Fast learning rates in statistical inference through aggregation

J.-Y. Audibert

Research Report 06-20
3 novembre 2006

Ecole Nationale
Des Ponts et Chaussées
Centre d'Enseignement et de Recherche
en Technologies de l'Information et Systèmes

CERTIS, ENPC,
77455 Marne la Vallee, France,

Fast learning rates in statistical inference through aggregation

Vitesses rapides d'apprentissage par méthodes d'agrégation

J.-Y. Audibert[1]

[1]CERTIS, ENPC, 77455 Marne la Vallee, France, http://www.enpc.fr/certis/

Abstract

We develop minimax optimal risk bounds for the general learning task consisting in predicting as well as the best function in a reference set $\mathcal{G}$ up to the smallest possible additive term, called the convergence rate. When the reference set is finite and when n denotes the size of the training data, we show that this minimax convergence rate is $\left(\frac{\log|\mathcal{G}|}{n}\right)^v$ where $0 < v \leq 1$ depends on the convexity of the loss function and on the level of noise in the output distribution.

The risk upper bounds are based on a sequential randomized algorithm, which at each step concentrates on functions having both low risk and low variance with respect to the previous step prediction function. Our analysis puts forward the links between the probabilistic and worst-case viewpoints, and allows to obtain risk bounds unachievable with the standard statistical learning approach. One of the key idea of this work is to use probabilistic inequalities with respect to appropriate (Gibbs) distributions on the prediction function space instead of using them with respect to the distribution generating the data.

The risk lower bounds are based on refinements of the Assouad's lemma taking particularly into account the properties of the loss function. Our key example to illustrate the upper and lower bounds is to consider the L_q-regression setting for which an exhaustive analysis of the convergence rates is given while q describes $[1; +\infty[$.

Résumé

Nous développons des bornes optimales au sens minimax pour le problème d'apprentissage statistique suivant : prédire aussi bien que la meilleure fonction de prédiction d'un ensemble de référence, au plus petit terme additif près appelé la vitesse de convergence. Lorsque l'ensemble de référence est fini et lorsque n désigne la taille de l'ensemble d'apprentissage, nous montrons que la vitesse de convergence est de l'ordre de $\left(\frac{\log|\mathcal{G}|}{n}\right)^v$ où $0 < v \leq 1$ dépend de la convexité de la fonction de perte et de la qualité des données d'apprentissage.

Les bornes supérieures sur l'erreur de généralisation sont basées sur un algorithme séquentiel, qui à chaque étape, se concentre sur les meilleures fonctions de prédiction et les plus proches de la fonction de prédiction construite à l'étape d'avant.

Notre étude met en avant les liens entre le point de vue probabiliste et l'analyse du pire des cas et permet d'obtenir des bornes sur l'erreur de généralisation que ne donnent pas l'approche standard de la théorie de l'apprentissage statistique. Une des idées fondamentales de ce travail est d'utiliser les bornes probabilistes par rapport à des distributions bien choisies sur l'espace des fonctions de prédiction au lieu de les employer par rapport à la distribution générant les données.

Les bornes inférieures sur l'erreur de généralisation sont obtenues par une amélioration du lemme d'Assouad qui notamment tient compte de la fonction de perte du problème d'apprentissage. Notre exemple clé pour illustrer nos bornes inférieures et supérieures est de considérer les fonctions de pertes L_q (i.e. $\ell(y, y') = |y - y'|^q$), $q \geq 1$, pour lesquelles nous donnons une analyse détaillée des vitesses de convergence.

Contents

1 Introduction

We are given a family $\mathcal{G}$ of functions and we want to learn from data a function that predicts as well as the best function in $\mathcal{G}$ up to some additive term called the convergence rate. Even when the set $\mathcal{G}$ is finite, this learning task is crucial since

- any continuous set of prediction functions can be viewed through its covering nets with respect to (w.r.t.) appropriate (pseudo-)distances and these nets are generally finite.

- one way of doing model selection among a finite family of submodels is to cut the training set into two parts, use the first part to learn the best prediction function of each submodel and use the second part to learn a prediction function which performs as well as the best of the prediction functions learned on the first part of the training set.

From this last item, our learning task for finite $\mathcal{G}$ is often referred to as model selection aggregation. It has two well-known variants. Instead of looking for a function predicting as well as the best in $\mathcal{G}$, these variants want to perform as well as the best convex combination of functions in $\mathcal{G}$ or as well as the best linear combination of functions in $\mathcal{G}$. These three aggregation tasks are linked in several ways (see [39] and references within).

Nevertheless, among these learning tasks, model selection aggregation has rare properties. First, in general an algorithm picking functions in the set $\mathcal{G}$ is not optimal (see e.g. [19, p.14]). This means that the estimator has to look at an enlarged set of prediction functions. Secondly, in the statistical community, the only known optimal algorithms are all based on a Cesaro mean of Bayesian estimators (also referred to as progressive mixture rule). Thirdly, the proof of their optimality is not achieved by the most prominent tool in statistical learning theory: bounds on the supremum of empirical processes (see [41], and refined works as [10, 35, 36, 15] and references within).

The idea of the proof, which comes back to Barron [8], is based on a chain rule and appeared to be successful for least square and entropy losses [18, 19, 9, 45, 17] and for general loss in [32].

In the online prediction with expert advice setting, without any probabilistic assumption on the generation of the data, appropriate weighting methods have been showed to behave as well as the best expert up to a minimax-optimal additive remainder term (see [37, 25] and references within). In this worst-case context, amazingly sharp constants have been found (see in particular [31, 23, 24, 46]). These results are expressed in cumulative loss and can be transposed to model selection aggregation to the extent that the expected risk of the randomized procedure based on sequential predictions is proportional to the expectation of the cumulative loss of the sequential procedure (see Lemma 4.3 for precise statement).

This work presents a sequential algorithm, which iteratively updates a prior distribution put on the set of prediction functions. Contrarily to previously mentioned works, these updates take into account the variance of the task. As a consequence, posterior distributions concentrate on simultaneously low risk functions and functions close to the previously drawn prediction function. This conservative law is not surprising in view of previous works on high dimensional statistical tasks, such as wavelet thresholding, shrinkage procedures, iterative compression schemes ([4]), iterative feature selection ([1]).

The paper is organized as follows. Section 2 introduces the notation and the existing algorithms. Section 3 proposes a unifying setting to combine worst-case analysis tight results and probabilistic tools. It details our randomized estimator and gives a sharp expectation bound. In Sections 4 and 5, we show how to apply our main result under assumptions coming respectively from sequential prediction and model selection aggregation. Section 6 contains algorithms that satisfy sharp standard-style generalization error bounds. To the author's knowledge, these bounds are not achievable with classical statistical learning approach based on supremum of empirical processes. Here the main trick is to use probabilistic inequalities w.r.t. appropriate distributions on the prediction function space instead of using them w.r.t. the distribution generating the data. Section 7 presents an improved bound for L_q-regression ($q > 1$) when the noise has just a bounded moment of order $s \geq q$. This last assumption is much weaker than the traditional exponential moment assumption. Section 8 refines Assouad's lemma in order to obtain sharp constants and to take into account the properties of the loss function of the learning task. We illustrate our results by providing lower bounds matching the upper bounds obtained in the previous sections.

2 Notation and existing algorithms

We assume that we observe n pairs $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ of input-output and that each pair has been independently drawn from the same unknown distribution denoted P. The input and output space are denoted respectively $\mathcal{X}$ and $\mathcal{Y}$, so that P is a probability distribution on the product space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. The target of a learning algorithm is to predict the output Y associated with an input X for pairs (X, Y) drawn from the distribution P. In this work, Z_{n+1} will denote a random variable independent from the training set $Z_1^n \triangleq (Z_1, \ldots, Z_n)$ and with the same distribution P. The quality of a prediction function $g : \mathcal{X} \to \mathcal{Y}$ is measured by the *risk* (also called expected loss or regret):

$$R(g) \triangleq \mathbb{E}_{Z \sim P}\, L(Z, g),$$

where $L(Z, g)$ assesses the loss of considering the prediction function g on the data $Z \in \mathcal{Z}$. The symbol $\triangleq$ is used to underline that the equality is a definition. When there is no ambiguity on the distribution that a random variable has, the expectation w.r.t. this distribution will simply be written by indexing the expectation sign $\mathbb{E}$ by the random variable. For instance, we can write $R(g) \triangleq \mathbb{E}_Z\, L(Z, g)$. We use $L(Z, g)$ rather than $L[Y, g(X)]$ to underline that our results are not restricted to non-regularized losses, where we call non-regularized loss a loss that can be written as $\ell[Y, g(X)]$ for some function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

For any $i \in \{0, \ldots, n\}$, the *cumulative loss* suffered by the prediction function g on the first i pairs of input-output, denoted Z_1^i for short, is

$$\Sigma_i(g) \triangleq \sum_{j=1}^{i} L(Z_j, g),$$

where by convention we take Σ_0 identically equal to zero. The symbol $\equiv$ is used to underline when a function is identical to a constant (e.g. $\Sigma_0 \equiv 0$). With slight abuse, a symbol denoting a constant function may be used to denote the value of this function.

The n-fold product of a distribution μ, which is the distribution of a vector consisting in n i.i.d. realizations of μ, is denoted $\mu^{\otimes n}$. For instance the distribution of $(Z_1, \ldots, Z_n)$ is $P^{\otimes n}$.

The symbol C will denote some positive constant whose value may differ from line to line. The set of non-negative real numbers is denoted $\mathbb{R}_+ = [0; +\infty[$. We define $\lfloor x \rfloor$ as the largest integer k such that $k \le x$. To shorten notation, any finite sequence $a_1, \ldots, a_n$ will occasionally be denoted a_1^n. For instance, the training set is Z_1^n.

To handle possibly continuous set $\mathcal{G}$, we consider that $\mathcal{G}$ is a measurable space and that we have some *prior distribution* π on it. The set of probability distributions on $\mathcal{G}$ will be denoted $\mathcal{M}$. The *Kullback-Leibler divergence* between a distribution $\rho \in \mathcal{M}$ and the prior distribution π is

$$K(\rho, \pi) \triangleq \begin{cases} \mathbb{E}_{g\sim\rho} \log\left(\frac{\rho}{\pi}(g)\right) & \text{if } \rho \ll \pi, \\ +\infty & \text{otherwise} \end{cases}$$

where $\frac{\rho}{\pi}$ denotes the density of ρ w.r.t. π when it exists (i.e. $\rho \ll \pi$). For any $\rho \in \mathcal{M}$, we have $K(\rho, \pi) \ge 0$ and when π is the uniform distribution on a finite set $\mathcal{G}$, we also have $K(\rho, \pi) \le \log|\mathcal{G}|$. The Kullback-Leibler divergence satisfies the duality formula (see e.g. [20, p.10]): for any real-valued measurable function h defined on $\mathcal{G}$,

$$\inf_{\rho\in\mathcal{M}} \left\{\mathbb{E}_{g\sim\rho}\, h(g) + K(\rho, \pi)\right\} = -\log \mathbb{E}_{g\sim\pi}\, e^{-h(g)}. \tag{2.1}$$

and that the infimum is reached for the *Gibbs distribution*

$$\pi_{-h}(dg) \triangleq \frac{e^{-h(g)}}{\mathbb{E}_{g'\sim\pi}\, e^{-h(g')}} \cdot \pi(dg). \tag{2.2}$$

Intuitively, the Gibbs distribution π_{-h} concentrates on prediction functions g that are close to minimizing the function $h : \mathcal{G} \to \mathbb{R}$.

For any $\rho \in \mathcal{M}$, $\mathbb{E}_{g\sim\rho}\, g : x \mapsto \mathbb{E}_{g\sim\rho}\, g(x) = \int g(x)\rho(dg)$ is called a mixture of prediction functions. When $\mathcal{G}$ is finite, a mixture is simply a convex combination. Throughout this work, whenever we consider mixtures of prediction functions, we implicitly assume that $\mathbb{E}_{g\sim\rho}\, g(x)$ belongs to $\mathcal{Y}$ for any x so that the mixture is a prediction function. This is typically the case when $\mathcal{Y}$ is an interval of $\mathbb{R}$.

We will say that the loss function is convex when the function $g \mapsto L(z, g)$ is convex for any $z \in \mathcal{Z}$, equivalently $L(z, \mathbb{E}_{g\sim\rho}\, g) \le \mathbb{E}_{g\sim\rho}\, L(z, g)$ for any $\rho \in \mathcal{M}$ and $z \in \mathcal{Z}$. In this work, we do not assume the loss function to be convex except when it is explicitly mentioned.

The algorithm used to prove optimal convergence rates for several different losses (see e.g. [18, 19, 9, 45, 17, 32]) is the following:

Algorithm A**:** Let $\lambda > 0$. Predict according to $\frac{1}{n+1}\sum_{i=0}^{n} \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}}\, g$, where we recall that Σ_i maps a function $g \in \mathcal{G}$ to its cumulative loss up to time i.

In other words, for a new input x, the prediction of the output given by Algorithm A is $\frac{1}{n+1}\sum_{i=0}^{n} \frac{\int g(x)e^{-\lambda\Sigma_i(g)}\pi(dg)}{\int e^{-\lambda\Sigma_i(g)}\pi(dg)}$. To give the optimal convergence rate, the parameter λ and the distribution π should be appropriately chosen. When $\mathcal{G}$ is finite, the estimator belongs to the convex hull of the set $\mathcal{G}$.

From Vovk, Haussler, Kivinen and Warmuth works ([43, 31, 44]) and the link between cumulative loss in online setting and expected risk in the batch setting (see later Lemma 4.3), an "optimal" algorithm is:

Algorithm B**:** Let $\lambda > 0$. For any $i \in \{0, \ldots, n\}$, let $\hat{h}_i$ be a prediction function such that

$$\forall\, z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \le -\frac{1}{\lambda} \log \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}}\, e^{-\lambda L(z,g)}.$$

If one of the $\hat{h}_i$ does not exist, the algorithm is said to fail. Otherwise it predicts according to $\frac{1}{n+1}\sum_{i=0}^{n} \hat{h}_i$.

In particular, for appropriate $\lambda > 0$, this algorithm does not fail when the loss function is the square loss (i.e. $L(z, g) = [y - g(x)]^2$) and when the output space is bounded. Algorithm B is based on the same Gibbs distribution $\pi_{-\lambda\Sigma_i}$ as Algorithm A. Besides, in [31, Example 3.13], it is shown that Algorithm A is not in general a particular case of Algorithm B, and that Algorithm B will not generally produce a prediction function in the convex hull of $\mathcal{G}$ unlike Algorithm

A. In Sections 4 and 5, we will see how both algorithms are connected to the generic algorithm presented in the next section.

We assume that the set, denoted $\bar{\mathcal{G}}$, of all prediction functions has been equipped with a σ-algebra. Let $\mathcal{D}$ be the set of all probability distributions on $\bar{\mathcal{G}}$. By definition, a randomized algorithm produces a prediction function drawn according to a probability in $\mathcal{D}$. Let $\mathcal{P}$ be a set of probability distributions on $\mathcal{Z}$ in which we assume that the true unknown distribution generating the data is. The learning task is essentially described by the 3-tuple $(\mathcal{G}, L, P)$ since we look for an estimator (or algorithm) $\hat{g}$ such that

$$\sup_{P\in\mathcal{P}} \left\{ \mathbb{E}_{Z_1^n} R(\hat{g}_{Z_1^n}) - \min_{g\in\mathcal{G}} \ R(g) \right\}$$

is minimized, where we recall that $R(g) \triangleq \mathbb{E}_{Z\sim P} \ L(Z,g)$. To shorten notation, when no confusion can arise, the dependence of $\hat{g}_{Z_1^n}$ w.r.t. the training sample Z_1^n will be dropped and we will simply write $\hat{g}$. This means that we use the same symbol for both the algorithm and the prediction function produced by the algorithm on a training sample.

Finally we implicitly assume that the quantities we manipulate are measurable: in particular, we assume that a prediction function is a measurable function from $\mathcal{X}$ to $\mathcal{Y}$, the mapping $(x,y,g) \mapsto L[(x,y),g]$ is measurable, the estimators considered in our lower bounds are measurable, ...

3 The algorithm and its generalization error bound

The aim of this section is to build an algorithm with the best possible convergence rate regardless of computational issues. For any $\lambda > 0$, let δ_λ be a real-valued function defined on $\mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}}$ that satisfies

$$\begin{aligned} &\forall \rho \in \mathcal{M} \quad \exists\, \hat{\pi}(\rho) \in \mathcal{D} \\ &\quad \sup_{P\in\mathcal{P}} \left\{ \mathbb{E}_{Z\sim P}\, \mathbb{E}_{g'\sim\hat{\pi}(\rho)} \ \log \mathbb{E}_{g\sim\rho} \ e^{\lambda\left[L(Z,g')-L(Z,g)-\delta_\lambda(Z,g,g')\right]} \right\} \le 0. \end{aligned} \tag{3.1}$$

Condition (3.1) is our probabilistic version of the generic algorithm condition in the online prediction setting (see [43, proof of Theorem 1] or more explicitly in [31, p.11]), in which we added the variance function δ_λ. Our results will be all the sharper as this variance function is small. To make (3.1) more readable, let us say for the moment that

- without any assumption on $\mathcal{P}$, for several usual "strongly" convex loss functions, we may take $\delta_\lambda \equiv 0$ provided that λ is a small enough constant (see Section 4).

- Inequality (3.1) can be seen as a "small expectation" inequality. The usual viewpoint is to control the quantity $L(Z, g)$ by its expectation w.r.t. Z and a variance term. Here, roughly, $L(Z, g)$ is mainly controlled by $L(Z, g')$ where g' is appropriately chosen through the choice of $\hat{\pi}(\rho)$, plus the additive term δ_λ. By definition this additive term does not depend on the particular probability distribution generating the data and leads to empirical compensation.

- in the examples we will be interested in throughout this work, $\hat{\pi}(\rho)$ will be either equal to ρ or to a Dirac distribution on some function, which is *not necessarily in $\mathcal{G}$*.

- for any loss function L, any set $\mathcal{P}$ and any $\lambda > 0$, one may choose $\delta_\lambda(Z, g, g') = \frac{\lambda}{2}\big[L(Z, g) - L(Z, g')\big]^2$ (see Section 6).

Our results concern the following algorithm, in which we recall that π is a prior distribution put on the set $\mathcal{G}$.

Generic Algorithm:

1. Let $\lambda > 0$. Define $\hat{\rho}_0 \triangleq \hat{\pi}(\pi)$ in the sense of (3.1) and draw a function $\hat{g}_0$ according to this distribution. Let $S_0(g) = 0$ for any $g \in \mathcal{G}$.

2. For any $i \in \{1, \ldots, n\}$, iteratively define

$$S_i(g) \triangleq S_{i-1}(g) + L(Z_i, g) + \delta_\lambda(Z_i, g, \hat{g}_{i-1}) \quad \text{for any } g \in \mathcal{G}. \tag{3.2}$$

   and

$$\hat{\rho}_i \triangleq \hat{\pi}(\pi_{-\lambda S_i}) \qquad \text{(in the sense of (3.1))}$$

   and draw a function $\hat{g}_i$ according to the distribution $\hat{\rho}_i$.

3. Predict with a function drawn according to the uniform distribution on the finite set $\{\hat{g}_0, \ldots, \hat{g}_n\}$.

Remark 3.1. When $\delta_\lambda(Z, g, g')$ does not depend on g, we recover a more standard-style algorithm to the extent that we then have $\pi_{-\lambda S_i} = \pi_{-\lambda \Sigma_i}$. Precisely our algorithm becomes the randomized version of Algorithm A. When $\delta_\lambda(Z, g, g')$ depends on g, the posterior distributions tend to concentrate on functions having small risk and small variance term. In Section 6, we will take $\delta_\lambda(Z, g, g') = \frac{\lambda}{2}\big[L(Z, g) - L(Z, g')\big]^2$. This choice implies a conservative mechanism: roughly, with high probability, among functions having low cumulative risk Σ_i, $\hat{g}_i$ will be chosen close to $\hat{g}_{i-1}$.

For any $i \in \{0, \ldots, n\}$, the quantities S_i, $\hat{\rho}_i$ and $\hat{g}_i$ depend on the training data only through Z_1^i, where we recall that Z_1^i denotes $(Z_1, \ldots, Z_i)$. Besides they are also random to the extent that they depend on the draws of the functions $\hat{g}_0, \ldots, \hat{g}_{i-1}$.

Our randomized algorithm produces a prediction function which has three causes of randomness: the training data, the way $\hat{g}_i$ is obtained (step 2) and the uniform draw (step 3). For fixed Z_1^i (i.e. conditional to Z_1^i), let Ω_i denote the joint distribution of $\hat{g}_0^i = (\hat{g}_0, \ldots, \hat{g}_i)$. Let $\hat{\mu}$ be the randomizing distribution of our algorithm, i.e. the distribution on $\bar{\mathcal{G}}$ corresponding to the last two causes of randomness. From the previous definitions, for any function $h : \bar{\mathcal{G}} \to \mathbb{R}$, we have $\mathbb{E}_{g\sim\hat{\mu}}\ h(g) = \mathbb{E}_{\hat{g}_0^n\sim\Omega_n}\ \frac{1}{n+1}\sum_{i=0}^n h(\hat{g}_i)$. Our main upper bound controls the expected risk $\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\ R(g)$ of our iteratively randomized generic procedure.

Theorem 3.1. *Let $\Delta_\lambda(g, g') \triangleq \mathbb{E}_{Z\sim P}\ \delta_\lambda(Z, g, g')$ for $g \in G$ and $g' \in \bar{\mathcal{G}}$, where we recall that δ_λ is defined as (3.1). The expected risk of the generic algorithm satisfies*

$$\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\hat{\mu}}\ R(g') \le \min_{\rho\in\mathcal{M}} \left\{ \mathbb{E}_{g\sim\rho}\ R(g) + \mathbb{E}_{g\sim\rho}\ \mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\hat{\mu}}\ \Delta_\lambda(g, g') + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \right\} \tag{3.3}$$

In particular, when $\mathcal{G}$ is finite and when the loss function L and the set $\mathcal{P}$ are such that $\delta_\lambda \equiv 0$, by taking π uniform on $\mathcal{G}$, we get

$$\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\ R(g) \le \min_{\mathcal{G}} R + \tfrac{\log|\mathcal{G}|}{\lambda(n+1)} \tag{3.4}$$

Proof. Let $\mathcal{E}$ denote the expected risk of the generic algorithm:

$$\mathcal{E} \triangleq \mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\ R(g) = \tfrac{1}{n+1}\sum_{i=0}^n \mathbb{E}_{Z_1^i}\mathbb{E}_{\hat{g}_0^i\sim\Omega_i}\ R(\hat{g}_i).$$

We recall that Z_{n+1} is a random variable independent from the training set Z_1^n and with the same distribution P. Let S_{n+1} be defined by (3.2) for $i = n+1$. To shorten formulae, let $\hat{\pi}_i \triangleq \pi_{-\lambda S_i}$ so that by definition we have $\hat{\rho}_i = \hat{\pi}(\hat{\pi}_i)$ in the sense of (3.1). Inequality (3.1) implies that

$$\mathbb{E}_{g'\sim\hat{\pi}(\rho)}\ R(g') \le -\tfrac{1}{\lambda}\mathbb{E}_Z\mathbb{E}_{g'\sim\hat{\pi}(\rho)}\ \log\mathbb{E}_{g\sim\rho}\ e^{-\lambda[L(Z,g)+\delta_\lambda(Z,g,g')]}.$$

So for any $i \in \{0, \ldots, n\}$, for fixed $\hat{g}_0^{i-1} = (\hat{g}_0, \ldots, \hat{g}_{i-1})$ and fixed Z_1^i, we have

$$\mathbb{E}_{g'\sim\hat{\rho}_i}\ R(g') \le -\tfrac{1}{\lambda}\mathbb{E}_{Z_{i+1}}\mathbb{E}_{g'\sim\hat{\rho}_i}\ \log\mathbb{E}_{g\sim\hat{\pi}_i}\ e^{-\lambda[L(Z_{i+1},g)+\delta_\lambda(Z_{i+1},g,g')]}$$

Taking the expectations w.r.t. $(Z_1^i, \hat{g}_0^{i-1})$, we get

$$\begin{aligned}\mathbb{E}_{Z_1^i}\mathbb{E}_{\hat{g}_0^i}R(\hat{g}_i) &= \mathbb{E}_{Z_1^i}\mathbb{E}_{\hat{g}_0^{i-1}}\mathbb{E}_{g'\sim\hat{\rho}_i}\ R(g') \\ &\le -\tfrac{1}{\lambda}\mathbb{E}_{Z_1^{i+1}}\mathbb{E}_{\hat{g}_0^i}\log\mathbb{E}_{g\sim\hat{\pi}_i}\ e^{-\lambda[L(Z_{i+1},g)+\delta_\lambda(Z_{i+1},g,\hat{g}_i)]}.\end{aligned}$$

Consequently, by the chain rule (i.e. cancellation in the sum of logarithmic terms; [8]) and by intensive use of Fubini's theorem, we get

$$\begin{aligned}
\mathcal{E} &= \tfrac{1}{n+1}\sum_{i=0}^{n}\mathbb{E}_{Z_1^i}\mathbb{E}_{\hat{g}_0^i}R(\hat{g}_i)\\
&\le -\tfrac{1}{\lambda(n+1)}\sum_{i=0}^{n}\mathbb{E}_{Z_1^{i+1}}\mathbb{E}_{\hat{g}_0^i}\log\mathbb{E}_{g\sim\hat{\pi}_i}\, e^{-\lambda[L(Z_{i+1},g)+\delta_\lambda(Z_{i+1},g,\hat{g}_i)]}\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{Z_1^{n+1}}\mathbb{E}_{\hat{g}_0^n}\sum_{i=0}^{n}\log\mathbb{E}_{g\sim\hat{\pi}_i}\, e^{-\lambda[L(Z_{i+1},g)+\delta_\lambda(Z_{i+1},g,\hat{g}_i)]}\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{Z_1^{n+1}}\mathbb{E}_{\hat{g}_0^n}\sum_{i=0}^{n}\log\left(\tfrac{\mathbb{E}_{g\sim\pi}\, e^{-\lambda S_{i+1}(g)}}{\mathbb{E}_{g\sim\pi}\, e^{-\lambda S_i(g)}}\right)\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{Z_1^{n+1}}\mathbb{E}_{\hat{g}_0^n}\log\left(\tfrac{\mathbb{E}_{g\sim\pi}\, e^{-\lambda S_{n+1}(g)}}{\mathbb{E}_{g\sim\pi}\, e^{-\lambda S_0(g)}}\right)\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{Z_1^{n+1}}\mathbb{E}_{\hat{g}_0^n}\log\mathbb{E}_{g\sim\pi}\, e^{-\lambda S_{n+1}(g)}
\end{aligned}$$

Now from the following lemma, we obtain

$$\begin{aligned}
\mathcal{E} &\le -\tfrac{1}{\lambda(n+1)}\log\mathbb{E}_{g\sim\pi}\, e^{-\lambda\mathbb{E}_{Z_1^{n+1}}\mathbb{E}_{\hat{g}_0^n}S_{n+1}(g)}\\
&= -\tfrac{1}{\lambda(n+1)}\log\mathbb{E}_{g\sim\pi}\, e^{-\lambda\left[(n+1)R(g)+\mathbb{E}_{Z_1^n}\mathbb{E}_{\hat{g}_0^n}\sum_{i=0}^{n}\Delta_\lambda(g,\hat{g}_i)\right]}\\
&= \min_{\rho\in\mathcal{M}}\left\{\mathbb{E}_{g\sim\rho}\, R(g)+\mathbb{E}_{g\sim\rho}\,\mathbb{E}_{Z_1^n}\mathbb{E}_{\hat{g}_0^n}\tfrac{\sum_{i=0}^{n}\Delta_\lambda(g,\hat{g}_i)}{n+1}+\tfrac{K(\rho,\pi)}{\lambda(n+1)}\right\}.
\end{aligned}$$

Lemma 3.2. *Let $\mathcal{W}$ be a real-valued measurable function defined on a product space $\mathcal{A}_1\times\mathcal{A}_2$ and let μ_1 and μ_2 be probability distributions on respectively $\mathcal{A}_1$ and $\mathcal{A}_2$ such that $\mathbb{E}_{a_1\sim\mu_1}\ \log\mathbb{E}_{a_2\sim\mu_2}\, e^{-\mathcal{W}(a_1,a_2)}<+\infty$. We have*

$$-\mathbb{E}_{a_1\sim\mu_1}\ \log\mathbb{E}_{a_2\sim\mu_2}\, e^{-\mathcal{W}(a_1,a_2)}\le-\log\mathbb{E}_{a_2\sim\mu_2}\, e^{-\mathbb{E}_{a_1\sim\mu_1}\ \mathcal{W}(a_1,a_2)}.$$

Proof. By using twice (2.1) and Fubini's theorem, we have

$$\begin{aligned}
-\mathbb{E}_{a_1}\log\mathbb{E}_{a_2}e^{-\mathcal{W}(a_1,a_2)} &= \mathbb{E}_{a_1}\inf_\rho\left\{\mathbb{E}_{a_2}\mathcal{W}(a_1,a_2)+K(\rho,\mu_2)\right\}\\
&\le \inf_\rho\mathbb{E}_{a_1}\left\{\mathbb{E}_{a_2}\mathcal{W}(a_1,a_2)+K(\rho,\mu_2)\right\}\\
&= -\log\mathbb{E}_{a_2}e^{-\mathbb{E}_{a_1}\mathcal{W}(a_1,a_2)}.
\end{aligned}$$

□

Inequality (3.4) is a direct consequence of (3.3). □

Theorem 3.1 bounds the expected risk of a randomized procedure, where the expectation is taken w.r.t. both the training set distribution and the randomizing distribution. From the following lemma, for convex loss functions, (3.4) implies

$$\mathbb{E}_{Z_1^n}\ R(\mathbb{E}_{g\sim\hat{\mu}}\ g)\le\min_{\mathcal{G}}\ R+\tfrac{\log|\mathcal{G}|}{\lambda(n+1)},\tag{3.5}$$

where we recall that $\hat{\mu}$ is the randomizing distribution of our generic algorithm and λ is a parameter whose typical value is the largest $\lambda>0$ such that $\delta_\lambda\equiv 0$.

Lemma 3.3. *For convex loss functions, the doubly expected risk of a randomized algorithm is greater than the expected risk of the deterministic version of the randomized algorithm, i.e. if $\hat{\rho}$ denotes the randomizing distribution, we have*

$$\mathbb{E}_{Z_1^n} R(\mathbb{E}_{g\sim\hat{\rho}}\, g) \leq \mathbb{E}_{Z_1^n} \mathbb{E}_{g\sim\hat{\rho}}\, R(g).$$

Proof. The result is a direct consequence of Jensen's inequality. □

In [23], the authors rely on worst-case analysis to recover standard-style statistical results such as Vapnik's bounds [42]. Theorem 3.1 can be seen as a complement to this pioneering work. Inequality (3.5) is the model selection bound that is well-known for least square regression and entropy loss, and that has been recently proved for general losses in [32].

Let us discuss the generalized form of the result. The r.h.s. of (3.3) is a classical regularized risk, which appears naturally in the PAC-Bayesian approach (see e.g. [21, 20, 6, 48]). An advantage of stating the result this way is to be able to deal with uncountable infinite $\mathcal{G}$. Even when $\mathcal{G}$ is countable, this formulation has some benefit to the extent that for any measurable function $h : \mathcal{G} \to \mathbb{R}$, $\min_{\rho\in\mathcal{M}}\{\mathbb{E}_{g\sim\rho}\, h(g) + K(\rho,\pi)\} \leq \min_{g\in\mathcal{G}} \{h(g) + \log \pi^{-1}(g)\}$.

Our generalization error bounds depend on two quantities λ and π which are the parameters of our algorithm. Their choice depends on the precise setting. Nevertheless, when $\mathcal{G}$ is finite and with no special structure a priori, a natural choice for π is the uniform distribution on $\mathcal{G}$.

Once the distribution π is fixed, an appropriate choice for the parameter λ is the minimizer of the r.h.s. of (3.3). This minimizer is unknown by the statistician, and it is an open problem to adaptively choose λ close to it.

4 Link with sequential prediction

This section aims at illustrating condition (3.1) and at clearly stating results coming from the online learning community in our batch setting. In [43, 31, 44], the loss function is assumed to satisfy: there are positive numbers η and c such that

$$\begin{aligned} &\forall\, \rho \in \mathcal{M} \quad \exists\, g_\rho : \mathcal{X} \to \mathcal{Y} \quad \forall\, x \in \mathcal{X} \quad \forall\, y \in \mathcal{Y} \\ &\qquad\qquad L[(x,y), g_\rho] \leq -\tfrac{c}{\eta} \log \mathbb{E}_{g\sim\rho}\, e^{-\eta L[(x,y),g]} \end{aligned} \tag{4.1}$$

Remark 4.1. If $g \mapsto e^{-\eta L(z,g)}$ is concave, then (4.1) holds for $c = 1$ (and one may take $g_\rho = \mathbb{E}_{g\sim\rho}\, g$).

Assumption (4.1) implies that condition (3.1) is satisfied both for $\lambda = \eta$ and $\delta_\lambda(Z, g, g') = -(1 - 1/c)L(Z, g')$ and for $\lambda = \eta/c$ and $\delta_\lambda(Z, g, g') = (c - 1)L(Z, g)$, and we may take in both cases $\hat{\pi}(\rho)$ as the Dirac distribution at g_ρ.

This leads to the *same* procedure that is described in the following straightforward corollary of Theorem 3.1.

Corollary 4.1. *Let $g_{\pi_{-\eta\Sigma_i}}$ be defined in the sense of (4.1) (for $\rho = \pi_{-\eta\Sigma_i}$). Consider the algorithm which predicts by drawing a function in $\{g_{\pi_{-\eta\Sigma_0}}, \dots, g_{\pi_{-\eta\Sigma_n}}\}$ according to the uniform distribution. Under Assumption (4.1), its expected risk $\mathbb{E}_{Z_1^n} \frac{1}{n+1} \sum_{i=0}^n R(g_{\pi_{-\eta\Sigma_i}})$ is upper bounded with*

$$c \min_{\rho \in \mathcal{M}} \left\{ \mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\eta(n+1)} \right\}. \tag{4.2}$$

This result is not surprising in view of the following two results. The first one comes from worst-case analysis in sequential prediction.

Theorem 4.2 (Haussler et al. [31], Theorem 3.8)**.** *Let $\mathcal{G}$ be countable. For any $g \in \mathcal{G}$, let $\Sigma_i(g) = \sum_{j=1}^i L(Z_j, g)$ (still) denote the cumulative loss up to time i of the expert which always predict according to function g. Under Assumption (4.1), the cumulative loss on Z_1^n of the strategy in which the prediction at time i is done according to $g_{\pi_{-\eta\Sigma_{i-1}}}$ in the sense of (4.1) (for $\rho = \pi_{-\eta\Sigma_{i-1}}$) is bounded by*

$$\inf_{g\in\mathcal{G}} \{c\Sigma_n(g) + \tfrac{c}{\eta} \log \pi^{-1}(g)\}. \tag{4.3}$$

The second result shows how the previous bound can be transposed into our model selection context by the following lemma.

Lemma 4.3. *Let $\mathcal{A}$ be a learning algorithm which produces the prediction function $\mathcal{A}(Z_1^i)$ at time $i+1$, i.e. from the data $Z_1^i = (Z_1, \dots, Z_i)$. Let $\mathcal{L}$ be the randomized algorithm which produces a prediction function $\mathcal{L}(Z_1^n)$ drawn according to the uniform distribution on $\{\mathcal{A}(\emptyset), \mathcal{A}(Z_1), \dots, \mathcal{A}(Z_1^n)\}$. The (doubly) expected risk of $\mathcal{L}$ is equal to $\frac{1}{n+1}$ times the expectation of the cumulative loss of $\mathcal{A}$ on the sequence $Z_1, \dots, Z_{n+1}$.*

Proof. By Fubini's theorem, we have

$$\begin{aligned} \mathbb{E}R[\mathcal{L}(Z_1^n)] &= \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \mathbb{E}_{Z_1^n} R[\mathcal{A}(Z_1^i)] \\ &= \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \mathbb{E}_{Z_1^{i+1}} L[Z_{i+1}, \mathcal{A}(Z_1^i)] \\ &= \tfrac{1}{n+1} \mathbb{E}_{Z_1^{n+1}} \textstyle\sum_{i=0}^n L[Z_{i+1}, \mathcal{A}(Z_1^i)]. \end{aligned}$$

□

For any $\eta > 0$, let $c(\eta)$ denote the infimum of the c for which (4.1) holds. Under weak assumptions, Vovk ([44]) proved that the infimum exists and studied the behavior of $c(\eta)$ and $a(\eta) = c(\eta)/\eta$, which are key quantities of (4.2) and (4.3). Under weak assumptions, and in particular in the examples given in Table

	Output space	Loss L(Z,g)	$c(\eta)$
Entropy loss [31, Example 4.3]	$\mathcal{Y}=[0;1]$	$Y\log\left(\frac{Y}{g(X)}\right)$ $+(1-Y)\log\left(\frac{1-Y}{1-g(X)}\right)$	$c(\eta)=1$ if $\eta\leq 1$ $c(\eta)=\infty$ if $\eta>1$
Absolute loss game [31, Section 4.2]	$\mathcal{Y}=[0;1]$	$\lvert Y-g(X)\rvert$	$\frac{\eta}{2\log[2/(1+e^{-\eta})]}$ $=1+\eta/4+o(\eta)$
Square loss [31, Example 4.4]	$\mathcal{Y}=[-B,B]$	$[Y-g(X)]^2$	$c(\eta)=1$ if $\eta\leq 1/(2B^2)$ $c(\eta)=+\infty$ if $\eta>1/(2B^2)$
L_q-loss (see p. 11)	$\mathcal{Y}=[-B,B]$	$\lvert Y-g(X)\rvert^q$ $q>1$	$c(\eta)=1$ if $\eta\leq\frac{q-1}{qB^q}(1\wedge 2^{2-q})$

Table 1: Value of $c(\eta)$ for different loss functions. Here B denotes a positive real.

1, the optimal constants in (4.3) are $c(\eta)$ and $a(\eta)$ ([44, Theorem 1]) and we have $c(\eta)\geq 1$, $\eta\mapsto c(\eta)$ nondecreasing and $\eta\mapsto a(\eta)$ nonincreasing. From these last properties, we understand the trade-off which occurs to choose the optimal η.

Table 1 specifies (4.2) in different well-known learning tasks. For instance, for bounded least square regression (i.e. when $|Y|\leq B$ for some $B>0$), the generalization error of the algorithm described in Corollary 4.1 when $\eta=1/(2B^2)$ is upper bounded with

$$\min_{\rho\in\mathcal{M}}\left\{\mathbb{E}_{g\sim\rho}\, R(g)+2B^2\frac{K(\rho,\pi)}{n+1}\right\}. \tag{4.4}$$

The constant appearing in front of the Kullback-Leibler divergence is much smaller than the ones obtained in unbounded regression setting even with gaussian noise and bounded regression function (see [17, 32] and [21, p.87]). The differences between these results partly comes from the absence of boundedness assumptions on the output and from the weighted average used in the aforementioned works. Indeed the weighted average prediction function, i.e. $\mathbb{E}_{g\sim\rho}\, g$, does not satisfy (4.1) for $c=1$ and $\eta=1/(2B^2)$ as was pointed out in [31, Example 3.13]. Nevertheless, it satisfies (4.1) for $c=1$ and $\eta\leq 1/(8B^2)$ (see Remark 4.1), which leads to similar but weaker bound (see (4.2)).

Case of the L_q-losses. To deal with these losses, we need the following slight generalization of the result given in Appendix A of [33].

Theorem 4.4. *Let $\mathcal{Y}=[a;b]$. We consider a non-regularized loss function, i.e. a loss function such that $L(Z,g)=\ell[Y,g(X)]$ for any $Z=(X,Y)\in\mathcal{Z}$ and some function $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. For any $y\in\mathcal{Y}$, let ℓ_y be the function $\big[y'\mapsto\ell(y,y')\big]$. If for any $y\in\mathcal{Y}$*

- *ℓ_y is continuous on $\mathcal{Y}$*
- *ℓ_y decreases on $[a;y]$, increases on $[y;b]$ and $\ell_y(y)=0$*

- *ℓ_y is twice differentiable on $]a;y[\cup]y;b[$,*

then (4.1) *is satisfied for $c = 1$ and*

$$\eta \leq \inf_{a \leq y_1 < y < y_2 \leq b} \frac{\ell'_{y_1}(y)\ell''_{y_2}(y) - \ell''_{y_1}(y)\ell'_{y_2}(y)}{\ell'_{y_1}(y)[\ell'_{y_2}(y)]^2 - [\ell'_{y_1}(y)]^2 \ell'_{y_2}(y)}, \tag{4.5}$$

where the infimum is taken w.r.t. y_1, y and y_2.

Proof. See Section 10.1. □

Remark 4.2. This result simplifies the original one to the extent that ℓ_y does not need to be twice differentiable at point y and the range of values for y in the infimum is $]y_1; y_2[$ instead of $]a; b[$.

Corollary 4.5. *For the L_q-loss, when $\mathcal{Y} = [-B; B]$ for some $B > 0$, condition* (4.1) *is satisfied for $c = 1$ and*

$$\eta \leq \frac{q-1}{qB^q}(1 \wedge 2^{2-q})$$

Proof. We apply Theorem 4.4. By simple computations, the r.h.s. of (4.5) is

$$\begin{aligned} & \inf_{-B \leq y_1 < y < y_2 \leq B} \frac{(q-1)(y_2-y_1)}{q(y-y_1)(y_2-y)[(y-y_1)^{q-1} + (y_2-y)^{q-1}]} \\ = \; & \frac{q-1}{q(2B)^q} \inf_{0<t<1} \frac{1}{t(1-t)[t^{q-1} + (1-t)^{q-1}]} \end{aligned}$$

For $1 < q \leq 2$, the infimum is reached for $t = 1/2$ and (4.5) can be written as $\eta \leq \frac{q-1}{qB^q}$. For $q \geq 2$, (4.5) is satisfied at least when $\eta \leq \frac{4(q-1)}{q\,(2B)^q}$. □

5 Model selection aggregation under Juditsky, Rigollet and Tsybakov assumptions ([32])

The main result of [32] relies on the following assumption on the loss function L and the set $\mathcal{P}$ of probability distributions on $\mathcal{Z}$ in which we assume that the true distribution is. There exist $\lambda > 0$ and a real-valued function ψ defined on $\mathcal{G} \times \mathcal{G}$ such that for any $P \in \mathcal{P}$

$$\begin{cases} \mathbb{E}_{Z\sim P}\, e^{\lambda[L(Z,g')-L(Z,g)]} \leq \psi(g',g) & \text{for any } g, g' \in \mathcal{G} \\ \psi(g,g) = 1 & \text{for any } g \in \mathcal{G} \\ \text{the function } \big[g \mapsto \psi(g',g)\big] \text{ is concave for any } g' \in \mathcal{G} \end{cases} \tag{5.1}$$

Theorem 3.1 gives the following result.

Corollary 5.1. *Consider the algorithm which draws uniformly its prediction function in the set* $\{\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_0}} g, \dots, \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_n}} g\}$*. Under Assumption (5.1), its expected risk* $\mathbb{E}_{Z_1^n} \frac{1}{n+1}\sum_{i=0}^n R(\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}} g)$ *is upper bounded with*

$$\min_{\rho\in\mathcal{M}} \left\{\mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\right\}. \tag{5.2}$$

Proof. See Section 10.2. □

In this context, our generic algorithm reduces to the randomized version of Algorithm A. From Lemma 3.3, for convex loss functions, (5.2) also holds for the risk of Algorithm A. Corollary 5.1 also shows that the risk bounds of [32, Theorem 3.2 and the examples of Section 4.2] hold with the same constants for our randomized algorithm (provided that the expected risk w.r.t. the training set distribution is replaced by the expected risk w.r.t. both training set and randomizing distributions).

On Assumption (5.1) we should say that it does not a priori require the function L to be convex. Nevertheless, any known relevant examples deal with "strongly" convex loss functions and we know that in general the assumption will not hold for the SVM (or hinge) loss function and for the absolute loss function. Indeed, without further assumption, one cannot expect rates better than $1/\sqrt{n}$ for these loss functions (see Section 8.4.2).

One can also recover the results in [32, Theorem 3.1 and Section 4.1] by taking the appropriate variance function $\delta_\lambda(Z,g,g')$. Once more the aggregation procedure is different because of the randomization step but the generalization error bounds are identical.

6 Standard-style statistical bounds

This section proposes new results of a different kind. In the previous sections, under convexity assumptions, we were able to achieve fast rates. Here we have assumption neither on the loss function nor on the probability generating the data. Nevertheless we show that our generic algorithm applied for $\delta_\lambda(Z,g,g') = \lambda[L(Z,g)-L(Z,g')]^2/2$ satisfies a sharp standard-style statistical bound.

This section contains two parts: the first one provides results in expectation (as in the preceding sections) whereas the second part provides deviation inequalities on the risk that requires advances on the sequential prediction analysis.

6.1 Bounds on the expected risk

6.1.1 Bernstein's type bound

Theorem 6.1. *Let* $V(g,g') = \mathbb{E}_Z\big\{[L(Z,g) - L(Z,g')]^2\big\}$. *Consider our generic algorithm (see Section 3) applied with* $\delta_\lambda(Z,g,g') = \lambda[L(Z,g) - L(Z,g')]^2/2$ *and* $\hat{\pi}(\rho) = \rho$. *Its expected risk* $\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\, R(g)$, *where we recall that* $\hat{\mu}$ *denotes the randomizing distribution, satisfies*

$$\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\hat{\mu}}\, R(g') \leq \min_{\rho\in\mathcal{M}} \Big\{\mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{\lambda}{2}\mathbb{E}_{g\sim\rho}\, \mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\hat{\mu}}\, V(g,g') + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big\} \tag{6.1}$$

Proof. See Section 10.3. □

To make (6.1) more explicit and to obtain a generalization error bound in which the randomizing distribution does not appear in the r.h.s. of the bound, the following corollary considers a widely used assumption that relates the variance term to the excess risk. Precisely, from Theorem 6.1, we obtain

Corollary 6.2. *Under the generalized Mammen and Tsybakov's assumption which states that there exist* $0 \leq \gamma \leq 1$ *and a prediction function* $\tilde{g}$ *(not necessarily in* $\mathcal{G}$*) such that* $V(g,\tilde{g}) \leq c[R(g) - R(\tilde{g})]^\gamma$ *for any* $g \in \mathcal{G}$, *the expected risk* $\mathcal{E} = \mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\, R(g)$ *of the generic algorithm used in Theorem 6.1 satisfies*

- *When* $\gamma = 1$,

  $$\mathcal{E} - R(\tilde{g}) \leq \min_{\rho\in\mathcal{M}} \Big\{\tfrac{1+c\lambda}{1-c\lambda}\big[\mathbb{E}_{g\sim\rho}\, R(g) - R(\tilde{g})\big] + \tfrac{K(\rho,\pi)}{(1-c\lambda)\lambda(n+1)}\Big\}$$

  In particular, for $\mathcal{G}$ *finite,* π *the uniform distribution,* $\lambda = 1/(2c)$, *when* $\tilde{g}$ *belongs to* $\mathcal{G}$, *we get* $\mathcal{E} \leq \min_{g\in\mathcal{G}} R(g) + \frac{4c\log|\mathcal{G}|}{n+1}$.

- *When* $\gamma < 1$, *for any* $0 < \beta < 1$ *and for* $\tilde{R}(g) \triangleq R(g) - R(\tilde{g})$,

  $$\mathcal{E} - R(\tilde{g}) \leq \Big\{\tfrac{1}{\beta}\Big(\mathbb{E}_{g\sim\rho}\,[\tilde{R}(g) + c\lambda\tilde{R}^\gamma(g)] + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big)\Big\} \vee \big(\tfrac{c\lambda}{1-\beta}\big)^{\frac{1}{1-\gamma}}.$$

Proof. See Section 10.4. □

To understand the sharpness of Theorem 6.1, we have to compare this result with the following one that comes from the traditional (PAC-Bayesian) statistical learning approach which relies on supremum of empirical processes. In the following theorem, we consider the estimator minimizing the uniform bound, i.e. the estimator for which we have the smallest upper bound on its generalization error.

Theorem 6.3. *We still use $V(g,g') = \mathbb{E}_Z\big\{[L(Z,g) - L(Z,g')]^2\big\}$. The generalization error of the algorithm which draws its prediction function according to the Gibbs distribution $\pi_{-\lambda\Sigma_n}$ satisfies*

$$\begin{aligned}
&\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, R(g') \\
&\quad\le \min_{\rho\in\mathcal{M}} \Big\{\mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)+1}{\lambda n} + \lambda\mathbb{E}_{g\sim\rho}\,\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, V(g,g') \\
&\quad + \lambda\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{g\sim\rho}\,\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, [L(Z_i,g) - L(Z_i,g')]^2\Big\}.
\end{aligned} \tag{6.2}$$

Let φ be the positive convex increasing function defined as $\varphi(t) \triangleq \frac{e^t-1-t}{t^2}$ and $\varphi(0) = \frac{1}{2}$ by continuity. When $\sup_{z\in\mathcal{Z},g\in\mathcal{G},g'\in\mathcal{G}} |L(z,g') - L(z,g)| \le B$, we also have

$$\begin{aligned}
\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, R(g') \le \min_{\rho\in\mathcal{M}} \Big\{&\mathbb{E}_{g\sim\rho}\, R(g) \\
&+\lambda\varphi(\lambda B)\mathbb{E}_{g\sim\rho}\,\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, V(g,g') + \tfrac{K(\rho,\pi)+1}{\lambda n}\Big\}.
\end{aligned} \tag{6.3}$$

Proof. See Section 10.5. □

As in Theorem 6.1, there is a variance term in which the randomizing distribution is involved. As in Corollary 6.2, one can convert (6.3) into a proper generalization error bound, that is a non trivial bound $\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_n}} R(g) \le \mathcal{B}(n,\pi,\lambda)$ where the training data do not appear in $\mathcal{B}(n,\pi,\lambda)$.

By comparing (6.3) and (6.1), we see that the classical approach requires the quantity $\sup_{g\in\mathcal{G},g'\in\mathcal{G}}|L(Z,g') - L(Z,g)|$ to be uniformly bounded and the unpleasing function φ appears. In fact, using technical small expectations theorems (see e.g. [3, Lemma 7.1]), exponential moments conditions on the above quantity would be sufficient.

The symmetrization trick used to prove Theorem 6.1 is performed in the prediction functions space. We do not call on the second virtual training set currently used in statistical learning theory (see [42]). Nevertheless both symmetrization tricks end up to the same nice property: we need no boundedness assumption on the loss functions. In our setting, symmetrization on training data leads to an unwanted expectation and to a constant four times larger (see the two variance terms of (6.2) and the discussion in [4, Section 8.3.3]).

In particular, deducing from Theorem 6.3 a corollary similar to Corollary 6.2 is only possible through (6.3) and provided that we have a boundedness assumption on $\sup_{z\in\mathcal{Z},g\in\mathcal{G},g'\in\mathcal{G}} |L(z,g') - L(z,g)|$. Indeed one cannot use (6.2) because of the last variance term in (6.2) (since Σ_n depends on Z_i).

6.1.2 Hoeffding's type bound

Contrary to generalization error bounds coming from Bernstein's inequality, (6.1) does not require any boundedness assumption. For bounded losses, without any variance assumption (i.e. roughly when the assumption used in Corollary (6.2) does not hold for $\gamma > 0$), tighter results are obtained by using Hoeffding's inequality, that is: for any random variable W satisfying $a \le W \le b$, then for any $\lambda > 0$

$$\mathbb{E}e^{\lambda(W-\mathbb{E}W)} \le e^{\lambda^2(b-a)^2/8}.$$

Theorem 6.4. *Assume that for any $z \in \mathcal{Z}$ and $g \in \mathcal{G}$, we have $a \le L(z,g) \le b$ for some reals a, b. Consider our generic algorithm (see Section 3) applied with $\delta_\lambda(Z,g,g') = \lambda(b-a)^2/8$ and $\hat{\pi}(\rho) = \rho$. Its expected risk $\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\, R(g)$, where we recall that $\hat{\mu}$ denotes the randomizing distribution, satisfies*

$$\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\, R(g) \le \min_{\rho\in\mathcal{M}} \left\{ \mathbb{E}_{g\sim\rho}\, R(g) + \frac{\lambda(b-a)^2}{8} + \frac{K(\rho,\pi)}{\lambda(n+1)} \right\} \tag{6.4}$$

In particular, when $\mathcal{G}$ is finite, by taking π uniform on $\mathcal{G}$ and $\lambda = \sqrt{\frac{8\log|\mathcal{G}|}{(b-a)^2(n+1)}}$, we get

$$\mathbb{E}_{Z_1^n}\mathbb{E}_{g\sim\hat{\mu}}\, R(g) - \min_{g\in\mathcal{G}} R(g) \le (b-a)\sqrt{\frac{\log|\mathcal{G}|}{2(n+1)}} \tag{6.5}$$

Proof. From Hoeffding's inequality, we have

$$\begin{aligned} \mathbb{E}_{g'\sim\hat{\pi}(\rho)}\, \log \mathbb{E}_{g\sim\rho}\, e^{\lambda[L(Z,g')-L(Z,g)]} &= \log \mathbb{E}_{g\sim\rho}\, e^{\lambda[\mathbb{E}_{g'\sim\hat{\pi}(\rho)}\, L(Z,g')-L(Z,g)]} \\ &\le \frac{\lambda^2(b-a)^2}{8}, \end{aligned}$$

hence (3.1) holds for $\delta_\lambda \equiv \lambda(b-a)^2/8$ and $\hat{\pi}(\rho) = \rho$. The result directly follows from Theorem 3.1. □

The standard point of view (see Appendix B) applies Hoeffding's inequality to the random variable $W = L(Z,g') - L(Z,g)$ for g and g' fixed and Z drawn according to the probability generating the data. The previous theorem uses it on the random variable $W = L(Z,g') - \mathbb{E}_{g\sim\rho}\, L(Z,g)$ for fixed Z and fixed probability distribution ρ but for g' drawn according to ρ. Here the gain is a multiplicative factor equal to 2 (see Appendix B).

6.2 Deviation inequalities

For the comparison between Theorem 6.1 and Theorem 6.3 to be fair, one should add that (6.3) and (6.2) comes from deviation inequalities that are not exactly obtainable to the author's knowledge with the arguments developed here. Precisely, consider the following adaptation of Lemma 5 of [47].

Lemma 6.5. *Let $\mathcal{A}$ be a learning algorithm which produces the prediction function $\mathcal{A}(Z_1^i)$ at time $i+1$, i.e. from the data $Z_1^i = (Z_1, \dots, Z_i)$. Let $\mathcal{L}$ be the* *randomized algorithm which produces a prediction function $\mathcal{L}(Z_1^n)$ drawn according to the uniform distribution on $\{\mathcal{A}(\emptyset), \mathcal{A}(Z_1), \dots, \mathcal{A}(Z_1^n)\}$. Assume that* $\sup_{z,g,g'} |L(z,g) - L(z,g')| \le B$ *for some $B > 0$. The expectation of the risk of $\mathcal{L}$ w.r.t. to the uniform draw is $\frac{1}{n+1}\sum_{i=0}^n R[\mathcal{A}(Z_1^i)]$ and satisfies: for any $\eta > 0$ and $\epsilon > 0$, for any reference prediction function $\tilde{g}$, with probability at least $1-\epsilon$ w.r.t. the distribution of $Z_1, \dots, Z_{n+1}$,*

$$\begin{aligned} \tfrac{1}{n+1}\textstyle\sum_{i=0}^n R[\mathcal{A}(Z_1^i)] - R(\tilde{g}) \le \tfrac{1}{n+1}\textstyle\sum_{i=0}^n \left\{ L[Z_{i+1}, \mathcal{A}(Z_1^i)] - L(Z_{i+1}, \tilde{g}) \right\} \\ + \eta\varphi(\eta B)\tfrac{1}{n+1}\textstyle\sum_{i=0}^n V[\mathcal{A}(Z_1^i), \tilde{g}] + \tfrac{\log(\epsilon^{-1})}{\eta(n+1)} \end{aligned} \tag{6.6}$$

where we still use $V(g,g') = \mathbb{E}_Z\left\{[L(Z,g) - L(Z,g')]^2\right\}$ for any prediction functions g and g' and $\varphi(t) \triangleq \frac{e^t - 1 - t}{t^2}$ for any $t > 0$.

Proof. See Section 10.6. □

We see that two variance terms appear. The first one comes from the worst-case analysis and is hidden in $\sum_{i=0}^n \left\{ L[Z_{i+1}, \mathcal{A}(Z_1^i)] - L(Z_{i+1}, \tilde{g}) \right\}$ and the second one comes from the concentration result (Lemma 10.1). The presence of this last variance term anneals the benefits of our approach in which we were manipulating variance terms much smaller than the traditional Bernstein's variance term.

To illustrate this point, consider for instance least square regression with bounded outputs: from Theorem 4.2 and Table 1, the hidden variance term is null while the second variance term prevents us to obtain deviation inequalities of order n^{-1} even though from (4.4) our procedure has n^{-1}-convergence rate in expectation.

To conclude, for deviation inequalities, we cannot expect to do better than the standard-style approach since at some point we use a Bernstein's type bound w.r.t. the distribution generating the data.

Remark 6.1. Lemma 6.5 should be compared with Lemma 4.3. The latter deals with results in expectation while the former concerns deviation inequalities. Note that Lemma 6.5 requires the loss function to be bounded and makes a variance term appear.

7 Application to L_q-regression

This section shows that Theorem 3.1 used jointly with the symmetrization idea developed in the previous section allows to obtain new convergence rates in heavy noise situation, i.e. when the output is not constrained to have a bounded exponential moment. As a warm-up exercise, we consider the absolute loss setting (Section 7.1). Section 7.2 then deals with the strongly convex loss functions (i.e. $q > 1$).

7.1 Case $q = 1$:

Here we just assume that the functions in the model $\mathcal{G}$ are uniformly bounded, i.e. $\sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \le b$ for some $b > 0$. From (6.1), using that $V(g,g') = \mathbb{E}_Z\big\{[|Y-g(X)|-|Y-g'(X)|]^2\big\} \le 4b^2$, for any $\lambda > 0$, there exists an estimator $\hat{g}$ such that $\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \le 2\lambda b^2 + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$. Choosing $\lambda = \sqrt{\frac{\log|\mathcal{G}|}{2b^2(n+1)}}$, i.e. the minimizer of the previous r.h.s., we obtain

$$\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \le 2b\sqrt{\tfrac{2\log|\mathcal{G}|}{n+1}}. \tag{7.1}$$

7.2 Case $q > 1$:

We start with the following theorem concerning general loss functions.

Theorem 7.1. *Let $B \ge b > 0$ and $\mathcal{Y} = \mathbb{R}$. Consider a loss function L which can be written as $L[(x,y),g] = \ell[y,g(x)]$, where the function $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ satisfies: there exists $\lambda_0 > 0$ such that for any $y \in [-B;B]$, the function $y' \mapsto e^{-\lambda_0 \ell(y,y')}$ is concave on $[-b;b]$. Let*

$$\Delta(y) \;=\; \sup_{|\alpha|\le b,|\beta|\le b} \big[\ell(y,\alpha) - \ell(y,\beta)\big].$$

For $\lambda \in]0;\lambda_0]$, consider the algorithm that draws uniformly its prediction function in the set $\{\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_0}}\, g, \dots, \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_n}}\, g\}$, and consider the deterministic version of this randomized algorithm. The expected risk of these algorithms satisfy

$$\begin{aligned}
&\mathbb{E}_{Z_1^n} R\big(\tfrac{1}{n+1}\textstyle\sum_{i=0}^n \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}}\, g\big)\\
&\qquad\le \mathbb{E}_{Z_1^n} \tfrac{1}{n+1}\textstyle\sum_{i=0}^n R\big(\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}}\, g\big)\\
&\qquad\le \min_{\rho\in\mathcal{M}} \Big\{\mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big\}\\
&\qquad\quad + \mathbb{E}\Big\{\tfrac{\lambda\Delta^2(Y)}{2}\mathbf{1}_{\lambda\Delta(Y)<1;|Y|>B} + \big[\Delta(Y) - \tfrac{1}{2\lambda}\big]\mathbf{1}_{\lambda\Delta(Y)\ge 1;|Y|>B}\Big\}.
\end{aligned} \tag{7.2}$$

Proof. See Section 10.7. □

Remark 7.1. For $y \in [-B;B]$, concavity of $y' \mapsto e^{-\lambda_0\ell(y,y')}$ on $[-b;b]$ for $\lambda_0 > 0$ implies convexity of $y' \mapsto \ell(y,y')$ on $[-b;b]$.

In particular, for least square regression, Theorem 7.1 can be simplified into:

Theorem 7.2. *Assume that $\sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \le b$ for some $b > 0$. For any $0 < \lambda \le 1/(8b^2)$, the expected risk of the algorithm that draws uniformly its prediction function among $\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_0}}\, g, \dots, \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_n}}\, g$ is upper bounded with*

$$\min_{\rho\in\mathcal{M}} \Big\{\mathbb{E}_{g\sim\rho}\, R(g) + 8\lambda b^2\mathbb{E}\big(Y^2\mathbf{1}_{|Y|>(2\lambda)^{-1/2}}\big) + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big\}. \tag{7.3}$$

Proof. For any $B \geq b$, for any $y \in [-B; B]$, straightforward computations show that $y' \mapsto e^{-\lambda_0 (y-y')^2}$ is concave on $[-b; b]$ for $\lambda_0 = \frac{1}{2(B+b)^2}$, so that we can apply Theorem 7.1. We have $\Delta(y) = 4b|y|$ so that by optimizing the parameter B, we obtain that the expected risk of the algorithm is upper bounded with

$$\begin{aligned} &\min_{\rho \in \mathcal{M}} \Big\{ \mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \Big\} + \mathbb{E}\big\{ \big(4b|Y| - \tfrac{1}{2\lambda}\big) \mathbf{1}_{|Y| \geq (4b\lambda)^{-1}} \big\} \\ &\qquad\qquad + \mathbb{E}\big\{ 8\lambda b^2 Y^2 \mathbf{1}_{(2\lambda)^{-1/2} - b < |Y| < (4b\lambda)^{-1}} \big\} \\ \leq\ &\min_{\rho \in \mathcal{M}} \Big\{ \mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \Big\} + \mathbb{E}\big\{ 8\lambda b^2 Y^2 \mathbf{1}_{|Y| \geq (4b\lambda)^{-1}} \big\} \\ &\qquad\qquad + \mathbb{E}\big\{ 8\lambda b^2 Y^2 \mathbf{1}_{(2\lambda)^{-1/2} - b < |Y| < (4b\lambda)^{-1}} \big\}, \end{aligned}$$

which gives the desired result. □

Theorem 7.2 improves [17, Theorem 1].

Corollary 7.3. *Under the assumptions*

$$\begin{cases} \sup_{g\in\mathcal{G}, x\in\mathcal{X}} |g(x)| \leq b & \text{for some } b > 0 \\ \mathbb{E}|Y|^s \leq A & \text{for some } s \geq 2 \text{ and } A > 0 \\ \mathcal{G} \text{ finite} & \end{cases}$$

for $\lambda = C_1 \big(\frac{\log|\mathcal{G}|}{n}\big)^{2/(s+2)}$ *where* $C_1 > 0$ *and* π *the uniform distribution on* $\mathcal{G}$*, the expected risk of the algorithm that draws uniformly its prediction function among* $\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_0}}\, g, \dots, \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_n}}\, g$ *is upper bounded with*

$$\min_{g\in\mathcal{G}} R(g) + C \Big(\frac{\log|\mathcal{G}|}{n}\Big)^{s/(s+2)} \tag{7.4}$$

for a quantity C *which depends only on* C_1*,* b*,* A *and* s*.*

Juditsky, Rigollet and Tsybakov proved that Corollary 7.3 can also be obtained through a simple adaptation of their original analysis (see [32, Section 4.1]).

Proof. The moment assumption on Y implies

$$\alpha^{s-q} \mathbb{E}|Y|^q \mathbf{1}_{|Y|\geq\alpha} \leq A \text{ for any } 0 \leq q \leq s \text{ and } \alpha \geq 0. \tag{7.5}$$

As a consequence, the second term in (7.3) is bounded with $8\lambda b^2 A (2\lambda)^{(s-2)/2}$, so that (7.3) is upper bounded with $\min_{g\in\mathcal{G}} R(g) + A 2^{2+s/2} b^2 \lambda^{s/2} + \frac{\log|\mathcal{G}|}{\lambda n}$, which gives the desired result. □

In particular, with the minimal assumption $\mathbb{E}Y^2 \leq A$ (i.e. $s = 2$), the convergence rate is of order $n^{-1/2}$, and at the opposite, when s goes to infinity, we recover the n^{-1} rate we have under exponential moment condition on the output.

Using Theorem 7.1, we can generalize Corollary 7.3 to L_q-regression and obtain the following result.

Corollary 7.4. *Let $q > 1$. Assume that*

$$\begin{cases} \sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \le b & \textit{for some } b > 0 \\ \mathbb{E}|Y|^s \le A & \textit{for some } s \ge q \textit{ and } A > 0 \\ \mathcal{G} \textit{ finite} \end{cases}$$

Let π be the uniform distribution on $\mathcal{G}$, $C_1 > 0$ and

$$\lambda = \begin{cases} C_1\left(\frac{\log|\mathcal{G}|}{n}\right)^{(q-1)/s} & \textit{when } q \le s < 2q-2 \\ C_1\left(\frac{\log|\mathcal{G}|}{n}\right)^{q/(s+2)} & \textit{when } s \ge 2q-2 \end{cases}.$$

The expected risk of the algorithm which draws uniformly its prediction function among $\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_0}}\, g, \ldots, \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_n}}\, g$ is upper bounded with

$$\begin{cases} \min_{g\in\mathcal{G}} R(g) + C\left(\frac{\log|\mathcal{G}|}{n}\right)^{1-\frac{q-1}{s}} & \textit{when } q \le s \le 2q-2 \\ \min_{g\in\mathcal{G}} R(g) + C\left(\frac{\log|\mathcal{G}|}{n}\right)^{1-\frac{q}{s+2}} & \textit{when } s \ge 2q-2 \end{cases}.$$

for a quantity C which depends only on C_1, b, A, q and s.

Proof. See Section 10.8. □

Remark 7.2. For $q > 2$, low convergence rates (that is $n^{-\gamma}$ with $\gamma < 1/2$) appear when the moment assumption is weak: $\mathbb{E}|Y|^s \le A$ for some $A > 0$ and $q \le s < 2q-2$. Convergence rates faster that the standard non parametric rates $n^{-1/2}$ are achieved for $s > 2q-2$. Fast convergence rates systematically occurs when $1 < q < 2$ since for these values of q, we have $s \ge q > 2q-2$. Surprisingly, for $q = 1$, the picture is completely different (see Section 8.4.2 for an exhaustive discussion and for the minimax optimality of the results of this section).

8 Lower bounds

The simplest way to assess the quality of an algorithm and of its expected risk upper bound is to prove a risk lower bound saying that no algorithm has better convergence rate. This section provides this kind of assertions. The lower bounds developed here have the same spirit as the ones in [16, 2, 13], [28, Chap. 15] and [5, Section 5] to the extent that it relies on the following ideas:

- The supremum of a quantity $\mathcal{Q}(P)$ when the distribution P belongs to some set $\mathcal{P}$ is larger than the supremum over a well chosen subset of P, which is larger than the mean of $\mathcal{Q}(P)$ when the distribution P is drawn uniformly in this subset.

- When the chosen subset is an hypercube of 2^m distributions (see Section 8.1), the design of a lower bound over the 2^m distributions reduces to the design of a lower bound over two distributions.

- When a sequence Z_1^n has similar likelihoods according to two different probability distributions, then no estimator will be accurate for both distributions: for this sequence, the risk of any estimator will be all the larger as the Bayes-optimal prediction associated with the two distributions are 'far away'.

We refer the reader to [14] and [40, Chap. 2] for lower bounds which applies for general set of probability distributions, not necessarily containing hypercubes as here. Our analysis focuses on hypercubes since in several settings they afford to obtain lower bounds with both the right convergence rate and close to optimal constants. Our contribution in this section is

- to provide results for general non-regularized loss functions (we recall that non-regularized loss functions are loss functions which can be written as $L[(x, y), g] = \ell[y, g(x)]$ for some function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$),

- to improve the upper bound on the variational distance appearing in Assouad's argument,

- to generalize the argument to asymmetrical hypercubes which is necessary to find the lower bound matching the upper bound of Corollary 7.4 for $q \leq s \leq 2q - 2$,

- to express the lower bounds in terms of similarity measures between two distributions characterizing the hypercube.

- to obtain lower bounds matching the upper bounds obtained in the previous sections.

Remark 8.1. In [31], the optimality of the constant in front of the $(\log |\mathcal{G}|)/n$ has been proved by considering the situation when both $|\mathcal{G}|$ and n goes to infinity. Note that this worst-case analysis constant is not necessary the same as our batch setting constant. This section shows that the batch setting constant is not "far" from the worst-case analysis constant.

Besides Lemma 4.3, which can be used to convert any worst-case analysis upper bounds into a risk upper bound in our batch setting, also means that any lower bounds for our batch setting leads to a lower bound in the sequential prediction setting. Indeed the cumulative loss on the worst sequence of data is bigger than the average cumulative loss when the data are taken i.i.d. from some probability distribution. As a consequence, the bounds developed in this section partially solve the open problem introduced in [31, Section 3.4] consisting in developing tight non-asymptotical lower bounds. For least square loss and entropy loss, our bounds are off by a multiplicative factor smaller than 4 (see Remarks 8.6 [p.35] and 8.7 [p.37]).

This section is organized as follows. Section 8.1 defines the quantities that characterize hypercubes of probability distributions and details the links between them. Section 8.2 defines a similarity measure between probability distributions coming from f-divergences (see [26]) and gives their main properties. We give our main lower bounds in Section 8.3. These bounds are illustrated in Section 8.4.

8.1 Hypercube of probability distributions

Let $m \in \mathbb{N}^*$. Consider a family of 2^m probability distributions on $\mathcal{Z}$

$$\{P_{\bar{\sigma}} : \bar{\sigma} \triangleq (\sigma_1, \dots, \sigma_m) \in \{-;+\}^m\}$$

having the same first marginal, denoted μ:

$$P_{\bar{\sigma}}(dX) = P_{(+,\dots,+)}(dX) \triangleq \mu(dX) \text{ for any } \bar{\sigma} \in \{-;+\}^m,$$

and such that there exists

- a partition $\mathcal{X}_0, \dots, \mathcal{X}_m$ of $\mathcal{X}$,
- functions h_1 and h_2 defined on $\mathcal{X} - \mathcal{X}_0$ taking their values in $\mathcal{Y}$
- functions p_+ and p_- defined on $\mathcal{X} - \mathcal{X}_0$ taking their values in $[0;1]$

for which for any $j \in \{1, \dots, m\}$, for any $x \in \mathcal{X}_j$, we have

$$P_{\bar{\sigma}}\big(Y = h_1(x)\big|X = x\big) = p_{\sigma_j}(x) = 1 - P_{\bar{\sigma}}\big(Y = h_2(x)\big|X = x\big), \tag{8.1}$$

and for any $x \in \mathcal{X}_0$, the distribution of Y knowing $X = x$ is independent from $\bar{\sigma}$ (i.e. the 2^m conditional distributions are identical).

In particular, (8.1) means that for any $x \in \mathcal{X} - \mathcal{X}_0$, the conditional probability of the output knowing the input x is concentrated on two values and that, under the distribution $P_{\bar{\sigma}}$, the disproportion between the probabilities of these two values is all the larger as $p_{\sigma_j}(x)$ is far from $1/2$ for j the integer such that $x \in \mathcal{X}_j$.

Remark 8.2. Equality (8.1) indirectly implies that for any x, $h_1(x) \neq h_2(x)$. This is not at all restricting since points for which we would have liked $h_1(x) = h_2(x)$ can be put in the "garbage" set $\mathcal{X}_0$.

Definition 8.1. The family of 2^m probability distributions will be referred to as an *hypercube* of distributions if and only if for any $j \in \{1, \dots, m\}$,

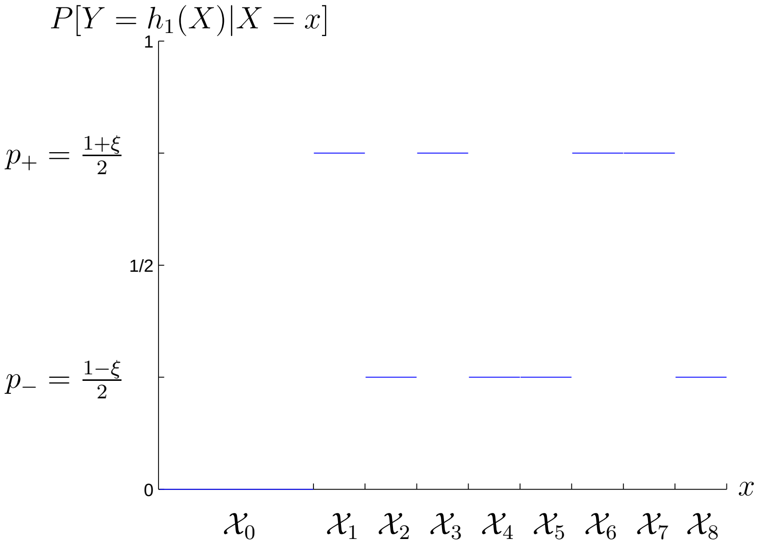

Figure 1: Representation of a probability distribution of the hypercube. Here the hypercube is a constant and symmetrical one (see Definition 8.2) with $m = 8$ and the probability distribution is characterized by $\bar{\sigma} = (+, -, +, -, -, +, +, -)$.

- the probability $\mu(\mathcal{X}_j) = \mu(X \in \mathcal{X}_j)$ is independent from j, i.e. $\mu(\mathcal{X}_1) = \cdots = \mu(\mathcal{X}_m)$,
- the law of $\big(p_+(X), p_-(X), h_1(X), h_2(X)\big)$ when X is drawn according to the conditional distribution $\mu(\cdot|X \in \mathcal{X}_j)$ is independent from j, i.e. the m conditional distributions are identical.

Remark 8.3. The typical situation in which we encounter hypercubes are when $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \geq 1$ and when we have translation invariance to the extent that there exist $t_2, \dots, t_d$ in $\mathbb{R}^d$ such that for any $j \in \{2, \dots, m\}$, $\mathcal{X}_j = \mathcal{X}_1 + t_j$ and for any $x \in \mathcal{X}_1$,

$$\big(p_+(x+t_j), p_-(x+t_j), h_1(x+t_j), h_2(x+t_j)\big) \\ = \big(p_+(x), p_-(x), h_1(x), h_2(x)\big).$$

A special hypercube is illustrated in Figure 1.

For any $p \in [0; 1]$, $y_1 \in \mathcal{Y}$, $y_2 \in \mathcal{Y}$ and $y \in \mathcal{Y}$, consider

$$\varphi_{p,y_1,y_2}(y) \triangleq p\ell(y_1, y) + (1-p)\ell(y_2, y) \tag{8.2}$$

When $y_1 \neq y_2$, this is the risk of the prediction function identically equal to y when the distribution generating the data satisfies $P[Y = y_1] = p = 1 - P[Y = y_2]$.

The case $y_1 = y_2$ corresponds to $P[Y = y_1 = y_2] = 1$and will not be of interest to us (since we will use this function for $y_1 = h_1(x) \neq h_2(x) = y_2$). .

Through this distribution, the quantity

$$\phi_{y_1,y_2}(p) \triangleq \inf_{y\in\mathcal{Y}} \varphi_{p,y_1,y_2}(y) \tag{8.3}$$

can be viewed as the risk of the best constant prediction function.

Remark 8.4. In the binary classification setting, i.e. when $|\mathcal{Y}| = 2$, from the Bayes rule, the function $\big[x \mapsto \text{a minimizer of } \varphi_{P(Y=1|X=x),-1,+1}\big]$ is the best prediction function to the extent that it minimizes the risk R (see typical examples of loss functions in Section 8.4.1 and an exhaustive study of their links in [11]).

For any p_+ and p_- in $[0;1]$, introduce

$$\begin{aligned}&\psi_{p_+,p_-,y_1,y_2}(\alpha)\\ &\quad\triangleq \phi_{y_1,y_2}[\alpha p_+ + (1-\alpha)p_-] - \alpha\phi_{y_1,y_2}(p_+) - (1-\alpha)\phi_{y_1,y_2}(p_-)\end{aligned} \tag{8.4}$$

Lemma 8.1. *1. For any $y_1 \in \mathcal{Y}$, $y_2 \in \mathcal{Y}$, $p_+ \in [0;1]$ and $p_- \in [0;1]$, the functions ϕ_{y_1,y_2} and ψ_{p_+,p_-,y_1,y_2} are concave, and consequently admit one-sided derivatives everywhere. The function ψ_{p_+,p_-,y_1,y_2} is non-negative.*

2. *Define the function K_α as $K_\alpha(t) = [(1-\alpha)t] \wedge [\alpha(1-t)]$. Let $q_- = p_- \wedge p_+$ and $q_+ = p_- \vee p_+$. Assume that the function ϕ_{y_1,y_2} is twice differentiable by parts on $[q_-;q_+]$ to the extent that there exist $q_- = \beta_0 < \beta_1 < \cdots < \beta_d < \beta_{d+1} = q_+$ such that for any $\ell \in \{0,\dots,d\}$, ϕ_{y_1,y_2} is twice differentiable on $]\beta_\ell;\beta_{\ell+1}[$. For any $\ell \in \{1,\dots,d\}$, let Δ_ℓ denote the difference between the right-sided and left-sided derivatives of ϕ_{y_1,y_2} at point β_ℓ. We have*

$$\begin{aligned}\psi_{p_+,p_-,y_1,y_2}(\alpha) = &-(p_+ - p_-)^2 \textstyle\int_0^1 K_\alpha(t)\phi''_{y_1,y_2}\big[tp_+ + (1-t)p_-\big]dt\\ &-|p_+ - p_-| \textstyle\sum_{\ell=1}^d K_\alpha\big(\frac{\beta_\ell - p_-}{p_+ - p_-}\big)\Delta_\ell\end{aligned}$$

In particular, we have

$$\begin{aligned}&\psi_{p_+,p_-,y_1,y_2}(1/2)\\ &\quad= \tfrac{(p_+-p_-)^2}{2} \textstyle\int_0^1 [t\wedge(1-t)]\big|\phi''_{y_1,y_2}\big[tp_+ + (1-t)p_-\big]\big|dt\\ &\quad+\tfrac{1}{2}\textstyle\sum_{\ell=1}^d \big(|p_+ - \beta_\ell| \wedge |\beta_\ell - p_-|\big)|\Delta_\ell|\end{aligned} \tag{8.5}$$

Proof. 1. The function ϕ_{y_1,y_2} is concave since it is the infimum of concave (affine) functions. As a direct consequence, the function ψ_{p_+,p_-,y_1,y_2} is concave and non-negative.

2. It suffices to apply the following lemma, that is proved in Section 10.9, to the function $f : t \mapsto \phi_{y_1,y_2}[tp_+ + (1-t)p_-]$, which has critical points on t_ℓ defined as $t_\ell p_+ + (1-t_\ell)p_- = \beta_\ell$.

Lemma 8.2. *Let $f : [0;1] \to \mathbb{R}$ be a function twice differentiable by parts to the extent that there exist $0 = t_0 < t_1 < \cdots < t_d < t_{d+1} = 1$ such that for any $\ell \in \{0, \dots, d\}$, f is twice differentiable on $]t_\ell; t_{\ell+1}[$. Assume that f is continuous and admits left-sided derivatives f'_l and right-sided derivatives $f'_r(t_\ell)$ at the critical points t_ℓ. For any $\alpha \in [0;1]$, we have*

$$f(\alpha) - \alpha f(1) - (1-\alpha) f(0) = - \int_0^1 K_\alpha(t) f''(t) dt \\ - \sum_{\ell=1}^d K_\alpha(t_\ell)[f'_r(t_\ell) - f'_l(t_\ell)]$$

□

Definition 8.2. Let $\{P_{\bar\sigma} : \bar\sigma \triangleq (\sigma_1, \dots, \sigma_m) \in \{-;+\}^m\}$ be an hypercube of distributions.

1. The positive integer m is called the *dimension* of the hypercube.
2. The probability $w \triangleq \mu(\mathcal{X}_1) = \cdots = \mu(\mathcal{X}_m)$ is called the *edge probability*.
3. The *characteristic function of the hypercube* is the function $\tilde\psi : \mathbb{R}_+ \to \mathbb{R}_+$ defined as for any $u \in \mathbb{R}_+$
$$\tilde\psi(u) = \tfrac{1}{2} m(u+1) \mathbb{E}_{X\sim\mu} \left\{ \mathbf{1}_{X\in\mathcal{X}_1} \psi_{p_+(X),p_-(X),h_1(X),h_2(X)}\left(\tfrac{u}{u+1}\right) \right\}.$$
4. The *edge discrepancies of type I* of the hypercube are
$$\begin{cases} d_{\mathrm{I}} \triangleq \frac{\tilde\psi(1)}{mw} = \mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)}\left[\psi_{p_+(X),p_-(X),h_1(X),h_2(X)}(1/2)\right] \\ d'_{\mathrm{I}} \triangleq \mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)}\{[p_+(X) - p_-(X)]^2\} \end{cases} \tag{8.6}$$
5. The *edge discrepancy of type II* of the hypercube is defined as
$$d_{\mathrm{II}} \triangleq \mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)}\left\{ \left(\sqrt{p_+(X)[1-p_-(X)]} - \sqrt{[1-p_+(X)]p_-(X)} \right)^2 \right\}. \tag{8.7}$$
6. A probability distribution P_0 on $\mathcal{Z}$ satisfying $P_0(dX) = \mu(dX)$ and for any $x \in \mathcal{X} - \mathcal{X}_0$, $P_0[Y = h_1(x)|X = x] = \frac{1}{2} = P_0[Y = h_2(x)|X = x]$ will be referred to as a *base of the hypercube*.
7. Let P_0 be a base of the hypercube. Consider distributions $P_{[\sigma]}, \sigma \in \{-,+\}$ admitting the following density w.r.t. P_0:
$$\frac{P_{[\sigma]}}{P_0}(x,y) = \begin{cases} 2p_\sigma(x) & \text{when } x \in \mathcal{X}_1 \text{ and } y = h_1(x) \\ 2[1-p_\sigma(x)] & \text{when } x \in \mathcal{X}_1 \text{ and } y = h_2(x) \\ 1 & \text{otherwise} \end{cases}$$
The distributions $P_{[-]}$ and $P_{[+]}$ will be referred to as the *representatives of the hypercube*.

8. When the functions p_+ and p_- are constant on $\mathcal{X}_1$, the hypercube will be said *constant*.

9. When the functions p_+ and p_- satisfies $p_+ = 1 - p_-$ on $\mathcal{X} - \mathcal{X}_0$, the hypercube will be said *symmetrical*. In this case, the function $2p_+ - 1$ will be denoted ξ so that
$$\begin{cases} p_+ &= \frac{1+\xi}{2} \\ p_- &= \frac{1-\xi}{2} \end{cases} \tag{8.8}$$
Otherwise it will be said *asymmetrical*.

The edge discrepancies are non-negative quantities that are all the smaller as p_- and p_+ become closer.

Let us introduce the following assumption.

Differentiability assumption. For any $x \in \mathcal{X}_1$, the function $\phi_{h_1(x),h_2(x)}$ is twice differentiable and satisfies for any $t \in [p_-(x) \wedge p_+(x); p_-(x) \vee p_+(x)]$,
$$\left|\phi''_{h_1(x),h_2(x)}(t)\right| \geq \zeta \tag{8.9}$$
for some $\zeta > 0$.

When $\mathcal{Y} \subseteq \mathbb{R}$, the differentiability assumption is typically fulfilled when for any $y_1 \neq y_2$, the functions $y \mapsto \ell(y_1, y)$ and $y \mapsto \ell(y_2, y)$ admit second derivatives lower bounded by a positive constant and when these functions are minimum for respectively $y = y_1$ and $y = y_2$. This is the case for least square loss and entropy loss, but it is not the case for hinge loss, absolute loss or classification loss. The following result gives the main properties of the characteristic function $\tilde{\psi}$ and useful lower bounds of it.

Lemma 8.3. *The characteristic function of the hypercube is a concave nondecreasing function and satisfies*

- $\tilde{\psi}(0) = 0$
- $\tilde{\psi}(u) \geq (u \wedge 1)\tilde{\psi}(1)$
- *Under the differentiability assumption (see* (8.9)*), we have*
$$\tilde{\psi}(u) \geq \frac{mw\zeta}{4} d'_{\mathrm{I}} \frac{u}{u+1} \geq \frac{mw\zeta}{8} d'_{\mathrm{I}}(u \wedge 1)$$
where we recall that $d'_{\mathrm{I}} \triangleq \mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)}\big\{[p_+(X) - p_-(X)]^2\big\}$*.In particular, we have*
$$d_{\mathrm{I}} \geq \frac{\zeta}{8} d'_{\mathrm{I}}. \tag{8.10}$$

Proof. From Lemma 8.1, the function ψ is non-negative and concave. Therefore the characteristic function is also concave and non-negative on $\mathbb{R}_+$. Consequently, it is nondecreasing. The remaining assertions of the lemma are then straightforward. □

To underline the link between the discrepancies of types I and II, one may consider (8.10) jointly with the following result

Lemma 8.4. *When an hypercube is constant and symmetrical, i.e. when on $\mathcal{X}_1$ $p_+ = \frac{1+\xi}{2}$ and $p_- = \frac{1-\xi}{2}$ for ξ constant, we have*

$$d'_{\mathrm{I}} = d_{\mathrm{II}} = \xi^2.$$

Finally, since the design of constant and symmetrical hypercubes is the key of numerous lower bounds, we use the following:

Definition 8.3. A $(\tilde{m}, \tilde{w}, \tilde{d}_{\mathrm{II}})$-hypercube is a constant and symmetrical $\tilde{m}$-dimensional hypercube with edge probability $\tilde{w}$ and edge discrepancy of type II equal to $\tilde{d}_{\mathrm{II}}$, and for which $p_+ > 1/2$ and h_1 and h_2 are constant functions.

For these hypercubes, we have $m = \tilde{m}$, $w = \tilde{w}$, $d_{\mathrm{II}} = \tilde{d}_{\mathrm{II}}$ and

$$\begin{cases} \xi & \equiv & \sqrt{d_{\mathrm{II}}} \\ p_- & \equiv & \frac{1-\sqrt{d_{\mathrm{II}}}}{2} \\ p_+ & \equiv & \frac{1+\sqrt{d_{\mathrm{II}}}}{2} \end{cases}$$

and from (8.5), when the function ϕ_{h_1,h_2} is twice differentiable on $]p_-;p_+[$,

$$d_{\mathrm{I}} = \frac{d_{\mathrm{II}}}{2} \int_0^1 [t \wedge (1-t)] \left| \phi''_{h_1,h_2}\left(\frac{1-\sqrt{d_{\mathrm{II}}}}{2} + \sqrt{d_{\mathrm{II}}}t \right) \right| dt. \tag{8.11}$$

8.2 f-similarity

Let us introduce a similarity measure between probability distributions. When a probability distribution $\mathbb{P}$ is absolutely continuous w.r.t. another probability distribution $\mathbb{Q}$, i.e. $\mathbb{P} \ll \mathbb{Q}$, $\frac{\mathbb{P}}{\mathbb{Q}}$ denotes the density of $\mathbb{P}$ w.r.t. $\mathbb{Q}$.

Definition 8.4. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function. The *f-similarity* between two probability distributions is defined as

$$\mathcal{S}_f(\mathbb{P}, \mathbb{Q}) = \begin{cases} \int f\left(\frac{\mathbb{P}}{\mathbb{Q}}\right) d\mathbb{Q} & \text{if } \mathbb{P} \ll \mathbb{Q} \\ f(0) & \text{otherwise} \end{cases} \tag{8.12}$$

Equivalently, if p and q denote the density of $\mathbb{P}$ and $\mathbb{Q}$ w.r.t. the probability distribution $(\mathbb{P}+\mathbb{Q})/2$, one can define the f-similarity as

$$\mathcal{S}_f(\mathbb{P},\mathbb{Q}) = \int_{q>0} f\left(\frac{p}{q}\right) d\mathbb{Q}$$

It is called f-similarity in reference to f-divergence (see [26]) to which it is closely related. Precisely, introduce the function $\tilde{f} = f(1) - f$. $\tilde{f}$ is convex and satisfies $\tilde{f}(1) = 0$. Thus it is associated with an $\tilde{f}$-divergence, which is defined as

$$D_{\tilde{f}}(\mathbb{P},\mathbb{Q}) = \begin{cases} \int \tilde{f}\left(\frac{\mathbb{P}}{\mathbb{Q}}\right) d\mathbb{Q} & \text{if } \mathbb{P} \ll \mathbb{Q} \\ \tilde{f}(0) & \text{otherwise} \end{cases} \tag{8.13}$$

Then we have $\mathcal{S}_f(\mathbb{P},\mathbb{Q}) = f(1) - D_{\tilde{f}}(\mathbb{P},\mathbb{Q})$.

Here we use f-similarities since they are the quantities that naturally appear when developing our lower bounds. As the f-divergence, the f-similarity is in general asymmetric in $\mathbb{P}$ and $\mathbb{Q}$. Nevertheless for a concave function $f : \mathbb{R}_+ \to \mathbb{R}_+$, one may define a concave function $f^* : \mathbb{R}_+ \to \mathbb{R}_+$ as $f^*(u) = uf(1/u)$ and $f^*(0) = \lim_{u\to 0} uf(1/u)$, and we have (see [26] for the equivalent result for f-divergence): when $\mathbb{P} \ll \mathbb{Q}$ and $\mathbb{Q} \ll \mathbb{P}$,

$$\mathcal{S}_f(\mathbb{P},\mathbb{Q}) = \mathcal{S}_{f^*}(\mathbb{Q},\mathbb{P}). \tag{8.14}$$

We will use the following properties of f-similarities.

Lemma 8.5. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on a measurable space $(\mathcal{E},\mathcal{B})$ such that $\mathbb{P} \ll \mathbb{Q}$.*

1. *Let f and g be non-negative concave functions defined on $\mathbb{R}_+$ and let a and b be non-negative real numbers. For any probability distributions $\mathbb{P}$ and $\mathbb{Q}$, we have $\mathcal{S}_{af+bg}(\mathbb{P},\mathbb{Q}) = a\mathcal{S}_f(\mathbb{P},\mathbb{Q}) + b\mathcal{S}_g(\mathbb{P},\mathbb{Q})$. Besides if $f \le g$, then $\mathcal{S}_f(\mathbb{P},\mathbb{Q}) \le \mathcal{S}_g(\mathbb{P},\mathbb{Q})$.*

2. *Let $A \in \mathcal{B}$ such that $\frac{\mathbb{P}}{\mathbb{Q}} = 1$ on A. Let $A^c = \mathcal{E} - A$. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function. Let $\mathbb{P}'$ and $\mathbb{Q}'$ be probability distributions on $\mathcal{E}$ such that*

   - $\mathbb{P}' \ll \mathbb{Q}'$
   - $\frac{\mathbb{P}'}{\mathbb{Q}'} = 1$ *on A.*
   - $\mathbb{P}' = \mathbb{P}$ *and* $\mathbb{Q}' = \mathbb{Q}$ *on A^c, i.e. for any $B \in \mathcal{B}$, $\mathbb{P}'(B\cap A^c) = \mathbb{P}(B\cap A^c)$ and $\mathbb{Q}'(B\cap A^c) = \mathbb{Q}(B\cap A^c)$.*

   We have

$$\mathcal{S}_f(\mathbb{P}',\mathbb{Q}') = \mathcal{S}_f(\mathbb{P},\mathbb{Q}).$$

3. *Let $\mathcal{E}'$ be a measurable space and μ be a σ-finite positive measure on $\mathcal{E}'$. Let $\{f_v\}_{v\in\mathcal{E}'}$ be a family of non-negative concave functions defined on $\mathbb{R}_+$ such that $(u,v) \mapsto f_v\big(\frac{\mathbb{P}}{\mathbb{Q}}(u)\big)$ is measurable. We have*

$$\int_{\mathcal{E}'} \mathcal{S}_{f_v}(\mathbb{P},\mathbb{Q})\mu(dv) = \mathcal{S}_{\int_{\mathcal{E}'} f_v\mu(dv)}(\mathbb{P},\mathbb{Q}). \tag{8.15}$$

Proof. 1. It directly follows from the definition of f-similarities.

2. We have

$$\begin{aligned}\mathcal{S}_f(\mathbb{P}',\mathbb{Q}') &= \int_A f\big(\tfrac{\mathbb{P}'}{\mathbb{Q}'}\big)d\mathbb{Q}' + \int_{A^c} f\big(\tfrac{\mathbb{P}'}{\mathbb{Q}'}\big)d\mathbb{Q}' \\ &= f(1)\mathbb{Q}'(A) + \int_{A^c} f\big(\tfrac{\mathbb{P}}{\mathbb{Q}}\big)d\mathbb{Q} \\ &= f(1)\mathbb{Q}(A) + \int_{A^c} f\big(\tfrac{\mathbb{P}}{\mathbb{Q}}\big)d\mathbb{Q} \\ &= \mathcal{S}_f(\mathbb{P},\mathbb{Q})\end{aligned}$$

3. This Fubini's type result follows from the definition of the integral of non-negative functions on a product space.

□

8.3 Generalized Assouad's lemma

We recall that the n-fold product of a distribution P is denoted $P^{\otimes n}$. We start this section with a general lower bound for hypercubes of distributions (as defined in Section 8.1). This lower bound is expressed in terms of a similarity (as defined in Section 8.2) between n-fold products of representatives of the hypercube.

Theorem 8.6. *Let $\mathcal{P}$ be a set of probability distributions containing an hypercube of distributions of characteristic function $\tilde{\psi}$ and representatives $P_{[-]}$ and $P_{[+]}$. For any training set size $n \in \mathbb{N}^*$ and any estimator $\hat{g}$, we have*

$$\sup_{P\in\mathcal{P}} \big\{\mathbb{E}R(\hat{g}) - \min_g R(g)\big\} \geq \mathcal{S}_{\tilde{\psi}}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) \tag{8.16}$$

where the minimum is taken over the space of prediction functions and $\mathbb{E}R(\hat{g})$ denotes the expected risk of the estimator $\hat{g}$ trained on a sample of size n: $\mathbb{E}R(\hat{g}) = \mathbb{E}_{Z_1^n\sim P^{\otimes n}} R(\hat{g}_{Z_1^n}) = \mathbb{E}_{Z_1^n\sim P^{\otimes n}} \mathbb{E}_{(X,Y)\sim P}\, \ell[Y, \hat{g}_{Z_1^n}(X)]$.

Proof. See Section 10.10. □

This theorem provides a lower bound holding for any estimator and expressed in terms of the hypercube structure. To obtain a tight lower bound associated with a particular learning task, it then suffices to find the hypercube in $\mathcal{P}$ for which the r.h.s. of (8.16) is the largest possible. By providing lower bounds of $\mathcal{S}_{\tilde{\psi}}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)$ that are more explicit w.r.t. the hypercube parameters, we obtain the following results that are more in a ready-to-use form than Theorem 8.6.

Theorem 8.7. *Let $\mathcal{P}$ be a set of probability distributions containing an hypercube of distributions characterized by its dimension m, its edge probability w and its edge discrepancies d_{I} and d_{II} (see Definition 8.2). For any training set size $n \in \mathbb{N}^*$ and any estimator $\hat{g}$, the following assertions hold.*

1. *We have*

$$\begin{aligned}\sup_{P\in\mathcal{P}}\big\{\mathbb{E}R(\hat{g}) - \min_g R(g)\big\} &\geq mwd_{\mathrm{I}}\big(1-\sqrt{1-[1-d_{\mathrm{II}}]^{nw}}\big)\\ &\geq mwd_{\mathrm{I}}\big(1-\sqrt{nwd_{\mathrm{II}}}\big).\end{aligned} \tag{8.17}$$

2. *When the hypercube is constant and symmetrical(see Definition 8.2) , we also have*

$$\sup_{P\in\mathcal{P}}\big\{\mathbb{E}R(\hat{g}) - \min_g R(g)\big\} \geq mwd_{\mathrm{I}}\Big\{\mathbb{P}\Big(|N| > \sqrt{\tfrac{nwd_{\mathrm{II}}}{1-d_{\mathrm{II}}}}\Big) - d_{\mathrm{II}}^{1/4}\Big\} \tag{8.18}$$

   for N a centered gaussian random variable with variance 1.

3. *When the hypercube satisfies $p_+ \equiv 1 \equiv 1 - p_-$, we also have*

$$\sup_{P\in\mathcal{P}}\big\{\mathbb{E}R(\hat{g}) - \min_g R(g)\big\} \geq mwd_{\mathrm{I}}(1-w)^n \tag{8.19}$$

4. *When the hypercube is constant and symmetrical and when the differentiability assumption (see (8.9)) holds, we also have*

$$\begin{aligned}\sup_{P\in\mathcal{P}}\big\{\mathbb{E}R(\hat{g}) - \min_g R(g)\big\} &\geq \tfrac{mw\zeta d_{\mathrm{II}}}{8}\\ &\times\Big\{1+\tfrac{1}{2}\big[1-\big(1-\sqrt{1-d_{\mathrm{II}}}\big)w\big]^n - \tfrac{1}{2}\big[1+\big(\tfrac{1+d_{\mathrm{II}}}{\sqrt{1-d_{\mathrm{II}}}}-1\big)w\big]^n\Big\}\end{aligned} \tag{8.20}$$

Proof. See Section 10.11. □

The lower bounds (8.17), (8.18) and (8.20) are of the same nature. (8.17) is the general lower bound having the simplest form. For constant and symmetrical hypercubes, it can be refined into (8.18) and, when the differentiability assumption holds, into (8.20). These refinements mainly concern constants as we will see in Section 8.4.3. Finally, (8.19) is less general but provide results with tight constants when convergence rate of order n^{-1} has to be proven (see Remarks 8.6 [p.35] and 8.7 [p.37]).

To better understand the link between (8.17), (8.18) and (8.20), the following corollary considers an asymptotic setting in which n goes to infinity and the parameters of the hypercube varies with n (which is the typical situation even for finite sample lower bounds).

Corollary 8.8. *Let $a > 0$ and N be a centered gaussian random variable with variance 1. Under the assumptions of item 4 of the previous theorem, we have*

$$d_{\mathrm{I}} \geq \tfrac{\zeta}{8} d_{\mathrm{II}} \tag{8.21}$$

and (8.17), (8.18) *and* (8.20) *respectively lead to*

$$\begin{aligned} \liminf_{d_{\mathrm{II}} \to 0, nwd_{\mathrm{II}} \to a} \ \tfrac{1}{mwd_{\mathrm{I}}} \sup_{P \in \mathcal{P}} \big\{ \mathbb{E}R(\hat{g}) - \min_g R(g) \big\} \ &\geq \ 1 - \sqrt{1 - e^{-a}} \\ &\geq \ 1 - \sqrt{a}, \end{aligned} \tag{8.22}$$

$$\liminf_{d_{\mathrm{II}} \to 0, nwd_{\mathrm{II}} \to a} \ \tfrac{1}{mwd_{\mathrm{I}}} \sup_{P \in \mathcal{P}} \big\{ \mathbb{E}R(\hat{g}) - \min_g R(g) \big\} \geq \mathbb{P}(|N| > \sqrt{a}) \tag{8.23}$$

and

$$\liminf_{d_{\mathrm{II}} \to 0, nwd_{\mathrm{II}} \to a} \ \tfrac{8}{mw\zeta d_{\mathrm{II}}} \sup_{P \in \mathcal{P}} \big\{ \mathbb{E}R(\hat{g}) - \min_g R(g) \big\} \geq 1 + \tfrac{e^{-a/2}}{2} - \tfrac{e^{3a/2}}{2} \tag{8.24}$$

Proof. It follows from Theorem 8.7 and $(1+x)^{1/x} \to e$ when $x \to 0$. □

Inequality (8.21) leads to (slightly) weakened versions of (8.22) and (8.23) that can be directly compared with (8.24) (see Figure 2). A numerical comparison of these bounds is given in Section 8.4.3.

Remark 8.5. The previous lower bounds consider deterministic estimators (or algorithms), i.e. functions from the training set space $\cup_{n \geq 0} \mathcal{Z}^n$ to the prediction function space $\bar{\mathcal{G}}$. They still hold for randomized estimators, i.e. functions from the training set space to the set $\mathcal{D}$ of probability distributions on $\bar{\mathcal{G}}$.

8.4 Examples

Theorem 8.7 gives a very simple strategy to obtain a lower bound for a given set $\mathcal{P}$ of probability distributions and a reference set $\mathcal{G}$ of prediction functions: it consists in looking for the hypercube contained in the set $\mathcal{P}$ and for which

- the lower bound is maximized,
- for any distribution of the hypercube, $\mathcal{G}$ contains a best prediction function, i.e. $\min_g R(g) = \min_{g \in \mathcal{G}} R(g)$.

In general, the order of the bound is given by the quantity mwd_{I} (or $mw\zeta d_{\mathrm{II}}$ in the case of (8.20)) and the quantities w and d_{II} are taken such that nwd_{II} is of order 1.

In this section, we apply this strategy in different learning tasks. Before giving these lower bounds (Section 8.4.2), Section 8.4.1 stresses on the influence of the loss function in the computations of the edge discrepancy d_{I} and the constant ζ of the differentiability assumption (8.9).

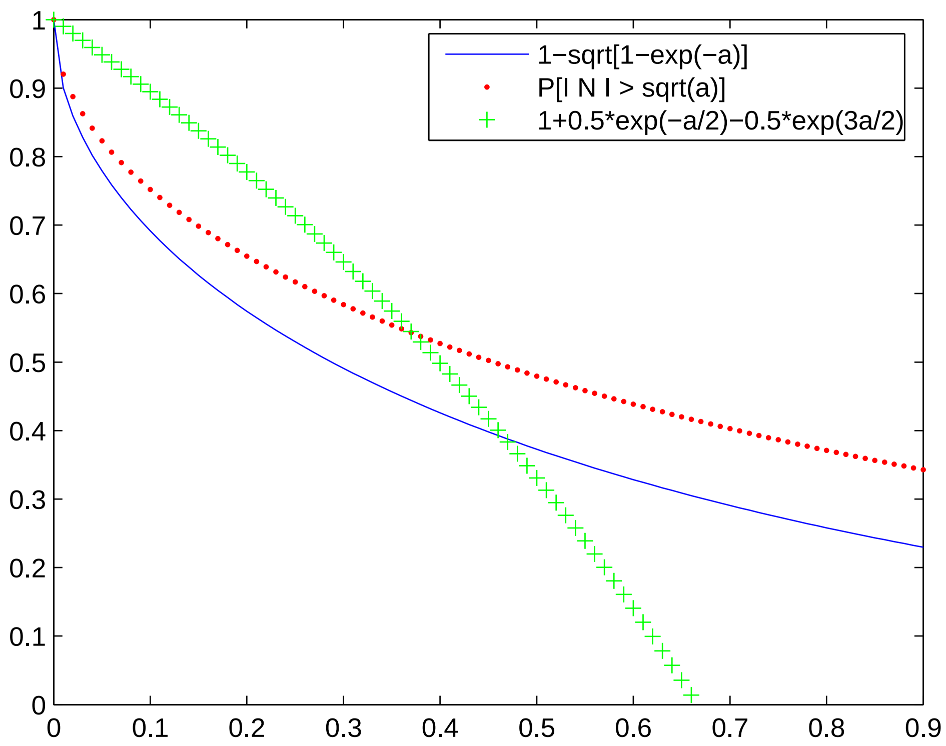

Figure 2: Comparison of the r.h.s. of (8.22), (8.23) and (8.24)

8.4.1 Edge discrepancy d_{I} and constant ζ of the differentiability assumption

All the previous lower bounds rely on either the edge discrepancy d_{I} or the constant ζ of the differentiability assumption (that has been introduced to control d_{I} in a simple way). The aim of this section is to provide more explicit formulas of these quantities for different loss functions.

To obtain the formula for d_{I}, we essentially use (8.6) and (8.5) jointly with the explicit computation of the second derivative of the function ϕ. To shorten notation, $\mu(\bullet|\mathcal{X}_1)$ denotes the conditional distribution $\mu(dX|X \in \mathcal{X}_j)$. For instance, (8.6) reduces to $d_{\mathrm{I}} = \mathbb{E}_{\mu(\bullet|\mathcal{X}_1)} \psi_{p_+,p_-,h_1,h_2}(1/2)$ and $d'_{\mathrm{I}} = \mathbb{E}_{\mu(\bullet|\mathcal{X}_1)}(p_+ - p_-)^2$.

Entropy loss. Here we consider $\mathcal{Y} = [0;1]$ and the loss for prediction y' instead of y is $\ell(y,y') = K(y,y')$, where $K(y,y')$ is the Kullback-Leibler divergence between Bernoulli distributions with respective parameters y and y', i.e. $K(y,y') = y \log\left(\frac{y}{y'}\right) + (1-y)\log\left(\frac{1-y}{1-y'}\right)$. Let $H(y)$ denote the Shannon's entropy of the Bernoulli distribution with parameter y, i.e.

$$H(y) = -y \log y - (1-y)\log(1-y). \tag{8.25}$$

Computations lead to: for any $p \in [0;1]$,

$$\phi_{y_1,y_2}(p) = H\big(py_1 + (1-p)y_2\big) - pH(y_1) - (1-p)H(y_2), \tag{8.26}$$

hence

$$\phi''_{y_1,y_2}(p) = -\frac{(y_1-y_2)^2}{[py_1+(1-p)y_2][(p(1-y_1)+(1-p)(1-y_2)]}.$$

This last equality is useful when one wants to compute ζ satisfying (8.9).

Classification loss. In this setting, we have $|\mathcal{Y}| < +\infty$ and the loss incurred by predicting y' instead of the true value y is $\ell(y,y') = \mathbf{1}_{y\neq y'}$. In this learning task, we have $\phi_{y_1,y_2}(p) = [p \wedge (1-p)]\mathbf{1}_{y_1\neq y_2}$ and $\phi''_{y_1,y_2} \equiv 0$ on $[0;1] - \{1/2\}$. Then (8.6), (8.5) and Remark 8.2 [p.22] lead to

$$d_{\mathrm{I}} = \mathbb{E}_{\mu(\bullet|\mathcal{X}_1)}\Big\{\big[\big|p_+ - \tfrac{1}{2}\big| \wedge \big|p_- - \tfrac{1}{2}\big|\big]\mathbf{1}_{(p_+-\frac{1}{2})(p_--\frac{1}{2})<0}\Big\}.$$

Binary classification losses (or regression losses when the output is binary). In this setting, we have $\mathcal{Y} = \mathbb{R}\cup\{-\infty;+\infty\}$, but we know that $P(Y \in \{-1;+1\}) = 1$. So we are only interested in hypercubes of distributions satisfying this constraint, i.e. such that for any $x \in \mathcal{X}$, $h_1(x)$ and $h_2(x)$ belong to $\{-1;+1\}$. In this setting, a best prediction function g, i.e. a measurable function from $\mathcal{X}$ to $\mathcal{Y}$ minimizing $R(g) = \mathbb{E}\,\ell[Y, g(X)]$, is determined by the regression function:

$$\eta(x) = P(Y = +1|X = x).$$

Let $\text{sign}(x) = \mathbf{1}_{x\geq 0} - \mathbf{1}_{x<0}$ be the sign function on $\mathbb{R}$.

- **$\mathbb{R}$-Classification loss**. The loss function is $\ell(y,y') = \mathbf{1}_{yy'<0}$ and a best prediction function is $g^*(x) = \text{sign}(\eta(x) - 1/2)$. Without surprise, we recover the same formulae as for the classification loss.

- **Hinge loss**. The loss function is $\ell(y,y') = (1-yy')_+ = \max\{0; 1-yy'\}$ and a best prediction function is $g^*(x) = \text{sign}(\eta(x) - 1/2)$. For any $y_1, y_2 \in \{-1;+1\}$, we have $\phi_{y_1,y_2}(p) = 2[p \wedge (1-p)]\mathbf{1}_{y_1\neq y_2}$ and $\phi''_{y_1,y_2} \equiv 0$ on $[0;1] - \{1/2\}$. Then (8.6), (8.5) and Remark 8.2 [p.22] lead to

$$d_{\mathrm{I}} = 2\mathbb{E}_{\mu(\bullet|\mathcal{X}_1)}\Big\{\big[\big|p_+ - \tfrac{1}{2}\big| \wedge \big|p_- - \tfrac{1}{2}\big|\big]\mathbf{1}_{(p_+-\frac{1}{2})(p_--\frac{1}{2})<0}\Big\},$$

- **Exponential loss (or AdaBoost loss)**. The loss function is $\ell(y,y')=e^{-yy'}$. For any $y_1\neq y_2\in\{-1;+1\}$ and any $p\in[0;1]$, the function φ_{p,y_1,y_2} is minimized for $y=\frac{y_1}{2}\log\left(\frac{p}{1-p}\right)$, so a best prediction function is $g^*(x)=\frac{1}{2}\log\left(\frac{\eta(x)}{1-\eta(x)}\right)$. We obtain $\phi_{y_1,y_2}(p)=2\sqrt{p(1-p)}\mathbf{1}_{y_1\neq y_2}$ and
$$\phi''_{y_1,y_2}(p)=-\tfrac{1}{2[p(1-p)]^{3/2}}\mathbf{1}_{y_1\neq y_2}.$$
In this setting, to obtain a lower bound of d_{I}, one has typically to compute ζ satisfying (8.9) and to use (8.10).

- **Logit loss**. The loss function is $\ell(y,y')=\log(1+e^{-yy'})$. For any $y_1\neq y_2\in\{-1;+1\}$ and any $p\in[0;1]$, the function φ_{p,y_1,y_2} is minimized for $y=y_1\log\left(\frac{p}{1-p}\right)$, so a best prediction function is $g^*(x)=\log\left(\frac{\eta(x)}{1-\eta(x)}\right)$. We obtain $\phi_{y_1,y_2}(p)=H(p)\mathbf{1}_{y_1\neq y_2}$, where $H(p)$ denote the Shannon's entropy of the Bernoulli distribution with parameter p (see (8.25)). We get
$$\phi''_{y_1,y_2}(p)=-\tfrac{1}{p(1-p)}\mathbf{1}_{y_1\neq y_2}.$$
Once more, to obtain a lower bound of d_{I}, one has typically to compute ζ satisfying (8.9) and to use (8.10).

L_q-loss. We consider $\mathcal{Y}=\mathbb{R}$ and the loss function is $\ell(y,y')=|y-y'|^q$ with $q\geq 1$. The values $q=1$ and $q=2$ respectively correspond to the absolute loss and the least square loss.

- *Case* $q=1$ *:* Due to the lack of strong convexity of the loss function, the absolute loss setting differs completely from what occurs for $q>1$ and appears to be similar to the classification and hinge losses settings. Indeed computations lead to $\phi_{y_1,y_2}(p)=[p\wedge(1-p)]|y_2-y_1|$ and $\phi''_{y_1,y_2}\equiv 0$ on $[0;1]-\{1/2\}$. Then (8.6) and (8.5) lead to
$$d_{\mathrm{I}}=\mathbb{E}_{\mu(\bullet|\mathcal{X}_1)}\Big\{|h_2-h_1|\big[\big|p_+-\tfrac{1}{2}\big|\wedge\big|p_--\tfrac{1}{2}\big|\big]\mathbf{1}_{(p_+-\frac{1}{2})(p_--\frac{1}{2})<0}\Big\}. \tag{8.27}$$

- *Case* $q>1$ *:* Tedious computations put in Appendix A lead to: for any $p\in[0;1]$,
$$\phi_{y_1,y_2}(p)=p(1-p)\frac{|y_2-y_1|^q}{\left[p^{\frac{1}{q-1}}+(1-p)^{\frac{1}{q-1}}\right]^{q-1}} \tag{8.28}$$
and
$$\phi''_{y_1,y_2}(p)=-\tfrac{q}{q-1}[p(1-p)]^{\frac{2-q}{q-1}}\frac{|y_2-y_1|^q}{\left[p^{\frac{1}{q-1}}+(1-p)^{\frac{1}{q-1}}\right]^{q+1}} \tag{8.29}$$
To obtain a lower bound of d_{I}, as for the entropy loss, one has to compute ζ satisfying (8.9) and to use (8.10).

- *Special case* $q = 2$ *:* For the least square setting, the formulae simplify into $\phi_{y_1,y_2}(p) = p(1-p)|y_2 - y_1|^2$ and $\phi''_{y_1,y_2}(p) = -2|y_2 - y_1|^2$. Then the edge discrepancy d_{I} can be written explicitly as
$$d_{\mathrm{I}} = \tfrac{1}{4}\mathbb{E}_{\mu(\bullet|\mathcal{X}_1)}\big\{(p_+ - p_-)^2(h_2 - h_1)^2\big\}. \tag{8.30}$$

8.4.2 Various learning lower bounds

Here we give learning lower bounds matching up to multiplicative constants the upper bounds developed in the previous sections. The logarithm in base 2 is denoted by $\log_2$ and $\lfloor x \rfloor$ denotes the largest integer k such that $k \le x$.

Entropy loss setting. We consider $\mathcal{Y} = [0;1]$ and $\ell(y, y') = K(y, y')$ (see p.32). We have seen in Section 4 that there exists an estimator $\hat{g}$ such that
$$\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \le \tfrac{\log|\mathcal{G}|}{n} \tag{8.31}$$

The following consequency of (8.19) shows that this result is tight.

Theorem 8.9. *For any training set size* $n \in \mathbb{N}^*$*, positive integer* d *and input space* $\mathcal{X}$ *containing at least* $\lfloor \log_2(2d) \rfloor$ *points, there exists a set* $\mathcal{G}$ *of* d *prediction functions such that: for any estimator* $\hat{g}$ *there exists a probability distribution on the data space* $\mathcal{X} \times [0;1]$ *for which*
$$\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \ge e^{-1}(\log 2)\big(1 \wedge \tfrac{\lfloor \log_2|\mathcal{G}|\rfloor}{n+1}\big)$$

Proof. We use a $\big(\tilde{m}, \frac{1}{n+1} \wedge \frac{1}{\tilde{m}}, 1\big)$-hypercube with $\tilde{m} = \lfloor \log_2|\mathcal{G}| \rfloor = \big\lfloor \frac{\log|\mathcal{G}|}{\log 2} \big\rfloor$, $h_1 \equiv 0$ and $h_2 \equiv 1$. From (8.4), (8.6) and (8.26), we have
$$d_{\mathrm{I}} = \psi_{1,0,0,1}(1/2) = \phi_{0,1}(1/2) = H(1/2) = \log 2.$$

From (8.19), we obtain
$$\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \ge \big(\tfrac{\lfloor \log_2|\mathcal{G}|\rfloor}{n+1} \wedge 1\big)(\log 2)\big(1 - \tfrac{1}{n+1} \wedge \tfrac{1}{\lfloor \log_2|\mathcal{G}|\rfloor}\big)^n$$

Then the result follows from $[1 - 1/(n+1)]^n \searrow e^{-1}$. □

Remark 8.6. For $|\mathcal{G}| < 2^{n+2}$, the lower bound matches the upper bound (8.31) up to the multiplicative factor $e \approx 2.718$. For $|\mathcal{G}| \ge 2^{n+2}$, the size of the model is too large and, without any extra assumption, no estimator can learn from the data. To prove the result, we consider distributions for which the output is deterministic when knowing the input. So the lower bound does not come from noisy situations but from situations in which different prediction functions are not separated by the data to the extent that no input data falls into the (small) subset on which they are different.

L_q**-regression with bounded outputs.** We consider $\mathcal{Y} = [-B; B]$ and $\ell(y, y') = |y - y'|^q$ (see p.34). The following two theorems are roughly summed up in Figure 3 that represents the optimal convergence rate for L_q-regression.

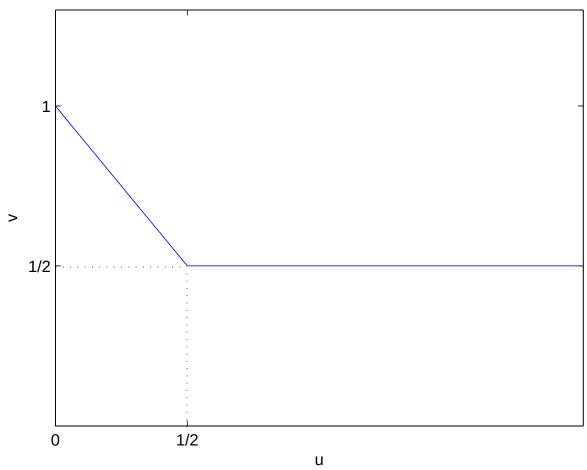

Figure 3: *Influence of the convexity of the loss on the optimal convergence rate.* Let $c > 0$. We consider L_q-losses with $q = 1 + c\big(\frac{\log|\mathcal{G}|}{n}\big)^u$ for $u \geq 0$. For such values of q, the optimal convergence rate of the associated learning task is of order $\big(\frac{\log|\mathcal{G}|}{n}\big)^v$ with $1/2 \leq v \leq 1$. This figure represents the value of u in abscissa and the value of v in ordinate. The value $u = 0$ corresponds to constant q greater than 1. For these q, the optimal convergence rate is of order n^{-1} while for $q = 1$ or "very close" to 1, the convergence rate is of order $n^{-1/2}$.

- *Case* $1 \leq q \leq 1 + \sqrt{\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{4n}} \wedge 1$*:* From (6.5), there exists an estimator $\hat{g}$ such that

$$\mathbb{E}R(\hat{g}) - \min_{g \in \mathcal{G}} R(g) \leq 2^{\frac{2q-1}{2}} B^q \sqrt{\frac{\log|\mathcal{G}|}{n}} \tag{8.32}$$

The following corollary of Theorem 8.7 shows that this result is tight.

Theorem 8.10. *Let $B > 0$ and $d \in \mathbb{N}^*$. For any training set size $n \in \mathbb{N}^*$ and any input space $\mathcal{X}$ containing at least $\lfloor \log_2 d \rfloor$ points, there exists a set $\mathcal{G}$ of d prediction functions such that: for any estimator $\hat{g}$ there exists a probability distribution on the data space $\mathcal{X} \times [-B; B]$ for which*

$$\mathbb{E}R(\hat{g}) - \min_{g \in \mathcal{G}} R(g) \geq \begin{cases} c_q B^q \sqrt{\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n}} & \text{if } |\mathcal{G}| < 2^{4n+1} \\ 2c_q B^q & \text{otherwise} \end{cases},$$

where

$$c_q = \begin{cases} 1/4 & \textit{if} \;\; q = 1 \\ q/40 & \textit{if} \;\; 1 < q \le 1 + \sqrt{\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{4n} \wedge 1} \end{cases}$$

Proof. See Section 10.12. □

- *Case* $q > 1 + \sqrt{\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{4n} \wedge 1}$ *:* We have seen in Section 4 that there exists an estimator $\hat{g}$ such that

$$\mathbb{E}R(\hat{g}) - \min_{g \in \mathcal{G}} R(g) \le \frac{q(1 \wedge 2^{q-2}) B^q}{q-1} (\log 2) \frac{\log_2 |\mathcal{G}|}{n} \tag{8.33}$$

The following corollary of Theorem 8.7 shows that this result is tight.

Theorem 8.11. *Let $B > 0$ and $d \in \mathbb{N}^*$. For any training set size $n \in \mathbb{N}^*$ and input space $\mathcal{X}$ containing at least $\lfloor \log_2(2d) \rfloor$ points, there exists a set $\mathcal{G}$ of d prediction functions such that: for any estimator $\hat{g}$ there exists a probability distribution on the data space $\mathcal{X} \times [-B; B]$ for which*

$$\mathbb{E}R(\hat{g}) - \min_{g \in \mathcal{G}} R(g) \ge \big(\tfrac{q}{90(q-1)} \vee e^{-1} \big) B^q \big(\tfrac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n+1} \wedge 1 \big).$$

Proof. See Section 10.12. □

Remark 8.7. For least square regression (i.e. q=2), Remark 8.6 [p.35] holds provided that the multiplicative factor becomes $2e \log 2 \approx 3.77$. More generally, the method used here gives close to optimal constants but not the exact ones. We believe that this limit is due to the use of the hypercube structure. Indeed, the reader may check that for hypercubes of distributions, the upper bounds used in this section are not constant-optimal since the simplifying step consisting in using $\min_{\rho \in \mathcal{M}} \cdots \le \min_{g \in \mathcal{G}} \ldots$ is loose.

The reader may check that there are essentially two classes of bounded losses: the ones which are not convex or not enough convex (typical examples are the classification loss, the hinge loss and the absolute loss) and the ones which are sufficiently convex (typical examples are the least square loss, the entropy loss, the logit loss and the exponential loss). For the first class of losses, the edge discrepancy of type I is proportional to $\sqrt{d_{\mathrm{II}}}$ for constant and symmetrical hypercubes and (8.17) leads to a convergence rate of $\sqrt{(\log |\mathcal{G}|)/n}$. For the second class, the convergence rate is $(\log |\mathcal{G}|)/n$ and the lower bound can be explained by the fact that, when two prediction functions are different on a set with low probability (typically n^{-1}), it often happens that the training data has no input points in this set. For such training data, it is impossible to consistently choose the right prediction function.

L_q-regression for unbounded outputs having finite moments.

- *Case* $q = 1$ *:* From (7.1), when $\sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \le b$ for some $b > 0$, there exists an estimator for which $\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \le 2b\sqrt{(2\log|\mathcal{G}|)/n}$. This means that even when the data are unbounded, $\sqrt{(\log|\mathcal{G}|)/n}$ is the minimax optimal convergence rate.

- *Case* $q > 1$ *:* First let us recall the upper bound. In Corollary 7.4, under the assumptions

$$\begin{cases} \sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \le b & \text{for some } b > 0 \\ \mathbb{E}|Y|^s \le A & \text{for some } s \ge q \text{ and } A > 0 \\ \mathcal{G} \text{ finite} \end{cases}$$

we have proposed an algorithm satisfying

$$R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \le \begin{cases} C\left(\frac{\log|\mathcal{G}|}{n}\right)^{1-\frac{q-1}{s}} & \text{when } q \le s \le 2q-2 \\ C\left(\frac{\log|\mathcal{G}|}{n}\right)^{1-\frac{q}{s+2}} & \text{when } s \ge 2q-2 \end{cases}.$$

for a quantity C which depends only on b, A, q and s.

The following corollary of Theorem 8.7 shows that this result is tight and is illustrated by Figure 4.

Theorem 8.12. *Let $d \in \mathbb{N}^*$, $s \ge q > 1$, $b > 0$ and $A > 0$. For any training set size $n \in \mathbb{N}^*$ and input space $\mathcal{X}$ containing at least $\lfloor\log_2(2d)\rfloor$ points, there exists a set $\mathcal{G}$ of d prediction functions uniformly bounded by b such that: for any estimator $\hat{g}$ there exists a probability distribution on the data space $\mathcal{X}\times\mathbb{R}$ for which $\mathbb{E}|Y|^s \le A$ and*

$$\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \ge \begin{cases} C\left(\frac{\log|\mathcal{G}|}{n} \wedge 1\right)^{1-\frac{q-1}{s}} \\ C\left(\frac{\log|\mathcal{G}|}{n} \wedge 1\right)^{1-\frac{q}{s+2}} \end{cases},$$

for a quantity C which depends only on the real numbers b, A, q and s.

Both inequalities simultaneously hold but the first one is tight for $q \le s \le 2q-2$ while the second one is tight for $s \ge 2q-2$. They are both based on (8.17) applied to a $\lfloor\log_2|\mathcal{G}|\rfloor$-dimensional hypercubes.

Contrary to other lower bounds obtained in this work, the first inequality is based on asymmetrical hypercubes. The use of this kind of hypercubes can be partially explained by the fact that the learning task is asymmetrical. Indeed all values of the output space do not have the same status since predictions are constrained to be in $[-b; b]$ while outputs are allowed to be in the whole real space (see the constraints on the hypercube in the proof given in Section 10.13).

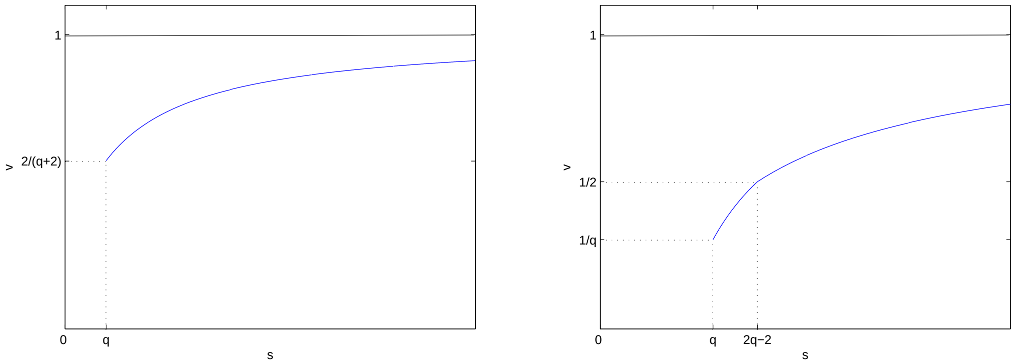

Figure 4: *Optimal convergence rates in L_q-regression when the output has a finite moment of order s (see Theorem 8.12).* The convergence rate is of order $\left(\frac{log|\mathcal{G}|}{n}\right)^v$ with $0 < v \leq 1$. The figure represents the value of s in abscissa and the value of v in ordinate. Two cases have to be distinguished. For $1 < q \leq 2$ (figure on the left), v depends smoothly on q. For $q > 2$ (figure on the right), two stages are observed depending whether s is larger than $2q - 2$.

8.4.3 Numerical comparison of the lower bounds (8.17), (8.18) and (8.20)

To compare (8.17), (8.18) and (8.20), we will compare their asymptotical version, i.e. (8.22), (8.23) and (8.24). Since the differentiability assumption does not hold in this setting, we only compare (8.17) and (8.18).

Classification in VC classes. We consider $|\mathcal{Y}| = 2$ and $\ell(y, y') = \mathbf{1}_{y \neq y'}$. Since the differentiability assumption does not hold in this setting, we only compare (8.17) and (8.18).

Theorem 8.13. *Let $\mathcal{P}$ be the set of all probability distributions on the data space $\mathcal{Z}$. Let $(\mathcal{G}_n)_{n\in\mathbb{N}}$ be a family of prediction function spaces of VC-dimension $V_n \geq 2$ satisfying $n/V_n \underset{n\to+\infty}{\longrightarrow} +\infty$. For any algorithm $\hat{g}$:*

$$\liminf_{n\to+\infty} \sqrt{\tfrac{n}{V_n}} \sup_{P\in\mathcal{P}} \left\{ \mathbb{E}_{Z_1^n} R(\hat{g}_{Z_1^n}) - \inf_{g\in\mathcal{G}_n} R(g) \right\} \geq \begin{cases} \alpha_1 & \textit{from } (8.22) \\ \alpha_2 & \textit{from } (8.23) \end{cases}$$

with $\begin{cases} \alpha_1 & = & \max_{a>0} \frac{\sqrt{a}(1-\sqrt{1-e^{-a}})}{2} \approx 0.135 \\ \alpha_2 & = & \max_{a>0} \sqrt{\frac{a}{2\pi}} \int_a^{+\infty} e^{-t^2/2} dt \approx 0.170 \end{cases}$.

In particular for a given set $\mathcal{G}$ of finite VC-dimension V, for n sufficiently large, any estimator $\hat{g}$ satisfies

$$\sup_{P\in\mathcal{P}} \left\{ \mathbb{E}R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \right\} \geq \tfrac{1}{6}\sqrt{V/n}.$$

Proof. It suffices to apply Corollary 8.8 to $(V_n, 1/V_n, aV_n/n)$-hypercubes, use that $d_{\mathrm{I}} = \sqrt{d_{\mathrm{II}}}/2$ and choose the real number $a > 0$ to maximize the lower bound. □

The two inequalities, coming from (8.22) and (8.23), simultaneously hold. They only differ by a multiplicative constant.

Least square regression with unbounded outputs satisfying $\mathbb{E}Y^2 \le A$ for some $A > 0$. We consider the context of Corollary 7.3 with $s = 2$. The best explicit constant in (7.4) is obtained from (7.3) for $\lambda = \sqrt{\frac{\log|\mathcal{G}|}{8b^2A(n+1)}} \wedge (8b^2)^{-1}$. When $\log|\mathcal{G}| \le An/(8b^2)$, we get

$$R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \le \sqrt{32}\, b\sqrt{A}\sqrt{\frac{\log|\mathcal{G}|}{n}}$$

Now let us give the associated lower bounds coming from (8.22), (8.23) and (8.24).

Theorem 8.14. *Let $\mathcal{X}$ be an infinite input space. Let $A > 0$ and $\mathcal{P}$ be the set of probability distributions on $\mathcal{X} \times \mathbb{R}$ such that $\mathbb{E}Y^2 \le A$. There exists a family of prediction function spaces $(\mathcal{G}_n)_{n\in\mathbb{N}}$ such that any prediction function in these sets is uniformly bounded by b, their sizes at most grows subexponentially, i.e. $n/\log|\mathcal{G}_n|$ goes to infinity when n goes to infinity, and for any algorithm $\hat{g}$*

$$\liminf_{n\to+\infty} \sqrt{\frac{n}{\log|\mathcal{G}_n|}} \sup_{P\in\mathcal{P}} \Big\{\mathbb{E}_{Z_1^n} R(\hat{g}_{Z_1^n}) - \min_g R(g)\Big\} \ge \begin{cases} \beta_1\, b\sqrt{A} & \textit{from } (8.22) \\ \beta_2\, b\sqrt{A} & \textit{from } (8.23) \\ \beta_3\, b\sqrt{A} & \textit{from } (8.24) \end{cases}$$

with

$$\begin{cases} \beta_1 &= (\log 2)^{-1} \max_{a>0} \sqrt{a}(1 - \sqrt{1-e^{-a}}) \approx 0.3897 \\ \beta_2 &= (\log 2)^{-1} \max_{a>0} \sqrt{\frac{2a}{\pi}} \int_a^{+\infty} e^{-t^2/2} dt \approx 0.4904 \\ \beta_3 &= (\log 2)^{-1} \max_{a>0} \sqrt{a}\big(1 + \frac{1}{2}e^{-a/2} - \frac{1}{2}e^{3a/2}\big) \approx 0.5154 \end{cases}.$$

Proof. See Section 10.14. □

The three inequalities, coming from (8.22), (8.23) and (8.24), only differ by a multiplicative constant. The one coming from (8.24) gives the tightest result. The difference between the upper bound and this lower bound is a multiplicative factor smaller than 11.

Remark 8.8. In this section, we have only considered constant hypercubes. The use of non-constant hypercubes can be required when smoothness assumptions are put on the regression function $\eta : x \mapsto P(Y = 1|X = x)$. This is typically the case in works on plug-in classifiers. For instance, the proof of [7, Theorems 3.5 and 4.1] relies on non-constant symmetrical hypercubes for which the function ξ (see Definition 8.2) is chosen such that it vanishes on the border of the partition cells, which ensures the regularity of the regression function η.

9 Conclusion and open problems

This work has developed minimax optimal risk bounds for the general learning task consisting in predicting as well as the best function in a reference set. It has proposed to summarize this learning problem by the variance function of the key condition (3.1). The generic algorithm based on this variance function leads to optimal convergence rates in the model selection aggregation problem, and our analysis gives a nice unified view to results coming from different communities.

Besides without any extra assumption on the learning task, we have obtained a Bernstein's type bound which has no known equivalent form when the loss function is not assumed to be bounded. When the loss function is bounded, the use of Hoeffding's inequality w.r.t. Gibbs distributions on the prediction function space instead of the distribution generating the data leads to an improvement by a factor 2 of the standard-style risk bound.

To prove that our bounds are minimax optimal, we have refined Assouad's lemma particularly by taking into account the properties of the loss function. We have also illustrated our upper and lower bounds by studying the influence of the noise of the output and of the convexity of the loss function.

Finally this work has the following limits. Our results concern expected risks and it is an open problem to provide corresponding tight exponential inequalities. Besides we should emphasize that our expected risk upper bounds hold only for our algorithm. This is quite different from the classical point of view that simultaneously gives upper bounds on the risk of any prediction function in the model. To our current knowledge, this classical approach has a flexibility that is not recovered in our approach. For instance, in several learning tasks, Dudley's chaining trick [29] is the only way to prove risk convergence with the optimal rate. So a natural question and another open problem is whether it is possible to combine the better variance control presented here with the chaining argument (or other localization argument used while exponential inequalities are available).

10 Proofs

10.1 Proof of Theorem 4.4

First, by a scaling argument, it suffices to prove the result for $a = 0$ and $b = 1$. For $\mathcal{Y} = [0;1]$, we modify the proof in Appendix A of [33]. Precisely, claims 1 and 2, with the notation used there, become:

1. If the function f is concave in $\alpha([p;q])$ then we have $A_t(q) \leq B_t(p)$,
2. If $c \geq R(z,p,q)$ for any $z \in]p;q[$, then the function f is concave in $\alpha([p;q])$.

Up to the missing α (typo), the difference is that we restrict ourselves to values of z in $[p;q]$. The proof of Claim 2 has no new argument. For claim 1, it suffices to modify the definition of $x'_{t,i}$ into $x'_{t,i} = q \wedge G^{-1}[\ell(p,x_{t,i})] \in [p;q]$. Then we have $L(p,x'_{t,i}) \leq L(p,x_{t,i})$ and $L(q,x'_{t,i}) \leq L(p,x_{t,i})$, hence $\alpha(x'_{t,i}) \geq \alpha(x_{t,i})$ and $\gamma(x'_{t,i}) \geq \gamma(x_{t,i})$. Now one can prove that f is decreasing on $\alpha([p;q])$. By using Jensen's inequality, we get

$$\begin{aligned}
\Delta_t(q) &= -c\log \textstyle\sum_{i=1}^n v_{t,i}\gamma(x_{t,i}) \\
&\geq -c\log \textstyle\sum_{i=1}^n v_{t,i}\gamma(x'_{t,i}) \\
&= -c\log \textstyle\sum_{i=1}^n v_{t,i} f[\alpha(x'_{t,i})] \\
&\geq -c\log f\Big[\textstyle\sum_{i=1}^n v_{t,i}\alpha(x'_{t,i})\Big] \\
&\geq -c\log f\Big[\textstyle\sum_{i=1}^n v_{t,i}\alpha(x_{t,i})\Big] \\
&= L[q, G^{-1}(\Delta_t(p))]
\end{aligned}$$

The end of the proof of claim 1 is then identical.

10.2 Proof of Corollary 5.1

We start by proving that condition (3.1) holds with $\delta_\lambda \equiv 0$, and that we may take $\hat{\pi}(\rho)$ be the Dirac distribution at the function $\mathbb{E}_{g\sim\rho}\, g$. By using Jensen's inequality and Fubini's theorem, Assumption (5.1) implies that

$$\begin{aligned}
&\mathbb{E}_{g'\sim\hat{\pi}(\rho)}\,\mathbb{E}_{Z\sim P}\,\log\mathbb{E}_{g\sim\rho}\, e^{\lambda[L(Z,g')-L(Z,g)]} \\
&\qquad = \mathbb{E}_{Z\sim P}\,\log\mathbb{E}_{g\sim\rho}\, e^{\lambda[L(Z,\mathbb{E}_{g'\sim\rho}\, g')-L(Z,g)]} \\
&\qquad \leq \log\mathbb{E}_{g\sim\rho}\,\mathbb{E}_{Z\sim P}\, e^{\lambda[L(Z,\mathbb{E}_{g'\sim\rho}\, g')-L(Z,g)]} \\
&\qquad \leq \log\mathbb{E}_{g\sim\rho}\,\psi(\mathbb{E}_{g'\sim\rho}\, g', g) \\
&\qquad \leq \log\psi(\mathbb{E}_{g'\sim\rho}\, g', \mathbb{E}_{g\sim\rho}\, g) \\
&\qquad = 0,
\end{aligned}$$

so that we can apply Theorem 3.1. It remains to note that in this context our generic algorithm is the one described in the corollary.

10.3 Proof of Theorem 6.1

To check that condition (3.1) holds, it suffices to prove that for any $z \in \mathcal{Z}$

$$\mathbb{E}_{g'\sim\rho} \log \mathbb{E}_{g\sim\rho} e^{\lambda[L(z,g')-L(z,g)]-\frac{\lambda^2}{2}[L(z,g')-L(z,g)]^2} \leq 0. \tag{10.1}$$

To shorten formulae, let $\alpha(g',g) \triangleq \lambda[L(z,g')-L(z,g)]$. By Jensen's inequality and the following symmetrization trick, (10.1) holds.

$$\begin{aligned}
&\mathbb{E}_{g'\sim\rho} \mathbb{E}_{g\sim\rho} e^{\alpha(g',g)-\frac{\alpha^2(g',g)}{2}} \\
&\leq \tfrac{1}{2}\mathbb{E}_{g'\sim\rho} \mathbb{E}_{g\sim\rho} e^{\alpha(g',g)-\frac{\alpha^2(g',g)}{2}} + \tfrac{1}{2}\mathbb{E}_{g'\sim\rho} \mathbb{E}_{g\sim\rho} e^{-\alpha(g',g)-\frac{\alpha^2(g',g)}{2}} \\
&\leq \mathbb{E}_{g'\sim\rho} \mathbb{E}_{g\sim\rho} \cosh\big(\alpha(g,g')\big)e^{-\frac{\alpha^2(g',g)}{2}} \\
&\leq 1
\end{aligned} \tag{10.2}$$

where in the last inequality we used the inequality $\cosh(t) \leq e^{t^2/2}$ for any $t \in \mathbb{R}$. The result then follows from Theorem 1.

10.4 Proof of Corollary 6.2

To shorten the following formula, let μ denote the law of the prediction function produced by our generic algorithm (w.r.t. simultaneously the training set and the randomizing procedure). Then (6.1) can be written as: for any $\rho \in \mathcal{M}$,

$$\mathbb{E}_{g'\sim\mu} R(g') \leq \mathbb{E}_{g\sim\rho} R(g) + \tfrac{\lambda}{2}\mathbb{E}_{g\sim\rho} \mathbb{E}_{g'\sim\mu} V(g,g') + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \tag{10.3}$$

Define $\tilde{R}(g) = R(g) - R(\tilde{g})$ for any $g \in \mathcal{G}$. Under the generalized Mammen and Tsybakov's assumption, for any $g, g' \in \mathcal{G}$, we have

$$\begin{aligned}
\tfrac{1}{2}V(g,g') &\leq \mathbb{E}_{Z\sim P}\left\{[L(Z,g)-L(Z,\tilde{g})]^2\right\} + \mathbb{E}_{Z\sim P}\left\{[L(Z,g')-L(Z,\tilde{g})]^2\right\} \\
&\leq c\tilde{R}^\gamma(g) + c\tilde{R}^\gamma(g'),
\end{aligned}$$

so that (10.3) leads to

$$\mathbb{E}_{g'\sim\mu} [\tilde{R}(g') - c\lambda\tilde{R}^\gamma(g)] \leq \mathbb{E}_{g\sim\rho} [\tilde{R}(g) + c\lambda\tilde{R}^\gamma(g)] + \tfrac{K(\rho,\pi)}{\lambda(n+1)}. \tag{10.4}$$

This gives the first assertion. For the second statement, let $u \triangleq \mathbb{E}_{g'\sim\mu} \tilde{R}(g')$ and $\chi(u) \triangleq u - c\lambda u^\gamma$. By Jensen's inequality, the l.h.s. of (10.4) is lower bounded by $\chi(u)$. By straightforward computations, for any $0 < \beta < 1$, when $u \geq \big(\frac{c\lambda}{1-\beta}\big)^{\frac{1}{1-\gamma}}$, $\chi(u)$ is lower bounded by βu, which implies the desired result.

10.5 Proof of Theorem 6.3

Let us prove (6.3). Let $r(g)$ denote the empirical risk of $g \in \mathcal{G}$, that is $r(g) = \frac{\Sigma_n(g)}{n}$. Let $\rho \in \mathcal{M}$ be some fixed distribution on $\mathcal{G}$. From [4, Section 8.1], with probability at least $1-\epsilon$ w.r.t. the training set distribution, for any $\mu \in \mathcal{M}$, we have

$$\begin{aligned}&\mathbb{E}_{g'\sim\mu}\, R(g') - \mathbb{E}_{g\sim\rho}\, R(g)\\ &\le \mathbb{E}_{g'\sim\mu}\, r(g') - \mathbb{E}_{g\sim\rho}\, r(g) + \lambda\varphi(\lambda B)\mathbb{E}_{g'\sim\mu}\,\mathbb{E}_{g\sim\rho}\, V(g,g') + \tfrac{K(\mu,\pi)+\log(\epsilon^{-1})}{\lambda n}.\end{aligned}$$

Since the Gibbs distribution $\pi_{-\lambda\Sigma_n}$ minimizes $\mu \mapsto \mathbb{E}_{g'\sim\mu}\, r(g') + \frac{K(\mu,\pi)}{\lambda n}$, we have

$$\begin{aligned}&\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, R(g')\\ &\qquad\le \mathbb{E}_{g\sim\rho}\, R(g) + \lambda\varphi(\lambda B)\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\,\mathbb{E}_{g\sim\rho}\, V(g,g') + \tfrac{K(\rho,\pi)+\log(\epsilon^{-1})}{\lambda n}.\end{aligned}$$

Then we apply the following inequality

$$\mathbb{E}W \le \mathbb{E}(W\vee 0) = \int_0^{+\infty} \mathbb{P}(W>u)du = \int_0^1 \epsilon^{-1}\mathbb{P}(W>\log(\epsilon^{-1}))d\epsilon$$

to the random variable

$$\begin{aligned}W = \lambda n\big[&\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\, R(g') - \mathbb{E}_{g\sim\rho}\, R(g)\\ &-\lambda\varphi(\lambda B)\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n}}\,\mathbb{E}_{g\sim\rho}\, V(g,g')\big] - K(\rho,\pi).\end{aligned}$$

We get $\mathbb{E}W \le 1$. At last we may choose the distribution ρ minimizing the upper bound to obtain (6.3). Similarly using [4, Section 8.3], we may prove (6.2).

10.6 Proof of Lemma 6.5

It suffices to apply the following adaptation of Lemma 5 of [47] to

$$\xi_i(Z_1,\dots,Z_i) = L[Z_i, \mathcal{A}(Z_1^{i-1})] - L(Z_i,\tilde{g}).$$

Lemma 10.1. *Let φ still denote the positive convex increasing function defined as $\varphi(t) \triangleq \frac{e^t-1-t}{t^2}$. Let b be a real number. For $i=1,\dots,n+1$, let $\xi_i : \mathcal{Z}^i \to \mathbb{R}$ be a function uniformly upper bounded by b. For any $\eta>0$, $\epsilon>0$, with probability at least $1-\epsilon$ w.r.t. the distribution of $Z_1,\dots,Z_{n+1}$, we have*

$$\begin{aligned}\textstyle\sum_{i=1}^{n+1} \xi_i(Z_1,\dots,Z_i) \le \sum_{i=1}^{n+1} &\mathbb{E}_{Z_i}\xi_i(Z_1,\dots,Z_i)\\ &+\eta\varphi(\eta b)\textstyle\sum_{i=1}^{n+1}\mathbb{E}_{Z_i}\xi_i^2(Z_1,\dots,Z_i) + \frac{\log(\epsilon^{-1})}{\eta},\end{aligned} \tag{10.5}$$

where $\mathbb{E}_{Z_i}$ denotes the expectation w.r.t. the distribution of Z_i only.

Remark 10.1. The same type of bounds without variance control can be found in [22].

Proof. For any $i \in \{0, \dots, n+1\}$, define

$$\psi_i = \psi_i(Z_1, \dots, Z_i) \triangleq \sum_{j=1}^{i} \xi_j - \sum_{j=1}^{i} \mathbb{E}_{Z_j} \xi_j - \eta\varphi(\eta b) \sum_{j=1}^{i} \mathbb{E}_{Z_j} \xi_j^2.$$

where ξ_j is the short version of $\xi_j(Z_1, \dots, Z_j)$. For any $i \in \{0, \dots, n\}$, we trivially have

$$\psi_{i+1} - \psi_i = \xi_{i+1} - \mathbb{E}_{Z_{i+1}} \xi_{i+1} - \eta\varphi(\eta b)\mathbb{E}_{Z_{i+1}} \xi_{i+1}^2. \tag{10.6}$$

Now for any $b \in \mathbb{R}$, $\eta > 0$ and any random variable W such that $W \le b$ a.s., we have

$$\mathbb{E} e^{\eta(W - \mathbb{E}W - \eta\varphi(\eta b)\mathbb{E}W^2)} \le 1. \tag{10.7}$$

Remark 10.2. The proof of (10.7) is standard and can be found e.g. in [3, Section 7.1.1]. We use (10.7) instead of the inequality used to prove Lemma 5 of [47], i.e. $\mathbb{E} e^{\eta[W - \mathbb{E}W - \eta\varphi(\eta b')\mathbb{E}(W - \mathbb{E}W)^2]} \le 1$ for $W - \mathbb{E}W \le b'$ since we are interested in excess risk bounds. Precisely, we will take W of the form $W = L(Z, g) - L(Z, g')$ for fixed functions g and g'. Then we have $W \le \sup_{z,g} L - \inf_{z,g} L$ while we only have $W - \mathbb{E}W \le 2\big(\sup_{z,g} L - \inf_{z,g} L\big)$. Besides the gain of having $\mathbb{E}(W - \mathbb{E}W)^2$ instead of $\mathbb{E}W^2$ is useless in the applications we develop here.

By combining (10.7) and (10.6), we obtain

$$\mathbb{E}_{Z_i} e^{\eta(\psi_{i+1} - \psi_i)} \le 1. \tag{10.8}$$

By using Markov's inequality, we upper bound the following probability w.r.t. the distribution of $Z_1, \dots, Z_{n+1}$:

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^{n+1} \xi_i > \sum_{i=1}^{n+1} \mathbb{E}_{Z_i} \xi_i + \eta\varphi(\eta b) \sum_{i=1}^{n+1} \mathbb{E}_{Z_i} \xi_i^2 + \tfrac{\log(\epsilon^{-1})}{\eta} \Big) \\
&\qquad = \mathbb{P}\big(\eta\psi_{n+1} > \log(\epsilon^{-1})\big) \\
&\qquad = \mathbb{P}\big(\epsilon e^{\eta\psi_{n+1}} > 1\big) \\
&\qquad \le \epsilon \mathbb{E} e^{\eta\psi_{n+1}} \\
&\qquad \le \epsilon \mathbb{E}_{Z_1}\big(e^{\eta(\psi_1 - \psi_0)} \mathbb{E}_{Z_2}\big(\cdots e^{\eta(\psi_n - \psi_{n-1})} \mathbb{E}_{Z_{n+1}} e^{\eta(\psi_{n+1} - \psi_n)}\big)\big) \\
&\qquad \le \epsilon
\end{aligned}$$

where the last inequality follows from recursive use of (10.8). □

10.7 Proof of Theorem 7.1

The first inequality follows from Jensen's inequality. Let us prove the second. According to Theorem 3.1, it suffices to check that condition (3.1) holds for $0 < \lambda \le \lambda_0$, $\hat{\pi}(\rho)$ the Dirac distribution at $\mathbb{E}_{g\sim\rho}\, g$ and

$$\begin{aligned}
\delta_\lambda[(x, y), g, g'] = \delta_\lambda(y) &\triangleq \min_{0 \le \zeta \le 1} \Big[\zeta\Delta(y) + \tfrac{(1-\zeta)^2 \lambda \Delta^2(y)}{2} \Big] \mathbf{1}_{|y| > B} \\
&= \tfrac{\lambda\Delta^2(y)}{2} \mathbf{1}_{\lambda\Delta(y) < 1; |y| > B} + \big[\Delta(y) - \tfrac{1}{2\lambda}\big] \mathbf{1}_{\lambda\Delta(y) \ge 1; |y| > B}.
\end{aligned}$$

- For any $z = (x, y) \in \mathcal{Z}$ such that $|y| \le B$, for any probability distribution ρ and for the above values of λ and δ_λ, by Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}_{g\sim\rho}\, e^{\lambda[L(z,\mathbb{E}_{g'\sim\rho}\, g')-L(z,g)-\delta_\lambda(z,g,g')]}\\
&\qquad\qquad\qquad = e^{\lambda L(z,\mathbb{E}_{g'\sim\rho}\, g')}\mathbb{E}_{g\sim\rho}\, e^{-\lambda\ell[y,g(x)]}\\
&\qquad\qquad\qquad \le e^{\lambda L(z,\mathbb{E}_{g'\sim\rho}\, g')}\Big(\mathbb{E}_{g\sim\rho}\, e^{-\lambda_0\ell[y,g(x)]}\Big)^{\lambda/\lambda_0}\\
&\qquad\qquad\qquad \le e^{\lambda\ell[y,\mathbb{E}_{g'\sim\rho}\, g'(x)]-\lambda\ell[y,\mathbb{E}_{g\sim\rho}\, g(x)]}\\
&\qquad\qquad\qquad = 1,
\end{aligned}$$

where the last inequality comes from the concavity of $y' \mapsto e^{-\lambda_0\ell(y,y')}$. This concavity argument goes back to [34, Section 4], and was also used in [17] and in some of the examples given in [32].

- For any $z = (x, y) \in \mathcal{Z}$ such that $|y| > B$, for any $0 \le \zeta \le 1$, by using twice Jensen's inequality and then by using the symmetrization trick presented in Section 6, we have

$$\begin{aligned}
&\mathbb{E}_{g\sim\rho}\, e^{\lambda[L(z,\mathbb{E}_{g'\sim\rho}\, g')-L(z,g)-\delta_\lambda(z,g,g')]}\\
&\quad = e^{-\delta_\lambda(y)}\mathbb{E}_{g\sim\rho}\, e^{\lambda[L(z,\mathbb{E}_{g'\sim\rho}\, g')-L(z,g)]}\\
&\quad \le e^{-\delta_\lambda(y)}\mathbb{E}_{g\sim\rho}\, e^{\lambda[\mathbb{E}_{g'\sim\rho}\, L(z,g')-L(z,g)]}\\
&\quad \le e^{-\delta_\lambda(y)}\mathbb{E}_{g\sim\rho}\, \mathbb{E}_{g'\sim\rho}\, e^{\lambda[L(z,g')-L(z,g)]}\\
&\quad = e^{-\delta_\lambda(y)}\mathbb{E}_{g\sim\rho}\, \mathbb{E}_{g'\sim\rho}\, \Big\{e^{\lambda(1-\zeta)[L(z,g')-L(z,g)]-\frac{1}{2}\lambda^2(1-\zeta)^2[L(z,g')-L(z,g)]^2}\\
&\qquad\qquad\qquad\qquad \times\, e^{\lambda\zeta[L(z,g')-L(z,g)]+\frac{1}{2}\lambda^2(1-\zeta)^2[L(z,g')-L(z,g)]^2}\Big\}\\
&\quad \le e^{-\delta_\lambda(y)}\mathbb{E}_{g\sim\rho}\, \mathbb{E}_{g'\sim\rho}\, \Big\{e^{\lambda(1-\zeta)[L(z,g')-L(z,g)]-\frac{1}{2}\lambda^2(1-\zeta)^2[L(z,g')-L(z,g)]^2}\\
&\qquad\qquad\qquad\qquad \times\, e^{\lambda\zeta\Delta(y)+\frac{1}{2}\lambda^2(1-\zeta)^2\Delta^2(y)}\Big\}\\
&\quad \le e^{-\delta_\lambda(y)}e^{\lambda\zeta\Delta(y)+\frac{1}{2}\lambda^2(1-\zeta)^2\Delta^2(y)}
\end{aligned}$$

Taking $\zeta \in [0; 1]$ minimizing the last r.h.s., we obtain that

$$\mathbb{E}_{g\sim\rho}\, e^{\lambda[L(z,\mathbb{E}_{g'\sim\rho}\, g')-L(z,g)-\delta_\lambda(z,g,g')]} \le 1$$

From the two previous computations, we obtain that for any $z \in \mathcal{Z}$,

$$\log \mathbb{E}_{g\sim\rho}\, e^{\lambda[L(z,\mathbb{E}_{g'\sim\rho}\, g')-L(z,g)-\delta_\lambda(z,g,g')]} \le 0,$$

so that condition (3.1) holds for the above values of λ, $\hat{\pi}(\rho)$ and δ_λ, and the result follows from Theorem 3.1.

10.8 Proof of Corollary 7.4

To apply Theorem 7.1, we will first determine λ_0 for which the function $\zeta : y' \mapsto e^{-\lambda_0|y-y'|^q}$ is concave. For any given $y \in [-B; B]$, for any $q > 1$, straightforward computations give

$$\zeta''(y') = \big[\lambda_0 q|y'-y|^q - (q-1)\big]\lambda_0 q|y'-y|^{q-2}e^{-\lambda_0|y-y'|^q}$$

for $y' \neq y$, hence $\zeta'' \leq 0$ on $[-b; b] - \{y\}$ for $\lambda_0 = \frac{q-1}{q(B+b)^q}$. Now since the derivative ζ' is defined at the point y, we conclude that the function ζ is concave on $[-b; b]$, so that we may use Theorem 7.1 with $\lambda_0 = \frac{q-1}{q(B+b)^q}$.

Contrary to the least square setting, we do not have a simple close formula for $\Delta(y)$, but for any $|y| \geq b$ we have

$$2bq(|y|-b)^{q-1} \leq \Delta(y) \leq 2bq(|y|+b)^{q-1}.$$

As a consequence, when $|y| \geq b + (2bq\lambda)^{-1/(q-1)}$, we have $\lambda\Delta(y) \geq 1$ and $\Delta(y) - 1/(2\lambda)$ can be upper bounded with $C'|y|^{q-1}$, where the quantity C' depends only on b and q.

For other values of $|y|$, i.e. when $b \leq |y| < b + (2bq\lambda)^{-1/(q-1)}$, we have

$$\begin{aligned}
\tfrac{\lambda\Delta^2(y)}{2}\mathbf{1}_{\lambda\Delta(y)<1;|y|>B} &+ \big[\Delta(y) - \tfrac{1}{2\lambda}\big]\mathbf{1}_{\lambda\Delta(y)\geq 1;|y|>B} \\
&= \min_{0\leq\zeta\leq 1}\Big[\zeta\Delta(y) + \tfrac{(1-\zeta)^2\lambda\Delta^2(y)}{2}\Big]\mathbf{1}_{|y|>B} \\
&\leq \tfrac{1}{2}\lambda\Delta^2(y)\mathbf{1}_{|y|>B} \\
&\leq 2\lambda b^2q^2(|y|+b)^{2q-2}\mathbf{1}_{|y|>B} \\
&\leq C''\lambda|y|^{2q-2}\mathbf{1}_{|y|>B},
\end{aligned}$$

where C'' depends only on b and q.

Therefore, from (7.2), for any $0 < b \leq B$ and $\lambda > 0$ satisfying $\lambda \leq \frac{q-1}{q(B+b)^q}$, the expected risk is upper bounded by

$$\begin{aligned}
\min_{\rho\in\mathcal{M}}\Big\{\mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big\} &+ \mathbb{E}\big\{C'|Y|^{q-1}\mathbf{1}_{|Y|\geq b+(2bq\lambda)^{-1/(q-1)};|Y|>B}\big\} \\
&+ \mathbb{E}\big\{C''\lambda|Y|^{2q-2}\mathbf{1}_{B<|Y|<b+(2bq\lambda)^{-1/(q-1)}}\big\}.
\end{aligned} \tag{10.9}$$

Let us take $B = \big(\frac{q-1}{q\lambda}\big)^{1/q} - b$ with λ small enough to ensure that $b \leq B \leq b + (2bq\lambda)^{-1/(q-1)}$. This means that λ should be taken smaller than some positive constant depending only on b and q. Then (10.9) can be written as

$$\begin{aligned}
\min_{\rho\in\mathcal{M}}\Big\{\mathbb{E}_{g\sim\rho}\, R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big\} &+ \mathbb{E}\big\{C'|Y|^{q-1}\mathbf{1}_{|Y|\geq b+(2bq\lambda)^{-1/(q-1)}}\big\} \\
&+ \mathbb{E}\big\{C''\lambda|Y|^{2q-2}\mathbf{1}_{(\frac{q-1}{q\lambda})^{1/q}-b<|Y|<b+(2bq\lambda)^{-1/(q-1)}}\big\}.
\end{aligned}$$

Now using (7.5), we can upper bound (10.9) with

$$\min_{g\in\mathcal{G}} R(g) + \frac{\log|\mathcal{G}|}{\lambda n} + C\lambda^{\frac{s+1-q}{q-1}} + C\lambda\Big(\lambda^{\frac{s-2q+2}{q}}\mathbf{1}_{s\geq 2q-2} + \lambda^{\frac{2-2q+s}{q-1}}\mathbf{1}_{s<2q-2}\Big)$$

where C depends only on b, A, q and s. So we get

$$\begin{aligned}
&\mathbb{E}_{Z_1^n}\tfrac{1}{n+1}\textstyle\sum_{i=0}^n R(\mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}} g) \\
&\qquad\qquad \leq \min_{g\in\mathcal{G}} R(g) + \frac{\log|\mathcal{G}|}{\lambda n} + C\lambda^{\frac{s+1-q}{q-1}} + C\lambda^{\frac{s-q+2}{q}}\mathbf{1}_{s\geq 2q-2} \\
&\qquad\qquad \leq \min_{g\in\mathcal{G}} R(g) + \frac{\log|\mathcal{G}|}{\lambda n} + C\lambda^{\frac{s+1-q}{q-1}}\mathbf{1}_{s<2q-2} + C\lambda^{\frac{s-q+2}{q}}\mathbf{1}_{s\geq 2q-2},
\end{aligned}$$

since $\frac{s+1-q}{q-1} \geq \frac{s-q+2}{q}$ is equivalent to $s \geq 2q-2$. By taking λ of order of the minimum of the r.h.s. (which implies that λ goes to 0 when $n/\log|\mathcal{G}|$ goes to infinity), we obtain the desired result.

10.9 Proof of Lemma 8.2

Let $\Delta_\ell = f_r'(t_\ell) - f_l'(t_\ell)$. We have

$$\begin{aligned}
&f(\alpha) - \alpha f(1) - (1-\alpha)f(0) \\
&\quad = f(\alpha) - f(0) - \alpha[f(1) - f(0)] \\
&\quad = \int_0^\alpha f'(u)du - \alpha\int_0^1 f'(u)du \\
&\quad = \int_0^\alpha \Big(\int_0^u f''(t)dt + \sum_{\ell:t_\ell\in]0;u[}\Delta_\ell\Big)du \\
&\qquad\qquad - \alpha\int_0^1 \Big(\int_0^u f''(t)dt + \sum_{\ell:t_\ell\in]0;u[}\Delta_\ell\Big)du \\
&\quad = \int_{[0;1]^2}\big(\mathbf{1}_{0<t<u<\alpha} - \alpha\mathbf{1}_{0<t<u<1}\big)f''(t)dtdu \\
&\qquad\qquad + \sum_{\ell:t_\ell\in]0;1[}[(\alpha-t_\ell)\mathbf{1}_{t_\ell<\alpha} - \alpha(1-t_\ell)]\Delta_\ell \\
&\quad = -\int_{[0;1]} K_\alpha(t)f''(t)dt - \sum_{\ell:t_\ell\in]0;1[} K_\alpha(t_\ell)\Delta_\ell
\end{aligned}$$

10.10 Proof of Theorem 8.6

The symbols $\sigma_1,\dots,\sigma_m$ still denote the coordinates of $\bar\sigma \in \{-;+\}^m$. For any $r \in \{-;0;+\}$, define $\bar\sigma_{j,r} \triangleq (\sigma_1,\dots,\sigma_{j-1},r,\sigma_{j+1},\dots,\sigma_m)$ as the vector deduced from $\bar\sigma$ by fixing its j-th coordinate to r. Since $\bar\sigma_{j,+}$ and $\bar\sigma_{j,-}$ belong to

$\{-;+\}^m$, we have already defined $P_{\bar{\sigma}_{j,+}}$ and $P_{\bar{\sigma}_{j,-}}$. We define the distribution $P_{\bar{\sigma}_{j,0}}$ as $P_{\bar{\sigma}_{j,0}}(dX) = \mu(dX)$ and

$$
\begin{aligned}
1 - P_{\bar{\sigma}_{j,0}}&(Y = h_2(X)|X) \\
&= P_{\bar{\sigma}_{j,0}}(Y = h_1(X)|X) = \begin{cases} \frac{1}{2} \text{ for any } X \in \mathcal{X}_j \\ P_{\bar{\sigma}}(Y = h_1(X)|X) \text{ otherwise} \end{cases} .
\end{aligned}
$$

The distribution $P_{\bar{\sigma}_{j,0}}$ differs from $P_{\bar{\sigma}}$ only by the conditional law of the output knowing that the input is in $\mathcal{X}_j$. We recall that $P^{\otimes n}$ denotes the n-fold product of a distribution P. For any $r \in \{-;+\}$, introduce the likelihood ratios for the data $Z_1^n = (Z_1, \dots, Z_n)$:

$$
\pi_{r,j}(Z_1^n) \triangleq \frac{P^{\otimes n}_{\bar{\sigma}_{j,r}}}{P^{\otimes n}_{\bar{\sigma}_{j,0}}}(Z_1^n) \quad = \quad \prod_{i=1}^n \big[1 + r\mathbf{1}_{X_i \in \mathcal{X}_j}(2\mathbf{1}_{Y_i = h_1(X_i)} - 1)\xi(X_i)\big].
$$

Note that this quantity is independent from the value of $\bar{\sigma}$. In the following, to shorten the notation, we will sometimes use h_1 for $h_1(X)$, h_2 for $h_2(X)$, p_+ for $p_+(X)$, p_- for $p_-(X)$. Let ν be the uniform distribution on $\{-,+\}$, i.e. $\nu\big(\{+\}\big) = 1/2 = 1 - \nu\big(\{-\}\big)$. In the following, $\mathbb{E}_{\bar{\sigma}}$ denotes the expectation when $\bar{\sigma}$ is drawn according to the m-fold product distribution of ν, and $\mathbb{E}_X = \mathbb{E}_{X\sim\mu}$. We have

$$
\begin{aligned}
&\sup_{P\in\mathcal{P}} \Big\{ \mathop{\mathbb{E}}_{Z_1^n \sim P^{\otimes n}} R(\hat{g}) - \min_g R(g) \Big\} \\
\geq\ & \sup_{\bar{\sigma}\in\{-;+\}^m} \Big\{ \mathbb{E}_{Z_1^n \sim P_{\bar{\sigma}}^{\otimes n}}\ \mathbb{E}_{Z\sim P_{\bar{\sigma}}}\ \ell[Y, \hat{g}(X)] - \min_g \mathbb{E}_{Z\sim P_{\bar{\sigma}}}\ \ell[Y, g(X)] \Big\} \\
=\ & \sup_{\bar{\sigma}\in\{-;+\}^m} \Big\{ \mathbb{E}_{Z_1^n \sim P_{\bar{\sigma}}^{\otimes n}}\ \mathbb{E}_{X\sim P_{\bar{\sigma}}(dX)} \Big[\mathbb{E}_{Y\sim P_{\bar{\sigma}}(dY|X)}\ \ell[Y, \hat{g}(X)] \\
& \qquad\qquad\qquad\qquad\qquad\qquad - \min_{y\in\mathcal{Y}}\ \mathbb{E}_{Y\sim P_{\bar{\sigma}}(dY|X)}\ \ell(Y, y) \Big] \Big\} \\
=\ & \sup_{\bar{\sigma}\in\{-;+\}^m} \Big\{ \mathbb{E}_{Z_1^n \sim P_{\bar{\sigma}}^{\otimes n}}\ \mathbb{E}_X \Big[\sum_{j=0}^m \mathbf{1}_{X\in\mathcal{X}_j} \\
& \qquad\qquad \times \Big(\varphi_{p_{\sigma_j}, h_1, h_2}[\hat{g}(X)] - \phi_{h_1,h_2}[p_{\sigma_j}] \Big) \Big] \Big\} \\
\geq\ & \mathbb{E}_{\bar{\sigma}} \mathbb{E}_{Z_1^n \sim P_{\bar{\sigma}}^{\otimes n}}\ \mathbb{E}_X \Big[\sum_{j=1}^m \mathbf{1}_{X\in\mathcal{X}_j} \\
& \qquad\qquad \times \Big(\varphi_{p_{\sigma_j}, h_1, h_2}[\hat{g}(X)] - \phi_{h_1,h_2}[p_{\sigma_j}] \Big) \Big] \\
=\ & \sum_{j=1}^m \mathbb{E}_X \Big\{ \mathbf{1}_{X\in\mathcal{X}_j} \mathbb{E}_{\bar{\sigma}} \mathbb{E}_{Z_1^n \sim P_{\bar{\sigma}_{j,0}}^{\otimes n}} \Big[\frac{P_{\bar{\sigma}}^{\otimes n}}{P_{\bar{\sigma}_{j,0}}^{\otimes n}}(Z_1^n) \\
& \qquad\qquad \times \Big(\varphi_{p_{\sigma_j}, h_1, h_2}[\hat{g}(X)] - \phi_{h_1,h_2}[p_{\sigma_j}] \Big) \Big] \Big\} \\
=\ & \sum_{j=1}^m \mathbb{E}_X \Big\{ \mathbf{1}_{X\in\mathcal{X}_j} \mathbb{E}_{\sigma_1,\dots,\sigma_{j-1},\sigma_{j+1},\dots,\sigma_m} \mathbb{E}_{Z_1^n \sim P_{\bar{\sigma}_{j,0}}^{\otimes n}} \\
& \qquad\qquad \mathbb{E}_{\sigma_j\sim\nu}\ \pi_{\sigma_j,j}(Z_1^n) \Big(\varphi_{p_{\sigma_j}, h_1, h_2}[\hat{g}(X)] - \phi_{h_1,h_2}[p_{\sigma_j}] \Big) \Big\}
\end{aligned}
\tag{10.10}
$$

The two inequalities in (10.10) are Assouad's argument ([2]). For any $x \in \mathcal{X}$, introduce

$$\tilde{\psi}_x(u) = \frac{1}{2}(u+1)\psi_{p_+(x),p_-(x),h_1(x),h_2(x)}\left(\frac{u}{u+1}\right).$$

Introduce

$$\alpha_j(Z_1^n) = \frac{\pi_{+,j}(Z_1^n)}{\pi_{+,j}(Z_1^n)+\pi_{-,j}(Z_1^n)}.$$

The last expectation in (10.10) is

$$\begin{aligned}
&\mathbb{E}_{\sigma\sim\nu}\, \pi_{\sigma,j}(Z_1^n)\Big(\varphi_{p_\sigma(X),h_1(X),h_2(X)}[\hat{g}(X)] - \phi_{h_1(X),h_2(X)}[p_\sigma(X)]\Big)\\
&\quad= \frac{1}{2}\big[\pi_{+,j}(Z_1^n)+\pi_{-,j}(Z_1^n)\big]\\
&\qquad\qquad\times\Big\{\alpha_j(Z_1^n)\varphi_{p_+,h_1,h_2}[\hat{g}(X)] + [1-\alpha_j(Z_1^n)]\varphi_{p_-,h_1,h_2}[\hat{g}(X)]\\
&\qquad\qquad\quad -\alpha_j(Z_1^n)\phi_{h_1,h_2}(p_+) - [1-\alpha_j(Z_1^n)]\phi_{h_1,h_2}(p_-)\Big\}\\
&\quad= \frac{1}{2}\big[\pi_{+,j}(Z_1^n)+\pi_{-,j}(Z_1^n)\big]\Big\{\varphi_{\alpha_j(Z_1^n)p_+ + [1-\alpha_j(Z_1^n)]p_-,h_1,h_2}[\hat{g}(X)]\\
&\qquad\qquad\quad -\alpha_j(Z_1^n)\phi_{h_1,h_2}(p_+) - [1-\alpha_j(Z_1^n)]\phi_{h_1,h_2}(p_-)\Big\}\\
&\quad\geq \frac{1}{2}\big[\pi_{+,j}(Z_1^n)+\pi_{-,j}(Z_1^n)\big]\Big\{\phi_{h_1,h_2}\Big(\alpha_j(Z_1^n)p_+ + [1-\alpha_j(Z_1^n)]p_-\Big)\\
&\qquad\qquad\quad -\alpha_j(Z_1^n)\phi_{h_1,h_2}(p_+) - [1-\alpha_j(Z_1^n)]\phi_{h_1,h_2}(p_-)\Big\}\\
&\quad= \frac{1}{2}\big[\pi_{+,j}(Z_1^n)+\pi_{-,j}(Z_1^n)\big]\psi_{p_+,p_-,h_1,h_2}[\alpha_j(Z_1^n)]\\
&\quad= \pi_{-,j}(Z_1^n)\tilde{\psi}_X\left(\frac{\pi_{+,j}(Z_1^n)}{\pi_{-,j}(Z_1^n)}\right)
\end{aligned} \tag{10.11}$$

so that

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}\Big\{\mathop{\mathbb{E}}_{Z_1^n\sim P^{\otimes n}} R(\hat{g}) - \min_g R(g)\Big\}\\
&\qquad\geq \sum_{j=1}^m \mathbb{E}_X\Big\{\mathbf{1}_{X\in\mathcal{X}_j}\mathbb{E}_{\bar{\sigma}}\mathbb{E}_{Z_1^n\sim P_{\bar{\sigma}_{j,0}}^{\otimes n}}\Big[\pi_{-,j}(Z_1^n)\tilde{\psi}_X\Big(\frac{\pi_{+,j}(Z_1^n)}{\pi_{-,j}(Z_1^n)}\Big)\Big]\Big\}\\
&\qquad= \sum_{j=1}^m \mathbb{E}_X\Big\{\mathbf{1}_{X\in\mathcal{X}_j}\mathbb{E}_{\bar{\sigma}}\mathcal{S}_{\tilde{\psi}_X}\big(P_{\bar{\sigma}_{j,+}}^{\otimes n}, P_{\bar{\sigma}_{j,-}}^{\otimes n}\big)\Big\}.
\end{aligned}$$

Now since we consider an hypercube, for any $j \in \{1,\dots,m\}$, all the terms in the sum are equal to the first one. Besides from the first part of Lemma 8.5, the last f-similarity does not depend on $\bar{\sigma}$, and in particular for $j = 1$, the f-similarity is equal to $\mathcal{S}_{\tilde{\psi}_X}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)$, where we recall that $P_{[+]}$ and $P_{[-]}$ denote the representatives of the hypercube (see Definition 8.2). Therefore we obtain

$$\begin{aligned}
\sup_{P\in\mathcal{P}}\Big\{\mathbb{E}R(\hat{g}) - \min_g R(g)\Big\} &\geq m\mathbb{E}_X\Big\{\mathbf{1}_{X\in\mathcal{X}_1}\mathcal{S}_{\tilde{\psi}_X}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)\Big\}\\
&= m\mathbb{E}_X\mathcal{S}_{\mathbf{1}_{X\in\mathcal{X}_1}\tilde{\psi}_X}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)\\
&= m\mathcal{S}_{\mathbb{E}_X(\mathbf{1}_{X\in\mathcal{X}_1}\tilde{\psi}_X)}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)\\
&= \mathcal{S}_{\tilde{\psi}}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)
\end{aligned}$$

where the second to last equality comes from the second part of Lemma 8.5.

10.11 Proof of Theorem 8.7

First, when the hypercube satisfies $p_+ \equiv 1 \equiv 1 - p_-$, from the definition of $d_{\rm I}$ given in (8.6), we have $\mathcal{S}_{\tilde\psi}\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) = mwd_{\rm I}(1-w)^n$ so that Theorem 8.6 implies (8.19).

Inequalities (8.17), (8.18) and (8.20) are deduced from Theorem 8.6 by lower bounding the $\tilde\psi$-similarity in different ways.

Since $u \mapsto u \wedge 1$ and $u \mapsto \frac{u}{u+1}$ are non-negative concave functions defined on $\mathbb{R}_+$, we may define the similarities

$$\begin{cases} \mathcal{S}_\wedge(\mathbb{P},\mathbb{Q}) \triangleq \int \big(\frac{\mathbb{P}}{\mathbb{Q}} \wedge 1\big) d\mathbb{Q} = \int \big(d\mathbb{P} \wedge d\mathbb{Q}\big) \\ \mathcal{S}_\bullet(\mathbb{P},\mathbb{Q}) \triangleq \int \frac{\mathbb{P}}{\mathbb{Q}} \frac{1}{\frac{\mathbb{P}}{\mathbb{Q}}+1} d\mathbb{Q} = \int \frac{d\mathbb{P}d\mathbb{Q}}{d\mathbb{P}+d\mathbb{Q}} \end{cases}.$$

where the second equality of both formulas introduces a formal (but intuitive) notation.

From Theorem 8.6, Lemma 8.3 and item 1 of Lemma 8.5, by using $\tilde\psi(1) = mwd_{\rm I}$, we obtain

Corollary 10.2. *Let $\mathcal{P}$ be a set of probability distributions containing an hypercube of distributions of characteristic function $\tilde\psi$ and representatives $P_{[-]}$ and $P_{[+]}$. For any estimator $\hat g$, we have*

$$\sup_{P\in\mathcal{P}} \big\{\mathbb{E}R(\hat g) - \min_g R(g)\big\} \quad\geq\quad mwd_{\rm I}\mathcal{S}_\wedge\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) \tag{10.12}$$

where the minimum is taken over the space of prediction functions. Besides if for any $x \in \mathcal{X}_1$ the function $\phi_{h_1(x),h_2(x)}$ is twice differentiable and satisfies for any $t \in [p_-(x) \wedge p_+(x); p_-(x) \vee p_+(x)]$, $-\phi''_{h_1(x),h_2(x)}(t) \geq \zeta$ for some $\zeta > 0$, then we have

$$\sup_{P\in\mathcal{P}} \big\{\mathbb{E}R(\hat g) - \min_g R(g)\big\} \geq \tfrac{mw\zeta}{4} d'_{\rm I}\, \mathcal{S}_\bullet\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big); \tag{10.13}$$

The following lemma and (10.12) imply (8.17) and (8.18).

Lemma 10.3. *We have*

$$\mathcal{S}_\wedge\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) \geq 1 - \sqrt{1-[1-d_{\rm II}]^{nw}} \geq 1 - \sqrt{nwd_{\rm II}}. \tag{10.14}$$

When the hypercube is symmetrical and constant, for N a centered gaussian random variable with variance 1, we have

$$\mathcal{S}_\wedge\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) \geq \mathbb{P}\Big(|N| > \sqrt{\tfrac{nwd_{\rm II}}{1-d_{\rm II}}}\Big) - d_{\rm II}^{1/4} \tag{10.15}$$

Proof. See Section 10.11.1. □

Remark 10.3. It is interesting to note that (10.15) is asymptotically optimal to the extent that for a $(m, w, d_{\rm II})$-hypercube (see Definition 8.3), we have

$$\Big|\mathcal{S}_\wedge\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) - \mathbb{P}\big(|N| > \sqrt{nwd_{\rm II}}\big)\Big| \underset{nw\to+\infty, d_{\rm II}\to 0}{\longrightarrow} 0, \tag{10.16}$$

[Proof in Appendix C.]

The following lemma and (10.13) imply (8.20).

Lemma 10.4. *When the hypercube is symmetrical and constant, we have*

$$\mathcal{S}_\bullet\Big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\Big) \geq \tfrac{1}{2}\Big\{1 + \tfrac{1}{2}\big[1-\big(1-\sqrt{1-d_{\rm II}}\big)w\big]^n - \tfrac{1}{2}\big[1+\big(\tfrac{1+d_{\rm II}}{\sqrt{1-d_{\rm II}}}-1\big)w\big]^n\Big\}$$

Proof. See Section 10.11.2. □

10.11.1 Proof of Lemma 10.3

For $\sigma \in \{-,+\}$, the conditional law of (X, Y) knowing $X \in \mathcal{X}_1$, when (X, Y) follows the law $P_{[\sigma]}$, is denoted $P_{\mathcal{X}_1,\sigma}$ and is called the *restricted representatives of the hypercube*. More explicitly, the probability distribution $P_{\mathcal{X}_1,\sigma}$ is such that $P_{\mathcal{X}_1,\sigma}(dX) = \mu(dX|X \in \mathcal{X}_1)$ and for any $x \in \mathcal{X}_1$

$$P_{\mathcal{X}_1,\sigma}\big(Y = h_1(x)\big|X = x\big) = p_\sigma(x) = 1 - P_{\mathcal{X}_1,\sigma}\big(Y = h_2(x)\big|X = x\big).$$

The following lemma relates the similarity between representatives of the hypercube and the similarity between restricted representatives.

Lemma 10.5. *Consider a convex function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\gamma(k) \leq \mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big)$$

for any $k \in \{0, \dots, n\}$, where by convention $\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes 0}, P_{\mathcal{X}_1,-}^{\otimes 0}\big) = 1$. For any estimator $\hat{g}$, we have

$$\mathcal{S}_\wedge\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) \geq \gamma(nw).$$

Proof. For any points $u_1, \dots, u_n$ in $\mathcal{X}$, let $\mathcal{C}(u_1, \dots, u_n)$ be the number of u_i belonging to $\mathcal{X}_1$. For any $k \in \{0, \dots, n\}$, let $B_k = \mathcal{C}^{-1}(\{k\})$ denote the subset of $\mathcal{X}^n$ for which exactly k points are in $\mathcal{X}_1$. We recall that there are $\binom{n}{k}$ possibilities of taking k elements among n and the probability of $X \in \mathcal{X}_1$ when X is drawn according to μ is $w = \mu(\mathcal{X}_1)$. We have

$$\begin{aligned}
&\mathcal{S}_\wedge\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big)\\
&\quad = \int 1 \wedge \Big(\frac{P_{[+]}^{\otimes n}}{P_{[-]}^{\otimes n}}(u_1,\dots,u_n)\Big) dP_{[-]}^{\otimes n}(u_1,\dots,u_n)\\
&\quad = \sum_{k=0}^n \int_{B_k} 1 \wedge \Big(\frac{P_{[+]}}{P_{[-]}}(u_1)\cdots\frac{P_{[+]}}{P_{[-]}}(u_n)\Big) dP_{[-]}(u_1)\cdots dP_{[-]}(u_n)\\
&\quad = \sum_{k=0}^n \binom{n}{k} \int_{(\mathcal{X}_1)^k\times(\mathcal{X}_1^c)^{n-k}} 1 \wedge \Big(\frac{P_{[+]}}{P_{[-]}}(u_1)\cdots\frac{P_{[+]}}{P_{[-]}}(u_n)\Big) dP_{[-]}^{\otimes n}(u_1,\dots,u_n)\\
&\quad = \sum_{k=0}^n \binom{n}{k} \int_{(\mathcal{X}_1)^k\times(\mathcal{X}_1^c)^{n-k}} 1 \wedge \Big(\frac{P_{[+]}}{P_{[-]}}(u_1)\cdots\frac{P_{[+]}}{P_{[-]}}(u_k)\Big) dP_{[-]}^{\otimes n}(u_1,\dots,u_n)\\
&\quad = \sum_{k=0}^n \binom{n}{k} \mu^{n-k}(\mathcal{X}_1^c) \int_{(\mathcal{X}_1)^k} \wedge \Big(\frac{P_{[+]}}{P_{[-]}}(u_1)\cdots\frac{P_{[+]}}{P_{[-]}}(u_k)\Big) dP_{[-]}^{\otimes n}(u_1,\dots,u_k)\\
&\quad = \sum_{k=0}^n \binom{n}{k} \mu^k(\mathcal{X}_1)\mu^{n-k}(\mathcal{X}_1^c)\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big)\\
&\quad \geq \sum_{k=0}^n \binom{n}{k} w^k(1-w)^{n-k}\gamma(k)\\
&\quad = \mathbb{E}\gamma(V)
\end{aligned} \tag{10.17}$$

where V is a Binomial distribution with parameters n and w. By Jensen's inequality, we have $\mathbb{E}\gamma(V) \geq \gamma[\mathbb{E}(V)] = \gamma(nw)$, which ends the proof. □

The interest of the previous lemma is to provide a lower bound on the similarity between representatives of the hypercube from a lower bound on the similarity between restricted representatives, restricted representatives being much simpler to study. The following result lower bounds the $\wedge$-similarity between the restricted representatives of the hypercube

Lemma 10.6. *For any non-negative integer k, we have*

$$\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) \geq 1 - \sqrt{1 - [1 - d_{\mathrm{II}}]^k} \geq 1 - \sqrt{k d_{\mathrm{II}}}, \tag{10.18}$$

When the hypercube is symmetrical and constant, for N a centered gaussian random variable with variance 1, *we have*

$$\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) \geq \mathbb{P}\Big(|N| > \sqrt{\tfrac{k d_{\mathrm{II}}}{1-d_{\mathrm{II}}}}\Big) - d_{\mathrm{II}}^{1/4} \tag{10.19}$$

Proof. First, we recall that P_0 denotes the base of the hypercube (see Definition 8.2). The conditional law of (X,Y) knowing $X \in \mathcal{X}_1$, when (X,Y) is drawn from P_0, is denoted $P_{\mathcal{X}_1,0}$.

For any $r \in \{-,0,+\}$, introduce $P_{r,x}$ the probability distribution on the output space such that $P_{r,x}(dY) = P_{\mathcal{X}_1,r}(dY|X=x)$. We have

$$\begin{aligned}
\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) &= \mathbb{E}_{Z_1^k\sim P_{\mathcal{X}_1,0}^{\otimes k}}\Big[\frac{P_{\mathcal{X}_1,+}^{\otimes k}}{P_{\mathcal{X}_1,0}^{\otimes k}}(Z_1^k) \wedge \frac{P_{\mathcal{X}_1,-}^{\otimes k}}{P_{\mathcal{X}_1,0}^{\otimes k}}(Z_1^k)\Big]\\
&= \mathbb{E}_{X_1^k\sim P_{\mathcal{X}_1,0}^{\otimes k}} \mathbb{E}_{Y_1^k\sim P_{\mathcal{X}_1,0}^{\otimes k}|X_1^k}\Big[\prod_{i=1}^k \frac{P_{+,X_i}}{P_{0,X_i}}(Y_i) \wedge \prod_{i=1}^k \frac{P_{-,X_i}}{P_{0,X_i}}(Y_i)\Big]\\
&= \mathbb{E}_{X_1^k\sim P_{\mathcal{X}_1,0}^{\otimes k}} \mathcal{S}_\wedge\Big(\otimes_{i=1}^k P_{+,X_i}, \otimes_{i=1}^k P_{-,X_i}\Big),
\end{aligned} \tag{10.20}$$

where $\otimes_{i=1}^k P_{r,X_i}$, $r \in \{-1; 1\}$ denotes the law of the k-tuple $(Y_1, \ldots, Y_k)$ when the Y_i are independently drawn from P_{r,X_i}.

To study divergences (or equivalently similarities) between k-fold product distributions, the standard way is to link the divergence (or similarity) of the product with the ones of base distributions. This lead to tensorization equalities or inequalities. To obtain a tensorization inequality for $\mathcal{S}_\wedge$, we introduce the similarity associated with the square root function (which is non-negative and concave):

$$\mathcal{S}_{\sqrt{}}(\mathbb{P}, \mathbb{Q}) \triangleq \int \sqrt{d\mathbb{P} d\mathbb{Q}}$$

and use the following lemmas:

Lemma 10.7. *For any probability distributions $\mathbb{P}$ and $\mathbb{Q}$, we have*

$$\mathcal{S}_\wedge(\mathbb{P}, \mathbb{Q}) \geq 1 - \sqrt{1 - \mathcal{S}_{\sqrt{}}^2(\mathbb{P}, \mathbb{Q})}.$$

Proof. Introduce the variational distance $V(\mathbb{P}, \mathbb{Q})$ as the f-divergence associated with the convex function $f : u \mapsto \frac{1}{2}|u - 1|$. From Scheffé's theorem, we have $\mathcal{S}_\wedge(\mathbb{P}, \mathbb{Q}) = 1 - V(\mathbb{P}, \mathbb{Q})$ for any distributions $\mathbb{P}$ and $\mathbb{Q}$. Introduce the Hellinger distance H, which is defined as $H(\mathbb{P}, \mathbb{Q}) \geq 0$ and $1 - \frac{H^2(\mathbb{P},\mathbb{Q})}{2} = \mathcal{S}_{\sqrt{}}(\mathbb{P}, \mathbb{Q})$ for any probability distributions $\mathbb{P}$ and $\mathbb{Q}$. The variational and Hellinger distances are known (see e.g. [40, Lemma 2.2]) to be related by

$$V(\mathbb{P}, \mathbb{Q}) \leq \sqrt{1 - \left(1 - \frac{H^2(\mathbb{P},\mathbb{Q})}{2}\right)^2},$$

hence the result. □

Lemma 10.8. *For any distributions $\mathbb{P}^{(1)}, \ldots, \mathbb{P}^{(k)}$, $\mathbb{Q}^{(1)}, \ldots, \mathbb{Q}^{(k)}$, we have*

$$\mathcal{S}_{\sqrt{}}(\mathbb{P}^{(1)} \otimes \cdots \otimes \mathbb{P}^{(k)}, \mathbb{Q}^{(1)} \otimes \cdots \otimes \mathbb{Q}^{(k)}) \\ = \mathcal{S}_{\sqrt{}}(\mathbb{P}^{(1)}, \mathbb{Q}^{(1)}) \times \cdots \times \mathcal{S}_{\sqrt{}}(\mathbb{P}^{(k)}, \mathbb{Q}^{(k)})$$

Proof. When it exists, the density of $\mathbb{P}^{(1)} \otimes \cdots \otimes \mathbb{P}^{(k)}$ w.r.t. $\mathbb{Q}^{(1)} \otimes \cdots \otimes \mathbb{Q}^{(k)}$ is the product of the densities of $\mathbb{P}^{(i)}$ w.r.t. $\mathbb{Q}^{(i)}$, $i = 1, \ldots, k$, hence the desired tensorization equality.

□

From the last two lemmas, we obtain

$$\mathcal{S}_\wedge\Big(\otimes_{i=1}^k P_{+,X_i}, \otimes_{i=1}^k P_{-,X_i}\Big) \geq 1 - \sqrt{1 - \prod_{i=1}^k S_{\sqrt{}}^2\Big(P_{+,X_i}, P_{-,X_i}\Big)} \quad (10.21)$$

From (10.20), (10.21) and Jensen's inequality, we obtain

$$
\begin{aligned}
&\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) \geq 1 - \mathbb{E}_{X_1^k \sim P_{\mathcal{X}_1,0}^{\otimes k}} \sqrt{1 - \prod_{i=1}^k S_{\sqrt{}}^2\Big(P_{+,X_i}, P_{-,X_i}\Big)} \\
&\geq 1 - \sqrt{1 - \mathbb{E}_{X_1^k \sim P_{\mathcal{X}_1,0}^{\otimes k}} \prod_{i=1}^k S_{\sqrt{}}^2\Big(P_{+,X_i}, P_{-,X_i}\Big)} \\
&= 1 - \sqrt{1 - \Big[\mathbb{E}_{X \sim P_{\mathcal{X}_1,0}} S_{\sqrt{}}^2\Big(P_{+,X}, P_{-,X}\Big)\Big]^k} \\
&= 1 - \sqrt{1 - \Big[\mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)} S_{\sqrt{}}^2\Big(P_{+,X}, P_{-,X}\Big)\Big]^k}
\end{aligned}
$$

Now we have

$$
\begin{aligned}
&\mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)} S_{\sqrt{}}^2\Big(P_{+,X}, P_{-,X}\Big) \\
&\qquad = \mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)} \big(\sqrt{p_+(X)p_-(X)} + \sqrt{[1-p_+(X)][1-p_-(X)]}\big)^2 \\
&\qquad = 1 - \mathbb{E}_{\mu(dX|X\in\mathcal{X}_1)} \big(\sqrt{p_+(X)[1-p_-(X)]} - \sqrt{[1-p_+(X)]p_-(X)}\big)^2 \\
&\qquad = 1 - d_{\mathrm{II}}
\end{aligned}
$$

So we get

$$
\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) \geq 1 - \sqrt{1 - (1 - d_{\mathrm{II}})^k} \geq 1 - \sqrt{k d_{\mathrm{II}}}, \tag{10.22}
$$

where the second inequality follows from the inequality $1 - x^k \leq k(1-x)$ that holds for any $0 \leq x \leq 1$ and $k \geq 1$. This ends the proof of (10.18).

For (10.19), since we assume that the hypercube is symmetrical and constant, we can tighten (10.22) for $k\sqrt{d_{\mathrm{II}}} \geq 1$. We have

$$
\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) = P_{\mathcal{X}_1,+}^{\otimes k}\Big(\frac{P_{\mathcal{X}_1,+}^{\otimes k}}{P_{\mathcal{X}_1,-}^{\otimes k}}(Z_1^k) \leq 1\Big) + P_{\mathcal{X}_1,-}^{\otimes k}\Big(\frac{P_{\mathcal{X}_1,+}^{\otimes k}}{P_{\mathcal{X}_1,-}^{\otimes k}}(Z_1^k) > 1\Big). \tag{10.23}
$$

Since $\frac{P_{\mathcal{X}_1,+}}{P_{\mathcal{X}_1,-}}(z) = \frac{P_{\mathcal{X}_1,+}(Y=y|X=x)}{P_{\mathcal{X}_1,-}(Y=y|X=x)} = \frac{P_{+,x}(y)}{P_{-,x}(y)}$ for any $z = (x,y) \in \mathcal{Z}$, we have

$$
\begin{aligned}
\frac{P_{\mathcal{X}_1,+}^{\otimes k}}{P_{\mathcal{X}_1,-}^{\otimes k}}(Z_1^k) &= \prod_{i=1}^k \frac{P_{+,X_i}(Y_i)}{P_{-,X_i}(Y_i)} \\
&= \prod_{i=1}^k \Big(\frac{p_+(X_i)}{p_-(X_i)}\Big)^{\mathbf{1}_{Y_i=h_1(X_i)}} \Big(\frac{1-p_+(X_i)}{1-p_-(X_i)}\Big)^{\mathbf{1}_{Y_i=h_2(X_i)}}.
\end{aligned} \tag{10.24}
$$

Using that the hypercube is symmetrical and constant, (10.24) leads to

$$
\begin{aligned}
\frac{P_{\mathcal{X}_1,+}^{\otimes k}}{P_{\mathcal{X}_1,-}^{\otimes k}}(Z_1^k) &= \prod_{i=1}^k \Big(\frac{p_+(X_i)}{1-p_+(X_i)}\Big)^{\mathbf{1}_{Y_i=h_1(X_i)} - \mathbf{1}_{Y_i=h_2(X_i)}} \\
&= \Big(\frac{p_+}{1-p_+}\Big)^{\sum_{i=1}^k [\mathbf{1}_{Y_i=h_1(X_i)} - \mathbf{1}_{Y_i=h_2(X_i)}]}.
\end{aligned} \tag{10.25}
$$

Without loss of generality, we may assume that $p_+ > 1/2$. Then we have $p_+ = 1 - p_- = \frac{1+\sqrt{d_{\mathrm{II}}}}{2}$. Introduce $W_i \triangleq \mathbf{1}_{Y_i=h_1(X_i)} - \mathbf{1}_{Y_i=h_2(X_i)}$. From (10.23) and (10.25), we obtain

$$\begin{aligned}\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) &= P_{\mathcal{X}_1,+}^{\otimes k}\big(\textstyle\sum_{i=1}^k W_i \le 0\big) + P_{\mathcal{X}_1,-}^{\otimes k}\big(\sum_{i=1}^k W_i > 0\big)\\ &= P_{\mathcal{X}_1,-}^{\otimes k}\big(\textstyle\sum_{i=1}^k W_i \ge 0\big) + P_{\mathcal{X}_1,-}^{\otimes k}\big(\sum_{i=1}^k W_i > 0\big)\end{aligned}$$

The law of $U \triangleq \sum_{i=1}^k W_i$ when $(X_1, Y_1), \dots, (X_n, Y_n)$ are independently drawn from $P_{\mathcal{X}_1,-}$ is the binomial distribution of parameter $\big(k, \frac{1-\sqrt{d_{\mathrm{II}}}}{2}\big)$. Let $\lfloor x \rfloor$ denote the largest integer k such that $k \le x$. We get

$$\begin{aligned}\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) &= \mathbb{P}(U > k/2) + \mathbb{P}(U \ge k/2)\\ &\ge 2\mathbb{P}\big(U \ge \lfloor k/2 \rfloor\big) - 2\mathbb{P}\big(U = \lfloor k/2 \rfloor\big)\end{aligned}$$

When $k\sqrt{d_{\mathrm{II}}} \ge 1$, this last r.h.s. can be lower bounded by Slud's theorem [38] for the first term and by using Stirling's formula for the second term (see e.g. [27, Appendix A.8]). It gives

$$\begin{aligned}\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) &\ge 2\mathbb{P}\Big(N \ge \frac{2\lfloor k/2 \rfloor - k(1-\sqrt{d_{\mathrm{II}}})}{\sqrt{k(1-d_{\mathrm{II}})}}\Big) - \sqrt{\frac{2}{k\pi}}\\ &\ge 2\mathbb{P}\Big(N \ge \sqrt{\frac{kd_{\mathrm{II}}}{1-d_{\mathrm{II}}}}\Big) - \sqrt{\frac{2}{\pi}}d_{\mathrm{II}}^{1/4},\end{aligned}$$

where we recall that N is a normalized gaussian random variable. Finally we have

$$\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) \ge \begin{cases} 1 - \sqrt{k}d_{\mathrm{II}}^{1/4} & \text{for any } k \ge 1\\ \mathbb{P}\Big(|N| > \sqrt{\frac{kd_{\mathrm{II}}}{1-d_{\mathrm{II}}}}\Big) - \sqrt{\frac{2}{\pi}}d_{\mathrm{II}}^{1/4} & \text{for any } k \ge \frac{1}{\sqrt{d_{\mathrm{II}}}}\end{cases}$$

which can be weakened into for any non-negative integer k

$$\mathcal{S}_\wedge\big(P_{\mathcal{X}_1,+}^{\otimes k}, P_{\mathcal{X}_1,-}^{\otimes k}\big) \ge \mathbb{P}\Big(|N| > \sqrt{\frac{kd_{\mathrm{II}}}{1-d_{\mathrm{II}}}}\Big) - d_{\mathrm{II}}^{1/4},$$

that is (10.19). □

By computing the second derivative of $u \mapsto \sqrt{1-e^{-u}}$ and $u \mapsto \int_0^{\sqrt{u}} e^{-t^2}dt$, we obtain that these functions are concave. So for any $a \in [0;1]$, the functions $x \mapsto 1 - \sqrt{1-a^x}$, $x \mapsto 1 - \sqrt{ax}$ and $x \mapsto \mathbb{P}\Big(|N| > \sqrt{\frac{xa}{1-a}}\Big) - a^{1/4}$ are convex. The convexity of these functions and Lemmas 10.5 and 10.6 imply Lemma 10.3.

10.11.2 Proof of Lemma 10.4

Let $\theta : u \mapsto u/(u+1)$ denote the non-negative concave function on which the similarity $\mathcal{S}_\bullet$ is defined. For any $u > 0$, we have

$$\begin{array}{rcl} \theta(u) & = & \frac{1}{4}\big(u+1-\frac{(u-1)^2}{u+1}\big) \\ & \geq & \frac{1}{4}\big(u+1-\frac{(u-1)^2}{2\sqrt{u}}\big) \\ & = & \frac{1}{4}\big(u+1+\sqrt{u}-\frac{u^{3/2}}{2}-\frac{1}{2\sqrt{u}}\big), \end{array}$$

hence for any probability distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$\begin{array}{rcl} \mathcal{S}_\bullet(\mathbb{P},\mathbb{Q}) & = & \int \theta\big(\frac{\mathbb{P}}{\mathbb{Q}}\big)d\mathbb{Q} \\ & \geq & \frac{1}{4}\int \big(d\mathbb{P}+d\mathbb{Q}+\sqrt{d\mathbb{P}d\mathbb{Q}}-\frac{d\mathbb{P}^{3/2}}{2d\mathbb{Q}^{1/2}}-\frac{d\mathbb{Q}^{3/2}}{2d\mathbb{P}^{1/2}}\big) \\ & = & \frac{1}{2}+\frac{1}{4}\int\sqrt{d\mathbb{P}d\mathbb{Q}}-\frac{1}{8}\int\frac{d\mathbb{P}^{3/2}}{d\mathbb{Q}^{1/2}}-\frac{1}{8}\int\frac{d\mathbb{Q}^{3/2}}{d\mathbb{P}^{1/2}} \end{array}$$

The goal of this bound is to obtain a form for which tensorization equalities hold. Precisely, let $I_1 \triangleq \int \sqrt{dP_{[+]}dP_{[-]}}$ and $I_2 \triangleq \int \frac{dP_{[+]}^{3/2}}{dP_{[-]}^{1/2}} = \int \frac{dP_{[-]}^{3/2}}{dP_{[+]}^{1/2}}$, where the last equality holds since the hypercube is symmetrical. We have

$$\mathcal{S}_\bullet\big(P_{[+]}^{\otimes n}, P_{[-]}^{\otimes n}\big) \geq \frac{1}{2}+\frac{1}{4}I_1^n-\frac{1}{4}I_2^n$$

Since the hypercube is symmetrical and constant, without loss of generality, we may assume that $p_+ \geq \frac{1}{2}$ on $\mathcal{X}_1$. Then we have $1-p_- = p_+ = (1+\sqrt{d_{\mathrm{II}}})/2$, hence $I_1 = 1-w+w\sqrt{1-d_{\mathrm{II}}}$ and

$$I_2 = 1-w+\frac{w}{2}\left(\frac{(1+\sqrt{d_{\mathrm{II}}})^{3/2}}{(1-\sqrt{d_{\mathrm{II}}})^{1/2}}+\frac{(1-\sqrt{d_{\mathrm{II}}})^{3/2}}{(1+\sqrt{d_{\mathrm{II}}})^{1/2}}\right) = 1-w+w\frac{1+d_{\mathrm{II}}}{\sqrt{1-d_{\mathrm{II}}}},$$

which gives the desired result.

10.12 Proof of Theorems 8.10 and 8.11

We consider a $(\tilde{m},\tilde{w},\tilde{d}_{\mathrm{II}})$-hypercube (see Definition 8.3 [p.27]) with

$$\tilde{m} = \lfloor \log_2 |\mathcal{G}| \rfloor,$$

$h_1 \equiv -B$ and $h_2 \equiv B$, and with $\tilde{w}$ and $\tilde{d}_{\mathrm{II}}$ to be taken in order to (almost) maximize the bound.

Case $q = 1$: From (8.27), we have $d_{\mathrm{I}} = \frac{\sqrt{d_{\mathrm{II}}}}{2}|h_2-h_1| = B\sqrt{d_{\mathrm{II}}}$ so that, choosing $\tilde{w} = 1/\tilde{m}$, (8.17) gives

$$\sup_{P\in\mathcal{H}}\big\{\mathbb{E}R(\hat{g})-\min_g R(g)\big\} \;\geq\; B\sqrt{d_{\mathrm{II}}}\big(1-\sqrt{nd_{\mathrm{II}}/\tilde{m}}\big).$$

Maximizing the lower bound w.r.t. d_{II}, we choose $d_{\mathrm{II}} = \frac{\tilde{m}}{4n}\wedge 1$ and obtain the announced result.

Case $1 < q \le 1 + \sqrt{\frac{\tilde m}{4n} \wedge 1}$ **:** From (8.11) and (8.29), for any $0 < \epsilon \le 1$, we have

$$\begin{array}{rcl} d_{\rm I} & \ge & \frac{d_{\rm II}}{2} \int_{\frac{1-\epsilon}{2}}^{\frac{1+\epsilon}{2}} [t \wedge (1-t)] \Big| \phi''_{h_1,h_2}\Big(\frac{1-\sqrt{d_{\rm II}}}{2} + \sqrt{d_{\rm II}} t\Big)\Big| dt \\ & \ge & \frac{d_{\rm II}}{2} \frac{\epsilon(2-\epsilon)}{4} \inf_{u \in \left[\frac{1-\epsilon\sqrt{d_{\rm II}}}{2};\frac{1+\epsilon\sqrt{d_{\rm II}}}{2}\right]} \big|\phi''_{h_1,h_2}(u)\big| \\ & \ge & \frac{\epsilon(2-\epsilon)}{8} d_{\rm II} \times \Big|\phi''_{-B,B}\Big(\frac{1-\epsilon\sqrt{d_{\rm II}}}{2}\Big)\Big| \\ & \ge & \frac{\epsilon(2-\epsilon)}{8} d_{\rm II} \times \frac{q}{q-1} \left[\frac{1-\epsilon^2 d_{\rm II}}{4}\right]^{\frac{2-q}{q-1}} \frac{(2B)^q}{2^{q+1}[(1+\epsilon\sqrt{d_{\rm II}})/2]^{\frac{q+1}{q-1}}} \\ & \ge & \frac{\epsilon(2-\epsilon)}{8} d_{\rm II} \times \frac{4qB^q}{q-1} (1-\epsilon\sqrt{d_{\rm II}})^{\frac{2-q}{q-1}} (1+\epsilon\sqrt{d_{\rm II}})^{\frac{1-2q}{q-1}} \\ & = & (1-\epsilon/2) q B^q \frac{\epsilon d_{\rm II}}{q-1} (1-\epsilon\sqrt{d_{\rm II}})^{\frac{2-q}{q-1}} (1+\epsilon\sqrt{d_{\rm II}})^{\frac{1-2q}{q-1}} \end{array}$$

Let $K = (1-\epsilon\sqrt{d_{\rm II}})^{\frac{2-q}{q-1}} (1+\epsilon\sqrt{d_{\rm II}})^{\frac{1-2q}{q-1}}$. From (8.17), taking $\tilde w = 1/\tilde m$, we get

$$\sup_{P \in \mathcal{H}} \big\{ \mathbb{E} R(\hat g) - \min_g R(g) \big\} \;\ge\; (1-\epsilon/2) K q B^q \frac{\epsilon d_{\rm II}}{q-1} \big(1 - \sqrt{n d_{\rm II}/\tilde m}\big). \quad (10.26)$$

This leads us to choose $d_{\rm II} = \frac{\tilde m}{4n} \wedge 1$ and $\epsilon = (q-1)\sqrt{\frac{n}{\tilde m} \vee \frac14} \le \frac12$ and obtain

$$\mathbb{E} R(\hat g) - \min_g R(g) \ge \frac{3qB^q}{8} K \Big\{ \Big(\frac14 \sqrt{\frac{\tilde m}{n}}\Big) \vee \Big(1 - \sqrt{\frac{n}{\tilde m}}\Big) \Big\}.$$

Since $1 < q \le 2$ and $\epsilon\sqrt{d_{\rm II}} = \frac{q-1}{2}$, we may check that $K \ge 0.29$ (to be compared with $\lim_{q \to 1} K = e^{-1} \approx 0.37$).

Case $q > 1 + \sqrt{\frac{\tilde m}{4n}}$ **:** We take $\tilde w = \frac{1}{n+1} \wedge \frac{1}{\tilde m}$. From (8.4), (8.6) and (8.28), we get $d_{\rm I} = \psi_{1,0,-B,B}(1/2) = \phi_{-B,B}(1/2) = B^q$. From (8.19), we obtain

$$\begin{array}{rcl} \mathbb{E} R(\hat g) - \min_{g \in \mathcal{G}} R(g) & \ge & \big(\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n+1} \wedge 1\big) B^q \big(1 - \frac{1}{n+1} \wedge \frac{1}{\lfloor \log_2 |\mathcal{G}| \rfloor}\big)^n \\ & \ge & e^{-1} B^q \big(\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n+1} \wedge 1\big), \end{array} \quad (10.27)$$

where the last inequality uses $[1 - 1/(n+1)]^n \searrow e^{-1}$.

Improvement when $1 + \sqrt{\frac{\tilde m}{4n} \wedge 1} < q < 2$ **:** From (10.26), by choosing $\epsilon = 1/2$ and introducing $K' \triangleq (1 - \sqrt{d_{\rm II}}/2)^{\frac{2-q}{q-1}} (1 + \sqrt{d_{\rm II}}/2)^{\frac{1-2q}{q-1}}$, we obtain

$$\sup_{P \in \mathcal{H}} \big\{ \mathbb{E} R(\hat g) - \min_g R(g) \big\} \;\ge\; \frac{3qB^q}{8} K' \frac{d_{\rm II}}{q-1} \big(1 - \sqrt{n d_{\rm II}/\tilde m}\big).$$

This leads us to choose $d_{\rm II} = \frac{4\tilde{m}}{9n} \wedge 1$. Since $\sqrt{\frac{\tilde{m}}{4n} \wedge 1} < q-1$, we have $\sqrt{d_{\rm II}} \le \frac{4}{3}(q-1)$, hence $K' \ge \big(1-\frac{2}{3}(q-1)\big)^{\frac{2-q}{q-1}}\big(1+\frac{2}{3}(q-1)\big)^{\frac{1-2q}{q-1}}$. For any $1<q<2$, this last quantity is greater than 0.2. So we have proved that for $1+\sqrt{\frac{\tilde{m}}{4n} \wedge 1} < q < 2$,

$$\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) \ge \frac{q}{90(q-1)} B^q \frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n}. \tag{10.28}$$

Theorem 8.11 follows from (10.27) and (10.28).

10.13 Proof of Theorem 8.12

10.13.1 Proof of the first inequality of Theorem 8.12.

Let $\tilde{m} = \lfloor \log_2 |\mathcal{G}| \rfloor$. Contrary to other lower bounds obtained in this work, this learning setting requires asymmetrical hypercubes of distributions. Here we consider a constant $\tilde{m}$-dimensional hypercube of distributions with edge probability $\tilde{w}$ such that $p_+ \equiv p$, $p_- \equiv 0$, $h_1 \equiv +B$ and $h_2 \equiv 0$, where $\tilde{w}$, p and B are positive real parameters to be chosen according to the strategy described at the beginning of Section 8.4. To have $\mathbb{E}|Y|^s \le A$, we need that $\tilde{m}\tilde{w}pB^s \le A$. To ensure that a best prediction function has infinite norm bounded by b, from the computations at the beginning of Appendix A, we need that

$$B \le \frac{p^{1/(q-1)}+(1-p)^{1/(q-1)}}{p^{1/(q-1)}}\, b.$$

This inequality is in particular satisfied for $B = Cp^{-1/(q-1)}$ for appropriate small constant C depending on b and q. From the definition of the edge discrepancy of type II, we have $d_{\rm II} = p$. In order to have the r.h.s. of (8.17) of order $mwd_{\rm I}$, we want to have $n\tilde{w}p \le C < 1$. All the previous constraints lead us to take the parameters $\tilde{w}, p$ and B such that

$$\begin{cases} B = Cp^{-1/(q-1)} \\ \tilde{m}\tilde{w}pB^s = A \\ n\tilde{w}p = 1/4 \end{cases}.$$

Let $Q = \frac{\tilde{m}}{n} \wedge 1$. This leads to $p = CQ^{(q-1)/s}$, $B = CQ^{-1/s}$ and $\tilde{w} = C\tilde{m}^{-1}Q^{1-(q-1)/s}$ with C small positive constants depending on b, A, q and s. Now from the definition of the edge discrepancy of type I and (8.5), we have

$$\begin{array}{rcl} d_{\rm I} &=& \frac{p^2}{2}\int_0^1 [t\wedge(1-t)]\big|\phi''_{0,B}(tp)\big|dt \\ &\ge& \frac{p^2}{2}\int_{1/4}^{3/4}\frac{1}{4}\min_{[p/4;3p/4]}\big|\phi''_{0,B}(tp)\big|dt \\ &\ge& Cp^2p^{\frac{2-q}{q-1}}B^q \\ &=& C \end{array}$$

where the last inequality comes from (8.29). From (8.17), we get

$$\sup_{P\in\mathcal{P}}\big\{\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g)\big\} \geq CQ^{1-\frac{q-1}{s}}.$$

10.13.2 Proof of the second inequality of Theorem 8.12.

We still use $\tilde{m} = \lfloor \log_2 |\mathcal{G}| \rfloor$. We consider a $(\tilde{m}, \tilde{w}, \tilde{d_{\mathrm{II}}})$-hypercube with $h_1 \equiv -B$ and $h_2 \equiv +B$, where $\tilde{w}, \tilde{d_{\mathrm{II}}}$ and B are positive real parameters to be chosen according to the strategy described at the beginning of Section 8.4. To have $\mathbb{E}|Y|^s \leq A$, we need that $\tilde{m}\tilde{w}B^s \leq A$. To ensure that a best prediction function has infinite norm bounded by b, from the computations at the beginning of Appendix A), we need that

$$B \leq \frac{[1+(\tilde{d_{\mathrm{II}}})^{1/2}]^{1/(q-1)}+[1-(\tilde{d_{\mathrm{II}}})^{1/2}]^{1/(q-1)}}{[1+(\tilde{d_{\mathrm{II}}})^{1/2}]^{1/(q-1)}-[1-(\tilde{d_{\mathrm{II}}})^{1/2}]^{1/(q-1)}}\, b. \tag{10.29}$$

For fixed q and b, this inequality essentially means that $B \leq C\tilde{d_{\mathrm{II}}}^{-1/2}$ since we intend to take $\tilde{d_{\mathrm{II}}}$ close to 0. In order to have the r.h.s. of (8.17) of order mwd_{I}, we want to have $n\tilde{w}\tilde{d_{\mathrm{II}}} \leq 1/4$ where, once more, this last constant is arbitrarily taken. The previous constraints lead us to choose

$$\begin{cases} B = C\tilde{d_{\mathrm{II}}}^{-1/2} \\ \tilde{m}\tilde{w}B^s = A \\ n\tilde{w}\tilde{d_{\mathrm{II}}} = 1/4 \end{cases}.$$

We still use $Q = \frac{\tilde{m}}{n} \wedge 1$. This leads to $\tilde{d_{\mathrm{II}}} = CQ^{2/(s+2)}$, $B = CQ^{-1/(s+2)}$ and $\tilde{w} = C\tilde{m}^{-1}Q^{s/(s+2)}$ with C small positive constants depending on b, A, q and s. Now from (8.29), the differentiability assumption is satisfied for $\zeta = CB^q = CQ^{-q/(s+2)}$. From (8.17) and (8.21), we obtain

$$\sup_{P\in\mathcal{P}}\big\{\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g)\big\} \geq CQ^{1-\frac{q}{s+2}}.$$

10.14 Proof of Theorem 8.14

The starting point is similar to the one in Section 10.13.2. Since $q = 2$, (10.29) simplifies into $B \leq b(\tilde{d_{\mathrm{II}}})^{-1/2}$. We take $B = b(\tilde{d_{\mathrm{II}}})^{-1/2}$ and $\tilde{w} = A/(\tilde{m}B^2)$ and we optimize the parameter $\tilde{d_{\mathrm{II}}}$ in order to maximize the lower bound. From (8.30), we get $\tilde{m}\tilde{w}d_{\mathrm{I}} = A\tilde{d_{\mathrm{II}}}$. Introducing $a \triangleq n\tilde{w}\tilde{d_{\mathrm{II}}} = \frac{nA}{\tilde{m}b^2}(\tilde{d_{\mathrm{II}}})^2$, we obtain $\tilde{m}\tilde{w}d_{\mathrm{I}} = b\sqrt{A\tilde{m}/n}\sqrt{a}$. The results then follow from Corollary 8.8 and the fact that the differentiability assumption (8.9) holds for $\zeta = 8B^2 = \frac{8d_{\mathrm{I}}}{d_{\mathrm{II}}}$.

A Computations of the second derivative of ϕ for the L_q-loss

Let y_1 and y_2 be fixed. We start with the computation of ϕ_{y_1,y_2}. For any $p \in [0;1]$, the quantity $\varphi_{p,y_1,y_2}(y) = p|y-y_1|^q + (1-p)|y-y_2|^q$ is minimized when $y \in [y_1 \wedge y_2; y_1 \vee y_2]$ and $pq(y-y_1)^{q-1} = (1-p)q(y_2-y)^{q-1}$. Introducing $r = \frac{1}{q-1}$ and $D = p^r + (1-p)^r$, the minimizer can be written as $y = \frac{p^r y_1 + (1-p)^r y_2}{D}$ and the minimum is

$$\begin{array}{rcl} \phi_{y_1,y_2}(p) & = & \Big(p\frac{(1-p)^{rq}}{D^q} + (1-p)\frac{p^{rq}}{D^q}\Big)|y_2-y_1|^q \\ & = & p(1-p)\frac{|y_2-y_1|^q}{D^{q-1}}, \end{array}$$

where we use the equality $rq = 1 + r$. We get

$$\begin{array}{rcl} \frac{1}{|y_2-y_1|^q}\phi'_{y_1,y_2} & = & \frac{1-2p}{D^{q-1}} + p(1-p)(1-q)rD^{-q}[p^{r-1} - (1-p)^{r-1}] \\ & = & D^{-q}\big\{(1-2p)[p^r + (1-p)^r] - (1-p)p^r + p(1-p)^r]\big\} \\ & = & D^{-q}\big\{(1-p)^{r+1} - p^{r+1}\big\}, \end{array}$$

hence

$$\begin{array}{rcl} \frac{1}{|y_2-y_1|^q}\phi''_{y_1,y_2} & = & -qrD^{-q-1}[p^{r-1} - (1-p)^{r-1}][(1-p)^{r+1} - p^{r+1}] \\ & & -qrD^{-q-1}[p^r - (1-p)^r]^2 \\ & = & -qrD^{-q-1}p^{r-1}(1-p)^{r-1} \\ & = & -\frac{q}{q-1}\frac{[p(1-p)]^{\frac{2-q}{q-1}}}{\Big[p^{\frac{1}{q-1}} + (1-p)^{\frac{1}{q-1}}\Big]^{q+1}}. \end{array}$$

B Expected risk bound from Hoeffding's inequality

Let $\lambda' > 0$ and ρ be a probability distribution on $\mathcal{G}$. Let $r(g)$ denote the empirical risk of a prediction function g, that is $r(g) = \frac{1}{n}\sum_{i=1}^n L(Z_i, g)$. Hoeffding's inequality applied to the random variable $W = \mathbb{E}_{g\sim\rho}\, L(Z,g) - L(Z,g') \in [-(b-a); b-a]$ for a fixed g' gives

$$\mathbb{E}_{Z\sim P}\, e^{\eta[W - \mathbb{E}W]} \le e^{\eta^2(b-a)^2/2}$$

for any $\eta > 0$. For $\eta = \lambda'/n$, this leads to

$$\mathbb{E}_{Z_1^n} e^{\lambda'[R(g') - \mathbb{E}_{g\sim\rho}\, R(g) - r(g') + \mathbb{E}_{g\sim\rho}\, r(g)]} \le e^{(\lambda')^2(b-a)^2/(2n)}$$

Consider the Gibbs distribution $\hat\rho = \pi_{-\lambda' r}$. This distribution satisfies

$$\mathbb{E}_{g'\sim\hat\rho}\, r(g') + K(\hat\rho, \pi)/\lambda' \le \mathbb{E}_{g\sim\rho}\, r(g) + K(\rho,\pi)/\lambda'.$$

We have

$$\begin{aligned}
& \mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\hat\rho}\, R(g') - \mathbb{E}_{g\sim\rho}\, R(g) \\
\le\; & \mathbb{E}_{Z_1^n}\Big\{\mathbb{E}_{g'\sim\hat\rho}\, \big[R(g') - \mathbb{E}_{g\sim\rho}\, R(g) - r(g') - \mathbb{E}_{g\sim\rho}\, r(g)\big] + \tfrac{K(\rho,\pi)-K(\hat\rho,\pi)}{\lambda'}\Big\} \\
\le\; & \tfrac{K(\rho,\pi)}{\lambda'} + \mathbb{E}_{Z_1^n}\tfrac{1}{\lambda'}\log \mathbb{E}_{g'\sim\pi}\, e^{\lambda'[R(g')-\mathbb{E}_{g\sim\rho}\, R(g)-r(g')-\mathbb{E}_{g\sim\rho}\, r(g)]} \\
\le\; & \tfrac{K(\rho,\pi)}{\lambda'} + \tfrac{1}{\lambda'}\log \mathbb{E}_{g'\sim\pi}\, \mathbb{E}_{Z_1^n} e^{\lambda'[R(g')-\mathbb{E}_{g\sim\rho}\, R(g)-r(g')-\mathbb{E}_{g\sim\rho}\, r(g)]} \\
\le\; & \tfrac{K(\rho,\pi)}{\lambda'} + \tfrac{\lambda'(b-a)^2}{2n}.
\end{aligned}$$

This proved that for any $\lambda > 0$, the generalization error of the algorithm which draws its prediction function according to the Gibbs distribution $\pi_{-\lambda\Sigma_n/2}$ satisfies

$$\mathbb{E}_{Z_1^n}\mathbb{E}_{g'\sim\pi_{-\lambda\Sigma_n/2}}\, R(g') \le \min_{\rho\in\mathcal{M}} \Big\{\mathbb{E}_{g\sim\rho}\, R(g) + 2\Big[\tfrac{\lambda(b-a)^2}{8} + \tfrac{K(\rho,\pi)}{\lambda n}\Big]\Big\},$$

where we use the change of variable $\lambda = 2\lambda'/n$ in order to underline the difference with (6.4).

C Proof of Inequality (10.16)

To prove (10.16), we need to uniformly control the difference between the tail of the sum of i.i.d. random variables and the gaussian approximate. This is done by the following result.

Theorem C.1 (Berry[12]-Esseen[30] inequality). *Let N be a centered gaussian variable of variance 1. Let $U_1,\dots,U_n$ be real-valued independent identically distributed random variables such that $\mathbb{E}U_1 = 0$, $\mathbb{E}U_1^2 = 1$ and $\mathbb{E}|U_1|^3 < +\infty$. Then*

$$\sup_{x\in\mathbb{R}} \Big|\mathbb{P}\big(n^{-1/2}\textstyle\sum_{i=1}^n U_i > x\big) - \mathbb{P}(N > x)\Big| \le Cn^{-1/2}\mathbb{E}|U_1|^3 \qquad \text{(C.1)}$$

for some universal positive constant C.

To shorten the notation, let $P^- = P_{[-]}^{\otimes n}$ and $P^+ = P_{[+]}^{\otimes n}$ be the n-fold product of the representatives of the hypercube. Since we have $p_+ > 1/2 > p_-$ (by definition of a (m, w, d_{II})-hypercube (p.27)), the set of sequences Z_1^n for which $\frac{P^+}{P^-}(Z_1^n) < 1$ is

$$E \triangleq \big\{ \textstyle\sum_{i=1}^n \mathbf{1}_{Y_i=h_1(X_i),X_i\in\mathcal{X}_1} < \sum_{i=1}^n \mathbf{1}_{Y_i=h_2(X_i),X_i\in\mathcal{X}_1}\big\}$$

Introduce the quantity

$$S_n \triangleq \textstyle\sum_{i=1}^n \big(2\mathbf{1}_{Y_i=h_1(X_i)} - 1\big)\mathbf{1}_{X_i\in\mathcal{X}_1}.$$

We have

$$\begin{aligned}\mathcal{S}_\wedge(P^+,P^-) &= 1-P^-(E)+P^+(E)\\ &= 1-P^-(S_n<0)+P^+(S_n<0)\\ &= P^-(S_n\geq 0)+P^-(S_n>0).\end{aligned} \tag{C.2}$$

Introduce

$$W_i = \big(2\mathbf{1}_{Y_i=h_1(X_i)}-1\big)\mathbf{1}_{X\in\mathcal{X}_1}.$$

From now on, we consider that the pairs $Z_i = (X_i, Y_i)$ are generated by P^-, so that $\mathbb{E}W_i$ and $\mathbb{V}\mathrm{ar}\, W_i$ simply denote the expectation and variance of W_i when (X_i, Y_i) is drawn according to $P_{-1,1,\dots,1}$. Define the normalized quantity

$$U_i \triangleq (W_i - \mathbb{E}W_i)/\sqrt{\mathbb{V}\mathrm{ar}\, W_i}.$$

We have

$$P^-(S_n>0) = P^-\big(n^{-1/2}\textstyle\sum_{i=1}^n U_i > t_n\big)$$

where $t_n \triangleq -\sqrt{\frac{n}{\mathbb{V}\mathrm{ar}\, W_1}}\mathbb{E}W_1$. By Berry-Esseen's inequality (Theorem C.1), we get

$$\big|P^-(S_n>0)-\mathbb{P}(N>t_n)\big| \leq Cn^{-1/2}\mathbb{E}|U_1|^3 \tag{C.3}$$

Let us now upper bound $n^{-1/2}\mathbb{E}|U_1|^3$.

Since we have $p_+ = (1+\xi)/2 = 1-p_-$ for a (m, w, d_{II})-hypercube (p.27)), the law of W_1 is described by

$$\begin{cases}\mathbb{P}(W_1=1)=w\frac{1-\xi}{2}\\ \mathbb{P}(W_1=0)=1-w\\ \mathbb{P}(W_1=-1)=w\frac{1+\xi}{2}\end{cases}$$

where w still denotes $\mu(\mathcal{X}_1)$. We get $\mathbb{E}W_1 = -w\xi$, $\mathbb{V}\mathrm{ar}\, W_1 = w(1-w\xi^2)$ and since $0<w\leq 1$ and $0<\xi\leq 1$

$$\begin{aligned}\mathbb{E}|W_1-\mathbb{E}W_1|^3 &= (1-w)(w\xi)^3 + w\tfrac{1-\xi}{2}(1+w\xi)^3 + w\tfrac{1+\xi}{2}(1-w\xi)^3\\ &\leq w + w[1+3(w\xi)^2]\\ &\leq 5w.\end{aligned} \tag{C.4}$$

We obtain

$$n^{-1/2}\mathbb{E}|U_1|^3 \leq n^{-1/2}\frac{5w}{[w(1-w\xi^2)]^{3/2}} \leq 5\frac{(nw)^{-1/2}}{(1-\xi^2)^{3/2}} \underset{nw\to+\infty,\xi\to 0}{\longrightarrow} 0.$$

From (C.3), we get that $\big|P^-(S_n>0)-\mathbb{P}(N>t_n)\big|$ converges to zero when nw goes to infinity and ξ goes to zero. Now the previous convergence also holds when

'$>$' is replaced with '$\geq$' (since it suffices to consider the random variables $-U_i$ in Theorem C.1). Using both convergence and (C.2), we obtain

$$\left|\mathcal{S}_\wedge(P^+, P^-) - \mathbb{P}(|N| > t_n)\right| \underset{nw\to+\infty,\xi\to 0}{\longrightarrow} 0,$$

which is the desired result since $t_n = -\sqrt{\frac{n}{\mathbb{V}\mathrm{ar}\, W_1}}\mathbb{E}W_1 = \sqrt{\frac{nw\xi^2}{1-w\xi^2}} = \sqrt{\frac{nwd_{\mathrm{II}}}{1-wd_{\mathrm{II}}}}$ and

$$\left|\mathbb{P}\left(|N| > \sqrt{\frac{nwd_{\mathrm{II}}}{1-wd_{\mathrm{II}}}}\right) - \mathbb{P}(|N| > \sqrt{nwd_{\mathrm{II}}})\right| \underset{nw\to+\infty,d_{\mathrm{II}}\to 0}{\longrightarrow} 0.$$

Acknowledgement. I would like to thank Nicolas Vayatis, Alexandre Tsybakov, Gilles Stoltz and Olivier Catoni for their helpful comments.

References

[1] P. Alquier. Iterative feature selection in least square regression estimation. Preprint n.1030, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2005, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2005.

[2] P. Assouad. Deux remarques sur l'estimation. *C.R. Acad. Sci. Paris*, 296:1021–1024, 1983.

[3] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 40(6):685–736, 2004.

[4] J.-Y. Audibert. A better variance control for PAC-Bayesian classification. Preprint n.905, http://www.proba.jussieu.fr/mathdoc/textes/PMA-905Bis.pdf, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004.

[5] J.-Y. Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint n.908, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2004, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004.

[6] J.-Y. Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004. http://cermics.enpc.fr/~audibert/ThesePack.zip.

[7] J.-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers under the margin condition, 2005. http://www.citebase.org/abstract?id=oai:arXiv.org:math/0507180.

[8] A. Barron. Are bayes rules consistent in information? In T.M. Cover and B. Gopinath, editors, *Open Problems in Communication and Computation*, pages 85–91. Springer, 1987.

[9] A. Barron and Y. Yang. Information-theoretic determination of minimax rates of convergence. *Ann. Stat.*, 27(5):1564–1599, 1999.

[10] P.L. Bartlett, O. Bousquet, and S. Mendelson. Localized rademacher complexities. In K. Kivinen, editor, *Proceedings of the 15th annual conference on Computational Learning Theory, Lecture Notes in Computer Science*, volume 2375. Springer-Verlag, 2002.

[11] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[12] A.C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. Amer. Math. Soc.*, 49:122–136, 1941.

[13] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65:181–237, 1983.

[14] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. on Information Theory*, 51(4):1611–1615, 2005.

[15] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: some recent advances. *ESAIM Probability & Statistics*, 9:323–375, 2005.

[16] J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.

[17] F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean, 2005. Technical report, Available from http://stat.fsu.edu/%7Eflori/ps/bnapril2005IEEE.pdf.

[18] O. Catoni. A mixture approach to universal model selection. preprint LMENS 97-30, Available from http://www.dma.ens.fr/edition/preprints/Index.97.html, 1997.

[19] O. Catoni. Universal aggregation rules with exact bias bound. Preprint n.510, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999, 1999.

[20] O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint n.840, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

[21] O. Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'été de Probabilités de Saint-Flour XXXI - 2001*. Lecture Notes in Mathematics. Springer Verlag, 2004.

[22] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[23] N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

[24] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Stat.*, 27(6):1865–1895, 1999.

[25] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[26] I. Csiszar. Information-type measures of difference of probability distributions and indirect observations. *Stud. Math. Hung.*, 2:299–318, 1967.

[27] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[28] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer-Verlag, 2000.

[29] R.M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6:899–929, 1978.

[30] C.-G. Esseen. On the Liapounoff limit of error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28A(9):19, 1942.

[31] D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. on Information Theory*, 44(5):1906–1925, 1998.

[32] A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. Preprint n.1034, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, http://arxiv.org/abs/math/0511468, 2006.

[33] J. Kivinen and M. K. Warmuth. Averaging expert predictions. 18 p., www.cse.ucsc.edu/~manfred/pubs/C50.pdf, 1999.

[34] J. Kivinen and M. K. Warmuth. Averaging expert predictions. *Lecture Notes in Computer Science*, 1572:153–167, 1999.

[35] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6), 2006.

[36] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse*, Math. 9(2):245–303, 2000.

[37] Merhav and Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44, 1998.

[38] E.V. Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, 5(3):404–412, 1977.

[39] A.B. Tsybakov. Optimal rates of aggregation. In B.Scholkopf and M.Warmuth, editors, *Computational Learning Theory and Kernel Machines, Lecture Notes in Artificial Intelligence*, volume 2777, pages 303–313. Springer, Heidelberg, 2003.

[40] A.B. Tsybakov. *Introduction à l'estimation non paramétrique*. Springer, 2004.

[41] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[42] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, second edition, 1995.

[43] V.G. Vovk. Aggregating strategies. In *COLT '90: Proceedings of the third annual workshop on Computational learning theory*, pages 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.

[44] V.G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, pages 153–173, 1998.

[45] Y. Yang. Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74:135–161, 2000.

[46] R. Yaroshinsky, R. El-Yaniv, and S.S. Seiden. How to better use expert advice. *Mach. Learn.*, 55(3):271–309, 2004.

[47] T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Proceedings of the 18th annual conference on Computational Learning Theory (COLT), Lecture Notes in Computer Science*, pages 173–187, 2005.

[48] T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory*, 52(4):1307–1321, 2006.