Classification
and
model selection

Olivier Catoni

DEA Probabilités et Applications
last revised February 13, 2003

CNRS – Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 (site Chevaleret), 4 place Jussieu – Case 188, 75 252 Paris Cedex 05.
CHAPTER 1

A PAC-Bayesian approach to adaptive inference

1. Introduction

In these lectures, we will prove what could be called localized PAC-Bayesian learning theorems and illustrate their use to solve classification problems. The setting will be the one of statistical learning theory: complex data have to be analyzed (e.g., images, speech, natural language, DNA, . . .), about which very little is known beforehand and some crudely approximate classification model has to be picked-up among a possibly huge number of candidates through some kind of robust and automated model selection mechanism.

The idea of PAC-Bayesian learning theorems, as introduced by D. McAllester, [24, 25] is to measure the complexity of models, and thereby their ability to generalize from observed examples to unknown situations, with the help of some prior probability measure defined on the parameter space. Here, we use for simplicity the term parameter space in a rather loose and unusual way, to talk about the union of all the parameters of all the models we envision (maybe the term model space would be more accurate: these parameters may be of finite or infinite dimension and we do not restrict the number of models, therefore we are definitely not describing a parametric statistical framework, but rather a non-parametric one!). The status of the prior measure has not to be misunderstood either: it does not represent the frequency according to which we expect to observe data produced by different probability distributions, nor does it stand for the belief we put in the accuracy of different possible distributions or different possible models. It is somehow equivalent to the choice of some representation of the parameter space (since it is possible to derive some coding scheme from a probability distribution, according to coding theory), and therefore is related to the Minimum Description Length approach of Rissanen and to the structural risk minimization approach of Vapnik. On a more technical level, it is meant to produce non asymptotic worst case bounds, (as opposed to a Bayesian study of the mean risk under the prior). It shares some common features with the use of mixture codes in lossless data compression theory [32].

2. Mathematical framework

Let us now sketch the mathematical framework of our study. We consider a product space $\mathcal{X} \times \mathcal{Y}$, where $(\mathcal{X}, \mathcal{B})$ is a measurable space and where $\mathcal{Y}$ is a finite set. In a classification application, the set $\mathcal{X}$ has to be thought of as the pattern space and $\mathcal{Y}$ as the label space. Patterns in $\mathcal{X}$ may be described by a combination of continuous and discrete parameters, however, except when it comes down to giving examples, we will capture the structure of $\mathcal{X}$ only through the use of a family of classification functions defined on $\mathcal{X}$, we will come back to this later.
The observation is made of an i.i.d. sample \((X_i, Y_i)_{i=1}^N\), drawn according to some product distribution \(P_{\times N}\), where \(P\) is a probability measure on \((\mathcal{X} \times \mathcal{Y}, \mathcal{B} \times \mathcal{B}')\), \(\mathcal{B}'\) being the algebra \(\{0,1\}^\mathcal{Y}\) of all the subsets of \(\mathcal{Y}\).

The relations between \(X\) and \(Y\) will be analyzed with the help of some prescribed set of classification rules

\[\mathcal{R} = \{ f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta \},\]

where \((\Theta, \mathcal{T})\) is some measurable parameter set and

\[(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X} \times \mathcal{B}) \to (\mathcal{Y}, \mathcal{B}')\]

is assumed to be measurable. As we have already explained, the set \(\mathcal{R}\) will in general not be a single parametric model, but rather the union of a large number of parametric models. From the technical point of view, our aim will be to produce non asymptotic bounds for the risk of properly designed estimators of \(Y\) given \(X\).

The risk of \(f_\theta : \mathcal{X} \to \mathcal{Y}\) will be measured as its error rate

\[R(\theta) = P(Y \neq f_\theta(X)).\]

Let us mention here that throughout these lectures the short notation \(P(W)\) will be used for the expectation of the random variable \(W\) under the distribution \(P\).

The PAC Bayesian approach could roughly be explained as follows: instead of bounding the supremum of the empirical risk

\[r(\theta) = \frac{1}{N} \sum_{k=1}^N \mathbb{I}[Y_k \neq f_\theta(X_k)],\]

with respect to the parameter \(\theta \in \Theta\), we study the deviations of the quantiles of \(r(\theta)\) with respect to some prior probability measure \(\pi \in \mathcal{M}_1(\Theta, \mathcal{T})\) defined on the parameter space.

More precisely, we cannot minimize \(R(\theta)\) in \(\theta\) as we would like to do, because \(R(\theta)\) is not observable: it depends on the unknown distribution \(P\). The next sensible attempt is to minimize \(r(\theta)\) instead. Unfortunately, although \(P(r(\theta)) = R(\theta)\), the fluctuations of the random process \(r(\theta) : \theta \in \Theta\) may be strong enough to make the solutions of the two minimization problems quite different, and even in many cases completely unrelated. An intensively studied way to get some control on this situation is to add a penalty term \(\gamma_N(\theta)\) and study the relations between \(\inf_\theta R(\theta) + \gamma_N(\theta)\) and \(\inf_\theta r(\theta) + \gamma_N(\theta)\). The penalty \(\gamma_N(\theta)\) has a regularization effect: it shrinks the size of the set of values of \(\theta\) where \(\inf_\theta r(\theta) + \gamma_N(\theta)\) is likely to be achieved and therefore provides a way to control the gap between \(P[\inf_\theta [r(\theta) + \gamma_N(\theta)]]\) and \(\inf_\theta R(\theta) + \gamma_N(\theta)\). The difficulty of this approach comes from the choice of \(\gamma_N(\theta)\), which has to depend on the “size” of the parameter space \(\Theta\), measured in a suitable way.

In the PAC-Bayesian approach, we circumvent this difficulty by measuring weights under some prior distribution \(\pi \in \mathcal{M}_1(\Theta, \mathcal{T})\) on the parameter space. This is an indirect way to make the size of \(\Theta\) come into play. Although we will not explicitly manipulate quantiles in the technical part of our study, we will introduce here the role of the prior \(\pi\) with the help of this familiar concept which gives us an opportunity to make a link with the maximum likelihood approach. Let us define
the $\alpha$ quantile of the empirical risk $r(\theta)$ as

$$q_\alpha(r) = \inf \left\{ \mu : \pi[r(\theta) \leq \mu] > \alpha \right\}.$$ 

It can be viewed as a probabilistic generalization of the essential infimum of $r(\theta)$ under $\pi$, since

$$\text{ess inf}_{\pi(d\theta)} r(\theta) = q_0(r).$$

This generalization is of practical interest to us, because, whereas $\text{ess inf}_{\pi(d\theta)} r(\theta)$ has fluctuations depending on the “size” (or more accurately the complexity) of the parameter space $\Theta$, the fluctuations of the quantile $q_\alpha(r)$ can be evaluated as a function of $\alpha$ only, as long as $\alpha > 0$. The reason is that a quantile with positive parameter $\alpha$ is separating two sets of parameters with positive $\pi$-weights, unlike the essential infimum which may separate a single point of null $\pi$-weight from the rest of the parameter space: to produce a random deviation of the quantile $q_\alpha(r)$, the values of $r(\theta)$ for a given proportion ($\alpha$, namely) of the parameters have to deviate from their typical values, whereas a lower deviation of the essential infimum may be the consequence of the behavior of the empirical risk on a set of parameters of arbitrarily small $\pi$-weight (and the behavior of the empirical risk at a single value of the parameter may of course be responsible for a lower deviation of the true infimum).

As shown by D. McAllester in his pioneering papers on the subject, the “hard threshold” vision of quantiles we explained above can be generalized to smoother objects, and indeed to any “posterior distribution” $\rho \in M^1_+(\Theta)$ on the parameter space. A posterior distribution here is simply a probability measure $\rho \in M^1_+(\Theta, T)$ on the parameter space which may depend on the observations $(X_i, Y_i)_{i=1}^N$ (therefore it is a random measure).

The random measures depending on the empirical risk $r(\theta)$ are a special case of posterior distributions. More precisely, we will make a heavy use of Gibbs posterior distributions of the form

$$(2.1) \quad d\rho(\theta) = d\pi \exp(-\beta r(\theta)) = \frac{\exp(-\beta r(\theta))}{\pi[\exp(-\beta r(\theta))]} d\pi(\theta).$$

The introduction of these posterior distributions, viewed as random objects whose fluctuations are easily manageable, leads us to consider randomized estimators: instead of picking some parameter $\hat{\theta}$ as a deterministic function of the observations $(X_i, Y_i)_{i=1}^N$, we choose it at random according to the posterior distribution $\rho$ (which itself depends on the observations). The resulting risk of this randomized estimation scheme is $\rho[R(\theta)]$, which plays the same role as $R(\hat{\theta})$ in the deterministic setting. Although it depends on the unknown and deterministic risk function $R$, it is still a random variable, due to the randomness of the posterior measure $\rho$, in the same way as $R(\hat{\theta})$ is a random variable due to the dependence of $\hat{\theta}$ on the observations. In some situations, it is natural to use randomized estimators, in others the support of $\rho$ will be concentrated around some deterministic estimator $\hat{\theta}$ in some sensible way and the introduction of randomized estimators should more likely be viewed as a technical steps in the study of more conventional estimation schemes.

In McAllester’s papers, the fluctuations of $\rho[r(\theta)]$ with respect to $\rho[R(\theta)]$ are controlled by some function of $\mathcal{K}(\rho, \pi)$, the Kullback divergence of the (random)
posterior measure $\rho$ with respect to the (fixed) prior measure $\pi$, defined as

$$K(\rho, \pi) = \left\{ \begin{array}{ll} \rho \left[ \log \frac{d\rho}{d\pi} \right], & \text{when } \rho \ll \pi, \\ +\infty, & \text{otherwise.} \end{array} \right.$$ 

In the present study, we will make an important step towards sharper bounds by replacing $K(\rho, \pi)$ with $K(\rho, \pi \exp(-\beta r))$, where $\pi \exp(-\beta r) \in \mathcal{M}_1^+(\Theta)$ is the Gibbs posterior built from $\pi$ and $r$ we already mentioned a few lines above.

We will start with simple PAC-Bayesian learning theorems, explain how they can be used, and introduce further improvements only in subsequent sections.

Then we will show how Vapnik’s statistical learning theory can be proved and strengthened using the PAC-Bayesian approach: the idea is to replace the use of a deterministic prior with the use of a data dependent one.

### 3. Low noise pattern classification

We will be interested here in the most favorable case of pattern recognition: the case when an i.i.d. sample $(X_i, Y_i)_{i=1}^N$ of classified patterns is observed, where the conditional distribution of the label $Y$ given the pattern $X$ is highly peaked on one label (which will of course be considered as the “true” label for pattern $X$). As already explained, $(X_i, Y_i)_{i=1}^N$ will be the canonical process on some space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$ endowed with a product measure $P \otimes \mathcal{N}$, where $P \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$. A set of classification rules $\mathcal{R} = \{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ is at our disposal, where $(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{T} \otimes \mathcal{B}) \to (\mathcal{Y}, \mathcal{B}')$ is measurable. We will not make any “low-noise” assumption, but it will just turn out that the bounds derived in this section will be sharp only when the empirical risk

$$r(\theta) = \frac{1}{N} \sum_{i=1}^N 1[Y_i \neq f_\theta(X_i)]$$

is such that $\inf_{\theta \in \Theta} r(\theta)$ is small with a high probability.

#### 3.1. A reminder of non-asymptotic deviation techniques: Bernstein’s inequality and the Legendre transform of the Kullback divergence function

We need a non-asymptotic deviation inequality for sums of independent random variables. For this purpose, a detailed formulation of Bernstein’s inequality is useful. It can be found in [26, p 203-204].

**Theorem 3.1.** Let $(\sigma_1, \ldots, \sigma_N)$ be independent real valued random variables and $\mathbb{P}$ their joint distribution. Let us assume that

$$\sigma_i - \mathbb{P}(\sigma_i) \leq b, \quad i = 1, \ldots, N.$$ 

Let

$$S = \frac{1}{N} \sum_{i=1}^N \sigma_i$$

be their normalized sum,

$$m = \mathbb{P}(S) = \frac{1}{N} \sum_{i=1}^N \mathbb{P}(\sigma_i)$$
its expectation and
\[ V = N \mathbb{P} \left[ (S - \mathbb{P}(S))^2 \right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P} \left[ (\sigma_i - \mathbb{P}(\sigma_i))^2 \right] \]
its renormalized variance. Let us introduce the increasing function
\[ g(x) = \frac{1}{x^2} (e^x - 1 - x). \]
The deviations of \( S \) are bounded, for any \( \lambda \in \mathbb{R}_+ \), any \( \eta \in \mathbb{R}_+ \), by
\[ \mathbb{P}(S - m \geq \eta) \leq \mathbb{P} \left[ \exp \left( -\lambda \eta + \lambda(S - m) \right) \right] \leq \exp \left( -\eta \lambda + g \left( \frac{b\lambda}{N} \frac{V}{N^2} \lambda^2 \right) \right), \]
moreover when \( \lambda \) is chosen to be
\[ \lambda = \frac{N}{b} \log \left( 1 + \frac{b\eta}{V} \right), \]
the right-hand side of the previous equation is itself bounded by
\[ \exp \left( -\eta \lambda + g \left( \frac{b\lambda}{N} \frac{V}{N^2} \lambda^2 \right) \right) \leq \exp \left( -\frac{3N\eta^2}{6V + 2b\eta} \right). \]

**Proof.** It can be easily checked that the function \( x \mapsto g(x) : \mathbb{R} \rightarrow \mathbb{R} \) is increasing. Moreover, it is clearly enough to prove the theorem when \( \mathbb{P}(\sigma_i) = 0 \), for any \( i = 1, \ldots, N \). Reminding that \( \log(1 + x) \leq x \), we see that
\[
\begin{align*}
\mathbb{E} \left[ \exp(\lambda S) \right] &= \exp \left\{ \sum_{i=1}^{N} \log \left[ \mathbb{E} \left[ \exp \left( \frac{\lambda}{N} \sigma_i \right) \right] - \mathbb{E} \left[ \frac{\lambda}{N} \sigma_i \right] \right] \right\} \\
&\leq \exp \left\{ \sum_{i=1}^{N} \mathbb{E} \left[ \exp \left( \frac{\lambda}{N} \sigma_i \right) - \frac{\lambda}{N} \sigma_i - 1 \right] \right\} \\
&= \exp \left\{ \sum_{i=1}^{N} \mathbb{E} \left[ g \left( \frac{\lambda}{N} \sigma_i \right) \frac{\sigma_i^2}{N^2} \right] \right\} \\
&\leq \exp \left\{ g \left( \frac{b\lambda}{N} \right) \sum_{i=1}^{N} \mathbb{E} \left[ \sigma_i^2 \right] \frac{\lambda^2}{N^2} \right\} \\
&= \exp \left\{ g \left( \frac{b\lambda}{N} \right) \frac{V}{N^2} \lambda^2 \right\}. 
\end{align*}
\]
This proves (3.2). The last statement of the theorem can be rewritten after a suitable change of variables as
\[ -\eta \lambda + \lambda^2 g(\lambda) \leq -\frac{3\eta^2}{6 + 2\eta} \quad \text{when} \quad \lambda = \log(1 + \eta). \]
This is equivalent to
\[ (1 + \eta) \log(1 + \eta) - \eta \geq \frac{3\eta^2}{6 + 2\eta}, \]
and therefore to
\[ (6 + 8\eta + 2\eta^2) \log(1 + \eta) - 5\eta^2 - 6\eta \geq 0. \]
This last inequality, which holds true when $\eta = 0$, can be checked to hold true for any positive value of $\eta$ by differentiating twice its left-hand side. □

Some background on the Legendre transform of the convex function $\rho \mapsto K(\rho, \pi)$ is also needed.

**Lemma 3.1.** Let us recall that for any measurable function $h : \Theta \to \mathbb{R}$,

\[
\log \left\{ \pi \left\{ \exp [h(\theta)] \right\} \right\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \rho[h(\theta)] - K(\rho, \pi),
\]

where the value of $\rho[h(\theta)]$ is defined by convention as

\[
\rho[h(\theta)] \overset{\text{def}}{=} \sup_{B \in \mathbb{R}} \rho \left[ \min \{ B, h(\theta) \} \right],
\]

and where it is also understood that

\[
\infty - \infty = \sup_{B \in \mathbb{R}} (B) - \infty = \sup_{B \in \mathbb{R}} (B - \infty) = -\infty.
\]

(In other words a priority is given to $-\infty$ in ambiguous cases: the expectation of a function whose negative part is not integrable will be assumed to be $-\infty$, even when its positive part integrates to $+\infty$.)

Moreover, when $h$ is upper bounded, for any $\rho \in \mathcal{M}_+^1(\Theta, \mathcal{T})$,

\[
\log \left\{ \pi \left\{ \exp [h(\theta)] \right\} \right\} + K(\rho, \pi) - \rho[h(\theta)] = K(\rho, \nu),
\]

where $d\nu(\theta) = \frac{\exp[h(\theta)]}{\pi[h(\theta)]} d\pi(\theta)$. (Equality is meant to hold in $\mathbb{R} \cup \{\infty\}$, meaning that $K(\rho, \nu) < \infty$ if and only if $K(\rho, \pi) < \infty$ and $-\rho[h(\theta)] < \infty$ and that in this case equality holds in $\mathbb{R}$.)

**Proof.** Let us give for the sake of completeness a short proof of this well known result. The second part of the lemma is a straightforward computation. Let us remark first that $\rho$ is absolutely continuous with respect to $\pi$ if and only if it is absolutely continuous with respect to $\nu$, because $\pi$ and $\nu$ have the same negligible measurable sets. Therefore if $\rho$ is singular with respect to $\pi$, then both members of (3.6) are equal to $\infty$. Let us assume now that $\rho$ is absolutely continuous with respect to $\pi$, and write from the definition of the divergence function

\[
K(\rho, \nu) = \rho \left\{ \log \left( \frac{d\rho}{d\pi} \right) - h(\theta) \right\} + \log \left\{ \pi \left[ \exp [h(\theta)] \right] \right\}.
\]

Remark that the negative part of $\log \left( \frac{d\rho}{d\pi} \right)$ is in $L^1(\rho)$, because $\frac{d\rho}{d\pi} \left[ \log \left( \frac{d\rho}{d\pi} \right) \right]_-$ is bounded and therefore in $L^1(\pi)$. As $-h$ is lower bounded, we can thus write in $\mathbb{R} \cup \{\infty\}$ that

\[
\rho \left\{ \log \left( \frac{d\rho}{d\pi} \right) - h(\theta) \right\} = \rho \left\{ \log \left( \frac{d\rho}{d\pi} \right) \right\} - \rho[h(\theta)].
\]

This is precisely (3.6).

In the case when $h$ is upper bounded, the first part of the lemma is a consequence of its second part, which shows moreover that the maximum in $\rho$ is attained.
when $\rho = \nu$. In the general case, we can write the following chain of equalities, where we have used the notation $\min\{B, h(\theta)\} = B \land h(\theta)$,

$$\log\left\{\pi\left\{\exp[h(\theta)]\right\}\right\} = \sup_{B \in \mathbb{R}} \log\left\{\pi\left\{\exp[B \land h(\theta)]\right\}\right\}$$

$$= \sup_{B \in \mathbb{R}} \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \left\{\rho[B \land h(\theta)] - \mathcal{K}(\rho, \pi)\right\}$$

$$= \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \sup_{B \in \mathbb{R}} \left\{\rho[B \land h(\theta)] - \mathcal{K}(\rho, \pi)\right\}$$

$$= \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \sup_{B \in \mathbb{R}} \left\{\rho[B \land h(\theta)] - \mathcal{K}(\rho, \pi)\right\} - \mathcal{K}(\rho, \pi)$$

$$= \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \rho[h(\theta)] - \mathcal{K}(\rho, \pi).$$

□

3.2. A non localized learning theorem for low-noise classification. We will apply the second inequality (3.2) of Bernstein’s theorem 3.1 successively to

$$\sigma_i \equiv \mathbb{I}\left(Y_i \neq f_\theta(X_i)\right)$$

and to

$$\sigma_i \equiv \mathbb{I}\left(Y_i \neq f_\theta(X_i)\right).$$

We will integrate both sides of the resulting inequality with respect to some prior $\pi \in \mathcal{M}_1^+(\Theta, \mathcal{T})$, to obtain a “learning” lemma which improves on the PAC-Bayesian bounds in [24, 25], which were derived from the weaker Hoeffding’s inequality.

**Lemma 3.2.** For any positive real parameter $\lambda \in \mathbb{R}_+$, any non negative real valued measurable function $\eta : \Theta \to \mathbb{R}_+$, any prior probability distribution $\pi \in \mathcal{M}_1^+(\Theta, \mathcal{T})$,

$$\begin{align*}
P^\otimes N\left\{\sup_{\rho \in \mathcal{M}_1^+(\Theta)} \lambda \rho[R(\theta)] - \lambda \rho[r(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi) \geq 0\right\} \\
\leq \pi\left\{\exp\left[\frac{\lambda^2}{N} g\left(\frac{\lambda R(\theta)}{N}\right) R(\theta) [1 - R(\theta)] - \eta(\theta)\right]\right\}.
\end{align*}$$

In the same way

$$\begin{align*}
P^\otimes N\left\{\sup_{\rho \in \mathcal{M}_1^+(\Theta)} \lambda \rho[r(\theta)] - \lambda \rho[R(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi) \geq 0\right\} \\
\leq \pi\left\{\exp\left[\frac{\lambda^2}{N} g\left(\frac{\lambda}{N}\right) R(\theta) [1 - R(\theta)] - \eta(\theta)\right]\right\}.
\end{align*}$$

**Proof.** According to lemma 3.1,

$$\begin{align*}
\sup_{\rho \in \mathcal{M}_1^+(\Theta)} \rho[\lambda R(\theta) - \lambda r(\theta) - \eta(\theta)] - \mathcal{K}(\rho, \pi) \\
= \log\left\{\pi\left\{\exp\left[\lambda R(\theta) - r(\theta) - \eta(\theta)\right]\right\}\right\}.
\end{align*}$$
Thus
\[ P^\otimes N \left\{ \sup_{\rho \in M^1_\pi(\Theta)} \lambda \rho [R(\theta)] - \lambda \rho [r(\theta)] - \rho [\eta(\theta)] - \mathcal{K}(\rho, \pi) \geq 0 \right\} \]
\[ = P^\otimes N \left\{ \pi \left\{ \exp \left[ \lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \geq 1 \right\} \]
\[ \leq P^\otimes N \left\{ \pi \left\{ \exp \left[ \lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \right\} \]
\[ = \pi \left\{ P^\otimes N \left\{ \exp \left[ \lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \right\} \]
\[ \leq \pi \left\{ \exp \left[ \frac{\lambda^2}{N} g \left( -\frac{\lambda R(\theta)}{N} \right) R(\theta) \left[ 1 - R(\theta) \right] - \eta(\theta) \right] \right\}. \]

Equality (3.11) is obtained by applying the Fubini theorem to the positive function \((\theta, X_1, Y_1, \ldots, X_N, Y_N) \mapsto \exp \left\{ \lambda [R(\theta) - r(\theta)] - \eta(\theta) \right\} \). Inequality (3.12) is obtained by applying inequality (3.2) of Bernstein’s theorem 3.1 for each value of the parameter \(\theta\).

The proof of the reverse inequality (3.7) is similar and is left to the reader. □

**Remark 3.1.** The last step of the proof (3.12) can be replaced with an equality depending on the unknown distribution \(P\), which is of a less practical interest but may bring some further understanding of the situation: indeed, it could be noticed that
\[ \pi \left\{ P^\otimes N \left\{ \exp \left[ \lambda [R(\theta) - r(\theta)] - \eta(\theta) \right] \right\} \right\} = \pi \left\{ \exp \left\{ N \mathcal{K}[P, P_{\exp(\lambda \sigma)}] - \eta(\theta) \right\} \right\}, \]
where \(\sigma(\theta, X, Y) = -1[Y \neq f_\theta(X)]\) and for any positive measurable function \(h : X \times Y \to \mathbb{R}_+^*\) we have introduced the notation
\[ dP_{h(X,Y)} = P[h(X,Y)]^{-1}dP(X,Y). \]
This is a simple application of equality (3.6) in another context.

In the sequel of this paper, we will state a series of more sophisticated learning lemmas. Therefore it may be of some help to stop for a moment and see what use can be made of this type of result and how it can be compared with more classical statistical theorems. The easiest way to build an estimator and estimate its use can be made of this type of result and how it can be compared with more classical statistical theorems. The easiest way to build an estimator and estimate its performance using lemma 3.2 is to apply it choosing \(\eta(\theta) = \log(\epsilon^{-1}) - \frac{\lambda^2}{N} g \left( \frac{\lambda}{N} \right) R(\theta)\), to get

**Corollary 3.1.** With \(P^\otimes N\) probability at least \(1 - \epsilon\), for any posterior distribution \(\rho \in M^1_\pi(\Theta)\),
\[ \rho [R(\theta)] \leq \left[ 1 - \frac{\lambda}{N} g \left( \frac{\lambda}{N} \right) \right]^{-1} \left\{ \rho [r(\theta)] + \frac{1}{\lambda} \left[ \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) \right] \right\}. \]

The above inequality is the kind of non-asymptotic empirical bound we will be after throughout these lectures. Let us show here that it provides in a natural way an estimator with a given level of confidence. Building a randomized estimator from an empirical bound is straightforward: it is obtained by minimizing the bound.
with respect to the posterior distribution \( \rho \). Let \( \hat{\rho} \) be this minimizing posterior. Although we will not use it in the following discussion, it may be interesting to notice here that \( \hat{\rho} \) can be explicit: namely it is the Gibbs posterior distribution 
\[
\hat{\rho} = \pi_{\exp(-\lambda r)}
\]
(where we have used the notation introduced by (2.1)). Its risk has an upper confidence bound \( B(\hat{\rho}, \epsilon) \) at level \( \epsilon \), where
\[
B(\rho, \epsilon) = \left(1 - \frac{\lambda}{N}\right)^{-1} \left\{ \rho \left[ r(\theta) \right] + \frac{1}{\lambda} K(\rho, \pi) + \frac{\log(\epsilon^{-1})}{\lambda} \right\},
\]
where we have put \( \kappa = g \left( \frac{\lambda}{N} \right) \approx \frac{1}{2} \) for short. In other words,
\[
P^{\otimes N} \left\{ \hat{\rho} \left[ R(\theta) \right] \geq B(\hat{\rho}, \epsilon) \right\} \leq \epsilon.
\]
This is satisfactory from the practical point of view, since \( B(\hat{\rho}, \epsilon) \) is computable from the observed sample \((X_i, Y_i)_{i=1}^N\). However, from a theoretical point of view, the reader may wonder about the performance of the estimator, that is about the link between \( B(\hat{\rho}, \epsilon) \) and \( \inf_{\theta \in \Theta} R(\theta) \). There is a standard way to deal with this question. Let us explain it here as a motivation for the following. For any fixed distribution \( \rho \in \mathcal{M}_1^*(\Theta, \mathcal{T}) \), the empirical (i.e. random) bound \( B(\rho, \epsilon) \) is up to some constant a sum of i.i.d. random variables, with mean \( \bar{B}(\rho, \epsilon) \) given by
\[
\bar{B}(\rho, \epsilon) = \left(1 - \frac{\lambda}{N}\right)^{-1} \left\{ \rho \left[ r(\theta) \right] + \frac{1}{\lambda} K(\rho, \pi) + \frac{\log(\epsilon^{-1})}{\lambda} \right\}.
\]
It is straightforward to estimate its deviations. We can for instance write that
\[
P^{\otimes N} \left\{ \rho \left[ r(\theta) \right] \geq \left(1 + \frac{\lambda}{N}\right) \rho \left[ R(\theta) \right] + \frac{\log(\epsilon^{-1})}{\lambda} \right\} \leq \epsilon,
\]
Moreover, from the construction of \( \hat{\rho} \), \( B(\hat{\rho}, \epsilon) \leq B(\rho, \epsilon) \). Thus for any \( \rho \in \mathcal{M}_1^*(\Theta, \mathcal{T}) \), with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \),
\[
B(\hat{\rho}, \epsilon) \leq B(\rho, \epsilon) \leq \left(1 - \frac{\lambda}{N}\right)^{-1} \left\{ \left(1 + \frac{\lambda}{N}\right) \rho \left[ R(\theta) \right] + \frac{1}{\lambda} K(\rho, \pi) + \frac{2\log(\epsilon^{-1})}{\lambda} \right\}.
\]
However, the right-hand side of this last inequality is non random, and therefore can legitimately be optimized in \( \rho \). Weakening a little the result to make it more readable, (and remembering that \( \hat{\rho} = \pi_{\exp(\beta r)} \)), we get

**Proposition 3.1.** For any \( \lambda \in \mathbb{R}_+ \), with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \),
\[
\pi_{\exp(-\lambda r)} \left[ R(\theta) \right] \leq \frac{1 + \frac{\lambda}{N}}{1 - \frac{\lambda}{N}} \left\{ \inf_{\rho \in \mathcal{M}_1^*(\Theta)} \frac{1}{\lambda} K(\rho, \pi) \right\} + \frac{2\log(\epsilon^{-1})}{(1 - \frac{\lambda^2}{N^2}) \lambda}.
\]
It is also possible to bound the mean risk \( P^{\otimes N} \left\{ \hat{\rho} \left[ R(\theta) \right] \right\} \). One standard way to achieve this is to start from inequality
\[
P^{\otimes N} \left\{ \hat{\rho} \left[ R(\theta) \right] \geq B(\rho, \epsilon) \right\} \leq \epsilon,
\]
where we have kept \( \rho \) non random, and to rewrite it as
\[
P^{\otimes N} \left\{ U \geq \alpha \right\} \leq \exp(-\lambda \alpha),
\]
where we have introduced the random variable

\[ U = \hat{\rho}[R(\theta)] - \left(1 - \frac{\lambda}{N}\right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\}. \]

We have

\[ P^\otimes N(U) \leq \int_0^{+\infty} P^\otimes N(U \geq \alpha) d\alpha \leq \frac{1}{\lambda} \left(1 - \frac{\lambda}{N}\right)^{-1}. \]

In other words,

\[ P^\otimes N \left\{ \hat{\rho}[R(\theta)] \right\} \leq \left(1 - \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_1^*(\Theta)} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) + \frac{1}{\lambda} \right\}. \]

A slight improvement is achieved if we come back to (3.10) and (3.12). With a proper choice of parameters, we get

\[ P^\otimes N \left\{ \pi \left[ \exp \left( \lambda(1 - \frac{\lambda}{N}) R(\theta) - \lambda r(\theta) \right) \right] \right\} \]

\[ = P^\otimes N \left\{ \exp \left[ \sup_{\rho \in \mathcal{M}_1^*(\Theta)} \lambda(1 - \frac{\lambda}{N}) \rho[R(\theta)] - \lambda \rho[r(\theta)] - \mathcal{K}(\rho, \pi) \right] \right\} \leq 1. \]

Using Jensen’s inequality for the (convex) exponential function, we see that

\[ P^\otimes N \left\{ \lambda(1 - \frac{\lambda}{N}) \hat{\rho}[R(\theta)] - \lambda \hat{\rho}[r(\theta)] - \mathcal{K}(\hat{\rho}, \pi) \right\} \leq 0. \]

Therefore

\[ P^\otimes N \left\{ \hat{\rho}[R(\theta)] \right\} \leq \left(1 - \frac{\lambda}{N}\right)^{-1} P^\otimes N \left\{ \inf_{\rho \in \mathcal{M}_1^*(\Theta)} \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} \]

\[ \leq \left(1 - \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_1^*(\Theta)} P^\otimes N \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} \]

\[ = \left(1 - \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_1^*(\Theta)} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\}. \]

**Proposition 3.1.** For any \( \lambda \in \mathbb{R}_+ \) such that \( \frac{\lambda}{N} < 1 \),

\[ P^\otimes N \left\{ \pi_{\exp(-\lambda r)}[R(\theta)] \right\} \]

\[ \leq \left(1 - \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_1^*(\Theta)} \left\{ \rho[R(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right\} \]

\[ = \left(1 - \frac{\lambda}{N}\right)^{-1} \frac{1}{\lambda} \log \left\{ \pi \left[ \exp \left[ -\lambda R(\theta) \right] \right] \right\}. \]

Another important remark is to notice that corollary 3.1 can also be optimized in \( \lambda \). A simple way to do this is to consider a countable (possibly dense) family \( \Lambda \subset \mathbb{R} \) and some probability measure \( \nu \) on \( \Lambda \). Then defining \( \lambda \) to be the minimizer in \( \lambda \in \Lambda \) of

\[ \left(1 - \frac{\lambda}{N}\right)^{-1} \inf_{\rho \in \mathcal{M}_1^*(\Theta)} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) - \frac{\log[c_\nu(\lambda)]}{\lambda} \right\}, \]
we get some estimator \( \hat{\rho}_\lambda \) satisfying with \( P^\otimes N \) probability at least \( 1 - \epsilon \)

\[
\hat{\rho}_\lambda[R(\theta)] \leq \inf_{\lambda \in \Lambda, \rho \in \mathcal{M}_1(\Theta)} \left( 1 - \frac{\kappa \lambda}{N} \right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) - \frac{1}{\lambda} \log [\nu(\lambda)] \right\}.
\]

A more sophisticated way to optimize in \( \lambda \) is to establish a learning lemma uniform in both \( \lambda \) and \( \rho \). Let \( \nu \in \mathcal{M}(\mathbb{R}_+, \mathcal{B}) \) be some prior on the positive real line equipped with the Borel sigma algebra. Similarly to what has been proved before

**Proposition 3.2.** With \( P^\otimes N \) probability at least \( 1 - \epsilon \) for any posterior distributions \( \mu \in \mathcal{M}_1(\mathbb{R}_+) \) and \( \rho \in \mathcal{M}_1(\Theta) \)

\[
\rho[R(\theta)] \leq \left( 1 - \frac{\mu(\kappa \lambda^2)}{\mu(\lambda) N} \right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\mu(\lambda)} \left[ \mathcal{K}(\rho, \pi) + \mathcal{K}(\mu, \nu) + \log(\epsilon^{-1}) \right] \right\}.
\]

(The union bound approach is the special case of this last inequality where \( \nu \) has a countable support and \( \mu \) is a Dirac mass).

Of course, the link previously made between empirical and theoretical bounds can be carried over to the empirical bounds optimized in \( \lambda \):

**Corollary 3.2.** If \( \hat{\mu} \) and \( \hat{\rho} = \pi_{\exp(-\hat{\mu}(\lambda)r)} \) are the optimizers of the empirical bound of proposition 3.2 at level of confidence \( \epsilon \), then with \( P^\otimes N \) probability at least \( 1 - \epsilon \)

\[
\hat{\rho}[R(\theta)] \leq \inf_{\mu \in \mathcal{M}_1(\mathbb{R}_+), \rho \in \mathcal{M}_1(\Theta)} \left\{ \left( 1 - \frac{\mu(\kappa \lambda^2)}{\mu(\lambda) N} \right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\mu(\lambda)} \left[ \mathcal{K}(\rho, \pi) + \mathcal{K}(\mu, \nu) + 2 \log(2) \right] + \log(\epsilon^{-1}) \right\} \right\}.
\]

Anyhow, we mentioned proposition 3.2 and its corollary rather as a curiosity, using the union bound on a grid (which is by the way a special case of this proposition) being quite sufficient in practice. Let us make this more explicit. Consider for some real parameter \( \alpha > 1 \) the grid

\[
\Lambda = \left\{ N \alpha^{-k} : k \in \mathbb{N}, 0 \leq k \leq \frac{\log(N)}{\log(\alpha)} \right\}.
\]

Let us consider the uniform probability distribution on \( \Lambda \), in other words, let us choose

\[
\nu = \frac{1}{|\Lambda|} \sum_{\lambda' \in \Lambda} \delta_{\lambda'}.
\]

We get with \( P^\otimes N \) probability at least \( 1 - \epsilon \) that for any posterior distribution \( \rho \in \mathcal{M}_1(\Theta) \),

\[
\rho[R(\theta)] \leq \inf_{\lambda' \in \Lambda} \left( 1 - \frac{\mu(\lambda')}{N} \right)^{-1} \left\{ \rho[r(\theta)] + \frac{1}{\lambda'} \left[ \mathcal{K}(\rho, \pi) + \log \left( \frac{\log(N)}{\log(\alpha)} + 1 \right) + \log(\epsilon^{-1}) \right] \right\}
\]
\( \lambda \) any posterior \( \hat{\lambda} \in \Lambda \) such that \( \lambda \leq \lambda' \leq \alpha \lambda \). Moreover the right-hand side of it is minimized for any fixed value of \( \lambda \) by the Gibbs distribution \( \pi_{\exp(-\lambda r)} \) (see (2.1)), which we will write for short as \( \hat{\rho}_\lambda = \pi_{\exp(-\lambda r)} \).

**Corollary 3.3.** For any \( \alpha > 1 \), with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any \( \lambda \in [1, N] \),

\[
\hat{\rho}_\lambda [R(\theta)] \leq \left( 1 - g\left( \frac{\alpha \lambda}{N} \right) \right)^{-1} \frac{1}{\lambda} \left\{ -\log \left\{ \pi \left[ \exp[-\lambda r(\theta)] \right] \right\} \right. \\
+ \left. \log \left[ \frac{\log(N)}{\log(\alpha)} + 1 \right] + \log(\epsilon^{-1}) \right\}.
\]

Let us remark that it can be useful to compute \( \log \left\{ \pi \left[ \exp[-\lambda r(\theta)] \right] \right\} \) to use the identity

\[-\log \left\{ \pi \left[ \exp[-\lambda r(\theta)] \right] \right\} = \int_0^\lambda \hat{\rho}_\beta [r(\theta)] d\beta, \]

and to notice that for any sequence \( \beta_0 = 0 < \beta_1 < \cdots < \beta_m = \lambda \),

\[
\sum_{k=1}^m (\beta_k - \beta_{k-1}) \hat{\rho}_{\beta_k} [r(\theta)] \leq \int_0^\lambda \hat{\rho}_\beta d\beta \leq \sum_{k=1}^m (\beta_k - \beta_{k-1}) \hat{\rho}_{\beta_{k-1}} [r(\theta)].
\]

(This comes from the fact that \( \beta \mapsto \hat{\rho}_\beta [r(\theta)] \) is decreasing, its derivative being the opposite of the variance of \( r(\theta) \) under \( \hat{\rho}_\beta \).)

Let us note also that the upper deviations of the risk \( R(\theta) \) under the Gibbs posterior \( \hat{\rho}_\lambda \) can easily be bounded. Indeed with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any \( \lambda \in [1, N] \), any \( \beta \in \mathbb{R}_+ \),

\[
\log \left\{ \hat{\rho}_\lambda \left[ \exp[\beta R(\theta)] \right] \right\} = \sup_{\rho \in \mathcal{M}_+^1 (\theta)} \beta \rho [R(\theta)] - \mathcal{K}(\rho, \hat{\rho}_\lambda) \\
= \sup_{\rho \in \mathcal{M}_+^1 (\theta)} \beta \rho [R(\theta)] - \mathcal{K}(\rho, \pi) \\
- \lambda \rho [r(\theta)] - \log \left\{ \pi \left[ \exp[-\lambda r(\theta)] \right] \right\} \\
\leq \beta \left( 1 - g\left( \frac{\alpha \lambda}{N} \right) \right)^{-1} \left\{ \rho [r(\theta)] \\
+ \frac{1}{\lambda} \left\{ \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) + \log \left[ \frac{\log(N)}{\log(\alpha)} \right] \right\} \right\} \\
- \mathcal{K}(\rho, \pi) - \lambda \rho [r(\theta)] - \log \left\{ \pi \left[ \exp[-\lambda r(\theta)] \right] \right\}. 
\]

Thus choosing \( \beta = \lambda \left( 1 - g\left( \frac{\alpha \lambda}{N} \right) \right) \), we get

\[
\log \left\{ \hat{\rho}_\lambda \left[ \exp[\beta R(\theta)] \right] \right\} \leq - \log \left\{ \pi \left[ \exp[-\lambda r(\theta)] \right] \right\} + \log(\epsilon^{-1}) + \log \left[ \frac{\log(N)}{\log(\alpha)} + 1 \right].
\]
This proves

**Corollary 3.4.** For any real constant $\alpha > 1$, with $P^{\otimes N}$ probability at least $1 - \epsilon$, for any $\lambda \in [1, N]$, with $\hat{\rho}_\lambda$ probability at least $1 - \eta$,

$$R(\theta) \leq \left(1 - g\left(\frac{\alpha \lambda}{N}\right)\right)^{-1} \frac{1}{\lambda} \left\{-\log\left(\pi\left[\exp\left[-\lambda r(\theta)\right]\right]\right) \right\}$$

$$- \log(\epsilon \eta) + \log \left[\frac{\log(N)}{\log(\alpha)} + 1\right].$$

Although we will not mention it any further, the same kind of upper deviation bounds with respect to the Gibbs posterior can be derived from the results presented in the sequel of these lectures.

**3.3. Example.** Before delving into improvements, let us illustrate the use of these simple bounds to build aggregated classifiers.

Let $\{f_\theta : X \to \{-1, +1\}; \theta \in \Theta\}$ be some family of classification rules in a two classes pattern recognition problem. Here the label space is equal to $Y = \{-1, +1\}$. For any probability measure $\rho \in \mathcal{M}_1^+(\Theta)$, we consider the aggregated classifier

$$f_\rho(x) = \text{sign}\left(\rho[f_\theta(x)]\right).$$

If $P$ is as previously the joint distribution of the patterns and labels, then the error rate of $f_\rho$ is

$$R(\rho) = P\{Y \rho[f_\theta(X)] < 0\},$$

and the corresponding empirical risk is

$$r(\rho) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{Y_i \rho[f_\theta(X_i)] < 0\}.$$ 

In this problem, the space of parameters is $\Theta' = \mathcal{M}_1^+(\Theta)$ and we need to consider some reference measure on this space to apply our method. One way to do this is to consider the mapping

$$\Psi : \Theta^M \to \Theta'$$

$$\theta_1^M \mapsto \frac{1}{M} \sum_{i=1}^{M} \delta_{\theta_i}. $$

Consider some reference probability measure $\pi \in \mathcal{M}_1^+(\Theta)$ and build a prior $\pi'$ belonging to $\mathcal{M}_1^+(\Theta')$ from the formula

$$\pi' = \pi^{\otimes M} \circ \Psi^{-1}.$$ 

**Lemma 3.3.** For any probability measure $\rho \in \mathcal{M}_1^+(\Theta^M)$, the posterior distribution $\rho' = \rho \circ \Psi^{-1}$ on $\mathcal{M}_1^+(\Theta')$ is such that

$$\mathcal{K}(\rho', \pi') \leq \mathcal{K}(\rho, \pi^{\otimes M}).$$

**Proof.** This is a consequence of the decomposition of the Kullback divergence function:

$$\mathcal{K}(\rho, \pi^{\otimes M}) = \mathcal{K}(\rho', \pi') + \rho\left\{\mathcal{K}\left[\rho[\delta\theta_1^M | \Psi(\theta_1^M)], \pi^{\otimes M}[\delta\theta_1^M | \Psi(\theta_1^M)]\right]\right\}.$$
CHAPTER 1. A PAC-BAYESIAN APPROACH TO ADAPTIVE INFERENCE

Note that equality holds when $\rho$ is a product measure. □

From corollary 3.3, a real parameter $\alpha > 1$ being chosen, with $P^\otimes N$ probability at least $1 - \epsilon$, for any $\lambda \in [1, N]$, any $\rho \in \mathcal{M}_1^+(\Theta^M)$,

$$\rho\left\{ R[\Psi(\theta^M_1)] \right\} \leq \left(1 - g\left(\frac{\alpha\lambda}{N}\right)\right)^{-1} \left\{ \rho\left\{ r[\Psi(\theta^M_1)] \right\} + \frac{1}{\lambda} \left[ \mathcal{K}(\rho, \pi^\otimes M) + \log(\epsilon^{-1}) + \log \left( \frac{\log(N)}{\log(\alpha)} + 1 \right) \right] \right\}.$$

Optimizing the right-hand side of this empirical inequality in $\rho$ gives a posterior $\hat{\rho}_\lambda$ defined by

$$d\hat{\rho}_\lambda(\theta^M_1) = \exp\left\{ -\lambda \frac{1}{N} \sum_{i=1}^{N} 1 \left[ Y_i \frac{1}{M} \sum_{j=1}^{M} f_{\theta_j}(X_i) < 0 \right] \right\} d\pi^\otimes M(\theta^M_1).$$

**Theorem 3.2.** For any real parameter $\alpha > 1$, with $P^\otimes N$ probability at least $1 - \epsilon$, for any choice of inverse temperature $\lambda \in [1, N]$,

$$\hat{\rho}_\lambda\left\{ R[\Psi(\theta^M_1)] \right\} \leq \left(1 - g\left(\frac{\alpha\lambda}{N}\right)\right)^{-1} \frac{1}{\lambda} \left\{ -\log\left( \pi^\otimes M \left[ \exp\left[ -\lambda r[\Psi(\theta^M_1)] \right] \right] \right) \right\}$$

$$+ \log(\epsilon^{-1}) + \log \left( \frac{\log(N)}{\log(\alpha)} + 1 \right) \right\}.$$

A union bound can furthermore be used to optimize the value of $M$.

The posterior $\hat{\rho}_\lambda$ can be simulated using a Metropolis algorithm at temperature $\lambda$. A simulated annealing scheme can be useful to compute an approximation of the right-hand side. Indeed, as we have already mentioned, we can write

$$-\log\left( \pi^\otimes M \left[ \exp\left[ -\lambda r[\Psi(\theta^M_1)] \right] \right] \right) = \int_{0}^{\lambda} \hat{\rho}_\gamma\left\{ r[\Psi(\theta^M_1)] \right\} d\gamma$$

$$\leq \sum_{k=1}^{m} (\beta_k - \beta_{k-1}) \hat{\rho}_{\beta_{k-1}}\left\{ r[\Psi(\theta^M_1)] \right\}$$

for any sequence of inverse temperatures $\beta_0 = 0 < \beta_1 < \cdots < \beta_m = \lambda$. This leads to the following computation scheme: estimate $\hat{\rho}_{\beta_k}\left\{ r[\Psi(\theta^M_1)] \right\}$ for increasing values of $\beta_k$ and compute the bound for $\hat{\rho}_{\beta_k}\left\{ R[\Psi(\theta^M_1)] \right\}$. Keep the temperature with the best bound. If we do not trust the constants in the bound, we can keep the highest temperature for which the bound is not more than a certain level above its minimum value. This could lead to less regularized estimators while keeping some warranty against overfitting.

In practice, one of the most successful rule for aggregating classification rules is the boosting algorithm. We refer to [19] for more informations on this topic, and to [20] which explains how the present approach can be used to study classification rules of the boosting type.
3.4. Comments. The results of this section have at least two weaknesses:
- the penalty $\mathcal{K}(\rho, \pi)$ is not as local as it could be;
- noisy samples are not handled properly.

We have also in mind to make some connection between the penalty terms presented here and Vapnik’s entropy. This is to be the subject of the three following sections.

4. Localized learning lemmas

The loss of localization in the use we made so far of lemma 3.2 came from the choice of $\eta(\theta)$: it was chosen to make the contribution of each $\theta$ in the level of confidence equal to $\epsilon$, whereas close to optimal values of $\theta$ may be expected to play a more critical role than others.

Better localization is achieved by choosing
$$
\eta(\theta) = \frac{\lambda^2}{N} g \left( \frac{\lambda}{N} \right) R(\theta) + \beta R(\theta) + \log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\} + \log(\epsilon^{-1}),
$$
leading to

$$(4.1) \quad P^\otimes N \left\{ \sup_{\rho \in \mathcal{M}_1^\Theta} \left( \lambda - \beta - \kappa \frac{\lambda^2}{N} \right) \rho [R(\theta)] - \lambda \rho [r(\theta)] 
- \mathcal{K}(\rho, \pi) - \log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\} \geq \log(\epsilon^{-1}) \right\} \leq \epsilon,$$

where we have put as usual $\kappa = g(\frac{\lambda}{N})$ for short. With the same choice of parameters, the reverse inequality reads as

$$(4.2) \quad P^\otimes N \left\{ \sup_{\rho \in \mathcal{M}_1^\Theta} \lambda \rho [r(\theta)] - \left( \lambda + \beta + \kappa \frac{\lambda^2}{N} \right) \rho [R(\theta)] 
- \mathcal{K}(\rho, \pi) - \log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\} \geq \log(\epsilon^{-1}) \right\} \leq \epsilon,$$

To exploit these inequalities, we need an empirical upper bound for $\log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\}$. This is where the reverse inequality (4.2) comes into play: with $P^\otimes N$ probability at least $1 - \epsilon$,

$$
\log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\} = \sup_{\rho \in \mathcal{M}_1^\Theta} -\beta \rho [R(\theta)] - \mathcal{K}(\rho, \pi)
\leq \sup_{\rho \in \mathcal{M}_1^\Theta} \beta \left( \lambda + \beta + \kappa \frac{\lambda^2}{N} \right)^{-1} \left\{ -\lambda \rho [r(\theta)] + \mathcal{K}(\rho, \pi) 
+ \log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\} \right\} + \log(\epsilon^{-1}) - \mathcal{K}(\rho, \pi)
$$

Putting $\xi = \frac{\beta}{\lambda + \kappa \frac{\lambda^2}{N}}$, this can be rewritten as

$$
\log \left\{ \pi \left[ \exp \left[ -\beta R(\theta) \right] \right] \right\} \leq \sup_{\rho \in \mathcal{M}_1^\Theta} \left\{ -\xi \lambda \rho [r(\theta)] - \mathcal{K}(\rho, \pi) \right\} + \xi \log(\epsilon^{-1})
= \log \left\{ \pi \left[ \exp \left[ -\xi \lambda r(\theta) \right] \right] \right\} + \xi \log(\epsilon^{-1}).
$$
Combining this result with (4.1), we get

**Lemma 4.1** (localized learning lemma, first form). For any \( \lambda \in \mathbb{R}_+ \) and \( \xi \in [0, 1] \)

\[
P \otimes N \left\{ \sup_{\rho \in \mathcal{M}_1^+} \left( (1 - \xi) \lambda - (1 + \xi) \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] - \lambda \rho[r(\theta)] - K(\rho, \pi) - \log \left\{ \pi \left[ \exp \left[ -\xi \lambda r(\theta) \right] \right] \right\} \geq (1 + \xi) \log \left( \frac{2}{\epsilon} \right) \right\} \leq \epsilon
\]

Another way to write this localized learning lemma is to remark that for any \( \rho \in \mathcal{M}_1^+ \),

\[
K(\rho, \pi) + \log \left\{ \pi \left[ \exp \left[ -\xi \lambda r(\theta) \right] \right] \right\} = K(\rho, \pi \exp \left[ -\xi \lambda r(\theta) \right]) - \xi \lambda \rho[r(\theta)],
\]

where

\[
d\pi \exp \left[ -\xi \lambda r(\theta) \right](\theta) = \frac{\exp \left[ -\xi \lambda r(\theta) \right]}{\pi \left[ \exp \left[ -\xi \lambda r(\theta) \right] \right]} d\pi(\theta).
\]

**Lemma 4.2** (localized learning lemma, second form). For any \( \lambda \in \mathbb{R}_+ \) and \( \xi \in [0, 1] \)

\[
P \otimes N \left\{ \sup_{\rho \in \mathcal{M}_1^+} \left( (1 - \xi) \lambda - (1 + \xi) \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] - (1 - \xi) \lambda \rho[r(\theta)] - K(\rho, \pi \exp \left[ -\xi \lambda r(\theta) \right]) \geq (1 + \xi) \log \left( \frac{2}{\epsilon} \right) \right\} \leq \epsilon
\]

Note that the newly introduced parameter \( \xi \) controls the level of localization of the bound: the value \( \xi = 0 \) corresponds to the non localized learning lemma (up to some minor loss in the confidence level).

Applying the first form of the localized learning lemma, we see that the optimal posterior \( \hat{\rho}_\lambda \) is of the same form as in the non localized case:

\[
d\hat{\rho}_\lambda(\theta) = \frac{\exp \left[ -\lambda r(\theta) \right]}{\pi \left[ \exp \left[ -\lambda r(\theta) \right] \right]} d\pi(\theta).
\]

It satisfies

**Corollary 4.1.** With \( P \otimes N \) probability at least \( 1 - \epsilon \),

\[
\hat{\rho}_\lambda[R(\theta)] \leq \left\{ (1 - \xi) \lambda - (1 + \xi) \kappa \frac{\lambda^2}{N} \right\}^{-1} \left\{ \log \left\{ \pi \left[ \exp \left[ -\lambda r(\theta) \right] \right] \right\} + \log \left\{ \pi \left[ \exp \left[ -\xi \lambda r(\theta) \right] \right] \right\} + (1 + \xi) \log \left( \frac{2}{\epsilon} \right) \right\}
\]

\[
= \left( 1 - \frac{1 + \xi}{1 - \xi} \kappa \frac{\lambda}{N} \right)^{-1} \left\{ \frac{1}{1 - \xi} \int_{\xi}^{1} \hat{\rho}_\beta r(\theta) d\beta + \frac{(1 + \xi)}{(1 - \xi)\lambda} \log \left( \frac{2}{\epsilon} \right) \right\}
\]

\[
\leq \left( 1 - \frac{1 + \xi}{1 - \xi} \kappa \frac{\lambda}{N} \right)^{-1} \left\{ \hat{\rho}_\lambda[r(\theta)] + \frac{(1 + \xi)}{(1 - \xi)\lambda} \log \left( \frac{2}{\epsilon} \right) \right\}
\]
Let us remark that this theorem is quite satisfactory from the point of view of localization. It says that the performance of the Gibbs randomized estimator on the observed sample used for training can be trusted to be the same as it will be on previously unseen patterns, up to some penalty factors which do not depend on the size of the model and some increase of the temperature from \( \frac{1}{\lambda} \) to \( \frac{1}{\xi \lambda} \): it can be said that the complexity of the model is taken into account by the Gibbs estimator in an automated way.

To get the corresponding theoretical bound, we can come back to (4.1) to see that for any fixed probability measure \( \rho \in \mathcal{M}_1^p(\Theta) \), with \( P^N \) probability at least \( 1 - \epsilon \),

\[
\hat{\rho}_\lambda [R(\theta)] \leq \left[ \lambda - \beta - \kappa \frac{\lambda^2}{N} \right]^{-1} \left\{ \lambda \rho[r(\theta)] + \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[ \exp \left[ \beta R(\theta) \right] \right] + \log(\epsilon^{-1}) \right\} \right.
\]

Moreover, from Bernstein’s inequality (3.2), with \( P^N \) probability at least \( 1 - \epsilon \),

\[
\lambda \rho[r(\theta)] \leq \left( \lambda + \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] + \log(\epsilon^{-1}).
\]

Thus, putting \( \beta = \xi \lambda \), with \( P^N \) probability at least \( 1 - \epsilon \),

\[
\hat{\rho}_\lambda [R(\theta)] \leq \left[ (1 - \xi) \lambda - \kappa \frac{\lambda^2}{N} \right]^{-1} \left\{ \left( \lambda + \kappa \frac{\lambda^2}{N} \right) \rho[R(\theta)] + \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[ \exp \left[ -\xi \lambda R(\theta) \right] \right] + 2 \log \left( \frac{2}{\epsilon} \right) \right\} \right.
\]

As explained in the case of non localized bounds, the right-hand side being non random can be optimized in \( \rho \), leading to

**Corollary 4.2.** With \( P^N \) probability at least \( 1 - \epsilon \),

\[
\hat{\rho}_\lambda [R(\theta)] \leq \inf_{\xi \in [0, 1]} \left( 1 - \frac{\kappa}{1 - \xi} \right)^{-1} \left\{ \frac{1}{1 - \xi} \int_{\xi}^{1 + \kappa \frac{\lambda}{N}} \pi \left[ \exp \left[ -\beta \lambda R(\theta) \right] \right] R(\theta) d\beta + \frac{2}{(1 - \xi) \lambda} \log \left( \frac{2}{\epsilon} \right) \right\}
\]

\[
\leq \inf_{\xi \in [0, 1]} \left( 1 - \frac{1}{1 - \xi} \right)^{-1} \left\{ \left( 1 + \kappa \frac{\lambda}{N} \right) \pi \left[ \exp \left[ -\xi \lambda R(\theta) \right] \right] [R(\theta)] + \frac{2}{(1 - \xi) \lambda} \log \left( \frac{2}{\epsilon} \right) \right\}.
\]

Let us show now how to make corollary 4.1 uniform in \( \lambda \) and \( \xi \), as it is desirable to optimize these two constants.

Let us first use a union bound on \( \lambda \) for a fixed value of \( \xi \). Let \( \zeta \) be some constant in \([\xi, 1]\), (we can for instance choose \( \zeta = \max \{\xi, \frac{1}{2} \} \)) and let

\[
\Lambda = \left\{ 2N \zeta^k, 0 \leq k < \frac{\log(2N)}{\log(\zeta^{-1})} \right\}.
\]

Olivier Catoni – February 13, 2003  Classification and model selection
CHAPTER 1. A PAC-BAYESIAN APPROACH TO ADAPTIVE INFERENCE

For any \( \lambda \in [1, 2N] \), let \( \lambda' \in \Lambda \) be such that \( \zeta \lambda' \leq \lambda \leq \lambda' \). From lemma 4.1 we deduce that with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any \( \lambda \in [1, 2N] \),

\[
\hat{\rho}_\lambda [R(\theta)] \leq \left( 1 - \xi (1 + (1 + \xi g(\frac{\lambda'}{N})) \right)^{-1} \left\{ \lambda' \hat{\rho}_\lambda [r(\theta)] + \mathcal{K}(\hat{\rho}_\lambda, \pi) \right\} + \log \left\{ \pi \left[ \exp \left[ -\xi \lambda' r(\theta) \right] \right] \right\} (1 + \xi) \log \left( \frac{2}{\epsilon} \frac{\log(2N)}{\epsilon \log(\zeta^{-1})} \right).
\]

We can now use the fact that

\[
\mathcal{K}(\hat{\rho}_\lambda, \pi) = -\log \left\{ \pi \left[ \exp \left[ -\lambda r(\theta) \right] \right] - \lambda \hat{\rho}_\lambda [r(\theta)] \right\}
\]

to get

\[
\hat{\rho}_\lambda [R(\theta)] \leq \left( 1 - \xi (1 + (1 + \xi g(\frac{\lambda'}{N})) \right)^{-1} \left\{ (1 - \frac{\lambda}{\lambda'}) \hat{\rho}_\lambda [r(\theta)] \right\} + \int_{\xi}^{\lambda/\lambda'} \hat{\rho}_{\lambda' \lambda} [r(\theta)]d\beta + \frac{(1 + \xi)}{\lambda'} \log \left( \frac{2}{\epsilon} \frac{\log(2N)}{\epsilon \log(\zeta^{-1})} \right).
\]

Let us remark now that

\[
\int_{\xi}^{\lambda/\lambda'} \hat{\rho}_{\lambda' \lambda} [r(\theta)]d\beta + \left( 1 - \frac{\lambda}{\lambda'} \right) \hat{\rho}_\lambda [r(\theta)] \leq \int_{\xi}^{\lambda/\lambda'} \hat{\rho}_{\lambda' \lambda} [r(\theta)]d\beta + \int_{\lambda/\lambda'}^{1} \hat{\rho}_{\lambda \lambda} [r(\theta)]d\beta = \int_{\xi}^{1} \hat{\rho}_{\lambda \lambda} [r(\theta)]d\beta.
\]

We have proved

**Corollary 4.3.** For any \( \xi \in [0, 1[ \), any \( \zeta \in [\xi, 1[ \), with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any \( \lambda \in [1, 2N] \),

\[
\hat{\rho}_\lambda [R(\theta)] \leq \frac{1}{1 - \xi} \int_{\xi}^{1} \hat{\rho}_{\lambda' \lambda} [r(\theta)]d\beta + \frac{(1 + \xi)}{(1 - \xi) \lambda} \log \left( \frac{2 \log(2N)}{\epsilon \log(\zeta^{-1})} \right) \frac{\log(N)}{\log(\alpha^{-1})}.
\]

Comparing this results with corollary 4.1 shows that gaining uniformity in \( \lambda \) is quite harmless to the quality of the bound. We can of course now go further by using a union bound for different values of \( \xi \). Since the bound explodes when \( \xi = 1 \) and the degree of localization is linked with the order of magnitude of \( \xi \), we would suggest a discretization set for \( \xi \) of the form

\[
\left\{ \alpha^k, 1 \leq k \leq \frac{\log(N)}{\log(\alpha^{-1})} \right\}.
\]
If we want to choose \( \alpha \) as a function of \( N \) and still avoid introducing \( \log(N) \) factors in the bound, we can for instance choose \( \alpha = 1 - \frac{1}{\log(N)} \).

5. Noisy pattern recognition

The mathematical setting is the same as previously, and we assume without further notice that there exists a regular version of the conditional probability measure \( P(Y \mid X) \). In this section, we are going to bring further improvements in the case when \( \inf_{\theta \in \Theta} R(\theta) > 0 \). This can result from various causes:

- The observed sample may be “noisy” in the sense that it is drawn according to a joint distribution \( P \) for which the best achievable error rate for pattern \( x \), \( \inf_{\theta \in \Theta} P(Y \neq y \mid X = x) \) is large for many patterns. This noise may come either from an inherently ambiguous classification task or from errors made in labeling the training examples.
- Even if the sample is not noisy, the best available classification rule may be poor.

The theoretical bounds in the previous section where at best of order \( \inf_{\theta \in \Theta} R(\theta) + \sqrt{\frac{R(\theta)}{N}} + \frac{c}{\sqrt{N}} \), leading to a convergence speed not faster than \( \frac{1}{\sqrt{N}} \) in the case of a noisy sample. We will improve this rate in the case when some classification rule \( f_\theta \) produces the most likely label among all the available rules for a strong majority of patterns.

To formulate this we will consider some distinguished classification rule \( f_\theta \).

The most favorable case is when \( R(\theta) = \inf_{\theta \in \Theta} R(\theta) \), but this condition will not be strictly imposed here. The case when \( \theta \notin \Theta \) makes no difference: it is covered by adding \( \theta \) to the parameter set \( \Theta \) and extending the prior \( \pi \) putting \( \pi(\theta) = 0 \). Of course \( \theta \), whose clever choice is bound to depend on \( P \), is not assumed to be known by the statistician!

Let us introduce the following relative quantities, where \( \text{Var}_P \) denotes the variance with respect to \( P \):

\[
\begin{align*}
\overline{R}(\theta) &= P[Y \neq f_\theta(X)] - P[Y \neq f_\theta(X)] \\
\pi(\theta) &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[Y_i \neq f_\theta(X_i)] - \mathbb{I}[Y_i \neq f_\theta(X_i)] \\
\overline{V}(\theta) &= \text{Var}_P \left\{ \mathbb{I}[Y \neq f_\theta(X)] - [Y \neq f_\theta(X)] \right\} \\
\overline{R}(\theta \mid X) &= P[Y \neq f_\theta(X) \mid X] - P[Y \neq f_\theta(X) \mid X] \\
\overline{V}(\theta \mid X) &= \text{Var}_P \left\{ \mathbb{I}[Y \neq f_\theta(X)] - [Y \neq f_\theta(X)] \mid X \right\}.
\end{align*}
\]

Let us define for any pattern \( x \in X \) the margin \( \alpha(x) \) of success of \( f_\theta(x) \) as

\[
\alpha(x) = \min \left\{ \overline{R}(\theta \mid x), \theta \in \Theta, f_\theta(x) \neq f_\theta(x) \right\}.
\]

(In this formula we assume that some realization of the conditional expectations has been chosen once for all). Note that \( \alpha(x) \) may be negative in the case when \( f_\theta(x) \) is not the most likely label for pattern \( x \).
CHAPTER 1. A PAC-BAYESIAN APPROACH TO ADAPTIVE INFERENCE

Thresholding the margin $\alpha(x)$ at level $\alpha$ defines some exceptional set $\Omega_\alpha$ of “$\alpha$-ambiguous” patterns:

$$\Omega_\alpha \overset{\text{def}}{=} \{ x \in X : \alpha(x) < \alpha \}.$$  

We introduce this notion of ambiguity to control the variance $\nabla(\theta)$ by the mean $R(\theta)$ of the relative error rate. Indeed

$$\nabla(\theta) = P[\nabla(\theta \mid X)] + \text{Var}[R(\theta \mid X)]$$

$$\leq P\left[ \frac{R(\theta \mid X)}{\alpha} \mathbb{1}(X \notin \Omega_\alpha) + \mathbb{1}(\Omega_\alpha) \right] + P[R(\theta \mid X)^2]$$

$$\leq \frac{1}{\alpha} \left[ R(\theta) + P(\Omega_0) \right] + P(\Omega_\alpha) + 2P(\Omega_0)$$

$$= aR(\theta) + b,$$

where we have put

$$a = \left( \frac{1}{\alpha} + 1 \right),$$

$$b = \left( \frac{1}{\alpha} + 2 \right) P(\Omega_0) + P(\Omega_\alpha).$$

Applying Bernstein’s theorem 3.1 in a way similar to what has already been done to establish lemma 3.2 in the previous section, we get some non localized learning lemma

**Lemma 5.1.** For any $\lambda \in \mathbb{R}_+$, any measurable function $\eta : \Theta \rightarrow \mathbb{R}$,

$$P \otimes \mathcal{N} \left\{ \sup_{\rho \in \mathcal{M}_1} \lambda \rho[R(\theta)] - \lambda \rho[r(\theta)] - K(\rho, \pi) - \eta(\theta) \geq 0 \right\}$$

$$\leq \pi \left\{ \exp \left[ g \left( \frac{1 + R(\theta)}{N} \right) \lambda \left( aR(\theta) + b \right) \frac{\lambda^2}{N} - \eta(\theta) \right] \right\}.$$  

In the same way

$$P \otimes \mathcal{N} \left\{ \sup_{\rho \in \mathcal{M}_1} \lambda \rho[R(\theta)] - \lambda \rho[r(\theta)] - K(\rho, \pi) - \eta(\theta) \geq 0 \right\}$$

$$\leq \pi \left\{ \exp \left[ g \left( \frac{1 - R(\theta)}{N} \right) \lambda \left( aR(\theta) + b \right) \frac{\lambda^2}{N} - \eta(\theta) \right] \right\}.$$  

**5.1. Non localized results.** Putting $\kappa = g \left( \frac{2\lambda}{N} \right)$ and taking

$$\eta(\theta) = \kappa [aR(\theta) + b] \frac{\lambda^2}{N} + \log(\epsilon^{-1}),$$

we get

**Corollary 5.1.** With $P \otimes \mathcal{N}$ probability at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_1(\Theta)$,

$$\rho[R(\theta)] \leq R(\tilde{\theta}) + \left( 1 - \kappa a \frac{\lambda}{N} \right)^{-1} \left\{ \rho[r(\theta)] - r(\tilde{\theta}) \right\}.$$
Therefore, we achieve a rate of convergence of $1/N$ and comparing values of $\lambda$

Using a grid $\Lambda = \{R \in \mathbb{R} : 0 \leq R \leq \log(N)/\log(2)\}$, a union bound for this grid, and comparing values of $\lambda \in [1, N]$ with the next value in the grid, we get

**Corollary 5.3.** With $P^\otimes N$ at least $1 - \epsilon$, for any posterior $\rho \in \mathcal{M}_1^+(\Theta)$,

$$
\rho[R(\theta)] \leq R(\bar{\theta}) + \inf_{\lambda \in [1, N]} \left(1 - \kappa a \frac{2\lambda}{N}\right)^{-1} \left\{ \rho[R(\theta)] - r(\bar{\theta}) + \frac{1}{\lambda} \left[ \mathcal{K}(\rho, \pi) + \log \left( \frac{\log(N)}{2\epsilon} \right) \right] + \kappa b \frac{\lambda}{N} \right\}.
$$

Note that to perform the optimization in $\lambda$ from empirical data, we need first to apply the empirical bound $-r(\bar{\theta}) \leq -\inf_{\theta \in \Theta} r(\theta)$.}

Olivier Catoni – February 13, 2003

Classification and model selection
5.2. Localized results. A localized learning lemma can be established exactly as explained in the previous section. It requires to choose
\[
\eta(\theta) = \kappa [aR(\theta) + b] \lambda^2 N + \beta R(\theta) + \log \left\{ \pi \left[ \exp \left( -\beta R(\theta) \right) \right] \right\} + \log(\epsilon^{-1}),
\]
where \( \kappa = g(\frac{\lambda^2}{N}) \).

Lemma 5.2. With \( P^{\otimes N} \) probability at least \( 1 - \epsilon \),
\[
\rho \left[ R(\theta) \right] \leq \left( \lambda - \beta - \kappa a \lambda^2 \right)^{-1} \left\{ -\lambda \rho \left[ \rho(\theta) \right] + \mathcal{K}(\rho, \pi) \right. \\
+ \left. \sup_{\rho \in \mathcal{M}_+^1(\theta)} \left\{ -\lambda \rho \left[ \rho(\theta) \right] + \mathcal{K}(\rho, \pi) \right\} \right\}.
\]

In the same way,
\[
-\rho \left[ R(\theta) \right] \leq \left( \lambda + \beta + \kappa a \lambda^2 \right)^{-1} \left\{ -\lambda \rho \left[ \rho(\theta) \right] + \mathcal{K}(\rho, \pi) \right. \\
+ \left. \sup_{\rho \in \mathcal{M}_+^1(\theta)} \left\{ -\lambda \rho \left[ \rho(\theta) \right] + \mathcal{K}(\rho, \pi) \right\} \right\}.
\]

Putting \( \xi = \frac{\beta}{\lambda(1 + \kappa a \lambda^2)} \), we see that with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \),
\[
\log \left\{ \pi \left[ \exp \left( -\beta R(\theta) \right) \right] \right\} = \sup_{\rho \in \mathcal{M}_+^1(\theta)} \left\{ -\beta \rho \left[ \rho(\theta) \right] - \mathcal{K}(\rho, \pi) \right\} \leq \sup_{\rho \in \mathcal{M}_+^1(\theta)} \frac{\xi}{1 + \xi} \left\{ -\lambda \rho \left[ \rho(\theta) \right] + \mathcal{K}(\rho, \pi) \right\} \right. \\
+ \left. \sup_{\rho \in \mathcal{M}_+^1(\theta)} \left\{ -\lambda \rho \left[ \rho(\theta) \right] + \mathcal{K}(\rho, \pi) \right\} \right\}.
\]

which can be rewritten as
\[
\log \left\{ \pi \left[ \exp \left( -\beta R(\theta) \right) \right] \right\} \leq \sup_{\rho \in \mathcal{M}_+^1(\theta)} \left\{ -\xi \lambda \rho \left[ \rho(\theta) \right] + \xi \log(\epsilon^{-1}) + \xi b \lambda^2 \right\}
\]
\[
= \log \left\{ \pi \left[ \exp \left( -\xi \lambda \rho \left[ \rho(\theta) \right] \right) \right] \right\} + \xi \log(\epsilon^{-1}) + \xi b \lambda^2 \right\}.
\]

Coming back to lemma 5.2, we obtain

Corollary 5.4. With \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any posterior \( \rho \in \mathcal{M}_+^1 \),
\[
\rho \left[ R(\theta) \right] - R(\tilde{\theta}) \leq \left( 1 - \frac{1 + \xi}{1 - \xi} \kappa a \lambda^2 \right)^{-1} \left\{ \rho \left[ r(\theta) \right] - r(\tilde{\theta}) \right\} \\
+ \frac{1}{(1 - \xi) \lambda} \left[ \mathcal{K}(\rho, \rho_{\xi \lambda}) + (1 + \xi) \log(\frac{\hat{\xi}}{\xi}) \right] + \kappa \frac{1 + \xi}{1 - \xi} b \lambda N \right\}
\]
\[
= \left( 1 - \frac{1 + \xi}{1 - \xi} \kappa a \lambda^2 \right)^{-1} \left\{ \frac{1}{(1 - \xi) \lambda} \left[ \lambda \rho \left[ r(\theta) \right] + \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[ \exp \left( -\xi \lambda r(\theta) \right) \right] \right\} \right] \\
- r(\tilde{\theta}) + \frac{1 + \xi}{1 - \xi} \frac{\log(\hat{\xi})}{\lambda} + \kappa b \lambda \right\}.
\]
The optimal posterior according to this bound is the Gibbs distribution \( \hat{\rho}_\lambda \). It is such that

\[
\hat{\rho}_\lambda[R(\theta)] - R(\tilde{\theta}) \leq \left( 1 - \frac{1 + \xi}{1 - \xi} \frac{\kappa a \lambda}{N} \right)^{-1} \left\{ \frac{1}{1 - \xi} \int_{\xi}^{1} \hat{\rho}_{\beta \lambda}[r(\theta)] d\beta \right\} \leq \hat{\rho}_{\lambda}[r(\theta)]
- r(\tilde{\theta}) + \frac{1 + \xi}{1 - \xi} \left[ \log(\xi) \frac{\lambda}{\lambda} + \kappa b \frac{\lambda}{N} \right].
\]

For a fixed value of \( \xi \), getting a uniform result in \( \lambda \) is achieved as in the case of corollary 4.3:

**Corollary 5.5.** For any \( \xi \in [0, 1] \), any \( \zeta \in [\xi, 1] \), with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any \( \lambda \in [1, N] \),

\[
\hat{\rho}_\lambda[R(\theta)] - R(\tilde{\theta}) \leq \left( 1 - \frac{1 + \xi}{1 - \xi} \frac{\kappa a \lambda}{\zeta N} \right)^{-1} \left\{ \frac{1}{1 - \xi} \int_{\xi}^{1} \hat{\rho}_{\beta \lambda}[r(\theta)] d\beta - r(\tilde{\theta}) \right\} \leq \hat{\rho}_{\lambda}[r(\theta)]
+ \frac{1 + \xi}{1 - \xi} \left[ \frac{1}{\lambda} \log \left( \frac{2 \log(N)}{\epsilon \log(\zeta^{-1})} \right) + \kappa b \frac{\lambda}{\zeta N} \right].
\]

The same remarks which were made about corollary 4.3 apply here: a union bound on different values of \( \xi \) can furthermore be performed. Let us also notice that optimizing the bound in \( \lambda \) from observations requires to use the empirical bound \( -r(\tilde{\theta}) \leq - \inf_{\theta \in \Theta} r(\theta) \).
CHAPTER 2

Learning with an exchangeable prior

1. The Vapnik Cervonenkis dimension of a family of subsets

Let us consider a set of \( n \) points \( X = \{x_1, \ldots, x_n\} \) and some set \( S \subset \{0, 1\}^X \) of subsets of \( X \). Let \( h(S) \) be the VC dimension of \( S \), defined as

\[
h(S) = \max \{|A| : A \subset X \text{ and } |A \cap S| = 2^{|A|}\},
\]

where by definition \( A \cap S = \{A \cap B : B \in S\} \). Let us notice that this definition does not depend on the choice of the reference set \( X \). Indeed \( X \) can be chosen to be \( \bigcup S \), the union of all the sets in \( S \). On the other hand, it can also be chosen to be some maybe infinite measurable set containing \( \bigcup S \).

This notion of VC dimension is useful because it can, as we will see about support vector machines, be computed in some important special cases. Let us prove here as an illustration that \( h(S) \leq d + 1 \) whenever \( X \subset \mathbb{R}^d \) and \( S \) is made of all the intersections of \( X \) with half spaces :

\[
S = \{A_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}, \text{ where } A_{w,b} = \{x \in X : \langle w, x \rangle \geq b\}.
\]

**Proposition 1.1.**

With the previous notations, \( h(S) \leq d + 1 \).

**Proof.** We have to prove that for any set \( A \subset \mathbb{R}^d \) of size \( |A| = d + 2 \), there is \( B \subset A \) such that \( B \not\subset (A \cap S) \). This will obviously be the case if the convex hulls of \( B \) and \( A \setminus B \) have a non empty intersection : indeed if a hyperplane separates two sets of points, it also separates their convex hulls. As \( |A| > d + 1 \), \( A \) is affine dependent : there is \( (\lambda_x)_{x \in A} \in \mathbb{R}^{d+2} \setminus \{0\} \) such that \( \sum_{x \in A} \lambda_x x = 0 \) and \( \sum_{x \in A} \lambda_x = 0 \). The set \( B = \{x \in A : \lambda_x > 0\} \) is non-empty, as well as its complement \( A \setminus B \), because \( \sum_{x \in A} \lambda_x = 0 \) and \( \lambda \neq 0 \). Moreover \( \sum_{x \in B} \lambda_x = \sum_{x \in A \setminus B} -\lambda_x > 0 \). The relation

\[
\sum_{x \in B} \lambda_x \sum_{x \in B} \lambda_x x = \frac{1}{\sum_{x \in B} \lambda_x} \sum_{x \in A \setminus B} -\lambda_x x
\]

shows that the convex hulls of \( B \) and \( A \setminus B \) have a non void intersection. \( \square \)

Let us introduce the function

\[
\Phi_n^h = \sum_{k=0}^h \binom{n}{k}
\]

Let us notice that \( \Phi \) can alternatively be defined by the relations :

\[
\Phi_n^h = \begin{cases} 2^n & \text{when } n \leq h, \\ \Phi_{n-1}^{h-1} + \Phi_{n-1}^h & \text{when } n > h. \end{cases}
\]

**Theorem 1.1.**

\[ |S| \leq \Phi \left( |\bigcup S|, h(S) \right). \]
Theorem 1.2.
\[ \Phi(n, h) \leq \exp(nH(\frac{h}{n})) \leq \exp\left[h\left(\log(\frac{h}{n}) + 1\right)\right], \]
where \( H(p) = -p \log(p) - (1-p) \log(1-p) \) is the Shannon entropy of the Bernoulli distribution with parameter \( p \).

Proof of Theorem 1.1. Let us prove this theorem by induction on \(|\bigcup S|\). It is easy to check that it holds true when \(|\bigcup S| = 1\). Let \( X = \bigcup S \), let \( x \in X \) and \( X' = X \setminus \{x\} \). Define
\[ S' = \{ A \in S : A \triangle \{x\} \in S \}, \]
\[ S'' = \{ A \in S : A \triangle \{x\} \notin S \}. \]
Clearly \( S = S' \cup S'' \) and \( S \cap X' = (S' \cap X') \cup (S'' \cap X') \). Moreover \(|S'| = 2|S' \cap X'|\) and \(|S''| = |S'' \cap X'|\). Thus \(|S| = |S'| + |S''| = 2|S' \cap X'| + |S''| = |S \cap X'| + |S' \cap X'|\). Obviously \( h(S \cap X') \leq h(S) \). Moreover \( h(S' \cap X') = h(S') - 1 \), because if \( A \subset X' \) is shattered by \( S' \) (or equivalently by \( S' \cap X' \)), then \( A \cup \{x\} \) is shattered by \( S' \) (we say that \( A \) is shattered by \( S \) when \( S \cap A = \{0,1\}^A \)). Using the induction hypothesis, we then see that \(|S \cap X'| \leq \Phi_{|X'|}^{|S'|} + \Phi_{|X'|}^{|S'| - 1}\). But as \(|X'| = |X| - 1\), the righthand side of this inequality is equal to \( \Phi_{|X|}^{|S|} \), according to the recurrence equation satisfied by \( \Phi \).

Proof of Theorem 1.2. This is the well known Chernoff bound for the deviation of sums of Bernoulli r.v.: let \( (\sigma_1, \ldots, \sigma_n) \) be i.i.d. Bernoulli r.v. with parameter 1/2. Let us notice that
\[ \Phi_n^h = 2^n \mathbb{P}\left( \sum_{i=1}^n \sigma_i \leq h \right). \]
For any positive real number \( \lambda \),
\[ \mathbb{P}\left( \sum_{i=1}^n \sigma_i \leq h \right) \leq \exp(\lambda h) \mathbb{E}\left[ \exp\left(-\lambda \sum_{i=1}^n \sigma_i\right) \right] = \exp\left\{ \lambda h + n \log\{\mathbb{E}[\exp(-\lambda \sigma_i)]\} \right\}. \]
Differentiating the right-hand side in \( \lambda \) shows that its minimal value is \( \exp\left[-n\mathcal{K}(\frac{h}{n}, \frac{1}{2})\right] \), where \( \mathcal{K}(p, q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right) \) is the Kullback divergence function between two Bernoulli distributions of parameters \( p \) and \( q \). The announced result then follows from the identity \( H(p) = \log(2) - \mathcal{K}(p, \frac{1}{2}) \).

2. Non localized bounds

In this chapter we assume that \( P_{2N} \) is some exchangeable distribution on \((X \times \mathcal{Y})^{2N}\), where \( (X, \mathcal{B}) \) is as previously a measurable space of patterns and \( \mathcal{Y} \) a finite set of labels. We assume that we observe \((X_1, \ldots, X_N), (Y_1, \ldots, Y_N)\) and possibly also \((X_{N+1}, \ldots, X_{2N})\). In other words half of the patterns are labeled and half of the patterns have to be labeled. Starting with a family \( \{f_\theta : X \to \mathcal{Y}; \theta \in \Theta\} \) of classification rules, we would like to minimize
\[ r_2(\theta) = \frac{1}{N} \sum_{k=N+1}^{2N} \mathbb{I}(Y_k \neq f_\theta(X_k)). \]
We can apply our PAC-Bayesian methodology in this situation, using an exchangeable prior. The interest of exchangeable priors is that they will provide a way to make a link between PAC-Bayesian theorems and Vapnik’s theory.

Let us first prove a deviation lemma based on the fact that $P_{2N}$ is exchangeable. Let

\[ r_1(\theta) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}[Y_k \neq f_\theta(X_k)] \]

\[ r_2(\theta) = \frac{1}{N} \sum_{k=N+1}^{2N} \mathbb{1}[Y_k \neq f_\theta(X_k)] \]

**Lemma 2.1.** For any exchangeable measurable function $\eta : (\mathcal{X} \times \mathcal{Y})^{2N} \times \Theta \rightarrow \mathbb{R}$, any $\theta \in \Theta$,

\[ P_{2N}\left\{ \exp \left[ \lambda r_2(\theta) - r_1(\theta) - \eta(\theta) \right] \right\} \leq P_{2N}\left\{ \exp \left[ \frac{\lambda^2}{2N} \left( r_1(\theta) + r_2(\theta) \right) - \eta(\theta) \right] \right\}. \]

**Proof.** Let us remember that $\log [\cosh(s)] \leq \frac{1}{2} s^2$ for any $s \in \mathbb{R}$. Let

\[ \sigma_k = \mathbb{1}[Y_k \neq f_\theta(X_k)] \]

Using the fact that $P_{2N}$ is assumed to be exchangeable, we get

\[ P_{2N}\left\{ \exp \left[ \lambda r_2(\theta) - r_1(\theta) - \eta(\theta) \right] \right\} \]

\[ = P_{2N}\left\{ \exp \left[ \frac{\lambda}{N} \sum_{k=1}^{N} [\sigma_{k+2N}(\theta) - \sigma_k(\theta)] - \eta(\theta) \right] \right\} \]

\[ = P_{2N}\left\{ \exp \left[ \sum_{k=1}^{N} \log \left\{ \cosh \left( \frac{\lambda}{N} [\sigma_{k+2N}(\theta) - \sigma_k(\theta)] \right) \right\} - \eta(\theta) \right] \right\} \]

\[ \leq P_{2N}\left\{ \exp \left[ \frac{\lambda^2}{2N^2} \sum_{k=1}^{N} [\sigma_{k+2N}(\theta) - \sigma_k(\theta)]^2 - \eta(\theta) \right] \right\} \]

\[ \leq P_{2N}\left\{ \exp \left[ \frac{\lambda^2}{2N^2} \sum_{k=1}^{N} [\sigma_{k+2N}(\theta) + \sigma_k(\theta)] - \eta(\theta) \right] \right\} \]

\[ = P_{2N}\left\{ \exp \left[ \frac{\lambda^2}{2N} \left( r_1(\theta) + r_2(\theta) \right) - \eta(\theta) \right] \right\} \]

□

Let us now consider some exchangeable random probability measure $\pi : (\mathcal{X} \times \mathcal{Y})^{2N} \rightarrow \mathcal{M}_1^+(\Theta)$. (We will assume that $(\Theta, \mathcal{T})$ is a Polish space and that $\pi$ is a regular conditional probability measure. Moreover, in practice, interesting exchangeable priors will depend only on $(X_1, \ldots, X_{2N})$, although the forthcoming bounds do not preclude them to depend also on $(Y_1, \ldots, Y_{2N})$.) Integrating the previous deviation lemma with respect to $\pi$, we get

**Lemma 2.2.** For any $\lambda \in \mathbb{R}_+$,

\[ P_{2N}\left\{ \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \theta) \geq 0 \right\} \]
where we have put

\[ \Psi(\theta, \phi) = \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] + \log(\epsilon^{-1}) \]

As in the preceding section, we can deduce from this lemma non-localized or localized results. Let us start with a non localized result.

Choosing \( \eta(\theta) = \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] \) we obtain

**Corollary 2.1.** With \( P_2 \) probability at least \( 1 - \epsilon \), for any posterior \( \rho \in M_0(\Theta) \),

\[
\rho[r_2(\theta)] \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left( \lambda + \frac{\lambda^2}{2N} \right) \rho[r_1(\theta)] + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}) \right\}.
\]

As a special case, we find a result similar to Vapnik’s bounds. Let

\[
N(X_i^2) = \left| \left\{ f_0(X_k) \right\}_{k=1}^{2N} : \theta \in \Theta \right|.
\]

**Corollary 2.2.** With \( P_2 \) probability at least \( 1 - \epsilon \), for any \( \theta \in \Theta \),

\[
r_2(\theta) \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left( \lambda + \frac{\lambda^2}{2N} \right) r_1(\theta) + \log[N(X_i^2)] + \log(\epsilon^{-1}) \right\}.
\]

Note that this is an improvement on classical Vapnik’s theory, since the complexity term \( \log[N(X_i^2)] \) is observable. Note also that we proved in a previous section that in the binary case when \( |Y| = 2 \),

\[
\log[N(X_i^2)] \leq 2NH \left( \frac{\hbar}{2N} \right) \leq h \left[ \log \left( \frac{2N}{\hbar} \right) + 1 \right],
\]

where \( H(p) = -p \log(p) - (1-p) \log(1-p) \) is the Shannon entropy of the Bernoulli distribution with parameter \( p \) and where

\[
h = \max \{|A| : A \subset \{ X_k : 1 \leq k \leq 2N \} \text{ and } |\{ A \cap f_0^{-1}(1) : \theta \in \Theta \}| = 2|A| \}
\]

is the VC dimension of the set \( \{ X_1, \ldots, X_{2N} \} \).

**Proof.** Let

\[ \Psi : \theta \mapsto \left\{ f_\theta(X_k) \right\}_{k=1}^{2N} \in Y^{2N}. \]

For each \( y \in \Psi(\Theta) \), let us choose \( \theta(y) \in \Psi^{-1}(y) \) to form a finite set \( \Theta' \subset \Theta \) of size \( N(X_i^2) \), as the collection \( \{ \Psi^{-1}(y) : y \in \Psi(\Theta) \} \) is an exchangeable function of \( X_i^2 \), the random set \( \Theta'(X_i^{2N}) \) can be chosen to be an exchangeable function of \( X_i^{2N} \). Let \( \pi \) be the uniform measure on \( \Theta' \). Then considering as posterior distribution the Dirac mass at \( \theta' \in \Theta' \), we see that with \( P_2 \) probability at least \( 1 - \epsilon \), for any \( \theta' \in \Theta' \),

\[
r_2(\theta') \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left( \lambda + \frac{\lambda^2}{2N} \right) r_1(\theta') + \log[N(X_i^{2N})] + \log(\epsilon^{-1}) \right\}.
\]

We end the proof with the remark that for any \( \theta \in \Theta \), there is \( \theta' \in \Theta' \) such that \( \Psi(\theta) = \Psi(\theta') \), and therefore such that \( r_1(\theta) = r_1(\theta') \) and \( r_2(\theta) = r_2(\theta') \).

It also makes sense to compare \( r_1(\theta) \) with

\[
R_2(\theta) = P_2 \left[ Y_{N+1} \neq f_\theta(X_{N+1}) \right| Z_{11}^N],
\]

where we have put \( Z_{11}^N = (X_k, Y_k)_{k=1}^N \) for short.

To this purpose we can use a variant of lemma 2.2
LEMMA 2.3. For any $\lambda \in \mathbb{R}_+$,
\[
P_{2N}\left\{P_{2N} \left[ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi)|Z_1^N] \right] \geq 0 \right\} \leq P_{2N}\left\{ \pi \left[ \exp\left( \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] - \eta(\theta) \right) \right] \right\}.
\]

**Proof.** Let
\[
U = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \lambda \rho[r_2(\theta)] - \lambda \rho[r_1(\theta)] - \rho[\eta(\theta)] - \mathcal{K}(\rho, \pi)
\]
and
\[
\epsilon = P_{2N}\left\{ \pi \left[ \exp\left( \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] - \eta(\theta) \right) \right] \right\}.
\]
Then as already proved, $P_{2N}[\exp(U)] \leq \epsilon$. But in the same time, from the convexity of the exponential function,
\[
P_{2N}\left\{ \exp\left( P_{2N}(U | Z_1^N) \right) \right\} \leq P_{2N}[\exp(U)],
\]
as required. \qed

Choosing in lemma 2.3
\[
\eta(\theta) = \frac{\lambda^2}{2N} [r_1(\theta) + r_2(\theta)] + \log(\epsilon^{-1})
\]
we get
\[
\text{COROLLARY 2.3. With } P_{2N}(dZ_1^N) \text{ (the distribution of } Z_1^N \text{ under } P_{2N} \text{) probability at least } 1 - \epsilon, \text{ for any regular conditional probability distribution } \rho : \mathcal{X}^{2N} \times \mathcal{Y}^{2N} \to \mathcal{M}_+^1(\Theta),
\]
\[
P_{2N}\left\{ \rho[r_2(\theta)] | Z_1^N \right\} \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left( \lambda + \frac{\lambda^2}{2N} \right) P_{2N}\left\{ \rho[r_1(\theta)] | Z_1^N \right\} + P_{2N}\left\{ \mathcal{K}(\rho, \pi)|Z_1^N \right\} + \log(\epsilon^{-1}) \right\}.
\]

The interesting case is of course when $\rho$ in fact does not depend on the non observed labels $Y_{N+1}^{2N}$. This may look cumbersome, but has a simple application.

**THEOREM 2.1.** With $P_{2N}(dZ_1^N)$ probability at least $1 - \epsilon$, for any estimator $\hat{\theta} : \mathcal{X}^N \times \mathcal{Y}^N \to \Theta$, assumed to be a measurable function,
\[
R_2(\hat{\theta}) \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1} \left\{ \left( \lambda + \frac{\lambda^2}{2N} \right) r_1(\hat{\theta}) + P_{2N}\left\{ \log[N(X_1^{2N})] | Z_1^N \right\} + \log(\epsilon^{-1}) \right\}.
\]

Note that in the independent case when $P_{2N} = P^\otimes 2N$, then $R_2(\hat{\theta}) = R(\hat{\theta})$ defined in the previous sections.

**Proof.** This is an integrated variant of corollary 2.2. With the same notations, we can define an estimator $\hat{\theta}'$ with values in $\Theta'$, such that $\Psi(\hat{\theta}) = \Psi(\hat{\theta}')$ everywhere. The proof for $\hat{\theta}'$ is a direct consequence of the preceding corollary, and the proof for $\hat{\theta}$ comes from the fact that $r_1(\hat{\theta}) = r_1(\hat{\theta}')$ and $r_2(\hat{\theta}) = r_2(\hat{\theta}')$ everywhere. \qed
Proving some variant of Vapnik’s theory is not the only possible use of corollary 2.1. Its right-hand side can also be optimized choosing for \( \rho \) the Gibbs distribution
\[
d\tilde{\rho}_\beta(\theta) = \frac{\exp[-\beta r_1(\theta)]}{\pi\{\exp[-\beta r_1(\theta)]\}}d\pi(\theta).
\]
(Note that \( \hat{\rho} \) depends not only on \( Z_1^N \) but also on \( X_{N+1}^{2N} \) through \( \pi \).)

**Corollary 2.4.** With \( P_{2N} \) probability at least \( 1 - \epsilon \),
\[
\hat{\rho}_{\lambda + \frac{\lambda^2}{2N}}[r_2(\theta)] \leq \left( \lambda - \frac{\lambda^2}{2N} \right)^{-1}\left\{ -\log\left[ \pi\{\exp[-(\lambda + \frac{\lambda^2}{2N}) r_1(\theta)]\} \right] + \log(\epsilon^{-1}) \right\}
\]
\[
= \frac{1 + \frac{\lambda^2}{2N}}{1 - \frac{\lambda^2}{2N}} \left\{ \frac{1}{\lambda + \frac{\lambda^2}{2N}} \int_0^{\lambda + \frac{\lambda^2}{2N}} \hat{\rho}_\beta[r_1(\theta)]d\beta \right\} + \log(\epsilon^{-1}) \frac{1}{\lambda - \frac{\lambda^2}{2N}}.
\]

3. Some possible applications of learning with an exchangeable prior

Before getting into more sophisticated bounds (localized or tailored for the noisy classification case), let us put forward that the choice of \( \pi \) as a function of \( \sum_{k=1}^{2N} \delta_{X_k} \) opens interesting possibilities.

**3.1. Compression schemes.** Let us explore first the idea of compression schemes, put forward by Littlestone and Warmuth [23, 17].

Let us consider some measurable training rule
\[
f^*: \bigcup_{n=1}^{+\infty}(X \times \mathcal{Y})^n \times X \rightarrow \mathcal{Y},
\]
which produces for any size of problem \( n \) and any training set \( Z' = (x'_i, y'_i)_{i=1}^n \in (X \times \mathcal{Y}) \) some classifier \( f_{Z'}: X \rightarrow \mathcal{Y} \). Let us assume that \( f_{Z'} \) is invariant under any permutation of the indices of the training set, as it is usually the case with estimators designed for i.i.d. or exchangeable samples.

A sample \( Z = (X_i, Y_i)_{i=1}^{2N} \) being given, it is natural to build in this case the following model: for any \( h = 1, \ldots, N \)
\[
\mathcal{R}_h = \{ f_{(x'_i, y'_i)_{i=1}^h} : x'_i : 1 \leq i \leq h \} \subset \{ X_i : 1 \leq i \leq 2N \}, (y'_i)_{i=1}^h \in \mathcal{Y}^h \}.
\]
We will consider the union of all these models: \( \mathcal{R} = \bigcup_{h=1}^N \mathcal{R}_h \). Our exchangeable prior will be uniform on each \( \mathcal{R}_h \), and we will put for some parameter \( \alpha \in [0, 1] \), \( \pi(\mathcal{R}_h) = (1 - \alpha)\alpha^h \) (obtaining in fact a subprobability, which is enough for our purpose).

It is easy to see that
\[
\log |\mathcal{R}_h| = \log \left( \frac{2N}{h} \right)^{|\mathcal{Y}|^h} \leq h \left[ \log \left( \frac{2N}{h} \right) + 1 + \log (|\mathcal{Y}|) \right].
\]

We are ready to apply theorem 2.1, in the case when we observe a training set \( (X_i, Y_i)_{i=1}^N \) under the distribution \( P^{\otimes N} \), where \( P \in \mathcal{M}_+(X) \). According to this theorem, applied to \( P_{2N} = P^{\otimes 2N} \), with \( P^{\otimes N} \) probability at least \( 1 - \epsilon \), for any \( h = 1, \ldots, 2N \), any \( f \in \mathcal{R}_h \),
\[
R(f) \leq \left( 1 - \frac{\lambda}{2N} \right)^{-1}\left\{ \left( 1 + \frac{\lambda}{2N} \right)^r(f) \right\}
\]
\[ + \frac{1}{\lambda} \left[ - \log(1 - \alpha) + h \left[ \log \left( \frac{N^2}{\lambda^2} \right) + 1 + \log(|y|) \right] - \log(\alpha) \right] + \log(\epsilon^{-1}) \right] \}, \]

where as usual \( R(f) = P(f(X) \neq Y) \) and \( r(f) = \frac{1}{N} \sum_{i=1}^{N} 1[f(X_i) \neq Y_i] \).

The next useful step is to make this statement uniform in \( \lambda \in [1, 2N] \). As explained earlier in these lectures, this can be achieved by considering for some real parameter \( \zeta > 1 \) a grid of value \( \Lambda = \{2N\zeta^{-k} : 0 \leq k \leq \frac{\log(2N)}{\log(\zeta)} \} \). Applying the previous inequality of any \( \lambda \in \Lambda \), we get

**Proposition 3.1.** With \( P^\otimes N \) probability at least \( 1 - \epsilon \) for any \( h = 1, \ldots, 2N \), for any \( f \in \mathcal{R}_h \)

\[ R(f) \leq \inf_{\lambda \in [1,2N]} B(\lambda, h, f), \]

where

\[ B(\lambda, h, f) = \left( 1 - \frac{\zeta \lambda}{2N} \right)^{-1} \left\{ 1 + \frac{\zeta \lambda}{2N} r(f) \right. \]

\[ + \frac{1}{\lambda} \left[ h \left[ \log \left( \frac{N^2}{\lambda^2} + 1 + \log(|y|) \right] - \log(\alpha) \right] + \log(\epsilon^{-1}) + \log(\frac{\log(2N)}{\log(\zeta)}) + 1 \right] \right\}. \]

An adaptive estimator \( \hat{f}_a \) can then be built by minimizing (3.1). Let \( \tilde{\mathcal{R}}_h \) be the observable part of \( \hat{\mathcal{R}}_h \), more precisely let

\[ \tilde{\mathcal{R}}_h = \{ \hat{f}(x'_i, y'_i)^h_{i=1} : (x'_i)_{i=1}^h \subset \{ X_i : 1 \leq i \leq N \}, (y'_i)_{i=1}^h \}. \]

Let us choose

\[ \hat{h} \in \arg \min_{h=1, \ldots, 2N} \inf_{f \in \mathcal{R}_h} \{ B(\lambda, h, f), \lambda \in [1, 2N], f \in \mathcal{R}_h \} \]

\[ \hat{f}_a \in \arg \min_{f \in \tilde{\mathcal{R}}_h} \inf_{\lambda \in [1,2N]} B(\lambda, \hat{h}, f). \]

**Proposition 3.2.** With the previous notations

\[ R(\hat{f}_a) \leq \inf \{ B(\lambda, h, f) : \lambda \in [1, 2N], h \in [1, N], f \in \tilde{\mathcal{R}}_h \}. \]

If we are in the so called “transductive learning” situation, where \( (X_i)_{i=1}^N \) and \( (Y_i)_{i=1}^N \) are observed (the estimator has to be applied to a known batch of test examples at least of the same size as the training set), we can use corollary 2.4 instead of theorem 2.1. Indeed, in this situation, for any \( h \), the whole model \( \mathcal{R}_h \) is observable, and therefore the Gibbs posterior distributions \( \hat{\rho}_\beta \) can be computed.

These Gibbs distributions can in particular be approximated using some Metropolis algorithm (see [16] for more details) where the coordinates of \( (x'_i)^h_{i=1} \) and \( (y'_i)^h_{i=1} \) are moved one at a time, and where additions and deletion of coordinates are also allowed to move from one \( \mathcal{R}_h \) to \( \mathcal{R}_{h-1} \) and \( \mathcal{R}_{h+1} \).

Note that this learning scheme is different from cross validation, since, although we should restrict ourselves to choosing \( \hat{f} \) as a function of \( (x'_i, y'_i)^h_{i=1} \) only, we are allowed to choose \( (x'_i, y'_i)^h_{i=1} \) as a function of the observed sample \( (X_1, \ldots, X_N), (Y_1, \ldots, Y_N) \), and also if we wish of \( (X_{N+1}, \ldots, X_{2N}) \) in any suitable way.

### 3.2. Pruning decision trees.

One possible use of compression schemes is to choose adaptively a (pruned) decision tree: given a set of questions \( (q_1, q_2, \ldots, q_n) \) and a small set of (hopefully “typical”) examples \( (x'_1, \ldots, x'_h) \) drawn from \( (X_1, \ldots, X_N) \), we may build a pruned decision tree by stopping to ask questions as soon as only one
example in \((x_i^h)_{i=1}^h\) matches the query. Using proposition 3.1 in this context leads to penalize the risk with a penalty proportional to the number of nodes, something we could have achieved through a different approach (like considering a Galton Watson process as a deterministic prior on trees). But we can do better: we can also prune inner nodes by deciding to remove questions which do not split \((x_i^h)_{i=1}^h\), and we can think about more clever strategies to choose the questions to be asked and the order in which they should be asked as a function of our “compression” set \((x_i^h)_{i=1}^h\) (for instance we can choose a set of questions leading to a balanced tree). We can also use the labels \((y_i^h)_{i=1}^h\) to prune the tree and select questions: indeed we can choose the decision tree in any way we like, as long as we build it in a unique way as a function of \((x_i^h, y_i^h)_{i=1}^h\) only. Then we can compare the performance of the obtained classifiers on the whole training sample \((X_1, Y_1, \ldots, X_N, Y_N)\) and retain the best typical compression set \((x_i^h, y_i^h)\) (using proposition 3.1). This gives a theoretical framework to guide the implementation of many algorithmic ideas in a data driven way.

4. Localization

We can localize our results for exchangeable priors as we had done in previously encountered situations. To achieve this, let us apply lemma 2.2 with

\[
\eta(\theta) = \left(\frac{\lambda^2}{2N} + \beta\right) [r_1(\theta) + r_2(\theta)] + \log \left\{ \pi \left[ \exp \left[ -\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} + \log(\epsilon^{-1}).
\]

We get

**Lemma 4.1.** With \(P_{2N}\) probability at least \(1 - \epsilon\), for any \(\rho \in \mathcal{M}_+^1(\Theta)\),

\[
\left(\lambda - \beta - \frac{\lambda^2}{2N}\right) \rho[r_2(\theta)] \leq \left(\lambda + \beta + \frac{\lambda^2}{2N}\right) \rho[r_1(\theta)]
\]

\[
+ \log \left\{ \pi \left[ \exp \left[ -\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}).
\]

Moreover we get with \(P_{2N}\) probability at least \(1 - \epsilon\),

\[
\log \left\{ \pi \left[ \exp \left[ -\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} = \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ -\beta \rho[r_1(\theta)] - \beta \rho[r_2(\theta)] - \mathcal{K}(\rho, \pi) \right\}
\]

\[
\leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ -\beta \rho[r_1(\theta)] - \mathcal{K}(\rho, \pi) - \beta \left(\lambda + \beta + \frac{\lambda^2}{2N}\right)^{-1} \right\}
\]

\[
- \log \left\{ \pi \left[ \exp \left[ -\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\} - \mathcal{K}(\rho, \pi) - \log(\epsilon^{-1}).
\]

Putting \(\xi = \frac{\beta}{\lambda + \frac{\lambda^2}{2N}}\), this can also be written as

\[
(4.1) \quad \log \left\{ \pi \left[ \exp \left[ -\beta [r_1(\theta) + r_2(\theta)] \right] \right] \right\}
\]

\[
\leq \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ -2\xi \lambda \rho[r_1(\theta)] - \mathcal{K}(\rho, \pi) + \xi \log(\epsilon^{-1}) \right\}
\]

\[
= \log \left\{ \pi \left[ \exp \left[ -2\xi \lambda r_1(\theta) \right] \right] \right\} + \xi \log(\epsilon^{-1}).
\]
This leads to

**Lemma 4.2.** With $P_{2N}$ probability at least $1 - \epsilon$, for any $\rho \in M_+^1$,

$$
\rho[r_2(\theta)] \leq \left[ (1 - \xi) \lambda - (1 + \xi) \frac{\lambda^2}{2N} \right]^{-1} \left\{ (1 + \xi) \lambda \left( 1 + \frac{\lambda}{2N} \right) \rho[r_1(\theta)] + \mathcal{K}(\rho, \pi) + \log \left\{ \pi \left[ \exp \left[ -2\xi \lambda r_1(\theta) \right] \right] \right\} + (1 + \xi) \log \left( \frac{2}{\epsilon} \right) \right\}
$$

$$
= \left[ (1 - \xi) \lambda - (1 + \xi) \frac{\lambda^2}{2N} \right]^{-1} \left\{ (1 - \xi) \lambda + (1 + \xi) \frac{\lambda^2}{2N} \right\} \rho[r_1(\theta)]
$$

$$
+ \mathcal{K}(\rho, \hat{\rho}_{2\xi \lambda}) + (1 + \xi) \log \left( \frac{2}{\epsilon} \right).
$$

Here again the right-hand side is minimized by a Gibbs distribution, leading to

**Corollary 4.1.** With $P_{2N}$ probability at least $1 - \epsilon$,

$$
\hat{\rho}_{(1+\xi)\lambda(1+\frac{2\xi}{N})}[r_2(\theta)] \leq \left[ (1 - \xi) \lambda - (1 + \xi) \frac{\lambda^2}{2N} \right]^{-1} \left\{ \int_{2\xi \lambda}^{(1+\xi)\lambda(1+\frac{2\xi}{N})} \hat{\rho}_\beta [r_1(\theta)] d\beta \right\}
$$

$$
+ (1 + \xi) \log \left( \frac{2}{\epsilon} \right)
$$

$$
\leq \left[ (1 - \xi) \lambda - (1 + \xi) \frac{\lambda^2}{2N} \right]^{-1} \left\{ (1 - \xi) \lambda + (1 + \xi) \frac{\lambda^2}{2N} \right\} \hat{\rho}_{2\xi \lambda}[r_1(\theta)]
$$

$$
+ (1 + \xi) \log \left( \frac{2}{\epsilon} \right)
$$

\}.
CHAPTER 3

Noisy classification with an exchangeable prior

1. Non localized bound

As in the case of deterministic priors treated before, we can derive bounds relative to a given reference classification rule which are sharper in the presence of noise. We will assume here that the distribution of patterns and labels is i.i.d. and therefore consider a product distribution \( P^\otimes 2N \) on \((\mathcal{X} \times \mathcal{Y})^2N, (\mathcal{B} \otimes \mathcal{B}')^\otimes 2N \). Similarly to what has been done in section 5 we consider some fixed (and unknown) parameter \( \tilde{\theta} \in \Theta \) and define

\[
\sigma_k(\theta) = 1[Y_k \neq f_\theta(X_k)]
\]

\[
\overline{r}_1(\theta) = \frac{1}{N} \sum_{k=1}^{N} \sigma_k(\theta) - \sigma_k(\tilde{\theta})
\]

\[
\overline{r}_2(\theta) = \frac{1}{N} \sum_{k=N+1}^{2N} \sigma_k(\theta) - \sigma_k(\tilde{\theta})
\]

\[
\overline{R}(\theta | X_k) = P[\sigma_k(\theta) - \sigma_k(\tilde{\theta}) | X_k]
\]

\[
r'_1(\theta) = \frac{1}{N} \sum_{k=1}^{N} \overline{R}(\theta | X_k),
\]

\[
r'_2(\theta) = \frac{1}{N} \sum_{k=N+1}^{2N} \overline{R}(\theta | X_k).
\]

For the sake of simplicity, we will assume that the rule \( f_\beta \) clearly outperforms the other rules for any pattern, in the sense that for some constant \( \alpha > 0 \) which will stay fixed in the remaining of this discussion, for any \( x \in \mathcal{X} \),

\[
\alpha(x) = \min \{ \overline{R}(\theta | x), \theta \in \Theta, f_\theta(x) \neq f_\beta(x) \} \geq \alpha.
\]

Let us consider two real numbers \( \beta > \lambda > 0 \) and put for short \( \kappa = \frac{1}{\alpha} \theta(\frac{2\beta}{N}) \). The following exponential inequalities will be helpful:

\[
P^\otimes 2N \left\{ \exp[\lambda \overline{r}_2(\theta) - \beta \overline{r}_1(\theta)] | X_1^{2N} \right\}
\]

\[
\leq P^\otimes 2N \left\{ \exp \left[ (\lambda + \kappa \frac{\lambda^2}{N}) r'_2(\theta) - (\beta - \kappa \frac{\beta^2}{N}) r'_1(\theta) \right] | X_1^{2N} \right\}.
\]

Moreover, putting

\[
\lambda' = \lambda + \kappa \frac{\lambda^2}{N}
\]

\[
\beta' = \beta - \kappa \frac{\beta^2}{N}.
\]
\[ P^{\otimes 2N} \left\{ \exp \left[ \lambda' r_2' - \beta' r_1' \right] \left| \sum_{k=1}^{2N} \delta_{X_k} \right\} \right. \]

\[ \leq P^{\otimes 2N} \left\{ \exp \left[ \sum_{k=1}^{2N} \left( \frac{1}{2N} \left( \lambda' + \beta' \right) - \frac{\lambda' - \lambda}{2} \right) [r_1' + r_2'] \right] \left| \sum_{k=1}^{2N} \delta_{X_k} \right\} \right. \]

Integrating these inequalities with respect to a random exchangeable prior distribution \( \pi : (\mathcal{X}^{2N}, \mathcal{B}^{\otimes 2N}) \rightarrow \mathcal{M}_+^1 (\Theta) \) we get

**Lemma 1.1.** With \( P^{\otimes 2N} \) probability at least \( 1 - \epsilon \), for any posterior \( \rho \in \mathcal{M}_+^1 (\Theta) \),

\[ \lambda \rho [\mathcal{F}_2 (\theta)] \leq \beta \rho [\mathcal{F}_1 (\theta)] + K (\rho, \pi) \]

\[ + \log \left\{ \pi \left[ \exp \left[ - \left( \frac{\beta' - \lambda'}{2} \right) [r_1' + r_2'] \right] \right) \right\} + \log (\epsilon^{-1}). \]

Moreover, with \( P^{\otimes 2N} \) probability at least \( 1 - \epsilon \), for any posterior \( \rho \in \mathcal{M}_+^1 (\Theta) \),

\[ \lambda' \rho [r_2' (\theta)] \leq \beta' \rho [r_1' (\theta)] + K (\rho, \pi) \]

\[ + \log \left\{ \pi \left[ \exp \left[ - \beta'' [r_1' (\theta) + r_2' (\theta)] \right] \right) \right\} + \log (\epsilon^{-1}), \]

(\text{where} \( \beta'' \) is defined in the previous equation).

To get a non localized learning theorem, we can choose for some parameter \( \mu \)

\[ \lambda' = \mu - \frac{1}{2N} \mu^2 = \lambda + \kappa \frac{\lambda^2}{N}, \]

\[ \beta' = \mu + \frac{1}{2N} \mu^2 = \beta - \kappa \frac{\beta^2}{N}, \]

and take advantage of the fact that \( r_1' (\theta) \) and \( r_2' (\theta) \) are all positive random variables (since we assumed that \( \tilde{\theta} \) was everywhere optimal).

**Theorem 1.1.** With \( P^{\otimes 2N} \) probability at least \( 1 - \epsilon \), for any posterior \( \rho \in \mathcal{M}_+^1 (\Theta) \),

\[ \rho [r_2 (\theta)] \leq r_2 (\tilde{\theta}) + \frac{\mu + \frac{\mu^2}{2N} + \kappa \frac{\beta^2}{N}}{\mu - \frac{\mu^2}{2N} - \kappa \frac{\beta^2}{N}} \left[ \rho [r_1 (\theta)] - r_1 (\tilde{\theta}) \right] \]

\[ + \frac{1}{\mu - \frac{\mu^2}{2N} - \kappa \frac{\beta^2}{N}} \left\{ K (\rho, \pi) + \log (\epsilon^{-1}) \right\}. \]

**2. Localized bound**

To get a localized learning theorem, we need an upper bound for

\[ \log \left\{ \pi \left[ \exp \left[ - \beta'' [r_1' (\theta) + r_2' (\theta)] \right] \right) \right\}. \]

We will achieve this in two steps. The first one is similar to the low noise case with an exchangeable prior, and compares the above quantity with \( \log \{\pi [\exp (\gamma r_1' (\theta))]\} \).
for a suitable choice of $\gamma$. Let us put
\[
\gamma' = \beta' \frac{\beta' + \lambda'}{\beta' - \beta''} = (\beta' - \lambda') \frac{1 - \frac{1}{N} (\lambda' + \beta')^2}{1 + \frac{1}{2N} \frac{\lambda' + \beta'}{2}}.
\]
\[
\xi = \frac{\beta''}{\beta' - \beta''} = \frac{\beta' - \lambda'}{\beta' + \lambda'} \leq \frac{\beta' - \lambda'}{\beta' + \lambda'}.
\]

The same computation that led to (4.1) shows that

**Lemma 2.1.** With $P_{\Theta}^{\otimes N}$ probability at least $1 - \epsilon$,
\[
\log\left\{ \pi\left[ \exp\left\{ -\beta'' [r_1(\theta) + r_2(\theta)] \right\} \right] \right\} \leq \log\left\{ \pi\left[ \exp\left\{ -\gamma' [r_1(\theta)] \right\} \right] \right\} + \xi \log(\epsilon^{-1}).
\]

Now we need to compare $\log\{\pi[\exp[-\gamma' r_1(\theta)]\}$ with $\log\{\pi[\exp[-\gamma r_1(\theta)]\}$ for some suitable value of $\gamma$. To achieve this, we use another learning lemma, derived from the inequality
\[
P_{\Theta}^{\otimes N}\left\{ \exp\left[ \lambda [\tau_1(\theta) - r_1(\theta)] - \mu r_1(\theta) \right] \right\} \leq \exp\left[ (\kappa \frac{\lambda^2}{N} - \mu) r_1(\theta) \right].
\]

**Lemma 2.2.** With $P_{\Theta}^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior probability distribution $\rho \in \mathcal{M}_1(\Theta)$,
\[
\lambda \rho[\tau_1(\theta)] \leq \left( \lambda + \gamma' + \kappa \frac{\lambda^2}{N} \right) \rho[r_1(\theta)] + \log\left\{ \pi\left[ \exp\left[ -\gamma' r_1(\theta) \right] \right] \right\} + \mathcal{K}(\rho, \pi) + \log(\epsilon^{-1}).
\]

Exactly as we derived (4.3), we can establish that with $P_{\Theta}^{\otimes N}$ probability at least $1 - \epsilon$,
\[
\log\left\{ \pi\left[ \exp\left[ -\gamma' r_1(\theta) \right] \right] \right\} \leq \log\left\{ \pi\left[ \exp\left[ -\gamma' \tau_1(\theta) \right] \right] \right\} + \frac{\gamma'}{\lambda + \kappa \frac{\lambda^2}{N}} \log(\epsilon^{-1}).
\]

Putting all these things together leads to a localized learning theorem for noisy classification using an exchangeable prior. Let us put
\[
\zeta = \left( 1 - \kappa \frac{\lambda^2 + \beta^2}{N(\beta - \lambda)} \right) \left( 1 - \frac{(\lambda' + \beta')^2}{4N(\beta' - \lambda')} \right) \left( 1 + \frac{\lambda' + \beta'}{4N} \right) \left( 1 - \kappa \frac{\lambda^2}{N} \right).
\]

**Theorem 2.1.** With the notations introduced in this section, with $P_{\Theta}^{\otimes N}$ probability at least $1 - \epsilon$, for any posterior distribution $\rho \in \mathcal{M}_1(\Theta)$,
\[
\rho[r_2(\theta)] \leq r_2(\tilde{\theta}) + \frac{\beta}{\lambda} \rho[r_1(\theta)] - r_1(\tilde{\theta})
\]
\[
+ \frac{1}{\lambda} \left\{ \mathcal{K}(\rho, \pi) + \log\left\{ \pi\left[ \exp\left[ -(\beta - \lambda)\zeta \tau_1(\theta) \right] \right] \right\} + \left( 1 + \frac{\beta', \lambda'}{\beta', \lambda'} \right) \log(\frac{\zeta}{\lambda}) \right\}
\]
\[
= r_2(\tilde{\theta}) + \left( \zeta + (1 - \zeta) \frac{\beta}{\lambda} \right) \left\{ \rho[r_1(\theta)] - r_1(\tilde{\theta}) \right\}
\]
\[
+ \frac{1}{\lambda} \left\{ \mathcal{K}(\rho, \rho_{(\beta - \lambda)c}) + \left( 1 + \frac{\beta', \lambda'}{\beta', \lambda'} \right) \log(\frac{\zeta}{\lambda}) \right\}.
\]
As a special case,

$$\hat{\rho}_\beta \left[ r_2(\theta) \right] \leq r_2(\bar{\theta}) + \frac{1}{\lambda} \int_{(\beta-\lambda)\zeta}^{\beta} \hat{\rho}_\gamma \left[ \gamma_1(\theta) \right] d\gamma + \frac{1}{\lambda} \left( 1 + \frac{\beta' - \lambda'}{\beta' + \lambda'} + \frac{\beta' - \lambda'}{\lambda'} \right) \log \left( \frac{3}{\epsilon} \right).$$
CHAPTER 4

Support Vector Machines

1. The canonical hyperplane

Let \( Z = (x_i, y_i)_{i=1}^N \in (\mathbb{R}^d \times \{-1, +1\})^N \) be some set of labelled examples (called the training set hereafter). Let

\[
I = \{1, \ldots, N\},
I_+ = \{i \in I : y_i = +1\},
I_- = \{i \in I : y_i = -1\},
\]

and consider

\[
A_Z = \{w \in \mathbb{R}^d : \sup_{b \in \mathbb{R}} \inf_{i \in I} (\langle w, x_i \rangle - b)y_i \geq 1 \}.
\]

Let us remark that this set of admissible separating directions can also be written as

\[
A_Z = \{w \in \mathbb{R}^d : \max_{i \in I_-} \langle w, x_i \rangle + 2 \leq \min_{i \in I_+} \langle w, x_i \rangle \}.
\]

As it is easily seen, the optimal value of \( b \) for a fixed value of \( w \), in other words the value of \( b \) which maximizes \( \inf_{i \in I} (\langle w, x_i \rangle - b)y_i \), is equal to

\[
b_w = \frac{1}{2} \left[ \max_{i \in I_-} \langle w, x_i \rangle + \min_{i \in I_+} \langle w, x_i \rangle \right].
\]

**Lemma 1.1.** When \( A_Z \neq \emptyset \), \( \inf_{i \in I} (\langle w, x_i \rangle - b)y_i \geq 1 \) is reached for only one value \( w_Z \) of \( w \).

**Proof.** The set \( A_Z \) is convex and \( w \rightarrow \|w\|^2 \) is strictly convex. \( \square \)

**Definition 1.1.** When \( A_Z \neq \emptyset \), the training set \( Z \) is said to be linearly separable. The hyperplane

\[
H = \{x \in \mathbb{R}^d : \langle w_Z, x \rangle - b_Z = 0\},
\]

where

\[
w_Z = \arg \min \{\|w\| : w \in A_Z\},
\]

\[
b_Z = b_{w_Z},
\]

is called the canonical separating hyperplane of the training set \( Z \). The quantity \( \|w_Z\|^{-1} \) is called the margin of the canonical hyperplane.

Note that as \( \min_{i \in I_+} \langle w_Z, x_i \rangle - \max_{i \in I_-} \langle w_Z, x_i \rangle = 2 \), the margin is also equal to half the distance between the projections on the direction \( w_Z \) of the positive and negative patterns.
2. Computation of the canonical hyperplane

Let us consider the convex hulls $X_+$ and $X_-$ of the positive and negative patterns:

$$X_+ = \left\{ \sum_{i \in I_+} \lambda_i x_i : (\lambda_i)_{i \in I_+} \in \mathbb{R}^I_+, \sum_{i \in I_+} \lambda_i = 1 \right\},$$

$$X_- = \left\{ \sum_{i \in I_-} \lambda_i x_i : (\lambda_i)_{i \in I_-} \in \mathbb{R}^I_-, \sum_{i \in I_-} \lambda_i = 1 \right\}.$$

Let us introduce the convex set

$$V = X_+ - X_- = \{ x_+ - x_- : x_+ \in X_+, x_- \in X_- \}.$$

As for any vector $w \in \mathbb{R}^d$,

$$\min_{i \in I_+} \langle w, x_i \rangle = \min_{x \in X_+} \langle w, x \rangle,$$

$$\max_{i \in I_-} \langle w, x_i \rangle = \max_{x \in X_-} \langle w, x \rangle,$$

we see that necessarily $\min_{i \in I_+} \langle w, x_i \rangle - \max_{i \in I_-} \langle w, x_i \rangle \leq 0$, which shows that $w$ cannot be in $A_Z$ and therefore that $A_Z$ is empty.

Let us assume now that $v^* \neq 0$, or equivalently that $X_+ \cap X_- = \emptyset$. Let us put

$$w^* = 2 \frac{v^*}{\|v^*\|^2} v^*.$$

As

$$\min_{i \in I_+} \langle w^*, x_i \rangle - \max_{i \in I_-} \langle w^*, x_i \rangle = \inf_{x \in X_+} \langle w^*, x \rangle - \sup_{x \in X_-} \langle w^*, x \rangle = \inf_{x \in X_+, x_+ \in X_-} \langle w^*, x_+ - x_- \rangle = \frac{2}{\|v^*\|^2} \inf_{v \in V} \langle v, v^* \rangle.$$

Let us now prove that $\inf_{v \in V} \langle v, v^* \rangle = \|v^*\|^2$. Some arbitrary $v \in V$ being fixed, consider the function

$$\beta \mapsto \|\beta v + (1 - \beta)v^*\|^2 : [0, 1] \to \mathbb{R}.$$

By definition of $v^*$, it reaches its minimum value for $\beta = 0$, and therefore has a non negative derivative at this point. Computing this derivative, we find that $\langle v - v^*, v^* \rangle \geq 0$, as claimed. We have proved that

$$\min_{i \in I_+} \langle w^*, x_i \rangle - \max_{i \in I_-} \langle w^*, x_i \rangle = 2,$$

and therefore that $w^* \in A_Z$. On the other hand, any $w \in A_Z$ is such that

$$2 \leq \min_{i \in I_+} \langle w, x_i \rangle - \max_{i \in I_-} \langle w, x_i \rangle = \inf_{v \in V} \langle w, v \rangle \leq \|w\| \inf_{v \in V} \|v\| = \|w\| \|v^*\|.$$

CHAPTER 4. SUPPORT VECTOR MACHINES
This proves that \( \|w^*\| = \inf \{ \|w\| : w \in A_Z \} \), and therefore that \( w^* = w_Z \) as claimed. \( \square \)

One way to compute \( w_Z \) would be therefore to compute \( v^* \) by minimizing

\[
\{ \| \sum_{i \in I} \lambda_i y_i x_i \|^2 : (\lambda_i)_{i \in I} \in \mathbb{R}_+^I, \sum_{i \in I} \lambda_i = 2 \sum_{i \in I} y_i \lambda_i = 0 \}.
\]

Although this is a tractable quadratic programming problem, a direct computation of \( w_Z \) through the following proposition is usually preferred.

**Proposition 2.1.** The canonical direction \( w_Z \) can be expressed as

\[
w_Z = \sum_{i=1}^{N} \alpha_i^* y_i x_i,
\]

where \( (\alpha_i^*)_{i=1}^{N} \) is obtained by minimizing

\[
\inf \{ F(\alpha) : \alpha \in A \},
\]

where

\[
A = \{ (\alpha_i)_{i \in I} \in \mathbb{R}_+^I, \sum_{i \in I} \alpha_i y_i = 0 \},
\]

and

\[
F(\alpha) = \left\| \sum_{i \in I} \alpha_i y_i x_i \right\|^2 - 2 \sum_{i \in I} \alpha_i.
\]

**Proof.** Let \( w(\alpha) = \sum_{i \in I} \alpha_i y_i x_i \) and let \( S(\alpha) = \frac{1}{2} \sum_{i \in I} \alpha_i \). We can express the function \( F(\alpha) \) as \( F(\alpha) = \|w(\alpha)\|^2 - 4S(\alpha) \). Moreover it is important to notice that for any \( s \in \mathbb{R}_+ \), \( \{ w(\alpha) : \alpha \in A, S(\alpha) = s \} = sV \). This shows that for any \( s \in \mathbb{R}_+ \), \( \inf \{ F(\alpha) : \alpha \in A, S(\alpha) = s \} \) is reached and that for any \( \alpha_s \in \{ \alpha \in A : S(\alpha) = s \} \) reaching this infimum, \( w(\alpha_s) = sv^* \). As \( s \mapsto s^2 \|v^*\|^2 - 4s : \mathbb{R}_+ \to \mathbb{R} \) reaches its infimum for only one value \( s^* \) of \( s \), namely at \( s^* = \frac{2}{\|v^*\|^2} \), this shows that \( F(\alpha) \) reaches its infimum on \( A \), and that for any \( \alpha^* \in A \) such that \( F(\alpha^*) = \inf \{ F(\alpha) : \alpha \in A \} \), \( w(\alpha^*) = \frac{2}{\|v^*\|^2} v^* = w_Z \). \( \square \)

### 3. Support vectors

**Definition 3.1.** The set of support vectors \( S \) is defined by

\[
S = \{ x_i : \langle w_Z, x_i \rangle - b_Z = y_i \}.
\]

**Proposition 3.1.** Any \( \alpha^* \) minimizing \( f(\alpha) \) on \( A \) is such that

\[
\{ x_i : \alpha_i^* > 0 \} \subset S.
\]

This implies that the representation \( w_Z = w(\alpha^*) \) involves in general only a limited number of non zero coefficients and that \( w_Z = w_{Z'} \), where \( Z' = \{ (x_i, y_i) : x_i \in S \} \).

**Proof.** Let us consider any given \( i \in I_+ \) and \( j \in I_- \), such that \( \alpha_i^* > 0 \) and \( \alpha_j^* > 0 \) (there exists at least one such index in each set \( I_- \) and \( I_+ \), since the sum of the components of \( \alpha^* \) on each of these sets are equal and since \( \sum_{k \in I} \alpha_k^* > 0 \)). For any \( t \in \mathbb{R} \), consider

\[
\alpha_k(t) = \alpha_k + t \mathbb{1}(k \in \{ i, j \}), \quad k \in I,
\]

where \( \mathbb{1} \) is the indicator function.
The vector \( \alpha(t) \) is in \( A \) for any value of \( t \) in some neighborhood of 0, therefore
\[
\frac{\partial}{\partial t}
|_{t=0} F[\alpha(t)] = 0.
\]
Computing this derivative, we find that
\[
y_{i}(w(\alpha^{*}), x_{i}) + y_{j}(w(\alpha^{*}), x_{j}) = 2.
\]
As \( y_{i} = -y_{j} \), this can also be written as
\[
y_{i}[(w(\alpha^{*}), x_{i}) - b_{Z}] + y_{j}[(w(\alpha^{*}), x_{j}) - b_{Z}] = 2.
\]
As \( w(\alpha^{*}) \in A_{Z} \)
\[
y_{k}[(w(\alpha^{*}), x_{k}) - b_{Z}] \geq 1, \quad k \in I,
\]
which implies necessarily as claimed that
\[
y_{i}[(w(\alpha^{*}), x_{i}) - b_{Z}] = y_{j}[(w(\alpha^{*}), x_{j}) - b_{Z}] = 1.
\]
\( \square \)

4. Support Vector Machines

**Definition 4.1.** The symmetric measurable kernel \( K \rightarrow \mathbb{R}_{+} \) is said to be positive (or more precisely positive semi-definite) if for any \( n \in \mathbb{N} \), any \( (x_{i})_{i=1}^{n} \in \mathcal{X}^{n} \),
\[
\inf_{\alpha \in \mathbb{R}_{+}^{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{i}K(x_{i}, x_{j})\alpha_{j} \geq 0.
\]

Let \( Z = (x_{i}, y_{i})_{i=1}^{N} \) be some training set. Let us consider as in the previous sections of this chapter
\[
\mathcal{A} = \{ \alpha \in \mathbb{R}_{+}^{N} : \sum_{i=1}^{N} \alpha_{i}y_{i} = 0 \}.
\]
Let
\[
F(\alpha) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{i}y_{i}K(x_{i}, x_{j})\alpha_{j} - 2 \sum_{i=1}^{N} \alpha_{i}.
\]

**Definition 4.2.** Let \( K \) be a positive symmetric kernel. The training set \( Z \) is said to be \( K \)-separable if
\[
\inf\{ F(\alpha) : \alpha \in \mathcal{A} \} > -\infty.
\]

**Lemma 4.1.** When \( Z \) is \( K \)-separable, \( \inf\{ F(\alpha) : \alpha \in \mathcal{A} \} \) is reached.

**Proof.** Consider the training set \( Z' = (x'_{i}, y_{i})_{i=1}^{N} \), where
\[
x'_{i} = \left\{ \left( K(x_{k}, x_{\ell}) \right)_{k=1, \ell=1}^{N} \right\}_{i=1}^{N}^{1/2} \in \mathbb{R}^{N}.
\]
We see that
\[
F(\alpha) = \left\| \sum_{i=1}^{N} \alpha_{i}y_{i}x'_{i} \right\|^{2} - 2 \sum_{i=1}^{N} \alpha_{i}. \]
We have proved in the previous section that \( Z' \) is linearly separable if and only if \( \inf\{ F(\alpha) : \alpha \in \mathcal{A} \} > -\infty \), and that the infimum is reached in this case. \( \square \)

**Proposition 4.1.** Let \( K \) be a symmetric positive kernel and let \( (Z_{i})_{i=1}^{N} \) be some \( K \)-separable training set. Let \( \alpha^{*} \in \mathcal{A} \) be such that \( F(\alpha^{*}) = \inf\{ F(\alpha) : \alpha \in \mathcal{A} \} \). Let
\[
I_{n}^{*} = \{ i \in \mathbb{N} : 1 \leq i \leq N, y_{i} = -1, \alpha_{i}^{*} > 0 \}
\]
\[
I_{n}^{*} = \{ i \in \mathbb{N} : 1 \leq i \leq N, y_{i} = +1, \alpha_{i}^{*} > 0 \}
\]
The classification rule \( f : \mathcal{X} \to \mathcal{Y} \) defined by the formula

\[
f(x) = \text{sign} \left( \sum_{i=1}^{N} \alpha^*_i y_i K(x_i, x) - b^* \right)
\]

is independent of the choice of \( \alpha^* \) and is called the support vector machine defined by \( K \) and \( Z \). The set \( S = \{ x_j : \sum_{i=1}^{N} \alpha^*_i y_i K(x_i, x_j) - b^* = y_j \} \) is called the set of support vectors. For any choice of \( \alpha^* \), \( \{ x_i : \alpha^*_i > 0 \} \subset S \).

**Proof.** The independence from the choice of \( \alpha^* \), which is not necessarily unique, is seen as follows. Let \( (x_i)_{i=1}^{N+1} \) and \( x \in \mathcal{X} \) be fixed. Let us put for ease of notations \( x_{N+1} = x \). Let \( M \) be the \((N + 1) \times (N + 1)\) symmetric semi-definite matrix defined by \( M(i, j) = K(x_i, x_j), i = 1, \ldots, N + 1, j = 1, \ldots, N + 1 \). Let us consider the mapping \( \Psi : \{ x_i : i = 1, \ldots, N + 1 \} \to \mathbb{R}^{N+1} \) defined by \( \Psi(x_i) = [M^1/2(i, j)]_{j=1}^{N+1} \in \mathbb{R}^{N+1} \). Let us consider the training set \( Z' = [\Psi(x_i), y_i]_{i=1}^{N} \). Then \( Z' \) is linearly separable,

\[
F(\alpha) = \left\| \sum_{i=1}^{N} \alpha_i y_i \Psi(x_i) \right\|^2 - 2 \sum_{i=1}^{N} \alpha_i,
\]

and we have proved that for any choice of \( \alpha^* \in A \) minimizing \( F(\alpha) \), \( w_{Z'} = \sum_{i=1}^{N} \alpha^*_i y_i \Psi(x_i) \). Thus the support vector machine defined by \( K \) and \( Z \) can also be expressed by the formula

\[
f(x) = \text{sign} \left[ \langle w_{Z'}, \Psi(x) \rangle - b_{Z'} \right]
\]

which does not depend on \( \alpha^* \). The definition of \( S \) is such that \( \Psi(S) \) is the set of support vectors defined in the linear case, where its stated property has already been proved. \( \square \)

### 5. Support vector machines seen as compression schemes

We can use support vector machines in the framework of compression schemes and apply proposition 3.1 of chapter 2. More precisely, given some positive symmetric kernel \( K \) on \( \mathcal{X} \), we may consider for any training set \( Z' = (x'_i, y'_i)_{i=1}^{N} \) the classifier \( \hat{f}_{Z'} : \mathcal{X} \to \mathcal{Y} \) which is equal to the support vector machine defined by \( K \) and \( Z' \) whenever \( Z' \) is \( K \)-separable, and which is equal to some constant classification rule otherwise (we take this convention to stick to the framework of section 3.1, we will only use \( \hat{f}_{Z'} \) in the \( K \)-separable case, so this extension of the definition is just a matter of presentation). In the application of proposition 3.1, in the case when the observed sample \( (X_i, Y_i)_{i=1}^{N} \) is \( K \)-separable, a natural (if not always optimal) choice of \( Z' \) is to choose for \( (x'_i) \) the set of support vectors defined by \( Z = (X_i, Y_i)_{i=1}^{N} \) and to choose for \( (y'_i) \) the corresponding values of \( Y \). This is justified by the fact that \( \hat{f}_Z = \hat{f}_{Z'} \), as shown in proposition 4.1. In the case when \( Z \) is not \( K \)-separable, in theory, we can flip the smallest number of values of \( Y_i \) to define a \( K \)-separable training set \( Z' = (X_i, y'_i)_{i=1}^{N} \), and then restrict this training set to its support vectors to see in which submodel this classification rule really lives. In practice, this may be time consuming, and various minimization criterion directly adapted to the
non $K$-separable case are of common use. We suggest [17] as a further reading on this topic.

6. Building kernels

The results of this section are drawn from [17].

**Proposition 6.1.** Let $K_1$ and $K_2$ be positive symmetric kernels on $X$. Then for any $a \in \mathbb{R}_+$

\[ (aK_1 + K_2)(x, x') \overset{\text{def}}{=} aK_1(x, x') + K_2(x, x') \]

and $(K_1 \cdot K_2)(x, x') \overset{\text{def}}{=} K_1(x, x')K_2(x, x')$

are also positive symmetric kernels. Moreover, for any measurable function $g : X \to \mathbb{R}$, $K_g(x, x') \overset{\text{def}}{=} g(x)g(x')$ is also a positive symmetric kernel.

**Proof.** It is enough to prove the proposition in the case when $X$ is finite and kernels are just ordinary symmetric matrices. Thus we can assume without loss of generality that $X = \{1, \ldots, n\}$. Then for any $\alpha \in \mathbb{R}^N$, using usual matrix notations,

\[ \langle \alpha, (aK_1 + K_2)\alpha \rangle = a\langle \alpha, K_1\alpha \rangle + \langle \alpha, K_2\alpha \rangle \geq 0, \]

\[ \langle \alpha, (K_1 \cdot K_2)\alpha \rangle = \sum_{i,j} \alpha_iK_1(i, j)K_2(i, j)\alpha_j \]

\[ = \sum_{i,j,k} \alpha_iK_1^{1/2}(i, k)K_1^{1/2}(k, j)K_2(i, j)\alpha_j \]

\[ = \sum_k \sum_{i,j} \left[ K_1^{1/2}(k, i)\alpha_iK_2(i, j)\left[ K_1^{1/2}(k, j)\alpha_j \right] \right] \geq 0, \]

\[ \langle \alpha, K_g\alpha \rangle = \sum_{i,j} \alpha_i g(i)g(j)\alpha_j = \left( \sum_i \alpha_i g(i) \right)^2 \geq 0. \]

\[ \square \]

**Proposition 6.2.** Let $K$ be some positive symmetric kernel on $X$. Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial with positive coefficients. Let $g : X \to \mathbb{R}^d$ be a measurable function. Then

\[ p(K)(x, x') \overset{\text{def}}{=} p[K(x, x')], \]

\[ \exp(K)(x, x') \overset{\text{def}}{=} \exp[K(x, x')] \]

and $G_g(x, x') \overset{\text{def}}{=} \exp(-\|g(x) - g(x')\|^2)$

are all positive symmetric kernels.

**Proof.** The first assertion is a direct consequence of the previous proposition. The second one comes from the fact that the exponential function is the pointwise limit of a sequence of polynomial functions with positive coefficients. The third one is seen from the second one and the decomposition

\[ G_g(x, x') = \left[ \exp(-\|g(x)\|^2) \exp(-\|g(x')\|^2) \right] \exp[2(g(x), g(x'))] \]

\[ \square \]
CHAPTER 5

VC dimension of linear rules with margin constraints

1. How far subsets can be separated

**Theorem 1.1.** Consider a family of points \((x_1, \ldots, x_n)\) in some Euclidean vector space \(E\) and a family of hyperplanes 
\[
\mathcal{H} = \{ H_{w,b} : w \in E, \|w\| = 1, b \in \mathbb{R} \},
\]
where 
\[
H_{w,b} = \{ x \in E : \langle x, w \rangle = b \}.
\]
Assume that for any \(H_{w,b} \in \mathcal{H}\),
\[
\inf_{i=1}^{n} |\langle x_i, w \rangle - b| \geq \gamma,
\]
and that for any subset of points \(S \subset \{x_1, \ldots, x_n\}\) there is \(H_{w,b} \in \mathcal{H}\) such that 
\[
S = \{ x_i : \langle w, x_i \rangle > b \}.
\]
Let us also introduce the empirical variance of \((x_i)_{i=1}^{n}\),
\[
\text{Var}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right|^2.
\]
In this case and with these notations,
\[
\frac{\text{Var}(x_1, \ldots, x_n)}{\gamma^2} \geq \begin{cases} 
\frac{n-1}{(n-1)n^2-1} & \text{when } n \text{ is even,} \\
\frac{n^2-1}{n^2} & \text{when } n \text{ is odd.}
\end{cases}
\]
Moreover, equality is reached when \(\gamma\) is optimal and \((x_1, \ldots, x_n)\) is a regular simplex (i.e. when \(2\gamma\) is the minimum distance between the convex hulls of any two subsets of \(\{x_1, \ldots, x_n\}\) and \(\|x_i - x_j\|\) does not depend on \(i \neq j\)).

**Proof.** Let \((s_i)_{i=1}^{n} \in \mathbb{R}^n\) be such that \(\sum_{i=1}^{n} s_i = 0\). Let \(\sigma\) be a uniformly distributed random variable with values in \(\mathfrak{S}_n\), the set of permutations of the first \(n\) integers \(\{1, \ldots, n\}\). For any value of \(\sigma\), there is a hyperplane \(H_{w,b} \in \mathcal{H}\) such that 
\[
\{ x_i : s_{\sigma(i)} > 0 \} = \{ x_i : \langle w, x_i \rangle > b \}.
\]
As a consequence 
\[
\left\langle \sum_{i=1}^{n} s_{\sigma(i)} x_i, w \right\rangle = \sum_{i=1}^{n} s_{\sigma(i)} \left( \langle x_i, w \rangle - b \right)
= \sum_{i=1}^{n} |s_{\sigma(i)}| |\langle x_i, w \rangle - b|
\]
\[ \geq \gamma \sum_{i=1}^{n} |s_{\sigma(i)}| \]
\[ = \gamma \sum_{i=1}^{n} |s_i|. \]

Therefore
\[ E \left( \left\| \sum_{i=1}^{n} s_{\sigma(i)} x_i \right\|^2 \right) \geq \gamma^2 \left( \sum_{i=1}^{n} |s_i| \right)^2. \]

On the other hand
\[ E \left( \left\| \sum_{i=1}^{n} s_{\sigma(i)} x_i \right\|^2 \right) = \sum_{i=1}^{n} E\left( s_{\sigma(i)}^2 \right) \| x_i \|^2 + \sum_{i \neq j} E\left( s_{\sigma(i)} s_{\sigma(j)} \right) \langle x_i, x_j \rangle. \]

Moreover
\[ E\left( s_{\sigma(i)}^2 \right) = \frac{1}{n} \sum_{i=1}^{n} s_{\sigma(i)}^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2. \]

In the same way, for any \( i \neq j, \)
\[ E\left( s_{\sigma(i)} s_{\sigma(j)} \right) = \frac{1}{n(n-1)} E\left( \sum_{i \neq j} s_{\sigma(i)} s_{\sigma(j)} \right) \]
\[ = \frac{1}{n(n-1)} \sum_{i \neq j} s_i s_j \]
\[ = \frac{1}{n(n-1)} \left[ \left( \sum_{i=1}^{n} s_i \right)^2 - \sum_{i=1}^{n} s_i^2 \right]_{=0} \]
\[ = - \frac{1}{n(n-1)} \sum_{i=1}^{n} s_i^2. \]

Thus
\[ E \left( \left\| \sum_{i=1}^{n} s_{\sigma(i)} x_i \right\|^2 \right) = \left( \sum_{i=1}^{n} s_i^2 \right) \left[ \frac{1}{n} \sum_{i=1}^{n} \| x_i \|^2 - \frac{1}{n(n-1)} \sum_{i \neq j} \langle x_i, x_j \rangle \right] \]
\[ = \left( \sum_{i=1}^{n} s_i^2 \right) \left[ \left( \frac{1}{n} + \frac{1}{n(n-1)} \right) \sum_{i=1}^{n} \| x_i \|^2 \right. \]
\[ - \frac{1}{n(n-1)} \left( \sum_{i=1}^{n} x_i \right)^2 \]
We have proved that

\[ \text{Var}(x_1, \ldots, x_n) \geq \frac{(n-1) \left( \sum_{i=1}^{n} |s_i| \right)^2}{n \sum_{i=1}^{n} s_i^2}. \]

This can be used with \( s_i = 1 \) (\( i \leq \frac{n}{2} \)) \(-\frac{2}{n+1} \) (\( i > \frac{n}{2} \)) in the case when \( n \) is even and \( s_i = 2 \) (\( i \leq \frac{n-1}{2} \)) \(-2n+1 \) (\( i > \frac{n-1}{2} \)) in the case when \( n \) is odd to establish the first inequality (1.1) of the theorem.

Checking that equality is reached for the simplex is an easy computation when the simplex \( (x_i)_{i=1}^{n} \in \mathbb{R}^n \) is parametrized in such a way that

\[ x_i(j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise}. \end{cases} \]

Indeed the distance between the convex hulls of any two subsets of the simplex is the distance between their mean values (i.e. centers of mass).

\[ \square \]

2. Application to support vector machines

We are going to apply theorem 1.1 to support vector machines in the transductive case. So let us consider \( (X_i, Y_i)_{i=1}^{2N} \) distributed according to some exchangeable distribution \( P_{2N} \) and assume that \( (X_i)_{i=1}^{2N} \) and \( (Y_i)_{i=1}^{N} \) are observed. Let us consider some positive kernel \( K \) on \( \mathcal{X} \). For any \( K \)-separable training set of the form \( Z' = (X_i, y'_i)_{i=1}^{2N} \), where \( (y'_i)_{i=1}^{2N} \in \mathbb{Y}^{2N} \), let \( f_{Z'} \) be the support vector machine defined by \( K \) and \( Z' \) and let \( \gamma(Z') \) be its margin. Let

\[ R = \max_{i=1, \ldots, 2N} K(x_i, x_i) + \frac{1}{4N^2} \sum_{j=1}^{2N} \sum_{k=1}^{2N} K(x_j, x_k) - \frac{1}{N} \sum_{j=1}^{2N} K(x_i, x_j). \]

Let us define for any integer \( h \) the margins

\[ \gamma_{2h} = \frac{R}{\sqrt{2h-1}} \]

and \( \gamma_{2h+1} = \frac{R}{\sqrt{2h \left( 1 - (2h+1)^2 \right)}} \).

Let us consider for any \( h = 1, \ldots, N \) the exchangeable model

\[ \mathcal{R}_h = \{ f_{Z'} : Z' = (X_i, y'_i)_{i=1}^{2N} \text{ is } K\text{-separable and } \gamma(Z') \geq \gamma_h \}. \]

The family of models \( \mathcal{R}_h, h = 1, \ldots, N \) is nested, and we know from theorem 1.1 that

\[ |\mathcal{R}_h| \leq h \left[ \log \left( \frac{2N}{h} \right) + 1 \right]. \]

We can then consider on the large model \( \mathcal{R} = \bigsqcup_{h=1}^{N} \mathcal{R}_h \) (the disjoint union of the submodels) an exchangeable prior \( \pi \) which is uniform on each \( \mathcal{R}_h \) and is such that \( \pi(\mathcal{R}_h) \leq (1 - \alpha) \alpha^h \) for some parameter \( \alpha \in [0, 1] \). Applying corollary 2.1 of chapter 2, and taking as posterior all the Dirac masses, we get
Proposition 2.1. With \( P_{2N} \) probability at least \( 1 - \epsilon \), for any \( h = 1, \ldots, N \), any support vector machine \( f \in \mathcal{R}_h \),

\[
    r_2(f) \leq \left( 1 - \frac{\lambda}{2N} \right)^{-1} \left\{ \left( 1 + \frac{\lambda}{2N} \right) r_1(f) \right. \\
    \left. + \frac{1}{\lambda} \left[ h \left[ \log \left( \frac{2N}{h} \right) + 1 - \log(\alpha) \right] - \log(1 - \alpha) - \log(\epsilon) \right] \right\}.
\]

(This proposition could of course be made uniform in \( \lambda \) in the standard way many times explained in these lectures.)
Appendix

1. Decomposition of the Kullback divergence function

**Proposition 1.1.** Let \( \rho \) and \( \mu \) be two probability distributions defined on some measurable set \((\Theta, \mathcal{F})\). Let \( \{P_\theta \in \mathcal{M}_1(X, \mathcal{F}) : \theta \in \Theta\} \) and \( \{Q_\theta \in \mathcal{M}_1(X, \mathcal{F}) : \theta \in \Theta\} \) be two families of probability distributions defined on some other measurable set \((X, \mathcal{F})\). Let us assume that for any measurable subset \( B \subset X \) the functions \( \theta \mapsto Q_\theta(B) : \Theta \to \mathbb{R}_+ \) and \( \theta \mapsto P_\theta(B) : \Theta \to \mathbb{R}_+ \) are measurable.

For any subset \( F \subset \Theta \times X \) which is measurable with respect to the product \( \sigma \)-algebra \( \mathcal{F} \otimes \mathcal{F} \), we let \( F_\theta = \{(\theta, x) : (\theta, x) \in F\} \) be the trace of \( F \) with respect to \( \theta \). In this case \( F_\theta \in \mathcal{F} \), and the functions
\[
P : \mathcal{F} \otimes \mathcal{F} \to \mathbb{R}_+ \quad \quad Q : \mathcal{F} \otimes \mathcal{F} \to \mathbb{R}_+
\]
\[
F \mapsto \int_{\Theta} P_\theta(F_\theta) \rho 
\]
\[
F \mapsto \int_{\Theta} Q_\theta(F_\theta) \mu
\]
are probability distributions on \((\Theta \times X, \mathcal{F} \otimes \mathcal{X})\). Let us assume moreover that one of the following conditions are satisfied:

1. \( \mathcal{F} = \sigma(\mathcal{G}) \), where \( \mathcal{G} \) is a countable family of subsets of \( X \).
2. \( \Theta \) is countable.

In this case, the function
\[
\theta \mapsto \mathcal{K}(P_\theta, Q_\theta) : \Theta \to [0, \infty]
\]
is “pseudo-integrable” with respect to \( \rho \), in the sense that its upper integral is equal to its lower integral: letting \( \mathcal{L}_+(\Theta, \mathcal{F}) \) denote the set of measurable functions with values ranging in \([0, \infty]\),
\[
\sup \left\{ \int_{\Theta} f \rho : f \in \mathcal{L}_+(\Theta, \mathcal{F}) \text{ and } f(\theta) \leq \mathcal{K}(P_\theta, Q_\theta) \text{ for any } \theta \in \Theta \right\} = \inf \left\{ \int_{\Theta} f \rho : f \in \mathcal{L}_+(\Theta, \mathcal{F}) \text{ and } f(\theta) \geq \mathcal{K}(P_\theta, Q_\theta) \text{ for any } \theta \in \Theta \right\}.
\]
(With this definition, any function equal to \(+\infty\) on a set of positive measure is pseudo-integrable. Moreover a pseudo-integrable function with a finite integral is measurable with respect to the completion of the \( \sigma \)-algebra \( \mathcal{F} \) with respect to the measure \( \rho \).)

Under these hypotheses, the Kullback divergence function satisfies the following decomposition formula:
\[
\mathcal{K}(P, Q) = \mathcal{K}(\rho, \mu) + E_{\rho(d\theta)} \left[ \mathcal{K}(P_\theta, Q_\theta) \right].
\]
Proof. The fact that \( F_\theta \) is measurable is a classical preliminary to the proof of Fubini's theorem: the sets \( F \in \mathcal{T} \otimes \mathcal{F} \) which satisfy this property form a \( \sigma \)-algebra containing the rectangles. It can then be established that for any \( F \in \mathcal{T} \otimes \mathcal{F} \)
\[
\theta \mapsto P_\theta(F_\theta)
\]
is measurable. Indeed, this property holds for disjoint unions of rectangles, it is also stable when increasing and decreasing limits of subsets are taken (a pointwise limit of measurable functions being measurable), therefore it holds for any set of the product \( \sigma \)-algebra, according to the monotone class theorem.

The fact that \( P \) and \( Q \) are measures on \((\Theta \times X, \mathcal{T} \otimes \mathcal{F})\) is then a consequence of the monotone convergence theorem.

Another consequence is that for any function \( h(\theta, x) \in L_+(\Theta \times X, \mathcal{T} \otimes \mathcal{F}) \),
\[
\theta \mapsto \int_X h(\theta, x) P_\theta(dx)
\]
is measurable. Indeed, in the case when \( h \) is bounded, it can be uniformly approximated by a sequence of simple functions (that is by finite linear combinations of indicator functions of measurable sets). The property is then extended to non bounded functions \( h \): any such function is the increasing limit of a sequence of bounded functions, to which the monotone convergence theorem can be applied for each value of \( P_\theta \).

Let us also notice that for any function \( h(\theta, x) \in L_+(\Theta \times X, \mathcal{T} \otimes \mathcal{F}) \),
\[
\int_{A \times B} 1_{\{P_\theta(\omega) = 0\}} Q(d\omega)
\]
and that the same holds for \( Q \). Indeed, in the case when \( h \) is bounded, it can be uniformly approximated by simple functions. The case when \( h \) is not bounded is dealt with the help of the monotone convergence theorem.

Let us discuss first the case when \( P \ll Q \). As \( \rho(A) = P(A \times X) \) and \( \mu(A) = Q(A \times X) \), it follows that \( \rho \ll \mu \). Let \( \frac{P}{Q} \) and \( \frac{\rho}{\mu} \) be some versions of the Radon Nikodym density of the corresponding measures. Let \( f(\theta, x) = \frac{P}{Q} \left[ \frac{\rho}{\mu} \right]^{-1} 1_{\{\frac{\rho}{\mu} \neq 0\}} \).

We are going to show for \( \rho \) almost all \( \theta \) that \( P_\theta \ll Q_\theta \) and that \( x \mapsto f(\theta, x) \) is a version of the density of \( P_\theta \) with respect to \( Q_\theta \). Indeed, for any \( A \in \mathcal{T} \) and any \( B \in \mathcal{F} \),
\[
\int_A \int_B f(\theta, x) Q_\theta(dx) d\rho(d\theta)
\]
\[
= \int_A \left\{ \int_B \frac{P}{Q}(\theta, x) \left[ \frac{\rho}{\mu}(\theta) \right]^{-1} 1_{\{\frac{\rho}{\mu}(\theta) \neq 0\}} Q_\theta(dx) \right\} \frac{\rho}{\mu}(\theta) d\mu(d\theta)
\]
\[
= \int_A \left\{ \int_B \frac{P}{Q}(\theta, x) 1_{\{\frac{\rho}{\mu} \neq 0\}} Q_\theta(dx) \right\} \mu(d\theta)
\]
\[
= \int_{A \times B} \frac{P}{Q} 1_{\{\frac{\rho}{\mu} = 0\}} Q.
\]

Moreover
\[
\int_{A \times B} \frac{P}{Q} 1_{\{\frac{\rho}{\mu} = 0\}} Q = \int_{A \times B} 1_{\{\frac{\rho}{\mu}(\theta) = 0\}} P(d(\theta, x))
\]
\[= \int_{A} \mathbb{1}(\frac{\nu}{\mu}(\theta) = 0) P_\theta(B) \rho(d\theta)\]
\[= \int_{A} \frac{\nu}{\mu}(\theta) \mathbb{1}(\frac{\nu}{\mu}(\theta) = 0) P_\theta(B) \mu(d\theta)\]
\[= 0.\]

This shows that
\[\int_{A} \left[ \int_{B} f(\theta, x) Q_\theta(dx) \right] \rho(d\theta) = \int_{A \times B} \frac{P}{Q} Q = P(A \times B) = \int_{A} P_\theta(B) \rho(d\theta).\]

Consequently, for any \(B \in \mathcal{F}\)
\[(1.1) \quad \rho\left( \int_{B} f(\theta, x) Q_\theta(dx) \neq P_\theta(B) \right) = 0.\]

When \(\Theta\) is countable, putting \(N = \{\theta \in \Theta : \rho(\{\theta\}) = 0\}\), one concludes that for any \(B \in \mathcal{F}\) and any \(\theta \in \Theta \setminus N\),
\[\int_{B} f(\theta, x) Q_\theta(dx) = P_\theta(B),\]
and therefore that for all these values of \(\theta, x \mapsto f(\theta, x)\) is a version of \(\frac{P_\theta}{Q_\theta}\).

On the other hand, if it is assumed that \(\mathcal{F} = \sigma(\mathcal{G})\) where \(\mathcal{G}\) is countable, the following reasoning can be carried. Let \(\mathcal{G}' = \{B \in \mathcal{F} : \mathcal{X} \setminus B \in \mathcal{G}\}\), let \(\mathcal{G}''\) be the set of finite intersections of sets of \(\mathcal{G}'\) and \(\mathcal{G}'''\) the finite unions of sets of \(\mathcal{G}''\). It is easily seen that these three sets are countable, because a countable union of countable sets is countable. Moreover, \(\mathcal{G}'''\) is the algebra generated by \(\mathcal{G}\): it is closed with respect to taking finite intersections, finite unions and complements (because any Boolean expression can be put in disjunctive normal form). Identity (1.1) is true for any \(B \in \mathcal{G}'''\). As \(\mathcal{G}'''\) is countable, the union of these events is still of null probability with respect to \(\rho\):
\[\rho\left( \int_{B} f(\theta, x) Q_\theta(dx) = P_\theta(B) \right. \left. \text{for any } B \in \mathcal{G}''' \right) = 1.\]

But for any fixed value of \(\theta\), \(\{C \in \mathcal{F} : \int_{C} f Q_\theta = P_\theta(C)\}\) is a monotone class of \(\mathcal{F}\) (from the monotone convergence theorem). It is therefore, according to the monotone class theorem, equal to the whole \(\sigma\)-algebra \(\mathcal{F}\), since it contains \(\mathcal{G}'''\). This proves that for some \(N \in \mathcal{F}\) such that \(\rho(N) = 0\), for any \(\theta \in \Theta \setminus N\),
\[\int_{B} f Q_\theta = P_\theta(B) \text{ for any } B \in \mathcal{F}.\]
Consequently, for any \(\theta \in \Theta \setminus N, x \mapsto f(\theta, x)\) is a version of \(\frac{P_\theta}{Q_\theta}\). Thus
\[\mathcal{K}(P_\theta, Q_\theta) = \int \log \left[ f(\theta, x) \right] P_\theta(dx)\]
\[= \int \log \left[ f(\theta, x) \right]_+ P_\theta(dx) - \int \log \left[ f(\theta, x) \right]_- P_\theta(dx), \quad \theta \in \Theta \setminus N,\]
where \([r]_+ = \max\{r, 0\}\) and \([r]_- = -\min\{r, 0\}\). Let us notice that the last equality makes sense because \(x \mapsto \log \left[ f(\theta, x) \right]_-\) belongs to \(L^1(P_\theta)\):
\[\int \log \left[ f(\theta, x) \right]_- P_\theta(dx) = \int f(\theta, x) \log \left[ f(\theta, x) \right]_- Q_\theta(dx) \leq e^{-1}, \quad \theta \in \Theta \setminus N,\]
(because \( r \log[r]_\leq e^{-1} \)). This shows that the function \( \theta \mapsto \mathcal{K}(P_\theta, Q_\theta) \) is measurable on \( \Theta \setminus N \), and therefore that it is measurable on \( \Theta \) with respect to the completion of the \( \sigma \)-algebra \( \mathcal{T} \) with respect to the measure \( \rho \).

Let us now prove that the function \( [\log(f)]_- \) belongs to \( \mathcal{L}^1(P) \). Indeed

\[
\int [\log(f)]_- P = \int_\Theta \left\{ \int_X [\log(f)]_- \rho \right\} \rho(d\theta) = \int_\Theta \left\{ \int_X \log\left( \frac{P_\theta}{Q_\theta} \right) \right\} \rho(d\theta) = \int_\Theta \left\{ \int_X \frac{P_\theta}{Q_\theta} \left[ \log\left( \frac{P_\theta}{Q_\theta} \right) \right] \right\} \rho(d\theta) \leq e^{-1}.
\]

In the same way \( \theta, x \mapsto \log\left( \frac{P_\theta}{Q_\theta} \right)_- \) is in \( \mathcal{L}^1(P) \). As \( P(\frac{\rho}{\mu} = 0) = 0 \), we can write

\[
\mathcal{K}(P, Q) = \int \mathbb{1}(\frac{\rho}{\mu} \neq 0) \log\left( \frac{\rho}{\mu} \right) P = \int \left\{ \frac{\rho}{\mu} \neq 0 \right\} \left[ \log(f) + \log\left( \frac{\rho}{\mu} \right) \right] P = \int \left\{ \frac{\rho}{\mu} \neq 0 \right\} \log(f) P + \int \left\{ \frac{\rho}{\mu} \neq 0 \right\} \log\left( \frac{\rho}{\mu} \right) P = \int \left\{ \frac{\rho}{\mu} \neq 0 \right\} \left\{ \int_X \log\left( f(\theta, x) \right) P_\theta(dx) \right\} \rho(d\theta) + \int \log\left( \frac{\rho}{\mu} \right) \rho(d\theta) = \int_\theta \mathcal{K}(P_\theta, Q_\theta) \rho(d\theta) + \mathcal{K}(\rho, \mu).
\]

Let us discuss now the case when \( P \) is not absolutely continuous with respect to \( Q \). Let us introduce the auxiliary probability distribution

\[
Q'(F) = \int_\Theta Q_\theta(F_\theta) \rho(d\theta).
\]

In the case when \( P \ll Q' \), then \( \theta \mapsto \mathcal{K}(P_\theta, Q_\theta) \) is \( \mathcal{T}_\rho \) measurable (where \( \mathcal{T}_\rho \) is the completion of \( \mathcal{T} \) with respect to the measure \( \rho \)). In the case when \( P \) is not absolutely continuous with respect to \( Q' \), there exists \( F \in \mathcal{T} \otimes \mathcal{X} \) such that \( P(F) > 0 \) and \( Q'(F) = 0 \). As \( P(F) = \int_\Theta P_\theta(F_\theta) \rho(d\theta) \), there is some set \( A \in \mathcal{T} \) such that \( \rho(A) > 0 \) and \( P_\theta(F_\theta) > 0 \) for any \( \theta \in A \). On the other hand \( Q'(F) = \int_\Theta Q_\theta(F_\theta) \rho(d\theta) \), which proves that \( Q_\theta(F_\theta) = 0 \) for any \( \theta \in \Theta \setminus N \) where \( N \in \mathcal{F} \) is some set such that \( \rho(N) = 0 \). It follows that \( P_\theta \) is not absolutely continuous with respect to \( Q_\theta \) at any point \( \theta \) in the set \( A \setminus N \) of measure \( \rho(A \setminus N) > 0 \). This implies that \( \theta \mapsto \mathcal{K}(P_\theta, Q_\theta) \) is equal to \( +\infty \) on the set \( A \setminus N \) of positive measure \( \rho(A \setminus N) \). It is therefore pseudo-integrable with respect to \( \rho \) in the sense indicated in the proposition.

Now that measurability and integrability issues are settled, it remains to prove the identity \( \mathcal{K}(P, Q) = \mathcal{K}(\rho, \mu) + \int_\theta \mathcal{K}(P_\theta, Q_\theta) \rho(d\theta) \). The lefthand side being equal to \( +\infty \), it is to be shown that one of the two terms on the righthand side is necessarily also equal to \( +\infty \) (as they both take their values in \([0, +\infty]\), there is no ambiguity about the definition of the sum). In the case when \( \mathcal{K}(\rho, \mu) = +\infty \), we are done, thus we can assume that \( \mathcal{K}(\rho, \mu) < +\infty \). In this latter case, \( \rho \ll \mu \), and as it was assumed that \( P \not\ll Q \), there exists \( F \in \mathcal{T} \otimes \mathcal{F} \) such that \( P(F) > 0 \)
and $Q(F) = 0$. The identities $P(F) = \int_\Theta P_\theta(F_\theta) \rho(d\theta)$ and $Q(F) = \int_\Theta Q_\theta(F_\theta) \mu(d\theta)$ show that there are two $\mathcal{T}$ measurable sets $A$ and $N$ such that $\rho(A) > 0$, $\mu(N) = 0$, $P_\theta(F_\theta) > 0$ for any $\theta \in A$, and $Q_\theta(F_\theta) = 0$ for any $\theta \in \Theta \setminus N$. As $\rho \ll \mu$, we also have $\rho(N) = 0$, and therefore $\rho(A \setminus N) > 0$. On this last set $\mathcal{K}(P_\theta, Q_\theta) = +\infty$, because $P_\theta \not\ll Q_\theta$. The conclusion is that $\int_\Theta \mathcal{K}(P_\theta, Q_\theta) \rho(d\theta) = +\infty$. \qed
Bibliography


CHAPTER 5. BIBLIOGRAPHY