Chapter 4

Support Vector Machines

4.1. How to build them

4.1.1. The canonical hyperplane. Support Vector Machines, of wide use and renown, were conceived by V. Vapkik (Vapnik, 1998). Before introducing them, we will study as a prerequisite the separation of points by hyperplanes in a finite dimensional Euclidean space. Support Vector Machines perform the same kind of linear separation after an implicit change of pattern space. The preceding PAC-Bayesian results provide a fit framework to analyse their generalization properties.

In this section we deal with the classification of points in $\mathbb{R}^d$ in two classes. Let $Z = (x_i, y_i)_{i=1}^N \in \big(\mathbb{R}^d \times \{-1, +1\}\big)^N$ be some set of labelled examples (called the training set hereafter). Let us split the set of indices $I = \{1, \dots, N\}$ according to the labels into two subsets

$$I_+ = \{i \in I : y_i = +1\},$$
$$I_- = \{i \in I : y_i = -1\}.$$

Let us then consider the set of admissible separating directions

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

which can also be written as

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

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

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

Lemma 4.1.1. *When $A_Z \neq \varnothing$, $\inf\{\|w\|^2 : w \in A_Z\}$ is reached for only one value w_Z of w.*

Proof. Let $w_0 \in A_Z$. The set $A_Z \cap \{w \in \mathbb{R}^d : \|w\| \leq \|w_0\|\}$ is a compact convex set and $w \mapsto \|w\|^2$ is strictly convex and therefore has a unique minimum on this set, which is also obviously its minimum on A_Z. □

Definition 4.1.1. When $A_Z \neq \varnothing$, 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

$$\begin{aligned} w_Z &= \arg\min\{\|w\| : w \in A_Z\}, \\ b_Z &= b_{w_Z}, \end{aligned}$$

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

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.

4.1.2. Computation of the canonical hyperplane. Let us consider the convex hulls X_+ and X_- of the positive and negative patterns:

$$\begin{aligned} \mathcal{X}_+ &= \Big\{\sum_{i\in I_+}\lambda_i x_i : (\lambda_i)_{i\in I_+} \in \mathbb{R}_+^{I_+}, \sum_{i\in I_+}\lambda_i = 1\Big\}, \\ \mathcal{X}_- &= \Big\{\sum_{i\in I_-}\lambda_i x_i : (\lambda_i)_{i\in I_-} \in \mathbb{R}_+^{I_-}, \sum_{i\in I_-}\lambda_i = 1\Big\}. \end{aligned}$$

Let us introduce the closed convex set

$$\mathcal{V} = \mathcal{X}_+ - \mathcal{X}_- = \big\{x_+ - x_- : x_+ \in \mathcal{X}_+, x_- \in \mathcal{X}_-\big\}.$$

As $v \mapsto \|v\|^2$ is strictly convex, with compact lower level sets, there is a unique vector v^* such that

$$\|v^*\|^2 = \inf_{v\in\mathcal{V}}\big\{\|v\|^2 : v \in \mathcal{V}\big\}.$$

Lemma 4.1.2. *The set A_Z is non-empty (i.e. the training set Z is linearly separable) if and only if $v^* \neq 0$. In this case*

$$w_Z = \frac{2}{\|v^*\|^2}v^*,$$

and the margin of the canonical hyperplane is equal to $\frac{1}{2}\|v^*\|$.

This lemma proves that the distance between the convex hulls of the positive and negative patterns is equal to twice the margin of the canonical hyperplane.

Proof. Let us assume first that $v^* = 0$, or equivalently that $\mathcal{X}_+ \cap \mathcal{X}_- \neq \varnothing$. For any vector $w \in \mathbb{R}^d$,

$$\begin{aligned} \min_{i\in I_+}\langle w, x_i\rangle &= \min_{x\in\mathcal{X}_+}\langle w, x\rangle, \\ \max_{i\in I_-}\langle w, x_i\rangle &= \max_{x\in\mathcal{X}_-}\langle w, x\rangle, \end{aligned}$$

so $\min_{i\in I_+}\langle w, x_i\rangle - \max_{i\in I_-}\langle w, x_i\rangle \leq 0$, 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 $\mathcal{X}_+ \cap \mathcal{X}_- = \varnothing$. Let us put $w^* = 2v^*/\|v^*\|^2$. Let us remark first that

$$\begin{aligned}\min_{i\in I_+}\langle w^*, x_i\rangle - \max_{i\in I_-}\langle w^*, x_i\rangle &= \inf_{x\in\mathcal{X}_+}\langle w^*, x\rangle - \sup_{x\in\mathcal{X}_-}\langle w^*, x\rangle \\ &= \inf_{x_+\in\mathcal{X}_+, x_-\in\mathcal{X}_-}\langle w^*, x_+ - x_-\rangle \\ &= \frac{2}{\|v^*\|^2}\inf_{v\in\mathcal{V}}\langle v^*, v\rangle.\end{aligned}$$

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

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

This proves that $\|w^*\| = \inf\{\|w\| : w \in A_Z\}$, and therefore that $w^* = w_Z$ as claimed. □

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

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

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

PROPOSITION 4.1.3. *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

$$\inf\{F(\alpha) : \alpha \in \mathcal{A}\}$$

where

$$\mathcal{A} = \left\{(\alpha_i)_{i\in I} \in \mathbb{R}_+^I, \sum_{i\in I}\alpha_i y_i = 0\right\},$$

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 $s \in \mathbb{R}_+$, $\{w(\alpha) : \alpha \in \mathcal{A}, S(\alpha) = s\} = s\mathcal{V}$. This shows that for any $s \in \mathbb{R}_+$, $\inf\{F(\alpha) : \alpha \in \mathcal{A}, S(\alpha) = s\}$ is reached and that for any $\alpha_s \in \{\alpha \in \mathcal{A} : S(\alpha) = s\}$ 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 $\mathcal{A}$, and that for any $\alpha^* \in \mathcal{A}$ such that $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$, $w(\alpha^*) = \frac{2}{\|v^*\|^2}v^* = w_Z$. □

4.1.3. SUPPORT VECTORS.

DEFINITION 4.1.2. The set of support vectors $\mathcal{S}$ is defined by

$$\mathcal{S} = \{x_i : \langle w_Z, x_i\rangle - b_Z = y_i\}.$$

PROPOSITION 4.1.4. *Any α^* minimizing $F(\alpha)$ on $\mathcal{A}$ is such that*

$$\{x_i : \alpha_i^* > 0\} \subset \mathcal{S}.$$

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 $Z' = \{(x_i, y_i) : x_i \in \mathcal{S}\}$.

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 α^* on each of these sets are equal and since $\sum_{k\in I} \alpha_k^* > 0$. For any $t \in \mathbb{R}$, consider

$$\alpha_k(t) = \alpha_k^* + t\mathbb{1}(k \in \{i, j\}), \quad k \in I.$$

The vector $\alpha(t)$ is in $\mathcal{A}$ for any value of t in some neighbourhood of 0, therefore $\frac{\partial}{\partial t}_{|t=0} F\big[\alpha(t)\big] = 0$. Computing this derivative, we find that

$$y_i\langle w(\alpha^*), x_i\rangle + y_j\langle w(\alpha^*), x_j\rangle = 2.$$

As $y_i = -y_j$, this can also be written as

$$y_i\big[\langle w(\alpha^*), x_i\rangle - b_Z\big] + y_j\big[\langle w(\alpha^*), x_j\rangle - b_Z\big] = 2.$$

As $w(\alpha^*) \in A_Z$,

$$y_k\big[\langle w(\alpha^*), x_k\rangle - b_Z\big] \geq 1, \qquad k \in I,$$

which implies necessarily as claimed that

$$y_i\big[\langle w(\alpha^*), x_i\rangle - b_Z\big] = y_j\big[\langle w(\alpha^*), x_j\rangle - b_Z\big] = 1.$$

□

4.1.4. THE NON-SEPARABLE CASE. In the case when the training set $Z = (x_i, y_i)_{i=1}^N$ is not linearly separable, we can define a noisy canonical hyperplane as follows: we can choose $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ to minimize

$$C(w, b) = \sum_{i=1}^{N} \big[1 - \big(\langle w, x_i\rangle - b\big)y_i\big]_+ + \tfrac{1}{2}\|w\|^2, \tag{4.1}$$

where for any real number r, $r_+ = \max\{r, 0\}$ is the positive part of r.

THEOREM 4.1.5. *Let us introduce the dual criterion*

$$F(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\Big\|\sum_{i=1}^{N} y_i\alpha_i x_i\Big\|^2$$

and the domain $\mathcal{A}' = \Big\{\alpha \in \mathbb{R}_+^N : \alpha_i \le 1, i = 1,\dots,N, \sum_{i=1}^{N} y_i\alpha_i = 0\Big\}$. *Let* $\alpha^* \in \mathcal{A}'$ *be such that* $F(\alpha^*) = \sup_{\alpha\in\mathcal{A}'} F(\alpha)$. *Let* $w^* = \sum_{i=1}^{N} y_i\alpha_i^* x_i$. *There is a threshold* b^* *(whose construction will be detailed in the proof), such that*

$$C(w^*, b^*) = \inf_{w\in\mathbb{R}^d, b\in\mathbb{R}} C(w, b).$$

COROLLARY 4.1.6. (SCALED CRITERION) *For any positive real parameter* λ *let us consider the criterion*

$$C_\lambda(w, b) = \lambda^2 \sum_{i=1}^{N} \big[1 - (\langle w, x_i\rangle - b)y_i\big]_+ + \tfrac{1}{2}\|w\|^2$$

and the domain

$$\mathcal{A}'_\lambda = \Big\{\alpha \in \mathbb{R}_+^N : \alpha_i \le \lambda^2, i = 1,\dots,N, \sum_{i=1}^{N} y_i\alpha_i = 0\Big\}.$$

For any solution α^* *of the minimization problem* $F(\alpha^*) = \sup_{\alpha\in\mathcal{A}'_\lambda} F(\alpha)$, *the vector* $w^* = \sum_{i=1}^{N} y_i\alpha_i^* x_i$ *is such that*

$$\inf_{b\in\mathbb{R}} C_\lambda(w^*, b) = \inf_{w\in\mathbb{R}^d, b\in\mathbb{R}} C_\lambda(w, b).$$

In the separable case, the scaled criterion is minimized by the canonical hyperplane for λ large enough. This extension of the canonical hyperplane computation in dual space is often called *the box constraint*, for obvious reasons.

PROOF. The corollary is a straightforward consequence of the scale property $C_\lambda(w, b, x) = \lambda^2 C(\lambda^{-1}w, b, \lambda x)$, where we have made the dependence of the criterion in $x \in \mathbb{R}^{dN}$ explicit. Let us come now to the proof of the theorem.

The minimization of $C(w, b)$ can be performed in dual space extending the couple of parameters (w, b) to $\overline{w} = (w, b, \gamma) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+^N$ and introducing the dual multipliers $\alpha \in \mathbb{R}_+^N$ and the criterion

$$G(\alpha, \overline{w}) = \sum_{i=1}^{N} \gamma_i + \sum_{i=1}^{N} \alpha_i\big\{\big[1 - (\langle w, x_i\rangle - b)y_i\big] - \gamma_i\big\} + \tfrac{1}{2}\|w\|^2.$$

We see that

$$C(w, b) = \inf_{\gamma\in\mathbb{R}_+^N} \sup_{\alpha\in\mathbb{R}_+^N} G\big[\alpha, (w, b, \gamma)\big],$$

and therefore, putting $\overline{\mathcal{W}} = \big\{(w, b, \gamma) : w \in \mathbb{R}^d, b \in \mathbb{R}, \gamma \in \mathbb{R}_+^N\big\}$, we are led to solve the minimization problem

$$G(\alpha_*, \overline{w}_*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha, \overline{w}),$$

whose solution $\overline{w}_* = (w_*, b_*, \gamma_*)$ is such that $C(\overline{w}_*, b_*) = \inf_{(w,b)\in\mathbb{R}^{d+1}} C(w,b)$, according to the preceding identity. As for any value of $\alpha' \in \mathbb{R}^N_+$,

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}^N_+} G(\alpha, \overline{w}) \geq \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha', \overline{w}),$$

it is immediately seen that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}^N_+} G(\alpha, \overline{w}) \geq \sup_{\alpha\in\mathbb{R}^N_+} \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha, \overline{w}).$$

We are going to show that there is no duality gap, meaning that this inequality is indeed an equality. More importantly, we will do so by exhibiting a saddle point, which, solving the dual minimization problem will also solve the original one.

Let us first make explicit the solution of the dual problem (the interest of this dual problem precisely lies in the fact that it can more easily be solved explicitly). Introducing the admissible set of values of α,

$$\mathcal{A}' = \big\{\alpha \in \mathbb{R}^N : 0 \leq \alpha_i \leq 1, i = 1, \ldots, N, \sum_{i=1}^N y_i\alpha_i = 0\big\},$$

it is elementary to check that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha, \overline{w}) = \begin{cases} \inf_{w\in\mathbb{R}^d} G\big[\alpha, (w,0,0)\big], & \alpha \in \mathcal{A}', \\ -\infty, & \text{otherwise.} \end{cases}$$

As

$$G\big[\alpha, (w,0,0)\big] = \tfrac{1}{2}\|w\|^2 + \sum_{i=1}^N \alpha_i\big(1 - \langle w, x_i\rangle y_i\big),$$

we see that $\inf_{w\in\mathbb{R}^d} G\big[\alpha, (w,0,0)\big]$ is reached at

$$w_\alpha = \sum_{i=1}^N y_i\alpha_i x_i.$$

This proves that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha, \overline{w}) = F(\alpha).$$

The continuous map $\alpha \mapsto \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha, \overline{w})$ reaches a maximum α^*, not necessarily unique, on the compact convex set $\mathcal{A}'$. We are now going to exhibit a choice of $\overline{w}^* \in \overline{\mathcal{W}}$ such that $(\alpha^*, \overline{w}^*)$ is a *saddle point*. This means that we are going to show that

$$G(\alpha^*, \overline{w}^*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha^*, \overline{w}) = \sup_{\alpha\in\mathbb{R}^N_+} G(\alpha, \overline{w}^*).$$

It will imply that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}^d_+} G(\alpha, \overline{w}) \leq \sup_{\alpha\in\mathbb{R}^N_+} G(\alpha, \overline{w}^*) = G(\alpha^*, \overline{w}^*)$$

on the one hand and that

$$\inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}^d_+} G(\alpha, \overline{w}) \geq \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha^*, \overline{w}) = G(\alpha^*, \overline{w}^*)$$

on the other hand, proving that

$$G(\alpha^*, \overline{w}^*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha, \overline{w})$$

as required.
Construction of $\overline{w}^*$.

- Let us put $w^* = w_{\alpha^*}$.
- If there is $j \in \{1, \dots, N\}$ such that $0 < \alpha_j^* < 1$, let us put

$$b^* = \langle x_j, w^*\rangle - y_j.$$

  Otherwise, let us put

$$b^* = \sup\{\langle x_i, w^*\rangle - 1 : \alpha_i^* > 0, y_i = +1, i = 1, \dots, N\}.$$

- Let us then put

$$\gamma_i^* = \begin{cases} 0, & \alpha_i^* < 1, \\ 1 - (\langle w^*, x_i\rangle - b^*)y_i, & \alpha_i^* = 1. \end{cases}$$

If we can prove that

$$1 - (\langle w^*, x_i\rangle - b^*)y_i \begin{cases} \le 0, & \alpha_i^* = 0, \\ = 0, & 0 < \alpha_i^* < 1, \\ \ge 0, & \alpha_i^* = 1, \end{cases} \tag{4.2}$$

it will show that $\gamma^* \in \mathbb{R}_+^N$ and therefore that $\overline{w}^* = (w^*, b^*, \gamma^*) \in \overline{\mathcal{W}}$. It will also show that

$$G(\alpha, \overline{w}^*) = \sum_{i=1}^N \gamma_i^* + \sum_{i,\alpha_i^*=0} \alpha_i\big[1 - (\langle \overline{w}^*, x_i\rangle - b^*)y_i\big] + \tfrac{1}{2}\|\overline{w}^*\|^2,$$

proving that $G(\alpha^*, \overline{w}^*) = \sup_{\alpha\in\mathbb{R}_+^N} G(\alpha, \overline{w}^*)$. As obviously $G(\alpha^*, \overline{w}^*) = G\big[\alpha^*, (w^*, 0, 0)\big]$, we already know that $G(\alpha^*, \overline{w}^*) = \inf_{\overline{w}\in\overline{\mathcal{W}}} G(\alpha^*, \overline{w})$. This will show that $(\alpha^*, \overline{w}^*)$ is the saddle point we were looking for, thus ending the proof of the theorem. □

Proof of equation (4.2). Let us deal first with the case when there is $j \in \{1, \dots, N\}$ such that $0 < \alpha_j^* < 1$.

For any $i \in \{1, \dots, N\}$ such that $0 < \alpha_i^* < 1$, there is $\epsilon > 0$ such that for any $t \in (-\epsilon, \epsilon)$, $\alpha^* + ty_ie_i - ty_je_j \in \mathcal{A}'$, where $(e_k)_{k=1}^N$ is the canonical base of $\mathbb{R}^N$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + ty_ie_i - ty_je_j) = 0$. Computing this derivative, we obtain

$$\begin{aligned} \frac{\partial}{\partial t}_{|t=0} F(\alpha^* + ty_ie_i - ty_je_j) &= y_i - \langle w^*, x_i\rangle + \langle w^*, x_j\rangle - y_j \\ &= y_i\big[1 - \big(\langle w, x_i\rangle - b^*\big)y_i\big]. \end{aligned}$$

Thus $1 - \big(\langle w, x_i\rangle - b^*\big)y_i = 0$, as required. This shows also that the definition of b^* does not depend on the choice of j such that $0 < \alpha_j^* < 1$.

For any $i \in \{1, \dots, N\}$ such that $\alpha_i^* = 0$, there is $\epsilon > 0$ such that for any $t \in (0, \epsilon)$, $\alpha^* + te_i - ty_iy_je_j \in \mathcal{A}'$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + te_i - ty_iy_je_j) \leq 0$, showing that $1 - \big(\langle w^*, x_i\rangle - b^*\big)y_i \leq 0$ as required.

For any $i \in \{1, \dots, N\}$ such that $\alpha_i^* = 1$, there is $\epsilon > 0$ such that $\alpha^* - te_i + ty_iy_je_j \in \mathcal{A}'$. Thus $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* - te_i + ty_iy_je_j) \leq 0$, showing that $1 - \big(\langle w^*, x_i\rangle - b^*\big)y_i \geq 0$ as required. This shows that $(\alpha^*, \overline{w}^*)$ is a saddle point in this case.

Let us deal now with the case where $\alpha^* \in \{0, 1\}^N$. If we are not in the trivial case where the vector $(y_i)_{i=1}^N$ is constant, the case $\alpha^* = 0$ is ruled out. Indeed, in this case, considering $\alpha^* + te_i + te_j$, where $y_iy_j = -1$, we would get the contradiction $2 = \frac{\partial}{\partial t}_{|t=0} F(\alpha^* + te_i + te_j) \leq 0$.

Thus there are values of j such that $\alpha_j^* = 1$, and since $\sum_{i=1}^N \alpha_i y_i = 0$, both classes are present in the set $\{j : \alpha_j^* = 1\}$.

Now for any $i, j \in \{1, \dots, N\}$ such that $\alpha_i^* = \alpha_j^* = 1$ and such that $y_i = +1$ and $y_j = -1$, $\frac{\partial}{\partial t}_{|t=0} F(\alpha^* - te_i - te_j) = -2 + \langle w^*, x_i\rangle - \langle w^*, x_j\rangle \leq 0$. Thus

$$\sup\{\langle w^*, x_i\rangle - 1 : \alpha_i^* = 1, y_i = +1\} \leq \inf\{\langle w^*, x_j\rangle + 1 : \alpha_j^* = 1, y_j = -1\},$$

showing that

$$1 - \big(\langle w^*, x_k\rangle - b^*\big)y_k \geq 0, \alpha_k^* = 1.$$

Finally, for any i such that $\alpha_i^* = 0$, for any j such that $\alpha_j^* = 1$ and $y_j = y_i$, we have

$$\frac{\partial}{\partial t}_{|t=0} F(\alpha^* + te_i - te_j) = y_i\langle w^*, x_i - x_j\rangle \leq 0,$$

showing that $1 - \big(\langle w^*, x_i\rangle - b^*\big)y_i \leq 0$. This shows that $(\alpha^*, \overline{w}^*)$ is always a saddle point.

4.1.5. Support Vector Machines.

Definition 4.1.3. The symmetric measurable kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 previously

$$\mathcal{A} = \left\{ \alpha \in \mathbb{R}_+^N \ : \sum_{i=1}^N \alpha_i y_i = 0 \right\}.$$

Let

$$F(\alpha) = \sum_{i=1}^N \sum_{j=1}^N \alpha_i y_i K(x_i, x_j) y_j \alpha_j - 2\sum_{i=1}^N \alpha_i.$$

Definition 4.1.4. Let K be a positive symmetric kernel. The training set Z is said to be K-separable if

$$\inf\big\{F(\alpha) \ : \alpha \in \mathcal{A}\big\} > -\infty.$$

LEMMA 4.1.7. *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

$$x'_i = \left\{ \left[\left\{ K(x_k, x_\ell) \right\}_{k=1,\ell=1}^{N \quad N} \right]^{1/2} (i,j) \right\}_{j=1}^N \in \mathbb{R}^N.$$

We see that $F(\alpha) = \|\sum_{i=1}^N \alpha_i y_i x'_i\|^2 - 2\sum_{i=1}^N \alpha_i$. We 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. □

PROPOSITION 4.1.8. *Let K be a symmetric positive kernel and let $Z = (x_i, y_i)_{i=1}^N$ be some K-separable training set. Let $\alpha^* \in \mathcal{A}$ be such that $F(\alpha^*) = \inf\{F(\alpha) : \alpha \in \mathcal{A}\}$. Let*

$$I_-^* = \{i \in \mathbb{N} : 1 \le i \le N, y_i = -1, \alpha_i^* > 0\}$$
$$I_+^* = \{i \in \mathbb{N} : 1 \le i \le N, y_i = +1, \alpha_i^* > 0\}$$
$$b^* = \frac{1}{2}\Big\{ \sum_{j=1}^N \alpha_j^* y_j K(x_j, x_{i_-}) + \sum_{j=1}^N \alpha_j^* y_j K(x_j, x_{i_+}) \Big\}, \qquad i_- \in I_-^*, i_+ \in I_+^*,$$

where the value of b^* does not depend on the choice of i_- and i_+. The classification rule $f : \mathcal{X} \to \mathcal{Y}$ defined by the formula

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^N \alpha_i^* y_i K(x_i, x) - b^* \right)$$

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

An important consequence of this proposition is that the support vector machine defined by K and Z is also the support vector machine defined by K and $Z' = \{(x_i, y_i) : \alpha_i^* > 0, 1 \le i \le N\}$, since this restriction of the index set contains the value α^* where the minimum of F is reached.

PROOF. The independence of the choice of α^*, which is not necessarily unique, is seen as follows. Let $(x_i)_{i=1}^N$ and $x \in \mathcal{X}$ be fixed. Let us put for ease of notation $x_{N+1} = x$. Let M be the $(N+1) \times (N+1)$ symmetric semi-definite matrix defined by $M(i,j) = K(x_i, x_j)$, $i = 1, \dots, N+1$, $j = 1, \dots, N+1$. Let us consider the mapping $\Psi : \{x_i : i = 1, \dots, N+1\} \to \mathbb{R}^{N+1}$ defined by

$$\Psi(x_i) = \left[M^{1/2}(i,j) \right]_{j=1}^{N+1} \in \mathbb{R}^{N+1}. \tag{4.3}$$

Let us consider the training set $Z' = \left[\Psi(x_i), y_i \right]_{i=1}^N$. Then Z' is linearly separable,

$$F(\alpha) = \Big\| \sum_{i=1}^N \alpha_i y_i \Psi(x_i) \Big\|^2 - 2\sum_{i=1}^N \alpha_i,$$

and we have proved that for any choice of $\alpha^* \in \mathcal{A}$ 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

$$f(x) = \operatorname{sign}\Big[\langle w_{Z'}, \Psi(x) \rangle - b_{Z'} \Big]$$

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

We can in the same way use the box constraint and show that any solution $\alpha^* \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i \le \lambda^2, i = 1, \dots, N\}$ minimizes

$$\inf_{b\in\mathbb{R}} \lambda^2 \sum_{i=1}^{N} \Big[1 - \Big(\sum_{j=1}^{N} y_j \alpha_j K(x_j, x_i) - b\Big) y_i\Big]_+ + \frac{1}{2} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j). \tag{4.4}$$

4.1.6. Building kernels. Except the last, the results of this section are drawn from Cristianini et al. (2000). We have no reference for the last proposition of this section, although we believe it is well known. We include them for the convenience of the reader.

Proposition 4.1.9. *Let K_1 and K_2 be positive symmetric kernels on $\mathcal{X}$. Then for any $a \in \mathbb{R}_+$*

$$(aK_1 + K_2)(x, x') \stackrel{\text{def}}{=} aK_1(x, x') + K_2(x, x')$$
$$\text{and } (K_1 \cdot K_2)(x, x') \stackrel{\text{def}}{=} K_1(x, x')K_2(x, x')$$

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

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

$$\begin{aligned}
\langle \alpha, (aK_1 + K_2)\alpha\rangle &= a\langle \alpha, K_1\alpha\rangle + \langle \alpha, K_2\alpha\rangle \ge 0,\\
\langle \alpha, (K_1 \cdot K_2)\alpha\rangle &= \sum_{i,j} \alpha_i K_1(i,j) K_2(i,j) \alpha_j\\
&= \sum_{i,j,k} \alpha_i K_1^{1/2}(i,k) K_1^{1/2}(k,j) K_2(i,j) \alpha_j\\
&= \sum_{k} \underbrace{\sum_{i,j} \big[K_1^{1/2}(k,i)\alpha_i\big] K_2(i,j) \big[K_1^{1/2}(k,j)\alpha_j\big]}_{\ge 0} \ge 0,\\
\langle \alpha, K_g\alpha\rangle &= \sum_{i,j} \alpha_i g(i) g(j) \alpha_j = \Big(\sum_i \alpha_i g(i)\Big)^2 \ge 0.
\end{aligned}$$

□

Proposition 4.1.10. *Let K be some positive symmetric kernel on $\mathcal{X}$. Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial with positive coefficients. Let $g : \mathcal{X} \to \mathbb{R}^d$ be a measurable function. Then*

$$p(K)(x, x') \stackrel{def}{=} p\big[K(x, x')\big],$$

$$\exp(K)(x,x') \stackrel{def}{=} \exp\big[K(x,x')\big]$$
$$\textit{and } G_g(x,x') \stackrel{def}{=} \exp\big(-\|g(x)-g(x')\|^2\big)$$

are all positive symmetric kernels.

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

$$G_g(x,x') = \Big[\exp\big(-\|g(x)\|^2\big)\exp\big(-\|g(x')\|^2\big)\Big]\exp\big[2\langle g(x),g(x')\rangle\big]$$

□

Proposition 4.1.11. *With the notation of the previous proposition,* any *training set* $Z=(x_i,y_i)_{i=1}^N \in \big(\mathcal{X}\times\{-1,+1\}\big)^N$ *is* G_g*-separable as soon as* $g(x_i)$, $i=1,\dots,N$ *are distinct points of* $\mathbb{R}^d$.

Proof. It is clearly enough to prove the case when $\mathcal{X}=\mathbb{R}^d$ and g is the identity. Let us consider some other generic point $x_{N+1}\in\mathbb{R}^d$ and define Ψ as in (4.3). It is enough to prove that $\Psi(x_1),\dots,\Psi(x_N)$ are affine independent, since the simplex, and therefore any affine independent set of points, can be split in any arbitrary way by affine half-spaces. Let us assume that $(x_1,\dots,x_N)$ are affine dependent; then for some $(\lambda_1,\dots,\lambda_N)\neq 0$ such that $\sum_{i=1}^N\lambda_i=0$,

$$\sum_{i=1}^N\sum_{j=1}^N \lambda_i G(x_i,x_j)\lambda_j = 0.$$

Thus, $(\lambda_i)_{i=1}^{N+1}$, where we have put $\lambda_{N+1}=0$ is in the kernel of the symmetric positive semi-definite matrix $G(x_i,x_j)_{i,j\in\{1,\dots,N+1\}}$. Therefore

$$\sum_{i=1}^N \lambda_i G(x_i,x_{N+1}) = 0,$$

for any $x_{N+1}\in\mathbb{R}^d$. This would mean that the functions $x\mapsto\exp(-\|x-x_i\|^2)$ are linearly dependent, which can be easily proved to be false. Indeed, let $n\in\mathbb{R}^d$ be such that $\|n\|=1$ and $\langle n,x_i\rangle$, $i=1,\dots,N$ are distinct (such a vector exists, because it has to be outside the union of a finite number of hyperplanes, which is of zero Lebesgue measure on the sphere). Let us assume for a while that for some $(\lambda_i)_{i=1}^N\in\mathbb{R}^N$, for any $x\in\mathbb{R}^d$,

$$\sum_{i=1}^N \lambda_i \exp(-\|x-x_i\|^2) = 0.$$

Considering $x=tn$, for $t\in\mathbb{R}$, we would get

$$\sum_{i=1}^N \lambda_i \exp(2t\langle n,x_i\rangle - \|x_i\|^2) = 0, \qquad t\in\mathbb{R}.$$

Letting t go to infinity, we see that this is only possible if $\lambda_i=0$ for all values of i.
□

4.2. Bounds for Support Vector Machines

4.2.1. Compression scheme bounds. We can use Support Vector Machines in the framework of compression schemes and apply Theorem 3.3.3 (page 125). More precisely, given some positive symmetric kernel K on $\mathcal{X}$, we may consider for any training set $Z' = (x'_i, y'_i)_{i=1}^h$ 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 described on page 117, 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 Theorem 3.3.3 in the case when the observed sample $(X_i, Y_i)_{i=1}^N$ is K-separable, a natural if perhaps sub-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 4.1.8 (page 139). If Z is not K-separable, we can train a Support Vector Machine with the box constraint, then remove all the errors to obtain a K-separable sub-sample $Z' = \{(X_i, Y_i) : \alpha_i^* < \lambda^2, 1 \leq i \leq N\}$, using the same notation as in equation (4.4) on page 140, and then consider its support vectors as the compression set. Still using the notation of page 140, this means we have to compute successively $\alpha^* \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i \leq \lambda^2\}$, and $\alpha^{**} \in \arg\min\{F(\alpha) : \alpha \in \mathcal{A}, \alpha_i = 0 \text{ when } \alpha_i^* = \lambda^2\}$, to keep the compression set indexed by $J = \{i : 1 \leq i \leq N, \alpha_i^{**} > 0\}$, and the corresponding Support Vector Machine $\widehat{f}_J$. Different values of λ can be used at this stage, producing different candidate compression sets: when λ increases, the number of errors should decrease, on the other hand when λ decreases, the margin $\|w\|^{-1}$ of the separable subset Z' increases, supporting the hope for a smaller set of support vectors, thus we can use λ to monitor the number of errors on the training set we accept from the compression scheme. As we can use whatever heuristic we want while selecting the compression set, we can also try to threshold in the previous construction α_i^{**} at different levels $\eta \geq 0$, to produce candidate compression sets $J_\eta = \{i : 1 \leq i \leq N, \alpha_i^{**} > \eta\}$ of various sizes.

As the size $|J|$ of the compression set is random in this construction, we must use a version of Theorem 3.3.3 (page 125) which handles compression sets of arbitrary sizes. This is done by choosing for each k a k-partially exchangeable posterior distribution π_k which weights the compression sets of all dimensions. We immediately see that we can choose π_k such that $-\log\left[\pi_k(\Delta_k(J))\right] \leq \log\left[|J|(|J|+1)\right] + |J| \log\left[\frac{(k+1)eN}{|J|}\right]$.

If we observe the shadow sample patterns, and if computer resources permit, we can of course use more elaborate bounds than Theorem 3.3.3, such as the transductive equivalent for Theorem 1.3.15 (page 30) (where we may consider the submodels made of all the compression sets of the same size). Theorems based on relative bounds, such as Theorem 2.2.4 (page 72) or Theorem 2.3.9 (page 107) can also be used. Gibbs distributions can be approximated by Monte Carlo techniques, where a Markov chain with the proper invariant measure consists in appropriate local perturbations of the compression set.

Let us mention also that the use of compression schemes based on Support Vector Machines can be tailored to perform some kind of *feature aggregation*. Imagine that the kernel K is defined as the scalar product in $L_2(\pi)$, where $\pi \in \mathcal{M}_+^1(\Theta)$. More precisely let us consider for some set of soft classification rules $\left\{f_\theta : \mathcal{X} \to \mathbb{R}\,; \theta \in \Theta\right\}$

the kernel

$$K(x,x') = \int_{\theta\in\Theta} f_\theta(x) f_\theta(x') \pi(d\theta).$$

In this setting, the Support Vector Machine applied to the training set $Z = (x_i, y_i)_{i=1}^N$ has the form

$$f_Z(x) = \operatorname{sign}\left(\int_{\theta\in\Theta} f_\theta(x) \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \pi(d\theta) - b \right)$$

and, if this is too burdensome to compute, we can replace it with some finite approximation

$$\widetilde{f}_Z(x) = \operatorname{sign}\left(\frac{1}{m} \sum_{k=1}^m f_{\theta_k}(x) w_k - b \right),$$

where the set $\{\theta_k,\, k = 1,\dots,m\}$ and the weights $\{w_k,\, k = 1,\dots,m\}$ are computed in some suitable way from the set $Z' = (x_i, y_i)_{i,\alpha_i>0}$ of support vectors of f_Z. For instance, we can draw $\{\theta_k,\, k = 1,\dots,m\}$ at random according to the probability distribution proportional to

$$\left| \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \right| \pi(d\theta),$$

define the weights w_k by

$$w_k = \operatorname{sign}\left(\sum_{i=1}^N y_i \alpha_i f_{\theta_k}(x_i) \right) \int_{\theta\in\Theta} \left| \sum_{i=1}^N y_i \alpha_i f_\theta(x_i) \right| \pi(d\theta),$$

and choose the smallest value of m for which this approximation still classifies Z' without errors. Let us remark that we have built $\widetilde{f}_Z$ in such a way that

$$\lim_{m\to+\infty} \widetilde{f}_Z(x_i) = f_Z(x_i) = y_i, \quad \text{a.s.}$$

for any support index i such that $\alpha_i > 0$.

Alternatively, given Z', we can select a finite set of features $\Theta' \subset \Theta$ such that Z' is $K_{\Theta'}$ separable, where $K_{\Theta'}(x,x') = \sum_{\theta\in\Theta'} f_\theta(x) f_\theta(x')$ and consider the Support Vector Machines $f_{Z'}$ built with the kernel $K_{\Theta'}$. As soon as Θ' is chosen as a function of Z' only, Theorem 3.3.3 (page 125) applies and provides some level of confidence for the risk of $f_{Z'}$.

4.2.2. The Vapnik–Cervonenkis dimension of a family of subsets. Let us consider some set X and some set $S \subset \{0,1\}^X$ of subsets of X. Let $h(S)$ be the Vapnik–Cervonenkis dimension of S, defined as

$$h(S) = \max\Big\{ |A| : A \subset X, |A| < \infty \text{ and } A \cap S = \{0,1\}^A \Big\},$$

where by definition $A \cap S = \{A \cap B : B \in S\}$ and $|A|$ is the number of points in A. 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 or any bigger set. Let us notice also that for any set B, $h(B \cap S) \le h(S)$, the reason being that $A \cap (B \cap S) = B \cap (A \cap S)$.

This notion of Vapnik–Cervonenkis dimension is useful because, as we will see for Support Vector Machines, it can be computed in some important special cases. Let us prove here as an illustration that $h(S) = d+1$ when $X = \mathbb{R}^d$ and S is made of all the 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 4.2.1. *With the previous notation, $h(S) = d+1$.*

PROOF. Let $(e_i)_{i=1}^{d+1}$ be the canonical base of $\mathbb{R}^{d+1}$, and let X be the affine subspace it generates, which can be identified with $\mathbb{R}^d$. For any $(\epsilon_i)_{i=1}^{d+1} \in \{-1,+1\}^{d+1}$, let $w = \sum_{i=1}^{d+1} \epsilon_i e_i$ and $b = 0$. The half space $A_{w,b} \cap X$ is such that $\{e_i\,; i = 1,\dots,d+1\} \cap (A_{w,b} \cap X) = \{e_i\,; \epsilon_i = +1\}$. This proves that $h(S) \geq d+1$.

To prove that $h(S) \leq d+1$, we have to show that for any set $A \subset \mathbb{R}^d$ of size $|A| = d+2$, there is $B \subset A$ such that $B \notin (A \cap S)$. Obviously this will be the case if the convex hulls of B and $A \backslash B$ 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 $B = \{x \in A : \lambda_x > 0\}$ and its complement $A \setminus B$ are non-empty, 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

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

shows that the convex hulls of B and $A \setminus B$ have a non-void intersection. □

Let us introduce the function of two integers

$$\Phi_n^h = \sum_{k=0}^{h} \binom{n}{k},$$

which 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 4.2.2. *Whenever $\bigcup S$ is finite,*

$$|S| \leq \Phi\left(\left|\bigcup S\right|, h(S)\right).$$

THEOREM 4.2.3. *For any $h \leq n$,*

$$\Phi_n^h \leq \exp\left[nH\left(\tfrac{h}{n}\right)\right] \leq \exp\left[h\left(\log\left(\tfrac{n}{h}\right)+1\right)\right],$$

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 4.2.2. 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 ($\triangle$ denoting the symmetric difference of two sets)

$$S' = \{A \in S : A \triangle \{x\} \in S\},$$
$$S'' = \{A \in S : A \triangle \{x\} \notin S\}.$$

Clearly, $\sqcup$ denoting the disjoint union, $S = S' \sqcup S''$ and $S \cap X' = (S' \cap X') \sqcup (S'' \cap X')$. 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 $A \cap S = \{0,1\}^A$). Using the induction hypothesis, we then see that $|S \cap X'| \leq \Phi^{h(S)}_{|X'|} + \Phi^{h(S)-1}_{|X'|}$. But as $|X'| = |X| - 1$, the right-hand side of this inequality is equal to $\Phi^{h(S)}_{|X|}$, according to the recurrence equation satisfied by Φ.

Proof of theorem 4.2.3: This is the well-known Chernoff bound for the deviation of sums of Bernoulli random variables: let $(\sigma_1, \ldots, \sigma_n)$ be i.i.d. Bernoulli random variables with parameter $1/2$. Let us notice that

$$\Phi^h_n = 2^n \mathbb{P}\left(\sum_{i=1}^n \sigma_i \leq h \right).$$

For any positive real number λ ,

$$\begin{aligned} \mathbb{P}\Big(\sum_{i=1}^n \sigma_i \leq h \Big) &\leq \exp(\lambda h) \mathbb{E}\Big[\exp\Big(-\lambda \sum_{i=1}^n \sigma_i \Big) \Big] \\ &= \exp\Big\{ \lambda h + n \log\big\{ \mathbb{E}\big[\exp\big(-\lambda \sigma_1 \big) \big] \big\} \Big\}. \end{aligned}$$

Differentiating the right-hand side in λ shows that its minimal value is $\exp\big[-n \mathcal{K}(\frac{h}{n}, \frac{1}{2}) \big]$, where $\mathcal{K}(p,q) = p \log(\frac{p}{q}) + (1-p) \log(\frac{1-p}{1-q})$ is the Kullback divergence function between two Bernoulli distributions B_p and B_q of parameters p and q. Indeed the optimal value λ^* of λ is such that

$$h = n \frac{\mathbb{E}\big[\sigma_1 \exp(-\lambda^* \sigma_1) \big]}{\mathbb{E}\big[\exp(-\lambda^* \sigma_1) \big]} = n B_{h/n}(\sigma_1).$$

Therefore, using the fact that two Bernoulli distributions with the same expectations are equal,

$$\log\big\{ \mathbb{E}\big[\exp(-\lambda^* \sigma_1) \big] \big\} = -\lambda^* B_{h/n}(\sigma_1) - \mathcal{K}(B_{h/n}, B_{1/2}) = -\lambda^* \tfrac{h}{n} - \mathcal{K}(\tfrac{h}{n}, \tfrac{1}{2}).$$

The announced result then follows from the identity

$$\begin{aligned} H(p) &= \log(2) - \mathcal{K}(p, \tfrac{1}{2}) \\ &= p \log(p^{-1}) + (1-p) \log(1 + \frac{p}{1-p}) \leq p\big[\log(p^{-1}) + 1 \big]. \end{aligned}$$

4.2.3. Vapnik–Cervonenkis dimension of linear rules with margin. The proof of the following theorem was suggested to us by a similar proof presented in Cristianini et al. (2000).

Theorem 4.2.4. *Consider a family of points $(x_1, \ldots, x_n)$ in some Euclidean vector space E and a family of affine functions*

$$\mathcal{H} = \big\{ g_{w,b} : E \to \mathbb{R} \,;\, w \in E, \|w\| = 1, b \in \mathbb{R} \big\},$$

where

$$g_{w,b}(x) = \langle w, x\rangle - b, \qquad x \in E.$$

Assume that there is a set of thresholds $(b_i)_{i=1}^n \in \mathbb{R}^n$ *such that for any* $(y_i)_{i=1}^n \in \{-1,+1\}^n$*, there is* $g_{w,b} \in \mathcal{H}$ *such that*

$$\inf_{i=1}^n \big(g_{w,b}(x_i) - b_i\big) y_i \geq \gamma.$$

Let us also introduce the empirical variance of $(x_i)_{i=1}^n$*,*

$$\mathbb{V}\mathrm{ar}(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^n \left\| x_i - \frac{1}{n}\sum_{j=1}^n x_j \right\|^2.$$

In this case and with this notation,

$$\frac{\mathbb{V}\mathrm{ar}(x_1,\dots,x_n)}{\gamma^2} \geq \begin{cases} n-1 & \text{when } n \text{ is even,} \\ (n-1)\frac{n^2-1}{n^2} & \text{when } n \text{ is odd.} \end{cases} \tag{4.5}$$

Moreover, equality is reached when γ *is optimal,* $b_i = 0$*,* $i = 1,\dots,n$ *and* $(x_1,\dots,x_n)$ *is a regular simplex (i.e. when* 2γ *is the minimum distance between the convex hulls of any two subsets of* $\{x_1,\dots,x_n\}$ *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 σ be a uniformly distributed random variable with values in $\mathfrak{S}_n$, the set of permutations of the first n integers $\{1,\dots,n\}$. By assumption, for any value of σ, there is an affine function $g_{w,b} \in \mathcal{H}$ such that

$$\min_{i=1,\dots,n} \big[g_{w,b}(x_i) - b_i\big]\big[2\mathbb{1}(s_{\sigma(i)} > 0) - 1\big] \geq \gamma.$$

As a consequence

$$\begin{aligned}\left\langle \sum_{i=1}^n s_{\sigma(i)} x_i, w \right\rangle &= \sum_{i=1}^n s_{\sigma(i)}\big(\langle x_i, w\rangle - b - b_i\big) + \sum_{i=1}^n s_{\sigma(i)} b_i \\ &\geq \sum_{i=1}^n \gamma |s_{\sigma(i)}| + s_{\sigma(i)} b_i.\end{aligned}$$

Therefore, using the fact that the map $x \mapsto \Big(\max\big\{0, x\big\}\Big)^2$ is convex,

$$\begin{aligned}\mathbb{E}\left(\left\|\sum_{i=1}^n s_{\sigma(i)} x_i\right\|^2\right) &\geq \mathbb{E}\left[\left(\max\left\{0, \sum_{i=1}^n \gamma|s_{\sigma(i)}| + s_{\sigma(i)} b_i\right\}\right)^2\right] \\ &\geq \left(\max\left\{0, \sum_{i=1}^n \gamma \mathbb{E}\big(|s_{\sigma(i)}|\big) + \mathbb{E}\big(s_{\sigma(i)}\big) b_i\right\}\right)^2 = \gamma^2 \left(\sum_{i=1}^n |s_i|\right)^2,\end{aligned}$$

where $\mathbb{E}$ is the expectation with respect to the random permutation σ. On the other hand

$$\mathbb{E}\left(\left\|\sum_{i=1}^n s_{\sigma(i)} x_i\right\|^2\right) = \sum_{i=1}^n \mathbb{E}(s_{\sigma(i)}^2)\|x_i\|^2 + \sum_{i\neq j} \mathbb{E}(s_{\sigma(i)} s_{\sigma(j)})\langle x_i, x_j\rangle.$$

Moreover

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

In the same way, for any $i \neq j$,

$$\begin{aligned}
\mathbb{E}\left(s_{\sigma(i)}s_{\sigma(j)}\right) &= \frac{1}{n(n-1)}\mathbb{E}\left(\sum_{i\neq j} s_{\sigma(i)}s_{\sigma(j)}\right) \\
&= \frac{1}{n(n-1)}\sum_{i\neq j} s_i s_j \\
&= \frac{1}{n(n-1)}\left[\underbrace{\left(\sum_{i=1}^n s_i\right)^2}_{=0} - \sum_{i=1}^n s_i^2\right] \\
&= -\frac{1}{n(n-1)}\sum_{i=1}^n s_i^2.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}\left(\left\|\sum_{i=1}^n s_{\sigma(i)}x_i\right\|^2\right) &= \left(\sum_{i=1}^n s_i^2\right)\left[\frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - \frac{1}{n(n-1)}\sum_{i\neq j}\langle x_i, x_j\rangle\right] \\
&= \left(\sum_{i=1}^n s_i^2\right)\left[\left(\frac{1}{n} + \frac{1}{n(n-1)}\right)\sum_{i=1}^n \|x_i\|^2\right. \\
&\quad \left. - \frac{1}{n(n-1)}\left\|\sum_{i=1}^n x_i\right\|^2\right] \\
&= \frac{n}{n-1}\left(\sum_{i=1}^n s_i^2\right)\mathbb{V}\mathrm{ar}(x_1, \ldots, x_n).
\end{aligned}$$

We have proved that

$$\frac{\mathbb{V}\mathrm{ar}(x_1, \ldots, x_n)}{\gamma^2} \geq \frac{(n-1)\left(\sum_{i=1}^n |s_i|\right)^2}{n\sum_{i=1}^n s_i^2}.$$

This can be used with $s_i = \mathbb{1}(i \leq \frac{n}{2}) - \mathbb{1}(i > \frac{n}{2})$ in the case when n is even and $s_i = \frac{2}{(n-1)}\mathbb{1}(i \leq \frac{n-1}{2}) - \frac{2}{n+1}\mathbb{1}(i > \frac{n-1}{2})$ in the case when n is odd, to establish the first inequality (4.5) of the theorem.

Checking that equality is reached for the simplex is an easy computation when the simplex $(x_i)_{i=1}^n \in (\mathbb{R}^n)^n$ 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). □

4.2.4. Application to Support Vector Machines. We are going to apply Theorem 4.2.4 (page 145) to Support Vector Machines in the transductive case. Let $(X_i, Y_i)_{i=1}^{(k+1)N}$ be distributed according to some partially exchangeable distribution $\mathbb{P}$ and assume that $(X_i)_{i=1}^{(k+1)N}$ 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 $Z' = (X_i, y_i')_{i=1}^{(k+1)N}$, where $(y_i')_{i=1}^{(k+1)N} \in \mathcal{Y}^{(k+1)N}$, let $\hat{f}_{Z'}$ be the Support Vector Machine defined by K and Z' and let $\gamma(Z')$ be its margin. Let

$$R^2 = \max_{i=1,\dots,(k+1)N} K(X_i, X_i) + \frac{1}{(k+1)^2 N^2} \sum_{j=1}^{(k+1)N} \sum_{k=1}^{(k+1)N} K(X_j, X_k) - \frac{2}{(k+1)N} \sum_{j=1}^{(k+1)N} K(X_i, X_j).$$

This is an easily computable upper-bound for the radius of some ball containing the image of $(X_1, \dots, X_{(k+1)N})$ in feature space.

Let us define for any integer h the margins

$$\gamma_{2h} = (2h-1)^{-1/2} \quad \text{and} \quad \gamma_{2h+1} = \left[2h\left(1 - \frac{1}{(2h+1)^2}\right)\right]^{-1/2}. \tag{4.6}$$

Let us consider for any $h = 1, \dots, N$ the exchangeable model

$$\mathcal{R}_h = \left\{\hat{f}_{Z'} : Z' = (X_i, y_i')_{i=1}^{(k+1)N} \text{ is } K\text{-separable and } \gamma(Z') \geq R\gamma_h\right\}.$$

The family of models $\mathcal{R}_h$, $h = 1, \dots, N$ is nested, and we know from Theorem 4.2.4 (page 145) and Theorems 4.2.2 (page 144) and 4.2.3 (page 144) that

$$\log\big(|\mathcal{R}_h|\big) \leq h \log\big(\tfrac{(k+1)eN}{h}\big).$$

We can then consider on the large model $\mathcal{R} = \bigsqcup_{h=1}^{N} \mathcal{R}_h$ (the disjoint union of the sub-models) an exchangeable prior π which is uniform on each $\mathcal{R}_h$ and is such that $\pi(\mathcal{R}_h) \geq \frac{1}{h(h+1)}$. Applying Theorem 3.2.3 (page 116) we get

Proposition 4.2.5. *With $\mathbb{P}$ probability at least $1-\epsilon$, for any $h = 1, \dots, N$, any Support Vector Machine $f \in \mathcal{R}_h$,*

$$r_2(f) \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{1 - \exp\left[-\frac{\lambda}{N} r_1(f) - \frac{h}{N}\log\left(\frac{e(k+1)N}{h}\right) - \frac{\log[h(h+1)] - \log(\epsilon)}{N}\right]}{1 - \exp(-\frac{\lambda}{N})} - \frac{r_1(f)}{k}.$$

Searching the whole model $\mathcal{R}_h$ to optimize the bound may require more computer resources than are available, but any heuristic can be applied to choose f, since the bound is uniform. For instance, a Support Vector Machine f' using a box constraint can be trained from the training set $(X_i, Y_i)_{i=1}^{N}$ and then $(y_i')_{i=1}^{(k+1)N}$ can be set to $y_i' = \operatorname{sign}(f'(X_i))$, $i = 1, \dots, (k+1)N$.

4.2.5. Inductive margin bounds for Support Vector Machines. In order to establish inductive margin bounds, we will need a different combinatorial lemma. It is due to Alon et al. (1997). We will reproduce their proof with some tiny improvements on the values of constants.

Let us consider the finite case when $\mathcal{X} = \{1, \dots, n\}$, $\mathcal{Y} = \{1, \dots, b\}$ and $b \geq 3$. The question we will study would be meaningless when $b \leq 2$. Assume as usual that we are dealing with a prescribed set of classification rules $\mathcal{R} = \big\{f : \mathcal{X} \to \mathcal{Y}\big\}$. Let us say that a pair (A, s), where $A \subset \mathcal{X}$ is a non-empty set of shapes and $s : A \to \{2, \dots, b-1\}$ a threshold function, is *shattered* by the set of functions $F \subset \mathcal{R}$ if for any $(\sigma_x)_{x \in A} \in \{-1, +1\}^A$, there exists some $f \in F$ such that $\min_{x \in A} \sigma_x \big[f(x) - s(x)\big] \geq 1$.

Definition 4.2.1. Let the *fat shattering dimension* of $(\mathcal{X}, \mathcal{R})$ be the maximal size $|A|$ of the first component of the pairs which are shattered by $\mathcal{R}$.

Let us say that a subset of classification rules $F \subset \mathcal{Y}^{\mathcal{X}}$ is *separated* whenever for any pair $(f, g) \in F^2$ such that $f \neq g$, $\|f - g\|_\infty = \max_{x \in \mathcal{X}} |f(x) - g(x)| \geq 2$. Let $\mathfrak{M}(\mathcal{R})$ be the maximum size $|F|$ of separated subsets F of $\mathcal{R}$. Note that if F is a separated subset of $\mathcal{R}$ such that $|F| = \mathfrak{M}(\mathcal{R})$, then it is a 1-net for the $\mathcal{L}_\infty$ distance: for any function $f \in \mathcal{R}$ there exists $g \in F$ such that $\|f - g\|_\infty \leq 1$ (otherwise f could be added to F to create a larger separated set).

Lemma 4.2.6. *With the above notation, whenever the fat shattering dimension of $(\mathcal{X}, \mathcal{R})$ is not greater than h,*

$$\log\big[\mathfrak{M}(\mathcal{R})\big] < \log\big[(b-1)(b-2)n\big] \left\{ \frac{\log\big[\sum_{i=1}^{h} \binom{n}{i} (b-2)^i\big]}{\log(2)} + 1 \right\} + \log(2)$$
$$\leq \log\big[(b-1)(b-2)n\big] \left\{ \left[\log\Big[\tfrac{(b-2)n}{h}\Big] + 1 \right] \frac{h}{\log(2)} + 1 \right\} + \log(2).$$

Proof. For any set of functions $F \subset \mathcal{Y}^{\mathcal{X}}$, let $t(F)$ be the number of pairs (A, s) shattered by F. Let $t(m, n)$ be the minimum of $t(F)$ over all *separated* sets of functions $F \subset \mathcal{Y}^{\mathcal{X}}$ of size $|F| = m$ (n is here to recall that the shape space $\mathcal{X}$ is made of n shapes). For any m such that $t(m, n) > \sum_{i=1}^{h} \binom{n}{i} (b-2)^i$, it is clear that any separated set of functions of size $|F| \geq m$ shatters at least one pair (A, s) such that $|A| > h$. Indeed, from its definition $t(m, n)$ is clearly a non-decreasing function of m, so that $t(|F|, n) > \sum_{i=1}^{h} \binom{n}{i} (b-2)^i$. Moreover there are only $\sum_{i=1}^{h} \binom{n}{i} (b-2)^i$ pairs (A, s) such that $|A| \leq h$. As a consequence, whenever the fat shattering dimension of $(\mathcal{X}, \mathcal{R})$ is not greater than h we have $\mathfrak{M}(\mathcal{R}) < m$.

It is clear that for any $n \geq 1$, $t(2, n) = 1$.

Lemma 4.2.7. *For any $m \geq 1$, $t\big[mn(b-1)(b-2), n\big] \geq 2t\big[m, n-1\big]$, and therefore $t\big[2n(n-1)\cdots(n-r+1)(b-1)^r(b-2)^r, n\big] \geq 2^r$.*

Proof. Let $F = \{f_1, \dots, f_{mn(b-1)(b-2)}\}$ be some separated set of functions of size $mn(b-1)(b-2)$. For any pair (f_{2i-1}, f_{2i}), $i = 1, \dots, mn(b-1)(b-2)/2$, there is $x_i \in \mathcal{X}$ such that $|f_{2i-1}(x_i) - f_{2i}(x_i)| \geq 2$. Since $|\mathcal{X}| = n$, there is $x \in \mathcal{X}$ such that $\sum_{i=1}^{mn(b-1)(b-2)/2} \mathbb{1}(x_i = x) \geq m(b-1)(b-2)/2$. Let $I = \{i : x_i = x\}$. Since there are $(b-1)(b-2)/2$ pairs $(y_1, y_2) \in \mathcal{Y}^2$ such that $1 \leq y_1 < y_2 - 1 \leq b-1$, there

is some pair (y_1, y_2), such that $1 \leq y_1 < y_2 \leq b$ and such that $\sum_{i \in I} \mathbb{1}(\{y_1, y_2\} = \{f_{2i-1}(x), f_{2i}(x)\}) \geq m$. Let $J = \big\{i \in I : \{f_{2i-1}(x), f_{2i}(x)\} = \{y_1, y_2\}\big\}$. Let

$$F_1 = \{f_{2i-1} : i \in J, f_{2i-1}(x) = y_1\} \cup \{f_{2i} : i \in J, f_{2i}(x) = y_1\},$$
$$F_2 = \{f_{2i-1} : i \in J, f_{2i-1}(x) = y_2\} \cup \{f_{2i} : i \in J, f_{2i}(x) = y_2\}.$$

Obviously $|F_1| = |F_2| = |J| = m$. Moreover the restrictions of the functions of F_1 to $\mathcal{X} \setminus \{x\}$ are separated, and it is the same with F_2. Thus F_1 strongly shatters at least $t(m, n-1)$ pairs (A, s) such that $A \subset \mathcal{X} \setminus \{x\}$ and it is the same with F_2. Finally, if the pair (A, s) where $A \subset \mathcal{X} \setminus \{x\}$ is both shattered by F_1 and F_2, then $F_1 \cup F_2$ shatters also $(A \cup \{x\}, s')$ where $s'(x') = s(x')$ for any $x' \in A$ and $s'(x) = \lfloor \frac{y_1+y_2}{2} \rfloor$. Thus $F_1 \cup F_2$, and therefore F, shatters at least $2t(m, n-1)$ pairs (A, s). □

Resuming the proof of lemma 4.2.6, let us choose for r the smallest integer such that $2^r > \sum_{i=1}^{h} \binom{n}{i}(b-2)^i$, which is no greater than

$$\left\{ \frac{\log\left[\sum_{i=1}^{h} \binom{n}{i}(b-2)^i\right]}{\log(2)} + 1 \right\}.$$

In the case when $1 \leq n \leq r$,

$$\log(\mathfrak{M}(\mathcal{R})) < |\mathcal{X}| \log(|\mathcal{Y}|) = n \log(b) \leq r \log(b) \leq r \log\big[(b-1)(b-2)n\big] + \log(2),$$

which proves the lemma. In the remaining case $n > r$,

$$\begin{aligned} t\big[2n^r(b-1)^r(b-2)^r, n\big] &\geq t\big[2n(n-1)\dots(n-r+1)(b-1)^r(b-2)^r, n\big] \\ &> \sum_{i=1}^{h} \binom{n}{i}(b-2)^i. \end{aligned}$$

Thus $|\mathfrak{M}(\mathcal{R})| < 2\Big[(b-2)(b-1)n\Big]^r$ as claimed. □

In order to apply this combinatorial lemma to Support Vector Machines, let us consider now the case of separating hyperplanes in $\mathbb{R}^d$ (the generalization to Support Vector Machines being straightforward). Assume that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. For any sample $(X)_{i=1}^{(k+1)N}$, let

$$R(X_1^{(k+1)N}) = \max\{\|X_i\| : 1 \leq i \leq (k+1)N\}.$$

Let us consider the set of parameters

$$\Theta = \big\{(w, b) \in \mathbb{R}^d \times \mathbb{R} : \|w\| = 1\big\}.$$

For any $(w, b) \in \Theta$, let $g_{w,b}(x) = \langle w, x \rangle - b$. Let h be some fixed integer and let $\gamma = R(X_1^{(k+1)N})\gamma_h$, where γ_h is defined by equation (4.6, page 148).

Let us define $\zeta : \mathbb{R} \to \mathbb{Z}$ by

$$\zeta(r) = \begin{cases} -5 & \text{when} \quad r \leq -4\gamma, \\ -3 & \text{when} \quad -4\gamma < r \leq -2\gamma, \\ -1 & \text{when} \quad -2\gamma < r \leq 0, \\ +1 & \text{when} \quad 0 < r \leq 2\gamma, \\ +3 & \text{when} \quad 2\gamma < r \leq 4\gamma, \\ +5 & \text{when} \quad 4\gamma < r. \end{cases}$$

Let $G_{w,b}(x) = \zeta\big[g_{w,b}(x)\big]$. The fat shattering dimension (as defined in 4.2.1) of

$$\Big(X_1^{(k+1)N}, \big\{(G_{w,b}+7)/2 : (w,b) \in \Theta\big\}\Big)$$

is not greater than h (according to Theorem 4.2.4, page 145), therefore there is some set $\mathcal{F}$ of functions from $X_1^{(k+1)N}$ to $\{-5,-3,-1,+1,+3,+5\}$ such that

$$\log\big(|\mathcal{F}|\big) \leq \log\big[20(k+1)N\big]\left\{\frac{h}{\log(2)}\left[\log\left(\frac{4(k+1)N}{h}\right)+1\right]+1\right\}+\log(2).$$

and for any $(w,b) \in \Theta$, there is $f_{w,b} \in \mathcal{F}$ such that $\sup\big\{|f_{w,b}(X_i) - G_{w,b}(X_i)| : i = 1,\ldots,(k+1)N\big\} \leq 2$. Moreover, the choice of $f_{w,b}$ may be required to depend on $(X_i)_{i=1}^{(k+1)N}$ in an exchangeable way. Similarly to Theorem 3.2.3 (page 116), it can be proved that for any partially exchangeable probability distribution $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $f_{w,b} \in \mathcal{F}$,

$$\begin{aligned}
&\frac{1}{kN}\sum_{i=N+1}^{(k+1)N} \mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big] \\
&\qquad \leq \frac{k+1}{k}\inf_{\lambda\in\mathbb{R}_+}\big[1-\exp(-\tfrac{\lambda}{N})\big]^{-1}\Bigg\{1- \\
&\qquad\qquad \exp\Bigg[-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big] - \frac{\log\big(|\mathcal{F}|\big)-\log(\epsilon)}{N}\Bigg]\Bigg\} \\
&\qquad\qquad\qquad - \frac{1}{kN}\sum_{i=1}^{N}\mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big].
\end{aligned}$$

Let us remark that

$$\mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b}(X_i)\geq 0\big]-1 \neq Y_i\Big\} = \mathbb{1}\big[G_{w,b}(X_i)Y_i < 0\big] \leq \mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big]$$

and

$$\mathbb{1}\big[f_{w,b}(X_i)Y_i \leq 1\big] \leq \mathbb{1}\big[G_{w,b}(X_i)Y_i \leq 3\big] \leq \mathbb{1}\big[g_{w,b}(X_i)Y_i \leq 4\gamma\big].$$

This proves the following theorem.

THEOREM 4.2.8. *Let us consider the sequence $(\gamma_h)_{h\in\mathbb{N}^*}$ defined by equation (4.6, page 148). With $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b) \in \Theta$,*

$$\begin{aligned}
&\frac{1}{kN}\sum_{i=N+1}^{(k+1)N}\mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b}(X_i)\geq 0\big]-1 \neq Y_i\Big\} \\
&\qquad \leq \frac{k+1}{k}\inf_{\lambda\in\mathbb{R}_+,h\in\mathbb{N}^*}\big[1-\exp(-\tfrac{\lambda}{N})\big]^{-1}\Bigg\{1- \\
&\qquad\quad \exp\Bigg[-\frac{\lambda}{N^2}\sum_{i=1}^{N}\mathbb{1}\big[g_{w,b}(X_i)Y_i \leq 4R\gamma_h\big] \\
&\qquad\qquad -\frac{\log\big[20(k+1)N\big]\Big\{\frac{h}{\log(2)}\log\Big(\frac{4e(k+1)N}{h}\Big)+1\Big\}+\log\Big[\frac{2h(h+1)}{\epsilon}\Big]}{N}\Bigg]\Bigg\}
\end{aligned}$$

$$- \frac{1}{kN} \sum_{i=1}^{N} \mathbb{1}\big[g_{w,b}(X_i)Y_i \leq 4R\gamma_h\big].$$

Properly speaking this theorem is not a margin bound, but more precisely a *margin quantile* bound, since it covers the case where some fraction of the training sample falls within the region defined by the margin parameter γ_h which optimizes the bound.

As a consequence though, we get a true (weaker) margin bound: with $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b) \in \Theta$ such that

$$\gamma = \min_{i=1,\ldots,N} g_{w,b}(X_i)Y_i > 0,$$

$$\frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \mathbb{1}\big[g_{w,b}(X_i)Y_i < 0\big] \leq \frac{k+1}{k}\left\{1 - \exp\left[-\frac{\log\big[20(k+1)N\big]}{N}\left\{\frac{16R^2+2\gamma^2}{\log(2)\gamma^2}\log\left(\frac{e(k+1)N\gamma^2}{4R^2}\right)+1\right\} + \frac{1}{N}\log(\tfrac{\epsilon}{2})\right]\right\}.$$

This inequality compares favourably with similar inequalities in Cristianini et al. (2000), which moreover do not extend to the margin quantile case as this one.

Let us also mention that it is easy to circumvent the fact that R is not observed when the test set $X_{N+1}^{(k+1)N}$ is not observed.

Indeed, we can consider the sample obtained by projecting $X_1^{(k+1)N}$ on some ball of fixed radius $R_{\max}$, putting

$$t_{R_{\max}}(X_i) = \min\left\{1, \frac{R_{\max}}{\|X_i\|}\right\} X_i.$$

We can further consider an atomic prior distribution $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$ bearing on $R_{\max}$, to obtain a uniform result through a union bound. As a consequence of the previous theorem, we have

Corollary 4.2.9. *For any atomic prior $\nu \in \mathcal{M}_+^1(\mathbb{R}_+)$, for any partially exchangeable probability measure $\mathbb{P} \in \mathcal{M}_+^1(\Omega)$, with $\mathbb{P}$ probability at least $1-\epsilon$, for any $(w,b) \in \Theta$, any $R_{\max} \in \mathbb{R}_+$,*

$$\frac{1}{kN} \sum_{i=N+1}^{(k+1)N} \mathbb{1}\Big\{2\mathbb{1}\big[g_{w,b} \circ t_{R_{\max}}(X_i) \geq 0\big] - 1 \neq Y_i\Big\} \leq \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+, h \in \mathbb{N}^*} \big[1 - \exp(-\tfrac{\lambda}{N})\big]^{-1} \left\{1 - \exp\left[-\frac{\lambda}{N^2}\sum_{i=1}^{N} \mathbb{1}\big[g_{w,b} \circ t_{R_{\max}}(X_i)Y_i \leq 4R_{\max}\gamma_h\big] - \frac{\log\big[20(k+1)N\big]\left\{\frac{h}{\log(2)}\log\left(\frac{4e(k+1)N}{h}\right)+1\right\} + \log\left[\frac{2h(h+1)}{\epsilon\nu(R_{\max})}\right]}{N}\right]\right\}$$

$$- \frac{1}{kN} \sum_{i=1}^{N} \mathbb{1}\big[g_{w,b} \circ t_{R_{\max}}(X_i) Y_i \leq 4 R_{\max} \gamma_h\big].$$

Let us remark that $t_{R_{\max}}(X_i) = X_i$, $i = N+1, \ldots, (k+1)N$, as soon as we consider only the values of $R_{\max}$ not smaller than $\max_{i=N+1,\ldots,(k+1)N} \|X_i\|$ in this corollary. Thus we obtain a bound on the transductive generalization error of the unthresholded classification rule $2\mathbb{1}\big[g_{w,b}(X_i) \geq 0\big] - 1$, as well as some incitation to replace it with a thresholded rule when the value of $R_{\max}$ minimizing the bound falls below $\max_{i=N+1,\ldots,(k+1)N} \|X_i\|$.