

CS492(F) - Computational Learning Theory.
Chapter 4 Model Selection.

1. Motivation

- (1) So far we studied results that let us use sample-based estimates, in particular, empirical error \hat{R}_{SCh} , and express properties about what we are ultimately interested in, namely \rightarrow generalisation error R_{Ch} . The next natural question is how to exploit these results for designing an algorithm and analysing it. $\xrightarrow{\text{rich-enough}}$
- (2) A reasonable-looking answer is to pick a hypothesis set \mathcal{H} and to pick $h \in \mathcal{H}$ for a given sample $S \in (\mathbb{X} \times \mathbb{Y})^m$ that minimises the empirical risk \hat{R}_{SCh} . This approach is called empirical risk minimization (ERM). At first look you may feel that this is as good as we can get.
- (3) But it turns out that we can do better than ERM. In a sense. The idea is to impose a structure on \mathcal{H} by assuming a family of hypothesis sets $\{\mathcal{H}_k\}_{k \in \mathbb{N}}$ with $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$ and to perform search at two levels, one at the level of a hypothesis set \mathcal{H}_k and the other at the level of hypothesis h . The search for a good hypothesis set \mathcal{H}_k is called model selection, the topic of this chapter. (\mathcal{H}_k is called a model). By the way, this two-level search can be done one after the other (as in m -folds cross validation) or simultaneously (as in structural risk minimization).

- (4) What should we do in order to do model selection well? We should define a good success measure of model, i.e., hypothesis set \mathcal{H}_k , which also permits an efficient approximate optimiser. In this chapter, we will study such measures and derive idealised learning algorithms, for instance, that search for a good model \hat{f}_{lk} using those measures and for a good hypothesis in the model \mathcal{H}_k .
- (5) By the way, I used "idealised" above to mean that derived algorithms might not be implementable because, for instance, they require search over infinitely many entities and their search objectives cannot be computed. However, these algorithms serve as a basis for real implementable algorithms which typically perform further approximation. As you expect, idealised algorithms are easier to analyse mathematically than real implementable algorithms.

2 Analysis of error.

- (1) Let $D \in \Pr(\mathcal{X} \times \mathcal{Y})$ with $\mathcal{Y} = \{-1, +1\}$.

Recall that the best error we can achieve is

$$R^* = \mathbb{E}_{x \sim D_x} [\min(\Pr[y=1|x], \Pr[y=-1|x])]$$

which is called Bayes error.

- (2) for a hypothesis set $\mathcal{H} \subseteq \{\hat{f} \rightarrow \mathcal{Y}\}$ and $h \in \mathcal{H}$, the excess error of h is $R(h) - R^*$.

It can be decomposed into two parts as follows:

$$R(\hat{h}) - R^* = R(\hat{h}) - \inf_{\hat{h} \in \mathcal{H}_{\text{eff}}} R(\hat{h}') + \inf_{\hat{h}' \in \mathcal{H}_{\text{eff}}} R(\hat{h}') - R^*$$

called estimation error. called approximation error.

(3) Note that if \mathcal{H} is rich, the approximation error is small but the estimation error is large. The opposite situation arises when \mathcal{H} is too simple.

(4) The structural risk minimisation (SRM) uses a success measure on hypo. sets \mathcal{H}_K (or models) that balances est. error and approx. error., as we will see shortly.

3. Empirical risk minimization

(1) Before studying algorithms w/o model selection, let's look at the most basic learning algorithm, empirical risk minimization.

(2) ERM algorithm for a hypo. set \mathcal{H} .

Input $S \in (\mathcal{X} \times \mathcal{Y})^m$

*we assume that
the minimum
exists.*

Output $h_S^{\text{ERM}} = \arg \min_{h \in \mathcal{H}} \widehat{R}_S(h)$.

(3) This algorithm has the following guarantee on estimation error.

Thm. $\mathbb{P} [\inf_{h \in \mathcal{H}} R(h_S^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon] \leq 2 \exp \left(-2m \left(\frac{\varepsilon}{2} - R_m(\mathcal{H}) \right)^2 \right)$

Equivalently, for all $S \geq 0$,

$$\mathbb{P} [\inf_{h \in \mathcal{H}} R(h_S^{\text{ERM}}) \leq \inf_{h \in \mathcal{H}} R(h) + 2R_m(\mathcal{H}) + 2\sqrt{\frac{\log 2/\delta}{2m}}] \geq 1 - \delta.$$

why? Because of Prop. 1 and the Rademacher-complexity bound.

Prop 4.1

$$P \left[R(h_{\text{S}}^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right]$$

$$\leq P \left[\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_{\text{S}}(h)| > \frac{\varepsilon}{2} \right]$$

Prop. 1

Proof of thm.

$$P \left[R(h_{\text{S}}^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right]$$

$$\leq P \left[\exists h \in \mathcal{H}, R(h) > \hat{R}_{\text{S}}(h) + \frac{\varepsilon}{2} \right] \text{ or.}$$

$$\text{union-bound} \quad R(h) < \hat{R}_{\text{S}}(h) - \frac{\varepsilon}{2} \quad].$$

$$\leq P \left[\exists h \in \mathcal{H}, R(h) > \hat{R}_{\text{S}}(h) + \frac{\varepsilon}{2} \right] +$$

$$P \left[\exists h \in \mathcal{H}, R(h) < \hat{R}_{\text{S}}(h) - \frac{\varepsilon}{2} \right].$$

$$\leq s + s' \quad \text{By Thm 3.5 and its dual lower-bound version.}$$

$$\text{for all } s, s' \geq \exp \left(-2m \left(\frac{\varepsilon}{2} - R_m(\mathcal{H}) \right)^2 \right) \quad (\text{which we didn't work out})$$

Here we use the fact that

$h_{\text{S}}^{\text{ERM}}$ is a minimizer.

□.

Proof of Prop 4.1

Pick $\varepsilon > 0$. Then, $\exists h_{\varepsilon} \in \mathcal{H}$ st. $R(h_{\varepsilon}) \leq \inf_{h \in \mathcal{H}} R(h) + \varepsilon$.

$$R(h_{\text{S}}^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) = R(h_{\text{S}}^{\text{ERM}}) - R(h_{\varepsilon}) + R(h_{\varepsilon}) - \inf_{h \in \mathcal{H}} R(h).$$

$$\leq R(h_{\text{S}}^{\text{ERM}}) - R(h_{\varepsilon}) + \varepsilon.$$

$$\leq R(h_{\text{S}}^{\text{ERM}}) - \hat{R}_{\text{S}}(h_{\text{S}}^{\text{ERM}}) + \hat{R}_{\text{S}}(h_{\varepsilon}) - R(h_{\varepsilon}) + \varepsilon.$$

($\because \hat{R}_{\text{S}}(h_{\text{S}}^{\text{ERM}}) \leq \hat{R}_{\text{S}}(h)$ for all $h \in \mathcal{H}$)

$$\leq 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_{\text{S}}(h)| + \varepsilon.$$

Since ε is arbitrary, $R(h_{\text{S}}^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) \leq 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_{\text{S}}(h)|$.

Thus, $P \left[R(h_{\text{S}}^{\text{ERM}}) - \inf_{h \in \mathcal{H}} R(h) > \varepsilon \right] \leq P \left[2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_{\text{S}}(h)| > \varepsilon \right]$. □.

(4) ERM performs poorly and h_S^{ERM} is difficult to compute.

Complex ff -- large estimation error indicated by the bound in the theorem that we just proved.

Simple ff -- large approximation error.

(5) The algorithms that we will look at next overcome this poor performance issue using model selection.

4. Structural risk minimization (SRM).

(1) We assume a countable family of hypo. sets $\{f_{fk}\}_{k \in \mathbb{N}}^{\text{st.}}$.
 $f_{lk} \subseteq f_{lk+1}$ for all $k \in \mathbb{N}$.

The SRM solves the following opt. pb:

$$h_S^{\text{SRM}} = \arg \min_{k \geq 1, h \in f_{lk}} F_k(h)$$

where

$$F_k(h) = \widehat{R}_S(h) + R_m(f_{lk}) + \sqrt{\frac{\log k}{m}}$$

(2) Note the term $R_m(f_{lk}) + \sqrt{\frac{\log k}{m}}$. If we didn't have this term, h_S^{SRM} would be the same as h_S^{ERM} over the hypo. set $\text{ff} = \bigcup_k f_{lk}$.

(3) Where does this term come from? It comes from the generalisation bound: for all $D \in \Pr(\mathcal{X} \times \mathcal{Y})$, $m \in \mathbb{N}$, $\delta > 0$,

$$\Pr [\forall k \in \mathbb{N}, \forall h \in f_{lk}, R(h) \leq \widehat{R}_S(h) + R_m(f_{lk}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}] \geq 1 - \delta.$$

which we will show soon.

(4) By using f_{lk} with large k , we can make $\hat{R}_{S(h)}$ small by choosing an appropriate $h \in f_{lk}$. But then we have to pay the price for this in the remaining term $R_m(f_{lk}) + \sqrt{\frac{\log k}{m}}$, which will become large. $\hat{R}_{S(h)}$ is related to approximation error, and $R_m(f_{lk}) + \sqrt{\frac{\log k}{m}}$ estimation error. Thus, $F_k(h) = \hat{R}_{S(h)} + R_m(f_{lk}) + \sqrt{\frac{\log k}{m}}$. From the generalization bound considers both types of errors and optimizing it lets us find a good balancing point between these two errors.

(5) For $h \in \mathcal{F} = \bigcup_{k \in \mathbb{N}} f_{lk}$, let $h(k) = \min\{h \in \mathcal{F}_k \mid h \in f_{lk}\}$.

Lemma: $\forall D \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. $\forall m \in \mathbb{N}$. $\forall \delta > 0$.

$$\boxed{\mathbb{P} \left[\forall k \exists h \in f_{lk}. R(h) \leq \hat{R}_{S(h)} + R_m(f_{lk}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log^2/\delta}{2m}} \right] \geq 1 - \delta}$$

Proof. Union bound

$$\begin{aligned} & \mathbb{P} \left[\exists k \exists h \in f_{lk}. R(h) > \hat{R}_{S(h)} + R_m(f_{lk}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log^2/\delta}{2m}} \right] \\ & \leq \sum_{k=1}^{m^D} \mathbb{P} \left[\exists h \in f_{lk}. R(h) > \hat{R}_{S(h)} + R_m(f_{lk}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log^2/\delta}{2m}} \right] \\ & \leq \sum_k \exp \left(-2m \left(\sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log^2/\delta}{2m}} \right)^2 \right) \\ & \stackrel{\text{Thm 3.5.8}}{=} \sum_k \frac{1}{k^2} \left(\frac{\log k}{m} + \frac{\log^2/\delta}{2m} \right) \\ & = \sum_k \frac{1}{k^2} \left(S/2 \right) = \frac{\pi^2}{6} (S/2) \leq S. \end{aligned}$$

D.

Thm 4.2 $\forall D \in \Pr(\mathcal{X} \times \mathcal{Y}) \quad \forall m \in \mathbb{N} \quad \forall S > 0$

$$\mathbb{P}_{\text{sub}}[\underset{h \in \mathcal{H}}{\text{RCh}_S^{\text{SRM}}} \leq \inf_{h \in \mathcal{H}} (\text{R}(h) + 2\text{R}_m(\mathcal{F}_{k(h)}) + 2\sqrt{\frac{\log k}{m}}) + \sqrt{\frac{2\log \frac{4}{\delta}}{m}}] \geq 1 - \delta.$$

$(= \bigcup_{k \in \mathbb{N}} \mathcal{F}_k)$ smaller (often much smaller) than $\text{R}_m(\mathcal{F})$

X ERM bound for $\mathcal{F} = \bigcup_k \mathcal{F}_k$. from Prob. 1. & then there

$$\mathbb{P}_{\text{sub}}[\underset{h \in \mathcal{H}}{\text{RCh}_S^{\text{ERM}}} \leq \inf_{h \in \mathcal{H}} (\text{R}(h) + 2\text{R}_m(\mathcal{F})) + \sqrt{\frac{2\log \frac{2}{\delta}}{m}}] \geq 1 - \delta.$$

Proof of Thm 4.2.

Lemmas in the previous page.
with $\sqrt{\frac{\log 2/\delta}{2m}} = \frac{\varepsilon}{2}$.

Let $\varepsilon = \sqrt{(2\log 4/\delta)/m}$. We have to show that

$$\mathbb{P}_{\text{sub}}[\underset{h \in \mathcal{H}}{\text{RCh}_S^{\text{SRM}}} - \inf_{h \in \mathcal{H}} (\text{R}(h) + 2\text{R}_m(\mathcal{F}_{k(h)}) + 2\sqrt{\frac{\log k}{m}}) \geq \varepsilon] \leq \delta.$$

This holds (by dominated convergence) if for all $h \in \mathcal{H}$,

$$\mathbb{P}[\text{RCh}_S^{\text{SRM}} - (\text{R}(h) + 2\text{R}_m(\mathcal{F}_{k(h)}) + 2\sqrt{\frac{\log k}{m}}) \geq \varepsilon] \leq \delta.$$

We will show this sufficient condition. Let $h_S = h_S^{\text{SRM}}$ and $k_S = k(h_S)$

$$\begin{aligned} & \mathbb{P}[\text{R}(h_S) - (\text{R}(h) + 2\text{R}_m(\mathcal{F}_{k(h)}) + 2\sqrt{\frac{\log k}{m}}) \geq \varepsilon] \\ &= \mathbb{P}[(\text{R}(h_S) - F_{k_S}(h_S)) + (F_{k_S}(h_S) - \text{R}(h)) - 2\text{R}_m(\mathcal{F}_{k(h)}) - 2\sqrt{\frac{\log k}{m}} \geq \varepsilon] \\ &\leq \mathbb{P}[\text{R}(h_S) - F_{k_S}(h_S) \geq \frac{\varepsilon}{2}] + \mathbb{P}[\underbrace{(F_{k_S}(h_S) - \text{R}(h)) - 2\text{R}_m(\mathcal{F}_{k(h)}) - 2\sqrt{\frac{\log k}{m}}}_{\geq \frac{\varepsilon}{2}} \geq \frac{\varepsilon}{2}] \\ &\leq \mathbb{P}[\text{R}(h_S) - F_{k_S}(h_S) \geq \frac{\varepsilon}{2}] + \mathbb{P}[\text{R}(h) - F_{k(h)}(h) - 2\text{R}_m(\mathcal{F}_{k(h)}) - 2\sqrt{\frac{\log k}{m}} \geq \frac{\varepsilon}{2}] \\ &\leq 2 \exp(-2m \cdot \left(\frac{\varepsilon}{2}\right)^2) + \mathbb{P}[\text{R}(h) - F_{k(h)}(h) - 2\text{R}_m(\mathcal{F}_{k(h)}) - 2\sqrt{\frac{\log k}{m}} \geq \frac{\varepsilon}{2}] \\ &\leq 2 \exp(-2m \left(\frac{\varepsilon}{2}\right)^2) + 2 \exp\left[-2m \left(\frac{\varepsilon}{2}\right)^2\right] = 4 \exp\left(-\frac{m\varepsilon^2}{2}\right) = \delta. \end{aligned}$$

↑ Dual version of the lemma in the previous page, i.e.,
 $\mathbb{P}[\forall k \forall h \in \mathcal{H}. \text{R}(h) \geq \text{R}(h_S) - \text{R}_m(\mathcal{F}_k) - \sqrt{\frac{\log k}{m}} - \sqrt{\frac{\log 2/\delta}{2m}}] \geq 1 - \delta.$

D.

- (b) Drawback of SRM:
- ① Difficult computationally.
 - $Rm(f)$ is hard to compute.
 - $\widehat{R}_S(h)$ is not easy to optimise.
 - ② Works only when the family $\{f_k\}_{k \in \mathbb{N}}$ is countable and increasing.

5. Cross Validation (CV)

- Here we mean
- (1) Cross validation used for model selection. It avoids the computation of $Rm(f)$, which is hard or impossible.
- (2) Idea: Use two samples S_1 and S_2 , the former ^{independent} for finding an optimal solution for each hypo. Set f_k , and the latter for estimating the generalisation errors of those solutions. Then, pick one based on the estimated errors.
- (3) Algorithm description.

- we assume a family $\{f_k\}_{k \in \mathbb{N}}$ s.t. $f_k \subseteq f_{k+1}$ for all k .
- Input : $S \in (\mathcal{X} \times \mathcal{Y})^m$.

Divide S into a sample S_1 of size $(1-\delta)m$ and the sample S_2 of the remaining δm pairs.

$$k^* = \underset{k \geq 1}{\operatorname{argmin}} \widehat{R}_{S_2}(h_{S_1, k}^{\text{ERM}}),$$

↑ note that we use S_2 , not S_1 .

where $h_{S_1, k}^{\text{ERM}}$ is an ERM solution wrt. f_k and S_1 .

Return h_S^{CV} .

- (4) Note the simplicity of CV. For instance, we don't need to compute $Rm(f_k)$.

(5) How well does CV work? The next result answers this question.

Thm 4.4 $\forall D \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \forall m \in \mathbb{N} \forall \delta > 0$

$$\mathbb{P}_{\text{sub}}^m [R(\hat{h}_S^{\text{CV}}) - R(h_{S_1}^{\text{SRM}}) \leq 2 \sqrt{\frac{\log \max(k h_S^{\text{CV}}, k h_{S_1}^{\text{SRM}})}{dm}} + 2 \sqrt{\frac{\log(4/\delta)}{dm}}] \geq 1 - \delta.$$

where S_1 is a subsequence of S of size $(1-\delta)m$

that is computed by CV.

① The thm shows that CV and SRM are related.

Note that by combining Thm 4.4 and Thm 4.2, we can get the following guarantee on CV:

$$\mathbb{P}_{\text{sub}}^m [R(\hat{h}_S^{\text{CV}}) \leq \inf_{h \in \mathcal{H}} (R(h) + 2R_{\text{distr}}(\text{ERM}(h)) + 2 \sqrt{\frac{\log t(h)}{dm}}) + 2 \sqrt{\frac{\log \max(k h_S^{\text{CV}}, k h_{S_1}^{\text{SRM}})}{dm}} / dm + (2 + \sqrt{2}) \sqrt{\frac{\log(4/\delta)}{dm}}] \geq 1 - \delta.$$

Not perfect due to the SRM-dependent term.

② How to prove Thm 4.4? We use Prop. 4.3 below:

Prop 4.3. $\forall D \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \forall m \in \mathbb{N} \forall \delta > 0 \forall \varepsilon > 0$

$$\mathbb{P}_{\text{sub}}^m [\sup_{k \geq 1} | R(h_{S_1, k}^{\text{SRM}}) - \hat{R}_{S_1}(h_{S_1, k}^{\text{SRM}}) | > \varepsilon + \sqrt{\frac{\log k}{dm}}] \leq 4 \exp(-2dm\varepsilon^2).$$

Where does it come from?

$\downarrow = 2 \times 2$ union-bound over all k and the term $\sqrt{\frac{\log k}{dm}}$.

Hoeffding for $1 - 1$.

Proof.

$$\begin{aligned} & \mathbb{P} \left[\sup_{k \geq 1} \left| R(h_{S_1, k}) - \hat{R}_{S_2}^{ERM}(h_{S_1, k}) \right| > \varepsilon + \sqrt{\frac{\log k}{dm}} \right] \\ & \leq \sum_{k=1}^{\infty} \mathbb{P} \left[\left| R(h_{S_1, k}) - \hat{R}_{S_2}^{ERM}(h_{S_1, k}) \right| > \varepsilon + \sqrt{\frac{\log k}{dm}} \right] \\ & = \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{P} \left[\cdot \mid S_1 \right] \right] \\ & \leq \sum_{k=1}^{\infty} \mathbb{E} \left[2 \exp \left(-\frac{2(dm(\varepsilon + \sqrt{\frac{\log k}{dm}}))^2}{dm} \right) \right] \end{aligned}$$

Hoeffding.

$$\begin{aligned} & \leq \sum_{k=1}^{\infty} 2 \exp(-2dm\varepsilon^2 - 2\log k) = \sum_{k=1}^{\infty} \frac{2}{k^2} \exp(-2dm\varepsilon^2) \\ & = 2 \times \frac{\pi^2}{6} \exp(-2dm\varepsilon^2) \leq 4 \exp(-2dm\varepsilon^2) \end{aligned}$$

D.

Let's rewrite the conclusion of Prop 4.3 in terms of δ , instead of ε . We set $\delta = 4 \exp(-2dm\varepsilon^2)$.

Then, $\varepsilon = \sqrt{\frac{\log 4/\delta}{2dm}}$. Thus, Prop 4.3 implies:

$$\mathbb{P} \left[\sup_{k \geq 1} \left| R(h_{S_1, k}) - \hat{R}_{S_2}^{ERM}(h_{S_1, k}) \right| \leq \sqrt{\frac{\log k}{dm}} + \sqrt{\frac{\log 4/\delta}{2dm}} \right]$$

$\geq 1 - \delta$. We use this bound in our proof of Thm 4.4.

Proof. of Thm 4.4 If $\sup_{k \geq 1} \left| R(h_{S_1, k}) - \hat{R}_{S_2}^{ERM}(h_{S_1, k}) \right| \leq \sqrt{\frac{\log k}{dm}} + \sqrt{\frac{\log 4/\delta}{2dm}}$,

$$\begin{aligned} R(h_S^{CV}) & \leq \hat{R}_{S_2}^{CV}(h_S^{CV}) + \sqrt{\frac{\log k(h_S^{CV})}{dm}} + \sqrt{\frac{\log 4/\delta}{2dm}} \leq \hat{R}_{S_2}^{SRM}(h_{S_1}^{SRM}) + \sqrt{\frac{\log k(h_S^{CV})}{dm}} + \sqrt{\frac{\log 4/\delta}{2dm}} \\ & \leq R(h_{S_1}^{SRM}) + \sqrt{\frac{\log k(h_{S_1}^{SRM})}{dm}} + \sqrt{\frac{\log k(h_S^{CV})}{dm}} + 2\sqrt{\frac{\log 4/\delta}{2dm}} \\ & \leq R(h_{S_1}^{SRM}) + 2 \sqrt{\frac{\log(\max_{dm} \{k(h_{S_1}^{SRM}), k(h_S^{CV})\})}{dm}} + 2\sqrt{\frac{\log 4/\delta}{2dm}} \end{aligned}$$

Thus, the thm follows from the above. A version of Prop 4.3.

(b) Any problem with CV? We waste examples in S . If d is big, S_1 has a small # of examples. Thus, our ERM solutions might not be that good. On the other hand, if d is small, $|S_1|$ is small, and the generalisation bounds that we proved are loose.

Is there a way to avoid using all examples during ERM while doing some form of cross validation?

N -fold cross validation is such a way.

6. N -fold cross validation.

(1) Idea: ① use S twice, first to find a good model or a hypothesis set \mathcal{H}_K , and then to find a good hypothesis h in \mathcal{H}_K . ② for the first, estimate the quality of generalisation of each \mathcal{H}_K using multiple splits of S based on n folds.

(2) Algorithm.

① Split S into n folds. S_1, S_2, \dots, S_n .

$$(|S| = \sum_{i=1}^n m_i \text{ where } m_i = |S_i|)$$

$$\text{② } \widehat{R}_{CV}(\mathcal{H}_K) = \frac{1}{n} \sum_{i=1}^n \widehat{R}_{S_i}(h_{S_{\setminus i}, K}^{ERM})$$

where $S_{\setminus i}$ means S without the i -th fold S_i .

③ Pick K that maximizes $\widehat{R}_{CV}(\mathcal{H}_K)$.

④ Compute $h_{S_{\setminus K}}^{ERM}$, and return it.

⑤ Note that all the examples in S are used to compute $h_{S_{\setminus K}}^{ERM}$.

model selection based on generalization

hypo. Selection in a selected model.

(4) When $m_1 = \dots = m_n$,

$\hat{R}_{cv}(f)$ is an unbiased estimate of

$$\mathbb{E} [\underset{\text{Sur}^m}{\text{RC}} h_{S,f}^{\text{ERM}}]$$

which measures the quality of generalization of the ERM algorithm over f .

7 Regularization and Convex Surrogate Loss.

(1) Besides the use of $R_{cv}(f)$ in the objective, there are two other issues in SRM. First, it assumes a countable family of hypo. sets $\{f_k\}_{k \in \mathbb{N}}$, which often limits the expressiveness of $f = \bigcup_{k \in \mathbb{N}} f_k$. If possible, we want the family to be uncountable, as in $\{f_r\}_{r \in \mathbb{R}}$ so that $f = \bigcup_{r \in \mathbb{R}} f_r$ becomes highly expressive.

Second, $\hat{R}_{cv}(f)$ is often hard to optimize. Of course, we are not optimizing $\hat{R}_{cv}(f)$ directly in SRM, but the difficulty of optimizing it may cause the opt. of SRM to be challenging. Note that this is not an issue specific to SRM. All of ERM, CV, n-folds CV use $\hat{R}_{cv}(f)$, so that they all get affected by the difficulty of optimizing $\hat{R}_{cv}(f)$.

(2). What should we do? A common approach for the first issue is to use regularization. In a simplest incarnation of the approach,

we consider an uncountable family of typ. sets.

$$\{f_{\gamma}r\}_{r \in \mathbb{R}}$$

$$f = \bigcup_{\gamma > 0} f_{\gamma}.$$

s.t. $f_{\gamma} = \{x \mapsto \underbrace{\text{sgn}(\langle w, \Phi(x) \rangle)}_{\begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases})} \mid \|w\|_p \leq \gamma \text{ and } w \in \mathbb{R}^k\}$

where k and p are fixed natural numbers.
 Φ is a function from \mathcal{X} to \mathbb{R}^k , and $\|w\|_p = \left(\sum_{i=1}^k (w_i)^p \right)^{1/p}$
 is the p -th norm of w . Then, we solve the following optimization problem:

$$\underset{h \in f}{\operatorname{argmin}} \quad \tilde{R}(h) + \lambda \|w\|_p$$

↑ w parameter of h .

↑ called regularisation.

where λ is a hyper-parameter at a very high level,
 (that is, it is a fixed constant).
 Just like SRM, this regularisation-based algorithm can be
 viewed as searching over hypothesis sets and hypotheses
 at the same time.

(3) A popular approach for addressing the second issue is to use a convex surrogate loss. It applies to a setting where hypotheses h of the following form:

$$h: \mathcal{X} \rightarrow \mathcal{Y} = \{-1, 1\}.$$

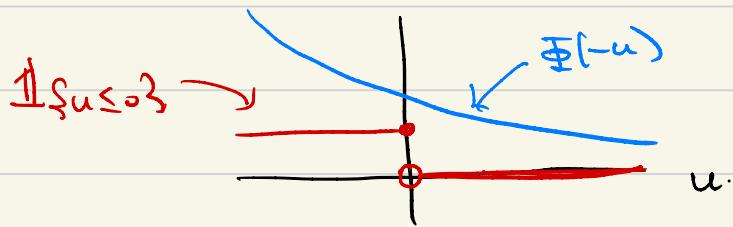
$$h(x) = \begin{cases} +1 & \text{if } f(x) \geq 0 \\ -1 & \text{otherwise.} \end{cases} = \text{sgn}(f(x)).$$

for some $f: \mathcal{X} \rightarrow \mathbb{R}$. We write $h_f^{(x)}$ to mean $\text{sgn}(f(x))$.

To use the approach, we first have to pick a convex and non-increasing function $\tilde{\Phi}: \mathbb{R} \rightarrow \mathbb{R}$ s.t.

$$\mathbb{1}_{\{u \leq 0\}} \leq \Phi(-u) \quad \text{for all } u \in \mathbb{R},$$

which, in picture, means:



Next, we define:

$$L_\Phi : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$$

$$L_\Phi(r, y) = \Phi(-y_r)$$

$$R_S^{\Phi}(h_f) = \frac{1}{m} \sum_{i=1}^m \Phi(-f(x_i) y_i). \quad (S = ((x_1, y_1), \dots, (x_m, y_m)))$$

Finally, we substitute $R_S^{\Phi}(h_f)$ for $R_S(h_f)$, in an optimisation objective, which we saw or will see, and solve the resulting new optimisation problem.

Lemma: $R_S(h_f) \leq R_S^{\Phi}(h_f)$

Proof: $R_S(h_f) \leq \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{h_f(x_i) \neq y_i\}}$ $(S = ((x_1, y_1), \dots, (x_m, y_m)))$

$$= \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{\text{sgn}(f(x_i)) \neq y_i\}}$$

$$\leq \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{y_i f(x_i) \leq 0\}} \leq \frac{1}{m} \sum_{i=1}^m \Phi(-y_i f(x_i))$$

$$= \frac{1}{m} \sum_{i=1}^m L_\Phi(f(x_i), y_i) = R_S^{\Phi}(h_f)$$

Lemma: R_S^{Φ} is convex: for $\alpha \in (0, 1)$, $f, g: \mathcal{X} \rightarrow \mathbb{R}$,

$$R_S^{\Phi}(h_{f+(1-\alpha)g}) \leq \alpha R_S^{\Phi}(h_f) + (1-\alpha) R_S^{\Phi}(h_g)$$

Prf.:

$$R_S^{\Phi}(h_{\alpha f + (1-\alpha)g})$$

$$= \frac{1}{m} \sum_{i=1}^m \Phi(-y_i (\alpha f(x_i) + (1-\alpha)g(x_i)))$$

$$\leq \frac{1}{m} \sum_{i=1}^m \Phi(\alpha(-y_i f(x_i)) + (1-\alpha)(-y_i g(x_i))) \stackrel{\Phi \text{ is convex}}{\leq} \frac{1}{m} \sum_{i=1}^m \alpha \Phi(-y_i f(x_i)) + (1-\alpha) \Phi(-y_i g(x_i))$$

(4) Examples.

- ① Hinge loss : $\Phi(u) = \max(0, 1-u)$
- ② Exponential loss : $\Phi(u) = \exp(u)$
- ③ Logistic loss : $\Phi(u) = \log_2(1+e^u)$

(5) Theoretical analysis.

- ① Bayes scoring fn ... $f^*(x) = \eta(x) - \frac{1}{2}$
where $\eta(x) = P[y=1|x]$.
Note that $h_{f^*} \in h_{\text{Bayes}}$, the Bayes hypothesis
that achieves the minimum risk called Bayes risk R^* .

- ② Let $R(f) = R(h_f) = \mathbb{E}_{x \sim D_f} [\mathbb{1}_{\{\text{sgn}(f(x)) \neq y\}}]$.

Note R^* (Bayes risk) = $R(f^*)$.

Lemma 4.5 $R(f) - R^* \leq 2 \mathbb{E}_{x \sim D_f} [|f^*(x)| \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}]$
 $= R(h_f) - R^*$.
 i.e. excess error of h_f .

Prf. Let $\eta(x) = P[y=1|x]$.

$$\begin{aligned} R(f) &= \mathbb{E}_{x \sim D_f} [\eta(x) \mathbb{1}_{\{f(x) < 0\}} + (1-\eta(x)) \mathbb{1}_{\{f(x) \geq 0\}}] \\ &= \mathbb{E}_{x \sim D_f} [\eta(x) \mathbb{1}_{\{f(x) < 0\}} + (1-\eta(x)) (1 - \mathbb{1}_{\{f(x) < 0\}})] \\ &= \mathbb{E}_x [2\eta(x) - 1] \mathbb{1}_{\{f(x) < 0\}} + (1-\eta(x)). \\ &= \mathbb{E}_x [2f^*(x)] \mathbb{1}_{\{f(x) < 0\}} + (1-\eta(x)). \end{aligned}$$

Thus, $R(f) - R(f^*) = \mathbb{E}_x [2f^*(x) [\mathbb{1}_{\{f(x) < 0\}} - \mathbb{1}_{\{f^*(x) < 0\}}]]$

case analysis $\leq \mathbb{E}_x [2f^*(x) \text{sgn}(f^*(x)) \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}]$
 on $f(x) < 0$ and $f^*(x) < 0$. $= 2 \mathbb{E}_x [|f^*(x)| \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}]$ $\neq (0, 0) \exists$.

③ Bayes scoring function wrt. \mathbb{E} .

$$f_{\mathbb{E}}^*(x) = \arg \min_{u \in [-\infty, \infty]} \eta(x) \mathbb{E}[-u] + (1 - \eta(x)) \mathbb{E}[u]$$

This minimizes the new risk induced by \mathbb{E} (which we explained in ③) as shown below.

↙ new risk:

$$R^{\mathbb{E}}(f) = \mathbb{E}_{(x,y) \sim D} [\mathbb{E}(-yf(x))].$$

$R^{\mathbb{E},*}$

Lemma: $R^{\mathbb{E}}(f^*) = \min_{f \in \mathcal{F} \rightarrow \mathbb{R}} R^{\mathbb{E}}(f)$

Proof: $R^{\mathbb{E}}(f) = \mathbb{E}_{(x,y) \sim D} [\mathbb{E}(-yf(x))]$

$$= \mathbb{E}_{x \sim D} [\eta(x) \mathbb{E}(-f(x)) + (1 - \eta(x)) \mathbb{E}(f(x))]$$

$$\geq \mathbb{E}_{x \sim D} \left[\min_{u \in [-\infty, \infty]} (\eta(x) \mathbb{E}(-u) + (1 - \eta(x)) \mathbb{E}(u)) \right]$$

$$= R^{\mathbb{E}}(f^*)$$

measurable. \mathbb{R}

Lemma: $R(f) \leq R^{\mathbb{E}}(f)$ for all $f: \mathcal{X} \rightarrow \mathbb{R}$.

Proof

$$R(f) = \mathbb{E}_{(x,y) \sim D} [\mathbb{1}_{\{ \text{sgn}(f(x)) \neq y \}}]$$

$$\leq \mathbb{E}_{(x,y) \sim D} [\mathbb{1}_{\{ yf(x) \leq 0 \}}] \stackrel{\mathbb{E}_{(x,y) \sim D} [\mathbb{E}(-yf(x))]}{=} R^{\mathbb{E}}(f)$$

Thm 4.7

Φ .. convex, non-decreasing.

$\exists s \geq 1$ and $c > 0$ s.t. $\forall x \in \mathcal{X}$.

$$|f^*(x)|^s = |\eta(x) - \frac{1}{2}|^s \leq c^s [\Phi(0) - \mathbb{E} [\Phi(-y f_{\Phi}^*(x))]]_{P[y|x]}$$

\Rightarrow for all $f: \mathcal{X} \rightarrow \mathbb{R}$,

$$(R(f) - R^*)^s \leq (2c)^s [R^{\Phi}(f) - R^{\Phi,*}]$$

Proof By the convexity of Φ ,

$$\Phi(-2f^*(x)f(x)) = \Phi((1-2\eta(x))f(x))$$

$$= \Phi(-\eta(x)f(x) + (1-\eta(x))f(x))$$

$$\stackrel{\text{Jensen}}{\leq} \eta(x)\Phi(-f(x)) + (1-\eta(x))\Phi(f(x)) = \mathbb{E} [\Phi(-y f(x))]_{P[y|x]}$$

Lemma 4.5.

$$R(f) - R(f^*) \leq \mathbb{E} [2|f^*(x)|^{\frac{1}{s}} \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}].$$

$$= \mathbb{E} [(|2\eta(x)-1|^s \mathbb{1}_{\{f(x)f^*(x) \leq 0\}})^{\frac{1}{s}}]$$

$$\leq \mathbb{E} [|2\eta(x)-1|^s \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}]^{\frac{1}{s}}$$

$$\leq \mathbb{E} [2^s c^s [\Phi(0) - \mathbb{E} [\Phi(-y f_{\Phi}^*(x))]]] \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}^{\frac{1}{s}}$$

$$\leq 2c \mathbb{E} [[\Phi(-2f^*f(x))] - \mathbb{E} [\Phi(-y f_{\Phi}^*(x))]] \mathbb{1}_{\{f(x)f^*(x) \leq 0\}}^{\frac{1}{s}}$$

$$\leq 2c \mathbb{E} [\mathbb{E} [\Phi(-y f(x))] - \mathbb{E} [\Phi(-y f_{\Phi}^*(x))]]^{\frac{1}{s}}$$

$$\leq 2c \mathbb{E} [\sqrt{\mathbb{E} [\Phi(-y f(x))]} - \mathbb{E} [\Phi(-y f_{\Phi}^*(x))]]^{\frac{1}{s}}$$

$$\leq 2c (R^{\Phi}(f) - R^{\Phi,*})^{\frac{1}{s}}.$$

X Hinge ... $s=1$, $c=\frac{1}{2}$, Exponential ... $s=2$, $c=\frac{1}{\sqrt{2}}$,
Logistic ... $s=2$, $c=\frac{1}{\sqrt{2}}$.