $R(h_{S_1}^{\mathrm{SRM}})$ be the generalization error of the SRM solution using a sample S_1 of size $(1-\alpha m)$ and $R(h_S^{\mathrm{CV}},S)$ the generalization error of the cross-validation solution using a sample S of size m. Then, using Proposition 4.3, the following learning guarantee can be derived which compares the error of the CV method to that of SRM.

Theorem 4.4 (Cross-validation versus SRM) *For any $\delta>0$, with probability at least $1-\delta$, the following holds:*

$$R(h_S^{\mathrm{CV}})-R(h_{S_1}^{\mathrm{SRM}})\leq 2\sqrt{\frac{\log\max(k(h_S^{\mathrm{CV}}),k(h_{S_1}^{\mathrm{SRM}}))}{\alpha m}}+2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}},$$

where, for any h, $k(h)$ denotes the smallest index of a hypothesis set containing h.

Proof: By Proposition 4.3 and Theorem 4.2, using the property of h_S^{CV} as a minimizer, for any $\delta>0$, with probability at least $1-\delta$, the following inequalities

hold:

$$\begin{aligned}
R(h_S^{\mathrm{CV}}) &\leq \widehat{R}_{S_2}(h_S^{\mathrm{CV}}) + \sqrt{\frac{\log(k(h_S^{\mathrm{CV}}))}{\alpha m}} + \sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} \\
&\leq \widehat{R}_{S_2}(h_{S_1}^{\mathrm{SRM}}) + \sqrt{\frac{\log(k(h_S^{\mathrm{CV}}))}{\alpha m}} + \sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} \\
&\leq R(h_{S_1}^{\mathrm{SRM}}) + \sqrt{\frac{\log(k(h_S^{\mathrm{CV}}))}{\alpha m}} + \sqrt{\frac{\log(k(h_{S_1}^{\mathrm{SRM}}))}{\alpha m}} + 2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}} \\
&\leq R(h_{S_1}^{\mathrm{SRM}}) + 2\sqrt{\frac{\log(\max(k(h_S^{\mathrm{CV}}), k(h_{S_1}^{\mathrm{SRM}}))}{\alpha m}} + 2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}},
\end{aligned}$$

which completes the proof. □

The learning guarantee just proven shows that, with high probability, the generalization error of the CV solution for a sample of size m is close to that of the SRM solution for a sample of size $(1-\alpha)m$. For α relatively small, this suggests a guarantee similar to that of SRM, which, as previously discussed, is very favorable. However, in some unfavorable regimes, an algorithm (here SRM) trained on $(1-\alpha)m$ points may have a significantly worse performance than when trained on m points (avoiding this phase transition issue is one of the main motivations behind the use of the n-fold cross-validation method in practice, see section 4.5). Thus, the bound suggests in fact a trade-off: α should be chosen sufficiently small to avoid the unfavorable regimes just mentioned and yet sufficiently large for the right-hand side of the bound to be small and thus informative.

The learning bound for CV can be made more explicit in some cases in practice. Assume for example that the hypothesis sets $\mathcal{H}_k$ are nested and that the empirical errors of the ERM solutions $h_{S_1,k}^{\mathrm{ERM}}$ are decreasing before reaching zero: for any k, $\widehat{R}_{S_1}(h_{S_1,k+1}^{\mathrm{ERM}}) < \widehat{R}_{S_1}(h_{S_1,k}^{\mathrm{ERM}})$ for all k such that $\widehat{R}_{S_1}(h_{S_1,k}^{\mathrm{ERM}}) > 0$ and $\widehat{R}_{S_1}(h_{S_1,k+1}^{\mathrm{ERM}}) \leq \widehat{R}_{S_1}(h_{S_1,k}^{\mathrm{ERM}})$ otherwise. Observe that $\widehat{R}_{S_1}(h_{S_1,k}^{\mathrm{ERM}}) > 0$ implies at least one error for $h_{S_1,k}^{\mathrm{ERM}}$, therefore $\widehat{R}_{S_1}(h_{S_1,k}^{\mathrm{ERM}}) > \frac{1}{m}$. In view of that, we must then have $\widehat{R}_{S_1}(h_{S_1,n}^{\mathrm{ERM}}) = 0$ for all $n \geq m+1$. Thus, we have $h_{S_1,n}^{\mathrm{ERM}} = h_{S_1,m+1}^{\mathrm{ERM}}$ for all $n \geq m+1$ and we can assume that $k(f_{CV}) \leq m+1$. Since the complexity of $\mathcal{H}_k$ increases with k we also have $k(f_{SRM}) \leq m+1$. In view of that, we obtain the following more explicit learning bound for cross-validation:

$$R(f_{CV}, S) - R(f_{SRM}, S_1) \leq 2\sqrt{\frac{\log(\frac{4}{\delta})}{2\alpha m}} + 2\sqrt{\frac{\log(m+1)}{\alpha m}}.$$

4.5 n-Fold cross-validation

In practice, the amount of labeled data available is often too small to set aside a validation sample since that would leave an insufficient amount of training data. Instead, a widely adopted method known as *n-fold cross-validation* is used to exploit the labeled data both for *model selection* and for training.

Let $\boldsymbol{\theta}$ denote the vector of free parameters of the algorithm. For a fixed value of $\boldsymbol{\theta}$, the method consists of first randomly partitioning a given sample S of m labeled examples into n subsamples, or folds. The ith fold is thus a labeled sample $((x_{i1}, y_{i1}), \ldots, (x_{im_i}, y_{im_i}))$ of size m_i. Then, for any $i \in [n]$, the learning algorithm is trained on all but the ith fold to generate a hypothesis h_i, and the performance of h_i is tested on the ith fold, as illustrated in figure 4.5a. The parameter value $\boldsymbol{\theta}$ is evaluated based on the average error of the hypotheses h_i, which is called the *cross-validation error*. This quantity is denoted by $\widehat{R}_{\mathrm{CV}}(\boldsymbol{\theta})$ and defined by

$$\widehat{R}_{\mathrm{CV}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{m_i} \sum_{j=1}^{m_i} L(h_i(x_{ij}), y_{ij})}_{\text{error of } h_i \text{ on the } i\text{th fold}} .$$

The folds are generally chosen to have equal size, that is $m_i = m/n$ for all $i \in [n]$. How should n be chosen? The appropriate choice is subject to a trade-off. For a large n, each training sample used in n-fold cross-validation has size $m - m/n = m(1 - 1/n)$ (illustrated by the right vertical red line in figure 4.5b), which is close to m, the size of the full sample, and also implies all training samples are quite similar. At the same time, the ith fold used to measure the error is relatively small and thus the cross-validation error tends to have a small bias but a large variance. In contrast, smaller values of n lead to more diverse training samples but their size (shown by the left vertical red line in figure 4.5b) is significantly less than m. In this regime, the ith fold is relatively large and thus the cross-validation error tends to have a smaller variance but a larger bias.

In applications, n is typically chosen to be 5 or 10. n-fold cross-validation is used as follows in model selection. The full labeled data is first split into a training and a test sample. The training sample of size m is then used to compute the n-fold cross-validation error $\widehat{R}_{\mathrm{CV}}(\boldsymbol{\theta})$ for a small number of possible values of $\boldsymbol{\theta}$. The free parameter $\boldsymbol{\theta}$ is next set to the value $\boldsymbol{\theta}_0$ for which $\widehat{R}_{\mathrm{CV}}(\boldsymbol{\theta})$ is smallest and the algorithm is trained with the parameter setting $\boldsymbol{\theta}_0$ over the full training sample of size m. Its performance is evaluated on the test sample as already described in the previous section.

The special case of n-fold cross-validation where $n = m$ is called *leave-one-out cross-validation*, since at each iteration exactly one instance is left out of the train-

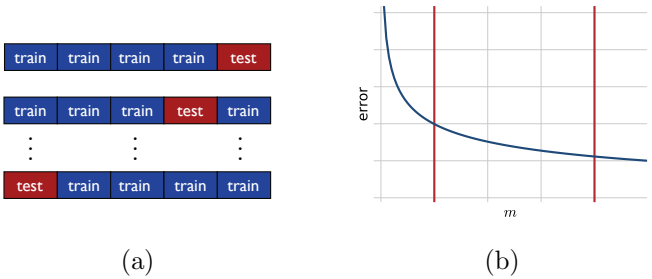

Figure 4.5
n-fold cross-validation. (a) Illustration of the partitioning of the training data into 5 folds. (b) Typical plot of a classifier's prediction error as a function of the size of the training sample m: the error decreases as a function of the number of training points. The red line on the left side marks the region for small values of n, while the red line on the right side marks the region for large values of n.

ing sample. As shown in chapter 5, the average leave-one-out error is an approximately unbiased estimate of the average error of an algorithm and can be used to derive simple guarantees for some algorithms. In general, the leave-one-out error is very costly to compute, since it requires training m times on samples of size $m-1$, but for some algorithms it admits a very efficient computation (see exercise 11.9).

In addition to model selection, n-fold cross-validation is also commonly used for performance evaluation. In that case, for a fixed parameter setting $\boldsymbol{\theta}$, the full labeled sample is divided into n random folds with no distinction between training and test samples. The performance reported is the n-fold cross-validation error on the full sample as well as the standard deviation of the errors measured on each fold.

4.6 Regularization-based algorithms

A broad family of algorithms inspired by the SRM method is that of *regularization-based algorithm*. This consists of selecting a very complex family $\mathcal{H}$ that is an uncountable union of nested hypothesis sets $\mathcal{H}_\gamma$: $\mathcal{H} = \bigcup_{\gamma>0} \mathcal{H}_\gamma$. $\mathcal{H}$ is often chosen to be dense in the space of continuous functions over $\mathcal{X}$. For example, $\mathcal{H}$ may be chosen to be the set of all linear functions in some high-dimensional space and $\mathcal{H}_\gamma$ the subset of those functions whose norm is bounded by γ: $\mathcal{H}_\gamma = \{x \mapsto \mathbf{w} \cdot \boldsymbol{\Phi}(x) \colon \|\mathbf{w}\| \leq \gamma\}$. For some choices of $\boldsymbol{\Phi}$ and the high-dimensional space, it can be shown that $\mathcal{H}$ is indeed dense in the space of continuous functions over $\mathcal{X}$.

Given a labeled sample S, the extension of the SRM method to an uncountable union would then suggest selecting h based on the following optimization problem:

$$\operatorname*{argmin}_{\gamma>0, h\in H_\gamma} \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_\gamma) + \sqrt{\frac{\log \gamma}{m}},$$

where other penalty terms $\operatorname{pen}(\gamma, m)$ can be chosen in lieu of the specific choice $\operatorname{pen}(\gamma, m) = \mathfrak{R}_m(\mathcal{H}_\gamma) + \sqrt{\frac{\log \gamma}{m}}$. Often, there exists a function $\mathcal{R}\colon \mathcal{H} \to \mathbb{R}$ such that, for any $\gamma > 0$, the constrained optimization problem $\operatorname{argmin}_{\gamma>0, h\in H_\gamma} \widehat{R}_S(h) + \operatorname{pen}(\gamma, m)$ can be equivalently written as the unconstrained optimization problem

$$\operatorname*{argmin}_{h\in\mathcal{H}} \widehat{R}_S(h) + \lambda \mathcal{R}(h),$$

for some $\lambda > 0$. $\mathcal{R}(h)$ is called a *regularization term* and $\lambda > 0$ is treated as a hyperparameter since its optimal value is often not known. For most algorithms, the regularization term $\mathcal{R}(h)$ is chosen to be an increasing function of $\|h\|$ for some choice of the norm $\|\cdot\|$, when $\mathcal{H}$ is the subset of a Hilbert space. The variable λ is often called a *regularization parameter*. Larger values of λ further penalize more complex hypotheses, while, for λ close or equal to zero, the regularization term has no effect and the algorithm coincides with ERM. In practice, λ is typically selected via cross-validation or using n-fold cross-validation.

When the regularization term is chosen to be $\|h\|_p$ for some choice of the norm and $p \geq 1$, then it is a convex function of h, since any norm is convex. However, for the zero-one loss, the first term of the objective function is non-convex, thereby making the optimization problem computationally hard. In practice, most regularization-based algorithms instead use a convex upper bound on the zero-one loss and replace the empirical zero-one term with the empirical value of that convex surrogate. The resulting optimization problem is then convex and therefore admits more efficient solutions than SRM. The next section studies the properties of such convex surrogate losses.

4.7 Convex surrogate losses

The guarantees for the estimation error that we presented in previous sections hold either for ERM or for SRM, which itself is defined in terms of ERM. However, as already mentioned, for many choices of the hypothesis set $\mathcal{H}$, including that of linear functions, solving the ERM optimization problem is NP-hard mainly because the zero-one loss function is not convex. One common method for addressing this problem consists of using a convex surrogate loss function that upper bounds the zero-one loss. This section analyzes learning guarantees for such surrogate losses in terms of the original loss.

The hypotheses we consider are real-valued functions $h\colon \mathcal{X} \to \mathbb{R}$. The sign of h defines a binary classifier $f_h\colon \mathcal{X} \to \{-1, +1\}$ defined for all $x \in \mathcal{X}$ by

$$f_h(x) = \begin{cases} +1 & \text{if } h(x) \geq 0 \\ -1 & \text{if } h(x) < 0. \end{cases}$$

The loss or error of h at point $(x, y) \in \mathcal{X} \times \{-1, +1\}$ is defined as the binary classification error of f_h:

$$1_{f_h(x) \neq y} = 1_{yh(x) < 0} + 1_{h(x)=0 \wedge y=-1} \leq 1_{yh(x) \leq 0}.$$

We will denote by $R(h)$ the expected error of h: $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[1_{f_h(x)\neq y}\right]$. For any $x \in \mathcal{X}$, let $\eta(x)$ denote $\eta(x) = \mathbb{P}[y = +1|x]$ and let $\mathcal{D}_{\mathcal{X}}$ denote the marginal distribution over $\mathcal{X}$. Then, for any h, we can write

$$\begin{aligned} R(h) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[1_{f_h(x)\neq y}\right] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)>0} + (1-\eta(x))1_{h(x)=0}\right] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)\geq 0}\right]. \end{aligned}$$

In view of that, the Bayes classifier can be defined as assigning label $+1$ to x when $\eta(x) \geq \frac{1}{2}$, -1 otherwise. It can therefore be induced by the function h^* defined by

$$h^*(x) = \eta(x) - \tfrac{1}{2}. \tag{4.9}$$

We will refer to $h^*\colon \mathcal{X} \to \mathbb{R}$ as the *Bayes scoring function* and will denote by R^* the error of the Bayes classifier or Bayes scoring function: $R^* = R(h^*)$.

Lemma 4.5 *The excess error of any hypothesis $h\colon \mathcal{X} \to \mathbb{R}$ can be expressed as follows in terms of η and the Bayes scoring function h^*:*

$$R(h) - R^* = 2 \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[|h^*(x)|\, 1_{h(x)h^*(x)\leq 0}\right].$$

Proof: For any h, we can write

$$\begin{aligned} R(h) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\eta(x)1_{h(x)<0} + (1-\eta(x))1_{h(x)\geq 0}\right] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\eta(x)1_{h(x)<0} + (1-\eta(x))(1-1_{h(x)<0})\right] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[[2\eta(x)-1]1_{h(x)<0} + (1-\eta(x))\right] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[2h^*(x)1_{h(x)<0} + (1-\eta(x))\right], \end{aligned}$$

where we used for the last step equation (4.9). In view of that, for any h, the following holds:

$$\begin{aligned} R(h) - R(h^*) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[2[h^*(x)](1_{h(x)\leq 0} - 1_{h^*(x)\leq 0})\Big] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[2[h^*(x)]\operatorname{sgn}(h^*(x))1_{(h(x)h^*(x)\leq 0)\wedge((h(x),h^*(x))\neq(0,0))}\Big] \\ &= 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[|h^*(x)|\,1_{h(x)h^*(x)\leq 0}\Big], \end{aligned}$$

which completes the proof, since $R(h^*) = R^*$. □

Let $\Phi\colon \mathbb{R}\to\mathbb{R}$ be a convex and non-decreasing function so that for any $u\in\mathbb{R}$, $1_{u\leq 0}\leq \Phi(-u)$. The Φ-loss of a function $h\colon \mathcal{X}\to\mathbb{R}$ at point $(x,y)\in\mathcal{X}\times\{-1,+1\}$ is defined as $\Phi(-yh(x))$ and its expected loss given by

$$\begin{aligned} \mathcal{L}_\Phi(h) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\big[\Phi(-yh(x))\big] \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\eta(x)\Phi(-h(x)) + (1-\eta(x))\Phi(h(x))\Big]. \end{aligned} \tag{4.10}$$

Notice that since $1_{yh(x)\leq 0}\leq\Phi(-yh(x))$, we have $R(h)\leq\mathcal{L}_\Phi(h)$. For any $x\in\mathcal{X}$, let $u\mapsto L_\Phi(x,u)$ be the function defined for all $u\in\mathbb{R}$ by

$$L_\Phi(x,u) = \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u).$$

Then, $\mathcal{L}_\Phi(h) = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[L_\Phi(x,h(x))]$. Since Φ is convex, $u\mapsto L_\Phi(x,u)$ is convex as a sum of two convex functions. Define $h^*_\Phi\colon\mathcal{X}\to[-\infty,+\infty]$ as the *Bayes solution for the loss function* L_Φ. That is, for any x, $h^*_\Phi(x)$ is a solution of the following convex optimization problem:

$$\begin{aligned} h^*_\Phi(x) &= \operatorname*{argmin}_{u\in[-\infty,+\infty]} L_\Phi(x,u) \\ &= \operatorname*{argmin}_{u\in[-\infty,+\infty]} \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u). \end{aligned}$$

The solution of this optimization is in general not unique. When $\eta(x) = 0$, $h^*_\Phi(x)$ is a minimizer of $u\mapsto\Phi(u)$ and since Φ is non-decreasing, we can choose $h^*_\Phi(x) = -\infty$ in that case. Similarly, when $\eta(x) = 1$, we can choose $h^*_\Phi(x) = +\infty$. When $\eta(x) = \frac{1}{2}$, $L_\Phi(x,u) = \frac{1}{2}[\Phi(-u)+\Phi(u)]$, thus, by convexity, $L_\Phi(x,u)\geq\Phi(-\frac{u}{2}+\frac{u}{2}) = \Phi(0)$. Thus, we can choose $h^*_\Phi(x) = 0$ in that case. For all other values of $\eta(x)$, in case of non-uniqueness, an arbitrary minimizer is chosen in this definition. We will denote by $\mathcal{L}^*_\Phi$ the Φ-loss of h^*_Φ: $\mathcal{L}^*_\Phi = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\Phi(-yh^*_\Phi(x))\big]$.

Proposition 4.6 *Let Φ be a convex and non-decreasing function that is differentiable at 0 with $\Phi'(0) > 0$. Then, the minimizer of Φ defines the Bayes classifier: for any $x\in\mathcal{X}$, $h^*_\Phi(x) > 0$ iff $h^*(x) > 0$ and $h^*(x) = 0$ iff $h^*_\Phi(x) = 0$, which implies $\mathcal{L}^*_\Phi = R^*$.*

Proof: Fix $x \in \mathcal{X}$. If $\eta(x) = 0$, then $h^*(x) = -\frac{1}{2}$ and $h^*_\Phi(x) = -\infty$, thus $h^*(x)$ and $h^*_\Phi(x)$ admit the same sign. Similarly, if $\eta(x) = 1$, then $h^*(x) = +\frac{1}{2}$ and $h^*_\Phi(x) = +\infty$, and $h^*(x)$ and $h^*_\Phi(x)$ admit the same sign.

Let u^* denote the minimizer defining $h^*_\Phi(x)$. u^* is a minimizer of $u \mapsto L_\Phi(x, u)$ iff the subdifferential of that function at u^* contains 0, that is, since $\partial L_\Phi(x, u^*) = -\eta(x)\partial\Phi(-u^*) + (1 - \eta(x))\partial\Phi(u^*)$, iff there exist $v_1 \in \partial\Phi(-u^*)$ and $v_2 \in \partial\Phi(u^*)$ such that

$$\eta(x)v_1 = (1 - \eta(x))v_2. \tag{4.11}$$

If $u^* = 0$, by the differentiability of Φ at 0 we have $v_1 = v_2 = \Phi'(0) > 0$ and thus $\eta(x) = \frac{1}{2}$, that is $h^*(x) = 0$. Conversely, If $h^*(x) = 0$, that is $\eta(x) = \frac{1}{2}$, then, by definition, we have $h^*_\Phi(x) = 0$. Thus, $h^*(x) = 0$ iff $h^*_\Phi(x) = 0$ iff $\eta(x) = \frac{1}{2}$.

We can assume now that $\eta(x)$ is not in $\{0, 1, \frac{1}{2}\}$. We first show that for any $u_1, u_2 \in \mathbb{R}$ with $u_1 < u_2$, and any two choices of the subgradients at u_1 and u_2, $v_1 \in \partial\Phi(u_1)$ and $v_2 \in \partial\Phi(u_2)$, we have $v_1 \leq v_2$. By definition of the subgradients at u_1 and u_2, the following inequalities hold:

$$\Phi(u_2) - \Phi(u_1) \geq v_1(u_2 - u_1) \qquad \Phi(u_1) - \Phi(u_2) \geq v_2(u_1 - u_2).$$

Summing up these inequalities yields $v_2(u_2 - u_1) \geq v_1(u_2 - u_1)$ and thus $v_2 \geq v_1$, since $u_1 < u_2$.

Now, if $u^* > 0$, then we have $-u^* < u^*$. By the property shown above, this implies $v_1 \leq v_2$. We cannot have $v_1 = v_2 \neq 0$ since (4.11) would then imply $\eta(x) = \frac{1}{2}$. We also cannot have $v_1 = v_2 = 0$ since by the property shown above, we must have $\Phi'(0) \leq v_2$ and thus $v_2 > 0$. Thus, we must have $v_1 < v_2$ with $v_2 > 0$, which, by (4.11), implies $\eta(x) > 1 - \eta(x)$, that is $h^*(x) > 0$.

Conversely, if $h^*(x) > 0$ then $\eta(x) > 1 - \eta(x)$. We cannot have $v_1 = v_2 = 0$ or $v_1 = v_2 \neq 0$ as already shown. Thus, since $\eta(x) \neq 1$, by (4.11), this implies $v_1 < v_2$. We cannot have $u^* < -u^*$ since, by the property shown above, this would imply $v_2 \leq v_1$. Thus, we must have $-u^* \leq u^*$, that is $u^* \geq 0$, and more specifically $u^* > 0$ since, as already shown above, $u^* = 0$ implies $h^*(x) = 0$. □

Theorem 4.7 *Let Φ be a convex and non-decreasing function. Assume that there exists $s \geq 1$ and $c > 0$ such that the following holds for all $x \in \mathcal{X}$:*

$$|h^*(x)|^s = \left|\eta(x) - \tfrac{1}{2}\right|^s \leq c^s \left[L_\Phi(x, 0) - L_\Phi(x, h^*_\Phi(x))\right].$$

Then, for any hypothesis h, the excess error of h is bounded as follows:

$$R(h) - R^* \leq 2c \left[\mathcal{L}_\Phi(h) - \mathcal{L}^*_\Phi\right]^{\frac{1}{s}}$$

Proof: We will use the following inequality which holds by the convexity of Φ:

$$\begin{aligned}\Phi\big(-2h^*(x)h(x)\big) &= \Phi\big((1-2\eta(x))h(x)\big)\\ &= \Phi\big(\eta(x)(-h(x)) + (1-\eta(x))h(x)\big)\\ &\leq \eta(x)\Phi((-h(x))) + (1-\eta(x))\Phi(h(x)) = L_\Phi(x, h(x)). \quad (4.12)\end{aligned}$$

By Lemma 4.5, Jensen's inequality, and $h^*(x) = \eta(x) - \frac{1}{2}$, we can write

$$\begin{aligned}&R(h) - R(h^*)\\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[|2\eta(x)-1|\,1_{h(x)h^*(x)\leq 0}\Big]\\ &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[|2\eta(x)-1|^s\,1_{h(x)h^*(x)\leq 0}\Big]^{\frac{1}{s}} && \text{(Jensen's ineq.)}\\ &\leq 2c \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\big[\Phi(0) - L_\Phi(x, h^*_\Phi(x))\big]\,1_{h(x)h^*(x)\leq 0}\Big]^{\frac{1}{s}} && \text{(assumption)}\\ &\leq 2c \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\big[\Phi\big(-2h^*(x)h(x)\big) - L_\Phi(x, h^*_\Phi(x))\big]\,1_{h(x)h^*(x)\leq 0}\Big]^{\frac{1}{s}} && (\Phi \text{ non-decreasing})\\ &\leq 2c \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[[L_\Phi(x, h(x)) - L_\Phi(x, h^*_\Phi(x))]\,1_{h(x)h^*(x)\leq 0}\Big]^{\frac{1}{s}} && \text{(convexity ineq. (4.12))}\\ &\leq 2c \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[L_\Phi(x, h(x)) - L_\Phi(x, h^*_\Phi(x))\Big]^{\frac{1}{s}},\end{aligned}$$

which completes the proof, since $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[L_\Phi(x, h^*_\Phi(x))] = L^*_\Phi$. □

The theorem shows that, when the assumption holds, the excess error of h can be upper bounded in terms of the excess Φ-loss. The assumption of the theorem holds in particular for the following convex loss functions:

- Hinge loss, where $\Phi(u) = \max(0, 1+u)$, with $s = 1$ and $c = \frac{1}{2}$.
- Exponential loss, where $\Phi(u) = \exp(u)$, with $s = 2$ and $c = \frac{1}{\sqrt{2}}$.
- Logistic loss, where $\Phi(u) = \log_2(1+e^u)$, with $s = 2$ and $c = \frac{1}{\sqrt{2}}$.

They also hold for the square loss and the squared Hinge loss (see Exercises 4.2 and 4.3).

4.8 Chapter notes

The structural risk minimization (SRM) technique is due to Vapnik [1998]. The original penalty term used by Vapnik [1998] is based on the VC-dimension of the hypothesis set. The version of SRM with Rademacher complexity-based penalties that we present here leads to finer data-dependent learning guarantees. Penalties based on alternative complexity measures can be used similarly leading to learning bounds in terms of the corresponding complexity measure [Bartlett et al., 2002a].

An alternative model selection theory of *Voted Risk Minimization* (VRM) has been recently developed by Cortes, Mohri, and Syed [2014] and other related publications [Kuznetsov et al., 2014, DeSalvo et al., 2015, Cortes et al., 2015].

Theorem 4.7 is due to Zhang [2003a]. The proof given here is somewhat different and simpler.

4.9 Exercises

4.1 For any hypothesis set $\mathcal{H}$, show that the following inequalities hold:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\widehat{R}_S\left(h_S^{\mathrm{ERM}}\right)\right] \leq \inf_{h\in\mathcal{H}} R(h) \leq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[R\left(h_S^{\mathrm{ERM}}\right)\right]. \tag{4.13}$$

4.2 Show that for the squared loss, $\Phi(u) = (1+u)^2$, the statement of Theorem 4.7 holds with $s = 2$ and $c = \frac{1}{2}$ and therefore that the excess error can be upper bounded as follows:

$$R(h) - R^* \leq \left[\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*\right]^{\frac{1}{2}}.$$

4.3 Show that for the squared Hinge loss, $\Phi(u) = \max(0, 1+u)^2$, the statement of Theorem 4.7 holds with $s = 2$ and $c = \frac{1}{2}$ and therefore that the excess error can be upper bounded as follows:

$$R(h) - R^* \leq \left[\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*\right]^{\frac{1}{2}}.$$

4.4 In this problem, the loss of $h\colon \mathcal{X} \to \mathbb{R}$ at point $(x, y) \in \mathcal{X} \times \{-1, +1\}$ is defined to be $1_{yh(x)\leq 0}$.

(a) Define the Bayes classifier and a Bayes scoring function h^* for this loss.

(b) Express the excess error of h in terms of h^* (counterpart of Lemma 4.5, for loss considered here).

(c) Give a counterpart of the result of Theorem 4.7 for this loss.

4.5 Same questions as in Exercise 4.5 with the loss of $h\colon \mathcal{X} \to \mathbb{R}$ at point $(x, y) \in \mathcal{X} \times \{-1, +1\}$ defined instead to be $1_{yh(x)<0}$.