Universités de Paris 6 & Paris 7 - CNRS (UMR 7599)

PRÉPUBLICATIONS DU LABORATOIRE

DE PROBABILITÉS & MODÈLES ALÉATOIRES

4, place Jussieu - Case 188 - 75 252 Paris cedex 05

http://www.proba.jussieu.fr

Classification under polynomial entropy and margin assumptions and randomized estimators

J.-Y. AUDIBERT

AVRIL 2004

Prépublication n^o 908

Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris VI & Université Paris VII,
4, place Jussieu, Case 188, F-75252 Paris Cedex 05.

&

CREST, ENSAE, Laboratoire de Finance et Assurance,
15 Bd Gabriel Péri, F-92245 Malakoff Cedex.

CLASSIFICATION UNDER POLYNOMIAL ENTROPY AND MARGIN ASSUMPTIONS AND RANDOMIZED ESTIMATORS

J.-Y. AUDIBERT

Université Paris VI and CREST

Abstract. The aim of this paper is two-fold. First we want to develop the PAC-Bayesian point of view [13, 3, 4, 1] and show how the efficiency of a Gibbs estimator relies on the weights given by the prior distribution to the balls centered at the best function in the model and associated with the pseudo-distance $(f_1, f_2) \mapsto \mathbb{P}[f_1(X) \neq f_2(X)]$.

Secondly, we show how to recover and improve results under empirical and non empirical polynomial entropy assumptions and Tsybakov's margin assumption. We also study the links between empirical and non empirical nets and give an observable version of the integral entropy [6, 9, 14].

Contents

2000 *Mathematics Subject Classification.* Primary 62H30, Secondary 68Q32.

Key words and phrases. Gibbs classifiers, entropy assumptions, margin assumptions, PAC-Bayesian bounds, chaining, oracle inequalities, VC theory.

I would like to thank very deeply Professor Olivier Catoni, my PhD advisor, for the numerous hours he spent with me. I am also very grateful to Olivier Bousquet and Professor Alexandre Tsybakov for the numerous discussions which have fed this work. I would also like to thank Professor Bernhard Schölkopf who invited me twice at the Max Planck Institute of Tuebingen. Meeting people there has been a key point for widening my scope of interest during my PhD.

1. Setup and notations

We assume that we observe an i.i.d. sample $Z_1^N \triangleq (X_i, Y_i)_{i=1}^N$ of random variables distributed according to a product probability measure $\mathbb{P}^{\otimes N}$, where $\mathbb{P}$ is a probability distribution on $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}}) \triangleq (\mathcal{X} \otimes \mathcal{Y}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Y}})$, $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ is a measurable space called the pattern space, $\mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$ is the (finite) label space and $\mathcal{B}_{\mathcal{Y}}$ is the sigma algebra of all subsets of $\mathcal{Y}$. Let $\mathbb{P}(dY|X)$ denote a regular version of the conditional probabilities (which we will use in the following without further mention).

Let $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ denote the set of all measurable functions mapping $\mathcal{X}$ into $\mathcal{Y}$. The aim of a classification procedure is to build a function $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ from the learning sample such that $f(X)$ predicts the label Y associated with X. The quality of the

prediction is measured by the expected risk

$$R(f) \triangleq \mathbb{P}[Y \neq f(X)].$$

A function $f^*_{\mathbb{P}}$ such that for any $x \in \mathcal{X}$,

$$f^*_{\mathbb{P}}(x) \in \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbb{P}(Y = y | X = x)$$

minimizes the expected risk. This function is not necessarily unique. We assume that there exists a measurable one. We will once for all fix it, refer to it as the Bayes classifier and often denote it f^* to shorten. Since we have no prior information about the distribution $\mathbb{P}$ of (X, Y), this classifier is unknown.

Since there is generally no measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$ such that

$$\lim_{N \to +\infty} \sup_{\mathbb{P} \in \mathcal{M}^1_+(\mathcal{Z})} \left\{ \mathbb{P}^{\otimes (N+1)} \left[Y_{N+1} \neq \hat{f}(Z_1^N)(X_{N+1}) \right] - \inf_{f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} \mathbb{P}[Y \neq f(X)] \right\} = 0,$$

we have to work with a prescribed set of classification functions $\mathcal{F}$, called the model. This set is just some subset of the set of all measurable functions $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. Let us denote $\tilde{f}$ the best function in the model, i.e. a function minimizing the expected risk:

$$\tilde{f} \in \operatorname*{argmin}_{\mathcal{F}} R.$$

For sake of simplicity, we assume that it exists[1]. Let

$$\bar{\mathbb{P}} \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{(X_i, Y_i)}$$

be the empirical distribution. The empirical risk

$$r(f) \triangleq \bar{\mathbb{P}}[Y \neq f(X)]$$

gives an estimate of the expected risk : from the law of large numbers, for any measurable function, it tends to the expected risk almost surely. An estimator which minimizes the empirical risk

$$\hat{f}_{\mathrm{ERM}} \in \operatorname*{argmin}_{\mathcal{F}} r$$

is called an ERM[2]-classifier. The regression function will be denoted

$$\eta^*(x) \triangleq \mathbb{P}(Y | X = x).$$

In the binary classification setting ($\mathcal{Y} = \{0; 1\}$), we have $\eta^*(x) = \mathbb{P}(Y = 1 | X = x)$.

Since we will study randomized estimators, we assume that we have a $\sigma-$algebra $\mathcal{T}$ such that $(\mathcal{F}, \mathcal{T})$ is a measurable space containing the sets $\{f\}$ for any $f \in \mathcal{F}$ and such that the function

$$\begin{array}{ccc} \mathcal{F} \times \mathcal{X} & \to & \mathcal{Y} \\ (f, x) & \mapsto & f(x) \end{array}$$

is measurable. A randomized estimator consists in drawing a function in $\mathcal{F}$ according to some random distribution $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\mathcal{F})$, where $\mathcal{M}^1_+(\mathcal{F})$ is the set of probability distributions on the measurable space $(\mathcal{F}, \mathcal{T})$.

[1]Otherwise we would have to introduce some small positive real β and consider $\tilde{f}$ as an estimator minimizing the expected risk up to β. This real β would then appear in all the equations related to this function and make things needlessly messy.

[2]ERM = Empirical Risk Minimization

To shorten, we will use μh to denote the expectation of the random variable h under the probability distribution μ: $\mu h \triangleq \int h(x)d\mu(x)$. The Kullback-Leibler divergence between two probability distributions is defined as $K(\mu,\nu) = \mu \log \frac{d\mu}{d\nu}$ when μ is absolutely continuous with respect to ν and $K(\mu,\nu) = +\infty$ otherwise.

The symbol C will denote a positive universal constant whose value may differ from line to line whereas the symbol $\breve{C}$ will denote a positive constant whose value depends on other constants and may also differ from line to line.

We define

$$\pi_h \triangleq \frac{\exp(h)}{\pi \exp(h)} \cdot \pi$$

for any measurable real function h such that $\exp(h)$ is π-integrable. The randomized estimators associated with the posterior distributions π_{-Cr} will be called the standard Gibbs estimators with temperature $\frac{1}{C}$.

1.1. **Measurability.** Finally, to circumvent some measurability problems, we will consider inner and outer expectations. Let $(A, \mathcal{A}, \mu)$ be a measure space and $\mathcal{C}(A;\mathbb{R})$ be the class of real measurable functions. For any (measurable or not) function f, its inner and outer expectation wrt μ are respectively $\mu_*(h) \triangleq \sup\big\{\mu(g) : g \in \mathcal{C}(A;\mathbb{R}), g \le h\big\}$ and $\mu^*(h) \triangleq \inf\big\{\mu(g) : g \in \mathcal{C}(A;\mathbb{R}), g \ge h\big\}$. Naturally, for any set $B \subset A$, $\mu_*(B)$ and $\mu^*(B)$ are defined by $\mu_*(B) = \mu_*(\mathbb{1}_B)$ and $\mu^*(B) = \mu^*(\mathbb{1}_B)$. Note that μ_* and μ^* are not measures but satisfy $\mu^*(B) + \mu_*(B^c) = 1$ and $\mu^*(B_1 \cup B_2) \le \mu^*(B_1) + \mu^*(B_2)$. Besides, if $\mu^*(h) < +\infty$, then there exists a random variable h^* such that $\mu^*(h) = \mu(h^*)$. For more details on properties of inner and outer expectations, see [17].

1.2. **Covering, packing and bracketing nets and entropies.** Let $\mathbb{Q}$ denote a probability distribution on the measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. The mapping $\mathbb{Q}_{\cdot,\cdot}$ from $\mathcal{F} \times \mathcal{F}$ into $\mathbb{R}_+$ defined as

$$\mathbb{Q}_{f_1,f_2} \triangleq \mathbb{Q}[f_1(X) \neq f_2(X)] \qquad \text{for any } f_1, f_2 \in \mathcal{F}$$

is a pseudo-distance. For any $u \ge 0$, a set of measurable functions $\mathcal{G} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$ such that

$$\sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{G}} \mathbb{Q}_{f,g} \le u$$

is called a $u-$covering net of the set $\mathcal{F}$ wrt the pseudo-distance $\mathbb{Q}$.

The log-cardinal $H(u,\mathcal{F},\mathbb{Q}_{\cdot,\cdot})$ of the smallest $u-$covering net (possibly infinite) is called the $u-$covering entropy. A $u-$covering net with log-cardinal equal to $H(u,\mathcal{F},\mathbb{Q}_{\cdot,\cdot})$ is called a minimal $u-$covering net[3].

In bracketing nets, we require in addition that any function in $\mathcal{F}$ can be encapsulated by two functions of the net. Specifically, for any $u \ge 0$, a set of measurable functions $\mathcal{G} \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$ such that for any function $f \in \mathcal{F}$, there exist $f_L, f_U \in \mathcal{G}$ satisfying $f_L \le f \le f_U$ and $\mathbb{Q}_{f_L,f_U} \le u$, is called a $u-$bracketing net of the set $\mathcal{F}$ wrt the pseudo-distance $\mathbb{Q}$. The log-cardinal $H^{[]}(u,\mathcal{F},\mathbb{Q}_{\cdot,\cdot})$ of the smallest $u-$bracketing net (possibly infinite) is called the $u-$bracketing entropy. A $u-$bracketing net with log-cardinal equal to $H^{[]}(u,\mathcal{F},\mathbb{Q}_{\cdot,\cdot})$ is called a minimal $u-$bracketing net.

[3]Here the functions in the net can be taken outside $\mathcal{F}$. This is not so important since it is well-known that a $2u-$covering net with functions in $\mathcal{F}$ can be constructed from any $u-$covering net.

Packing nets are covering nets such that for any functions f_1, f_2 in the net, we have $\mathbb{Q}_{f_1,f_2} > u$. The packing entropy $H_{\mathrm{p}}(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$ is the log-cardinal of a minimal packing net.

We have $H(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot}) \leq H_{\mathrm{p}}(u, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot}) \leq H(\frac{u}{2}, \mathcal{F}, \mathbb{Q}_{\cdot,\cdot})$. Any $u-$bracketing net is a $u-$covering net. The converse is false since it is easy to find a set $\mathcal{F}$ with finite $u-$covering entropy and infinite $u-$bracketing entropy.

Finally, we will say that a family of u_N-nets, $N \in \mathbb{N}$, is almost minimal when the log-cardinal of the size of the u_N-net has the same order as the $(u_N, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-entropy.

The paper is organized as follows. Section 2 recalls some PAC-Bayesian concentration inequalities which are extracted from [1]. In Section 3, we assume that we have Tsybakov's margin assumption and that the $\mathbb{P}_{\cdot,\cdot}$-entropies are polynomial. In this setting, we study the convergence rate of standard Gibbs estimators and classifiers minimizing the empirical risk on $\mathbb{P}_{\cdot,\cdot}$-covering nets. In particular, we stresses on the influence of the chaining trick and the differences between bracketing and covering entropy assumptions. Section 4 tries to answer the questions: what happens when we relieve the polynomial $\mathbb{P}_{\cdot,\cdot}$-entropy assumption? Can we give an empirical equivalent (i.e. with $\overline{\mathbb{P}}_{\cdot,\cdot}$-entropies) of the previous results? Section 5 gives a version of Assouad's lemma dedicated to classification. The proofs are gathered in Section 6.

2. Known PAC-Bayesian bounds

In this section, we recall some results of [1] which will be useful in this paper.

Theorem 2.1. *Let $g(u) \triangleq \frac{\exp(u)-1-u}{u^2}$ for any $u > 0$. For any $\lambda > 0$, $\epsilon > 0$ and $\pi_1, \pi_2 \in \mathcal{M}_+^1(\mathcal{F})$, with $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-\epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$, we have*

(2.1)
$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)(\rho_1 \otimes \rho_2)\mathbb{P}_{\cdot,\cdot} + \tfrac{K(\rho_1,\pi_1)+K(\rho_2,\pi_2)+\log(\epsilon^{-1})}{\lambda}$$

As a consequence, with $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-\epsilon$, for any $\rho_1, \rho_2 \in \mathcal{M}_+^1(\mathcal{F})$,

(2.2)
$$\rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r \leq \min_{\lambda \in [\sqrt{N}; N]} \Big\{ 0.8 \tfrac{\lambda}{N}(\rho_1 \otimes \rho_2)\mathbb{P}_{\cdot,\cdot} + 1.7 \tfrac{K(\rho_1,\pi_1)+K(\rho_2,\pi_2)+\log[\log(eN)\epsilon^{-1}]}{\lambda} \Big\}.$$

Besides, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be finite subsets of $\mathcal{F}$, with $\left(\mathbb{P}^{\otimes N}\right)_*$-probability at least $1-\epsilon$, for any $(f_1, f_2) \in \mathcal{S}_1 \times \mathcal{S}_2$, we have

$$R(f_2) - R(f_1) + r(f_1) - r(f_2) \leq \sqrt{\tfrac{2\log(|\mathcal{S}_1||\mathcal{S}_2|\epsilon^{-1})\mathbb{P}_{f_1,f_2}}{N}} + \tfrac{\log(|\mathcal{S}_1||\mathcal{S}_2|\epsilon^{-1})}{3N}. \tag{2.3}$$

Proof. The first part comes from Theorem 4.8 in [1]. Then the second part is obtained by a union bound on the set of parameters $\Lambda \triangleq \left\{\sqrt{N}e^{k/2}; 0 \leq k \leq \log N\right\}$ (see Section 4.2 in [1] for details). The third part comes from Theorem 8.1 in [1] applied to $\mathcal{W}\big[(f_1, f_2), Z\big] = \mathbb{1}_{Y \neq f_2(X)} - \mathbb{1}_{Y \neq f_1(X)}$ and ν equal to the uniform measure on $\mathcal{S}_1 \times \mathcal{S}_2$. □

The following theorem ([1, Theorem 6.4]) brackets the efficiency of a standard Gibbs classifier

Theorem 2.2. *For any $\lambda > 0$ and $0 < \chi \leq 1$, we have*

$$\pi_{-(1+\chi)\lambda R} R - \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \leq \pi_{-\lambda r} R \leq \pi_{-(1-\chi)\lambda R} R + \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda},$$

and for any $\epsilon > 0$, $0 < \gamma < \frac{1}{2}$ and $0 < \lambda \leq 0.39\,\gamma N$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, we have

$$\begin{aligned} K(\pi_{-\lambda r}, \pi_{-\lambda R}) &\leq \tfrac{4}{1-\gamma} \log \pi_{-\lambda R} \exp\left(\tfrac{4.1\,\lambda^2}{\gamma N} \pi_{-\lambda R} \mathbb{P}_{\cdot,\cdot}\right) + \tfrac{5\gamma}{1-\gamma} \log(4\epsilon^{-1}) \\ &\leq 16 \log \pi_{-\lambda R} \exp\left(\tfrac{4.1\,\lambda^2}{\gamma N} \mathbb{P}_{\cdot,\tilde{f}}\right) + 10\gamma \log(4\epsilon^{-1}). \end{aligned} \tag{2.4}$$

3. Convergence rate of classifiers under complexity and margin assumptions

3.1. **Complexity and margin assumptions.** The following assumptions have the same form as the one used in the pioneering work of Mammen and Tsybakov ([11]). The margin assumption appears to be the key assumption to obtain fast rates of convergence (i.e. $N^{-\beta}$ with $\beta > \frac{1}{2}$).

3.1.1. *Complexity assumptions.* Let $q \geq 0$. Define

$$h_q(u) \triangleq \begin{cases} \log(eu^{-1}) & \text{when} \quad q = 0 \\ u^{-q} & \text{when} \quad q > 0 \end{cases}.$$

We will alternatively use the following complexity assumptions.
(CA1) : there exists $C' > 0$ such that the covering entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{\cdot,\cdot}$ satisfies for any $u > 0$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq C' h_q(u)$.
(CA2) : there exists $C' > 0$ such that the bracketing entropy of the model $\mathcal{F}$ for the distance $\mathbb{P}_{\cdot,\cdot}$ satisfies for any $u > 0$, $H^{[]}(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq C' h_q(u)$.
(CA3) : there exist $C' > 0$ and $\pi \in \mathcal{M}_+^1(\mathcal{F})$ such that for any $t > 0$, for any $f' \in \mathcal{F}$, we have $\pi\big(\mathbb{P}_{\cdot,f'} \leq t\big) \geq \exp\left[-C' h_q(t)\right]$.

We have[4]: *(CA2)* $\Rightarrow$ *(CA1)* $\Leftrightarrow$ *(CA3)*. Let t and C' be positive reals. We will say that a probability distribution π satisfies (t, C')-*(CA3)* when we have

$$\pi\big(\mathbb{P}_{\cdot,\tilde{f}} \leq t\big) \geq \exp\left[-C' h_q(t)\right].$$

Note that this last assumption is, unlike the others, a local complexity assumption.

3.1.2. *Margin assumptions.* We will consider variants of Tsybakov's margin assumption ([11, 15]). Let $\alpha \in \mathbb{R}_+ \cup \{+\infty\}$ and $\kappa \in [1; +\infty]$. We define

$$\Delta R(f) \triangleq R(f) - R(\tilde{f}).$$

(MA1) : $\mathcal{Y} = \{0; 1\}$ and there exists $C'' > 0$ such that for any $t > 0$,

$$\mathbb{P}\big(0 < |\eta^*(X) - 1/2| \leq t\big) \leq C'' t^{\alpha}.$$

(MA2) : there exists $C'' > 0$ such that for any function $f \in \mathcal{F}$,

$$\mathbb{P}_{f,\tilde{f}} \leq C'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}}.$$

(MA3) : there exist $c'', C'' > 0$ such that for any function $f \in \mathcal{F}$,

$$c'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}} \leq \mathbb{P}_{f,\tilde{f}} \leq C'' \big[\Delta R(f)\big]^{\frac{1}{\kappa}}. \tag{3.1}$$

[4]To prove *(CA1)* $\Rightarrow$ *(CA3)*: for any $k \in \mathbb{N}^*$, introduce π_k the uniform distribution on a $(2^{-k}, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})-$minimal covering net. The prior distribution $\pi \triangleq \sum_{k \geq 1} \frac{\pi_k}{k(k+1)}$ satisfies the claim.

(MA4) : there exist $c'', C'' > 0$ such that $\mathbb{P}_{\cdot,\tilde{f}} \leq C''[\Delta R]^{\frac{1}{\kappa}}$, and for any $t > 0$, $\pi(\Delta R \leq t) \geq c''\pi\big(\mathbb{P}_{\cdot,\tilde{f}} \leq C''t^{\frac{1}{\kappa}}\big)$[5].

This last assumption makes sense only in the bayesian context where a prior distribution π is put on the model. It is easy to check the following implications: *(MA3)* $\Rightarrow$ *(MA4)* $\Rightarrow$ *(MA2)*. Besides when $f^* \in \mathcal{F}$ (= no bias assumption), we have: *(MA1)* $\Rightarrow$ *(MA2)* for $\kappa = \frac{1+\alpha}{\alpha}$. When $\kappa = +\infty$, Assumption *(MA2)* is empty[6] and Assumptions *(MA3)* and *(MA4)* are not satisfied by non-trivial models. The margin Assumptions *(MA1)* and *(MA2)* are all the stronger as κ is small. When $\kappa = 1$, the lower bound in inequality (3.1) holds trivially for $c'' = 1$ and we have: *(MA3)* $\Leftrightarrow$ *(MA4)* $\Leftrightarrow$ *(MA2)*.

Remark 3.1. For sake of simplicity, we have assumed that there exists a function $\tilde{f} \in \mathcal{F}$ such that $R(\tilde{f}) = \inf_{\mathcal{F}} R$. Then, under Assumption *(MA2)*, this function needs to be unique. In fact this is not more necessary than the existence of the minimum. To be more specific, the results in this paper under Assumption *(MA2)* will still hold when this assumption is replaced with: there exists $k \in \mathbb{N}^*$ such that for any $\beta > 0$, there exists $f_1, \ldots, f_k \in \mathcal{F}$

$$\forall f \in \mathcal{F},\ \exists i \in \{1, \ldots, k\},\ \mathbb{P}_{f,f_i} \leq C''\big[R(f) - \inf_{\mathcal{F}} R\big]^{\frac{1}{\kappa}} + \beta.$$

Note that this implies that for any $i \in \{1, \ldots, k\}$, $R(f_i) - \inf_{\mathcal{F}} R \leq \beta$. Similarly, we can give weakened versions of Assumptions *(MA3)* and *(MA4)*. Naturally, the value of k will influence the value of the constants in the results under Assumption *(MA2)*.

3.2. **Gibbs classifier.**

3.2.1. *Under Assumptions* (MA4) *and* (CA3) *for* $q > 0$. In this paper, we will often consider prior distributions $\pi^{(N)}$ which may depend on N. To shorten, we will simply write it π. The following lemma guarantees the efficiency of the standard Gibbs estimator for a temperature appropriately chosen.

Lemma 3.1. *Let* π *be a probability distribution such that*

$$\pi\big[\Delta R \leq \check{C}_1 N^{-\frac{\kappa}{2\kappa-1+q}}\big] \geq e^{-\check{C}_2 N^{\frac{q}{2\kappa-1+q}}} \tag{3.2}$$

and λ_N *have the same order as* $N^{\frac{\kappa+q}{2\kappa-1+q}}$*, i.e. such that*

$$\check{C}_3 N^{\frac{\kappa+q}{2\kappa-1+q}} \leq \lambda_N \leq \check{C}_4 N^{\frac{\kappa+q}{2\kappa-1+q}} \tag{3.3}$$

for some positive constants $\check{C}_i, i = 1, \ldots, 4$*. Then, under Assumption* (MA2)*, the standard Gibbs classifier in which the prediction function is drawn according to the posterior distribution* $\pi_{-\lambda_N r}$ *has the convergence rate* $N^{-\frac{\kappa}{2\kappa-1+q}}$ *to the extent that*

$$\mathbb{P}^{\otimes N}\pi_{-\lambda_N r}R - R(\tilde{f}) \leq \breve{C}N^{-\frac{\kappa}{2\kappa-1+q}}$$

for some constant $\breve{C} > 0$ *(depending only on* c'', C'', κ *and* $\check{C}_i, i = 1, \ldots, 4$*).*

More precisely, with $\mathbb{P}^{\otimes N}$*-probability at least* $1 - \epsilon$*, with* $\pi_{-\lambda_N r}$*-probability at least* $1 - \epsilon$*, we have*

$$R - R(\tilde{f}) \leq \breve{C}\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}, \tag{3.4}$$

for some constant $\breve{C} > 0$ *(depending on* C'', κ *and* $\check{C}_i, i = 1, \ldots, 4$*).*

---

[5] As a consequence, $\pi(\Delta R \leq t)$ has the same order as $\pi(\mathbb{P}_{\cdot,\tilde{f}} \leq C''t^{\frac{1}{\kappa}})$.

[6] since the inequality trivially holds for $C'' = 1$

Proof. See Section 6.1. □

Theorem 3.2. *Let π be a distribution satisfying Assumptions* (MA4) *and* $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \check{C}_2\big)$-(CA3) *for $q > 0$ and λ_N be a real satisfying inequality* (3.3) *for given positive constants $\check{C}_i, i = 1, \dots, 4$. Then we have*

$$\mathbb{P}^{\otimes N}\pi_{-\lambda_N r}R - R(\tilde{f}) \le \breve{C} N^{-\frac{\kappa}{2\kappa-1+q}}$$

for some constant $\breve{C} > 0$ (depending only on c'', C'', κ and $\check{C}_i, i = 1, \dots, 4$).

Proof. It suffices to check that, under these assumptions, we can apply Lemma 3.1. □

Remark 3.2. To make the link with previous works about non randomized sieve estimators, one can choose π as the uniform distribution on an almost minimal $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}\big)$-covering net. Then Assumption *(CA1)* implies that the distribution π satisfies Assumption $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \check{C}_2\big)$-*(CA3)* for some constant $\check{C}_2 > 0$ $\big($depending on $\check{C}_1$, on the almost minimality constant and on the constant C' involved in *(CA1)*$\big)$. Note that, as in Mammen and Tsybakov's work ([11, 15]), the computation of the estimator requires that, without knowing $\mathbb{P}(dX)$ exactly, one can construct a $(t, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-net with log-cardinality of order $H(t, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$.

The convergence rate of the standard Gibbs estimator in Theorem 3.2 is optimal since the following lower bound holds.

Theorem 3.3. *Let $q \ge 0$ and $\kappa \in [1; +\infty]$. There exist an input space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, a model $\mathcal{F}$ and a set $\mathcal{P}$ be the set of probability distributions satisfying*

- *for any $\mathbb{P} \in \mathcal{P}$, $f^*_{\mathbb{P}} \in \mathcal{F}$*
- *Assumptions* (CA2), (MA3) *and* (MA1) *with $\alpha = \frac{1}{\kappa-1} \in [0; +\infty]$*

such that for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$,

$$\sup_{\mathbb{P}\in\mathcal{P}} \big\{\mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f})\big\} \ge C N^{-\frac{\kappa}{2\kappa-1+q}}.$$

Proof. See Section 6.2. □

Remark 3.3. In [15], the same result is proved (using the classes of boundary fragments) for the set of probability distributions such that the Bayes classifier is in the model and Assumptions *(CA2)* and *(MA1)* with $\alpha = \frac{1}{\kappa-1}$ hold.

Remark 3.4. The previous theorem is stronger than what is required to prove that the convergence rate obtained in Theorem 3.2 is optimal since the set $\mathcal{P}$ in Theorem 3.3 is smaller than the set of probability distributions $\mathbb{P}$ such that there exists a distribution π satisfying Assumptions *(MA4)* and $\big(\check{C}_1 N^{-\frac{1}{2\kappa-1+q}}, \check{C}_2\big)$-*(CA3)*.

3.2.2. *Under Assumptions* (MA2) *and* (CA3) *for $q = 0$.* Using the same tools as in the previous section, we can prove

Theorem 3.4. *Let π be a distribution satisfying Assumption* $\big(\check{C}_1 N^{-\frac{\kappa}{2\kappa-1}}, \check{C}_2\big)$-(CA3) *for $q = 0$ and λ_N be a real satisfying*

$$\check{C}_3 N^{\frac{\kappa}{2\kappa-1}} \le \lambda_N \le \check{C}_4 N^{\frac{\kappa}{2\kappa-1}} \tag{3.5}$$

for given positive constants $\check{C}_i, i = 1, \dots, 4$. Under Assumption (MA2), *we have*

$$\mathbb{P}^{\otimes N}\pi_{-\lambda_N r}R - R(\tilde{f}) \le \breve{C}(\log N) N^{-\frac{\kappa}{2\kappa-1}}$$

for some constant $\breve{C} > 0$ (depending only on C'', κ and $\check{C}_i, i = 1, \dots, 4$).

Proof. We use Lemma 6.1 and the inequalities

$$\left\{\begin{array}{rcl} T_2(\pi) & \leq & \breve{C}\lambda\big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}} \\ T_1(\pi) & \leq & -\log\Big[\pi\big(\Delta R \leq \check{C}_1 N^{-\frac{\kappa}{2\kappa-1}}\big)\Big] + \lambda\check{C}_1 N^{-\frac{\kappa}{2\kappa-1}} \\ & \leq & -\log\Big[\pi\big(\mathbb{P}_{\cdot,\tilde{f}} \leq \check{C}_1 N^{-\frac{\kappa}{2\kappa-1}}\big)\Big] + \lambda\check{C}_1 N^{-\frac{\kappa}{2\kappa-1}} \end{array}\right. .$$

□

From Theorem 3.3, since we have *(CA2)* ⇒ *(CA3)* and *(MA3)* ⇒ *(MA2)*, this convergence rate is optimal up to the $\log N$ factor.

3.2.3. *Under Assumption* (MA2) *and a local complexity assumption.* The following theorem considers a local complexity assumption and its first and second parts respectively complete Theorem 3.4 and Lemma 3.1.

Theorem 3.5. *Let $\epsilon > 0$, $s \geq 0$, $C' > 0$, $C''' \in \mathbb{R}$ and $1 \leq \kappa \leq +\infty$. Consider λ depending on N such that $\lambda \underset{N\to+\infty}{\to} +\infty$ and that Assumption* (MA2) *holds.*

First, assume that $\log \pi^{-1}\big\{R - R(\tilde{f}) \leq x\big\} = -C' \log x + C''' + \underset{x\to 0}{\mathrm{o}}(x^s)$. Then we have

- *for $\lambda = \underset{N\to+\infty}{\mathrm{o}}\big(N^{\frac{\kappa}{2\kappa-1}}\big)$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,*

$$\pi_{-\lambda r} R = \frac{C'}{\lambda} + \frac{1}{\lambda}\Big\{\underset{N\to+\infty}{\mathrm{o}}\big(\lambda^{-\frac{s}{2}}\big) + \underset{N\to+\infty}{\mathrm{O}}\Big(\lambda^{\frac{(2\kappa-1)C'}{2(2\kappa C'+\kappa-1)}} N^{-\frac{\kappa C'}{2(2\kappa C'+\kappa-1)}}\Big)\log(e\epsilon^{-1})\Big\}$$

- *when $\lambda = cN^{\frac{\kappa}{2\kappa-1}}$: for any $\beta > 0$, there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :*

$$\tfrac{C'-\beta}{\lambda} \leq \pi_{-\lambda r} R \leq \tfrac{C'+\beta}{\lambda}.$$

Secondly, assume that $\log \pi^{-1}\big\{R - R(\tilde{f}) \leq x\big\} = C' x^{-\frac{q}{\kappa}} + C''' + \underset{x\to 0}{\mathrm{o}}(1)$ with $q > 0$ and $\kappa \neq +\infty$. Then we have

- *for $\lambda = \underset{N\to+\infty}{\mathrm{o}}\big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\big)$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,*

$$\pi_{-\lambda r} R = \big(\tfrac{qC'+\mathrm{o}(1)}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} + \underset{N\to+\infty}{\mathrm{O}}(1)\tfrac{\log(\epsilon^{-1})}{\lambda}$$

- *when $\lambda = cN^{-\frac{\kappa+q}{2\kappa-1+q}}$: for any $0 < \beta \leq qC'$, there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :*

$$\big(\tfrac{qC'-\beta}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} \leq \pi_{-\lambda r} R \leq \big(\tfrac{qC'+\beta}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}.$$

Proof. See Section 6.3. □

It is interesting to note that this asymptotic behaviour only depends on the local complexity given by the weight of the sets $\big\{f \in \mathcal{F} : R(f) - R(\tilde{f}) \leq x\big\}$ when $x \to 0$. Had we had $\mathbb{P}_{f,\tilde{f}} \underset{x\to 0}{\sim} C''\big[R - R(\tilde{f})\big]^{\frac{1}{\kappa}}$ on these sets, the complexity assumption would be similar to the ones introduced in Section 3.1.1 to the extent that we would have $\log \pi^{-1}\big(\mathbb{P}_{\cdot,\tilde{f}} \leq x\big) \underset{x\to 0}{\sim} \breve{C} h_q(x)$.

In Theorems 3.2 and 3.4, we have seen how to choose the parameter λ depending on N such that the Gibbs classifier has the optimal convergence rate. The previous result shows that for λ smaller than these "optimal" parameters and a slightly modified complexity assumption, we can tightly bracket the efficiency of standard

Gibbs classifiers. For larger λ, the picture is not clear: it seems that the KL-divergence term in Theorem 2.2 becomes the leading term. This KL-divergence will in general explode for $\lambda \gg N$, and finally we just know that

$$\pi_{-\lambda r} R \underset{\lambda \to +\infty}{\to} \pi|_{r=\min_{\mathcal{F}} r} R \triangleq \frac{\int_{\mathcal{F}} \mathbb{1}_{r(f)=\min_{\mathcal{F}} r} R(f) d\pi(f)}{\int_{\mathcal{F}} \mathbb{1}_{r(f)=\min_{\mathcal{F}} r} d\pi(f)}.$$

Remark 3.5. The confidence level ϵ does not appear in the main terms of the expansions of $\pi_{-\lambda r} R$, hence the asymptotic orders of $\pi_{-\lambda r} R$ hold with exponential probability.

3.2.4. *Adaptive choice of the temperature.* Here we consider that Assumption *(MA3)* holds for an unknown margin parameter κ and we prove that under assumption *(CA3)* a standard Gibbs classifier with an appropriately chosen temperature is adaptive wrt this parameter, i.e. without prior knowledge of κ, the generalization error of the randomized estimator is upper bounded by $\breve{C} N^{-\frac{\kappa}{2\kappa-1+q}}$ when $q > 0$ and by $\breve{C}(\log N) N^{-\frac{\kappa}{2\kappa-1}}$ when $q = 0$. The adaptation to the margin problem has also been studied in [15, 16]. In particular, in [16], Tsybakov and van de Geer proposed an adaptive penalized classifier using wavelets.

Theorem 3.6. *Under Assumptions* (MA3) *and* (CA3)*, the algorithm given in Section* 3.4.2 *of* [1] *achieves an adaptive choice of the temperature of the standard Gibbs classifier wrt the margin parameter* κ*.*

Proof. See Section 6.4. □

3.3. **Empirical risk minimization on nets.**

3.3.1. *Under Assumptions* (MA3) *and* (CA1) *for* $q > 0$. This section shows that, by using inequality (2.1), we can recover results on sieve estimators given in [11, 15]. These results have to be compared with the ones in Section 3.2.1 (recall that *(MA3)* $\Rightarrow$ *(MA4)* and *(CA1)* $\Leftrightarrow$ *(CA3)*).

Theorem 3.7. *Under Assumptions* (MA3) *and* (CA1) *for* $q > 0$*, for any classifier* $\hat{f}$ *minimizing the empirical risk among a* u_N*-covering net* $\mathcal{N}_{u_N}$ *such that*

$$\check{C}_1 N^{-\frac{1}{2\kappa-1+q}} \le u_N \le \check{C}_2 N^{-\frac{1}{2\kappa-1+q}} \tag{3.6}$$

and

$$\log \left|\mathcal{N}_{u_N}\right| \le \check{C}_3 h_q(u_N) \tag{3.7}$$

for some positive constants $\check{C}_i, i = 1, \dots, 3$*, we have*

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - R(\tilde{f})\big] \le \breve{C} N^{-\frac{\kappa}{2\kappa-1+q}}$$

for some constant $\breve{C} > 0$ *(depending only on* C', c'', C'' *and* $\check{C}_i, i = 1, \dots, 3$*).*

Proof. See Section 6.5. □

Remark 3.6. Inequality (3.7) just says that the net $\mathcal{N}_{u_N}$ is almost minimal.

3.3.2. *Under Assumptions* (MA2) *and* (CA1) *for* $q = 0$. Since *(CA1)* $\Leftrightarrow$ *(CA3)*, this section gives results to be compared with the ones in Section 3.2.2.

Theorem 3.8. *Under Assumptions* (MA2) *and* (CA1) *for* $q = 0$, *for any classifier* $\hat{f}$ *minimizing the empirical risk among a* u_N*-covering net* $\mathcal{N}_{u_N}$ *such that*

$$N^{-\check{C}_1} \leq u_N \leq \check{C}_2 (\log N) N^{-\frac{\kappa}{2\kappa-1}} \tag{3.8}$$

and

$$\log \left| \mathcal{N}_{u_N} \right| \leq \check{C}_3 h_0(u_N) \tag{3.9}$$

for some positive constants $\check{C}_i, i = 1, \ldots, 3$, *we have*

$$\mathbb{P}^{\otimes N} \big[R(\hat{f}) - R(\tilde{f}) \big] \leq \breve{C} (\log N) N^{-\frac{\kappa}{2\kappa-1}}$$

for some constant $\breve{C} > 0$ *(depending only on* c'', C'' *and* $\check{C}_i, i = 1, \ldots, 3$*).*

Proof. It follows the lines of Section 6.5. This time, we take $\left(\frac{\lambda}{N}\right)^{\frac{\kappa}{\kappa-1}}$ and $\frac{\log(eu^{-1})}{\lambda}$ of the same order and greater than u. This is realized when inequality (3.8) is satisfied and $\lambda = N^{\frac{\kappa}{2\kappa-1}}$. □

3.4. **Chaining.** When a class of functions has a polynomial entropy, there is a trick called the chaining ([6]) which allows us to improve the previous results. This technique is used to get tighter upper bounds of the difference $R(f_1) - R(f_2)$ between the expected risk at two different functions f_1 and f_2. It is based on finer and finer approximations of these functions. The advantage of considering rough approximation of these functions is that the set of all possible rough approximations is small (in other words, has a small complexity). On the contrary, the set of fine approximations is big, but the distance between the fine approximation and the function approximated is small. So there is a kind of bias/variance trade-off and for polynomial entropy classes of functions, it is interesting to have this trade-off on a sequence of links and not directly on the big link $f_1 \cdots f_2$.

Let us give some results due to this technique. Consider the context of Theorem 3.7. Let us see what happens if we replace the margin Assumption *(MA3)* with Assumption *(MA2)*. Then we can no longer upper bound ΔR with $\mathrm{Cst}\, \mathbb{P}_{\cdot,\tilde{f}}^{\kappa}$ (inequality which is used to obtain (6.8)). We only have $\Delta R \leq \mathbb{P}_{\cdot,\tilde{f}}$. This leads to the convergence rate $N^{-\frac{\kappa}{2\kappa-1+q\kappa}}$ instead of $N^{-\frac{\kappa}{2\kappa-1+q}}$. Using the chaining trick, we will prove (see Theorem 3.10) that this rate is suboptimal and that, by minimizing the empirical risk on well chosen nets, we can still reach the rate $N^{-\frac{\kappa}{2\kappa-1+q}}$ when $0 < q < 1$ and the rate $N^{-\frac{1}{1+q}}$ when $q > 1$.

Remark 3.7. The convergence rate $N^{-\frac{\kappa}{2\kappa-1+q\kappa}}$ is optimal under Assumption *(MA2)* and the complexity assumption $H(u_N) \leq C' h_q(u_N)$ for the radius $u_N = N^{-\frac{\kappa}{2\kappa-1+q\kappa}}$. The lower bound comes from Lemma 5.1 applied to a $\big(N^{\frac{q\kappa}{2\kappa-1+q\kappa}}, N^{-\frac{1+q\kappa}{2\kappa-1+q\kappa}}, \frac{1}{2} N^{-\frac{\kappa-1}{2\kappa-1+q\kappa}}\big)$-constant hypercube. By slightly modifying the proof of Theorem 3.7, we can obtain that, under the previous margin and complexity assumptions, any classifier $\hat{f}$ minimizing the empirical risk among a u_N-almost minimal net satisfies

$$\mathbb{P}^{\otimes N} \big[R(\hat{f}) - R(\tilde{f}) \big] \leq \breve{C} N^{-\frac{\kappa}{2\kappa-1+q\kappa}}.$$

The chaining technique appears to be the only tool which allows to take into account an entropy assumption which holds for any radius such as *(CA1)* and *(CA2)*.

The chaining trick may also be used to prove that the empirical risk can be minimized on tighter nets (provided that they are still minimal, or at least almost minimal, under polynomial entropy assumptions).

Before giving results concerning nets, one can illustrate the chaining technique by considering randomized posteriors concentrated on small balls of fixed radius. For any $u>0$, introduce $B_{f,u} \triangleq \{f' \in \mathcal{F} : \mathbb{P}_{f,f'} \le u\}$ and $\pi_{f,u} \triangleq \frac{\mathbb{1}_{B_{f,u}}}{\pi(B_{f,u})} \cdot \pi$. Define $h(v) \triangleq \sup_{f\in\mathcal{F}} \log \pi^{-1}(B_{f,v})$ and $h_+(v) \triangleq h(v) \vee 1$.

Theorem 3.9. *Let $u>0$, $L \triangleq \frac{\log(2/u)}{\log 2}$, $C_1 \triangleq \sqrt{\frac{4h_+(u)}{3Nu}}$ and $C_0 \triangleq 2\sqrt{3}[1+2g(C_1)]$. For any $\epsilon>0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \mathcal{F}$ such that $\mathbb{P}_{f_1,f_2} > u$, we have*

$$\begin{aligned}\pi_{f_2,u}R - \pi_{f_1,u}R + \pi_{f_1,u}r - \pi_{f_2,u}r & \\ &\le \frac{C_0}{\sqrt{N}} \sum_{k\in\mathbb{N}: u2^k < \mathbb{P}_{f_1,f_2}} \sqrt{u2^k h_+(u2^k)} + 6\sqrt{\frac{\mathbb{P}_{f_1,f_2}}{N}} \log[L\epsilon^{-1}] \\ &\le \frac{2C_0}{\sqrt{N}} \int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h_+(v)}{v}} dv + 6\sqrt{\frac{\mathbb{P}_{f_1,f_2}}{N}} \log[L\epsilon^{-1}].\end{aligned}$$

Proof. See Section 6.6. □

Had we not chained inequality (2.1), we would have obtained

$$\pi_{f_2,u}R - \pi_{f_1,u}R + \pi_{f_1,u}r - \pi_{f_2,u}r \le \tfrac{\lambda}{N} g\big(\tfrac{\lambda}{N}\big)\big(\mathbb{P}_{f_1,f_2} + 2u\big) + \tfrac{2h(u)+\log(\epsilon^{-1})}{\lambda}$$

This upper bound is greater than $\inf_{\lambda>0}\big\{\frac{\lambda}{2N}\mathbb{P}_{f_1,f_2} + \frac{2h(u)}{\lambda}\big\} = 2\sqrt{\frac{\mathbb{P}_{f_1,f_2} h(u)}{N}}$, which is much bigger than the chained bound for polynomial entropies $h(u) \approx u^{-q}$, $q>0$ when[7] $N^{-\frac{1}{1+q}} \le u \ll \mathbb{P}_{f_1,f_2}$.

The following result is an extension of Theorems 3.7 and 3.8.

Theorem 3.10. *We assume that Assumptions* (MA2) *and* (CA1) *hold. When Assumption* (MA3) *also holds, we define*

$$(v_N, a_N) \triangleq \begin{cases} \Big(\big[\frac{\log N}{N}\big]^{\frac{\kappa}{2\kappa-1}}, \exp\Big\{-\check{C}_1 (\log N)^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}}\Big\}\Big) & \text{for } q=0 \\ \Big(N^{-\frac{\kappa}{2\kappa-1+q}}, \check{C}_1 N^{-\frac{(\kappa-1)\mathbb{1}_{q\le 1}+q}{q(2\kappa-1+q)}}\Big) & \text{for } q>0 \end{cases}$$

and $b_N \triangleq \check{C}_2 (v_N)^{\frac{1}{\kappa}}$.

When Assumption (MA3) *does not hold, we define*

$$(v_N, a_N) \triangleq \begin{cases} \Big(\big[\frac{\log(eN^{1/\kappa})}{N}\big]^{\frac{\kappa}{2\kappa-1}}, \exp\Big\{-\check{C}_1 (\log[eN^{1/\kappa}])^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}}\Big\}\Big) & \text{for } q=0 \\ \Big(N^{-\frac{\kappa}{2\kappa-1+q}}, \check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}}\Big) & \text{for } 0<q<1 \\ \Big((\log N) N^{-\frac12}, \check{C}_1 (\log N)^{-\frac12} N^{-\frac12}\Big) & \text{for } q=1 \\ \Big(N^{-\frac{1}{1+q}}, \check{C}_1 N^{-\frac{1}{1+q}}\Big) & \text{for } q>1 \end{cases}$$

and $b_N \triangleq \check{C}_2 v_N$.

[7] The quantity C_0 behaves as a constant only when $\frac{h_+(u)}{Nu} \le C$, so when $u \ge CN^{-\frac{1}{1+q}}$.

For any classifier minimizing the empirical risk among a u_N-covering net $\mathcal{N}_{u_N}$ such that

$$a_N \leq u_N \leq b_N \tag{3.10}$$

and

$$\log\left|\mathcal{N}_{u_N}\right| \leq \check{C}_3 h_q(u_N) \tag{3.11}$$

for some positive constants $\check{C}_i, i = 1, \ldots, 3$, we have

$$\mathbb{P}^{\otimes N}\big[R(\hat{f}) - R(\tilde{f})\big] \leq C v_N$$

for some constant $C > 0$ $\big($depending on C'', $\check{C}_i, i = 1, \ldots, 3$ [and also on c'' under Assumption (MA3)]$\big)$.

Proof. See Section 6.7. □

Remark 3.8. When $q = 0$ and $\kappa = +\infty$ (i.e. no margin assumption), the $\log N$ factor in $\log(eN^{1/\kappa})$ disappears. The suppression of the logarithmic factor, obtained by chaining, is similar to what occurs for VC classes (see Corollary 4.6). The difference is just that the complexity assumption concerns $\mathbb{P}$-nets here instead of empirical nets.

From Theorems 3.3 and the following theorem, these convergence rates are optimal (up to the logarithmic factor when we have $q \in \{0; 1\}$).

Theorem 3.11. *There exist an input space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, a model $\mathcal{F}$ and a set $\mathcal{P}$ of probability distributions satisfying Assumptions* (CA2) *and* (MA2) *such that for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$,*

$$\sup_{\mathbb{P}\in\mathcal{P}} \big\{\mathbb{P}^{\otimes N} R(\hat{f}) - R(\tilde{f})\big\} \geq C N^{-\frac{1}{1+q}}.$$

Proof. Apply Lemma 5.1 for a set $\mathcal{P}$ equal to a $\big(N^{\frac{q}{1+q}}, N^{-1}, \frac{1}{2}\big)$-constant hypercube and take $\mathcal{F} \triangleq \big\{f^*_{\mathbb{P}} : \mathbb{P} \in \mathcal{P}\big\}$. □

In Theorem 3.10, we consider classifiers which minimize the empirical risk on an almost minimal net $\mathcal{N}$. The following result just asserts that the same convergence rate holds for randomized estimators which "roughly" minimizes the empirical risk.

Theorem 3.12. *For any randomized classifier $\hat{\rho} : \mathcal{Z}^N \to \mathcal{M}^1_+(\mathcal{N}_{u_N})$ such that there exists a function $\breve{f} \in \mathcal{F}$ satisfying*

- $\mathbb{P}_{\breve{f},\tilde{f}} \leq C u_N$,
- *for any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, $\hat{\rho} r \leq r(\breve{f}) + C \log(e\epsilon^{-1}) v_N$,*

we have

$$\mathbb{P}^{\otimes N} \hat{\rho} R - R(\tilde{f}) \leq \breve{C} v_N.$$

Proof. It suffices to modify slightly the proof of Theorem 3.10. Let $\tilde{f}_{\mathcal{N}}$ be the nearest neighbour of $\breve{f}$ in $\mathcal{N}_{u_N}$. We have $\mathbb{P}_{\tilde{f}_{\mathcal{N}},\breve{f}} \leq \breve{C} u_N$. From inequality (2.3) with $\mathcal{S}_1 = \{\breve{f}\}$ and $\mathcal{S}_2 = \{\tilde{f}_{\mathcal{N}}\}$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, we have

$$r(\breve{f}) \leq r(\tilde{f}_{\mathcal{N}}) + \breve{C}\sqrt{\tfrac{u_N \log(\epsilon^{-1})}{N}} + \breve{C}\tfrac{\log(\epsilon^{-1})}{N} + \sup_{\mathbb{P}_{\cdot,\tilde{f}} \leq \breve{C} u_N} \Delta R,$$

hence $r(\breve{f}) \le r(\tilde{f}_{\mathcal{N}}) + \breve{C}\log(e\epsilon^{-1})v_N$. We obtain that $\hat{\rho}r \le r(\tilde{f}_{\mathcal{N}}) + \breve{C}\log(e\epsilon^{-1})v_N$. From this inequality and by using the last bound in Corollary 6.6, we obtain a new version of inequality (6.12) from which the convergence rate follows. □

As a consequence, the Gibbs estimators $\pi_{-\lambda r}$ in which the prior distribution is the uniform distribution on a net $\mathcal{N}$ perform as well as an $(\mathrm{ERM}, \mathcal{N})$-algorithm (i.e. a classifier which minimizes the empirical risk on the net $\mathcal{N}$) as soon as the inverse temperature parameter λ is sufficiently large. This is not surprising to the extent that the Gibbs estimator $\pi_{-\lambda r}$ when $\lambda \to +\infty$ classifies, roughly speaking, as an $(\mathrm{ERM}, \mathcal{N})$-algorithm.

The following theorem completes Theorem 3.10.

Theorem 3.13. *Let $\mathcal{N}_{u_N}$ be a u_N-covering net such that inequations* (3.10) *and* (3.11) *hold, let $\lambda_N \ge \breve{C}_4 \frac{h_q(u_N)}{v_N}$, and let π be a probability distribution on the net $\mathcal{N}_{u_N}$ satisfying $(u_N, \breve{C}_5)$-*(CA3) *for some positive constants $\breve{C}_i, i = 1, \dots, 5$. Then we have*

$$\mathbb{P}^{\otimes N}\big[\pi_{-\lambda_N r} R - R(\tilde{f})\big] \le \breve{C} v_N$$

for some constant $\breve{C} > 0$ $\big($depending on C'', $\breve{C}_i, i = 1, \dots, 5$ [and also on c'' under Assumption (MA3)$]\big)$.

Proof. Introduce the function $\tilde{f}_{\mathcal{N}}$ in the net $\mathcal{N}_{u_N}$ such that $\mathbb{P}_{\tilde{f}, \tilde{f}_{\mathcal{N}}} \le u_N$. By Assumption $(u_N, \breve{C}_5)$-*(CA3)* and inequality (3.11), we can choose the function $\tilde{f}_{\mathcal{N}}$ such that we also have $\pi(\{\tilde{f}_{\mathcal{N}}\}) \ge e^{-(\breve{C}_5 + \breve{C}_3) h_q(u_N)}$. So we have

$$\pi_{-\lambda_N r} r - r(\tilde{f}_{\mathcal{N}}) \le \frac{\log[\pi(\tilde{f}_{\mathcal{N}})^{-1})]}{\lambda_N} \le \breve{C}\frac{h_q(u_N)}{\lambda_N} \le \breve{C} v_N.$$

The result then follows from Theorem 3.12. □

3.5. **Bracketing entropy.** To minimize the empirical risk over all the model $\mathcal{F}$ can lead to inconsistency even for models with small covering entropy. For instance, define the set $\mathcal{X} = [0; 1]$, the functions $f_0 \equiv 0$ and $f_1 \equiv 1$, and the probability distribution $\mathbb{P}$ such that $\mathbb{P}(dX) = \mathcal{U}([0; 1])(dX)$ (uniform law over $\mathcal{X}$) and $Y = \mathbb{1}_{X \ge \frac{3}{4}}$. Consider the model formed by f_1 and all the functions equal to f_0 except on a finite number of points. For any $u < 1$, we have $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) = \log 2$. However, in general, the ERM-algorithm will classify poorly[8]. (On the contrary, the classifier based on the ERM-principle over a $(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-net for small u is efficient). This phenomenon occurs since the covering entropy does not suitably measures the complexity of models. In this section, we will see that the bracketing entropy does not suffer from this drawback.

Under polynomial bracketing entropy conditions, the empirical data contain what happens in expectation to the extent that two functions close for the distance $\mathbb{P}_{\cdot,\cdot}$ are also close for the distance $\bar{\mathbb{P}}_{\cdot,\cdot}$.

Recall that if $\mathcal{G}$ is a u-bracketing net of the set $\mathcal{F}$, then for any function $f \in \mathcal{F}$, there exist $f_L, f_U \in \mathcal{G}$ satisfying $f_L \le f \le f_U$ and $\mathbb{P}_{f_L, f_U} \le u$. Let us define the mappings $n_L, n_U : \mathcal{F} \to \mathcal{G}$ such that $n_L(f) = f_L$ and $n_U(f) = f_U$ (from the axiom of choice, they exist).

The following theorem, to be compared with Theorem 3.10, shows the influence of considering bracketing entropy assumptions instead of covering ones.

[8]That is why, in Theorem 3.10, we need to consider almost *minimal* nets (inequality (3.7)).

Theorem 3.14. *Let us define*

$$w_N \triangleq \begin{cases} \left[\frac{\log(eN^{1/\kappa})}{N}\right]^{\frac{\kappa}{2\kappa-1}} & \text{under Assumptions (MA2)+(CA2) for } q=0 \\ N^{-\frac{\kappa}{2\kappa-1+q}} & \text{under Assumptions (MA2)+(CA2) for } 0<q<1 \\ (\log N)N^{-\frac{1}{2}} & \text{under Assumptions (MA2)+(CA2) for } q=1 \\ N^{-\frac{1}{1+q}} & \text{under Assumptions (MA2)+(CA2) for } q>1 \end{cases}.$$

For any classifier $\hat{f}_{ERM,\mathcal{N}}$ minimizing the empirical risk in a $u_N \triangleq \check{C}_1 w_N$-covering net $\mathcal{N}$ for some positive constant $\check{C}_1$, we have

$$\mathbb{P}^{\otimes N}\left[R(\hat{f}_{ERM,\mathcal{N}}) - R(\tilde{f})\right] \leq \breve{C} w_N$$

for some constant $\breve{C} > 0$ $\big($depending on C', C'' and $\check{C}_1\big)$.

Proof. Let $\mathcal{N}'$ be a u_N-minimal bracketing net of the net $\mathcal{N}$. Let $\tilde{f}_{\mathcal{N}'}$ be the nearest neighbour of the function $\tilde{f}$ in the net $\mathcal{N}'$. By definition of the set $\mathcal{N}'$,

- we have $\log|\mathcal{N}'| \leq C' h_q(u_N)$,
- there exists a function $f_{\mathcal{N}}$ such that $n_L(f_{\mathcal{N}}) = \tilde{f}_{\mathcal{N}'}$ or $n_U(f_{\mathcal{N}}) = \tilde{f}_{\mathcal{N}'}$; consequently, we have $r(f_{\mathcal{N}}) \leq r(\tilde{f}_{\mathcal{N}'}) + u_N$,
- there exists a classifier $\hat{f}_{\mathcal{N}'} : \mathcal{Z}^N \to \mathcal{N}'$ $\big(\hat{f}_{\mathcal{N}'} \triangleq n_L(\hat{f}_{\mathrm{ERM},\mathcal{N}})$ for instance$\big)$ such that we have $r(\hat{f}_{\mathcal{N}'}) \leq r(\hat{f}_{\mathrm{ERM},\mathcal{N}}) + u_N$ and

$$R(\hat{f}_{\mathrm{ERM},\mathcal{N}}) \leq R(\hat{f}_{\mathcal{N}'}) + u_N. \tag{3.12}$$

So the estimator $\hat{f}_{\mathcal{N}'} : \mathcal{Z}^N \to \mathcal{N}'$ satisfies

$$r(\hat{f}_{\mathcal{N}'}) \leq r(\hat{f}_{\mathrm{ERM},\mathcal{N}}) + u_N \leq r(f_{\mathcal{N}}) + u_N \leq r(\tilde{f}_{\mathcal{N}'}) + 2u_N.$$

Then the result follows from Theorem 3.12 and inequality (3.12). □

Remark 3.9. Since we have *(CA2) ⇒ (CA1)*, Theorem 3.10 can be applied when the assumptions of Theorem 3.14 hold. We see that, under bracketing entropy assumptions, the ERM on nets containing a huge (possibly infinite) number of functions has also the optimal convergence rate. This was not the case under covering entropy assumptions.

Remark 3.10. The same convergence rate holds for classifiers minimizing the empirical risk up to an additive factor Cw_N.

The following theorem completes the previous one.

Theorem 3.15. *Let $\lambda_N \geq \check{C}_1 \frac{h_q(w_N)}{w_N}$ and π be a probability distribution satisfying $(\check{C}_2 w_N, \check{C}_3)$-*(CA3) *for some positive constants $\check{C}_i, i = 1, \ldots, 3$. Then we have*

$$\mathbb{P}^{\otimes N}\left[\pi_{-\lambda_N r} R - R(\tilde{f})\right] \leq \breve{C} w_N$$

for some constant $\breve{C} > 0$ (depending on C'', $\check{C}_i, i = 1, \ldots, 3$).

Proof. See Section 6.8. □

Remark 3.11. From the previous theorem, the inverse temperature parameter λ_N should be taken as

$$\lambda_N \geq C \begin{cases} (\log N) N^{\frac{\kappa}{2\kappa-1}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q=0 \\ N^{\frac{\kappa(1+q)}{2\kappa-1+q}} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } 0<q<1 \\ \frac{N}{(\log N)^2} & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q=1 \\ N & \text{under Assumptions } \textit{(MA2)+(CA2)} \text{ for } q>1 \end{cases}.$$

The threshold value is all the smaller as the model is small[9] (i.e for small q) and the margin assumption is weak[10] (i.e for large κ).

Finally, the following theorem shows that, under polynomial bracketing entropy assumption, with high probability, the empirical covering nets are similar to the covering nets wrt the pseudo-distance $\mathbb{P}(dX)$.

Theorem 3.16. *Let $\breve{C}$ be positive constant and define*

$$(\alpha_q, \beta_q) = \begin{cases} \left(\frac{1}{N}, \frac{\log N}{N}\right) & \text{when } q = 0 \\ \left(\exp\left\{-N^{\frac{q}{1+q}}\right\}, N^{-\frac{1}{1+q}}\right) & \text{when } q > 0 \end{cases}.$$

With $\mathbb{P}^{\otimes N}$-probability at least $1 - (\alpha_q)^{\breve{C}}$, there exists $\breve{C}_1, \breve{C}_2, \breve{C}_3, \breve{C}_4 > 0$ such that for any $u \geq \breve{C}_1 \beta_q$,

- *a $(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-covering net is a $(\breve{C}_3 u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$-covering net,*
- *a $(u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$-covering net is a $(\breve{C}_2 u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-covering net,*
- *$H(u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) \leq \breve{C}_4 h_q(u)$.*

Proof. See Section 6.9. □

Therefore under polynomial bracketing entropy assumption, we can classify optimally by using the minimizer of the empirical risk on an *empirical* net of radius less than Cw_N. Note that another way of proving this result consists in saying that this classifier minimizes the empirical risk on the set $\mathcal{F}$ up to an additive Cw_N factor.

4. Classification under empirical complexity assumptions

In this section, we will see that if we replace the complexity assumption concerning $\mathbb{P}$-entropies with a similar assumption on the empirical entropies, the same kind of convergence rates appear. VC-classes are a special case in which for any $u > 0$ and any training set, we have $H(u, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot}) \leq CVh_0(u)$ where V is the VC-dimension of $\mathcal{F}$.

4.1. Concentration of the empirical entropies.

In general, the link between the $\mathbb{P}$-entropies and $\bar{\mathbb{P}}$-entropies is not known. However, thanks to recent work by Boucheron, Bousquet, Lugosi and Massart, we are able to prove that the empirical entropies are concentrated.

Theorem 4.1. *For any $\epsilon > 0$ and $u \geq 0$*

- *with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, we have*

(4.1)

$$H(u, \mathcal{F}, \bar{\mathbb{P}}) \leq \mathbb{P}^{\otimes N} H(u, \mathcal{F}, \bar{\mathbb{P}}) + \frac{(\log 2)\log(\epsilon^{-1})}{3}\left(1 + \sqrt{1 + \frac{18\mathbb{P}^{\otimes N} H(u, \mathcal{F}, \bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}}\right)$$

[9]This might be explained by looking at the size of the sets $\{f \in \mathcal{F} : r(f) - \min_{\mathcal{F}} r = \frac{k}{N}\}$. Indeed, when the model becomes larger and larger, the weight on these sets increases much more for small k than for very small k, hence we need to have larger λ to get rid of functions having a not-so-small empirical risk.

[10]This is not surprising since the stronger the margin assumption is, the smaller the optimal convergence rate is, and consequently the more selective we need to be.

equivalently

$$\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}) \geq H(u,\mathcal{F},\bar{\mathbb{P}}) + \frac{2(\log 2)\log(\epsilon^{-1})}{3}\left(1-\sqrt{1+\frac{9H(u,\mathcal{F},\bar{\mathbb{P}})}{2(\log 2)\log(\epsilon^{-1})}}\right)$$

- *with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$,*

$$H(u,\mathcal{F},\bar{\mathbb{P}}) \geq \mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}) - \sqrt{2(\log 2)\log(\epsilon^{-1})\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}})}$$

equivalently

$$\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}) \leq H(u,\mathcal{F},\bar{\mathbb{P}}) + (\log 2)\log(\epsilon^{-1})\left(1+\sqrt{1+\frac{2H(u,\mathcal{F},\bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}}\right) \tag{4.2}$$

- *with $\mathbb{P}^{\otimes 2N}$-probability at least $1-2\epsilon$,*

$$H(u,\mathcal{F},\bar{\mathbb{P}}') \leq H(u,\mathcal{F},\bar{\mathbb{P}}) + 2(\log 2)\log(\epsilon^{-1})\left(\frac{6}{5}+\sqrt{1+\frac{2H(u,\mathcal{F},\bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}}\right) \tag{4.3}$$

Proof. See Section 6.10. □

The previous result shows that the empirical entropies behave with high probability as the non empirical quantity $\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}})$. Specifically, by using a union bound on the different possible radius, we obtain that for any $\breve{C}' > 0$ there exists $\breve{C} > 0$ such that with probability at least $1 - \frac{1}{N^{\breve{C}'}}$, for any $u > 0$, we have[11]

$$\left\{\begin{array}{lcl} \mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) & \leq & \breve{C}[H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + \log N] \\ H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) & \leq & \breve{C}[\mathbb{P}^{\otimes N} H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + \log N] \\ H(u,\mathcal{F},\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}) & \leq & \breve{C}[H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + \log N] \\ H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) & \leq & H(u/2,\mathcal{F},\bar{\bar{\mathbb{P}}}_{\cdot,\cdot}) \end{array}\right. .$$

4.2. **Chaining empirical quantities...**

4.2.1. *...in the transductive learning.* In this section, we assume that we possess two samples of size N. The first sample is labeled: $\{(X_1,Y_1),\ldots,(X_N,Y_N)\}$. The second one $\{X_{N+1},\ldots,X_{2N}\}$ has to be labeled: the outputs $\{Y_{N+1},\ldots,Y_{2N}\}$ are unknown. We will use the following notations:

$$\left\{\begin{array}{lcl} \bar{\mathbb{P}} & \triangleq & \frac{1}{N}\sum_{i=1}^{N}\delta_{(X_i,Y_i)} \\ \bar{\mathbb{P}}' & \triangleq & \frac{1}{N}\sum_{i=N+1}^{2N}\delta_{(X_i,Y_i)} \\ \bar{\bar{\mathbb{P}}} & \triangleq & \frac{1}{2N}\sum_{i=1}^{2N}\delta_{(X_i,Y_i)} \\ r(f) & \triangleq & \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{Y_i\neq f(X_i)} = \bar{\mathbb{P}}[Y\neq f(X)] \\ r'(f) & \triangleq & \frac{1}{N}\sum_{i=N+1}^{2N}\mathbb{1}_{Y_i\neq f(X_i)} = \bar{\mathbb{P}}'[Y\neq f(X)] \end{array}\right.$$

Let us start with a basic result which is not "chained".

Lemma 4.2. *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two finite sets of functions from $\mathcal{X}$ into $\mathcal{Y}$ possibly depending on the data Z_1^{2N} in an exchangeable way. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any functions $f_1 \in \mathcal{S}_1$ and $f_2 \in \mathcal{S}_2$, we have*

$$r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \leq \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{f_1,f_2}\log(|\mathcal{S}_1||\mathcal{S}_2|\epsilon^{-1})}{N}}$$

[11] For the third inequality, we use the inequality $H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) \leq H(u,\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + H(u,\mathcal{F},\bar{\mathbb{P}}'_{\cdot,\cdot})$ and inequality (4.3). The fourth inequality always holds.

Proof. The result comes from inequality (8.7) in [1] in which we take ν equal to the uniform distribution on $\mathcal{S}_1 \times \mathcal{S}_2$ and $\mathcal{W}[(f_1,f_2),Z] = \mathbb{1}_{Y \neq f_2(X)} - \mathbb{1}_{Y \neq f_1(X)}$. □

By chaining this inequality, we obtain:

Theorem 4.3. *Let $U \in \mathbb{N}^*$ and $u \triangleq 2^{-U}$. Let $\mathcal{N}$ be an u-minimal covering net. For any $k \in \mathbb{N}^*$, let H_k be an upper bound of $H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}}) \vee 1$. For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \mathcal{N}$,*

$$r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \leq \sum_{k \in \mathbb{N}^* : u \leq 2^{-k} \leq \bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u} 4\sqrt{\frac{6 \times 2^{-k}\{2H_k + \log[3k(k+1)] + \log(\epsilon^{-1})\}}{N}}.$$

Proof. See Section 6.11. □

Remark 4.1. The previous result can also be written in terms of integral. For instance, for $H_k = H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}}) \vee 1$, the previous RHS is upper bounded by[12]

$$\frac{28}{\sqrt{N}} \int_{\frac{u}{2}}^{(\bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u) \wedge \frac{1}{2}} \left(\sqrt{\frac{H(x,\mathcal{F},\bar{\bar{\mathbb{P}}}) \vee 1}{x}} + \sqrt{\frac{\log(4 \log x^{-1})}{x}} \right) dx + 34 \sqrt{\frac{(\bar{\bar{\mathbb{P}}}_{f_1,f_2} \vee u) \log(3\epsilon^{-1})}{N}}.$$

4.2.2. *...in the inductive learning.* The empirical bound for the inductive learning is derived from the one for the transductive learning and from the concentration properties of the pseudo-distances and the empirical entropies.

Theorem 4.4. *Let $\epsilon > 0$ and $H_\infty \triangleq 16 \log N(X_1^N) + 20 \log(5 \log N) + 12 \log(\epsilon^{-1})$. With $\mathbb{P}^{\otimes 2N}$-probability $1 - 3\epsilon$, for any functions f_1 and f_2 in the set $\mathcal{F}$, we have*

$$r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \leq \frac{10}{\sqrt{N}} \sum_{\substack{k \in \mathbb{N}^* : \\ \frac{1}{2N} \leq 2^{-k} \leq \frac{5}{4} \bar{\mathbb{P}}_{f_1,f_2} + \frac{H_\infty}{N}}} \sqrt{8H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{.,.}) + 6\log[k(k+1)\epsilon^{-1}] + 1}\,.$$

Proof. See Section 6.12. □

The previous theorem gives, for instance, a guarantee of misclassification rate of the ERM-classifier on N new input data to classify. We recall that the leading term in the square root is generally the entropy one. Once more, we can upper bound the associated sum with the integral entropy

$$\frac{C}{\sqrt{N}} \int_{\frac{1}{4N}}^{\frac{5}{4}\bar{\mathbb{P}}_{f_1,f_2} + \frac{H_\infty}{N}} \sqrt{\frac{H(x,\mathcal{F},\bar{\mathbb{P}}_{.,.})}{x}} dx$$

Note that this result is less general than the one for transductive learning since the integral starts from $\frac{1}{4N}$, which means that the largest complexity terms are taken into account. In Section 3, we have seen that for polynomial entropies with $q \geq 1$, the optimal convergence rate (which was of order $N^{-\frac{1}{1+q}}$ up to the logarithmic factor) was proved since the largest complexities were not in the integral entropy.

On the contrary, for $q < 1$, we can recover the same convergence rates under the assumption $H(u,\mathcal{F},\bar{\mathbb{P}}) \leq C' h_q(u)$ for any $u > 0$, as under the polynomial bracketing entropy assumption. The following section deals with a special case of the case $q = 0$.

[12] Proof at the end of Section 6.11.

4.3. **Application to VC-classes.** At first sight, it is not obvious that Theorem 4.3 gives a tighter bound than Lemma 4.2 applied to $(\mathcal{S}_1, \mathcal{S}_2) = (\mathcal{N}, \mathcal{N})$. We will see in this section that for VC-classes, the two bounds gives the same convergence rate for the ERM-classifier, except when we have no margin assumption. In this last case, the chained result allows to get rid of a logarithmic factor.

Let us consider the binary classification setting: $\mathcal{Y} = \{0; 1\}$. Introduce the shattering number $N(X_1^{2N}) \triangleq \big|\big\{[f(X_k)]_{k=1}^{2N} : f \in \mathcal{F}\big\}\big| = H(u, \mathcal{F}, \bar{\bar{\mathbb{P}}})$ for any $u < \frac{1}{2N}$. Let V be the VC-dimension of the set $\mathcal{F}$

$$V \triangleq \max\big\{|A| : A \in \mathcal{X}^{2N} \text{ such that } |\{A \cap f^{-1}(1) : f \in \mathcal{F}\}| = 2^{|A|}\big\}.$$

The empirical entropies satisfy[13]

$$H(u, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \leq \left\{ \begin{array}{l} V \log\big(\frac{2Ne}{V}\big) \\ V \log\big(\frac{4e}{u}\big) \end{array} \right. .$$

Let $\hat{f}_{\mathrm{ERM}}$ be the minimizer of the empirical risk on the set $\mathcal{F}$ and $\tilde{f}'$ be the minimizer on $\mathcal{F}$ of either r' or R. From Lemma 4.2, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$, we have

$$r'(\hat{f}_{\mathrm{ERM}}) \leq \inf_{f \in \mathcal{F}} \left\{ r'(f) + 4\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, f}\big[V \log\big(\frac{2eN}{V}\big) + \frac{1}{2}\log(\epsilon^{-1})\big]}{N}} \right\},$$

and consequently, after some standard computations:

$$\mathbb{P}^{\otimes N} R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}) \leq 4\sqrt{\frac{V \mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, \tilde{f}}}{N} \log\big(\tfrac{2eN}{V}\big)} + 2\sqrt{\frac{2\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}}, \tilde{f}}}{N}}. \tag{4.4}$$

To compare, from Theorem 4.3, we obtain

Corollary 4.5. *For any $\epsilon > 0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1 - \epsilon$, we have*

$$r'(\hat{f}_{ERM}) \leq \inf_{f \in \mathcal{F}} \left\{ r'(f) + 47\sqrt{\frac{(V+1)\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, f}}{N} \log\big(\frac{8e}{\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, f}}\big)} + 34\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, f} \log(\epsilon^{-1})}{N}} \right\}$$

and, consequently,

$$\begin{aligned} \mathbb{P}^{\otimes N} R(\hat{f}_{ERM}) - R(\tilde{f}) \quad \leq \quad & 47\sqrt{\frac{(V+1)\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}}}{N} \log\left(\frac{8e}{\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}}}\right)} \\ & + 34\sqrt{\frac{\mathbb{P}^{\otimes 2N} \bar{\bar{\mathbb{P}}}_{\hat{f}_{ERM}, \tilde{f}}}{N}}. \end{aligned} \tag{4.5}$$

Proof. See Section 6.13. □

As a consequence, we obtain

Corollary 4.6. *Under assumption* (MA2)*, for any set $\mathcal{F}$ of VC-dimension V, the ERM-classifier satisfies*

$$\mathbb{P}^{\otimes N} R(\hat{f}_{ERM}) - R(\tilde{f}) \leq \breve{C} \left\{ \begin{array}{ll} \big(\frac{V}{N} \log N\big)^{\frac{\kappa}{2\kappa - 1}} & \text{when } 1 \leq \kappa < +\infty \\ \sqrt{\frac{V}{N}} & \text{when } \kappa = +\infty \end{array} \right. .$$

Proof. See section 6.14. □

[13]The first inequality is well-known consequence of Sauer's lemma; the second one comes from Haussler's formula ([7]), which asserts that for any $u > 0$, $H(u, \mathcal{F}, \bar{\bar{\mathbb{P}}}) \leq V \log\big(\frac{2e}{u}\big) + \log[e(V+1)]$.

Remark 4.2. This is an improvement of Massart and Nédélec results [12, Corollary 2.2] to the extent we do not have an extra additive term $R(\tilde{f}) - R(f^*)$, where f^* is the Bayes classifier. The second part of the corollary is a well-known result which is given with a simple proof in [5, p.31].

Remark 4.3. By comparing Inequalities (4.4) and (4.5), we see that the constants in chained inequalities are not satisfactory. The gap between the upper bound (6.23) and the lower bound (see Theorem 5.2) is the factor 8×83 ! We do not know how to chain inequalities with significantly tighter constants.

5. Assouad's lemma

Definition 5.1. Let $m \in \mathbb{N}^*$, $w \in]0;1]$, $b \in]0;1]$ and $b' \in]0;1]$. A (m, w, b, b')-*hypercube* of probability distributions is a family

$$\big\{\mathbb{P}_{\vec{\sigma}} \in \mathcal{M}_+^1(\mathcal{Z}) : \vec{\sigma} \triangleq (\sigma_1, \dots, \sigma_m) \in \{-1;+1\}^m\big\}$$

of 2^m probability distributions having the same first marginal:

$$\mathbb{P}_{\vec{\sigma}}(dX) = \mathbb{P}_{(+1,\dots,+1)}(dX) \triangleq \mu \text{ for any } \vec{\sigma} \in \{-1;+1\}^m,$$

and such that there exists a partition $\mathcal{X}_0, \dots, \mathcal{X}_m$ of $\mathcal{X}$ satisfying

- for any $j \in \{1, \dots, m\}$, we have $\mu(\mathcal{X}_j) = w$
- for any $j \in \{0, \dots, m\}$, for any $X \in \mathcal{X}_j$, we have

$$\mathbb{P}_{\vec{\sigma}}(Y = 1|X) = \tfrac{1+\sigma_j \xi(X)}{2} = 1 - \mathbb{P}_{\vec{\sigma}}(Y = 0|X),$$

where $\sigma_0 \triangleq 1$ and $\xi : \mathcal{X} \to [0;1]$ is such that for any $j \in \{1, \dots, m\}$,

$$\left\{ \begin{array}{rcl} b & = & \sqrt{1 - \big(\mu\big[\sqrt{1-\xi^2(X)}\,\big|X \in \mathcal{X}_j\big]\big)^2} \\ b' & = & \mu\big[\xi(X)|X \in \mathcal{X}_j\big] \end{array} \right. .$$

When ξ is constant on $\mathcal{X}_j, j = 1, \dots, m$ (which implies $\xi \equiv b' = b$ on $\mathcal{X} - \mathcal{X}_0$), the hypercube will be said a (m, w, b)-constant hypercube. The hypercube will be said noiseless when $\xi \equiv 1$ on $\mathcal{X}_0$.

The following lemma is Assouad's lemma adapted to the classification framework.

Lemma 5.1. *If a set $\mathcal{P}$ of probability distributions contains a (m, w, b, b')-hypercube, then for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we have*

$$\sup_{\mathbb{P} \in \mathcal{P}} \big\{\mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f_{\mathbb{P}}^*)\big\} \geq \tfrac{1-b\sqrt{Nw}}{2} mwb'.$$

Proof. See Section 6.15. □

Lemma 5.1 gives a very simple strategy to obtain a lower bound for a given set $\mathcal{P}$ of probability distributions: it consists in looking for the (m, w, b, b')-hypercube which is contained in the set $\mathcal{P}$ and for which $\frac{1-b\sqrt{Nw}}{2} mwb'$ is maximized.

In general, the order of the bound is given by the quantity mwb' and w, b are taken such that $\sqrt{Nw}b = \text{Cst} < 1$. To obtain this order, we do not need the sophisticated computations detailed in the proof of the lemma. We can use two well-known lemmas instead (Birgé's lemma and Huber's lemma) as it is proved in Appendix E.

Lemma 5.1 implies lower bounds for VC-classes with decent constants. The following result is to be compared with Theorems 14.1 and 14.5 in [5].

Theorem 5.2. *For any model $\mathcal{F}$, define $\mathcal{P}_L$ as the set of probability distributions such that $\inf_{f\in\mathcal{F}} R_{\mathbb{P}}(f) = L$ for a fixed $L \in [0; 1/2]$.*

- *When $L = 0$:*

for any classification model $\mathcal{F}$ of VC-dimension $V \geq 2$, for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_0} \left\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \right\} \geq \begin{cases} \frac{V-1}{2e(N+1)} & \text{when } N \geq (V-2) \vee 1 \\ \frac{1}{2}\big(1 - \frac{1}{V-1}\big)^N \end{cases} .$$

- *When $0 < L \leq 1/2$:*

for any classification model $\mathcal{F}$ of VC-dimension $V \geq 2$, for any measurable estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_L} \left\{ \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \right\} \geq \begin{cases} \sqrt{\frac{L(V-1)}{32N}} \vee \frac{V-1}{16N} & \text{when } \frac{(1-2L)^2 N}{V-1} \geq \frac{1}{4} \\ (\frac{1}{2} - L)\sqrt{\frac{L}{2}} & \text{otherwise} \end{cases} .$$

Proof. See Appendix F. □

Remark 5.1. It is a well known result that, when $\inf_{f\in\mathcal{F}} R_{\mathbb{P}}(f)$ is of order $1/N$ and when the complexity of the class is not too high, there exists an estimator such that $\mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - \inf_{f\in\mathcal{F}} R_{\mathbb{P}}(f) = \mathrm{O}\big(\frac{1}{N}\big)$. The previous theorem gives a corresponding lower bound.

6. Proofs

6.1. **Proof of Lemma 3.1.** Let $T_1(\pi) \triangleq -\log \pi \exp(-\lambda \Delta R)$ and

$$T_2(\pi) \triangleq 0 \vee \log \pi \exp\Big(8.2\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}} - \lambda\Delta R\Big).$$

We start with the following lemma.

Lemma 6.1. *For any $\epsilon > 0$ and $0 < \lambda \leq 0.19N$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - \epsilon$, we have $\pi_{-\lambda r} R \leq \frac{C}{\lambda}\big[T_1(\pi) + T_2(\pi) + \log(4\epsilon^{-1})\big]$.*

Proof. Taking $\chi = \frac{1}{2}$ and $\gamma = \frac{1}{2}$ in Theorem 2.2, we get

$$\begin{aligned} \pi_{-\lambda r}\Delta R &\leq \pi_{-\frac{\lambda}{2}R}\Delta R + \tfrac{2}{\lambda}\Big[16 \log \pi_{-\lambda R} \exp\Big(\tfrac{8.2\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) + 5\log(4\epsilon^{-1})\Big] \\ &\leq -\tfrac{2}{\lambda}\log \pi \exp\big(-\tfrac{\lambda}{2}\Delta R\big) + \tfrac{32}{\lambda}\log \pi_{-\lambda R}\exp\Big(\tfrac{8.2\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) + \tfrac{10}{\lambda}\log(4\epsilon^{-1}) \\ &\leq -\tfrac{34}{\lambda}\log \pi \exp(-\lambda\Delta R) + \tfrac{32}{\lambda}\log \pi \exp\Big(-\lambda\Delta R + \tfrac{8.2\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}}\Big) \\ &\qquad + \tfrac{10}{\lambda}\log(4\epsilon^{-1}). \end{aligned}$$

□

Remark 6.1. In order to explain the assumptions used in Lemma 3.1, let us give upper bounds for the quantities T_1 and T_2 using the strong complexity and margin Assumptions *(CA1)* and *(MA3)* for a well chosen distribution π. Under Assumption *(CA1)* (which is equivalent to Assumption *(CA3)*), there exists a distribution $\pi^{(t)}$ such that for any $f' \in \mathcal{F}$, $\pi^{(t)}\big(\mathbb{P}_{\cdot,f'} \leq t\big) \geq e^{-C' t^{-q}}$.

For any $0 < t < 1$, we have

$$\begin{aligned} T_1\Big[\pi^{\left(c'' t^{1/\kappa}\right)}\Big] &\leq -\log\Big[\pi^{\left(c'' t^{1/\kappa}\right)}(\Delta R \leq t) e^{-\lambda t}\Big] && \text{(by Markov's inequality)} \\ &\leq -\log\Big[\pi^{\left(c'' t^{1/\kappa}\right)}(\mathbb{P}_{\cdot,\tilde{f}} \leq c'' t^{\frac{1}{\kappa}})\Big] + \lambda t && \big(\text{according to } \textit{(MA3)}\big) \\ &\leq C' c''^{-q} t^{-\frac{q}{\kappa}} + \lambda t && \big(\text{by definition of } \pi^{(t)}\big). \end{aligned}$$

Assumption *(MA2)* (recall that *(MA3)* $\Rightarrow$ *(MA2)*) implies that for any $\lambda > 0$

$$\begin{aligned} 8.2\tfrac{\lambda^2}{N}\mathbb{P}_{\cdot,\tilde{f}} - \lambda\Delta R &\leq 8.2C''\tfrac{\lambda^2}{N}(\Delta R)^{\frac{1}{\kappa}} - \lambda\Delta R \\ &\leq \lambda \sup_{x\geq 0}\Big\{8.2C''\tfrac{\lambda}{N}x^{\frac{1}{\kappa}} - x\Big\} \\ &= (\kappa-1)\lambda\Big(\tfrac{8.2C''\lambda}{\kappa N}\Big)^{\frac{\kappa}{\kappa-1}}, \end{aligned}$$

hence $T_2(\pi) \leq \breve{C}\lambda\big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}}$ for any distribution π and a constant $\breve{C} > 0$ depending on C'' and κ. (Note that for the limit case $\kappa = 1$, we have $T_2 = 0$ for any $\lambda \leq \frac{N}{8.2C''}$.) Therefore, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1 - 4\epsilon$, we have

$$\big\{\pi^{\left(c''t^{1/\kappa}\right)}\big\}_{-\lambda r}R - R(\tilde{f}) \leq \breve{C}\Big[t + \frac{t^{-\frac{q}{\kappa}} + \log(\epsilon^{-1})}{\lambda} + \Big(\frac{\lambda}{N}\Big)^{\frac{\kappa}{\kappa-1}}\Big],$$

where the constant $\breve{C} > 0$ depends on C', c'' and κ. The sum $t + \frac{t^{-\frac{q}{\kappa}}}{\lambda} + \big(\frac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}}$ has the minimal order $N^{-\frac{\kappa}{2\kappa-1+q}}$ when λ has the order of $N^{\frac{\kappa+q}{2\kappa-1+q}}$ and t has the order of $N^{-\frac{\kappa}{2\kappa-1+q}}$. This computation explains the choice of Assumptions (3.2) and (3.3).

From inequality (3.2), we have $T_1(\pi) \leq \breve{C}N^{\frac{q}{2\kappa-1+q}} + \lambda_N N^{-\frac{\kappa}{2\kappa-1+q}}$. From Assumption *(MA2)*, we have seen in the previous remark that $T_2(\pi) \leq \breve{C}\lambda_N\big(\frac{\lambda_N}{N}\big)^{\frac{\kappa}{\kappa-1}}$. From inequality (3.3), we obtain the desired convergence rate.

Now let us prove the sharper result: inequality (3.4). Let $a(\lambda) \triangleq \frac{\lambda}{N}g\big(\frac{\lambda}{N}\big)$. From Theorem 6.2 in [1] and the same computations as for the quantity $\mathcal{L}''$ in Section 9.12 of [1] to upper bound $-\log \pi \exp\big\{-\lambda[r - r(\tilde{f})]\big\}$, we obtain :

Lemma 6.2. *For any $\epsilon > 0$, $\lambda > 0$ and $\xi > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1 - 2\epsilon$, with $\pi_{-\lambda r}$-probability at least $1 - \epsilon$, we have*

$$\Delta R \leq a(\lambda)\mathbb{P}_{\cdot,\tilde{f}} + \frac{-\log \pi \exp\big\{-\lambda\Delta R - \lambda a\big(\frac{\lambda}{\xi}\big)\mathbb{P}_{\cdot,\tilde{f}}\big\} + (2+\xi)\log(\epsilon^{-1})}{\lambda}.$$

Taking $\xi = 1$ and $\lambda = \lambda_N$, using the margin assumption $\mathbb{P}_{\cdot,\tilde{f}} \leq C''(\Delta R)^{\frac{1}{\kappa}}$ and noting that $a(\lambda_N) \leq g(\breve{C}_4)\frac{\lambda}{N}$, we get

$$\Delta R \leq \breve{C}\tfrac{\lambda_N}{N}(\Delta R)^{\frac{1}{\kappa}} + \tfrac{3\log(\epsilon^{-1})}{\lambda_N} + \sup_{x\geq 0}\big\{\breve{C}\tfrac{\lambda_N}{N}x^{\frac{1}{\kappa}} - x\big\} - \tfrac{\log \pi \exp(-2\lambda_N\Delta R)}{\lambda_N},$$

where the constant $\breve{C} > 0$ only depends on C'' and $\breve{C}_4$. Now from the same computations as in Remark 6.1, when the Inequalities (3.2) and (3.3) hold, we get

$$\Delta R \leq \breve{C}\Big[\frac{\lambda_N}{N}(\Delta R)^{\frac{1}{\kappa}} + N^{-\frac{\kappa}{2\kappa-1+q}}\Big] + \frac{3\log(\epsilon^{-1})}{\lambda_N}.$$

We obtain successively

$$\Delta R \leq \breve{C}\big[N^{\frac{1-\kappa}{2\kappa-1+q}}(\Delta R)^{\frac{1}{\kappa}} + \log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}\big]$$

and

$$\Delta R \leq \breve{C}\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}.$$

6.2. **Proof of Theorem 3.3.** A standard idea to prove lower bounds is to consider an adequate hypercube of probability distributions and to use Assouad's lemma (see Section 5 for the definition of the hypercube of distributions).

Consider a $\big(N^{\frac{q}{2\kappa-1+q}}, N^{-\frac{1+q}{2\kappa-1+q}}, aN^{-\frac{\kappa-1}{2\kappa-1+q}}\big)$-constant noiseless hypercube of probability distributions $\big\{\mathbb{P}_{\vec{\sigma}} : \vec{\sigma} \in \{-1;+1\}^m\big\}$, where $a > 0$ is a constant which will be chosen later.

Had we replaced Assumption *(MA3)* with Assumption *(MA2)* in Theorem 3.3, the result would have been a direct consequence of Lemma 5.1 applied to this hypercube with $a = \frac{1}{2}$.

In this proof, we will not apply Assouad's lemma but Fano's lemma since Assumption *(MA3)* is not satisfied by the whole hypercube. First let us state the following classical result on the hypercube which is a refined version of Varshamov-Gilbert bound (1962).

Lemma 6.3 (Huber,[8, p.256])**.** *Let $\delta(\Sigma, \Sigma')$ denote the Hamming distance between Σ and Σ' in $\{-1,1\}^m$: $\delta(\Sigma, \Sigma') \triangleq \sum_{i=1}^m \mathbb{1}_{\Sigma_i \neq \Sigma'_i}$. There exists a subset $\mathcal{S}$ of the hypercube $\{-1,1\}^m$ such that*

- *for any $\Sigma \neq \Sigma'$ in $\mathcal{S}$, we have $\delta(\Sigma, \Sigma') \geq \frac{m}{4}$*
- $\log|\mathcal{S}| \geq \frac{m}{8}$.

Proof. It suffices to upper bound the number of points in the ball centered at a point σ of the hypercube and of radius $\frac{m}{4}$. Consider the uniform distribution $\nu(d\Sigma)$ on the hypercube $\{-1,1\}^m$. Specifically, we have

$$\nu\bigg(\delta(\Sigma,\sigma) \leq \frac{m}{4}\bigg) \leq \nu e^{\frac{m}{4}-\delta(\Sigma,\sigma)} = e^{\frac{m}{4}}\bigg(\nu e^{-\mathbb{1}_{\Sigma_i \neq \sigma_i}}\bigg)^m = \bigg(\frac{e^{\frac{1}{4}}(1+e^{-1})}{2}\bigg)^m \leq e^{-\frac{m}{8}},$$

which leads to the desired result. □

Let $\mathcal{S} \subset \{-1;+1\}^m$ such that $|\mathcal{S}| = \lfloor e^{\frac{m}{8}} \rfloor$ and for any $\Sigma \neq \Sigma'$ in $\mathcal{S}$, $\delta(\Sigma, \Sigma') \geq \frac{m}{4}$. From inequality (5.1) in [2], Birgé's version of Fano's lemma can be stated as

Lemma 6.4. *Given a non-trivial (i.e. cardinal ≥ 2) finite family $\mathcal{D}$ of probability measures on some measurable set (E, ξ) and a random variable $\bar{X}$ with an unknown distribution in the family, we have*

$$\inf_{\hat{T}} \sup_{\mathbb{P}\in\mathcal{D}} \mathbb{P}\big[\hat{T}(\bar{X}) \neq \mathbb{P}\big] \geq 0.36 \wedge \bigg(1 - \frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\bigg),$$

where $K_{\mathcal{D}} \triangleq \inf_{\mathbb{P}\in\mathcal{D}} \sum_{\mathbb{Q}\neq\mathbb{P}} K(\mathbb{Q}, \mathbb{P})$ and the infimum is taken over all measurable (possibly randomized) estimators based on $\bar{X}$ with values in the finite set $\mathcal{D}$.

Define $\mathcal{D}' \triangleq \big\{\mathbb{P}_{\vec{\sigma}} : \vec{\sigma} \in \mathcal{S}\big\}$. Let us apply Birgé's lemma to the set of probability distributions

$$\mathcal{D} \triangleq \big\{\mathbb{P}^{\otimes N} : \mathbb{P} \in \mathcal{D}'\big\}.$$

With any estimator $\hat{f} : \mathcal{Z}^N \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, we can associate an estimator $\hat{T} : \mathcal{Z}^N \to \mathcal{D}$ defined as $\hat{T}(Z_1^N) = \mathbb{P}^{\otimes N}$, where $\mathbb{P} \in \mathcal{D}'$ minimizes $\mu[\xi(X)\mathbb{1}_{f^*_{\mathbb{P}}(X) \neq \hat{f}(Z_1^N)(X)}]$, where $f^*_{\mathbb{P}}$ denotes the Bayes classifier associated with the distribution $\mathbb{P}$.

By Birgé's lemma, we have $\sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{Q}\big[\hat{T}(Z_1^N) \neq \mathbb{Q}\big] \geq 0.36 \wedge \Big(1 - \frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)$. Now, when $\hat{T}(Z_1^N) \neq \mathbb{P}^{\otimes N}$, we have $\mu(f^*_{\mathbb{P}} \neq \hat{f}) \geq \frac{m}{8}w$, hence $R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \geq \frac{m}{8}w\beta$.

Therefore, we get

$$\sup_{\mathbb{P}^{\otimes N}\in\mathcal{D}} \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \geq \frac{m}{8} w\beta \Big[0.36 \wedge \Big(1 - \frac{K_{\mathcal{D}}}{|\mathcal{D}| \log |\mathcal{D}|} \Big) \Big]. \tag{6.1}$$

For any $\mathbb{P} \neq \mathbb{Q} \in \mathcal{D}$, we have $K(\mathbb{P},\mathbb{Q}) \leq Nmw\beta \log\big(\frac{1+\beta}{1-\beta}\big)$. Since we have $|\mathcal{D}| = \lfloor e^{\frac{m}{8}} \rfloor$, we obtain $\frac{K_{\mathcal{D}}}{|\mathcal{D}| \log |\mathcal{D}|} \leq \frac{1}{\log \lfloor e^{\frac{m}{8}} \rfloor} Nmw\beta \log\big(\frac{1+\beta}{1-\beta}\big) \leq 20 Nw\beta^2$ for m large enough and β small enough. In our case, we have $m = N^{\frac{q}{2\kappa-1+q}}$, $w = N^{-\frac{1+q}{2\kappa-1+q}}$ and $\beta = aN^{-\frac{\kappa-1}{2\kappa-1+q}}$. So for N large enough, we have $\frac{K_{\mathcal{D}}}{|\mathcal{D}| \log |\mathcal{D}|} \leq 20a^2$. Let us choose a such that $20a^2 = 0.64$. We obtain $\sup_{\mathbb{P}\in\mathcal{D}'} \mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}}) \geq 0.008 N^{-\frac{\kappa}{2\kappa-1+q}}$ for N sufficiently large.

Finally it remains to check that the set of distributions $\mathcal{D}'$ is included in $\mathcal{P}$. For any $\mathbb{P} \in \mathcal{D}'$, the complexity Assumption *(CA2)* is satisfied since

- for $u < mw$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) \leq \log |\mathcal{F}| \leq Cu^{-q}$ for some constant $C > 0$.
- for $u \geq mw$, $H(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot}) = 0 \leq Cu^{-q}$.

For any $\mathbb{P} \in \mathcal{D}'$, the margin Assumption *(MA3)* is satisfied since for any functions $f \in \mathcal{F} - \{\tilde{f}\}$, $\mathbb{P}_{f,\tilde{f}}$ has the order of $mw = N^{-\frac{1}{2\kappa-1+q}}$ and $\Delta R(f)$ has the order of $mw\beta = aN^{-\frac{\kappa}{2\kappa-1+q}}$. The margin Assumption *(MA1)* also holds since we have

$$\mathbb{P}\left(0 < |\eta^*(x) - \tfrac{1}{2}| \leq t\right) = \begin{cases} 0 & \text{when } t < \beta \\ mw & \text{when } \beta \leq t < \frac{1}{2} \end{cases}$$

Remark 6.2. The proof also holds when $q = 0$. In this case, we take $m = 1$, $w = N^{-\frac{1}{2\kappa-1}}$ and $\beta = \sqrt{\frac{0.64}{20}} N^{-\frac{\kappa-1}{2\kappa-1}}$.

6.3. **Proof of Theorem 3.5.**

6.3.1. *First case:* $\log \pi^{-1}(\Delta R \leq x) = -C' \log x + C''' + o(x^s)$. Since we have[14]

$$\pi_{-\lambda R} \Delta R = \frac{C' + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda}, \tag{6.2}$$

from Theorem 2.2, for any $0 \leq \chi < 1$, we get

$$\begin{cases} \pi_{-\lambda r} R & \leq & \frac{C' + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}(\chi)}{\lambda} + \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \\ \pi_{-\lambda r} R & \geq & \frac{C' + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}(\chi)}{\lambda} - \frac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \end{cases}.$$

Taking $\chi = \sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}$, we obtain

$$\lambda \pi_{-\lambda r} R = C' + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}\Big(\sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}\Big). \tag{6.3}$$

First subcase: $\lambda = \mathrm{o}\big(N^{\frac{\kappa}{2\kappa-1}}\big)$. Assume that $\lambda = \underset{N\to+\infty}{\mathrm{o}}\big(N^{\frac{\kappa}{2\kappa-1}}\big)$. Then there exists $\gamma \in]0; \frac{1}{2}]$ such that $\gamma = \underset{N\to+\infty}{\mathrm{o}}(1)$ and $\lambda\big(\frac{\lambda}{\gamma N}\big)^{\frac{\kappa}{\kappa-1}} = \underset{N\to+\infty}{\mathrm{o}}(1)$. We have[15]

$$\log \pi_{-\lambda R} \exp\Big\{ C \frac{\lambda^2}{\gamma N} (\Delta R)^{\frac{1}{\kappa}} \Big\} = \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}\Big(\Big[\lambda \big(\frac{\lambda}{\gamma N}\big)^{\frac{\kappa}{\kappa-1}} \Big]^{\frac{(\kappa-1)C'}{\kappa C' + \kappa - 1}} \Big). \tag{6.4}$$

[14] See Appendix A.

[15] Proof in Appendix B.

Let $L \triangleq \log(e\epsilon^{-1})$. From Theorem 2.2, for any $0 < \gamma < \frac{1}{2}$ and $0 < \lambda \leq 0.39\,\gamma N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$\begin{aligned} K(\pi_{-\lambda r}, \pi_{-\lambda R}) &\leq \log \pi_{-\lambda R} \exp\left\{ C\tfrac{\lambda^2}{\gamma N}(\Delta R)^{\frac{1}{\kappa}} \right\} + C\gamma L \\ &= \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{N\to+\infty}{\mathrm{O}}\left(\left[\lambda \big(\tfrac{\lambda}{N}\big)^{\frac{\kappa}{\kappa-1}} \right]^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}} \gamma^{-\frac{\kappa C'}{\kappa C'+\kappa-1}} + \gamma L \right). \end{aligned}$$

Taking $\gamma = \lambda^{\frac{(2\kappa-1)C'}{2\kappa C'+\kappa-1}} N^{-\frac{\kappa C'}{2\kappa C'+\kappa-1}} L^{-\frac{\kappa C'+\kappa-1}{2\kappa C'+\kappa-1}}$, we obtain

(6.5)
$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) = \underset{N\to+\infty}{\mathrm{O}}\left(\lambda^{\frac{(2\kappa-1)C'}{2\kappa C'+\kappa-1}} N^{-\frac{\kappa C'}{2\kappa C'+\kappa-1}} L^{\frac{\kappa C'}{2\kappa C'+\kappa-1}} \right) + \underset{N\to+\infty}{\mathrm{o}}(\lambda^{-s}),$$

which, combined with equality (6.3), gives the desired result.

Second subcase: $\lambda = cN^{\frac{\kappa}{2\kappa-1}}$ for a small enough c.

Then the previous computations can be adapted and we obtain that for any $\beta > 0$ there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$ with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :

$$\tfrac{C'-\beta}{\lambda} \leq \pi_{-\lambda r} R \leq \tfrac{C'+\beta}{\lambda}.$$

6.3.2. *Second case:* $\log \pi^{-1}(\Delta R \leq x) = C' x^{-\frac{q}{\kappa}} + C''' + \mathrm{o}(1)$. Since we have[16]

$$\pi_{-\lambda R} \Delta R \underset{\lambda\to+\infty}{\sim} \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}, \tag{6.6}$$

from Theorem 2.2, for any $0 \leq \chi < 1$, we get

$$\begin{cases} \pi_{-\lambda r} R &\leq \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} \big[1 + \underset{N\to+\infty}{\mathrm{o}}(1) + \underset{N\to+\infty}{\mathrm{O}}(\chi)\big] + \tfrac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \\ \pi_{-\lambda r} R &\geq \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} \big[1 + \underset{N\to+\infty}{\mathrm{o}}(1) + \underset{N\to+\infty}{\mathrm{O}}(\chi)\big] - \tfrac{K(\pi_{-\lambda r}, \pi_{-\lambda R})}{\chi\lambda} \end{cases}.$$

Taking $\chi = \lambda^{-\frac{q}{2(\kappa+q)}} \sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}$, we obtain

$$\pi_{-\lambda r} R = \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} \Big[1 + \underset{N\to+\infty}{\mathrm{o}}(1) + \underset{N\to+\infty}{\mathrm{O}}\Big(\lambda^{-\frac{q}{2(\kappa+q)}} \sqrt{K(\pi_{-\lambda r}, \pi_{-\lambda R})}\Big)\Big].$$

First subcase: $\lambda = \mathrm{o}\big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\big)$. From Theorem 2.2, for any $0 < \lambda \leq 0.19\,N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have

$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) \leq \log \pi_{-\lambda R} \exp\left\{ C\tfrac{\lambda^2}{N}(\Delta R)^{\frac{1}{\kappa}} \right\} + C \log(e\epsilon^{-1}).$$

We can prove[17] that for any $\alpha \leq \breve{c}\lambda^{-\frac{\kappa-1}{\kappa+q}}$ for $\breve{c}$ small enough, we have

$$\log \pi_{-\lambda R} \exp\left\{ \lambda\alpha(\Delta R)^{\frac{1}{\kappa}} \right\} = \underset{\lambda\to+\infty}{\mathrm{O}}\Big(\lambda^{\frac{q}{\kappa+q}} \alpha \lambda^{\frac{\kappa-1}{\kappa+q}}\Big), \tag{6.7}$$

Let us assume that $\lambda = \underset{N\to+\infty}{\mathrm{o}}\Big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\Big)$. Then we have $\frac{\lambda}{N}\lambda^{\frac{\kappa-1}{\kappa+q}} = \underset{N\to+\infty}{\mathrm{o}}(1)$, hence

$$K(\pi_{-\lambda r}, \pi_{-\lambda R}) \leq \underset{N\to+\infty}{\mathrm{o}}\big(\lambda^{\frac{q}{\kappa+q}}\big) + C \log(e\epsilon^{-1}).$$

[16] See Appendix C.

[17] See Appendix D.

So we obtain that for $\lambda = \underset{N\to+\infty}{\mathrm{o}}\Big(N^{-\frac{\kappa+q}{2\kappa-1+q}}\Big)$,

$$\pi_{-\lambda r}R = \big(\tfrac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}\Big[1+\underset{N\to+\infty}{\mathrm{o}}(1)\Big].$$

Second subcase: $\lambda = cN^{-\frac{\kappa+q}{2\kappa-1+q}}$ for a small enough c.

Once more, the previous computations can be adapted in order to obtain that for any $\beta > 0$ there exist $c > 0$ and $N_0 > 0$ such that for any $N > N_0$ with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, :

$$\big(\tfrac{qC'-\beta}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}} \le \pi_{-\lambda r}R \le \big(\tfrac{qC'+\beta}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}.$$

6.4. **Proof of Theorem 3.6.** For any $0 \le j \le \log N$, introduce $\lambda_j \triangleq 0.19\sqrt{N}e^{\frac{j}{2}}$. Define $L \triangleq \log[\log(eN)\epsilon^{-1}]$. In [1, Section 3.4.2], an algorithm is proposed to choose the temperature of the standard Gibbs classifier. The associated generalization error is bounded by

$$\mathbb{G} \triangleq \min_{1\le j\le \log N}\left\{\pi_{-\lambda_{j-1}R}R + \frac{\sup_{0\le i\le j}\big\{\log \pi_{-\lambda_i R}\otimes\pi_{-\lambda_i R}\exp\big(\frac{C\lambda_i^2}{N}\mathbb{P}_{\cdot,\cdot}\big)\big\}}{\lambda_j} + C\frac{L}{\lambda_j}\right\}.$$

Under Assumptions *(MA3)* and *(CA1)*, for any $1 \le j \le \log N$ and $t > 0$, by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{G} &\le -\frac{\log \pi\exp(-\lambda_{j-1}R)}{\lambda_{j-1}} + \frac{\sup_{0\le i\le j}\big\{\log \pi_{-\lambda_i R}\exp\big(\frac{C\lambda_i^2}{N}\mathbb{P}_{\cdot,\tilde f}\big)\big\}}{\lambda_j} + C\frac{L}{\lambda_j}\\
&\le -\frac{\log \pi\exp(-\lambda_{j-1}R)}{\lambda_{j-1}} + \frac{\sup_{0\le i\le j}\big\{\log \pi\exp\big(-\lambda_i\Delta R + \frac{C\lambda_i^2}{N}(\Delta R)^{\frac{1}{\kappa}}\big)\big\}}{\lambda_j}\\
&\qquad + \sup_{0\le i\le j}\Big\{-\frac{\log \pi\exp(-\lambda_i\Delta R)}{\lambda_j}\Big\} + C\frac{L}{\lambda_j}\\
&\le R(\tilde f) - 2\sqrt{e}\frac{\log \pi\exp(-\lambda_j R)}{\lambda_j} + \frac{\sup_{0\le i\le j;x\ge 0}\big\{-\lambda_i x + \frac{C\lambda_i^2}{N}x^{\frac{1}{\kappa}}\big\}}{\lambda_j} + C\frac{L}{\lambda_j}\\
&\le R(\tilde f) - 2\sqrt{e}\frac{\log[\pi(\Delta R\le t)\exp(-\lambda_j t)]}{\lambda_j} + C\big(\tfrac{\lambda_j}{N}\big)^{\frac{\kappa}{\kappa-1}} + C\frac{L}{\lambda_j}\\
&\le R(\tilde f) + C\frac{h_q(t^{1/\kappa})}{\lambda_j} + Ct + C\frac{L}{\lambda_j}.
\end{aligned}$$

Taking j such that λ_j is of order $N^{\frac{\kappa+q}{2\kappa-1+q}}$ and t minimizing $C\frac{h_q(t^{1/\kappa})}{\lambda_j} + Ct$, we obtain the desired rates (the ones given in Theorems 3.2 and 3.4). So the algorithm is adaptive wrt the margin parameter κ.

6.5. **Proof of Theorem 3.7.** We will prove the result for a minimal net. It is easy to generalize it to almost minimal nets. Let $u > 0$. Let π be the uniform distribution on a minimal $(u, \mathcal{F}, \mathbb{P}_{\cdot,\cdot})$-net denoted $\mathcal{N}_u$. Let $\tilde f_u$ be the the nearest neighbour of $\tilde f$ in the net $\mathcal{N}_u$. Define $a(\lambda) \triangleq \frac{\lambda}{N}g\big(\frac{\lambda}{N}\big)$. From inequality (2.1) for $\big(\rho_2,\pi_2,\rho_1,\pi_1\big) = \big(\delta_{\hat f},\pi,\delta_{\tilde f_u},\delta_{\tilde f_u}\big)$, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, we have

$$R(\hat f) - R(\tilde f_u) + r(\tilde f_u) - r(\hat f) \le a(\lambda)\mathbb{P}_{\hat f,\tilde f_u} + \frac{H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(\epsilon^{-1})}{\lambda},$$

When $\hat f = \hat f_{\mathrm{ERM},u}$ minimizes the empirical risk over the net $\mathcal{N}_u$, we obtain

$$R(\hat f) - R(\tilde f_u) \le a(\lambda)\mathbb{P}_{\hat f,\tilde f_u} + \frac{H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(\epsilon^{-1})}{\lambda},$$

hence

$$R(\hat{f}) - R(\tilde{f}) \leq R(\tilde{f}_u) - R(\tilde{f}) + a(\lambda)\mathbb{P}_{\hat{f},\tilde{f}} + a(\lambda)\mathbb{P}_{\tilde{f},\tilde{f}_u} + \frac{H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(\epsilon^{-1})}{\lambda}.$$

Let $\breve{C} > 0$ denote a constant (possibly depending on c'', C'', C', $\breve{C}_1$ and $\breve{C}_2$) whose value may differ from line to line. For any $0 < \lambda \leq N$, we have

$$\begin{aligned} \Delta R(\hat{f}) &\leq \breve{C}\Big(\mathbb{P}^{\kappa}_{\tilde{f},\tilde{f}_u} + \tfrac{\lambda}{N}\Delta R^{\frac{1}{\kappa}}(\hat{f}) + \tfrac{\lambda}{N}\mathbb{P}_{\tilde{f},\tilde{f}_u} + \tfrac{u^{-q}+\log(\epsilon^{-1})}{\lambda}\Big) \\ &\leq \breve{C}\Big(\tfrac{\lambda}{N}\Delta R^{\frac{1}{\kappa}}(\hat{f}) + u^{\kappa} + \tfrac{\lambda}{N}u + \tfrac{u^{-q}+\log(\epsilon^{-1})}{\lambda}\Big). \end{aligned} \tag{6.8}$$

Let us take u and λ such that $\frac{\lambda}{N}u$, u^{κ} and $\frac{u^{-q}}{\lambda}$ have the same orders. This is realized when Inequalities (3.6) hold and $\lambda = N^{\frac{\kappa+q}{2\kappa-1+q}}$. We obtain

$$\Delta R(\hat{f}) \leq \breve{C}\big[N^{-\frac{\kappa-1}{2\kappa-1+q}}\Delta R^{\frac{1}{\kappa}}(\hat{f}) + \log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}\big].$$

Simple computations lead to

$$\Delta R(\hat{f}) \leq \breve{C}\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}$$

and, then, to $\mathbb{P}^{\otimes N}\Delta R(\hat{f}) \leq \breve{C}N^{-\frac{\kappa}{2\kappa-1+q}}$.

6.6. **Proof of Theorem 3.9.** The chaining idea comes from [6] and is well presented also, for instance, in [5, p.19-21]. Let $\partial(\rho_1,\rho_2) \triangleq \rho_2 R - \rho_1 R + \rho_1 r - \rho_2 r$. Let $u_k = u2^k$. Let $c_k \triangleq h_+(u_k)$. To shorten, denote $\pi_{i,k} \triangleq \pi_{f_i,u_k}$. Let K be the nonnegative integer such that $\frac{\mathbb{P}_{f_1,f_2}}{2} \leq u_K < \mathbb{P}_{f_1,f_2}$. The integer K exists as soon as $\mathbb{P}_{f_1,f_2} > u$. Let L' be the nonnegative integer such that $\frac{1}{2} \leq u_{L'} < 1$. Let $\lambda_1,\ldots,\lambda_{L'+1}$ be real positive parameters to be chosen. We apply inequality (2.1) for this $L'+1$ parameters and for $\pi_1 = \pi_2 = \pi$.

With $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-(L'+1)\epsilon$, we have

$$\begin{aligned} & \partial(\pi_{1,0},\pi_{2,0}) \\ = \;& \partial(\pi_{1,K},\pi_{2,K}) + \textstyle\sum_{k=1}^{K}\big\{\partial(\pi_{1,k-1},\pi_{1,k}) + \partial(\pi_{2,k},\pi_{2,k-1})\big\} \\ \leq \;& \tfrac{\lambda_{K+1}}{N}g\big(\tfrac{\lambda_{K+1}}{N}\big)4u_K + \tfrac{K(\pi_{1,K},\pi)+K(\pi_{2,K},\pi)+\log(\epsilon^{-1})}{\lambda_{K+1}} \\ & + \textstyle\sum_{k=1}^{K}\Big\{2\tfrac{\lambda_k}{N}g\big(\tfrac{\lambda_k}{N}\big)(u_{k-1}+u_k) + \tfrac{2\log(\epsilon^{-1})+\sum_{i=1}^{2}\sum_{k'=k-1}^{k}K(\pi_{i,k'},\pi)}{\lambda_k}\Big\} \\ \leq \;& \tfrac{\lambda_{K+1}}{N}g\big(\tfrac{\lambda_{K+1}}{N}\big)4u_K + \tfrac{2c_K+\log(\epsilon^{-1})}{\lambda_{K+1}} \\ & + \textstyle\sum_{k=1}^{K}\Big\{6\tfrac{\lambda_k}{N}g\big(\tfrac{\lambda_k}{N}\big)u_{k-1} + \tfrac{2\log(\epsilon^{-1})+4c_{k-1}}{\lambda_k}\Big\} \\ \leq \;& \textstyle\sum_{k=1}^{K+1}\Big\{6\tfrac{\lambda_k}{N}g\big(\tfrac{\lambda_k}{N}\big)u_{k-1} + \tfrac{2\log(\epsilon^{-1})+4c_{k-1}}{\lambda_k}\Big\}. \end{aligned}$$

Let us choose the λ_k's such that they do not depend on ϵ and they roughly minimize the RHS of the last bound. Taking $\lambda_k = \sqrt{\frac{4Nc_{k-1}}{3u_{k-1}}}$ for $k \geq 1$, we obtain

$$\begin{aligned} \partial(\pi_{1,0},\pi_{2,0}) &\leq \textstyle\sum_{k=1}^{K+1}\big[1+2g\big(\tfrac{\lambda_k}{N}\big)\big]\sqrt{\tfrac{12c_{k-1}u_{k-1}}{N}} + \sum_{k=1}^{K+1}\tfrac{2\log(\epsilon^{-1})}{\lambda_k} \\ &\leq \textstyle\sum_{k=1}^{K+1}\big[1+2g\big(\tfrac{\lambda_k}{N}\big)\big]\sqrt{\tfrac{12c_{k-1}u_{k-1}}{N}} + 2\log(\epsilon^{-1})\sqrt{\tfrac{3u}{4N}}\textstyle\sum_{k=1}^{K+1}\sqrt{2}^{k-1}. \end{aligned}$$

For any $k \in \{1, \dots, L'+1\}$, we have $\frac{\lambda_k}{N} \le \sqrt{\frac{4c_0}{3Nu}} \triangleq C_1$. Since $C_0 = 2\sqrt{3}[1+2g(C_1)]$, we obtain

$$\begin{aligned}\partial(\pi_{1,0},\pi_{2,0}) &\le \tfrac{2C_0}{\sqrt{N}} \textstyle\sum_{k=1}^{K+1}(u_{k-1}-u_{k-2})\sqrt{\frac{c_{k-1}}{u_{k-1}}} + \sqrt{\frac{6\mathbb{P}_{f_1,f_2}}{N}}\frac{\log(\epsilon^{-1})}{\sqrt{2}-1} \\ &\le \tfrac{2C_0}{\sqrt{N}} \textstyle\int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h_+(v)}{v}}dv + 6\sqrt{\frac{\mathbb{P}_{f_1,f_2}}{N}}\log(\epsilon^{-1}).\end{aligned}$$

Remark 6.3. We have used the "global bayesian entropy" $h(v) \triangleq \sup_{f\in\mathcal{F}} \log \pi^{-1}(B_{f,v})$ since it was convenient to have an (almost) optimal λ's which do not depend on the functions f_1 and f_2. Had we done a union bound on the parameters λ, we would have been able to make it depend on the functions f_1, f_2. Then the global bayesian entropy would have been replaced with the local ones $h(v,f_1)$ and $h(v,f_2)$ where $h(v,f) \triangleq \log \pi^{-1}(B_{f,v})$. In other words, the quantity $\partial(\pi_{1,0},\pi_{2,0})$ is mainly driven by the two integrals $\int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h(v,f_1)}{vN}}dv$ and $\int_{u/2}^{\mathbb{P}_{f_1,f_2}} \sqrt{\frac{h(v,f_2)}{vN}}dv$.

6.7. **Proof of Theorem 3.10.**

6.7.1. *First step: upper bounds due to the chaining technique.* We start with the following chained result which is slightly different from Theorem 3.9 to the extent that we chained functions belonging to covering nets instead of chaining balls. Had we been interested in results for packing nets, Theorem 3.9 applied to an appropriate prior distribution[18] would have been sufficient. Let $H(u) \triangleq H(u,\mathcal{F},\mathbb{P}_{\cdot,\cdot})$.

Theorem 6.5. *Let $u>0$, $\mathcal{N}$ a minimal $u-$covering net and $L \triangleq \frac{\log(2u^{-1})}{\log 2}$. We have*

- *for any $\epsilon>0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any $f_1, f_2 \in \mathcal{N}_u$,*

$$\begin{aligned}&R(f_2)-R(f_1)+r(f_1)-r(f_2) \\ &\qquad \le 8\sqrt{\tfrac{3}{N}}\textstyle\int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u}\sqrt{\frac{H(v)}{v}}dv + \frac{8}{3N}\int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u}\frac{H(v)}{v}dv \\ &\qquad\quad +17\sqrt{\tfrac{\log(3L\epsilon^{-1})}{N}}\sqrt{\mathbb{P}_{f_1,f_2}\vee u} + \tfrac{2L\log(3L\epsilon^{-1})}{3N},\end{aligned} \tag{6.9}$$

- *for any $\epsilon>0$, for any $f_1 \in \mathcal{N}_u$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any $f_2 \in \mathcal{N}_u$,*

$$\begin{aligned}&R(f_2)-R(f_1)+r(f_1)-r(f_2) \\ &\qquad \le 4\sqrt{\tfrac{3}{N}}\textstyle\int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u}\sqrt{\frac{H(v)}{v}}dv + \frac{4}{3N}\int_{u/2}^{\mathbb{P}_{f_1,f_2}\vee u}\frac{H(v)}{v}dv \\ &\qquad\quad +8.5\sqrt{\tfrac{\log(2L\epsilon^{-1})}{N}}\sqrt{\mathbb{P}_{f_1,f_2}\vee u} + \tfrac{L\log(2L\epsilon^{-1})}{3N}\end{aligned}$$

Proof. The proof is similar to the one of Theorem 3.9. Instead of chaining balls, we will chain on covering nets. Let $\partial(f_1,f_2) \triangleq R(f_2)-R(f_1)+r(f_1)-r(f_2)$, $u_k = u2^k$ and $c_k \triangleq H_+(u_k)$. Introduce $P \triangleq \mathbb{P}_{f_1,f_2}\vee u$ and let $0 \le K \le L'$ be integers such that $\frac{\mathbb{P}_{f_1,f_2}}{2} < u_K \le P$ and $\frac{1}{2} < u_{L'} \le 1$.

Consider the family $\big(\mathcal{N}_k\big)_{k=\{0,\dots,L'\}}$ of minimal nets of radius u_k. For any $(j,k) \in \{1,2\}\times\{0,\dots,L'\}$, introduce $f_{j,k} \in \operatorname{argmin}_{\mathcal{N}_{u2^k}} \mathbb{P}_{\cdot,f_j}$ a nearest neighbour of f_j in

[18] Let $\mathcal{N}_{\mathrm{p}}$ be a u-packing net. Using the notation of Section 6.6, an appropriate prior distribution is $\pi = \frac{1}{L'+1}\sum_{k=0}^{L'}\pi_k$, where π_k is the uniform distribution on a u_k-minimal packing net of the set $\mathcal{F}$ built using points in $\mathcal{N}_{\mathrm{p}}$. The log-cardinal of such a set is upper bounded by $H(u_{k-1},\mathcal{F},\mathbb{P}_{\cdot,\cdot})$, hence $h(u_k) \le H(u_{k-1},\mathcal{F},\mathbb{P}_{\cdot,\cdot}) + \log(L'+1)$.

$\mathcal{N}_{u2^k}$. Since $f_1, f_2 \in \mathcal{N}_u$, we have $f_{1,0} = f_1$ and $f_{2,0} = f_2$. Let π_k be the uniform distribution on the net $\mathcal{N}_k$.

Let $l \triangleq \log[(3L'+1)\epsilon^{-1}]$. By applying $3L'+1$ times inequality (2.3), we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any functions f_1, f_2 in $\mathcal{N}_u$, we have

(6.10)

$$\begin{aligned}
\partial(f_1,f_2) &= \partial(f_{1,K},f_{2,K}) + \textstyle\sum_{k=1}^K \big\{\partial(f_{1,k-1},f_{1,k}) + \partial(f_{2,k},f_{2,k-1})\big\} \\
&\le \sqrt{\tfrac{2[2H(u_K)+l]\mathbb{P}_{f_{1,K},f_{2,K}}}{N}} + \tfrac{2H(u_K)+l}{3N} \\
&\quad + \textstyle\sum_{k=1}^K \Big\{ \sqrt{\tfrac{2[H(u_{k-1})+H(u_k)+l]\mathbb{P}_{f_{1,k-1},f_{1,k}}}{N}} + \tfrac{H(u_{k-1})+H(u_k)+l}{3N} \\
&\qquad + \sqrt{\tfrac{2[H(u_{k-1})+H(u_k)+l]\mathbb{P}_{f_{2,k-1},f_{2,k}}}{N}} + \tfrac{H(u_{k-1})+H(u_k)+l}{3N} \Big\} \\
&\le 2\textstyle\sum_{k=1}^{K+1} \Big\{ \sqrt{\tfrac{6[2H(u_{k-1})+l]u_{k-1}}{N}} + \tfrac{2H(u_{k-1})+l}{3N} \Big\} \\
&\le 2\textstyle\sum_{k=1}^{K+1} \Big\{ \sqrt{\tfrac{12H(u_{k-1})u_{k-1}}{N}} + \sqrt{\tfrac{6lu_{k-1}}{N}} + \tfrac{2H(u_{k-1})}{3N} \Big\} + \tfrac{2(K+1)l}{3N} \\
&\le 2\sqrt{\tfrac{6lu}{N}}\tfrac{\sqrt{2}^{K+1}}{\sqrt{2}-1} + \tfrac{2(K+1)l}{3N} + 2\textstyle\sum_{k=1}^{K+1} \Big\{ \sqrt{\tfrac{12H(u_{k-1})u_{k-1}}{N}} + \tfrac{2H(u_{k-1})}{3N} \Big\}
\end{aligned}$$

Now the last sum can be upper bounded using integrals since the function $v \mapsto H(v)$ is non increasing on $\mathbb{R}_+^*$. We obtain

$$\partial(f_1,f_2) \le \tfrac{4}{\sqrt{2}-1}\sqrt{\tfrac{3lP}{N}} + \tfrac{2l}{3N}\tfrac{\log(\frac{2P}{u})}{\log 2} + 8\sqrt{\tfrac{3}{N}}\textstyle\int_{u/2}^P \sqrt{\tfrac{H(v)}{v}}dv + \tfrac{8}{3N}\int_{u/2}^P \tfrac{H(v)}{v}dv.$$

For the second part of Theorem 6.5, it suffices to modify slightly the previous argument. This time, the functions $f_{2,k}$ are defined as previously. The functions $f_{1,k}$ are defined as $f_{1,k} \triangleq f_1$. Therefore we have $\partial(f_{1,k-1}, f_{1,k}) = 0$, hence the modification of the constants. □

Consider that Assumption *(CA1)* holds. Let $c_q > 0$ such that for any $0 < u \le 1$, $\sum_{k=0}^{L'} 3e^{-c_q h_q(u_k)} \le 1$. In the previous proof, we used an uniform union bound over the $3L'+1$ inequalities coming from (2.3). If we are just interested in the order of the bounds, we can weight the inequalities associated with $\partial(f_{1,k-1}, f_{1,k})$ and $\partial(f_{2,k}, f_{2,k-1})$ with $e^{-c_q h_q(u_{k-1})}$ and those corresponding to $\partial(f_{2,k}, f_{2,k-1})$ with at least weight $e^{-c_q h_q(u_k)}$.

Then, in Inequalities (6.10), we may replace $2H(u_K)+l$ and $H(u_{k-1})+H(u_k)+l$ with respectively $2H(u_K)+c_q h_q(u_K)+\log(\epsilon^{-1})$ and $H(u_{k-1})+H(u_k)+c_q h_q(u_{k-1})+\log(\epsilon^{-1})$, so that we obtain

(6.11)

$$\begin{aligned}
\partial(f_1,f_2) &\le \breve{C}\textstyle\sum_{k=1}^{K+1} \Big\{ \sqrt{\tfrac{[h_q(u_{k-1})+\log(\epsilon^{-1})]u_{k-1}}{N}} + \tfrac{h_q(u_{k-1})+\log(\epsilon^{-1})}{N} \Big\} \\
&\le \breve{C}\sqrt{\tfrac{P\log(\epsilon^{-1})}{N}} + \breve{C}\tfrac{\log(eu^{-1})\log(\epsilon^{-1})}{N} + \breve{C}\textstyle\int_{u/2}^P \Big(\sqrt{\tfrac{h_q(v)}{Nv}} + \tfrac{h_q(v)}{Nv}\Big)dv.
\end{aligned}$$

Corollary 6.6. *Let $\mathcal{N}$ denote a minimal $u-$net, where u is a positive real. Define $\mathcal{U} \triangleq \sup_{f:\mathbb{P}_{f,\tilde{f}} \le u} \big\{R(f) - R(\tilde{f})\big\}$. Introduce a function $\tilde{f}_{\mathcal{N}} \in \mathcal{N}$ such that $\mathbb{P}_{\tilde{f}_{\mathcal{N}},\tilde{f}} \le u$. Let $\gamma_u :]u;1] \to \mathbb{R}$ and $\Gamma_u :]u;1] \to \mathbb{R}$ be non decreasing concave functions respectively upper bounding the functions $\int_{\frac{u}{2}}^{\cdot} \sqrt{\frac{H(v)}{v}}dv$ and $\int_{\frac{u}{2}}^{\cdot} \frac{H(v)}{v}dv$..*

For any $\epsilon > 0$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any function $f \in \mathcal{N}$, we have

$$\begin{aligned} R(f) - R(\tilde{f}) \quad &\leq \quad r(f) - r(\tilde{f}_{\mathcal{N}}) + \tfrac{\breve{C}}{\sqrt{N}}\Big\{\gamma_u(\mathbb{P}_{f,\tilde{f}} + u) + \sqrt{(\mathbb{P}_{f,\tilde{f}} + u)\log(\epsilon^{-1})}\Big\} \\ &\quad + \tfrac{\breve{C}}{N}\Big\{\Gamma_u\big(\mathbb{P}_{f,\tilde{f}} + u\big) + \log(eu^{-1})\log(\epsilon^{-1})\Big\} + \mathcal{U} \end{aligned}$$

Consequently, for any $\rho \in \mathcal{M}_+^1(\mathcal{N})$, we have

$$\begin{aligned} \rho R - R(\tilde{f}) \quad &\leq \quad \rho r - r(\tilde{f}_{\mathcal{N}}) + \tfrac{\breve{C}}{\sqrt{N}}\Big\{\gamma_u(\rho\mathbb{P}_{\cdot,\tilde{f}} + u) + \sqrt{(\rho\mathbb{P}_{\cdot,\tilde{f}} + u)\log(\epsilon^{-1})}\Big\} \\ &\quad + \tfrac{\breve{C}}{N}\Big\{\Gamma_u\big(\rho\mathbb{P}_{\cdot,\tilde{f}} + u\big) + \log(eu^{-1})\log(\epsilon^{-1})\Big\} + \mathcal{U} \end{aligned}$$

Proof. The first inequality comes mainly from inequalities (6.11) and the decomposition: $R(f) - R(\tilde{f}) = R(f) - R(\tilde{f}_{\mathcal{N}}) + R(\tilde{f}_{\mathcal{N}}) - R(\tilde{f})$. The second inequality is then deduced from Jensen's inequality. □

6.7.2. *Second step: determining the radius of the nets.* Corollary 6.6 implies that for any $\epsilon > 0$, for any classifier $\hat{f}$ minimizing the empirical risk over the net $\mathcal{N}$, with $\big(\mathbb{P}^{\otimes N}\big)_*$-probability at least $1-\epsilon$, we have

$$\begin{aligned} \Delta R(\hat{f}) \quad &\leq \quad \mathcal{U} + \tfrac{\breve{C}}{\sqrt{N}}\Big\{\gamma_u\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big) + \sqrt{\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big)\log(\epsilon^{-1})}\Big\} \\ &\quad + \tfrac{\breve{C}}{N}\Big\{\Gamma_u\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big) + \log(eu^{-1})\log(\epsilon^{-1})\Big\}. \end{aligned} \tag{6.12}$$

Now we have

$$\mathcal{U} \leq \begin{cases} \breve{C}u^{\kappa} & \text{under Assumption } \textit{(MA3)} \\ 2u & \text{in any case} \end{cases},$$

and we can take

$$\gamma_u(x) \triangleq \begin{cases} \breve{C}\sqrt{\log(e^2x^{-1})x} & \text{under Assumption } \textit{(CA1)} \text{ for } q = 0 \\ \breve{C}x^{\frac{1-q}{2}} & \text{under Assumption } \textit{(CA1)} \text{ for } 0 < q < 1 \\ \breve{C}\log\big(\frac{2x}{u}\big) & \text{under Assumption } \textit{(CA1)} \text{ for } q = 1 \\ \breve{C}u^{\frac{1-q}{2}} & \text{under Assumption } \textit{(CA1)} \text{ for } q > 1 \end{cases}$$

and

$$\Gamma_u(x) \triangleq \begin{cases} \breve{C}[\log(eu^{-1})]^2 & \text{under Assumption } \textit{(CA1)} \text{ for } q = 0 \\ \breve{C}u^{-q} & \text{under Assumption } \textit{(CA1)} \text{ for } q > 0 \end{cases}.$$

Then we have eight cases corresponding to the different complexity and margin assumptions. When we have $q > 0$, inequality (6.12) implies

$$\Delta R(\hat{f}) \leq \mathcal{U} + \breve{C}\tfrac{\log(e\epsilon^{-1})}{\sqrt{N}}\gamma_u\big(\mathbb{P}_{\hat{f},\tilde{f}} + u\big) + \breve{C}\tfrac{\log(e\epsilon^{-1})}{N}u^{-q}.$$

Under Assumptions (MA2) *and* (CA1) *for $q = 0$*

Let $\Delta \triangleq \Delta R(\hat{f})$ to shorten. Inequality (6.12) becomes

$$\Delta \leq \breve{C}\Big[\log(e\epsilon^{-1})N^{-\frac{1}{2}}\Big(\sqrt{\log(e^2\Delta^{-\frac{1}{\kappa}})}\Delta^{\frac{1}{2\kappa}} + \sqrt{\log(e^2u^{-1})u}\Big) + u + \tfrac{[\log(eu^{-1})]^2}{N}\Big].$$

We obtain $\Delta \leq \breve{C}\log(e\epsilon^{-1})(\log[eN^{1/\kappa}])^{\frac{\kappa}{2\kappa-1}}N^{-\frac{\kappa}{2\kappa-1}}$ when[19]

$$\sqrt{\tfrac{\log(e^2u^{-1})u}{N}} + u + \tfrac{[\log(eu^{-1})]^2}{N} \leq \breve{C}(\log[eN^{1/\kappa}])^{\frac{\kappa}{2\kappa-1}}N^{-\frac{\kappa}{2\kappa-1}},$$

[19]We use $\log[eN^{1/\kappa}]$ since the logarithmic factor disappears for $\kappa = +\infty$. For $\kappa < +\infty$, the factor $\log[eN^{1/\kappa}]$ can be simplified into $\log N$ for $N \geq 2$.

hence when there exists $\check{C}_1, \check{C}_2 > 0$ such that

$$\exp\Big\{ -\check{C}_1(\log[eN^{1/\kappa}])^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}} \Big\} \leq u \leq \check{C}_2(\log[eN^{1/\kappa}])^{\frac{\kappa}{2\kappa-1}} N^{-\frac{\kappa}{2\kappa-1}}.$$

Under Assumptions (MA3) *and* (CA1) *for* $q = 0$ *and* $\kappa < +\infty$

Inequality (6.12) gives

$$\Delta \leq \breve{C}\Big[\log(e\epsilon^{-1})N^{-\frac{1}{2}}\Big(\sqrt{\log(e^2\Delta^{-\frac{1}{\kappa}})}\Delta^{\frac{1}{2\kappa}} + \sqrt{\log(e^2u^{-1})u}\Big) + u^\kappa + \tfrac{[\log(eu^{-1})]^2}{N}\Big].$$

We obtain $\Delta \leq \breve{C}\log(e\epsilon^{-1})(\log N)^{\frac{\kappa}{2\kappa-1}} N^{-\frac{\kappa}{2\kappa-1}}$ when

$$\sqrt{\tfrac{\log(e^2u^{-1})u}{N}} + u^\kappa + \tfrac{[\log(eu^{-1})]^2}{N} \leq \breve{C}(\log N)^{\frac{\kappa}{2\kappa-1}} N^{-\frac{\kappa}{2\kappa-1}},$$

so when there exists $\check{C}_1, \check{C}_2 > 0$ such that

$$\exp\Big\{ -\check{C}_1(\log N)^{\frac{\kappa}{4\kappa-2}} N^{\frac{\kappa-1}{4\kappa-2}} \Big\} \leq u \leq \check{C}_2(\log N)^{\frac{1}{2\kappa-1}} N^{-\frac{1}{2\kappa-1}}.$$

Under Assumptions (MA2) *and* (CA1) *for* $0 < q < 1$

Inequality (6.12) becomes

$$\Delta \leq \breve{C}\Big[u + \log(e\epsilon^{-1})N^{-\frac{1}{2}}\Big(\Delta^{\frac{1-q}{2\kappa}} + u^{\frac{1-q}{2}}\Big) + \breve{C}\frac{\log(e\epsilon^{-1})}{N}u^{-q}\Big].$$

This leads to $\Delta \leq \breve{C}\log(e\epsilon^{-1})N^{-\frac{\kappa}{2\kappa-1+q}}$ when the inequality

$$N^{-\frac{1}{2}}u^{\frac{1-q}{2}} + u + \breve{C}\tfrac{\log(e\epsilon^{-1})}{N}u^{-q} \leq \breve{C}N^{-\frac{\kappa}{2\kappa-1+q}}$$

holds, hence when there exist $\check{C}_1, \check{C}_2$ such that $\check{C}_1 N^{-\frac{\kappa-1+q}{q(2\kappa-1+q)}} \leq u \leq \check{C}_2 N^{-\frac{\kappa}{2\kappa-1+q}}$.

Similarly, we deal with the five other cases. To finish the proof, we just have to notice that, when for any $\epsilon > 0$ and some real function ϕ, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have $\Delta \leq \log(e\epsilon^{-1})\phi(N)$, then we have $\mathbb{P}^{\otimes N}\Delta \leq 2\phi(N)$.

Remark 6.4. Once more, for sake of simplicity, we have done the proof for minimal nets without explicit values of the constants. It is easy to adapt the proof to almost minimal nets and to get an explicit constant $\breve{C}$ in terms of the other constants of the problem.

6.8. **Proof of Theorem 3.15.** Let $u_N = \check{C}_2 w_N$. Let $\mathcal{N}'$ be a u_N-minimal bracketing net of the model $\mathcal{F}$. Let $A \triangleq \{f \in \mathcal{F} : \mathbb{P}_{f,\bar{f}} \leq u_N\}$. There exists a posterior distribution $\hat{\rho}_{\mathcal{N}'} : \mathcal{Z}^N \to \mathcal{M}_+^1(\mathcal{N}')$ $\big($for instance, $\hat{\rho}_{\mathcal{N}'} \triangleq \pi_{-\lambda_N r} \circ n_L^{-1}\big)$ such that we have $\hat{\rho}_{\mathcal{N}'} r \leq \pi_{-\lambda_N r} r + u_N$ and

$$\pi_{-\lambda_N r} R \leq \hat{\rho}_{\mathcal{N}'} R + u_N. \tag{6.13}$$

We have

$$\hat{\rho}_{\mathcal{N}'} r \leq \pi_{-\lambda_N r} r + u_N \leq \pi|_A r + \frac{K(\pi|_A, \pi)}{\lambda_N} + u_N \tag{6.14}$$

and

$$K(\pi|_A, \pi) = \log[\pi(A)^{-1}] \leq \breve{C} h_q(u_N) \leq \breve{C}\lambda_N u_N. \tag{6.15}$$

From inequality (2.1) for $\big(\rho_2,\pi_2,\rho_1,\pi_1\big)=\big(\delta_{\tilde f},\delta_{\tilde f},\pi|_A,\pi|_A\big)$ and $\lambda=N$, with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we have $\pi|_A r-r(\tilde f)\le \pi|_A R-R(\tilde f)+u_N+\frac{\log(\epsilon^{-1})}{N}$, hence

$$\pi|_A r-r(\tilde f)\le 2u_N+\frac{\log(\epsilon^{-1})}{N}\le C\log(e\epsilon^{-1})w_N. \tag{6.16}$$

Combining Inequalities (6.14), (6.15) and (6.16), with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, we obtain $\hat\rho_{\mathcal{N}'}r\le r(\tilde f)+\breve C\log(e\epsilon^{-1})w_N$. The result follows from Theorem 3.12 and inequality (6.13).

6.9. **Proof of Theorem 3.16.** Let $\breve C_1>0$, $u\ge \breve C_1\beta_q$ and $\mathcal{N}$ be a $\big(u,\mathcal{F},\mathbb{P}\big)$-minimal bracketing net. Let π be the uniform distribution on this net. From inequality (8.2) in [1] for $\mathcal{W}\big((f_1,f_2),X\big)=\mathbb{1}_{f_1(X)\neq f_2(X)}$ and $\nu=\pi\otimes\pi$, we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any function f_1',f_2' in the net $\mathcal{N}$ and any $\lambda>0$, we have

$$\bar{\mathbb{P}}_{f_1',f_2'}\le\big[1+\tfrac{\lambda}{N}g\big(\tfrac{\lambda}{N}\big)\big]\mathbb{P}_{f_1',f_2'}+\tfrac{2\log|\mathcal{N}|+\log(\epsilon^{-1})}{\lambda}$$

Recall that $\breve C$ is a constant (possibly depending on the other constants of the problem) which value may differ from line to line. Taking $\epsilon=(\alpha_q)^C$ and $\lambda=N$, we obtain that with probability at least $1-(\alpha_q)^C$, for any functions f_1,f_2 in the set $\mathcal{F}$, we have

$$\begin{aligned}
\bar{\mathbb{P}}_{f_1,f_2} &\le \bar{\mathbb{P}}_{f_1,n_L(f_1)}+\bar{\mathbb{P}}_{n_L(f_1),n_L(f_2)}+\bar{\mathbb{P}}_{n_L(f_2),f_2}\\
&\le \bar{\mathbb{P}}_{n_L(f_1),n_U(f_1)}+\bar{\mathbb{P}}_{n_L(f_1),n_L(f_2)}+\bar{\mathbb{P}}_{n_L(f_2),n_U(f_2)}\\
&\le (e-1)\mathbb{P}_{n_L(f_1),n_L(f_2)}+\tfrac{6C'h_q(\breve C_1\beta_q)+3C\log(\alpha_q^{-1})}{N}+2u\\
&\le (e-1)\mathbb{P}_{n_L(f_1),n_L(f_2)}+\breve C\beta_q+2u\\
&\le (e-1)\mathbb{P}_{f_1,f_2}+\breve Cu.
\end{aligned}$$

By applying inequality (8.2) in [1] to $\mathcal{W}\big((f_1,f_2),X\big)=-\mathbb{1}_{f_1(X)\neq f_2(X)}$, we can similarly proved that with probability at least $1-(\alpha_q)^C$, for any functions f_1,f_2 in the set $\mathcal{F}$, we have $(3-e)\mathbb{P}_{f_1,f_2}\le\bar{\mathbb{P}}_{f_1,f_2}+\breve Cu$. (The constants $e-1$ and $3-e$ have nothing fundamental and we can make them as close as 1 as we want provided that we change the other constants.) These two inequalities allows to prove respectively the first two items of the theorem for one radius $u\ge\breve C_1\beta_q$. To get a result uniform wrt the radius, it suffices to make a union bound for radius in a geometric grid of $[\breve C_1\beta_q;1]$.

For the last item, when we have $u\ge\breve C_1\beta_q$ for a sufficiently large $\breve C_1$, there exists a small constant $\breve C'$ satisfying

$$H(u,\mathcal{F},\bar{\mathbb{P}})\le H_{\mathrm{p}}(u,\mathcal{F},\bar{\mathbb{P}})\le H_{\mathrm{p}}(\breve C'u,\mathcal{F},\mathbb{P})\le H(\breve C'u/2,\mathcal{F},\mathbb{P})\le\breve Ch_q(u).$$

6.10. **Proof of Theorem 4.1.** The proof is adapted from the proof of the concentration of $N(X_1^N)$ [10, p.42]. First, we prove that for any $k\in\mathbb{N}$, the quantity $\frac{H(\frac{k}{N},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot})}{\log 2}$ is a self-bounded quantity in the sense given in [10, p.23]. Let $\mathcal{N}_k$ be a $\big(\frac{k}{N},\mathcal{F},\bar{\mathbb{P}}\big)$-minimal net. Define the probability distribution

$$\bar{\mathbb{P}}^{(i)}\triangleq\tfrac{\delta_{Z_1}+\cdots+\delta_{Z_{i-1}}+\delta_{Z_{i+1}}+\cdots+\delta_{Z_N}}{N-1}.$$

Let $H^{(i)}$ be the logarithm of the cardinal of the smallest $\big(\frac{k}{N-1},\mathcal{F},\bar{\mathbb{P}}^{(i)}\big)$-net using only functions in the net $\mathcal{N}_k$. We have

$$0\le H(\tfrac{k}{N},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot})-H^{(i)}\le H(\tfrac{k}{N},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot})-H(\tfrac{k}{N-1},\mathcal{F},\bar{\mathbb{P}}^{(i)})\le\log 2.$$

Let $\mathbb{V} = \big(f(X_1), \dots, f(X_N)\big)$ be the random vector induced by the uniform distribution on the net $\mathcal{N}_k$. The Shannon entropy of this vector is $\log|\mathcal{N}_k| = H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$. Define $\mathbb{V}^{(i)} \triangleq \big(f(X_1), \dots, f(X_{i-1}), f(X_{i+1}), \dots, f(X_N)\big)$. Since the uniform distribution maximizes the entropy, we have $H^{(i)} \geq H(\mathbb{V}^{(i)})$. Then, using Han's inequality (see for instance [10, p.31]), we obtain that $\frac{H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})}{\log 2}$ is a self-bounded quantity.

Therefore, we can apply Corollary 5 in [10, p.43]. To shorten, we write until the end of the proof $H(k)$ for $H(\frac{k}{N}, \mathcal{F}, \bar{\mathbb{P}}_{\cdot,\cdot})$. For any $\eta > 0$, we have

$$\mathbb{P}^{\otimes N}\Big[H(k) \geq \mathbb{P}^{\otimes N} H(k) + (\log 2)\eta\Big] \leq e^{-\frac{\eta^2}{2\mathbb{P}^{\otimes N} H(k) + \frac{2\eta}{3}}} \tag{6.17}$$

and

$$\mathbb{P}^{\otimes N}\Big[H(k) \leq \mathbb{P}^{\otimes N} H(k) - (\log 2)\eta\Big] \leq e^{-\frac{\eta^2}{2\mathbb{P}^{\otimes N} H(k)}}. \tag{6.18}$$

Introducing $\epsilon = e^{-\frac{\eta^2}{2\mathbb{P}^{\otimes N} H(k) + \frac{2\eta}{3}}}$, equivalently

$$\eta^2 - \left(\frac{2\mathbb{P}^{\otimes N} H(k)}{\log 2} + \frac{2\eta}{3}\right)\log(\epsilon^{-1}) = 0,$$

$$\eta = \frac{\log(\epsilon^{-1})}{3}\left(1 + \sqrt{1 + \frac{18\mathbb{P}^{\otimes N} H(k)}{(\log 2)\log(\epsilon^{-1})}}\right),$$

we obtain that for any $\epsilon > 0$,

$$\mathbb{P}^{\otimes N}\left\{H(k) \geq \mathbb{P}^{\otimes N} H(k) + \frac{(\log 2)\log(\epsilon^{-1})}{3}\left(1 + \sqrt{1 + \frac{18\mathbb{P}^{\otimes N} H(k)}{(\log 2)\log(\epsilon^{-1})}}\right)\right\} \leq \epsilon,$$

which is the first assertion of the lemma. The second inequality of the lemma is a direct consequence of inequality (6.18). The following two inequalities in the lemma comes similarly from inequality (6.17). Finally, inequality (4.3) comes from combining Inequalities (4.1) and (4.2).

6.11. **Proof of Theorem 4.3.** For any $k \in \{0, \dots, U\}$, let $\mathcal{N}_k$ be a 2^{-k}-minimal covering net of $\mathcal{F}$ for the pseudo-distance $\bar{\bar{\mathbb{P}}}$. For any $(i,k) \in \{1,2\} \times \{0, \dots, U\}$, let $f_{i,k}$ be a nearest neighbour of f_i in the set $\mathcal{N}_k$. Let $0 \leq K \leq U$ be the integer satisfying $\frac{\bar{\bar{\mathbb{P}}}_{f_1, f_2} \vee u}{2} < 2^{-K} \leq \bar{\bar{\mathbb{P}}}_{f_1, f_2} \vee u$.

Since we have

$$\partial(f_1; f_2) = \partial(f_{1,K}; f_{2,K}) + \textstyle\sum_{k=K+1}^{U} \Big\{\partial(f_{1,k}; f_{1,k-1}) + \partial(f_{2,k-1}; f_{2,k})\Big\},$$

we need to apply Lemma 4.2 to $(\mathcal{S}_1, \mathcal{S}_2) \in \cup_{1 \leq k \leq U}\big\{(\mathcal{N}_{k-1}, \mathcal{N}_k) \cup (\mathcal{N}_k, \mathcal{N}_{k-1}) \cup (\mathcal{N}_k, \mathcal{N}_k)\big\}$ and to do a union bound on the associated $3U$ inequalities. Let w_k, $k \in \mathbb{N}^*$ be positive integers such that $\sum_{k \geq 1} w_k = 1$. With probability at least $1 - \epsilon$, for any $k \in \mathbb{N}^*$, for any $(f_1', f_2') \in (\mathcal{N}_{k-1} \times \mathcal{N}_k) \cup (\mathcal{N}_k \times \mathcal{N}_{k-1}) \cup (\mathcal{N}_k \times \mathcal{N}_k)$, we have

$$\partial(f_1'; f_2') \leq \sqrt{\frac{8\bar{\bar{\mathbb{P}}}_{f_1', f_2'} \log(3|\mathcal{N}_k|^2 w_k^{-1} \epsilon^{-1})}{N}}$$

For any $(i,k) \in \{1,2\} \times \{1, \dots, U\}$, we have $\mathbb{P}_{f_{i,k-1}, f_{i,k}} \leq 3 \times 2^{-k}$. Denote

$$B_k \triangleq \sqrt{\frac{24 \times 2^{-k} \log(3|\mathcal{N}_k|^2 w_k^{-1} \epsilon^{-1})}{N}}$$

We have $\mathbb{P}_{f_{1,K},f_{2,K}} \leq 4\times 2^{-K}$ and $f_{1,0}=f_{2,0}$. Chaining the inequalities, we obtain that, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$,

$$\partial(f_1;f_2) \leq 2B_K \mathbb{1}_{K>0} + 2\sum_{k=K+1}^{U} B_k \leq 2\sum_{k=K\vee 1}^{U} B_k,$$

hence

$$\partial(f_1;f_2) \leq 4\sum_{k=K\vee 1}^{U} \sqrt{\frac{6\times 2^{-k}[2H_k+\log(3w_k^{-1})+\log(\epsilon^{-1})]}{N}}.$$

We want that the union bound term $\log(3w_k^{-1})$ remains negligible wrt the complexity term $2H_k$. This leads to choose, for instance, $w_k = \frac{1}{k(k+1)}$ since $\sum_{k\geq 1}\frac{1}{k(k+1)}=1$ and for small classes of functions (i.e. VC-classes), the entropy H_k has the order of k, hence $\log(3w_k^{-1}) \ll H_k$. We obtain

$$\partial(f_1;f_2) \leq \sum_{k\in\mathbb{N}^*:u\leq 2^{-k}\leq \bar{\mathbb{P}}_{f_1,f_2}\vee u} 4\sqrt{\frac{6\times 2^{-k}\{2H_k+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}}.$$

Remark 6.5. The previous result can also be written in terms of integral. Introduce the set $E \triangleq \left\{k\in\mathbb{N} : 2^{-k} \leq (\bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u)\wedge\frac{1}{2}\right\}$ and take $H_k = H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}})\vee 1$. We get

$$\begin{aligned}
\partial(f_1;f_2) &\leq 16\sqrt{\tfrac{3}{N}}\sum_{k\in E}\sqrt{\tfrac{H_k}{2^{-k}}}\big(2^{-k}-2^{-k-1}\big) + 4\sqrt{\tfrac{6\log(3\epsilon^{-1})}{N}}\sum_{k\in E}(\sqrt{2})^{-k} \\
&\qquad +4\sqrt{\tfrac{6}{N}}\sum_{k\in E}(\sqrt{2})^{-k}\sqrt{\log[k(k+1)]} \\
&\leq 16\sqrt{\tfrac{3}{N}}\int_{\frac{u}{2}}^{(\bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u)\wedge\frac{1}{2}}\sqrt{\tfrac{H(x,\mathcal{F},\bar{\bar{\mathbb{P}}})\vee 1}{x}}\,dx \\
&\qquad +4\sqrt{\tfrac{6\log(3\epsilon^{-1})}{N}}\tfrac{\sqrt{2}}{\sqrt{2}-1}\sqrt{\bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u} + 4\sqrt{\tfrac{6}{N}}\sum_{k\in E}(\sqrt{2})^{-k}\sqrt{2\log(k+1)}.
\end{aligned}$$

Let $\varphi(x) \triangleq \frac{1}{\sqrt{x}}\sqrt{\log\left(e\frac{\log x^{-1}}{\log 2}\right)}$ for any $0<x\leq\frac{1}{2}$. The function φ is decreasing on $\left[0;\frac{1}{2}\right]$. The last term can be written as

$$8\sqrt{\tfrac{3}{N}}\sum_{k\in E}\frac{2\left(2^{-k}-2^{-k-1}\right)}{\sqrt{2^{-k}}}\sqrt{\log(k+1)} \leq 16\sqrt{\tfrac{3}{N}}\int_{\frac{u}{2}}^{(\bar{\bar{\mathbb{P}}}_{f_1,f_2}\vee u)\wedge\frac{1}{2}}\varphi(x)dx.$$

6.12. **Proof of Theorem 4.4.** Let us take $U\in\mathbb{N}$ such that $2^{-U}<\frac{1}{2N}$. From Theorem 4.3, for any $\epsilon>0$, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any functions f_1 and f_2 in the set $\mathcal{F}$,

(6.19)

$$\begin{aligned}
&r'(f_2)-r'(f_1)+r(f_1)-r(f_2) \\
&\qquad\leq \sum_{k\in\mathbb{N}:\frac{1}{2N}\leq 2^{-k}\leq\bar{\bar{\mathbb{P}}}_{f_1,f_2}\wedge\frac{1}{2}} 4\sqrt{\frac{6\times 2^{-k}\{2H_k+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}}.
\end{aligned}$$

Let $\mathcal{N}$ be a $(\frac{1}{3N},\mathcal{F},\bar{\bar{\mathbb{P}}}_{\cdot,\cdot})$-minimal covering net. From Theorem 8.4 in [1] apply to $\mathcal{W}\big((f_1,f_2),X\big)=\mathbb{1}_{f_1(X)\neq f_2(X)}$, we obtain that with $\mathbb{P}^{\otimes N}$-probability at least $1-\epsilon$, for any function f_1,f_2 in the net $\mathcal{F}$, we have

$$\bar{\mathbb{P}}'_{f_1,f_2} \leq \bar{\mathbb{P}}_{f_1,f_2} + 2\sqrt{\frac{\bar{\bar{\mathbb{P}}}_{f_1,f_2}\log[N^2(X_1^{2N})\epsilon^{-1}]}{N}},$$

hence

$$\bar{\bar{\mathbb{P}}}_{f_1,f_2} \leq \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\frac{\bar{\bar{\mathbb{P}}}_{f_1,f_2}\log[N^2(X_1^{2N})\epsilon^{-1}]}{N}}.$$

Let $\check{\mathcal{K}} \triangleq \frac{2\log N(X_1^{2N})+\log(\epsilon^{-1})}{N}$. By solving the previous inequation, we obtain

$$\bar{\bar{\mathbb{P}}}_{f_1,f_2} \le \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\check{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2} + \frac{\check{\mathcal{K}}^2}{4}} + \frac{\check{\mathcal{K}}}{2}. \tag{6.20}$$

From inequality (4.3), we have

$$\begin{aligned} H(u,\mathcal{F},\bar{\mathbb{P}}') &\le H(u,\mathcal{F},\bar{\mathbb{P}}) + (\log 2)\log(\epsilon^{-1})\left(\tfrac{12}{5} + 1 + \tfrac{2H(u,\mathcal{F},\bar{\mathbb{P}})}{(\log 2)\log(\epsilon^{-1})}\right) \\ &= 3H(u,\mathcal{F},\bar{\mathbb{P}}) + \tfrac{17\log 2}{5}\log(\epsilon^{-1}), \end{aligned}$$

hence

$$H(u,\mathcal{F},\bar{\bar{\mathbb{P}}}) \le 4H(u,\mathcal{F},\bar{\mathbb{P}}) + \tfrac{17\log 2}{5}\log(\epsilon^{-1}).$$

Taking an union bound with weight $\frac{1}{k(k+1)}$, we obtain that with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, for any $k \ge 1$, we have

$$H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}}) \le 4H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + 2.4\log[k(k+1)\epsilon^{-1}]. \tag{6.21}$$

Let $H'_k \triangleq 4H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + 2.4\log[k(k+1)\epsilon^{-1}]$. Rigorously, we cannot apply Theorem 4.3 for $H_k = H'_k$ since H'_k is not always an upper bound of $H(2^{-k},\mathcal{F},\bar{\bar{\mathbb{P}}}) \vee 1$. However we can modify the proof of Theorem 4.3 to take into a "probably approximatively correct" inequality. Therefore, combining Inequalities (6.19), (6.20) and (6.21), letting $\bar{\mathcal{K}} \triangleq \frac{2H'_U+\log(\epsilon^{-1})}{N}$ and

$$\bar{E} \triangleq \left\{k \in \mathbb{N}^* : \tfrac{1}{2N} \le 2^{-k} \le \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\bar{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2} + \frac{\bar{\mathcal{K}}^2}{4}} + \frac{\bar{\mathcal{K}}}{2}\right\},$$

we obtain that with $\mathbb{P}^{\otimes 2N}$-probability at least $1-3\epsilon$, for any functions f_1, f_2 in $\mathcal{F}$, we have

$$r'(f_2) - r'(f_1) + r(f_1) - r(f_2) \le \sum_{k\in\bar{E}} 4\sqrt{\tfrac{6\times 2^{-k}\{2H'_k+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}}.$$

To obtain the announced result, we simplify this formula by using

$$\bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\bar{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2} + \frac{\bar{\mathcal{K}}^2}{4}} + \frac{\bar{\mathcal{K}}}{2} \le \bar{\mathbb{P}}_{f_1,f_2} + \sqrt{\bar{\mathcal{K}}\bar{\mathbb{P}}_{f_1,f_2}} + \bar{\mathcal{K}} \le \tfrac{5}{4}\bar{\mathbb{P}}_{f_1,f_2} + 2\bar{\mathcal{K}}, \tag{6.22}$$

$$\log[U(U+1)] \le \log\left[\left(2+\tfrac{\log N}{\log 2}\right)\left(3+\tfrac{\log N}{\log 2}\right)\right] \le \log 6 + 2\log\left(\tfrac{e}{2\log 2}\log N\right)$$

and

$$2H'_k + \log[3k(k+1)\epsilon^{-1}] \le 8H(2^{-k},\mathcal{F},\bar{\mathbb{P}}_{\cdot,\cdot}) + 6\log[k(k+1)\epsilon^{-1}] + 1.$$

6.13. **Proof of Corollary 4.5.** Let $f \in \mathcal{F}$. If $\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f} = 0$, then we trivially have $r'(\hat{f}_{\mathrm{ERM}}) \le r'(f)$. Otherwise, we have $\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f} \ge \frac{1}{2N}$. Let $K \in \mathbb{N}$ such that $\frac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{2} < 2^{-K} \le \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}$. From Theorem 4.3, with $\mathbb{P}^{\otimes 2N}$-probability at least $1-\epsilon$, we have

$$\begin{aligned} r'(\hat{f}_{\mathrm{ERM}}) - r'(f) &\le \sum_{k\ge K} 4\sqrt{\tfrac{6\times 2^{-k}\{2V\log(e2^{k+2})+\log[3k(k+1)]+\log(\epsilon^{-1})\}}{N}} \\ &\le \sum_{k\ge K} 4\sqrt{\tfrac{6\times 2^{-k}\{(2V+1)\log(e2^{k+2})+\log(\epsilon^{-1})\}}{N}} \\ &\le 4\sqrt{\tfrac{6(2V+1)}{N}}\sum_{k\ge K}\sqrt{2^{-k}\log(e2^{k+2})} + 4\sqrt{\tfrac{6\log(\epsilon^{-1})}{N}}\sum_{k\ge K}(\sqrt{2})^{-k} \end{aligned}$$

Now, for any $k \geq K \geq 0$ and $V \geq 1$, we have $\frac{\log(e2^{k+2})}{\log(e2^{K+2})} \leq \frac{(k-K)(\log 2)}{1+2\log 2}+1$. Therefore we get

$$\begin{aligned} r'(\hat{f}_{\mathrm{ERM}}) - r'(f) &\leq 4\sqrt{\tfrac{6(2V+1)2^{-K}\log(e2^{K+2})}{N}} \sum_{k\geq 0} \sqrt{2^{-k}\big(\tfrac{k\log 2}{1+2\log 2}+1\big)} \\ &\qquad +4\sqrt{\tfrac{6\log(\epsilon^{-1})}{N}}(\sqrt{2})^{-K}\tfrac{\sqrt{2}}{\sqrt{2}-1} \\ &\leq 4\sqrt{\tfrac{6(2V+1)2^{-K}\log(e2^{K+2})}{N}} \sum_{k\geq 0} \sqrt{2^{-k}\big(\tfrac{k\log 2}{1+2\log 2}+1\big)} \\ &\qquad +4\sqrt{\tfrac{6\log(\epsilon^{-1})}{N}}(\sqrt{2})^{-K}\tfrac{\sqrt{2}}{\sqrt{2}-1} \\ &\leq 47\sqrt{\tfrac{V+1}{N}}\sqrt{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}\log\Big(\tfrac{8e}{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}\Big)} + 34\sqrt{\tfrac{\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}\log(\epsilon^{-1})}{N}} \end{aligned}$$

For the second assertion of the corollary, we use Jensen's inequality and the concavity of $x \mapsto \sqrt{x\log(8ex^{-1})}$ in order to obtain that for any function $f \in \mathcal{F}$,

$$\begin{aligned} \mathbb{P}^{\otimes N}R(\hat{f}_{\mathrm{ERM}}) &\leq R(f) + 47\sqrt{\tfrac{(V+1)\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{N}\log\Big(\tfrac{8e}{\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}\Big)} \\ &\qquad +34\sqrt{\tfrac{\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},f}}{N}}. \end{aligned}$$

6.14. **Proof of Corollary 4.6.** For $\kappa = +\infty$ (i.e. no margin assumption), the result comes from inequality (4.5) since the function $x \mapsto x\log(8e/x)$ is an increasing function on $[0;1]$, hence upper bounded by its value for $x = 1$. Specifically, we obtain

$$\mathbb{P}^{\otimes N}R(\hat{f}_{\mathrm{ERM}}) - R(\tilde{f}) \leq 83\sqrt{\tfrac{V+1}{N}} + \tfrac{34}{\sqrt{N}}. \tag{6.23}$$

Note that it is thanks to the chaining that we get rid of the $\log N$ factor. For $\kappa < +\infty$, chained and unchained results lead to the same convergence rate: $\big(\frac{V}{N}\log N\big)^{\frac{\kappa}{2\kappa-1}}$.

To obtain this rate from the previous bounds, we just need to link the variance term $\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}$ with $\mathbb{P}^{\otimes N}\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}}$ in order to use the margin assumption.

Combining Inequalities (6.20) and (6.22), we obtain

$$\begin{aligned} \bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} &\leq \tfrac{5}{4}\bar{\mathbb{P}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} + \tfrac{4\log N(X_1^{2N})+2\log(\epsilon^{-1})}{N} \\ &\leq \tfrac{5}{4}\bar{\mathbb{P}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} + \tfrac{4V\log\left(\frac{2eN}{V}\right)+2\log(\epsilon^{-1})}{N}, \end{aligned}$$

hence

$$\mathbb{P}^{\otimes 2N}\bar{\bar{\mathbb{P}}}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} \leq \tfrac{5}{4}\mathbb{P}^{\otimes N}\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} + \tfrac{4V\log\left(\frac{2eN}{V}\right)+2}{N}. \tag{6.24}$$

Now, by the margin assumption and Jensen's inequality, we have

$$\mathbb{P}^{\otimes N}\mathbb{P}_{\hat{f}_{\mathrm{ERM}},\tilde{f}} \leq C''\mathbb{P}^{\otimes N}\big(\Delta^{\frac{1}{\kappa}}\big) \leq C''\left(\mathbb{P}^{\otimes N}\Delta\right)^{\frac{1}{\kappa}}. \tag{6.25}$$

The convergence rate then follows from (6.24), (6.25) and either (4.4) or (4.5).

6.15. **Proof of Lemma 5.1.** Let $\vec{\sigma}_{j,r} \triangleq (\sigma_1,\ldots,\sigma_{j-1},r,\sigma_{j+1},\ldots,\sigma_m)$ for any $r \in \{-1,0,+1\}$. The distribution $\mathbb{P}_{\vec{\sigma}_{j,0}}$ is such that $\mathbb{P}_{\vec{\sigma}_{j,0}}(dX) = \mu(dX)$ and

$$\mathbb{P}_{\vec{\sigma}_{j,0}}(Y=1|X) = \begin{cases} \frac{1}{2} \text{ for any } X \in \mathcal{X}_j \\ \mathbb{P}_{\vec{\sigma}}(Y=1|X) \text{ otherwise} \end{cases}.$$

Introduce the quantities $\pi_{r,j} \triangleq \frac{\mathbb{P}_{\vec{\sigma}_{j,r}}^{\otimes N}}{\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes N}}(Z_1^N) = \prod_{i=1}^N \big[1 + r\mathbb{1}_{X_i \in \mathcal{X}_j}(2Y_i - 1)\xi(X_i)\big]$ for any $r \in \{-1; +1\}$. Let ν denote the distribution of a Rademacher variable:

$$\nu(\sigma = +1) = \nu(\sigma = -1) = \frac{1}{2}.$$

The variational distance between two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as $V(\mathbb{P}_1, \mathbb{P}_2) \triangleq \sup_{A \text{ measurable set}} \big\{\mathbb{P}_1(A) - \mathbb{P}_2(A)\big\}$. When the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ are absolutely continuous wrt a probability distribution $\mathbb{P}_0$, we have

$$V(\mathbb{P}_1, \mathbb{P}_2) = \tfrac{1}{2}\int \big|\tfrac{\mathbb{P}_1}{\mathbb{P}_0} - \tfrac{\mathbb{P}_2}{\mathbb{P}_0}\big| d\mathbb{P}_0 = 1 - \int \big(\tfrac{\mathbb{P}_1}{\mathbb{P}_0} \wedge \tfrac{\mathbb{P}_2}{\mathbb{P}_0}\big) d\mathbb{P}_0.$$

We have successively

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{P}} \Big\{\mathbb{P}^{\otimes N}\mathbb{P}[\hat{f}(X) \neq Y] - \mathbb{P}[f^*_{\mathbb{P}}(X) \neq Y)\Big\} \\
\geq\ & \sup_{\vec{\sigma}\in\{-1;+1\}^m} \Big\{\big(\mathbb{P}_{\vec{\sigma}}^{\otimes N}\big)\mathbb{P}_{\vec{\sigma}}[\hat{f}(X) \neq Y] - \mathbb{P}_{\vec{\sigma}}[f^*_{\mathbb{P}_{\vec{\sigma}}} \neq Y]\Big\} \\
=\ & \sup_{\vec{\sigma}\in\{-1;+1\}^m} \left\{\big(\mathbb{P}_{\vec{\sigma}}^{\otimes N}\big)\mathbb{P}_{\vec{\sigma}}\Big(\xi(X)\mathbb{1}_{\hat{f}(X)\neq f^*_{\mathbb{P}_{\vec{\sigma}}}(X)}\Big)\right\} \\
=\ & \sup_{\vec{\sigma}\in\{-1;+1\}^m} \left\{\mathbb{P}_{\vec{\sigma}}^{\otimes N}\Big(\textstyle\sum_{j=1}^m \mu\big[\xi(X)\mathbb{1}_{\hat{f}(X)\neq\frac{1+\sigma_j}{2};X\in\mathcal{X}_j}\big]\Big)\right\} \\
\geq\ & \mathbb{E}_{\nu^{\otimes m}} \textstyle\sum_{j=1}^m \mathbb{P}_{\vec{\sigma}}^{\otimes N}\Big[\mu\big[\xi(X)\mathbb{1}_{\hat{f}(X)\neq\frac{1+\sigma_j}{2};X\in\mathcal{X}_j}\big]\Big] \\
=\ & \mathbb{E}_{\nu^{\otimes m}} \textstyle\sum_{j=1}^m \mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes N}\left(\frac{\mathbb{P}_{\vec{\sigma}}^{\otimes N}}{\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes N}}\mu\big[\xi(X)\mathbb{1}_{\hat{f}(X)\neq\frac{1+\sigma_j}{2};X\in\mathcal{X}_j}\big]\right) \\
=\ & \mathbb{E}_{\nu^{\otimes(m-1)}(d\sigma_1,\ldots,d\sigma_{j-1},d\sigma_{j+1},\ldots,d\sigma_m)} \textstyle\sum_{j=1}^m \mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes N}\mathbb{E}_{\nu(d\sigma_j)} \\
&\qquad\qquad \left(\frac{\mathbb{P}_{\vec{\sigma}}^{\otimes N}}{\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes N}}\mu\big[\xi(X)\mathbb{1}_{\hat{f}(X)\neq\frac{1+\sigma_j}{2};X\in\mathcal{X}_j}\big]\right) \\
\geq\ & \mathbb{E}_{\nu^{\otimes(m-1)}(d\sigma_1,\ldots,d\sigma_{j-1},d\sigma_{j+1},\ldots,d\sigma_m)} \textstyle\sum_{j=1}^m \mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes N} \\
&\qquad\qquad \left[(\pi_{-1,j} \wedge \pi_{+1,j})\mathbb{E}_{\nu(d\sigma_j)}\mu\big[\xi(X)\mathbb{1}_{\hat{f}(X)\neq\frac{1+\sigma_j}{2};X\in\mathcal{X}_j}\big]\right] \\
=\ & \mathbb{E}_{\nu^{\otimes(m-1)}(d\sigma_1,\ldots,d\sigma_{j-1},d\sigma_{j+1},\ldots,d\sigma_m)} \textstyle\sum_{j=1}^m \\
&\qquad\qquad \tfrac{1}{2}\mu\big[\xi(X)\mathbb{1}_{X\in\mathcal{X}_j}\big]\Big[1 - V\Big(\mathbb{P}_{\vec{\sigma}_{j,-1}}^{\otimes N}, \mathbb{P}_{\vec{\sigma}_{j,+1}}^{\otimes N}\Big)\Big] \\
=\ & \tfrac{mw}{2}\mu\big[\xi(X) \,\big|\, X \in \mathcal{X}_j\big]\Big[1 - V\Big(\mathbb{P}_{-1,1,\ldots,1}^{\otimes N}, \mathbb{P}_{1,1,\ldots,1}^{\otimes N}\Big)\Big].
\end{aligned} \tag{6.26}$$

Now let us prove

$$V\Big(\mathbb{P}_{-1,1,\ldots,1}^{\otimes N}, \mathbb{P}_{1,1,\ldots,1}^{\otimes N}\Big) \leq b\sqrt{Nw}. \tag{6.27}$$

First, we have

$$V\Big(\mathbb{P}_{-1,1,\ldots,1}^{\otimes N}, \mathbb{P}_{1,1,\ldots,1}^{\otimes N}\Big) = \sum_{l=1}^N \binom{N}{l} w^l (1-w)^{N-l}\mathcal{V}_l, \tag{6.28}$$

where $\mathcal{V}_l \triangleq V\Big(\mathbb{P}_{-1}^{\otimes l}, \mathbb{P}_{+1}^{\otimes l}\Big)$ and $\mathbb{P}_\sigma \triangleq \mathbb{P}_{\sigma,1,\ldots,1}(\cdot|X \in \mathcal{X}_1)$ for any $\sigma \in \{-1, +1\}$. By simple computations, we get $\mathcal{V}_1 = \mu[\xi(X)|X \in \mathcal{X}_j]$. From Jensen's inequality, by the concavity of $x \mapsto \sqrt{1-x^2}$, we have

$$\sqrt{1-b^2} = \mu\big[\sqrt{1-\xi^2(X)} \,\big|\, X \in \mathcal{X}_j\big] \leq \sqrt{1 - \{\mu[\xi(X)|X \in \mathcal{X}_j]\}^2},$$

hence $\mathcal{V}_1 \leq b$.

For $l \geq 2$, we upper bound the variational distance by the Hellinger distance. By definition, the Hellinger distance $H(\mathbb{P},\mathbb{Q})$ satisfies $1 - \frac{H^2(\mathbb{P},\mathbb{Q})}{2} = \int \sqrt{d\mathbb{P}}\sqrt{d\mathbb{Q}}$. Hence the tensorization equality is $1 - \frac{H^2(\mathbb{P}^{\otimes l},\mathbb{Q}^{\otimes l})}{2} = \big(1 - \frac{H^2(\mathbb{P},\mathbb{Q})}{2}\big)^l$. We have

$$\mathcal{V}_l \leq H\Big(\mathbb{P}_{-1}^{\otimes l}, \mathbb{P}_{+1}^{\otimes l}\Big) = \sqrt{2\Big(1 - \Big[1 - \tfrac{H^2(\mathbb{P}_{-1},\mathbb{P}_{+1})}{2}\Big]^l\Big)}$$

and $1 - \frac{H^2(\mathbb{P}_{-1},\mathbb{P}_{+1})}{2} = \mu\big[\sqrt{1-\xi^2(X)}|X \in \mathcal{X}_1\big] = \sqrt{1-b^2}$, by definition of b. Now, for any $l \geq 2$ and $x \geq 0$, $2\Big(1 - \big[1-x^2\big]^{\frac{l}{2}}\Big) \leq lx^2$. Finally, for any $l \geq 1$, we have $\mathcal{V}_l \leq b\sqrt{l}$. Putting this result in equality (6.28), we get

$$V\Big(\mathbb{P}_{-1,1,\ldots,1}^{\otimes N}, \mathbb{P}_{1,1,\ldots,1}^{\otimes N}\Big) \leq b\sum_{l=0}^{N} \mathbb{P}\Big(\sum_{i=1}^{N}\epsilon_i = l\Big)\sqrt{l},$$

where the ϵ_i are i.i.d. random variables such that $\mathbb{P}(\epsilon_i = 1) = w = 1 - \mathbb{P}(\epsilon_i = 0)$. So we have $V\Big(\mathbb{P}_{-1,1,\ldots,1}^{\otimes N}, \mathbb{P}_{1,1,\ldots,1}^{\otimes N}\Big) \leq b\mathbb{P}\sqrt{\sum_{i=1}^N \epsilon_i} \leq b\sqrt{\mathbb{P}\sum_{i=1}^N \epsilon_i} = b\sqrt{Nw}$.

Remark 6.6. The last inequality in (6.26) is an equality when for any $j \in \{1,\ldots,m\}$, $\hat{f} = \underset{r\in\{-1;+1\}}{\operatorname{argmax}} \mathbb{P}^{\otimes N}{}_{\vec{\sigma}_{j,r}}$ on $\mathcal{X}_j$, i.e. when $\hat{f}$ is the maximum likelyhood estimator on the set $\mathcal{X} - \mathcal{X}_0$.

Appendix A. Proof of inequality (6.2)

For any $r > 0$, define $\Gamma(r) \triangleq \int_0^{+\infty} u^{r-1}\exp(-u)du$. Integrating by parts, we obtain the well-known property $\Gamma(r+1) = r\Gamma(r)$.

• We have

$$\begin{aligned}\pi\exp\big(-\lambda\Delta R\big) &= \int_0^{+\infty} \pi\big\{\exp\big(-\lambda\Delta R\big) \geq u\big\}du \\ &= \exp(-\lambda) + \int_{\exp(-\lambda)}^{1} \pi\big\{\exp\big(-\lambda\Delta R\big) \geq u\big\}du \\ &= \exp(-\lambda) + \int_0^1 \lambda\exp\big(-\lambda x\big)\pi\big(\Delta R \leq x\big)dx\end{aligned} \tag{A.1}$$

Let us introduce $A' \triangleq \exp(-C''')$. Since we have $\pi(\Delta R \leq x) = x^{C'}\big[A' + \eta(x)\big]$ with $\eta(x) = \underset{x\to 0}{\mathrm{o}}(x^s)$ and $\eta(x) \leq x^{-C'}$ and $\int_0^{+\infty}\lambda\exp\big(-\lambda x\big)x^{C'}dx = \frac{\Gamma(C'+1)}{\lambda^{C'}}$, we get

$$\begin{aligned}&\Big|\pi\exp\big(-\lambda\Delta R\big) - A'\tfrac{\Gamma(C'+1)}{\lambda^{C'}} - \exp(-\lambda)\Big| \\ &\qquad = \int_0^1 \lambda\exp\big(-\lambda x\big)x^{C'}\eta(x)dx + A'\int_1^{+\infty}\lambda\exp\big(-\lambda x\big)x^{C'}dx\end{aligned}$$

Since we have

$$\begin{aligned}&\int_0^1 \lambda\exp\big(-\lambda x\big)x^{C'}\eta(x)dx \\ &\qquad = \int_0^{\frac{1}{\sqrt{\lambda}}} \lambda\exp\big(-\lambda x\big)x^{C'+s}\underset{x\to 0}{\mathrm{o}}(1)dx + \int_{\frac{1}{\sqrt{\lambda}}}^1 \lambda\exp\big(-\lambda x\big)x^{C'}\eta(x)dx \\ &\qquad \leq \underset{\lambda\to+\infty}{\mathrm{o}}\big(\lambda^{-(C'+s)}\big) + \int_{\frac{1}{\sqrt{\lambda}}}^1 \lambda\exp\big(-\lambda x\big)dx \\ &\qquad = \underset{\lambda\to+\infty}{\mathrm{o}}\big(\lambda^{-(C'+s)}\big)\end{aligned}$$

and $\int_1^{+\infty}\lambda\exp\big(-\lambda x\big)x^{C'}dx = \underset{\lambda\to+\infty}{\mathrm{o}}\big(\lambda^{-(C'+s)}\big)$, we obtain

$$\pi\exp\big(-\lambda\Delta R\big) = A'\frac{\Gamma(C'+1) + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'}} \tag{A.2}$$

• From equalities (A.1), we have

(A.3)
$$\pi\big[\Delta R\exp\big(-\lambda\Delta R\big)\big] = \exp(-\lambda) + \int_0^1 (\lambda x - 1)\exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx.$$

We have just seen that $\int_0^1 \exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx = A'\frac{\Gamma(C'+1)+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'+1}}$. Besides, from the same argument as above, we have

$$\int_0^1 \lambda x\exp\big(-\lambda x\big)\pi\big(\Delta R \le x\big)dx = A'\frac{\Gamma(C'+2)+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'+1}}.$$

Since $\Gamma(C'+2) = (C'+1)\Gamma(C'+1)$, we obtain

$$\pi\big[\Delta R\exp\big(-\lambda\Delta R\big)\big] = A'\frac{C'\Gamma(C'+1)+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda^{C'+1}}.$$

• Combining the previous results, we obtain $\pi_{-\lambda R}\Delta R = \frac{C'+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\lambda}$.

Appendix B. Proof of inequality (6.4)

Let $\alpha > 0$ depend on λ such that $\lambda\alpha^{\frac{\kappa}{\kappa-1}} \underset{\lambda\to+\infty}{\to} 0$. Then there exists $0 < \zeta < 1$ depending on λ such that $\zeta \underset{\lambda\to+\infty}{\to} 1$ and $\lambda\big(\frac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}} \underset{\lambda\to+\infty}{\to} 0$. Let $h_\alpha(x) \triangleq x - \alpha x^{\frac{1}{\kappa}}$ and $x_0 \triangleq \big(\frac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}$. The function h decreases on $[0; x_0]$ and increases on $[x_0; +\infty]$. We have

$$\begin{aligned}
\pi\exp\big[-\lambda h_\alpha(\Delta R)\big] &= \pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)\le\zeta\Delta R}\Big\} \\
&\qquad +\pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)>\zeta\Delta R}\Big\} \\
&\le \exp\big[-\lambda h_\alpha(x_0)\big]\pi\Big\{\Delta R \le \big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}}\Big\} + \pi\exp\big[-\lambda\zeta\Delta R\big] \\
&\le \exp\big[(\kappa-1)\lambda\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}\big]\big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{C'\kappa}{\kappa-1}}\Big[1 + \underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big] \\
&\qquad +\pi\exp\big[-\lambda\zeta\Delta R\big]
\end{aligned}$$

From equality (A.2), we get

$$\begin{aligned}
&\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big) \\
&\quad = \exp\Big[(\kappa-1)\lambda\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}\Big]\lambda^{C'}\big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{C'\kappa}{\kappa-1}}\Big[\tfrac{1}{\Gamma(C'+1)} + \underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big] + \frac{1+\underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s})}{\zeta^{C'}} \\
&\quad = \underset{\lambda\to+\infty}{\mathrm{O}}\Big(\lambda^{C'}\big[\tfrac{\alpha}{1-\zeta}\big]^{\frac{\kappa C'}{\kappa-1}}\Big) + 1 + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s}) + \underset{\lambda\to+\infty}{\mathrm{O}}\big(1-\zeta\big)
\end{aligned}$$

Taking $\zeta = 1 - \big(\lambda\alpha^{\frac{\kappa}{\kappa-1}}\big)^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}}$, we obtain

$$\log\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big) = \underset{\lambda\to+\infty}{\mathrm{O}}\Big(\big[\lambda\alpha^{\frac{\kappa}{\kappa-1}}\big]^{\frac{(\kappa-1)C'}{\kappa C'+\kappa-1}}\Big) + \underset{\lambda\to+\infty}{\mathrm{o}}(\lambda^{-s}).$$

Appendix C. Proof of inequality (6.6)

We start with the following lemma.

Lemma C.1. *Let $h : \mathbb{R}^* \to \mathbb{R}$ be a C^3 convex function such that there exists $u_0 > 0$ satisfying $h'(u_0) = 0$ and $h''(u_0) > 0$. Let $\phi : \mathbb{R}^* \to \mathbb{R}$ be a continuous non negative*

function such that $\phi(u_0>0)$ and $u\mapsto \phi(u)\exp(-t_0u)$ integrable for some $t_0>0$. Then for any $A>u_0$, we have

$$\int_0^A \phi(u)\exp\big[-th(u)\big]du \underset{t\to+\infty}{\sim} \phi(u_0)\exp\big[-th(u_0)\big]\sqrt{\tfrac{2\pi}{th''(u_0)}}.$$

Proof. • Since the function h'' is non negative, continuous, $h''(u_0)>0$ and $h'(u_0)=0$, there exists $c>0$ such that for any $u\in[0;A]$, $h(u)-h(u_0)\geq c(u-u_0)^2$. Let $\alpha_t \triangleq t^{-p}$ with $\frac{1}{3}<p<\frac{1}{2}$. We have

$$\begin{aligned}\int_{[0;u_0-\alpha_t]\cup[u_0+\alpha_t;A]} \phi(u)\exp\big[-th(u)\big]du \\ \leq \exp\big[-th(u_0)\big]\int_{[0;u_0-\alpha_t]\cup[u_0+\alpha_t;A]}\phi(u)\exp\big[-tc(u-u_0)^2\big]du \\ = \exp\big[-th(u_0)\big]\underset{t\to+\infty}{\mathrm{O}}\big(\exp\big[-ct{\alpha_t}^2\big]\big).\end{aligned}$$

• From Taylor's theorem, for any $u\in[u_0-\alpha_t;u_0+\alpha_t]$, there exists $u^*\in[u_0-\alpha_t;u_0+\alpha_t]$ such that

$$h(u)=h(u_0)+\tfrac{h''(u_0)}{2}(u-u_0)^2+\tfrac{h'''(u^*)}{6}(u-u_0)^3$$

Let $A''\triangleq \sup_{[\frac{u_0}{2};A]}\big|\frac{h'''(u)}{6}\big|$ and $I_t\triangleq\int_{[u_0-\alpha_t;u_0+\alpha_t]}\phi(u)\exp\big[-t\frac{h''(u_0)}{2}(u-u_0)^2\big]du$ We get

$$\int_{[u_0-\alpha_t;u_0+\alpha_t]}\phi(u)\exp\big[-th(u)\big]du\leq \exp\big[A''t\alpha_t^3\big]\exp\big[-th(u_0)\big]I_t$$

and

$$\int_{[u_0-\alpha_t;u_0+\alpha_t]}\phi(u)\exp\big[-th(u)\big]du\geq \exp\big[-A''t\alpha_t^3\big]\exp\big[-th(u_0)\big]I_t.$$

We have

$$\begin{aligned}\Big|I_t-\int_{-\infty}^{+\infty}\phi(u_0)\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du\Big| \\ \leq \int_{[u_0-\alpha_t;u_0+\alpha_t]}|\phi(u)-\phi(u_0)|\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du \\ +\int_{]-\infty;u_0-\alpha_t]\cup[u_0+\alpha_t;+\infty[}\phi(u_0)\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du \\ \leq \underset{t\to+\infty}{\mathrm{o}}\Big(\int_{-\infty}^{+\infty}\exp\big[-t\tfrac{h''(u_0)}{2}(u-u_0)^2\big]du\Big) \\ +\underset{t\to+\infty}{\mathrm{O}}\Big(\exp\big[-\tfrac{h''(u_0)}{2}t{\alpha_t}^2\big]du\Big)\end{aligned}$$

Since we have $\int_{-\infty}^{+\infty}\exp\big[-t\frac{h''(u_0)}{2}(u-u_0)^2\big]du=\sqrt{\frac{2\pi}{th''(u_0)}}$, we obtain

$$I_t=\big[\phi(u_0)+\underset{t\to+\infty}{\mathrm{o}}(1)\big]\sqrt{\tfrac{2\pi}{th''(u_0)}}.$$

• Combining the previous results, we obtain

$$\int_0^A\phi(u)\exp\big[-th(u)\big]du=\big[\phi(u_0)+\underset{t\to+\infty}{\mathrm{o}}(1)\big]\exp\big[-th(u_0)\big]\sqrt{\tfrac{2\pi}{th''(u_0)}}.$$

□

By assumption, we may write $\pi\big(\Delta R\leq x\big)=\exp\big(-C'x^{-\frac{q}{\kappa}}-C'''\big)[1+\eta(x)]$ with $\eta(x)=\underset{x\to0}{\mathrm{o}}(1)$. Let $A'\triangleq\exp(-C''')$, $u_0\triangleq\operatorname{argmin}_{x>0}\big(x+C'x^{-\frac{q}{\kappa}}\big)$, $H\triangleq u_0+C'{u_0}^{-\frac{q}{\kappa}}$ and $\theta\triangleq 2H\lambda^{-\frac{\kappa}{\kappa+q}}$.

From inequality (A.1), we have

$$\begin{aligned}\pi\exp\big(-\lambda\Delta R\big) &= \exp(-\lambda)+\int_0^1\lambda\exp\big(-\lambda x\big)\pi\big(\Delta R\leq x\big)dx \\ &\leq \exp(-\lambda\theta)+\int_0^\theta\lambda\exp\big(-\lambda x\big)\pi\big(\Delta R\leq x\big)dx\end{aligned}\tag{C.1}$$

Besides, we have

$$\begin{aligned}\int_0^\theta \exp\big(-\lambda x\big)\pi\big(\Delta R\le x\big)dx&\\ &= A'\int_0^\theta \exp\big(-\lambda x - C'x^{-\frac{q}{\kappa}}\big)[1+\eta(x)]dx\\ &= A'\int_0^{2H} \exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x+C'x^{-\frac{q}{\kappa}}\big]\big)\big[1+\eta\big(\lambda^{-\frac{\kappa}{\kappa+q}}x\big)\big]dx\end{aligned}$$

For any $\beta>0$, there exists λ_0 such that for any $\lambda>\lambda_0$ and any $x\le\theta$, we have $\big|\eta\big(\lambda^{-\frac{\kappa}{\kappa+q}}x\big)\big|\le\beta$. We obtain

$$\left|\frac{\int_0^\theta \exp(-\lambda x)\pi(\Delta R\le x)dx}{A'\int_0^{2H}\exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x+C'x^{-\frac{q}{\kappa}}\big]\big)dx}-1\right|\le\beta.$$

Using Lemma C.1, we get

(C.2)

$$\begin{aligned}\int_0^\theta \exp\big(-\lambda x\big)\pi\big(\Delta R\le x\big)dx &\underset{\lambda\to+\infty}{\sim} A'\int_0^{2H}\exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x+C'x^{-\frac{q}{\kappa}}\big]\big)dx\\ &\underset{\lambda\to+\infty}{\sim} \exp\big(-\lambda^{\frac{q}{\kappa+q}}H\big)\end{aligned}$$

So inequality (C.1) implies

$$\pi\exp\big(-\lambda\Delta R\big)\underset{\lambda\to+\infty}{\sim}\lambda\exp\big(-\lambda^{\frac{q}{\kappa+q}}H\big). \tag{C.3}$$

From equality (A.3), we have

$$\begin{aligned}\pi\big[\Delta R\exp\big(-\lambda\Delta R\big)\big]=\exp(-\lambda)-\int_0^1\exp\big(-\lambda x\big)\pi\big(\Delta R\le x\big)dx&\\ +\lambda\int_0^1 x\exp\big(-\lambda x\big)\pi\big(\Delta R\le x\big)dx.&\end{aligned}$$

Using similar computations to the one used to prove (C.2) and from the equality

$$\begin{aligned}\int_0^\theta x\exp\big(-\lambda x-C'x^{-\frac{q}{\kappa}}\big)[1+\eta(x)]&\\ =\int_0^{2H}\lambda^{-\frac{\kappa}{\kappa+q}}x\exp\big(-\lambda^{\frac{q}{\kappa+q}}\big[x+C'x^{-\frac{q}{\kappa}}\big]\big)\big[1+\eta\big(\lambda^{-\frac{\kappa}{\kappa+q}}x\big)\big]dx,&\end{aligned}$$

we obtain

$$\pi\big[\Delta R\exp\big(-\lambda\Delta R\big)\big]\underset{\lambda\to+\infty}{\sim}\lambda^{\frac{q}{\kappa+q}}u_0\exp\big(-\lambda^{\frac{q}{\kappa+q}}H\big).$$

Consequently, we have proved $\pi_{-\lambda R}\Delta R\underset{\lambda\to+\infty}{\sim}u_0\lambda^{-\frac{\kappa}{\kappa+q}}$. By definition of u_0, we get $\pi_{-\lambda R}\Delta R\underset{\lambda\to+\infty}{\sim}\big(\frac{qC'}{\kappa\lambda}\big)^{\frac{\kappa}{\kappa+q}}$.

Appendix D. Proof of inequality (6.7)

Let $0<\alpha\le\breve{c}\lambda^{-\frac{\kappa-1}{\kappa+q}}$ for some constant $\breve{c}>0$ to be determined, and $\frac12<\zeta<1$. We use once more the function $h_\alpha(x)\triangleq x-\alpha x^{\frac1\kappa}$ which is minimum at $x_0\triangleq\big(\frac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}$. Let $\nu\triangleq\eta\big\{\big(\frac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}}\big\}$. We have

$$\begin{aligned}\pi\exp\big[-\lambda h_\alpha(\Delta R)\big]&=\pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)\le\zeta\Delta R}\Big\}\\ &\quad+\pi\Big\{\exp\big[-\lambda h_\alpha(\Delta R)\big]\mathbb{1}_{h_\alpha(\Delta R)>\zeta\Delta R}\Big\}\\ &\le\exp\big[-\lambda h_\alpha(x_0)\big]\pi\Big\{\Delta R\le\big(\tfrac{\alpha}{1-\zeta}\big)^{\frac{\kappa}{\kappa-1}}\Big\}+\pi\exp\big[-\lambda\zeta\Delta R\big]\\ &\le\exp\Big\{\lambda(\kappa-1)\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}-C'\big(\tfrac{1-\zeta}{\alpha}\big)^{\frac{q}{\kappa-1}}-C'''\Big\}(1+\nu)\\ &\quad+\pi\exp\big[-\lambda\zeta\Delta R\big]\end{aligned}$$

Using (C.3), we obtain

$$\begin{aligned}\pi_{-\lambda R}&\exp\big(\lambda\alpha\Delta R\big)\\ &\le \lambda^{-1}\exp\Big(H\lambda^{\frac{q}{\kappa+q}}+\lambda(\kappa-1)\big(\tfrac{\alpha}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}-C'\big(\tfrac{1-\zeta}{\alpha}\big)^{\frac{q}{\kappa-1}}\Big)\Big[\breve{C}+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\\ &\quad+\zeta\exp\Big(\lambda^{\frac{q}{\kappa+q}}H\big[1-\zeta^{\frac{q}{\kappa+q}}\big]\Big)\Big[1+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\end{aligned}$$

Let $\zeta=1-\big(\frac{2H}{C'}\big)^{\frac{\kappa-1}{q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}$ so that $C'\big(\frac{1-\zeta}{\alpha}\big)^{\frac{q}{\kappa-1}}=2H\lambda^{\frac{q}{\kappa+q}}$ and let $\breve{c}>0$ such that $(\kappa-1)\big(\frac{\breve{c}}{\kappa}\big)^{\frac{\kappa}{\kappa-1}}\le H$. Then we have

$$\begin{aligned}\pi_{-\lambda R}&\exp\big(\lambda\alpha\Delta R\big)\\ &\le \lambda^{-1}\Big[\breve{C}+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]+\exp\Big\{\lambda^{\frac{q}{\kappa+q}}\tfrac{q}{\kappa+q}H\mathrm{O}(1-\zeta)\Big\}\Big[1+\underset{\lambda\to+\infty}{\mathrm{o}}(1)\Big]\\ &=\underset{\lambda\to+\infty}{\mathrm{O}}\Big(\exp\Big\{\breve{C}\lambda^{\frac{q}{\kappa+q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}\Big\}\Big),\end{aligned}$$

hence

$$\log\pi_{-\lambda R}\exp\big(\lambda\alpha\Delta R\big)=\underset{\lambda\to+\infty}{\mathrm{O}}\Big(\lambda^{\frac{q}{\kappa+q}}\alpha\lambda^{\frac{\kappa-1}{\kappa+q}}\Big).$$

Appendix E. Another way of getting the right order

This section proves that by using well-known results, we can obtain a lower bound having the same spirit as Lemma 5.1 but without proper constants.

Applying Lemma 6.4 to the set of probability distributions

$$\mathcal{D}\triangleq\big\{\mathbb{P}^{\otimes N}:\mathbb{P}\in\mathcal{D}'\big\}$$

where $\mathcal{D}'\triangleq\big\{\mathbb{P}_{\sigma_1^m}:\sigma_1^m\in\mathcal{S}\subset\{-1;+1\}^m\big\}$ and $\mathcal{S}$ satisfies $\delta(\Sigma,\Sigma')\ge\frac{m}{4}$ for any $\Sigma\ne\Sigma'\in\mathcal{S}$ and $|\mathcal{S}|=\lfloor e^{\frac{m}{8}}\rfloor$. From Lemma 6.3, such a set $\mathcal{S}$ exists. With any estimator $\hat{f}:\mathcal{Z}^N\to\mathcal{F}(\mathcal{X},\mathcal{Y})$, we can associate an estimator $\hat{T}:\mathcal{Z}^N\to\mathcal{D}$ defined as $\hat{T}(Z_1^N)=\mathbb{P}^{\otimes N}$, where $\mathbb{P}\in\mathcal{D}'$ minimizes $\mu[\xi(X)\mathbb{1}_{f^*_{\mathbb{P}}(X)\ne\hat{f}(Z_1^N)(X)}]$.

By Birgé's lemma, we have $\sup_{\mathbb{P}\in\mathcal{D}'}\mathbb{P}^{\otimes N}\big[\hat{T}(Z_1^N)\ne\mathbb{P}\big]\ge 0.36\wedge\Big(1-\frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)$. Now, when $\hat{T}(Z_1^N)\ne\mathbb{P}$, we have $R_{\mathbb{P}}(\hat{f})-R_{\mathbb{P}}(f^*_{\mathbb{P}})\ge\frac{m}{8}wb'$. Therefore, we get

$$\sup_{\mathbb{P}\in\mathcal{D}}\big\{\mathbb{P}^{\otimes N}R_{\mathbb{P}}(\hat{f})-R_{\mathbb{P}}(f^*_{\mathbb{P}})\big\}\ge\frac{m}{8}wb'\Big[0.36\wedge\Big(1-\frac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\Big)\Big]. \tag{E.1}$$

For any $\mathbb{P}\ne\mathbb{Q}\in\mathcal{D}'$, we have

$$K(\mathbb{P},\mathbb{Q})\le N\mu\Big[\xi(X)\log\Big(\frac{1+\xi(X)}{1-\xi(X)}\Big)\mathbb{1}_{X\notin\mathcal{X}_0}\Big]\le Nmwb'\log\Big(\frac{1+B}{1-B}\Big),$$

where $B\triangleq\sup_{x\in\mathcal{X}-\mathcal{X}_0}\xi(x)$. If we assume that $B\le Cb<1$, we get

$$K(\mathbb{P},\mathbb{Q})\le CNmwb^2$$

for some constant $C>0$. Since we have $|\mathcal{D}|=\lfloor e^{\frac{m}{8}}\rfloor$, we obtain

$$\tfrac{K_{\mathcal{D}}}{|\mathcal{D}|\log|\mathcal{D}|}\le\tfrac{C}{\log\lfloor e^{\frac{m}{8}}\rfloor}Nmwb^2\le C'Nwb^2$$

for m large enough and some constant $C'>0$. So we obtain the right order to the extent that when the quantity Nwb^2 is small enough, the order of the bound is given by the product mwb.

Appendix F. Proof of Theorem 5.2

• When $L = 0$: let $x_0, x_1, \ldots, x_{V-1}$ denote the V points shattered by the model. Let us take

$$\left\{\begin{array}{l} m = V-1 \\ \mathcal{X}_0 = \mathcal{X} - \{x_1, \ldots, x_{V-1}\} \\ \mathcal{X}_j = \{x_j\} \\ \mu(\mathcal{X}_j) = w \qquad \text{for any } j \in \{1, \ldots, m\} \\ \mu(\{x_0\}) = 1 - mw \\ b = 1 \qquad (\xi \equiv 1) \end{array}\right. ,$$

where w is a free positive parameter which satisfies $mw \leq 1$ (since μ is a probability distribution). By noticing that

$$1 - V\Big(\mathbb{P}^{\otimes N}_{-1,1,\ldots,1}, \mathbb{P}^{\otimes N}_{1,1,\ldots,1}\Big) = \mu^{\otimes N}\big(\text{for any } i \in \{1, \ldots, N\}, X_i \notin \mathcal{X}_j\big) = (1-w)^N$$

and using inequality (6.26), we obtain

$$\sup_{\mathbb{P}\in\mathcal{P}}\Big\{\mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*_{\mathbb{P}})\Big\} \quad \geq \quad \tfrac{V-1}{2} \sup_{w \leq \frac{1}{V-1}} \left\{ w\Big(1-w\Big)^N \right\}$$

This supremum is attained for $w = \frac{1}{N+1}$ when $N \geq (V-2) \vee 1$ and for $w = \frac{1}{V-1}$ otherwise.

• When $0 \leq L \leq \frac{1}{2}$: once more, $x_0, x_1, \ldots, x_{V-1}$ denote the V points shattered by the model. This time, we take

$$\left\{\begin{array}{l} m = V-1 \\ \mathcal{X}_0 = \mathcal{X} - \{x_1, \ldots, x_{V-1}\} \\ \mathcal{X}_j = \{x_j\} \\ \mu(\mathcal{X}_j) = w \qquad \text{for any } j \in \{1, \ldots, m\} \\ \mu(\{x_0\}) = 1 - mw \\ \xi(x) = \left\{\begin{array}{l} b_0 \text{ when } x \in \mathcal{X}_0 \\ b \text{ otherwise} \end{array}\right. \end{array}\right. ,$$

where w is a free positive parameter which satisfies $mw \leq 1$ (since μ is a probability distribution) and b and b_0 belong to $[0;1]$. Since we have

$$L = \frac{1}{2} mw(1-b) + \frac{1}{2}(1-mw)(1-b_0)$$

and $b_0 \in [0;1]$, the parameters m, w and b should satisfy

$$mw(1-b) \leq 2L \leq 1 - mwb.$$

Since this condition implies that $mw \leq 1$, we have the following lower bound

$$\sup_{\mathbb{P}\in\mathcal{P}}\Big\{\mathbb{P}^{\otimes N} R_{\mathbb{P}}(\hat{f}) - R_{\mathbb{P}}(f^*)\Big\} \quad \geq \quad \sup_{\substack{w \geq 0 \\ 0 \leq b \leq 1 \\ mw(1-b) \leq 2L \leq 1 - mw\beta}} \tfrac{1}{2} mwb\big[1 - b\sqrt{Nw}\big].$$

From this lower bound, one can recover the first assertion of Theorem 5.2 with a constant slightly worsened (due to the upper bound (6.27)). We will now slightly weaken this result in order to get a simple lower bound. Introduce $x = b^2 wN$. The

previous supremum can be written as

$$
\begin{aligned}
&\sup_{\substack{x\geq 0\\ 0<b\leq 1\\ \frac{(V-1)x}{bN}\frac{1-b}{b}\leq 2L\\ \frac{(V-1)x}{bN}\leq 1-2L}} \tfrac{1}{2}\tfrac{(V-1)x}{bN}\big[1-\sqrt{x}\big] \\
&\qquad\geq \sup_{\substack{x\geq 0\\ 0<b<1\\ \frac{(V-1)x}{bN}\frac{1-b}{b}=2L\\ 2L\frac{b}{1-b}\leq 1-2L}} \tfrac{1}{2}\tfrac{(V-1)x}{bN}\big[1-\sqrt{x}\big] \\
&\qquad= \sup_{\substack{x>0\\ b=\frac{2}{1+\sqrt{1+\frac{8LN}{(V-1)x}}}\\ b\leq 1-2L}} \tfrac{1}{2}\tfrac{(V-1)x}{bN}\big[1-\sqrt{x}\big] \\
&\qquad= \sup_{0<x\leq\frac{(1-2L)^2N}{V-1}} \tfrac{(V-1)x}{4N}\Big(1+\sqrt{1+\tfrac{8LN}{(V-1)x}}\Big)\big[1-\sqrt{x}\big] \\
&\qquad> \sup_{0<x\leq\frac{(1-2L)^2N}{V-1}} \sqrt{\tfrac{L(V-1)x}{2N}}\big[1-\sqrt{x}\big] \qquad \text{(A)} \\
&\qquad= \begin{cases} \sqrt{\frac{L(V-1)}{32N}} & \text{when } \frac{(1-2L)^2N}{V-1}\geq\frac{1}{4} \\ \sqrt{\frac{L(1-2L)^2}{8}} & \text{otherwise.}\end{cases}
\end{aligned}
$$

Note that the step (A) prevents us to have a good lower bound when $L=\mathrm{o}\big(\frac{V-1}{N}\big)$. In this last case, the lower bound (A) can be replaced with $\frac{V-1}{2N}x(1-\sqrt{x})$ which, by taking $x=\frac{1}{4}$, leads to the desired bound $\frac{V-1}{16N}$.

References

1. J.-Y. Audibert, *Data-dependent generalization error bounds for (noisy) classification: the PAC-Bayesian approach*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2004.
2. L. Birgé, *A new look at an old result: Fano's lemma*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2001.
3. O. Catoni, *Statistical learning theory and stochastic optimization,* Lecture notes, Saint-Flour summer school on Probability Theory, 2001, Springer, to appear.
4. ———, *A PAC-Bayesian approach to adaptive classification*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2003.
5. L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*, Springer-Verlag, 2000.
6. R.M. Dudley, *Central limit theorems for empirical measures*, Ann. Probab. **6** (1978), 899–929.
7. D. Haussler, *Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension*, Journal of Combinatorial Theory **69** (1995), 217–232, Series A.
8. C. Huber, *Lower bounds for function estimation*, Research paper in probability and statistics: Festschrift for Lucien Le Cam (1996), 245–258.
9. V.I. Koltchinskii, *On the central limit theorem for empirical measures*, Theory Probab. Math. Stat. **24** (1981), 71–82.
10. G. Lugosi, *Concentration-of-measure inequalities*, 2003, Lecture notes, Machine Learning Summer School, Canberra.
11. E. Mammen and A.B. Tsybakov, *Smooth discrimination analysis*, Ann. Stat. **27** (1999), 1808–1829.
12. P. Massart and E. Nédélec, *Risk bounds for statistical learning*, Available from http://www.math.u-psud.fr/∼massart/margin.pdf, 2003.
13. D. A. McAllester, *PAC-Bayesian model averaging*, Morgan Kaufmann Publishers, 1999.
14. D. Pollard, *A central limit theorem for empirical measures*, J. Aust. Math. Soc., Ser. A **33** (1982), 235–248.

15. A.B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Stat. **32** (2004), no. 1.
16. A.B. Tsybakov and S. van de Geer, *Square root penalty: adaptation to the margin in classification and in edge estimation*, Preprint, Laboratoire de Probabilité et Modelès Aléatoires, 2004.
17. A. van der Vaart and J. Wellner, *Weak convergence and empirical processes with application to statistics*, John Wiley & Sons, New York, 1996.