8

Statistical Learning

8.1 Introduction

Suppose that one observes independent variables $\xi_1, \ldots, \xi_n$ taking their values in some measurable space Ξ. Let us furthermore assume, for the sake of simplicity, that these variables are identically distributed with common distribution P. The two main frameworks that we have in mind are respectively the classification and the bounded regression frameworks. In those cases, for every i the variable $\xi_i = (X_i, Y_i)$ is a copy of a pair of random variables (X, Y), where X takes its values in some measurable space $\mathcal{X}$ and Y is assumed to take its values in $[0, 1]$. In the classification case, the response variable Y is assumed to belong to $\{0, 1\}$. One defines the regression function η as

$$\eta(x) = \mathbb{E}[Y \mid X = x] \tag{8.1}$$

for every $x \in \mathcal{X}$. In the regression case, one is interested in the estimation of η while in the classification case, one wants to estimate the Bayes classifier s^*, defined for every $x \in \mathcal{X}$ by

$$s^*(x) = \mathbb{1}_{\eta(x) \geq 1/2}. \tag{8.2}$$

One of the most commonly used method to estimate η or s^* or more generally to estimate a quantity of interest s depending on the unknown distribution P is the so called empirical risk minimization by Vapnik which is a special instance of minimum contrast estimation.

Empirical Risk Minimization

Basically one considers some set $\mathcal{S}$ which is known to contain s, think of $\mathcal{S}$ as being the set of all measurable functions from $\mathcal{X}$ to $[0, 1]$ in the regression case or to $\{0, 1\}$ in the classification case. Then we consider some *loss* (*or contrast*) function

$$\gamma \text{ from } \mathcal{S} \times \Xi \text{ to } [0, 1]$$

which is well adapted to our estimation problem of s in the sense that the *expected loss* $\mathbb{E}[\gamma(t,\xi_1)]$ achieves a minimum at point s when t varies in $\mathcal{S}$. In other words the *relative expected loss* ℓ defined by

$$\ell(s,t) = \mathbb{E}[\gamma(t,\xi_1) - \gamma(s,\xi_1)], \text{ for all } t \in \mathcal{S} \tag{8.3}$$

is nonnegative. In the regression or the classification cases, one can take

$$\gamma(t,(x,y)) = (y - t(x))^2$$

since η (resp. s^*) is indeed the minimizer of $\mathbb{E}\left[(Y - t(X))^2\right]$ over the set of measurable functions t taking their values in $[0,1]$ (resp. $\{0,1\}$). The heuristics of empirical risk minimization (or minimum contrast estimation) can be described as follows. If one substitutes the empirical risk

$$\gamma_n(t) = P_n[\gamma(t,.)] = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,\xi_i),$$

to its expectation $P[\gamma(t,.)] = \mathbb{E}[\gamma(t,\xi_1)]$ and minimizes γ_n on some subset S of $\mathcal{S}$ (that we call a *model*), there is some hope to get a sensible estimator $\widehat{s}$ of s, at least if s belongs (or is close enough) to the model S.

8.2 Model Selection in Statistical Learning

The purpose of this section is to provide an other look at the celebrated Vapnik's method of *structural risk minimization* (initiated in [121]) based on concentration inequalities. In the next section, we shall present an alternative analysis which can lead to improvements of Vapnik's method for the classification problem. Let us consider some countable or finite (but possibly depending on n) collection of models $\{S_m\}_{m\in\mathcal{M}}$ and the corresponding collection of empirical risk minimizers $\{\widehat{s}_m\}_{m\in\mathcal{M}}$. For every $m \in \mathcal{M}$ an empirical risk minimizer within model S_m is defined by

$$\widehat{s}_m = \operatorname{argmin}_{t\in S_m}\gamma_n(t).$$

Given some penalty function pen: $\mathcal{M} \to \mathbb{R}_+$ and let us define $\widehat{m}$ as a minimizer of

$$\gamma_n(\widehat{s}_m) + \operatorname{pen}(m) \tag{8.4}$$

over $\mathcal{M}$ and finally estimate s by the *penalized estimator*

$$\widetilde{s} = \widehat{s}_{\widehat{m}}.$$

Since some problems can occur with the existence of a solution to the previous minimization problems, it is useful to consider approximate solutions (note that even if $\widehat{s}_m$ does exist, it is relevant from a practical point of view

to consider approximate solutions since $\widehat{s}_m$ will typically be approximated by some numerical algorithm). Therefore, given $\rho \geq 0$ (in practice, taking $\rho = n^{-2}$ makes the introduction of an approximate solution painless), we shall consider for every $m \in \mathcal{M}$ some approximate empirical risk minimizer $\widehat{s}_m$ satisfying

$$\gamma_n\left(\widehat{s}_m\right) \leq \gamma_n\left(t\right) + \rho$$

and say that $\widetilde{s}$ is a *ρ-penalized estimator* of s if

$$\gamma_n\left(\widetilde{s}\right) + \mathrm{pen}\left(\widehat{m}\right) \leq \gamma_n\left(t\right) + \mathrm{pen}\left(m\right) + \rho, \, \forall m \in \mathcal{M} \text{ and } \forall t \in S_m. \tag{8.5}$$

To analyze the statistical performance of this procedure, the key is to take $\ell\left(s,t\right)$ as a loss function and notice that the definition of the penalized procedure leads to a very simple but fundamental control for $\ell\left(s,\widetilde{s}\right)$. Indeed, by the definition of $\widetilde{s}$ we have, whatever $m \in \mathcal{M}$ and $s_m \in S_m$,

$$\gamma_n\left(\widetilde{s}\right) + \mathrm{pen}\left(\widehat{m}\right) \leq \gamma_n\left(s_m\right) + \mathrm{pen}\left(m\right) + \rho,$$

and therefore

$$\gamma_n\left(\widetilde{s}\right) \leq \gamma_n\left(s_m\right) + \mathrm{pen}\left(m\right) - \mathrm{pen}\left(\widehat{m}\right) + \rho. \tag{8.6}$$

If we introduce the centered empirical process

$$\overline{\gamma}_n\left(t\right) = \gamma_n\left(t\right) - \mathbb{E}\left[\gamma\left(t,\xi_1\right)\right], \, t \in \mathcal{S}$$

and notice that $\mathbb{E}\left[\gamma\left(t,\xi_1\right)\right] - \mathbb{E}\left[\gamma\left(u,\xi_1\right)\right] = \ell\left(s,t\right) - \ell\left(s,u\right)$ for all $t,u \in \mathcal{S}$, we readily get from (8.6)

$$\ell\left(s,\widetilde{s}\right) \leq \ell\left(s,s_m\right) + \overline{\gamma}_n\left(s_m\right) - \overline{\gamma}_n\left(\widetilde{s}\right) - \mathrm{pen}\left(\widehat{m}\right) + \mathrm{pen}\left(m\right) + \rho. \tag{8.7}$$

8.2.1 A Model Selection Theorem

Let us first see what can be derived from (8.7) by using only the following boundedness assumption on the contrast function γ

A1 For every t belonging to some set $\mathcal{S}$, one has $0 \leq \gamma\left(t,.\right) \leq 1$.

In order to avoid any measurability problem, let us first assume that each of the models S_m is countable. Given some constant Σ, let us consider some preliminary collection of nonnegative weights $\left\{x_m\right\}_{m \in \mathcal{M}}$ such that

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq \Sigma$$

and let $z > 0$ be given. It follows from (5.7) (which was proved in Chapter 5 to be a consequence of Mc Diarmid's Inequality) that for every $m' \in \mathcal{M}$,

$$\mathbb{P}\left[\sup_{t \in S_{m'}} \left(-\overline{\gamma}_n\left(t\right)\right) \geq \mathbb{E}\left[\sup_{t \in S_{m'}} \left(-\overline{\gamma}_n\left(t\right)\right)\right] + \sqrt{\frac{x_{m'} + z}{2n}}\right] \leq e^{-x_{m'} - z},$$

and therefore, setting $\mathbb{E}\left[\sup_{t\in S_{m'}}\left(-\overline{\gamma}_n\left(t\right)\right)\right]=E_{m'}$, except on a set of probability not larger than Σe^{-z}, one has for every $m'\in\mathcal{M}$,

$$\sup_{t\in S_{m'}}\left(-\overline{\gamma}_n\left(t\right)\right)\leq E_{m'}+\sqrt{\frac{x_{m'}+z}{2n}}.$$

Hence, (8.7) implies that the following inequality holds, except on a set of probability not larger than Σe^{-z}:

$$\ell\left(s,\widetilde{s}\right)\leq\ell\left(s,s_m\right)+\overline{\gamma}_n\left(s_m\right)+E_{\widehat{m}}+\sqrt{\frac{x_{\widehat{m}}}{2n}}-\mathrm{pen}\left(\widehat{m}\right)+\mathrm{pen}\left(m\right)+\sqrt{\frac{z}{2n}}+\rho. \quad (8.8)$$

It is tempting to choose $\mathrm{pen}\left(m'\right)=E_{m'}+\sqrt{x_{m'}/2n}$ for every $m'\in\mathcal{M}$ but we should not forget that $E_{m'}$ typically depends on the unknown s. Thus, we are forced to consider some upper bound $\widetilde{E}_{m'}$ of $E_{m'}$ which does not depend on s. This upper bound can be either deterministic (we shall discuss below the drawbacks of this strategy) or random and in such a case we shall take benefit of the fact that it is enough to assume that $\widetilde{E}_{m'}\geq E_{m'}$ holds on a set with sufficiently high probability. More precisely, assuming that for some constant K and for every $m'\in\mathcal{M}$

$$\mathrm{pen}\left(m'\right)\geq E_{m'}+\sqrt{\frac{x_{m'}}{2n}}-K\sqrt{\frac{z}{2n}} \quad (8.9)$$

holds, except on set of probability not larger than $\exp\left(-x_{m'}-z\right)$, we derive from (8.8) and (8.9) that

$$\ell\left(s,\widetilde{s}\,\right)\leq\ell\left(s,s_m\right)+\overline{\gamma}_n\left(s_m\right)+\mathrm{pen}\left(m\right)+\left(1+K\right)\sqrt{\frac{z}{2n}}+\rho$$

holds except on a set of probability not larger than $2\Sigma e^{-z}$. Thus, integrating with respect to z leads to

$$\mathbb{E}\left[\left(\ell\left(s,\widetilde{s}\,\right)-\ell\left(s,s_m\right)-\overline{\gamma}_n\left(s_m\right)-\mathrm{pen}\left(m\right)-\rho\right)^+\right]\leq\Sigma\left(1+K\right)\sqrt{\frac{\pi}{2n}}$$

and therefore, since $\overline{\gamma}_n\left(s_m\right)$ is centered at expectation

$$\mathbb{E}\left[\ell\left(s,\widetilde{s}\,\right)\right]\leq\ell\left(s,s_m\right)+\mathbb{E}\left[\mathrm{pen}\left(m\right)\right]+\Sigma\left(1+K\right)\sqrt{\frac{\pi}{2n}}+\rho.$$

Hence, we have proven the following result.

Theorem 8.1 *Let $\xi_1,\ldots,\xi_n$ be independent observations taking their values in some measurable space Ξ and with common distribution P depending on some unknown parameter $s\in\mathcal{S}$. Let $\gamma:\mathcal{S}\times\Xi\to\mathbb{R}$ be some contrast function*

satisfying assumption ***A1****. Let* $\{S_m\}_{m\in\mathcal{M}}$ *be some at most countable collection of countable subsets of* $\mathcal{S}$ *and* $\rho \geq 0$ *be given. Consider some absolute constant* Σ*, some family of nonnegative weights* $\{x_m\}_{m\in\mathcal{M}}$ *such that*

$$\sum_{m\in\mathcal{M}} e^{-x_m} = \Sigma < \infty$$

and some (possibly data-dependent) penalty function $pen : \mathcal{M} \to \mathbb{R}_+$*. Let* $\widetilde{s}$ *be a* ρ*-penalized estimator of* s *as defined by (8.5). Then, if for some nonnegative constant* K*, for every* $m \in \mathcal{M}$ *and every positive* z

$$pen(m) \geq \mathbb{E}\left[\sup_{t\in S_m} (-\overline{\gamma}_n(t))\right] + \sqrt{\frac{x_m}{2n}} - K\sqrt{\frac{z}{2n}}$$

holds with probability larger than $1 - \exp(-x_m - z)$*, the following risk bound holds for all* $s \in \mathcal{S}$

$$\mathbb{E}\left[\ell(s,\widetilde{s})\right] \leq \inf_{m\in\mathcal{M}} \left(\ell(s,S_m) + \mathbb{E}\left[pen(m)\right]\right) + \Sigma(1+K)\sqrt{\frac{\pi}{2n}} + \rho, \quad (8.10)$$

where ℓ *is defined by (8.3) and* $\ell(s,S_m) = \inf_{t\in S_m} \ell(s,t)$*.*

It is not that easy to discuss whether this result is sharp or not in the generality where it is stated here. Nevertheless we shall see that, at the price of making an extra assumption on the contrast function γ, it is possible to improve on (8.10) by weakening the restriction on the penalty function. This will be the purpose of our next section.

Vapnik's Learning Theory Revisited

We would like here to explain how Vapnik's *structural minimization of the risk method* (as described in [121] and further developed in [122]) fits in the above framework of model selection via penalization. More precisely, we shall consider the *classification* problem and show how to recover (or refine in the spirit of [31]) some of Vapnik's results from Theorem 8.1. The data $\xi_1 = (X_1, Y_1), \ldots, \xi_n = (X_n, Y_n)$ consist of independent, identically distributed copies of the random variable pair (X, Y) taking values in $\mathcal{X} \times \{0,1\}$. Let the models $\{S_m\}_{m\in\mathcal{M}}$ being defined for every $m \in \mathcal{M}$ as

$$S_m = \{\mathbb{1}_C : C \in \mathcal{A}_m\},$$

where $\mathcal{A}_m$ is some countable class of subsets of $\mathcal{X}$. Let $\mathcal{S}$ be the set of measurable functions taking their values in $[0,1]$. In this case, the least squares contrast function fulfills condition **A1**. Indeed,since $\gamma(t,(x,y)) = (y - t(x))^2$,

A1 is fulfilled whenever $t \in \mathcal{S}$ and $y \in [0,1]$. For a function t taking only the two values 0 and 1, the least squares criterion also writes

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - t\left(X_i\right)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i \neq t(X_i)}$$

so that minimizing the least squares criterion means minimizing the number of misclassifications on the training sample $\xi_1 = (X_1, Y_1), \ldots, \xi_n = (X_n, Y_n)$. Each estimator $\widehat{s}_m$ represents some possible classification rule and the purpose of model selection is here to select what classification rule is the best according to some risk minimization criterion. At this stage it should be noticed that we have the choice here between two different definitions of the statistical object of interest s. Indeed, we can take s to be the minimizer of $t \to \mathbb{E}\left[Y - t\left(X\right)\right]^2$ subject or not to the restriction that t takes its values in $\{0,1\}$. On the one hand the Bayes classifier s^* as defined by (8.2) is a minimizer of $\mathbb{E}\left[Y - t\left(X\right)\right]^2$ under the restriction that t takes its values in $\{0,1\}$ and the corresponding loss function can be written as

$$\ell\left(s^*, t\right) = \mathbb{E}\left[s^*\left(X\right) - t\left(X\right)\right]^2 = \mathbb{P}\left[Y \neq t\left(X\right)\right] - \mathbb{P}\left[Y \neq s^*\left(X\right)\right].$$

On the other hand, the regression function η as defined by (8.1) minimizes $\mathbb{E}\left[Y - t\left(X\right)\right]^2$ without the restriction that t takes its values in $\{0,1\}$, and the corresponding loss function is simply $\ell\left(\eta, t\right) = \mathbb{E}\left[\eta\left(X\right) - t\left(X\right)\right]^2$. It turns out that the results presented below are valid for both definitions of s simultaneously. In order to apply Theorem 8.1, it remains to upper bound $\mathbb{E}\left[\sup_{t \in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]$. Let us introduce the (random) combinatorial entropy of $\mathcal{A}_m$

$$H_m = \ln\left|\left\{A \cap \left\{X_1, \ldots, X_n\right\},\, A \in \mathcal{A}_m\right\}\right|.$$

If we take some independent copy $(\xi_1', \ldots, \xi_n')$ of $(\xi_1, \ldots, \xi_n)$ and consider the corresponding copy γ_n' of γ_n, we can use the following standard symmetrization argument. By Jensen's inequality

$$\mathbb{E}\left[\sup_{t \in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right] \leq \mathbb{E}\left[\sup_{t \in S_m}\left(\gamma_n'\left(t\right) - \gamma_n\left(t\right)\right)\right],$$

so that, given independent random signs $(\varepsilon_1, \ldots, \varepsilon_n)$, independent of $(\xi_1, \ldots, \xi_n)$, one has,

$$\begin{aligned}\mathbb{E}\left[\sup_{t \in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right] &\leq \frac{1}{n}\mathbb{E}\left[\sup_{t \in S_m}\left(\sum_{i=1}^{n}\varepsilon_i\left(\mathbb{1}_{Y_i' \neq t\left(X_i'\right)} - \mathbb{1}_{Y_i \neq t\left(X_i\right)}\right)\right)\right] \\ &\leq \frac{2}{n}\mathbb{E}\left[\sup_{t \in S_m}\left(\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{Y_i \neq t\left(X_i\right)}\right)\right].\end{aligned}$$

Hence, by (6.3) we get

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]\leq\frac{2\sqrt{2}}{n}\mathbb{E}\left[\left(H_m\sup_{t\in S_m}\left(\sum_{i=1}^{n}\mathbb{1}_{Y_i\neq t(X_i)}\right)\right)^{1/2}\right],$$

and by Jensen's inequality

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]\leq 2\sqrt{\frac{2\mathbb{E}\left[H_m\right]}{n}}. \tag{8.11}$$

The trouble now is that $\mathbb{E}\left[H_m\right]$ is unknown. Two different strategies can be followed to overcome this difficulty. First, one can use the VC-dimension to upper bound $\mathbb{E}\left[H_m\right]$. Assume each $\mathcal{A}_m$ to be a VC-class with VC-dimension V_m (see Definition 6.2), one derives from (6.9) that

$$\mathbb{E}\left[H_m\right]\leq V_m\left(1+\ln\left(\frac{n}{V_m}\right)\right). \tag{8.12}$$

If $\mathcal{M}$ has cardinality not larger than n, one can take $x_m=\ln\left(n\right)$ for each $m\in\mathcal{M}$ which leads to a penalty function of the form

$$\mathrm{pen}\left(m\right)=2\sqrt{\frac{2V_m\left(1+\ln\left(n/V_m\right)\right)}{n}}+\sqrt{\frac{\ln\left(n\right)}{2n}},$$

and to the following risk bound for the corresponding penalized estimator $\widetilde{s}$, since then $\Sigma\leq 1$:

$$\mathbb{E}\left[\ell\left(s,\widetilde{s}\right)\right]\leq\inf_{m\in\mathcal{M}}\left(\ell\left(s,S_m\right)+\mathrm{pen}\left(m\right)\right)+\sqrt{\frac{\pi}{2n}}+\rho. \tag{8.13}$$

This approach has two main drawbacks:

- the VC-dimension of a given collection of sets is generally very difficult to compute or even to evaluate (see [6] and [69] for instance);
- even if the VC-dimension is computable (in the case of affine half spaces of $\mathbb{R}^d$ for instance), inequality (8.12) is too pessimistic and it would be desirable to define a penalty function from a quantity which is much closer to $\mathbb{E}\left[H_m\right]$ than the right-hand side of (8.12).

Following [31], the second strategy consists of substituting H_m to $\mathbb{E}\left[H_m\right]$ by using again a concentration argument. Indeed, by (5.22), for any positive z, one has $H_m\geq\mathbb{E}\left[H_m\right]-\sqrt{2\ln\left(2\right)\mathbb{E}\left[H_m\right]\left(x_m+z\right)}$, on a set of probability not less than $1-\exp\left(-x_m-z\right)$. Hence, since

$$\sqrt{2\ln\left(2\right)\mathbb{E}\left[H_m\right]\left(x_m+z\right)}\leq\frac{\mathbb{E}\left[H_m\right]}{2}+\ln\left(2\right)\left(x_m+z\right),$$

we have on the same set,

$$\mathbb{E}\left[H_m\right]\leq 2H_m+2\ln\left(2\right)\left(x_m+z\right),$$

which, by (8.11), yields

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]\leq 4\left(\sqrt{\frac{H_m}{n}}+\sqrt{\frac{\ln\left(2\right)x_m}{n}}+\sqrt{\frac{\ln\left(2\right)z}{n}}\right).$$

Taking $x_m=\ln\left(n\right)$ as before leads to the following choice for the penalty function

$$\text{pen}\left(m\right)=4\sqrt{\frac{H_m}{n}}+4.1\sqrt{\frac{\ln\left(n\right)}{n}},$$

which satisfies

$$\text{pen}\left(m\right)\geq\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]+\sqrt{\frac{\ln\left(n\right)}{2n}}-4\sqrt{\frac{\ln\left(2\right)z}{n}}.$$

The corresponding risk bound can be written as

$$\mathbb{E}\left[\ell\left(s,\widetilde{s}\right)\right]\leq\left[\inf_{m\in\mathcal{M}}\left(\ell\left(s,S_m\right)+\mathbb{E}\left[\text{pen}\left(m\right)\right]\right)+4\sqrt{\frac{\pi\ln\left(2\right)}{n}}+\rho\right],$$

and therefore, by Jensen's inequality

$$\mathbb{E}\left[\ell\left(s,\widetilde{s}\right)\right]\leq\left[\inf_{m\in\mathcal{M}}\left(\ell\left(s,S_m\right)+4\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}\right)+4.1\sqrt{\frac{\ln\left(n\right)}{n}}+\frac{6}{\sqrt{n}}+\rho\right]. \tag{8.14}$$

Note that if we take $s=s^*$, denoting by L_t the probability of misclassification of the rule t, i.e., $L_t=\mathbb{P}\left[Y\neq t\left(X\right)\right]$, the risk bound (8.14) can also be written as

$$\mathbb{E}\left[L_{\widetilde{s}}\right]\leq\inf_{m\in\mathcal{M}}\left(\inf_{t\in S_m}L_t+4\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}\right)+4.1\sqrt{\frac{\ln\left(n\right)}{n}}+\frac{6}{\sqrt{n}}+\rho,$$

which is maybe a more standard way of expressing the performance of a classifier in the statistical learning literature. Of course, if we follow the first strategy of penalization a similar bound can be derived from (8.13), namely

$$\mathbb{E}\left[L_{\widetilde{s}}\right]\leq\inf_{m\in\mathcal{M}}\left(\inf_{t\in S_m}L_t+2\sqrt{\frac{2V_m\left(1+\ln\left(n/V_m\right)\right)}{n}}\right)$$
$$+\sqrt{\frac{\ln\left(n\right)}{2n}}+\sqrt{\frac{\pi}{2n}}+\rho.$$

Note that the same conclusions would hold true (up to straightforward modifications of the absolute constants) if instead of the combinatorial entropy H_m, one would take as a random measure of complexity for the class S_m the Rademacher conditional mean

$$\frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{t\in S_m}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{Y_i\neq t(X_i)} \mid (X_i,Y_i)_{1\leq i\leq n}\right]$$

since we have indeed seen in Chapter 4 that this quantity obeys exactly to the same concentration inequality as H_m. This leads to risk bounds for Rademacher penalties of the same nature as those obtained by Bartlett, Boucheron and Lugosi (see [13]) or Koltchinskii (see [70]).

8.3 A Refined Analysis for the Risk of an Empirical Risk Minimizer

The purpose of this section is to provide a general upper bound for the relative expected loss between $\widehat{s}$ and s, where $\widehat{s}$ denotes the empirical risk minimizer over a given model S.

We introduce the centered empirical process $\overline{\gamma}_n$. In addition to the relative expected loss function ℓ we shall need another way of measuring the closeness between the elements of S which is directly connected to the variance of the increments of $\overline{\gamma}_n$ and therefore will play an important role in the analysis of the fluctuations of $\overline{\gamma}_n$. Let d be some pseudodistance on $\mathcal{S}\times\mathcal{S}$ (which may perfectly depend on the unknown distribution P) such that

$$P\left(\left(\gamma\left(t,.\right)-\gamma\left(s,.\right)\right)^2\right)\leq d^2\left(s,t\right), \text{ for every } t\in\mathcal{S}.$$

Of course, we can take d as the pseudodistance associated to the variance of γ itself, but it will more convenient in the applications to take d as a more intrinsic distance. For instance, in the regression or the classification setting it is easy to see that d can be chosen (up to constant) as the $\mathbb{L}_2\left(\mu\right)$ distance, where μ denotes the distribution of X. Indeed, for classification

$$\left|\gamma\left(t,(x,y)\right)-\gamma\left(s^*,(x,y)\right)\right|=\left|\mathbb{1}_{y\neq t(x)}-\mathbb{1}_{y\neq s^*(x)}\right|=\left|t\left(x\right)-s^*\left(x\right)\right|$$

and therefore setting $d^2\left(s^*,t\right)=\mathbb{E}\left[\left(t\left(X\right)-s^*\left(X\right)\right)^2\right]$ leads to

$$P\left(\left(\gamma\left(t,.\right)-\gamma\left(s^*,.\right)\right)^2\right)\leq d^2\left(s^*,t\right).$$

For regression, we write

$$\left[\gamma\left(t,(x,y)\right)-\gamma\left(\eta,(x,y)\right)\right]^2=\left[t\left(x\right)-\eta\left(x\right)\right]^2\left[2\left(y-\eta\left(x\right)\right)-t\left(x\right)+\eta\left(x\right)\right]^2.$$

Now

$$\mathbb{E}\left[Y-\eta\left(X\right)\mid X\right]=0 \text{ and } \mathbb{E}\left[\left(Y-\eta\left(X\right)\right)^2\mid X\right]\leq\frac{1}{4},$$

imply that

$$\mathbb{E}\left[\left[2\left(y-\eta\left(x\right)\right)-t\left(x\right)+\eta\left(x\right)\right]^{2}\mid X\right]=4\mathbb{E}\left[\left(Y-\eta\left(X\right)\right)^{2}\mid X\right]+\left(-t\left(X\right)+\eta\left(X\right)\right)^{2}\leq 2,$$

and therefore

$$P\left(\left(\gamma\left(t,.\right)-\gamma\left(\eta,.\right)\right)^{2}\right)\leq 2\mathbb{E}\left(t\left(X\right)-\eta\left(X\right)\right)^{2}. \tag{8.15}$$

Our main result below will crucially depend on two different moduli of uniform continuity:

- the stochastic modulus of uniform continuity of $\overline{\gamma}_n$ over S with respect to d,
- the modulus of uniform continuity of d with respect to ℓ.

The main tool that we shall use is Bousquet's version of Talagrand's inequality for empirical processes (see Chapter 4) which will allow us to control the oscillations of the empirical process $\overline{\gamma}_n$ by the modulus of uniform continuity of $\overline{\gamma}_n$ in expectation. Bousquet's version has the advantage of providing explicit constants and of dealing with one-sided suprema (in the spirit of [91], we could also work with absolute suprema but it is easier and somehow more natural to work with one-sided suprema).

8.3.1 The Main Theorem

We need to specify some mild regularity conditions that we shall assume to be verified by the moduli of continuity involved in our result.

Definition 8.2 *We denote by $\mathcal{C}_1$ the class of nondecreasing and continuous functions ψ from $\mathbb{R}_+$ to $\mathbb{R}_+$ such that $x \to \psi\left(x\right)/x$ is nonincreasing on $(0,+\infty)$ and $\psi\left(1\right)\geq 1$.*

Note that if ψ is a nondecreasing continuous and concave function on $\mathbb{R}_+$with $\psi\left(0\right)=0$ and $\psi\left(1\right)\geq 1$, then ψ belongs to $\mathcal{C}_1$. In particular, for the applications that we shall study below an example of special interest is $\psi\left(x\right)=Ax^{\alpha}$, where $\alpha\in[0,1]$ and $A\geq 1$.

In order to avoid measurability problems and to use the concentration tools, we need to consider some separability condition on S. The following one will be convenient

(M) There exists some countable subset S' of S such that for every $t\in S$, there exists some sequence $\left(t_k\right)_{k\geq 1}$ of elements of S' such that for every $\xi\in\Xi$, $\gamma\left(t_k,\xi\right)$ tends to $\gamma\left(t,\xi\right)$ as k tends to infinity.

We are now in a position to state our upper bound for the relative expected loss of any empirical risk minimizer on some given model S.

Theorem 8.3 *Let $\xi_1, \ldots, \xi_n$ be independent observations taking their values in some measurable space Ξ and with common distribution P. Let $\mathcal{S}$ be some set, $\gamma : \mathcal{S} \times \Xi \to [0,1]$ be a measurable function such that for every $t \in \mathcal{S}$, $x \to \gamma(t,x)$ is measurable. Assume that there exists some minimizer s of $P(\gamma(t,.))$ over $\mathcal{S}$ and denote by $\ell(s,t)$ the nonnegative quantity $P(\gamma(t,.)) - P(\gamma(s,.))$ for every $t \in \mathcal{S}$. Let γ_n be the empirical risk*

$$\gamma_n(t) = P_n(\gamma(t,.)) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t,\xi_i), \text{ for every } t \in \mathcal{S}$$

and $\overline{\gamma}_n$ be the centered empirical process defined by

$$\overline{\gamma}_n(t) = P_n(\gamma(t,.)) - P(\gamma(t,.)), \text{ for every } t \in \mathcal{S}.$$

Let d be some pseudodistance on $\mathcal{S} \times \mathcal{S}$ such that

$$P\left((\gamma(t,.) - \gamma(s,.))^2\right) \leq d^2(s,t), \text{ for every } t \in \mathcal{S}. \tag{8.16}$$

*Let ϕ and w belong to the class of functions $\mathcal{C}_1$ defined above and let S be a subset of $\mathcal{S}$ satisfying separability condition (**M**). Assume that on the one hand, for every $t \in \mathcal{S}$*

$$d(s,t) \leq w\left(\sqrt{\ell(s,t)}\right) \tag{8.17}$$

and that on the other hand one has for every $u \in S'$

$$\sqrt{n}\mathbb{E}\left[\sup_{t \in S', d(u,t) \leq \sigma} [\overline{\gamma}_n(u) - \overline{\gamma}_n(t)]\right] \leq \phi(\sigma) \tag{8.18}$$

*for every positive σ such that $\phi(\sigma) \leq \sqrt{n}\sigma^2$, where S' is given by assumption (**M**). Let ε_* be the unique solution of the equation*

$$\sqrt{n}\varepsilon_*^2 = \phi(w(\varepsilon_*)). \tag{8.19}$$

Let ρ be some given nonnegative real number and consider any ρ-empirical risk minimizer, i.e., any estimator $\widehat{s}$ taking its values in S and such that

$$\gamma_n(\widehat{s}) \leq \rho + \inf_{t \in S} \gamma_n(t).$$

Then, setting

$$\ell(s,S) = \inf_{t \in S} \ell(s,t),$$

there exists some absolute constant κ such that for every $y \geq 0$, the following inequality holds

$$\mathbb{P}\left[\ell(s,\widehat{s}) > 2\rho + 2\ell(s,S) + \kappa\left(\varepsilon_*^2 + \frac{(1 \wedge w^2(\varepsilon_*))}{n\varepsilon_*^2} y\right)\right] \leq e^{-y}. \tag{8.20}$$

In particular, the following risk bound is available

$$\mathbb{E}[\ell(s,\widehat{s})] \leq 2\left(\rho + \ell(s,S) + \kappa\varepsilon_*^2\right).$$

Comments.

Let us give some first comments about Theorem 8.3.

- The absolute constant 2 appearing in (8.20) has no magic meaning here, it could be replaced by any $C > 1$ at the price of making the constant κ depend on C.
- One can wonder whether an empirical risk minimizer over S does exist or not. Note that condition (**M**) implies that for every positive ρ, there exists some measurable choice of a ρ-empirical risk minimizer since then $\inf_{t \in S'} \gamma_n(t) = \inf_{t \in S} \gamma_n(t)$. If $\rho = 1/n$ for instance, it is clear that, according to (8.20), such an estimator performs as well as a strict empirical risk minimizer.
- For the computation of ϕ satisfying (8.18), since the supremum appearing in the left-hand side of (8.18) is extended to the countable set S' and not S itself, it will allow us to restrict ourself to the case where S is countable.
- It is worth mentioning that, assuming for simplicity that $s \in S$, (8.20) still holds if we consider the relative empirical risk $\gamma_n(s) - \gamma_n(\widehat{s})$ instead of the expected loss $\ell(s, \widehat{s})$. This is indeed a by-product of the proof of Theorem 8.3 below.

Proof. According to measurability condition (**M**), we may without loss of generality assume S to be countable. Suppose, first, for the sake of simplicity that there exists some point $\pi(s)$ belonging to S such that

$$\ell(s, \pi(s)) = \ell(s, S). \tag{8.21}$$

We start from the identity

$$\ell(s, \widehat{s}) = \ell(s, \pi(s)) + \gamma_n(\widehat{s}) - \gamma_n(\pi(s)) + \overline{\gamma}_n(\pi(s)) - \overline{\gamma}_n(\widehat{s}),$$

which, by definition of $\widehat{s}$ implies that

$$\ell(s, \widehat{s}) \leq \rho + \ell(s, \pi(s)) + \overline{\gamma}_n(\pi(s)) - \overline{\gamma}_n(\widehat{s}).$$

Let $x > 0$ with

$$x^2 = \kappa \left(\varepsilon_*^2 + \frac{\left(1 \wedge w^2(\varepsilon_*)\right) y}{n \varepsilon_*^2} \right),$$

where κ is a constant to be chosen later such that $\kappa \geq 1$, and

$$V_x = \sup_{t \in S} \frac{\overline{\gamma}_n(\pi(s)) - \overline{\gamma}_n(t)}{\ell(s,t) + x^2}.$$

Then,

$$\ell(s, \widehat{s}) \leq \rho + \ell(s, \pi(s)) + V_x \left(\ell(s, \widehat{s}) + x^2 \right)$$

and therefore, on the event $V_x < 1/2$, one has

$$\ell(s, \widehat{s}) < 2 \left(\rho + \ell(s, \pi(s)) \right) + x^2$$

yielding

$$\mathbb{P}\left[\ell\left(s,\widehat{s}\right)\geq 2\left(\rho+\ell\left(s,\pi\left(s\right)\right)\right)+x^2\right]\leq\mathbb{P}\left[V_x\geq\frac{1}{2}\right]. \tag{8.22}$$

Since ℓ is bounded by 1, we may always assume x (and thus ε_*) to be not larger than 1. Assuming that $x\leq 1$, it remains to control the variable V_x via Bousquet's inequality. By (8.16), (8.17), the definition of $\pi\left(s\right)$ and the monotonicity of w, we derive that for every $t\in S$

$$\mathrm{Var}_P\left(-\gamma\left(t,.\right)+\gamma\left(\pi\left(s\right),.\right)\right)\leq\left(d\left(s,t\right)+d\left(s,\pi\left(s\right)\right)\right)^2\leq 4w^2\left(\sqrt{\ell\left(s,t\right)}\right).$$

Hence, since γ takes its values in $[0,1]$, introducing the function $w_1=1\wedge 2w$, we derive that

$$\sup_{t\in S}\mathrm{Var}_P\left[\frac{\gamma\left(t,.\right)-\gamma\left(\pi\left(s\right),.\right)}{\ell\left(s,t\right)+x^2}\right]\leq\sup_{\varepsilon\geq 0}\frac{w_1^2\left(\varepsilon\right)}{\left(\varepsilon^2+x^2\right)^2}\leq\frac{1}{x^2}\sup_{\varepsilon\geq 0}\left(\frac{w_1\left(\varepsilon\right)}{\varepsilon\vee x}\right)^2.$$

Now the monotonicity assumptions on w imply that either $w\left(\varepsilon\right)\leq w\left(x\right)$ if $x\geq\varepsilon$ or $w\left(\varepsilon\right)/\varepsilon\leq w\left(x\right)/x$ if $x\leq\varepsilon$, hence one has in any case $w\left(\varepsilon\right)/\left(\varepsilon\vee x\right)\leq w\left(x\right)/x$ which finally yields

$$\sup_{t\in S}\mathrm{Var}_P\left[\frac{\gamma\left(t,.\right)-\gamma\left(\pi\left(s\right),.\right)}{\ell\left(s,t\right)+x^2}\right]\leq\frac{w_1^2\left(x\right)}{x^4}.$$

On the other hand since γ takes its values in $[0,1]$, we have

$$\sup_{t\in S}\left\|\frac{\gamma\left(t,.\right)-\gamma\left(\pi\left(s\right),.\right)}{\ell\left(s,t\right)+x^2}\right\|_\infty\leq\frac{1}{x^2}.$$

We can therefore apply (5.49) with $v=w_1^2\left(x\right)x^{-4}$ and $b=2x^{-2}$, which gives that, on a set Ω_y with probability larger than $1-\exp\left(-y\right)$, the following inequality holds

$$\begin{aligned}V_x&<\mathbb{E}\left[V_x\right]+\sqrt{\frac{2\left(w_1^2\left(x\right)x^{-2}+4\mathbb{E}\left[V_x\right]\right)y}{nx^2}}+\frac{y}{nx^2}\\&<3\mathbb{E}\left[V_x\right]+\sqrt{\frac{2w_1^2\left(x\right)x^{-2}y}{nx^2}}+\frac{2y}{nx^2}.\end{aligned} \tag{8.23}$$

Now since ε_* is assumed to be not larger than 1, one has $w\left(\varepsilon_*\right)\geq\varepsilon_*$ and therefore for every $\sigma\geq w\left(\varepsilon_*\right)$, the following inequality derives from the definition of ε_* by monotonicity

$$\frac{\phi\left(\sigma\right)}{\sigma^2}\leq\frac{\phi\left(w\left(\varepsilon_*\right)\right)}{w^2\left(\varepsilon_*\right)}\leq\frac{\phi\left(w\left(\varepsilon_*\right)\right)}{\varepsilon_*^2}=\sqrt{n}.$$

Hence (8.18) holds for every $\sigma\geq w\left(\varepsilon_*\right)$ and since $u\to\phi\left(2w\left(u\right)\right)/u$ is nonincreasing, by assumption (8.17) and (8.21) we can use Lemma 4.23 (and the triangle inequality for d) to get

$$\mathbb{E}\left[V_x\right]\leq\frac{4\phi\left(2w\left(x\right)\right)}{\sqrt{n}x^2}.$$

Hence, by monotonicity of $u \to \phi(u)/u$

$$\mathbb{E}[V_x] \leq \frac{8\phi(w(x))}{\sqrt{n}x^2}.$$

Since $\kappa \geq 1$ we note that $x \geq \sqrt{\kappa}\varepsilon_* \geq \varepsilon_*$. Thus, using the monotonicity of $u \to \phi(w(u))/u$, and the definition of ε_*, we derive that

$$\mathbb{E}[V_x] \leq \frac{8\phi(w(\varepsilon_*))}{\sqrt{n}x\varepsilon_*} = \frac{8\varepsilon_*}{x} \leq \frac{8}{\sqrt{\kappa}}. \tag{8.24}$$

Now, the monotonicity of $u \to w_1(u)/u$ implies that

$$\frac{w_1^2(x)}{x^2} \leq \frac{w_1^2(\varepsilon_*)}{\varepsilon_*^2}. \tag{8.25}$$

Plugging (8.24) and (8.25) into (8.23) implies that, on the set Ω_y,

$$V_x < \frac{24}{\sqrt{\kappa}} + \sqrt{\frac{2w_1^2(\varepsilon_*)\varepsilon_*^{-2}y}{nx^2}} + \frac{2y}{nx^2}.$$

Recalling that $\varepsilon_* \leq 1$, it remains to use the lower bound $4nx^2 \geq \kappa w_1^2(\varepsilon_*)\varepsilon_*^{-2}y$, noticing that $w_1^2(\varepsilon_*)\varepsilon_*^{-2} \geq 1$ to derive that, on the set Ω_y, the following inequality holds

$$V_x < \frac{24}{\sqrt{\kappa}} + \sqrt{\frac{8}{\kappa}} + \frac{8}{\kappa}.$$

Hence, choosing κ as a large enough numerical constant warrants that $V_x < 1/2$ on Ω_y. Thus

$$\mathbb{P}\left[V_x \geq \frac{1}{2}\right] \leq \mathbb{P}\left[\Omega_y^c\right] \leq e^{-y}, \tag{8.26}$$

and therefore (8.22) leads to

$$\mathbb{P}\left[\ell(s,\widehat{s}) \geq 2(\rho + \ell(s,\pi(s))) + x^2\right] \leq e^{-y}.$$

If a point $\pi(s)$ satisfying (8.21) does not exist we can use as well some point $\pi(s)$ satisfying $\ell(s,\pi(s)) \leq \ell(s,S)+\delta$ and get the required probability bound (8.20) by letting δ tend to zero. But since $\phi(u)/u \geq \phi(1) \geq 1$ for every $u \in [0,1]$, we derive from (8.19) and the monotonicity of ϕ and $u \to \phi(u)/u$ that

$$\frac{1 \wedge w^2(\varepsilon_*)}{\varepsilon_*^2} \leq \frac{\phi^2(1 \wedge w(\varepsilon_*))}{\varepsilon_*^2} \leq \frac{\phi^2(w(\varepsilon_*))}{\varepsilon_*^2}$$

and therefore

$$\frac{1 \wedge w^2(\varepsilon_*)}{\varepsilon_*^2} \leq n\varepsilon_*^2. \tag{8.27}$$

The proof can then be easily completed by integrating the tail bound (8.20) to get

$$\mathbb{E}\left[\ell\left(s,\widehat{s}\right)\right] \leq 2\left(\rho + \ell\left(s,S\right)\right) + \kappa\varepsilon_*^2 + \kappa\frac{1 \wedge w^2\left(\varepsilon_*\right)}{n\varepsilon_*^2}.$$

yielding the required upper bound on the expected risk via (8.27). ■

Even though the main motivation for Theorem 8.3 is the study of classification, it can also be easily applied to bounded regression. We begin the illustration of Theorem 8.3 with this framework which is more elementary than classification since in this case there is a clear connection between the expected loss and the variance of the increments.

8.3.2 Application to Bounded Regression

In this setting, the regression function $\eta : x \to \mathbb{E}\left[Y \mid X = x\right]$ is the target to be estimated, so that here $s = \eta$. We recall that for this framework we can take d to be the $\mathbb{L}_2\left(\mu\right)$ distance times $\sqrt{2}$. The connection between the loss function ℓ and d is especially simple in this case since

$$\left[\gamma\left(t,\left(x,y\right)\right) - \gamma\left(s,\left(x,y\right)\right)\right] = \left[-t\left(x\right) + s\left(x\right)\right]\left[2\left(y - s\left(x\right)\right) - t\left(x\right) + s\left(x\right)\right]$$

which implies since $\mathbb{E}\left[Y - s\left(X\right) \mid X\right] = 0$ that

$$\ell\left(s,t\right) = \mathbb{E}\left[\gamma\left(t,\left(X,Y\right)\right) - \gamma\left(s,\left(X,Y\right)\right)\right] = \mathbb{E}\left(t\left(X\right) - s\left(X\right)\right)^2.$$

Hence $2\ell\left(s,t\right) = d^2\left(s,t\right)$ and in this case the modulus of continuity w can simply be taken as $w\left(\varepsilon\right) = \sqrt{2}\varepsilon$. Note also that in this case, an empirical risk minimizer $\widehat{s}$ over some model S is a LSE. The quadratic risk of $\widehat{s}$ depends only on the modulus of continuity ϕ satisfying (8.18) and one derives from Theorem 8.3 that, for some absolute constant κ',

$$\mathbb{E}\left[d^2\left(s,\widehat{s}\right)\right] \leq 2d^2\left(s,S\right) + \kappa'\varepsilon_*^2$$

where ε_* is the solution of

$$\sqrt{n}\varepsilon_*^2 = \phi\left(\varepsilon_*\right).$$

To be more concrete, let us give an example where this modulus ϕ and the bias term $d^2\left(s,S\right)$ can be evaluated, leading to an upper bound for the minimax risk over some classes of regression functions.

Binary Images

Following [72], our purpose is to study the particular regression framework for which the variables X_i are uniformly distributed on $\left[0,1\right]^2$ and $s\left(x\right) = \mathbb{E}\left[Y \mid X = x\right]$ is of the form

$$s\left(x_1,x_2\right) = b \text{ if } x_2 \leq \partial s\left(x_1\right) \text{ and } a \text{ otherwise,}$$

where ∂s is some measurable map from $\left[0,1\right]$ to $\left[0,1\right]$ and $0 < a < b < 1$. The function ∂s should be understood as the parametrization of a boundary

fragment corresponding to some portion s of a binary image in the plane (a and b, representing the two level of colors which are taken by the image) and restoring this portion of the image from the noisy data $(X_1, Y_1), \ldots, (X_n, Y_n)$ means estimating s or equivalently ∂s. Let $\mathcal{G}$ be the set of measurable maps from $[0,1]$ to $[0,1]$. For any $f \in \mathcal{G}$, let us denote by χ_f the function defined on $[0,1]^2$ by

$$\chi_f(x_1, x_2) = b \text{ if } x_2 \leq f(x_1) \text{ and } a \text{ otherwise.}$$

From this definition we see that $\chi_{\partial s} = s$ and more generally if we define $\mathcal{S} = \{\chi_f : f \in \mathcal{G}\}$, for every $t \in \mathcal{S}$, we denote by ∂t the element of $\mathcal{G}$ such that $\chi_{\partial t} = t$. It is natural to consider here as an approximate model for s a model S of the form $S = \{\chi_f : f \in \partial S\}$, where ∂S denotes some subset of $\mathcal{G}$. In what follows, we shall assume condition (**M**) to be fulfilled which allows us to make as if S was countable. Denoting by $\|.\|_1$ (resp. $\|.\|_2$) the Lebesgue $\mathbb{L}_1$-norm (resp. $\mathbb{L}_2$-norm), one has for every $f, g \in \mathcal{G}$

$$\|\chi_f - \chi_g\|_1 = (b-a)\|f-g\|_1 \text{ and } \|\chi_f - \chi_g\|_2^2 = (b-a)^2 \|f-g\|_1$$

or equivalently for every $s, t \in \mathcal{S}$,

$$\|s-t\|_1 = (b-a)\|\partial s - \partial t\|_1 \text{ and } \|s-t\|_2^2 = (b-a)^2 \|\partial s - \partial t\|_1 .$$

Given $u \in S$ and setting $S_\sigma = \{t \in S,\, d(t,u) \leq \sigma\}$, we have to compute some function ϕ fulfilling (8.18) and therefore to upper bound $\mathbb{E}[W(\sigma)]$, where

$$W(\sigma) = \sup_{t \in S_\sigma} \overline{\gamma}_n(u) - \overline{\gamma}_n(t) .$$

This can be done using entropy with bracketing arguments. Indeed, let us notice that if g belongs to some ball with radius δ in $\mathbb{L}_\infty[0,1]$, then for some function $f \in \mathbb{L}_\infty[0,1]$, one has $f - \delta \leq g \leq f + \delta$ and therefore, defining $f_L = \sup(f - \delta, 0)$ and $f_U = \inf(f + \delta, 1)$

$$\chi_{f_L} \leq \chi_g \leq \chi_{f_U}$$

with $\|\chi_{f_L} - \chi_{f_U}\|_1 \leq 2(b-a)\delta$. This means that, defining $H_\infty(\delta, \partial S, \rho)$ as the supremum over $g \in \partial S$ of the $\mathbb{L}_\infty$-metric entropy for radius δ of the $\mathbb{L}_1$ ball centered at g with radius ρ in ∂S, one has for every positive ε

$$H_{[.]}(\varepsilon, S_\sigma, \mu) \leq H_\infty\left(\frac{\varepsilon}{2(b-a)}, \partial S, \frac{\sigma^2}{2(b-a)^2}\right).$$

Moreover if $[t_L, t_U]$ is a bracket with extremities in $\mathcal{S}$ and $\mathbb{L}_1(\mu)$ diameter not larger than δ and if $t \in [t_L, t_U]$, then

$$y^2 - 2t_U(x)\, y + t_L^2(x) \leq (y - t(x))^2 \leq y^2 - 2t_L(x)\, y + t_U^2(x) ,$$

which implies that $\gamma(t,.)$ belongs to a bracket with $\mathbb{L}_1(P)$-diameter not larger than

$$2\mathbb{E}\left[(t_U(X)-t_L(X))\left(Y+\frac{t_U(X)+t_L(X)}{2}\right)\right]\leq 4\delta.$$

Hence, if $\mathcal{F}=\{\gamma(t,.),t\in S \text{ and } d(t,u)\leq\sigma\}$, then

$$H_{[.]}(x,\mathcal{F},P)\leq H_\infty\left(\frac{x}{8(b-a)},\partial S,\frac{\sigma^2}{2(b-a)^2}\right)$$

and furthermore, if $d(t,u)\leq\sigma$

$$\mathbb{E}\left[\left|(Y-t(X))^2-(Y-u(X))^2\right|\right]\leq 2\|u-t\|_1=\frac{2\|u-t\|_2^2}{(b-a)}\leq\frac{\sigma^2}{(b-a)}.$$

We can therefore apply Lemma 6.5 to the class $\mathcal{F}$ and derive that, setting

$$\varphi(\sigma)=\int_0^{\sigma/\sqrt{b-a}}\left(H_\infty\left(\frac{x^2}{8(b-a)},\partial S,\frac{\sigma^2}{2(b-a)^2}\right)\right)^{1/2}dx,$$

one has

$$\sqrt{n}\mathbb{E}[W(\sigma)]\leq 12\varphi(\sigma),$$

provided that

$$4\varphi(\sigma)\leq\sqrt{n}\frac{\sigma^2}{(b-a)}. \tag{8.28}$$

The point now is that, whenever ∂S is part of a linear finite dimensional subspace of $\mathbb{L}_\infty[0,1]$, $H_\infty(\delta,\partial S,\rho)$ is typically bounded by $D[B+\ln(\rho/\delta)]$ for some appropriate constants D and B. If it is so then

$$\begin{aligned}\varphi(\sigma)&\leq\sqrt{D}\int_0^{\sigma/\sqrt{b-a}}\left(B+\ln\left(\frac{4\sigma^2}{x^2(b-a)}\right)\right)^{1/2}dx\\&\leq\frac{\sqrt{D}\sigma}{\sqrt{b-a}}\int_0^1\sqrt{B+2|\ln(2\delta)|}d\delta,\end{aligned}$$

which implies that for some absolute constant κ

$$\varphi(\sigma)\leq\kappa\sigma\sqrt{\frac{(1+B)D}{(b-a)}}.$$

The restriction (8.28) is a fortiori satisfied if $\sigma\sqrt{b-a}\geq 4\kappa\sqrt{(1+B)D/n}$. Hence if we take

$$\phi(\sigma)=12\kappa\sigma\sqrt{\frac{(1+B)D}{(b-a)}},$$

assumption (8.18) is satisfied. To be more concrete let us consider the example where ∂S is taken to be the set of piecewise constant functions on a regular partition with D pieces on $[0,1]$ with values in $[0,1]$. Then, it is shown in [12] that

$$H_\infty(\delta, \partial S, \rho) \leq D\left[\ln(\rho/\delta)\right]$$

and therefore the previous analysis can be used with $B = 0$. As a matter of fact this extends to piecewise polynomials with degree not larger than r via some adequate choice of B as a function of r but we just consider the histogram case here to be simple. As a conclusion, Theorem 8.3 yields in this case for the LSE $\widehat{s}$ over S

$$\mathbb{E}\left[\|\partial s - \partial\widehat{s}\|_1\right] \leq 2\inf_{t\in S}\|\partial s - \partial t\|_1 + C\frac{D}{(b-a)^3 n}$$

for some absolute constant C. In particular, if ∂s is Hölder smooth,

$$|\partial s(x) - \partial s(x')| \leq R|x - x'|^\alpha \tag{8.29}$$

with $R > 0$ and $\alpha \in (0,1]$, then

$$\inf_{t\in S}\|\partial s - \partial t\|_1 \leq RD^{-\alpha}$$

leading to

$$\mathbb{E}\left[\|\partial s - \partial\widehat{s}\|_1\right] \leq 2RD^{-\alpha} + C\frac{D}{(b-a)^3 n}.$$

Hence, if $\mathcal{H}(R,\alpha)$ denotes the set of functions from $[0,1]$ to $[0,1]$ satisfying (8.29), an adequate choice of D yields for some constant C' depending only on a and b

$$\sup_{\partial s\in\mathcal{H}(R,\alpha)}\mathbb{E}\left[\|\partial s - \partial\widehat{s}\|_1\right] \leq C'\left(R \vee \frac{1}{n}\right)^{\frac{1}{\alpha+1}} n^{-\frac{\alpha}{1+\alpha}}.$$

As a matter of fact, this upper bound is unimprovable (up to constants) from a minimax point of view (see [72] for the corresponding minimax lower bound).

8.3.3 Application to Classification

Our purpose is to apply Theorem 8.3 to the classification setting, assuming that the Bayes classifier is the target to be estimated, so that here $s = s^*$. We recall that for this framework we can take d to be the $\mathbb{L}_2(\mu)$ distance and $S = \{\mathbb{1}_A, A \in \mathcal{A}\}$, where $\mathcal{A}$ is some class of measurable sets. Our main task is to compute the moduli of continuity ϕ and w. In order to evaluate w, we need some margin type condition. For instance we can use Tsybakov's margin condition

$$\ell(s,t) \geq h^\theta d^{2\theta}(s,t), \text{ for every } t \in S, \tag{8.30}$$

where h is some positive constant (that we can assume to be smaller than 1 since we can always change h into $h \wedge 1$ without violating (8.30))and $\theta \geq 1$. As explained by Tsybakov in [115], this condition is fulfilled if the distribution of $\eta(X)$ is well behaved around $1/2$. A simple situation is the following. Assume that, for some positive number h, one has for every $x \in \mathcal{X}$

$$|2\eta(x) - 1| \geq h. \tag{8.31}$$

Then

$$\ell(s,t) = \mathbb{E}\left[|2\eta(X) - 1| \, |s(X) - t(X)|\right] \geq h d^2(s,t)$$

which means that Tsybakov's condition is satisfied with $\theta = 1$. Of course, Tsybakov's condition implies that the modulus of continuity w can be taken as

$$w(\varepsilon) = h^{-1/2} \varepsilon^{1/\theta}. \tag{8.32}$$

In order to evaluate ϕ, we shall consider two different kinds of assumptions on S which are well known to imply the Donsker property for the class of functions $\{\gamma(t,.), t \in S\}$ and therefore the existence of a modulus ϕ which tends to 0 at 0, namely a VC-condition or an entropy with bracketing assumption. Given $u \in S$, in order to bound the expectation of

$$W(\sigma) = \sup_{t \in S; d(u,t) \leq \sigma} \left(-\overline{\gamma}_n(t) + \overline{\gamma}_n(u)\right),$$

we shall use the maximal inequalities for empirical processes which are established in Chapter 6 via slightly different techniques according to the way the "size" of the class $\mathcal{A}$ is measured.

The VC-Case

Let us first assume for the sake of simplicity that $\mathcal{A}$ is countable. We use the definitions, notations and results of Section 6.1.2, to express ϕ in terms of the random combinatorial entropy or the VC-dimension of $\mathcal{A}$. Indeed, we introduce the classes of sets

$$\mathcal{A}_+ = \left\{\left\{(x,y) : \mathbb{1}_{y \neq t(x)} \leq \mathbb{1}_{y \neq u(x)}\right\}, t \in S\right\}$$

and

$$\mathcal{A}_- = \left\{\left\{(x,y) : \mathbb{1}_{y \neq t(x)} \geq \mathbb{1}_{y \neq u(x)}\right\}, t \in S\right\}$$

and define for every class of sets $\mathcal{B}$ of $\mathcal{X} \times \{0,1\}$

$$W_{\mathcal{B}}^+(\sigma) = \sup_{B \in \mathcal{B}, P(B) \leq \sigma^2} (P_n - P)(B), \; W_{\mathcal{B}}^-(\sigma) = \sup_{B \in \mathcal{B}, P(B) \leq \sigma^2} (P - P_n)(B).$$

Then,

$$\mathbb{E}\left[W(\sigma)\right] \leq \mathbb{E}\left[W_{\mathcal{A}_+}^+(\sigma)\right] + \mathbb{E}\left[W_{\mathcal{A}_-}^-(\sigma)\right] \tag{8.33}$$

and it remains to control $\mathbb{E}\left[W_{\mathcal{A}_+}^+(\sigma)\right]$ and $\mathbb{E}\left[W_{\mathcal{A}_-}^-(\sigma)\right]$ via Lemma 6.4.

Since the VC-dimension of $\mathcal{A}_+$ and $\mathcal{A}_-$ are not larger than that of $\mathcal{A}$, and that similarly, the combinatorial entropies of $\mathcal{A}_+$ and $\mathcal{A}_-$ are not larger than the combinatorial entropy of $\mathcal{A}$, denoting by $V_{\mathcal{A}}$ the VC-dimension of $\mathcal{A}$ (assuming that $V_{\mathcal{A}} \geq 1$), we derive from (8.33) and Lemma 6.4 that

$$\sqrt{n}\mathbb{E}\left[W\left(\sigma\right)\right] \leq \phi\left(\sigma\right)$$

provided that $\phi\left(\sigma\right) \leq \sqrt{n}\sigma^2$, where ϕ can be taken either as

$$\phi\left(\sigma\right) = K\sigma\sqrt{\left(1 \vee \mathbb{E}\left[H_{\mathcal{A}}\right]\right)} \tag{8.34}$$

or as

$$\phi\left(\sigma\right) = K\sigma\sqrt{V\left(1 + \ln\left(\sigma^{-1} \vee 1\right)\right)}. \tag{8.35}$$

In both cases, assumption (8.18) is satisfied and we can apply Theorem 8.3 with $w \equiv 1$ or w defined by (8.32). When ϕ is given by (8.34) the solution ε_* of equation (8.19) can be explicitly computed when w is given by (8.32) or $w \equiv 1$. Hence the conclusion of Theorem 8.3 holds with

$$\varepsilon_*^2 = \left(\frac{K^2\left(1 \vee \mathbb{E}\left[H_{\mathcal{A}}\right]\right)}{nh}\right)^{\theta/(2\theta-1)} \wedge \sqrt{\frac{K^2\left(1 \vee \mathbb{E}\left[H_{\mathcal{A}}\right]\right)}{n}}.$$

In the second case i.e., when ϕ is given by (8.35), $w \equiv 1$ implies by (8.19) that

$$\varepsilon_*^2 = K\sqrt{\frac{V}{n}}$$

while if $w\left(\varepsilon_*\right) = h^{-1/2}\varepsilon_*^{1/\theta}$ then

$$\varepsilon_*^2 = K\varepsilon_*^{1/\theta}\sqrt{\frac{V}{nh}}\sqrt{1 + \ln\left(\left(\sqrt{h}\varepsilon_*^{-1/\theta}\right) \vee 1\right)}. \tag{8.36}$$

Since $1 + \ln\left(\left(\sqrt{h}\varepsilon_*^{-1/\theta}\right) \vee 1\right) \geq 1$ and $K \geq 1$, we derive from (8.36) that

$$\varepsilon_*^2 \geq \left(\frac{V}{nh}\right)^{\theta/(2\theta-1)}. \tag{8.37}$$

Plugging this inequality in the logarithmic factor of (8.36) yields

$$\varepsilon_*^2 \leq K\varepsilon_*^{1/\theta}\sqrt{\frac{V}{nh}}\sqrt{1 + \frac{1}{2\left(2\theta-1\right)}\ln\left(\left(\frac{nh^{2\theta}}{V}\right) \vee 1\right)}$$

and therefore, since $\theta \geq 1$

$$\varepsilon_*^2 \leq K\varepsilon_*^{1/\theta}\sqrt{\frac{V}{nh}}\sqrt{1 + \ln\left(\left(\frac{nh^{2\theta}}{V}\right) \vee 1\right)}.$$

Hence

$$\varepsilon_*^2 \leq \left(\frac{K^2 V \left(1 + \ln\left(\left(nh^{2\theta}/V\right) \vee 1\right)\right)}{nh} \right)^{\theta/(2\theta-1)}$$
$$\leq K^2 \left(\frac{V \left(1 + \ln\left(\left(nh^{2\theta}/V\right) \vee 1\right)\right)}{nh} \right)^{\theta/(2\theta-1)}$$

and therefore the conclusion of Theorem 8.3 holds with

$$\varepsilon_*^2 = K^2 \left[\left(\frac{V \left(1 + \ln\left(\left(nh^{2\theta}/V\right) \vee 1\right)\right)}{nh} \right)^{\theta/(2\theta-1)} \wedge \sqrt{\frac{V}{n}} \right].$$

Of course, if S (and therefore $\mathcal{A}$) is not countable but fulfills condition (**M**), the previous arguments still apply for a conveniently countable subclass of $\mathcal{A}$ so that we have a fortiori obtained the following result.

Corollary 8.4 *Let $\mathcal{A}$ be a VC-class with dimension $V \geq 1$ and assume that s^* belongs to $S = \{\mathbb{1}_A, A \in \mathcal{A}\}$. Assuming that S satisfies to (**M**), there exists an absolute constant C such that if $\widehat{s}$ denotes an empirical risk minimizer over S, the following inequality holds*

$$\mathbb{E}\left[\ell\left(s^*, \widehat{s}\right)\right] \leq C \sqrt{\frac{V \wedge \left(1 \vee \mathbb{E}\left[H_{\mathcal{A}}\right]\right)}{n}}. \tag{8.38}$$

Moreover if $\theta \geq 1$ is given and one assumes that the margin condition (8.30) holds, then the following inequalities are also available

$$\mathbb{E}\left[\ell\left(s^*, \widehat{s}\right)\right] \leq C \left(\frac{\left(1 \vee \mathbb{E}\left[H_{\mathcal{A}}\right]\right)}{nh} \right)^{\theta/(2\theta-1)} \tag{8.39}$$

and

$$\mathbb{E}\left[\ell\left(s^*, \widehat{s}\right)\right] \leq C \left(\frac{V \left(1 + \ln\left(nh^{2\theta}/V\right)\right)}{nh} \right)^{\theta/(2\theta-1)}, \tag{8.40}$$

provided that $h \geq (V/n)^{1/2\theta}$.

Comments.

- The risk bound (8.38) is well known. Our purpose was just here to show how it can be derived from our approach.
- The risk bounds (8.39) and (8.40) are new and they perfectly fit with (8.38) when one considers the borderline case $h = (V/n)^{1/2\theta}$. They look very similar but are not strictly comparable since roughly speaking they differ from a logarithmic factor. Indeed it may happen that $\mathbb{E}\left[H_{\mathcal{A}}\right]$ turns out to be of the order of V (without any extra log-factor). This the case

when $\mathcal{A}$ is the family of all subsets of a given finite set with cardinality V. In such a case, $\mathbb{E}[H_{\mathcal{A}}] \leq V$ and (8.39) is sharper than (8.40). On the contrary, for some arbitrary VC-class, let us remember that the consequence (6.9) of Sauer's lemma tells us that $H_{\mathcal{A}} \leq V(1 + \ln(n/V))$. The logarithmic factor $1 + \ln(n/V)$ is larger than $1 + \ln\left(nh^{2\theta}/V\right)$ and turns out to be especially over pessimistic when h is close to the borderline value $(V/n)^{1/2\theta}$.

- For the sake of simplicity we have assumed s^* to belong to S in the above statement. Of course this assumption is not necessary (since our main Theorem does not require it). The price to pay if s^* does not belong to S is simply to add $2\ell(s^*, S)$ to the right hand side of the risk bounds above.

In [92] the optimality of (8.40) from a minimax point of view is discussed in the case where $\theta = 1$, showing that it is essentially unimprovable in that sense.

Bracketing Conditions

For the same reasons as in the previous section, let us make the preliminary assumption that S is countable (the final result will easily extend to the case where S satisfies (**M**) anyway). If t_1 and t_2 are measurable functions such that $t_1 \leq t_2$, the collection of measurable functions t such that $t_1 \leq t \leq t_2$ is denoted by $[t_1, t_2]$ and called bracket with lower extremity t_1 and upper extremity t_2. Recalling that μ denotes the probability distribution of the explanatory variable X, the $\mathbb{L}_1(\mu)$-diameter of a bracket $[t_1, t_2]$ is given by $\mu(t_2) - \mu(t_1)$. Recall that the $\mathbb{L}_1(\mu)$-entropy with bracketing of S is defined for every positive δ, as the logarithm of the minimal number of brackets with $\mathbb{L}_1(\mu)$-diameter not larger than δ which are needed to cover S and is denoted by $H_{[.]}(\delta, S, \mu)$. The point is that if $\mathcal{F}$ denotes the class of functions $\mathcal{F} = \{\gamma(t, .), t \in S \text{ with } d(u, t) \leq \sigma\}$, one has

$$H_{[.]}(\delta, \mathcal{F}, P) \leq H_{[.]}(\delta, S, \mu)$$

hence, we derive from (8.33) and Lemma 6.5 that, setting

$$\varphi(\sigma) = \int_0^\sigma H_{[.]}^{1/2}\left(x^2, S, \mu\right) dx,$$

the following inequality is available

$$\sqrt{n}\mathbb{E}[W(\sigma)] \leq 12\varphi(\sigma)$$

provided that $4\varphi(\sigma) \leq \sigma^2\sqrt{n}$. Hence, we can apply Theorem 8.3 with $\phi = 12\varphi$ and if we assume Tsybakov's margin condition (8.30) to be satisfied, then we can also take w according to (8.32) and derive from that the conclusions of Theorem 8.3 hold with ε_* solution of the equation

$$\sqrt{n}\varepsilon_*^2 = \phi\left(h^{-1/2}\varepsilon_*^{1/\theta}\right).$$

In particular, if

$$H_{[.]}(x, S, \mu) \leq Cx^{-r} \text{ with } 0 < r < 1, \tag{8.41}$$

then for some constant C' depending only on C, one has

$$\varepsilon_*^2 \leq C' \left[(1-r)^2 nh^{1-r} \right]^{-\frac{\theta}{2\theta-1+r}}. \tag{8.42}$$

If S' is taken as a δ_n-net (with respect to the $\mathbb{L}_2(\mu)$-distance d) of a bigger class S to which the target s^* is assumed to belong, then we can also apply Theorem 8.3 to the empirical risk minimizer over S' and since $H_{[.]}(x, S', \mu) \leq H_{[.]}(x, S, \mu)$, we still get the conclusions of Theorem 8.3 with ε_* satisfying (8.42) and $\ell(s^*, S') \leq \delta_n^2$. This means that if δ_n is conveniently chosen (in a way that δ_n is of lower order as compared to ε_*), for instance $\delta_n^2 = n^{-1/(1+r)}$, then, for some constant C'' depending only on C, one has

$$\mathbb{E}\left[\ell(s^*, \widehat{s})\right] \leq C'' \left[(1-r)^2 nh^{1-r} \right]^{-\frac{\theta}{2\theta-1+r}}.$$

This means that we have recovered Tsybakov's Theorem 1 in [115] (as a matter of fact our result is slightly more precise since it also provides the dependency of the risk bound with respect to the margin parameter h and not only on θ as in Tsybakov's Theorem). We refer to [85] for concrete examples of classes of sets with smooth boundaries satisfying (8.41) when μ is equivalent to the Lebesgue measure on some compact set of $\mathbb{R}^d$.

8.4 A Refined Model Selection Theorem

It is indeed quite easy to formally derive from (8.20) the following model selection version of Theorem 8.3.

Theorem 8.5 *Let $\xi_1, \ldots, \xi_n$ be independent observations taking their values in some measurable space Ξ and with common distribution P. Let $\mathcal{S}$ be some set, $\gamma : \mathcal{S} \times \Xi \to [0,1]$ be a measurable function such that for every $t \in \mathcal{S}$, $x \to \gamma(t,x)$ is measurable. Assume that there exists some minimizer s of $P(\gamma(t,.))$ over $\mathcal{S}$ and denote by $\ell(s,t)$ the nonnegative quantity $P(\gamma(t,.)) - P(\gamma(s,.))$ for every $t \in \mathcal{S}$. Let γ_n be the empirical risk*

$$\gamma_n(t) = P_n(\gamma(t,.)) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t, \xi_i), \text{ for every } t \in \mathcal{S}$$

and $\overline{\gamma}_n$ be the centered empirical process defined by

$$\overline{\gamma}_n(t) = P_n(\gamma(t,.)) - P(\gamma(t,.)), \text{ for every } t \in \mathcal{S}.$$

Let d be some pseudodistance on $\mathcal{S} \times \mathcal{S}$ such that (8.16) holds. Let $\{S_m\}_{m \in \mathcal{M}}$ be some at most countable collection of subsets of $\mathcal{S}$, each model S_m admitting

*some countable subset S'_m such that S_m satisfies to separability condition (**M**). Let w and ϕ_m belong to the class of functions $\mathcal{C}_1$ defined above for every $m \in \mathcal{M}$. Assume that on the one hand assumption (8.17) holds and that on the other hand one has for every $m \in \mathcal{M}$ and $u \in S'_m$*

$$\sqrt{n}\mathbb{E}\left[\sup_{t\in S'_m, d(u,t)\leq\sigma} [\overline{\gamma}_n(u) - \overline{\gamma}_n(t)]\right] \leq \phi_m(\sigma) \tag{8.43}$$

for every positive σ such that $\phi_m(\sigma) \leq \sqrt{n}\sigma^2$. Let ε_m be the unique solution of the equation

$$\sqrt{n}\varepsilon_m^2 = \phi_m(w(\varepsilon_m)). \tag{8.44}$$

Let ρ be some given nonnegative real number and consider $\widehat{s}_m$ taking its values in S_m and such that

$$\gamma_n(\widehat{s}_m) \leq \inf_{t\in S_m} \gamma_n(t) + \rho.$$

Let $\{x_m\}_{m\in\mathcal{M}}$ be some family of nonnegative weights such that

$$\sum_{m\in\mathcal{M}} e^{-x_m} = \Sigma < \infty,$$

$\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ *such that for every $m \in \mathcal{M}$*

$$\mathrm{pen}(m) \geq K\left(\varepsilon_m^2 + \frac{w^2(\varepsilon_m)x_m}{n\varepsilon_m^2}\right).$$

Then, if K is large enough, there almost surely exists some minimizer $\widehat{m}$ of

$$\gamma_n(\widehat{s}_m) + \mathrm{pen}(m). \tag{8.45}$$

and some constant $C(K)$ such that the penalized estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ satisfies the following inequality

$$\mathbb{E}[\ell(s,\widetilde{s})] \leq C(K)\left[\inf_{m\in\mathcal{M}}(\ell(s,S_m) + \mathrm{pen}(m)) + \frac{(\Sigma+1)}{n} + \rho\right].$$

Concerning the proof of Theorem 8.5, the hard work has been already done to derive (8.20). The proof of Theorem 8.5 can indeed be sketched as follows: start from exponential bound (8.26) (which as a matter of fact readily implies (8.20)) and use a union bound argument. The calculations are quite similar to those of the proof of Theorem 4.18. At this stage they can be considered as routine and we shall therefore skip them. From the point of view of model selection for classification, Theorem 8.5 is definitely disappointing and far from producing the result we could expect anyway. In this classification context, it should be considered as a formal exercise. Indeed, the classification framework was the main motivation for introducing the "margin" function w in Theorem 8.3. The major drawback of Theorem 8.5 is that the penalization procedure

involved *does require the knowledge* of w. Hence, apart from the situation where w can "legally" be assumed to be known (like for bounded regression where one can take $w(\varepsilon) = \sqrt{2}\varepsilon$), we cannot freely use Theorem 8.5 to build adaptive estimators as we did with the related model selection theorems in the other functional estimation frameworks that we have studied in the previous chapters (Gaussian white noise or density estimation). We shall come back to the classification framework in Section 8.5 below to design "margin adaptive" model selection strategies. For the moment we may at least use Theorem 8.5 in the bounded regression framework (note that more generally, when $w(\varepsilon) = C\varepsilon$ for a known absolute constant C, Theorem 8.5 is nothing more than Theorem 4.2. in [91]).

8.4.1 Application to Bounded Regression

As mentioned above, bounded regression is a typical framework for which the previous model selection theorem (Theorem 8.5) is relevant. Indeed, let us recall that in this setting, the regression function $\eta : x \to \mathbb{E}[Y \mid X = x]$ is the target s to be estimated and d may be taken as the $\mathbb{L}_2(\mu)$ distance times $\sqrt{2}$. The connection between the loss function ℓ and d is trivial since

$$\ell(s,t) = \mathbb{E}(t(X) - s(X))^2 = d^2(s,t)/2$$

and therefore w can simply be taken as $w(\varepsilon) = \sqrt{2}\varepsilon$. The penalized criterion given by (8.45) is a *penalized least squares criterion* and the corresponding penalized estimator $\widetilde{s}$ is merely a *penalized LSE* . It is not very difficult to study again the example of boundary images, showing this time that some adequate choice of the collection of models leads to adaptive properties for the penalized LSE on classes of binary images with smooth boundaries.

Binary Images

We consider the same framework as in Section 8.3.2, i.e., the variables X_i are uniformly distributed on $[0,1]^2$ and the regression function s is of the form

$$s(x_1,x_2) = b \text{ if } x_2 \leq \partial s(x_1) \text{ and } a \text{ otherwise,}$$

where ∂s is some measurable map from $[0,1]$ to $[0,1]$ and $0 < a < b < 1$. Let $\mathcal{G}$ be the set of measurable maps from $[0,1]$ to $[0,1]$ and, for any $f \in \mathcal{G}$, χ_f denotes the function defined on $[0,1]^2$ by

$$\chi_f(x_1,x_2) = b \text{ if } x_2 \leq f(x_1) \text{ and } a \text{ otherwise.}$$

Setting $\mathcal{S} = \{\chi_f : f \in \mathcal{G}\}$, for every $t \in \mathcal{S}$, ∂t denotes the element of $\mathcal{G}$ such that $\chi_{\partial t} = t$. Consider for every positive integer m, ∂S_m to be the set of piecewise constant functions on a regular partition of $[0,1]$ by m intervals and define $S_m = \{\chi_f : f \in \partial S_m\}$. We take $\{S_m\}_{m \in \mathbb{N}^*}$ as a collection of models. In order

to apply Theorem 8.5, given $u \in S_m$, we need to upper bound $\mathbb{E}[W_m(\sigma)]$ where

$$W_m(\sigma) = \sup_{t \in S_m; d(u,t) \leq \sigma} \overline{\gamma}_n(u) - \overline{\gamma}_n(t).$$

We derive from the calculations of Section 8.3.2 that for some absolute numerical constant κ'

$$\sqrt{n}\mathbb{E}[W_m(\sigma)] \leq \kappa' \sigma \sqrt{\frac{m}{(b-a)}}$$

so that we can take

$$\phi_m(\sigma) = \kappa' \sigma \sqrt{\frac{m}{(b-a)}}.$$

Hence the solution ε_m of (8.44) is given by

$$\varepsilon_m^2 = \frac{2m\kappa'^2}{n(b-a)}.$$

Choosing $x_m = m$, leads to $\Sigma < 1$ and therefore, applying Theorem 8.5, we know that for some adequate numerical constants K' and C', one can take

$$\mathrm{pen}(m) = K' \frac{m}{n(b-a)}$$

and the resulting penalized LSE $\widetilde{s}$ satisfies to

$$\mathbb{E}[\|\partial s - \partial \widetilde{s}\|_1] \leq C' \inf_{m \geq 1} \left\{ \inf_{t \in S_m} \|\partial s - \partial t\|_1 + \frac{m}{(b-a)^3 n} \right\}.$$

Assuming now that ∂s is Hölder smooth

$$|\partial s(x) - \partial s(x')| \leq R|x - x'|^\alpha$$

with $R > 0$ and $\alpha \in (0, 1]$, then

$$\inf_{t \in S_m} \|\partial s - \partial t\|_1 \leq R m^{-\alpha},$$

leading to

$$\mathbb{E}[\|\partial s - \partial \widetilde{s}\|_1] \leq C' \inf_{m \geq 1} \left\{ R m^{-\alpha} + \frac{m}{(b-a)^3 n} \right\}.$$

Hence, provided that $R \geq 1/n$, optimizing this bound with respect to m implies that

$$\sup_{\partial s \in \mathcal{H}(R,\alpha)} \mathbb{E}[\|\partial s - \partial \widetilde{s}\|_1] \leq C' R^{\frac{1}{\alpha+1}} n^{-\frac{\alpha}{1+\alpha}}.$$

Taking into account that the minimax risk is indeed of order $R^{\frac{1}{\alpha+1}} n^{-\frac{\alpha}{1+\alpha}}$ according to [72], this proves that the penalized LSE $\widetilde{s}$ is adaptive on each of the Hölder classes $\mathcal{H}(R,\alpha)$ such that $R \geq 1/n$ and $\alpha \in (0,1]$. Of course, with a little more efforts, the same kind of results could be obtained with collections of piecewise polynomials with variable degree, leading to adaptive estimators on Hölder classes $\mathcal{H}(R,\alpha)$ such that $R \geq 1/n$, for any positive value of α.

Selecting Nets

We can try to mimic the discretization strategies that we have developed in Chapter 4 and Chapter 7. As compared to the density estimation problem for instance, there is at least one noticeable difference. Indeed for density estimation, the dominating probability measure μ is assumed to be known. Here the role of this dominating measure is played by the distribution of the explanatory variables X_is. For some specific problems it makes sense to assume that μ is known (as we did in the previous boundary images estimation problem above), but most of the time one cannot make such an assumption. In such a situation there are at least two possibilities to overcome this difficulty: use $\mathbb{L}_\infty$ nets or empirical nets based on the empirical distribution of the variables X_is. Even if the second approach is more general than the first one, it would lead us to use extra technicalities that we prefer to avoid here. Constructing $\mathbb{L}_\infty$ nets concretely means that if the variables X_is take their values in $\mathbb{R}^d$ for instance, one has to assume that they are compactly supported and that we know their support. Moreover it also means that we have in view to estimate a rather smooth regression function s. Let us first state a straightforward consequence of Theorem 8.5 when applied to the selection of finite models problem in the regression framework.

Corollary 8.6 *Let $\{S_m\}_{m\in\mathcal{M}}$ be some at most countable collection of models, where for each $m \in \mathcal{M}$, S_m is assumed to be a finite set of functions taking their values in $[0,1]$. We consider a corresponding collection $(\widehat{s}_m)_{m\in\mathcal{M}}$ of LSE, which means that for every $m \in \mathcal{M}$*

$$\sum_{i=1}^{n} \left(Y_i - \widehat{s}_m(X_i)\right)^2 = \inf_{t\in S_m} \sum_{i=1}^{n} \left(Y_i - t(X_i)\right)^2 .$$

Let $\{\Delta_m\}_{m\in\mathcal{M}}$, $\{x_m\}_{m\in\mathcal{M}}$ be some families of nonnegative numbers such that $\Delta_m \geq \ln(|S_m|)$ for every $m \in \mathcal{M}$ and

$$\sum_{m\in\mathcal{M}} e^{-x_m} = \Sigma < \infty.$$

Define

$$\mathrm{pen}(m) = \frac{\kappa''(\Delta_m + x_m)}{n} \text{ for every } m \in \mathcal{M} \tag{8.46}$$

for some suitable numerical constant κ''. Then, if κ'' is large enough, there almost surely exists some minimizer $\widehat{m}$ of

$$\sum_{i=1}^{n} \left(Y_i - \widehat{s}_m\left(X_i\right)\right)^2 + \operatorname{pen}(m)$$

over $\mathcal{M}$. Moreover, for such a minimizer, the following inequality is valid whatever the regression function s

$$\mathbb{E}_s\left[d^2\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq C''\left(\inf_{m\in\mathcal{M}}\left(d^2\left(s, S_m\right) + \frac{\kappa''\left(\Delta_m + x_m\right)}{n}\right) + \frac{(1+\Sigma)}{n}\right). \tag{8.47}$$

Proof. It suffices to apply Theorem 8.5 with $w(\varepsilon) = \sqrt{2}\varepsilon$ and for each model $m \in \mathcal{M}$, check that by (6.4) the function ϕ_m defined by

$$\phi_m(\sigma) = 2\sigma\sqrt{\Delta_m}$$

does satisfy to (8.43). The result easily follows. ■

Let us see what kind of result is achievable when working with nets by considering the same type of example as in the Gaussian or the density estimation frameworks. Let us consider some collection of compact subsets $\{\mathcal{S}_{\alpha,R}, \alpha \in \mathbb{N}^*, R > 0\}$ of $\mathcal{S}$ with the following structure:

$$\mathcal{S}_{\alpha,R} = \mathcal{S} \cap R\mathcal{H}_{\alpha,1},$$

where $\mathcal{H}_{\alpha,1}$ is star-shaped at 0 and satisfies for some positive constant $C_2(\alpha)$ to

$$H_\infty\left(\delta, \mathcal{H}_{\alpha,1}\right) \leq C_2(\alpha)\,\delta^{-1/\alpha}$$

for every $\delta \leq 1$. We consider for every positive integers α and k some $\mathbb{L}_\infty$-net $S_{\alpha,k}$ of $\mathcal{S}_{\alpha,k/\sqrt{n}}$ with radius $\delta_{\alpha,k} = k^{1/(2\alpha+1)}/\sqrt{n}$, so that

$$\ln\left|S_{\alpha,k}\right| \leq C_2(\alpha)\left(\frac{k}{\sqrt{n}\delta_{\alpha,k}}\right)^{1/\alpha} \leq C_2(\alpha)\,k^{2/(2\alpha+1)}$$

and

$$d^2\left(s, S_{\alpha,k}\right) \leq 2\delta_{\alpha,k}^2, \text{ for all } s \in \mathcal{S}_{\alpha,k/\sqrt{n}}. \tag{8.48}$$

Applying Corollary 8.6 to the collection $\left(S_{\alpha,k}\right)_{\alpha\geq 1,k\geq 1}$ with

$$\Delta_{\alpha,k} = C_2(\alpha)\,k^{2/(2\alpha+1)} \text{ and } x_{\alpha,k} = 4\alpha k^{2/(2\alpha+1)}$$

leads to the penalty

$$\operatorname{pen}(\alpha, k) = K(\alpha)\,k^{2/(2\alpha+1)}\varepsilon^2$$

where $K(\alpha) = \kappa''\left(C_2(\alpha) + 4\alpha\right)$. Noticing that $x_{\alpha,k} \geq \alpha + 2\ln(k)$ one has

$$\Sigma = \sum_{\alpha,k} e^{-x_{\alpha,k}} \leq \left(\sum_{\alpha \geq 1} e^{-\alpha} \right) \left(\sum_{k \geq 1} k^{-2} \right) < 1$$

and it follows from (8.47) that if $\widetilde{s}$ denotes the penalized LSE one has whatever the regression function s

$$\mathbb{E}_s \left[d^2 \left(s, \widetilde{s} \right) \right] \leq C \left(\alpha \right) \inf_{\alpha,k} \left(d^2 \left(s, S_{\alpha,k} \right) + \frac{k^{2/(2\alpha+1)}}{n} \right).$$

In particular if $s \in \mathcal{S}_{\alpha,R}$ for some integer α and some real number $R \geq 1/\sqrt{n}$, setting $k = \lceil R\sqrt{n} \rceil$ we have $s \in \mathcal{S}_{\alpha,k/\sqrt{n}}$ and since $S_{\alpha,k}$ is a $k^{1/(2\alpha+1)}\varepsilon$-net of $\mathcal{S}_{\alpha,k\varepsilon}$, the previous inequality implies via (8.48) that

$$\begin{aligned} \sup_{s \in \mathcal{S}_{\alpha,R}} \mathbb{E}_s \left[d^2 \left(s, \widetilde{s} \right) \right] &\leq 3C \left(\alpha \right) \left(\frac{k^{2/(2\alpha+1)}}{n} \right) \\ &\leq 3C \left(\alpha \right) 2^{2/(2\alpha+1)} \frac{\left(R\sqrt{n} \right)^{2/(2\alpha+1)}}{n}. \end{aligned}$$

Since d^2 is upper bounded by 2, we finally derive that for some constant $C' \left(\alpha \right) \geq 1$

$$\sup_{s \in \mathcal{S}_{\alpha,R}} \mathbb{E}_s \left[d^2 \left(s, \widetilde{s} \right) \right] \leq C' \left(\alpha \right) \left(\left(R^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \right) \wedge 1 \right). \tag{8.49}$$

If $\mathcal{H}_{\alpha,R}$ is the Hölder class $\mathcal{H} \left(\alpha, R \right)$ defined in Section 7.5.1, we have already used the following property

$$H_\infty \left(\delta, \mathcal{H} \left(\alpha, R \right) \right) \leq C_2 \left(\alpha \right) \left(\frac{R}{\delta} \right)^{1/\alpha}.$$

Hence the previous approach applies to this case. Of course, nothing warrants that the above upper bound for the risk is minimax for arbitrary probability measures μ. For Hölder classes , it would not be difficult to show that this is indeed the case provided that one restricts to probability measures μ which are absolutely continuous with respect to Lebesgue measure with density f satisfying $0 < a \leq f \leq b < \infty$, for given positive constants a and b.

8.5 Advanced Model Selection Problems

All along the preceding Chapters, we have focused on model selection via penalization. It is worth noticing however, that some much simpler procedure can be used if one is ready to split the data into two parts, using the first half of the original simple to build the collection of estimators on each model and the second half to select among the family. This is the so-called *hold-out.* It should

be seen as some primitive version of the V-fold cross-validation method which is commonly used in practice when one deals with i.i.d. data as it is the case in this Section. The advantage of hold-out is that it is very easy to study from a mathematical point of view. Of course it would be very interesting to derive similar results for V-fold cross-validation but we do not see how to do it for the moment.

8.5.1 Hold-Out as a Margin Adaptive Selection Procedure

Our purpose is here to show that the hold-out is a naturally margin adaptive selection procedure for classification. More generally, for i.i.d. data we wish to understand what is the performance of the hold-out as a model selection procedure. Our analysis will be based on the following abstract selection theorem among some family of functions $\{f_m, m \in \mathcal{M}\}$. The reason for introducing an auxiliary family of functions $\{g_m, m \in \mathcal{M}\}$ in the statement of Theorem 8.7 below will become clear in the section devoted to the study of MLEs. At first reading it is better to consider the simplest case where $g_m = f_m$ for every $m \in \mathcal{M}$, which is indeed enough to deal with the applications to bounded regression or classification that we have in view.

Theorem 8.7 *Let $\{f_m, m \in \mathcal{M}\}$ be some at most countable collection of real-valued measurable functions defined on some measurable space $\mathcal{X}$. Let $\xi_1, \ldots, \xi_n$ be some i.i.d. random variables with common distribution P and denote by P_n the empirical probability measure based on $\xi_1, \ldots, \xi_n$. Assume that for some family of positive numbers $\{\sigma_m, m \in \mathcal{M}\}$ and some positive constant c, one has for every integer $k \geq 2$*

$$P\left(|f_m - f_{m'}|^k\right) \leq \frac{k!}{2} c^{k-2} \left(\sigma_m + \sigma_{m'}\right)^2 \text{ for every } m \in \mathcal{M}, m' \in \mathcal{M}. \quad (8.50)$$

Assume furthermore that $P(f_m) \geq 0$ for every $m \in \mathcal{M}$ and let w be some nonnegative and nondecreasing continuous function on $\mathbb{R}_+$ such that $w(x)/x$ is nonincreasing on $(0, +\infty)$ and

$$\sigma_m \leq w\left(\sqrt{P(f_m)}\right) \text{ for every } m \in \mathcal{M}. \quad (8.51)$$

Let $\{g_m, m \in \mathcal{M}\}$ be a family of functions such that $f_m \leq g_m$ and $\{x_m\}_{m \in \mathcal{M}}$ some family of nonnegative numbers such that

$$\sum_{m \in \mathcal{M}} e^{-x_m} = \Sigma < \infty.$$

Let $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ and consider some random variable $\widehat{m}$ such that

$$P_n\left(g_{\widehat{m}}\right) + \mathrm{pen}(\widehat{m}) = \inf_{m \in \mathcal{M}} \left(P_n(g_m) + \mathrm{pen}(m)\right).$$

Define δ_* as the unique positive solution of the equation

$$w(\delta) = \sqrt{n}\delta^2$$

and suppose that for some constant $\theta \in (0,1)$

$$\mathrm{pen}(m) \geq x_m \left(\frac{\delta_*^2}{\theta} + \frac{c}{n} \right) \text{ for every } m \in \mathcal{M}. \tag{8.52}$$

Then, setting

$$\mathcal{R}_{\min} = \inf_{m \in \mathcal{M}} \left(P(g_m) + \mathrm{pen}(m) \right),$$

one has

$$(1-\theta)\,\mathbb{E}\left[P\left(f_{\widehat{m}} \right) \right] \leq (1+\theta)\,\mathcal{R}_{\min} + \delta_*^2 \left(2\theta + \Sigma\theta^{-1} \right) + \frac{c\Sigma}{n}. \tag{8.53}$$

Moreover, if $f_m = g_m$ *for every* $m \in \mathcal{M}$*, the following exponential bound holds for every positive real number* x

$$\mathbb{P}\left[(1-\theta)\, P\left(f_{\widehat{m}} \right) > (1+\theta)\,\mathcal{R}_{\min} + \delta_*^2 \left(2\theta + x\theta^{-1} \right) + \frac{cx}{n} \right] \leq \Sigma e^{-x}. \tag{8.54}$$

Proof. We may always assume that the infimum of $P(g_m) + \mathrm{pen}(m)$ is achieved on $\mathcal{M}$ (otherwise we can always take m_ε such that $P(g_{m_\varepsilon}) + \mathrm{pen}(m_\varepsilon) \leq \inf_{m \in \mathcal{M}} (P(g_m) + \mathrm{pen}(m)) + \varepsilon$ and make ε tend to 0 in the resulting tail bound). So let m such that

$$P(g_m) + \mathrm{pen}(m) = \inf_{m' \in \mathcal{M}} \left(P(g_{m'}) + \mathrm{pen}(m') \right).$$

Our aim is to prove that, except on a set of probability less than Σe^{-x}, one has

$$(1-\theta)\, P\left(f_{\widehat{m}} \right) + U_m \leq (1+\theta) \left(P(g_m) + \mathrm{pen}(m) \right) + \delta_*^2 \left(2\theta + x\theta^{-1} \right) + \frac{cx}{n}, \tag{8.55}$$

where $U_m = (P - P_n)(g_m - f_m)$. Noticing that U_m is centered at expectation and is equal to 0 whenever $f_m = g_m$, this will achieve the proof of Theorem 8.7. Indeed (8.55) leads to (8.53) by integrating with respect to x and (8.55) means exactly (8.54) whenever $f_m = g_m$. To prove (8.55), let us notice that by definition of $\widehat{m}$

$$P_n\left(g_{\widehat{m}} \right) + \mathrm{pen}(\widehat{m}) \leq P_n(g_m) + \mathrm{pen}(m),$$

hence, since $f_{\widehat{m}} \leq g_{\widehat{m}}$

$$\begin{aligned} P\left(f_{\widehat{m}} \right) &= (P - P_n)\left(f_{\widehat{m}} \right) + P_n\left(f_{\widehat{m}} \right) \\ &\leq P_n(g_m) + \mathrm{pen}(m) + (P - P_n)\left(f_{\widehat{m}} \right) - \mathrm{pen}(\widehat{m}) \end{aligned}$$

and therefore

$$P\left(f_{\widehat{m}} \right) + U_m \leq P(g_m) + \mathrm{pen}(m) + (P - P_n)\left(f_{\widehat{m}} - f_m \right) - \mathrm{pen}(\widehat{m}). \tag{8.56}$$

It comes from Bernstein's inequality that for every $m' \in \mathcal{M}$ and every positive number $y_{m'}$, the following inequality holds, except on a set with probability less than $e^{-y_{m'}}$

$$(P - P_n)(f_{m'} - f_m) \leq \sqrt{\frac{2y_{m'}}{n}}(\sigma_m + \sigma_{m'}) + \frac{cy_{m'}}{n}.$$

Choosing $y_{m'} = x_{m'} + x$ for every $m' \in \mathcal{M}$, this implies that, except on some set Ω_x with probability less than Σe^{-x},

$$(P - P_n)\left(f_{\widehat{m}} - f_m\right) \leq \sqrt{\frac{2y_{\widehat{m}}}{n}}\left(\sigma_m + \sigma_{\widehat{m}}\right) + \frac{cy_{\widehat{m}}}{n}. \tag{8.57}$$

If u is some nonnegative real number, we derive from the monotonicity assumptions on w that

$$w\left(\sqrt{u}\right) \leq w\left(\sqrt{u + \delta_*^2}\right) \leq \sqrt{u + \delta_*^2}\frac{w(\delta_*)}{\delta_*}.$$

Hence, for every positive number y, we get by definition of δ_*

$$\sqrt{\frac{2y}{n}}w\left(\sqrt{u}\right) \leq \theta\left(u + \delta_*^2\right) + \theta^{-1}\frac{yw^2(\delta_*)}{2n\delta_*^2} \leq \theta\left(u + \delta_*^2\right) + \frac{y\delta_*^2\theta^{-1}}{2}.$$

Using this inequality with $y = y_{\widehat{m}}$ and successively $u = P(f_m)$ and $u = P\left(f_{\widehat{m}}\right)$, it comes from (8.51) that

$$\sqrt{\frac{2y_{\widehat{m}}}{n}}\left(\sigma_m + \sigma_{\widehat{m}}\right) \leq \delta_*^2\left(2\theta + y_{\widehat{m}}\theta^{-1}\right) + \theta P(f_m) + \theta P\left(f_{\widehat{m}}\right).$$

Combining this inequality with (8.57) and (8.52) implies that, except on Ω_x

$$(P - P_n)\left(f_{\widehat{m}} - f_m\right) \leq \mathrm{pen}(\widehat{m}) + \delta_*^2\left(2\theta + x\theta^{-1}\right) + \frac{cx}{n} + \theta P(f_m) + \theta P\left(f_{\widehat{m}}\right).$$

Plugging this inequality in (8.56) yields since $f_m \leq g_m$

$$(1 - \theta)P\left(f_{\widehat{m}}\right) + U_m \leq (1 + \theta)P(g_m) + \mathrm{pen}(m) + \delta_*^2\left(2\theta + x\theta^{-1}\right) + \frac{cx}{n}$$

which a fortiori implies that (8.55) holds. ■

Theorem 8.7 has a maybe more easily understandable corollary directly orientated towards the hold-out procedure *without* penalization in statistical learning.

Corollary 8.8 *Let $\{f_m, m \in \mathcal{M}\}$ be some finite collection of real-valued measurable functions defined on some measurable space $\mathcal{X}$. Let $\xi_1, \ldots, \xi_n$ be some i.i.d. random variables with common distribution P and denote by P_n the empirical probability measure based on $\xi_1, \ldots, \xi_n$. Assume that $f_m - f_{m'} \leq 1$*

for every $m, m' \in \mathcal{M}$. *Assume furthermore that* $P(f_m) \geq 0$ *for every* $m \in \mathcal{M}$ *and let* w *be some nonnegative and nondecreasing continuous function on* $\mathbb{R}_+$ *such that* $w(x)/x$ *is nonincreasing on* $(0, +\infty)$, $w(1) \geq 1$ *and*

$$P\left(f_m^2\right) \leq w^2\left(\sqrt{P(f_m)}\right) \text{ for every } m \in \mathcal{M}. \tag{8.58}$$

Consider some random variable $\widehat{m}$ *such that*

$$P_n\left(f_{\widehat{m}}\right) = \inf_{m \in \mathcal{M}} P_n(f_m).$$

Define δ_* *as the unique positive solution of the equation*

$$w(\delta) = \sqrt{n}\delta^2.$$

Then, for every $\theta \in (0,1)$

$$(1-\theta)\,\mathbb{E}\left[P\left(f_{\widehat{m}}\right)\right] \leq (1+\theta) \inf_{m \in \mathcal{M}} P(f_m) + \delta_*^2\left(2\theta + \ln\left(e\,|\mathcal{M}|\right)\left(\frac{1}{3} + \theta^{-1}\right)\right). \tag{8.59}$$

Proof. Noticing that since $w(1) \geq 1$, one has $\delta_*^2 \geq 1/n$, we simply apply Theorem 8.7 with $c = 1/3$, $x_m = \ln(|\mathcal{M}|)$ and

$$\mathrm{pen}(m) = \delta_*^2 \ln(|\mathcal{M}|)\left(\theta^{-1} + (1/3)\right).$$

Since $\Sigma = 1$, (8.53) leads to (8.59). ■

Hold-Out for Bounded Contrasts

Let us consider again the empirical risk minimization procedure. Assume that we observe $N + n$ random variables with common distribution P depending on some parameter s to be estimated. The first N observations $\xi_1', \ldots, \xi_N'$ are used to build some preliminary collection of estimators $\{\widehat{s}_m\}_{m \in \mathcal{M}}$ and we use the remaining observations $\xi_1, \ldots, \xi_n$ to select some estimator $\widehat{s}_m$ among the collection $\{\widehat{s}_m\}_{m \in \mathcal{M}}$. We more precisely consider here the situation where there exists some (bounded) loss or contrast

$$\gamma \text{ from } \mathcal{S} \times \Xi \text{ to } [0,1]$$

which is well adapted to our estimation problem of s in the sense that the expected loss $P[\gamma(t,.)]$ achieves a minimum at point s when t varies in $\mathcal{S}$. We recall that the relative expected loss is defined by

$$\ell(s,t) = P[\gamma(t,.) - \gamma(s,.)], \text{ for all } t \in \mathcal{S}.$$

In the bounded regression or the classification cases, we have already seen that one can take

$$\gamma(t,(x,y)) = (y - t(x))^2$$

since η (resp. s^*) is indeed the minimizer of $\mathbb{E}\left[(Y - t(X))^2\right]$ over the set of measurable functions t taking their values in $[0,1]$ (resp.$\{0,1\}$). The idea is now to apply the results of the previous section conditionally on the training sample $\xi'_1, \ldots, \xi'_N$. For instance, we can apply Corollary 8.8 to the collection of functions $\{f_m = \gamma(\widehat{s}_m, .) - \gamma(s, .),\ m \in \mathcal{M}\}$. Let us consider the case where $\mathcal{M}$ is finite and define $\widehat{m}$ as a minimizer of the empirical risk $P_n(\gamma(\widehat{s}_m, .))$ over $\mathcal{M}$. If $w \in \mathcal{C}_1$ is such that for all $t \in \mathcal{S}$

$$P\left((\gamma(t,.) - \gamma(s,.))^2\right) \leq w^2\left(\sqrt{\ell(s,t)}\right),$$

we derive from (8.59) that conditionally on $\xi'_1, \ldots, \xi'_N$, one has for every $\theta \in (0,1)$

$$(1-\theta)\,\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right) \mid \xi'\right] \leq (1+\theta) \inf_{m \in \mathcal{M}} \ell(s, \widehat{s}_m) + \delta_*^2\left(2\theta + \ln(e\,|\mathcal{M}|)\left(\frac{1}{3} + \theta^{-1}\right)\right), \tag{8.60}$$

where δ_* satisfies to $\sqrt{n}\delta_*^2 = w(\delta_*)$. The striking feature of this result is that the hold-out selection procedure provides an oracle type inequality involving the modulus of continuity w which is not known in advance. This is especially interesting in the classification framework for which w can be of very different natures according to the difficulty of the classification problem. The main issue is therefore to understand whether the term $\delta_*^2(1 + \ln(|\mathcal{M}|))$ appearing in (8.60) is indeed a remainder term or not. We cannot exactly answer to this question because it is hard to compare δ_*^2 with $\inf_{m \in \mathcal{M}} \ell(s, \widehat{s}_m)$. However, if $\widehat{s}_m$ is itself an empirical risk minimizer over some model S_m, we can compare δ_*^2 with $\inf_{m \in \mathcal{M}} \varepsilon_m^2$, where ε_m^2 is (up to constant) the upper bound for the expected risk $\mathbb{E}[\ell(s, \widehat{s}_m)]$ provided by Theorem 8.3. More precisely, taking for instance $\theta = 1/2$, we derive from (8.60) that

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq 3 \inf_{m \in \mathcal{M}} \mathbb{E}[\ell(s, \widehat{s}_m)] + \delta_*^2(3 + 2\ln(|\mathcal{M}|)).$$

By Theorem 8.3, setting $\ell(s, S_m) = \inf_{t \in S_m} \ell(s,t)$, one has for some absolute constant κ

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq 6 \inf_{m \in \mathcal{M}} \left(\ell(s, S_m) + \kappa\varepsilon_m^2\right) + \delta_*^2(3 + 2\ln(|\mathcal{M}|)), \tag{8.61}$$

where ε_m is defined by the equation

$$\sqrt{N}\varepsilon_m^2 = \phi_m(w(\varepsilon_m)).$$

Let ϕ_m belong to $\mathcal{C}_1$ controlling the modulus of continuity of the empirical process $(P'_N - P)(\gamma(t,.))$ over model S_m with respect to some pseudodistance d satisfying to (8.16) and let w satisfy to (8.17). If N and n are of the

same order of magnitude, $N = n$ say to be as simple as possible, then, since one can always assume that $w \leq 1$ (otherwise one can change w into $1 \wedge w$) one has $\phi_m(w(\varepsilon_m)) \geq w(\varepsilon_m)$ and therefore $\delta_* \leq \varepsilon_m$. This shows that in full generality, the risk of $\widehat{s}_{\widehat{m}}$ is at most of order

$$\ln(e|\mathcal{M}|) \inf_{m \in \mathcal{M}} \left(\ell(s, S_m) + \kappa \varepsilon_m^2\right). \tag{8.62}$$

Up to the unpleasant logarithmic factor $\ln(e|\mathcal{M}|)$, this is exactly what one could expect of a clever model selection procedure, i.e., it performs as well as if the margin function w was known. This is of course especially interesting in the classification setting. We were in fact over pessimistic when deriving (8.62) from (8.61). To see this, let us consider the classification framework and consider the VC case with margin function $w(\varepsilon) = 1 \wedge h^{-1/2}\varepsilon$, assuming that $|\mathcal{M}| \leq n$. Then, if V_m denotes the VC-dimension of S_m, combining (8.61) with Theorem 8.3 (in the spirit of Corollary 8.4) yields

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq 6 \inf_{m \in \mathcal{M}} \left(\ell(s, S_m) + C \ln(n) \left(\sqrt{\frac{V_m}{n}}\right) \wedge \left(\frac{V_m}{nh}\right)\right).$$

Hold-Out for the Maximum Likelihood Criterion

We consider here the maximum likelihood criterion. We can derive from Theorem 8.7 the following general result for penalized log-likelihood hold-out procedures. We recall that $\mathbf{K}$ (resp. $\mathbf{h}$) denote the Kullback–Leibler information number (resp. the Hellinger distance) as defined at the beginning of Chapter 7.

Theorem 8.9 *Assume that we observe $N+n$ random variables with common distribution P with density s with respect to some given positive σ-finite measure μ. The first N observations $\xi'_1, \ldots, \xi'_N$ are used to build some preliminary collection of estimators $\{\widehat{s}_m\}_{m \in \mathcal{M}}$ and we use the remaining observations $\xi_1, \ldots, \xi_n$ to select some estimator $\widehat{s}_m$ among the collection $\{\widehat{s}_m\}_{m \in \mathcal{M}}$. Let $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ and denoting by P_n the empirical probability measure based on $\xi_1, \ldots, \xi_n$ consider some random variable $\widehat{m}$ such that*

$$P_n\left(-\ln\left(\widehat{s}_{\widehat{m}}\right)\right) + \mathrm{pen}(\widehat{m}) = \inf_{m \in \mathcal{M}} \left(P_n(-\ln(\widehat{s}_m)) + \mathrm{pen}(m)\right).$$

Let $\{x_m\}_{m \in \mathcal{M}}$ be some family of nonnegative numbers such that

$$\sum_{m \in \mathcal{M}} e^{-x_m} = \Sigma < \infty,$$

and suppose that for some constant $\theta \in (0,1)$

$$\mathrm{pen}(m) \geq \frac{x_m}{n}\left(\frac{3}{\theta} + 2\right) \text{ for every } m \in \mathcal{M}. \tag{8.63}$$

Then,

$$(1-\theta)\,\mathbb{E}\left[\mathbf{K}\left(s,\frac{s+\widehat{s}_{\widehat{m}}}{2}\right)\right]\leq(1+\theta)\inf_{m\in\mathcal{M}}\left(\mathbb{E}\left[\mathbf{K}\left(s,\widehat{s}_m\right)\right]+\mathrm{pen}\left(m\right)\right)+\frac{3\left(2\theta+\Sigma\theta^{-1}\right)+2\Sigma}{n}. \tag{8.64}$$

Proof. We work conditionally to $\xi'_1,\ldots,\xi'_N$ and apply Theorem 8.7 to the family of functions

$$g_m=-\frac{1}{2}\ln\left(\frac{\widehat{s}_m}{s}\right)\text{ and }f_m=-\ln\left(\frac{s+\widehat{s}_m}{2s}\right),\ m\in\mathcal{M}.$$

By concavity of the logarithm, we indeed have $f_m\leq g_m$ for every $m\in\mathcal{M}$. Now we must check the moment condition (8.50). It comes from Lemma 7.26 that given two probability densities u and t, by the triangle inequality, the following moment inequality is available for every integer $k\geq 2$

$$P\left(\left|\ln\left(\frac{s+u}{s+t}\right)\right|^k\right)\leq 2^{k-2}k!\times\frac{9}{8}\left(\mathbf{h}\left(s,u\right)+\mathbf{h}\left(s,t\right)\right)^2.$$

Since $9/\left(8\left(2\ln\left(2\right)-1\right)\right)\leq 3$, combining this inequality with (7.103) leads to

$$P\left(\left|\ln\left(\frac{s+u}{s+t}\right)\right|^k\right)\leq 2^{k-2}k!\times 3\left(\sqrt{\mathbf{K}\left(s,\frac{s+u}{2}\right)}+\sqrt{\mathbf{K}\left(s,\frac{s+t}{2}\right)}\right)^2,$$

which means that (8.50) holds with $c=2$ and $\sigma_m^2=3\mathbf{K}\left(s,\left(s+\widehat{s}_m\right)/2\right)$. Hence, since

$$P\left(f_m\right)=\mathbf{K}\left(s,\frac{s+\widehat{s}_m}{2}\right),$$

we derive from the definition of σ_m^2 that assumption (8.51) holds true with $w\left(x\right)=\sqrt{3}x$. Hence, setting

$$\delta_*^2=\frac{3}{n}$$

(8.53) is valid (conditionally to $\xi'_1,\ldots,\xi'_N$), provided that condition (8.52) is satisfied. This clearly yields (8.64). ■

The oracle inequality above is expressed in terms of the unusual loss function $\mathbf{K}\left(s,\left(s+t\right)/2\right)$. It comes from Lemma 7.23 that this loss function is also linked to the square Hellinger distance, so that, up to some absolute constant (8.64) remains true for the square Hellinger loss $\mathbf{h}^2\left(s,t\right)$.

8.5.2 Data-Driven Penalties

It could seem a bit disappointing to discover that a very crude method like hold-out is working so well. This is especially true in the classification framework. It is indeed a really hard work in this context to design margin adaptive

penalties. Of course recent works on the topic (see [71] for a review), involving local Rademacher penalties for instance, provide at least some theoretical solution to the problem but still if one carefully looks at the penalties which are proposed in these works, they systematically involve constants which are typically unknown. In some cases, these constants are absolute constants which should nevertheless considered as unknown just because the numerical values coming from the theory are obviously over pessimistic. In some other cases, it is even worse since they also depend on nuisance parameters related to the unknown distribution (like for instance the infimum of the density of the explanatory variables). In any case these penalization methods are not ready to be implemented and remain far from being competitive with simple methods like hold out (or more generally with cross-validation methods).

Hence, two natural and connected questions emerge:

- Is there some room left for penalization methods?
- How to calibrate penalties to design efficient penalization criteria?

There are at least two reasons for which despite of the arguments against penalization that we have developed at the beginning of this Section, one should however keep interest for penalization methods. The first one is that for independent but not identically distributed observations (we typically think of Gaussian regression on a fixed design), hold out or more generally cross-validation may become irrelevant. The second one is that, talking about hold-out, since one uses part of the original data as testing data, one looses a bit of efficiency. A close inspection of the oracle inequalities presented in the preceding section shows that in the situation of half-sampling for instance one typically looses some factor 2 in the oracle inequality. Moreover hold-out is also known to be quite unstable and this is the reason why V-fold cross-validation is preferred to hold-out and widely used in practice. But now, concerning V-fold cross-validation, the question becomes how to choose V and what is the influence of this choice on the statistical performance of the method. This means that on the one hand, it remains to better understand cross-validation from a theoretical point of view and on the other hand that there is some room left for improvements. One can indeed hope to do better when using some direct method like penalization. But of course, since the opponent is strong, beating it requires to calibrate penalties sharply. This leads us to the second question raised above. We would like to conclude this Chapter by providing some possible answers to this last question, partly based on theoretical results which are already available and partly based on heuristics and thoughts which lead to some empirical rules and new theoretical problems.

A Practical Rule for Calibrating Penalties from the Data

To explain our idea which consists in guessing what is the right penalty to be used from the data themselves, let us come back to Gaussian model selection.

If we consider again the Gaussian model selection theorem for linear models, the following points can be made

- Mallows' C_p can underpenalize.
- Condition $K > 1$ in the statement of Theorem 4.2 is sharp.
- What penalty should be recommended? One can try to optimize the oracle inequality. The result is that roughly speaking, $K = 2$ is a good choice (see [24]).

In practice, the level of noise is unknown, but one can retain from the Gaussian theory the rule of thumb:

$$\textit{"optimal" penalty} = 2 \times \textit{"minimal" penalty.}$$

Interestingly the minimal penalty can be evaluated from the data because when the penalty is not heavy enough one systematically chooses models with large dimension. It remains to multiply by 2 to produce the desired (nearly) optimal penalty. This is a strategy for designing a data-driven penalty without knowing in advance the level of noise.

Practical implementation of penalization methods involves the extension to non Gaussian frameworks of the data-driven penalty choice strategy suggested above in the Gaussian case. It can roughly be described as follows

- Compute the minimum contrast estimator $\widehat{s}_D$ on the union of models defined by the same number D of parameters.
- Use the theory to guess the shape of the penalty $\mathrm{pen}(D)$, typically $\mathrm{pen}(D) = \alpha D$ (but other forms are also possible, like $\mathrm{pen}(D) = \alpha D(1 + \ln(n/D))$).
- Estimate α from the data by multiplying by 2 the smallest value for which the corresponding penalized criterion does not explode.

In the context of change points detection, this data-driven calibration method for the penalty has been successfully implemented and tested by Lebarbier (see [74]). In the non-Gaussian case, we believe that this procedure remains valid but theoretical justification is far from being trivial and remains open. As already mentioned at the beginning of this Section, this problem is especially challenging in the classification context since it is connected to the question of defining optimal classifiers *without* knowing in advance the margin condition on the underlying distribution, which is a topic attracting much attention in the statistical learning community at this moment (see [115], [116], [14] for instance and [71] for a review).

Some Heuristics

More generally, defining proper data-driven strategies for choosing a penalty offers a new field of mathematical investigation since future progress on the topic requires to understand in depth the behavior of $\gamma_n(\widehat{s}_D)$. Recent advances

involve new concentration inequalities. A first step in this direction is made in [32] and a joint work in progress with S. Boucheron is building upon the new moment inequalities proved in [30]. If one wants to better understand how to penalize optimally and the role that concentration inequalities could play in this matter, one has to come back to the root of the topic of model selection via penalization i.e., to Mallows' and Akaike's heuristics which are both based on the idea of estimating the risk in an unbiased way (at least asymptotically as far as Akaike's heuristics is concerned). The idea is the following.

Let us consider, in each model S_m some minimizer s_m of $t \to \mathbb{E}\left[\gamma_n\left(t\right)\right]$ over S_m (assuming that such a point does exist). Defining for every $m \in \mathcal{M}$,

$$\widehat{b}_m = \gamma_n\left(s_m\right) - \gamma_n\left(s\right) \text{ and } \widehat{v}_m = \gamma_n\left(s_m\right) - \gamma_n\left(\widehat{s}_m\right),$$

minimizing some penalized criterion

$$\gamma_n\left(\widehat{s}_m\right) + \mathrm{pen}\left(m\right)$$

over $\mathcal{M}$ amounts to minimize

$$\widehat{b}_m - \widehat{v}_m + \mathrm{pen}\left(m\right).$$

The point is that since $\widehat{b}_m$ is an unbiased estimator of the bias term $\ell\left(s, s_m\right)$. If we have in mind to use concentration arguments, one can hope that minimizing the quantity above will be approximately equivalent to minimize

$$\ell\left(s, s_m\right) - \mathbb{E}\left[\widehat{v}_m\right] + \mathrm{pen}\left(m\right).$$

Since the purpose of the game is to minimize the risk $\mathbb{E}\left[\ell\left(s, \widehat{s}_m\right)\right]$, an ideal penalty would therefore be

$$\mathrm{pen}\left(m\right) = \mathbb{E}\left[\widehat{v}_m\right] + \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m\right)\right].$$

In the Mallows' C_p case, the models S_m are linear and $\mathbb{E}\left[\widehat{v}_m\right] = \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m\right)\right]$ are explicitly computable (at least if the level of noise is assumed to be known). For Akaike's penalized log-likelihood criterion, this is similar, at least asymptotically. More precisely, one uses the fact that

$$\mathbb{E}\left[\widehat{v}_m\right] \approx \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m\right)\right] \approx D_m / \left(2n\right),$$

where D_m stands for the number of parameters defining model S_m. The conclusion of these considerations is that Mallows' C_p as well as Akaike's criterion are indeed both based on the unbiased risk estimation principle.

Our guess is that we can go further in this direction and that the approximation $\mathbb{E}\left[\widehat{v}_m\right] \approx \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m\right)\right]$ remains generally valid. If we believe in it then a good penalty becomes $2\mathbb{E}\left[\widehat{v}_m\right]$ or equivalently (having still in mind concentration arguments) $2\widehat{v}_m$. This in some sense explains the rule of thumb which is given in the preceding Section: the minimal penalty is $\widehat{v}_m$ while the

optimal penalty should be $\widehat{v}_m + \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m\right)\right]$ and their ratio is approximately equal to 2.

Of course, concentration arguments will work only if the list of models is not too rich. In practice this means that starting from a given list of models, one has first to decide to penalize in the same way the models which are defined by the same number of parameters. Then one considers a new list of models $(S_D)_{D\geq 1}$, where for each integer D, S_D is the union of those among the initial models which are defined by D parameters and then apply the preceding heuristics to this new list.