

Risk bounds for new M-estimation problems

Nabil Rachdi¹

nabil.rachdi@eads.net

Jean-Claude Fort²

jean-claude.fort@parisdescartes.fr

Thierry Klein³

thierry.klein@math.univ-toulouse.fr

Abstract

In this paper, we develop new algorithms for parameter estimation in the case of models type Input/Output in order to represent and to characterize a phenomenon Y . From experimental data Y_1, \dots, Y_n supposed to be i.i.d from Y , we prove a risk bound qualifying the proposed procedures in terms of the number of experimental data n , computing budget m and model complexity. The methods we present are general enough which should cover a wide range of applications.

1 Introduction

As in many statistical problems, we are interested to investigate the stochastic behavior of a random variable Y . We have at disposal an i.i.d sample Y_1, \dots, Y_n . These data come from experiments that could be real or the result of a computer code. In an industrial context, it is not rare that the size of the available set of data is small. This is due either to the cost of each real experiment or to the very long time needed for each run of a simulation code. It is encountered in various field of industry: meteorology, oil extraction, nuclear security, aeronautic, mechanical engineering etc...

Besides these costly experiments or codes, various reduced models are available. Even if they still are complicated, one can use them to simulate in a reasonable computing time and obtain large samples from simulations. Of course, these reduced models depend on parameters that are not well known and need to be estimated. So that the reduced models take the following form: $(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta \mapsto h(\mathbf{x}, \boldsymbol{\theta})$. It is important to note that when the model h varies, the set of input variables \mathcal{X} and parameters Θ may change too. Moreover, these variables are not directly related to the "conditions" leading to the "experimental" data Y_1, \dots, Y_n . Indeed, in our study, we don't suppose having the data $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$ which differs our framework from the classical regression one. That's why we assume that the available data reduced to Y_1, \dots, Y_n : this includes the cases where the experimental conditions are not available or where the input of the complex code, modeling the phenomenon, are not clearly related to the input of the reduced models.

Let us take an example of particular interest coming from EADS ⁴ Research department: the

¹Institut de Mathématiques de Toulouse - EADS Innovation Works, 12 rue Pasteur, 92152 Suresnes

²Université Paris Descartes, 45 rue des saints pères, 75006 Paris

³Institut de Mathématiques de Toulouse, 118 route de Narbonne F-31062 Toulouse

⁴EADS : European Aeronautic Defense and Space Company

effect of an electromagnetic field on the behavior of an aircraft. When lightning or an electromagnetic field strike an aircraft, sensors measure data corresponding to the intensity of such field in various part of the aircraft. The data recorded are dispersed due to the intrinsic variability of the phenomenon. In our framework, information of one sensor is represented by the sample Y_1, \dots, Y_n . On another side, we dispose of several computer codes h modeling the electromagnetic field in function of input variables \mathbf{z} . These input variables take the following form, $\mathbf{z} = (\mathbf{x}, \boldsymbol{\theta})$, where \mathbf{x} represents variables not well controlled and $\boldsymbol{\theta}$ a vector of parameters to be estimated, corresponding to the field properties (angles, atmospheric conditions etc...). The uncontrolled variables \mathbf{x} will be modeled by a random variable \mathbf{X} with distribution $P^{\mathbf{x}}$. This distribution may be known or not, we suppose at least having at disposal a sample $\mathbf{X}_1, \dots, \mathbf{X}_m$ where $m \gg n$.

The computer code are *complex systems*, i.e the result of interconnected disciplines providing a *granular* modeling. Actually, one disposes of a set of models \mathcal{H} covering all available models: from the simplest to the most complicated. Hence, another important issue would be to "select" a model among the set \mathcal{H} for a specific use. We don't treat this aspect in this paper, we work with one model h only.

So, shortly speaking, our goal is to construct a *Random Simulator*, $\mathbf{X} \mapsto \hat{h}(\mathbf{X}, \hat{\boldsymbol{\theta}})$ with \mathbf{X} some random variable, predicting as well as possible the observed data Y_1, \dots, Y_n . In this setting, it may be non-significant to talk about *function approximation*. For instance, suppose that $Y \sim \mathcal{U}([0, 1])$ (uniform distribution on $[0, 1]$) and consider the model $h(\mathbf{X}, \boldsymbol{\theta}) = \theta_1 + \theta_2 \mathbf{X}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\mathbf{X} \sim \mathcal{U}([0, 1])$. The cases $\boldsymbol{\theta}_1 = (0, 1)$ and $\boldsymbol{\theta}_2 = (1, -1)$, corresponding to models $h(\mathbf{x}, \boldsymbol{\theta}_1) = \mathbf{x}$ and $h(\mathbf{x}, \boldsymbol{\theta}_2) = 1 - \mathbf{x}$ respectively, produce the (same) *Random Simulator* $h(\mathbf{X}, \hat{\boldsymbol{\theta}}) \sim \mathcal{U}([0, 1])$ ($\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$). Hence, this *Random Simulator* predicts like the variable of interest Y but with two different models (the models or the parameters are not *identifiable*). Thus, the *function approximation* approach can be meaningless without preliminary precautions.

This paper is the theoretical part of a work on industrial applications in the field of "Uncertainty Management" [2]. We aim at constructing a data-dependent model which outputs are "close to" some observed data (*experimental data*). The results we present are theoretical in that the estimation and selection algorithms we propose don't include practical implementations. The same is true for the modeling aspect: we deal with (input/output) models without specifying what can be done in practice. For instance, we do not deal with the pertinence of the possible *metamodels* (see [10, 24, 17, 19]). Here, we don't talk about the impact of modeling technics, this is let for a forthcoming paper where we will apply some results obtained in this study in an industrial context.

The main tool of our development is the empirical processes theory. This theory constitutes the mathematical toolbox of asymptotics statistics and was first explored in the 1950's by the work on Functional Central Limit Theorem [4]. Along the years, the development of empirical processes theory increased successfully thanks to work of many contributors, R.M. Dudley [5], D. Pollard [16], P. Gaenssler [6], Galen R. Shorack and Jon A. Wellner [18] and others. More recently, many references give a general overview of this theory with its applications to statistics, for example [23, 21, 12]. Empirical processes give power tools for evaluating statistical estimation and inference problems. In particular, we use concentration inequalities to derive risk bounds following the work of [20], [15] and [11] among others.

Estimation based on minimizing a function was introduced by Huber in 1964 [8] where he proposed generalizing maximum likelihood estimation. The estimators resulting are called *M-estimator* ("M" for minimizing or maximizing) [9]. The class of M-estimators is a broad class because many estimation procedure can be viewed as M-estimation, maximum likelihood and least-squares estimators are some of the most important examples. Asymptotic properties of

these estimators were widely studied in a general context, and many authors like [21] or [22] used empirical processes theory which turn out to be a very valuable tool.

We present a general method where the criterion to minimize depends on both experimental and simulated data. This paper is divided into five parts. In Section 2 we describe our general framework. In Section 3 we establish Theorem 4.1 providing a risk bound for inverse problems based on both experimental and simulated data. In Section 6 we discuss about constants in Theorem 4.1.

2 General setting

2.1 The model

- *Probabilistic modeling.*

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We assume that all random variables are defined on this probability space.

Let a complex phenomenon modeled by a random real valued variable $Y \in \mathcal{Y}$, with distribution unknown Q and f the associated (Lebesgue) density function. Let assume that $\mathcal{Y} \subset [-M, M]$, $M > 0$.

Suppose that a n -sample Y_1, \dots, Y_n is available: we call *experimental data*.

Next, we suppose that this complex phenomenon can be represented by the outputs $h(\mathbf{x}, \boldsymbol{\theta})$ given by *models* h which belong to a set \mathcal{H}

$$\begin{aligned} h : \mathcal{X} \times \Theta &\longrightarrow \mathcal{Y} \\ (\mathbf{x}, \boldsymbol{\theta}) &\longmapsto h(\mathbf{x}, \boldsymbol{\theta}) \end{aligned}$$

where $\mathcal{X} \subset \mathbb{R}^d$ (*input space*), $\Theta \subset \mathbb{R}^k$ compact (*parameters space*).

We equip the input space \mathcal{X} with a probability measure $P^{\mathbf{x}}$ which forms a probability space $(\mathcal{X}, \mathcal{B}, P^{\mathbf{x}})$. The probability measure $P^{\mathbf{x}}$ is not supposed to be known, we will only dispose of a sample drawn from this distribution. In the case where $P^{\mathbf{x}}$ is known, without loss of generality, one can simply consider the uniform distribution on $[0, 1]$ provided to apply a well known probabilistic transformation.

The input vector is a random vector \mathbf{X} defined on this space, and so, the output vector $h(\mathbf{X}, \boldsymbol{\theta})$ is a random real valued variable, for each $\boldsymbol{\theta} \in \Theta$.

The space \mathcal{Y} is equipped with a σ -algebra \mathcal{E} so as to ensure the measurability of the functions

$$\begin{aligned} h(\cdot, \boldsymbol{\theta}) : (\mathcal{X}, \mathcal{B}, P^{\mathbf{x}}) &\longrightarrow (\mathcal{Y}, \mathcal{E}) \\ \mathbf{X} &\longmapsto h(\mathbf{X}, \boldsymbol{\theta}) \end{aligned}$$

Moreover, we suppose given m realizations of the input random vector \mathbf{X} ,

$$\mathbf{X}_1, \dots, \mathbf{X}_m$$

which provides m output *simulated data*

$$h(\mathbf{X}_1, \boldsymbol{\theta}), \dots, h(\mathbf{X}_m, \boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta.$$

Remark 2.1. In practice, the data $\mathbf{X}_1, \dots, \mathbf{X}_m$ may either arise from a data base (from experiments etc...) or simply arise from simulations of the random variable \mathbf{X} with known distribution $P^{\mathbf{x}}$.

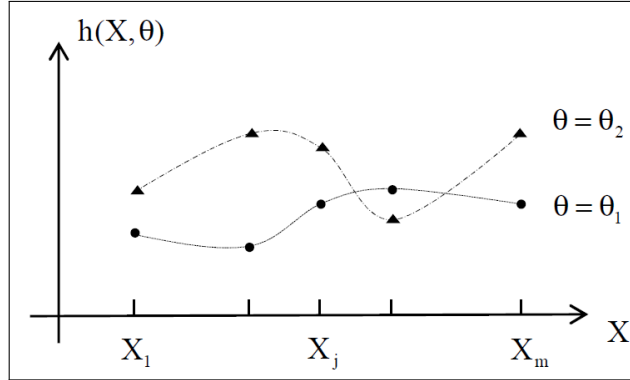


Figure 1: Example of model outputs with 2 different parameters.

In this paper, we develop a general method for estimating the parameter θ based on the *training data*

$$Y_1, \dots, Y_n; \mathbf{X}_1, \dots, \mathbf{X}_m.$$

The method we propose is general enough to include some specific problems met in practice. Indeed, two kinds of statistical analysis involving inverse problems can be considered: *Identification* and *Prediction*.

- *Identification*.

This analysis consists in estimating the "true" parameter θ^* . It aims at estimating "physical" parameters having a real signification like dimensions or material properties for instance.

- *Prediction*.

In prediction, one wants to estimate a parameter θ^* (not necessarily unique) in order to predict the random phenomenon Y . Informally, one hopes that

$$h^*(\mathbf{X}, \theta^*) \approx Y.$$

Here, the parameter θ^* may have no real signification. It is the case in model calibration for example.

2.2 Model performance

2.2.1 Tools for evaluating the model performance

Let introduce some tools to evaluate the quality of a model $h \in \mathcal{H}$ parameterized by $\theta \in \Theta$.

- *Feature of probability measure, model, contrast and Risk function*.

A *feature* of the distribution μ is defined following the Definition ?? in Chapter ??, that is, as a quantity $\rho_{\mathcal{F}}(\mu) \in \mathcal{F}$ where \mathcal{F} is called the *feature space*.

Notice that the feature space \mathcal{F} can be either a scalar space (mean, threshold probability, etc...) or a functional space (density distribution, cumulative distribution function).

We equip the feature space \mathcal{F} with the norm $\|\cdot\|_{\mathcal{F}}$ which can be either the absolute value norm $|\cdot|$ when $\mathcal{F} \subset \mathbb{R}$, or a L_r -norm ($r \geq 1$) when \mathcal{F} is a functional space (with functions define on \mathcal{Y}).

In all what follows, we denote by $\rho_h(\boldsymbol{\theta})$ a feature of the distribution of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$. In the previous chapter, we used the notation $\rho_{\mathcal{F}}(\boldsymbol{\theta})$, but especially in this chapter, we will use the writing $\rho_h(\boldsymbol{\theta})$ in order to emphasize the fact that the quantities of interest we deal with are relative to the numerical model h .

We call **model** (feature space) a subset $F \subset \mathcal{F}$. In particular, we will deal with a model induced by h given by

$$(1) \quad F_{h,\boldsymbol{\theta}} = \{\rho_h(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}.$$

Definition 2.1. Contrast and risk function.

A **contrast function** (with value in $L_1(Q)$) is any function

$$(2) \quad \begin{aligned} \Psi : \mathcal{F} &\longrightarrow L_1(Q) \\ \rho &\longmapsto \Psi(\rho, \cdot) : y \in \mathcal{Y} \longmapsto \Psi(\rho, y), \end{aligned}$$

such that

$$\rho^* = \underset{\rho \in \mathcal{F}}{\text{Argmin}} \mathbb{E}_Y \Psi(\rho, Y)$$

is *unique*.

We call **risk function** the application

$$\forall \rho \in \mathcal{F}, \quad \mathcal{R}_\Psi(\rho) := \mathbb{E}_Y \Psi(\rho, Y).$$

On the model $F_{h,\boldsymbol{\theta}} \subset \mathcal{F}$, we denote the risk by

$$(3) \quad \mathcal{R}_\Psi(h, \boldsymbol{\theta}) := \mathbb{E}_Y \Psi(\rho_h(\boldsymbol{\theta}), Y).$$

Next, for a random variable ξ , we use the notation \mathbb{E}_ξ for the expectation under the variable ξ .

Example 2.1. Some classical features and associated contrasts.

- $\mathcal{F} = \mathbb{R}$: we may consider $\rho(\mu) = \int u \mu(du) = \mathbb{E}_\mu(\xi)$ (mean), $\rho(\mu) = \int \mathbb{1}_{[s,+\infty[}(u) \mu(du) = \mu(\xi > s)$ (exceeding probability), etc...

Mean-contrast

$$\Psi(\rho, y) = (y - \rho)^2$$

- $\mathcal{F} = \{\text{set of density functions}\}$

log-contrast

$$\Psi(\rho, y) = -\log \rho(y)$$

L_2 -contrast

$$\Psi(\rho, y) = \|\rho\|_2^2 - 2\rho(y)$$

- etc...

See Table 1, page 11.

Example 2.2. Some classical risk functions.

By elementary calculus, we see that

- the mean-contrast gives a distance between means (up to a constant term)

$$\mathcal{R}_\Psi(h, \boldsymbol{\theta}) = (\mathbb{E}(Y) - \rho_h(\boldsymbol{\theta}))^2 + \text{Var}(Y)$$

- the log-contrast gives the Kullback-Leibler divergence (up to a constant term)

$$\mathcal{R}_\Psi(h, \boldsymbol{\theta}) = KL(f, \rho_h(\boldsymbol{\theta})) - \mathbb{E}(\log(Y)),$$

where $KL(g_1, g_2) = \int \log(\frac{g_1}{g_2})(y) g_1(y) dy$,

- the L_2 -contrast gives a L_2 distance between density functions (up to a constant term)

$$\mathcal{R}_\Psi(h, \boldsymbol{\theta}) = \|\rho_h(\boldsymbol{\theta}) - f\|_2^2 - \|f\|_2^2.$$

In view of that examples, it make sense to investigate models h or/and parameters $\boldsymbol{\theta}$ providing small risk values.

Let precise what we mean by *complex* models in view of statistical using.

- *Complex models.*

For $\boldsymbol{\theta} \in \Theta$, let consider a feature $\rho_h(\boldsymbol{\theta})$ of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$.

We say that h is complex if the feature $\rho_h(\boldsymbol{\theta})$ is analytically *unreachable* in $\boldsymbol{\theta}$.

For instance, if $\rho_h(\boldsymbol{\theta}) = \int_{\mathcal{X}} h(x, \boldsymbol{\theta}) P^{\mathbf{X}}(dx)$, this integral is not necessarily tractable, even if the probability measure $P^{\mathbf{X}}$ is known.

Complex models can arise from several ways. For example, the function $h(\cdot, \boldsymbol{\theta})$ can have a complicated form due to the high complexity of the modeling, or the function can be a *black box* function input/output and so, not with an analytical form.

This situation is very common in engineering, where complex models exist and are only known through simulations

$$(\mathbf{X}_1, h(\mathbf{X}_1, \boldsymbol{\theta})), \dots, (\mathbf{X}_m, h(\mathbf{X}_m, \boldsymbol{\theta})) \quad \text{for all } \boldsymbol{\theta} \in \Theta.$$

This aspect is the principal motivation of our work.

3 Inverse Problem.

Our goal is to compute a parameter $\boldsymbol{\theta} \in \Theta$ making the risk function $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$ as small as possible.

- *Oracle.*

We want to estimate a parameter $\boldsymbol{\theta}^*$ minimizing the risk (3), i.e

$$(4) \quad \boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \mathcal{R}_\Psi(h, \boldsymbol{\theta}).$$

In the literature, the parameter $\boldsymbol{\theta}^*$ is also called the *oracle*. This term was introduce by Donoho and Johnstone [3].

Notice that it may exist more than one parameter minimizing the risk $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$. The minimal risk we can reach is $\mathcal{R}_\Psi(h, \boldsymbol{\theta}^*)$, also called *ideal risk*.

However, the risk function $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$ is uncomputable (hence $\boldsymbol{\theta}^*$) for two reasons. First, the measure Q is unknown, and second, because we are dealing with complex models.

We aim at computing a parameter $\widehat{\boldsymbol{\theta}}$ that performs as well as the oracle $\boldsymbol{\theta}^*$, that is

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}) \approx \mathcal{R}_\Psi(h, \boldsymbol{\theta}^*).$$

In what follows, we establish a risk bound of the form

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}) \leq C \mathcal{R}_\Psi(h, \boldsymbol{\theta}^*) + \Delta.$$

We propose the following estimation procedure to built $\widehat{\boldsymbol{\theta}}$.

As Q is unknown, we replace it by its empirical version

$$Q_n := \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}$$

based on Y_1, \dots, Y_n . The approximation of the risk becomes

$$\frac{1}{n} \sum_{i=1}^n \Psi(\rho_h(\boldsymbol{\theta}), Y_i).$$

Then, it remains the feature $\rho_h(\boldsymbol{\theta})$ which is supposed analytically intractable (for each $\boldsymbol{\theta}$). We propose to estimate the feature as follows.

- *Plug-in estimator.*

We denote by $\rho_h^m(\boldsymbol{\theta})$ a *plug-in* estimator of $\rho_h(\boldsymbol{\theta})$ based on $h(\mathbf{X}_1, \boldsymbol{\theta}), \dots, h(\mathbf{X}_m, \boldsymbol{\theta})$. We suppose that $\rho_h^m(\boldsymbol{\theta})$ takes the following form

$$(5) \quad \rho_h^m(\boldsymbol{\theta}) := \frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))$$

where $\frac{1}{m} \tilde{\rho} : \mathcal{Y} \rightarrow \mathcal{F}$ is a *weight function* depending on the contrast Ψ considered. For simplicity, we may also call $\tilde{\rho}$ weight function.

Example 3.1. Examples of weight functions.

- *mean-contrast*

$$\frac{1}{m} \tilde{\rho}(y) = \frac{y}{m}$$

- *log-contrast or L_2 -contrast*

$$\frac{1}{m} \tilde{\rho}(y)(\cdot) = \frac{1}{m} K_b(\cdot - y)$$

where $K_b(\cdot - y) = \frac{1}{b} K(\frac{\cdot - y}{b})$ for a kernel $K(\cdot)$ and a bandwidth b .

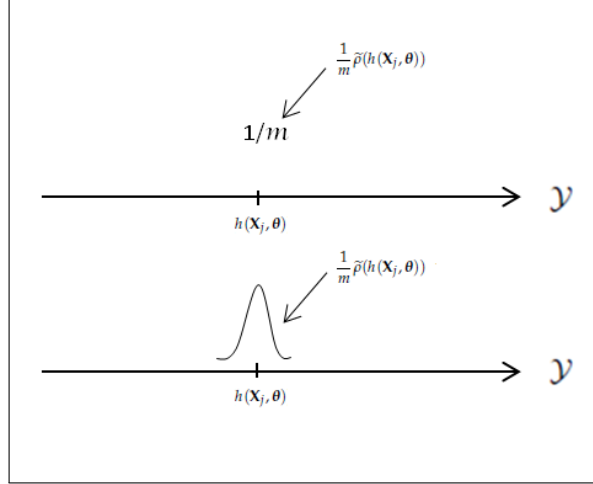


Figure 2: Example of weight function in the case of the mean (top) and the case of the density (bottom).

See Figure (2) for an illustration.

Remark 3.1. The weight function $\frac{1}{m} \tilde{\rho}(y)$ evaluated at $y \in \mathcal{Y}$ can be either a scalar value ($\frac{1}{m}$ for the mean) or a function (a kernel for the density), see Figure (2).

Without loss of generality, one can see the weight function $\frac{1}{m} \tilde{\rho}(y)$ at a point $y \in \mathcal{Y}$ as a function,

$$\tilde{\rho}(y) : \lambda \in \mathcal{Y} \mapsto \tilde{\rho}(y)(\lambda).$$

For instance, in the case where $\frac{1}{m} \tilde{\rho}(y) = \frac{y}{m}$, the function $\tilde{\rho}(y)(\lambda)$ is constant in λ .

For notation convenience, we may use the notation $\xi_{1..l}$ for a sample ξ_1, \dots, ξ_l of random variables, and $\mathbb{E}_{\xi_{1..l}}$ will be the expectation under the joint law of (ξ_1, \dots, ξ_l) .

Definition 3.1. We denote by $\sigma_h^m(\boldsymbol{\theta})$, called *simulation error*, the error committed estimating the feature $\rho_h(\boldsymbol{\theta})$ by the estimator $\rho_h^m(\boldsymbol{\theta})$,

$$\sigma_h^m(\boldsymbol{\theta}) := \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}}.$$

By triangular inequality and the fact that $\mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) = \mathbb{E}_{\mathbf{X}_{1..m}} \rho_h^m(\boldsymbol{\theta})$, it holds

$$\begin{aligned}
 \sigma_h^m(\boldsymbol{\theta}) &= \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
 &= \|\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}_{1..m}} \rho_h^m(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{X}_{1..m}} \rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
 &= \|\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) + \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
 &\leq \|\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))\|_{\mathcal{F}} + \|\mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
 (6) \quad &\leq \left\| \frac{1}{m} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))] \right\|_{\mathcal{F}} + b_h^m(\boldsymbol{\theta})
 \end{aligned}$$

with

$$(7) \quad b_h^m(\boldsymbol{\theta}) := \|\mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}}$$

the *bias error*. For example, in the case where $\tilde{\rho}(y)(\cdot) = K_b(\cdot - y)$, the bandwidth will depend on m (b_m).

The first term in the right hand side of inequality (6) is a *variance* (random) term, and the second is a *bias* (deterministic) term.

For our statistical analysis, the variability term

$$\left\| \frac{1}{m} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))] \right\|_{\mathcal{F}}$$

will play a crucial role.

Assumption 3.1. We assume that the plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ (5) is uniformly asymptotically unbiased, i.e it exists some constant $b_h(m)$ depending on h and m such that the bias error (7) satisfies

$$\sup_{\boldsymbol{\theta} \in \Theta} b_h^m(\boldsymbol{\theta}) < b_h(m) < \infty,$$

and $b_h(m) \rightarrow 0$ with m .

For instance, in the case where $\rho_h(\boldsymbol{\theta})$ is the density (of $h(\mathbf{X}, \boldsymbol{\theta})$), the quantity $\mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) = \mathbb{E}_{\mathbf{X}} K_b(\cdot - h(\mathbf{X}, \boldsymbol{\theta}))$ may be choose to be the expectation of a kernel K_b with bandwidth b . Under right conditions on the set of densities $\{\rho_h(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, one can hope that

$$\sup_{\boldsymbol{\theta} \in \Theta} b_h^m(\boldsymbol{\theta}) \xrightarrow{b \rightarrow 0} 0$$

See the example given in Subsection 6.2.

Finally, the criterion we propose to minimize has the form

$$\frac{1}{n} \sum_{i=1}^n \Psi \left(\frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})), Y_i \right),$$

which provides the estimator

$$(8) \quad \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^n \Psi \left(\frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})), Y_i \right),$$

or

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \sum_{i=1}^n \Psi \left(\frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})), Y_i \right).$$

We give some examples of estimators $\hat{\boldsymbol{\theta}}$.

Example 3.2. Examples of estimators.

- *mean-contrast*

$$\hat{\boldsymbol{\theta}}_M = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \sum_{i=1}^n \left(\sum_{j=1}^m (Y_i - h(\mathbf{X}_j, \boldsymbol{\theta})) \right)^2$$

- log-contrast

$$\hat{\boldsymbol{\theta}}_{\log} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{Argmin}} - \sum_{i=1}^n \log \left(\sum_{j=1}^m K_b(Y_i - h(\mathbf{X}_j, \boldsymbol{\theta})) \right)$$

- L_2 -contrast

$$\hat{\boldsymbol{\theta}}_{L_2} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{Argmin}} \left\{ \left\| \sum_{j=1}^m K_b(\cdot - h(\mathbf{X}_j, \boldsymbol{\theta})) \right\|_2^2 - \frac{2m}{n} \sum_{i=1}^n \sum_{j=1}^m K_b(Y_i - h(\mathbf{X}_j, \boldsymbol{\theta})) \right\}.$$

Remark 3.2. 1. The estimator $\hat{\boldsymbol{\theta}}$ depends on the model h , the number of experimental data n and the number of simulation data m .
2. The number of simulations m have to be thought greater than n (number of experimental data). It appears natural to think that experimental data are difficult to obtain whereas simulated data are more reachable.

We recall that the issue is the statistical properties of this procedure taking into account the two kinds of data: experimental and simulated data, which is non classical in statistics. Indeed, once we define the procedure for computing $\hat{\boldsymbol{\theta}}$, we have to qualify the *quality* of this procedure. It's the topic of the following section.

4 Main Result

In this section, we aim at establishing a risk bound which provides a qualification of the estimation procedure previously defined.

We recall that

$$\begin{aligned} \mathcal{R}_{\Psi}(h, \boldsymbol{\theta}) &= \mathbb{E}_Y \Psi(\rho_h(\boldsymbol{\theta}), Y), \\ \boldsymbol{\theta}^* &\in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{Argmin}} \mathcal{R}_{\Psi}(h, \boldsymbol{\theta}), \end{aligned}$$

and

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{Argmin}} \frac{1}{n} \sum_{i=1}^n \Psi \left(\frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})), Y_i \right).$$

Now, we give some definitions and notations useful for setting the Theorem 4.1.

Denote by

$$\mathbb{G}_n = \sqrt{n}(Q_n - Q)$$

and

$$\mathbb{K}_m^{\mathbf{x}} = \sqrt{m}(P_m^{\mathbf{x}} - P^{\mathbf{x}}),$$

the Q -empirical process (based on Y_1, \dots, Y_n) and $P^{\mathbf{x}}$ -empirical process (based on $\mathbf{X}_1, \dots, \mathbf{X}_m$), respectively.

Let the classes of functions

$$(9) \quad \mathcal{W}_{(\tilde{\rho}, \Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\tilde{\rho}(\lambda), y), \lambda \in \mathcal{Y}\},$$

$$(10) \quad \mathcal{P}_{(\tilde{\rho}, h)} = \{\mathbf{x} \in \mathcal{X} \mapsto \tilde{\rho}(h(\mathbf{x}, \boldsymbol{\theta}))(\lambda), (\boldsymbol{\theta}, \lambda) \in \Theta \times \mathcal{Y}\}.$$

	$\mathcal{W}_{(\bar{\rho}, \Psi)}$	$\mathcal{P}_{(\bar{\rho}, h)}$	A_Ψ
mean-contrast	$y \mapsto (y - \lambda)^2,$ $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta}),$ $\boldsymbol{\theta} \in \Theta$	$4M$
log-contrast	$y \mapsto -\log(K_b(y - \lambda)),$ $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto K_b(\lambda - h(\mathbf{x}, \boldsymbol{\theta})),$ $(\lambda, \boldsymbol{\theta}) \in \Theta \times \mathcal{Y}$	$\ f\ _2/\eta$
L_2 -contrast	$y \mapsto \ K_b(\cdot - \lambda)\ _2 - 2K_b(y - \lambda),$ $\lambda \in \mathcal{Y}$	<i>idem</i>	$2(\ f\ _2 + B)$

Table 1: Example of classes of functions and constant A_Ψ (see section (6.1)).

Next, we use the following notation: let P be some measure and \mathcal{G} a class of real valued functions. We denote by

$$Pg := \int g(u)P(du) \quad g \in \mathcal{G}$$

and

$$\|P\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} |Pg|.$$

With this notation, for a class of functions $\mathcal{G}_{\mathcal{Y}}, : \mathcal{Y} \rightarrow \mathbb{R}$ we have

$$\begin{aligned} \mathbb{G}_n g &= \int_{\mathcal{Y}} g(u) \mathbb{G}_n(du) \\ &= \sqrt{n} \int_{\mathcal{Y}} g(u) (Q_n - Q)(du) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(Y_i) - \mathbb{E}(g(Y))). \end{aligned}$$

Also, for a class of functions $\mathcal{G}_{\mathcal{X}}, : \mathcal{X} \rightarrow \mathbb{R}$

$$\mathbb{K}_m^{\mathbf{x}} g = \frac{1}{\sqrt{m}} \sum_{j=1}^m (g(\mathbf{X}_j) - \mathbb{E}(g(\mathbf{X}))).$$

Remark 4.1. The quantities $\|\mathbb{G}_n\|_{\mathcal{G}_{\mathcal{Y}}}$ and $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{G}_{\mathcal{X}}}$ are nonnegative real valued random variables.

In our applications, the class of functions $\mathcal{G}_{\mathcal{Y}}$ is $\mathcal{W}_{(\bar{\rho}, \Psi)}$ and $\mathcal{G}_{\mathcal{X}}$ is $\mathcal{P}_{(\bar{\rho}, h)}$, respectively defined in (9) and (10).

Definition 4.1. Tightness.

Let $(\xi_l)_{l \geq 1}$ be a sequence of real value random variables defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

This sequence is tight if for all $\varepsilon > 0$, it exists some compact $\mathcal{K}^\varepsilon \subset \mathbb{R}$ such that

$$\forall l \geq 1, \quad \mathbb{P}(\xi_l \in \mathcal{K}^\varepsilon) \geq 1 - \varepsilon.$$

In particular, if the ξ_l are nonnegative, the sequence is tight if for all $\varepsilon > 0$ it exists some constant $\bar{K}^\varepsilon \geq 0$ such that

$$\forall l \geq 1, \quad \mathbb{P}(\xi_l \leq \bar{K}^\varepsilon) \geq 1 - \varepsilon.$$

We make the following assumption.

Assumption 4.1. We assume that the contrast Ψ satisfies

- for all $y \in \mathcal{Y}$, the function $\rho \mapsto \Psi(\rho, y)$ is convex ,
- for all $y \in \mathcal{Y}$ and $\rho_1, \rho_2 \in \mathcal{F}$

$$|\Psi(\rho_1, y) - \Psi(\rho_2, y)| \leq L_\Psi(y) \|\rho_1 - \rho_2\|_{\mathcal{F}}$$

with $L_\Psi : \mathcal{Y} \rightarrow \mathbb{R}$ satisfying $A_\Psi := \mathbb{E}_Y L_\Psi(Y) < \infty$.

The function L_Ψ (hence the constant A_Ψ) doesn't depend on ρ_1 and ρ_2 .

The contrasts given in Example 2.1 satisfy this assumption under right conditions on the distribution of Y .

Theorem 4.1. Risk bound for Parameter Estimation.

Under the Assumptions (4.1) and (3.1), suppose that the sequences of random variables $\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}}$ and $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}}$ are tight. Denote by $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$ the associated constants, uniform (or decreasing) in n and m , respectively.

Let the feature space \mathcal{F} equipped with either the absolute value norm, or some L_r norm. Then, for all $\varepsilon > 0$, with probability at least $1 - 2\varepsilon$ it holds

$$\mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{K_{(\tilde{\rho}, \Psi)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m) \right)$$

where the constants $K_{(\tilde{\rho}, \Psi)}^\varepsilon, K_{(\tilde{\rho}, h)}^\varepsilon$ depend on $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon, \bar{K}_{(\tilde{\rho}, h)}^\varepsilon, A_\Psi, M$ and r . B_m is a bias factor depending on $b_h(m)$.

5 Some comments

It is of interest to compare the methodology we develop with the classical framework where the feature $\rho_h(\boldsymbol{\theta})$ of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$ is analytically tractable. In this case, the estimation procedure (8) is classically

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^n \Psi(\rho_h(\boldsymbol{\theta}), Y_i),$$

and we can derive immediately a risk bound.

Proposition 5.1. Basic risk bound.

It holds that

$$(11) \quad \mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \|\mathbb{G}_n\|_{\tilde{\mathcal{W}}_\Psi},$$

where

$$\tilde{\mathcal{W}}_\Psi = \{y \in \mathcal{Y} \mapsto \Psi(\rho_h(\boldsymbol{\theta}), y), \boldsymbol{\theta} \in \Theta\}.$$

Proof. The proof comes from a classical calculus in M-estimation, see for example [22] (p. 46) \square

Most of statistical procedures, as likelihood, regression, classification etc... can be written like (11). Such procedures have been widely studied with a large literature available. Recently, authors use the Empirical Processes theory (see [21, 22, 23, 12] among others) to derive limit theorems. Indeed, the asymptotic (and non-asymptotic) properties of the estimator $\hat{\boldsymbol{\theta}}_n$ can be given from the behavior of the residual term $\frac{2}{\sqrt{n}}\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi}$. In particular, for *identification* problem (i.e $\boldsymbol{\theta}^*$ is unique), consistency and rate of convergence are derived from the fluctuations of the random variable $\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi}$, see for example [21].

Suppose for a moment that it exists some constant (uniform in n) such that with high probability

$$\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi} \leq \frac{K}{2},$$

then by inequality (11), with high probability

$$(12) \quad \mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{K}{\sqrt{n}}.$$

Thus, depending on whether the constant K is sharp or not, one can bound properly the estimation error. To compute such (sharp) constant K is difficult in general, we can refer to [13, 20, 23, 14].

Inequality (11) can not be applied to our framework because the induced procedure $\hat{\boldsymbol{\theta}}_n$ involves the quantity $\rho_h(\boldsymbol{\theta})$ intractable for *complex models*.

The result of Theorem 4.1 is non-asymptotic, i.e valid for all $n \geq 1$ and $m \geq 1$ under mentioned assumptions. The fundamental point of this theorem is the "*concentration of the measure phenomenon*" (Ledoux [13], Billingsley [1]) presents in the assumptions, more precisely, when we supposed the tightness of the sequences of the random variables $\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}}$ ($Y_{1..n}$ -dependent) and $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}}$ ($\mathbf{X}_{1..m}$ -dependent). Moreover, we insist on the fact that the constants $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ (that bounds $\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}}$) and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$ (that bounds $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}}$) are uniform (or decreasing) in n and m , respectively. The advantage of this uniformity is the explicit expression of the *residual* term

$$(13) \quad \frac{K_{(\tilde{\rho}, \Psi)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m) \right)$$

depending on the data (n and m) on one hand, and on the constants $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$, $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$ and B_m on the other hand. However, although the existence of such constants are proved or supposed, their computation is more tedious. Indeed, we need results about tail bounds for Gaussian and Empirical Processes. We will discuss in Section 6.3 how to compute properly such constants using concentration inequalities. Let assume for a moment the existence of these constants.

We showed that the estimation procedure $\hat{\boldsymbol{\theta}}$ defined in (8) "mimic" the ideal risk $\mathcal{R}_\Psi(h, \boldsymbol{\theta}^*) = \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta}))$ up to the residual term (13). Making $m \rightarrow +\infty$, this residual becomes simply $\frac{K_{(\tilde{\rho}, \Psi)}^\varepsilon}{\sqrt{n}}$ which has the same form as those found in classical cases (12). We find the usual rate of convergence \sqrt{n} .

In our purpose, the factor

$$\left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m) \right) > 1$$

we call *simulation factor*, is due to simulation used estimating the feature $\rho_h(\boldsymbol{\theta})$ of the random output $h(\mathbf{X}, \boldsymbol{\theta})$ by a plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ we defined in (5).

Example 5.1. For unbiased plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ we have $B_m = 0$, so the simulation factor is simply

$$\left(1 + \sqrt{\frac{n}{m}} K_{(\tilde{\rho}, h)}^\varepsilon\right).$$

It appears that for fixed n , one should have a number of simulation data m greater than n . For instance, for some $\beta > 1$, if we have

$$m = n^\beta \quad \text{or} \quad n (\log(n))^\beta,$$

we can make the simulation factor close to 1.

Remark 5.1. The term $\inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta}))$ in Theorem 4.1 appears as the best (smaller) error one can make. This kind of error is commonly called *approximation error* or *systematic error*. It can be understood as the "distance" between the *a priori* knowledge one has with the observed phenomenon.

By Examples (2.2) and (3.2), we can write the risk bound in Theorem 4.1 in specific cases as follows.

Example 5.2. Risk bounds in specific cases.

- mean-contrast

$$\left(\mathbb{E}(Y) - \rho_h(\hat{\boldsymbol{\theta}}_M)\right)^2 \leq \inf_{\boldsymbol{\theta} \in \Theta} \left(\mathbb{E}(Y) - \rho_h(\boldsymbol{\theta})\right)^2 + \frac{K_{(\tilde{\rho}, M)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m)\right)$$

In practice, $B_m = 0$.

- log-contrast

$$KL(\rho_h(\hat{\boldsymbol{\theta}}_{\log}), f) \leq \inf_{\boldsymbol{\theta} \in \Theta} (KL(\rho_h(\boldsymbol{\theta}), f)) + \frac{K_{(\tilde{\rho}, \log)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m)\right)$$

- L_2 -contrast

$$\|\rho_h(\hat{\boldsymbol{\theta}}_{L_2}) - f\|_2^2 \leq \inf_{\boldsymbol{\theta} \in \Theta} (\|\rho_h(\boldsymbol{\theta}) - f\|_2^2) + \frac{K_{(\tilde{\rho}, L_2)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m)\right).$$

The terms $\inf_{\boldsymbol{\theta} \in \Theta} \left(\mathbb{E}(Y) - \rho_h(\boldsymbol{\theta})\right)^2$, $\inf_{\boldsymbol{\theta} \in \Theta} (KL(\rho_h(\boldsymbol{\theta}), f))$ and $\inf_{\boldsymbol{\theta} \in \Theta} (\|\rho_h(\boldsymbol{\theta}) - f\|_2^2)$ are the *ideal risks* $\inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta}))$ in different situations. These examples show clearly that these terms represent a "distance" between the "target" and the "best" information available, see Remark 5.1. If these terms are supposed equal to zero, it means that we believe that for instance the density f belongs to the family of densities $\{\rho_h(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$. In this case we obtain for example (L_2 -contrast)

$$\|\rho_h(\hat{\boldsymbol{\theta}}_{L_2}) - f\|_2^2 \leq \frac{K_{(\tilde{\rho}, L_2)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m)\right).$$

However, such *a priori* has to be made with precautions.

6 About the constants in Theorem 4.1

6.1 Constant A_Ψ

We will show how we obtain the constants A_Ψ in Table (1). Let recall that $\mathcal{Y} \in [-M, M]$.

- *mean-contrast.*

Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F} \subset \mathcal{Y}$. We have

$$\begin{aligned} |(y - \rho_1)^2 - (y - \rho_2)^2| &= |\rho_1 - \rho_2| |2y - (\rho_1 + \rho_2)| \\ &\leq |\rho_1 - \rho_2| 4M. \end{aligned}$$

- *log-contrast.*

Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F}$, with \mathcal{F} some set of density functions.

Moreover, suppose that it exists some $\eta > 0$ such that

$$\forall \rho \in \mathcal{F} \quad \rho > \eta$$

By Taylor Lagrange formula, it exists some $\tau \in (\rho_1(y), \rho_2(y))$ such that

$$\begin{aligned} |\log(\rho_1(y)) - \log(\rho_2(y))| &= \frac{1}{\tau} |\rho_1(y) - \rho_2(y)| \\ &\leq \frac{1}{\eta} |\rho_1(y) - \rho_2(y)| \end{aligned}$$

since $\rho > \eta$ for all $\rho \in \mathcal{F}$ and $\tau > \eta$.

Taking the expectation under the measure Q (with Lebesgue density f) involves the quantity $\mathbb{E}_Y(|\rho_1(Y) - \rho_2(Y)|)$ in the right member. By Cauchy-Schwarz inequality

$$\mathbb{E}_Y(|\rho_1(Y) - \rho_2(Y)|) \leq \|\rho_1 - \rho_2\|_2 \|f\|_2,$$

so

$$\mathbb{E}_Y |\log(\rho_1(Y)) - \log(\rho_2(Y))| \leq \frac{\|f\|_2}{\eta} \|\rho_1 - \rho_2\|_2.$$

- *L_2 -contrast.*

Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F}$, with \mathcal{F} some set of density functions.

Suppose that it exists some $B > 0$ such that

$$\sup_{\rho \in \mathcal{F}} \|\rho\|_2 < B.$$

By triangular inequality

$$\begin{aligned} |(\|\rho_1\|_2^2 - 2\rho_1(y)) - (\|\rho_2\|_2^2 - 2\rho_2(y))| &\leq | \|\rho_1\|_2^2 - \|\rho_2\|_2^2 | + 2|\rho_2(y) - \rho_1(y)| \\ &\leq \|\rho_1 - \rho_2\|_2^2 + 2|\rho_2(y) - \rho_1(y)|. \end{aligned}$$

Taking the expectation under Q and by Cauchy-Schwarz inequality (as before) yields

$$\begin{aligned} \mathbb{E}_Y |(\|\rho_1\|_2^2 - 2\rho_1(Y)) - (\|\rho_2\|_2^2 - 2\rho_2(Y))| &\leq \|\rho_1 - \rho_2\|_2^2 + 2\|\rho_1 - \rho_2\|_2 \|f\|_2 \\ &\leq \|\rho_1 - \rho_2\|_2 (\|\rho_1 - \rho_2\|_2 + 2\|f\|_2) \\ &\leq 2(B + \|f\|_2) \|\rho_1 - \rho_2\|_2 \end{aligned}$$

6.2 Constant $b_h(m)$

When the *plug-in* estimator $\rho_h^m(\boldsymbol{\theta})$ is unbiased, the bias term $b_h^m(\boldsymbol{\theta})$ defined in (7) is zero for all $\boldsymbol{\theta} \in \Theta$ and $m > 0$, hence $b_h(m) = 0$ too.

We study the example of the kernel estimator (biased), i.e when the weight function $\tilde{\rho}$ is a function of the form

$$\tilde{\rho}(y)(\cdot) = K_b(\cdot - y)$$

where $K_b(\cdot - y) = \frac{1}{b}K\left(\frac{\cdot - y}{b}\right)$ for some kernel $K(\cdot)$ and some bandwidth b .

Consider that $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_2$, then for all $\boldsymbol{\theta} \in \Theta$ we have

$$\begin{aligned} b_h^m(\boldsymbol{\theta}) &= \|\mathbb{E}_{\mathbf{X}}(K_b(\cdot - h(\mathbf{X}, \boldsymbol{\theta}))) - \rho_h(\boldsymbol{\theta})\|_2 \\ &= \left(\int_{\mathcal{Y}} \left(\int_{\mathcal{X}} (K_b(y - h(x, \boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})) P^{\mathbf{X}}(dx) \right)^2 dy \right)^{1/2}. \end{aligned}$$

Theorem (24.1) in [22] (p. 345) gives the following result.

Theorem 6.1. *Let $\xi_1, \dots, \xi_m \in \mathcal{Y}$ an i.i.d sample drawn from a probability density function g and $K : \mathcal{Y} \rightarrow \mathbb{R}^+$ some function (kernel). Denote by*

$$\hat{g}(y) = \frac{1}{m} \sum_{j=1}^m \frac{1}{b} K\left(\frac{y - \xi_m}{b}\right).$$

If the following assumptions are valid

- $\|g''\|_2 < +\infty$
- $\int y K(y) dy = 0$
- $I = \int y^2 K(y) dy < +\infty$,

then it exists a constant C_g such that for all $b > 0$

$$\mathbb{E}_{\xi_{1..m}} \|\hat{g} - g\|_2^2 \leq C_g \left(\frac{1}{mb} + b^4 \right).$$

In particular, the bias term $\|\mathbb{E}_{\xi_{1..m}} \hat{g} - g\|_2$ is bounded above by

$$\frac{I \|g''\|_2}{\sqrt{3}} b^2.$$

In our context, take $g = \rho_h(\boldsymbol{\theta})$ and suppose that the assumptions of this Theorem are satisfied, then

$$b_h^m(\boldsymbol{\theta}) \leq \frac{I \|\rho_h''(\boldsymbol{\theta})\|_2}{\sqrt{3}} b^2.$$

Moreover, if $\sup_{\boldsymbol{\theta} \in \Theta} \|\rho_h''(\boldsymbol{\theta})\|_2$ is finite, it justifies the existence of $b_h(m) = \sup_{\boldsymbol{\theta} \in \Theta} b_h^m(\boldsymbol{\theta})$.

6.3 Constants $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$

We detail the arguments for computing the constants $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$. Since these constants are tightness constants relative to some empirical processes (see the assumptions of Theorem 4.1), we will give arguments with a generic empirical process $\mathbb{W}_p = \sqrt{p}(W_p - W)$ indexed by a generic class of functions \mathcal{G} .

Now, the goal is to compute some constant $K(\varepsilon)$ such that

$$(14) \quad \mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \leq K(\varepsilon)) \geq 1 - \varepsilon \quad \text{for small } \varepsilon > 0.$$

For this, we propose to use the work of T. Klein and E. Rio [11], in particular Theorem 1.1, that deal with right hand side deviations of the empirical process. They show that for an empirical process \mathbb{W}_p indexed by a **countable** class of functions \mathcal{G} with values in $[-1, 1]$

$$(15) \quad \mathbb{P}\left(\sup_{g \in \mathcal{G}} \mathbb{W}_p(g) \geq \mathbb{E}(\sup_{g \in \mathcal{G}} \mathbb{W}_p(g)) + t\right) \leq \exp\left(-\frac{t^2}{2v + 3x/\sqrt{p}}\right),$$

for all positive t and some constant v . They also give left hand side deviations.

In our purpose, we don't really work with $\sup_{g \in \mathcal{G}} \mathbb{W}_p(g)$ but rather with $\sup_{g \in \mathcal{G}} |\mathbb{W}_p(g)| = \|\mathbb{W}_p\|_{\mathcal{G}}$ corresponding to a two-side control. Hence, according to the work of T. Klein and E. Rio [11], it exists some function $\varphi_{\mathcal{G}} : \mathbb{R}_+ \rightarrow [0, 1]$ decreasing to zero such that for all positive t

$$(16) \quad \mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + t) \leq \varphi_{\mathcal{G}}(t).$$

Another point is missing before we apply this result in our context, it is the fact that the result is valid for countable classes of functions, and so, we need to extend the Theorem 1.1 in [11]. We prove the following proposition.

Proposition 6.1. *Let \mathbb{W}_p be an empirical process indexed by a class of functions \mathcal{G} taking values in $[-1, 1]$ and parameterized by a **compact** set \mathcal{C} of \mathbb{R}^l , $l \geq 1$. Suppose that the application*

$$(17) \quad \lambda \in \mathcal{C} \mapsto g_\lambda \in \mathcal{G} \subset L_2$$

is continuous.

Then, it exists a function $\varphi_{\mathcal{G}}$ decreasing to zero (given by [11]) such that for all $t \geq 0$

$$(18) \quad \mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + t) \leq \varphi_{\mathcal{G}}(t).$$

Proof. For simplicity, we prove the proposition with $\mathcal{G} = \mathcal{W}_{(\tilde{\rho}, \Psi)}$ where

$$\mathcal{W}_{(\tilde{\rho}, \Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\tilde{\rho}(\lambda), y), \lambda \in \mathcal{Y}\}$$

(in fact we consider $\mathbb{W}_p = \mathbb{G}_p$) and take $\mathcal{Y} = [-M, M]$. Moreover, without loss of generality, suppose that the functions in $\mathcal{W}_{(\tilde{\rho}, \Psi)}$ take values in $[-1, 1]$.

We define the sets $\mathcal{Y}^s = \{y_1^s, \dots, y_{i_s}^s\}$ for $s \geq 1$ recursively, as follows:

- $\mathcal{Y}^1 = \{-M, 0, M\}$.

- Assume that the set $\mathcal{Y}^s = \{y_1^s, \dots, y_{i_s}^s\}$ is construct with increasing elements, i.e $y_1^s < \dots < y_{i_s}^s$.
For $j = 1, \dots, i_s - 1$, let

$$\tilde{y}_j^s = \frac{y_j^s + y_{j+1}^s}{2}$$

and

$$\tilde{\mathcal{Y}}^s = \{\tilde{y}_j^s, i = 1, \dots, i_{s-1} - 1\}.$$

- Define

$$\mathcal{Y}^{s+1} = \mathcal{Y}^s \cup \tilde{\mathcal{Y}}^s$$

with increasing elements.

Remark 6.1. One can verify that

$$\text{Card}(\mathcal{Y}^s) = 2^s + 1.$$

Now, define the classes of functions

$$\mathcal{W}_{(\tilde{\rho}, \Psi)}^s = \{y \in \mathcal{Y} \mapsto \Psi(\tilde{\rho}(\lambda), y), \lambda \in \mathcal{Y}_s\}$$

and notice that for all $s \geq 1$,

$$(19) \quad \mathcal{W}_{(\tilde{\rho}, \Psi)}^{s-1} \subsetneq \mathcal{W}_{(\tilde{\rho}, \Psi)}^s \subsetneq \mathcal{W}_{(\tilde{\rho}, \Psi)}.$$

By this previous display and the fact that $\bigcup_{s \geq 1} \mathcal{Y}^s$ is dense in $[-M, M]$ and by the continuous assumption (17), we have

$$(20) \quad \overline{\lim_{s \rightarrow \infty} \mathcal{W}_{(\tilde{\rho}, \Psi)}^s} = \overline{\bigcup_{s \geq 1} \mathcal{W}_{(\tilde{\rho}, \Psi)}^s} = \mathcal{W}_{(\tilde{\rho}, \Psi)}.$$

The classes of functions $\mathcal{W}_{(\tilde{\rho}, \Psi)}^s$, $s \geq 1$ are countable ($2^s + 1$ elements) with values in $[-1, 1]$. Finally, we apply the inequality (16) to the classes $\mathcal{W}_{(\tilde{\rho}, \Psi)}^s$, we get for all $t \geq 0$ and $s \geq 1$

$$(21) \quad \mathbb{P} \left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s}) + t \right) \leq \varphi_s(t).$$

We wish to prove that the left and right member of this last inequality converge when $s \rightarrow \infty$. Write the left member as follows

$$(22) \quad \begin{aligned} & \mathbb{P} \left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s}) + t \right) \\ &= \mathbb{E} \left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s}) + t} \right) \\ &= \mathbb{E} \left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s} - \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s}) \geq t} \right). \end{aligned}$$

The inclusions (19) yields

$$\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^{s-1}} \leq \|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}^s} \leq \|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}} \quad \forall s \geq 1,$$

so the sequence $\left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}^s}\right)_{s \geq 1}$ is increasing and bounded, thus it converges. By monotone convergence, we obtain that the sequence $\left(\mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}^s}\right)\right)_{s \geq 1}$ converges too provided that $\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}}) < \infty$. Thus, the sequence $\left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}^s} - \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}^s}\right)\right)_{s \geq 1}$ converges too, and by dominated convergence the quantity (22) converges to the wanted limit

$$\mathbb{E}\left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}} - \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}}) \geq t}\right) = \mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\bar{\rho}, \Psi)}}) + t\right).$$

For the right member of (21), by similar arguments, it can be shown that $\varphi_s(t) \rightarrow \varphi(t) = \varphi_{\mathcal{G}}(t)$. That concludes the proof. \square

Next, since the function $t \mapsto \varphi_{\mathcal{G}}(t)$ is decreasing from \mathbb{R}_+ into $[0, 1]$, then it exists a unique function $\kappa_{\mathcal{G}} : [0, 1] \rightarrow \mathbb{R}_+$ such that

$$(23) \quad \forall t \geq 0 \quad \kappa_{\mathcal{G}}^{-1}(t) = \varphi_{\mathcal{G}}(t).$$

Then, we can write (18) as follows, for all $\varepsilon \in]0, 1[$

$$\mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon)) \leq \varepsilon$$

or equivalently

$$\mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \leq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon)) \geq 1 - \varepsilon.$$

Thus, for a constant $K(\varepsilon)$ that should satisfy (14), i.e

$$\mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \leq K(\varepsilon)) \geq 1 - \varepsilon,$$

one can take $K(\varepsilon)$ equal to

$$\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon).$$

But, the quantity $\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}})$ remains not tractable. We propose to bound it.

Indeed, *maximal inequalities* allow to bound such quantities in terms of *entropy integrals* we will define. Although these methods are known to be not sharp, the bounds we will obtain are of interest for our purpose. Before, let recall some useful definitions.

Let \mathcal{G} be a class of functions and W some probability measure.

An *envelope function* of the class \mathcal{G} is a function $G : y \mapsto G(y)$ such that $|g(y)| \leq G(y)$, for all y and $g \in \mathcal{G}$.

Denote by

$$\|g\|_{2,W} = \left(\int g^2(y) W(dy)\right)^{1/2}.$$

The three following definitions are from [23] (p. 83-85).

Definition 6.1. $L_2(W)$ **Covering numbers and Entropy.**

The covering number $N(\varepsilon, \mathcal{G}, L_2(W))$ is the minimal number of balls $\{j, \|j - g\|_{2,W} < \varepsilon\}$ of radius ε needed to cover the class \mathcal{G} . The centers of the balls need not belong to \mathcal{G} , but they should have finite norm. The entropy is the logarithm of the covering number.

Definition 6.2. $L_2(W)$ **Bracketing numbers and Entropy with bracketing.**

Given two functions l, u , the bracket $[l, u]$ is the set of all functions g with $l \leq g \leq u$. An ε -bracket is a bracket $[l, u]$ with $\|u - l\|_{2,W} < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{G}, L_2(W))$ is the minimum number of ε -brackets needed to cover the class of functions \mathcal{G} .

The entropy with bracketing is the logarithm of the bracketing number.

The bracketing numbers measure the "size", the complexity of a class of functions. We also dispose of a definition providing at which "speed" the classes grow.

Definition 6.3. $L_2(W)$ **Bracketing integral.**

The bracketing integral is defined as

$$J_{[\cdot]}(\delta, \mathcal{G}, L_2(W)) := \int_0^\delta \sqrt{\log N_{[\cdot]}(\epsilon, \mathcal{G}, L_2(W))} d\epsilon.$$

Now we apply Corollary 19.35 of [22] (p. 288), it holds that

$$(24) \quad \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) \leq a_{\mathcal{G}} J_{[\cdot]}(\|G\|_{2,W}, \mathcal{G}, L_2(W)),$$

where

- $a_{\mathcal{G}}$ is some universal constant
- G is an envelop function of \mathcal{G} and

$$\|G\|_{2,W} = \left(\int G^2 W(dy) \right)^{1/2}.$$

Remark 6.2. The quantity $J_{[\cdot]}(\|G\|_{2,W}, \mathcal{G}, L_2(W))$ is computable if one has the bracketing numbers $N_{[\cdot]}(\epsilon, \mathcal{G}, L_2(W))$ ($\forall \epsilon > 0$), see examples in Section 7 below.

Finally, setting

$$(25) \quad K(\varepsilon) = a_{\mathcal{G}} J_{[\cdot]}(\|G\|_{2,Q}, \mathcal{G}_{(\tilde{\rho}, \Psi)}, L_2(W)) + \kappa_{\mathcal{G}}(\varepsilon)$$

provides the claimed constant. In particular, we should take $\mathcal{G} = \mathcal{W}_{(\tilde{\rho}, \Psi)}$ ($W = Q$) and $\mathcal{G} = \mathcal{P}_{(\tilde{\rho}, h)}$ ($W = P^x$) in order to compute $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$, respectively.

7 Constants $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$ in particular cases

7.1 $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ for the Mean-contrast

Recall that in this case

$$\mathcal{W}_{(\tilde{\rho}, \Psi)} = \{y \mapsto (y - \lambda)^2, \lambda \in \mathcal{Y}\}.$$

This class is uniformly bounded by $4M^2$, we take the envelop function $G = 4M^2$. Then, we have

$$|(y - \lambda_1)^2 - (y - \lambda_2)^2| \leq |\lambda_1 - \lambda_2| F(y),$$

with $F(y) = |2y + 2M|$, and by Theorem (2.7.11) in [23] (p. 164) it holds that

$$N_{[\cdot]}(\epsilon, \mathcal{W}_{(\tilde{\rho}, \Psi)}, L_2(Q)) \leq N\left(\frac{\epsilon}{2\|F\|_{2,Q}}, \mathcal{Y}, |\cdot|\right)$$

Notice that $\|F\|_{2,Q} \leq 4M$. Since $\mathcal{Y} \subset [-M, M]$, we have

$$N\left(\frac{\epsilon}{2\|F\|_{2,Q}}, \mathcal{Y}, |\cdot|\right) \leq N\left(\frac{\epsilon}{8M}, [-M, M], |\cdot|\right).$$

The covering number in the right member is bounded by $16 M^2/\epsilon$, so that we finally get

$$N_{[]}(\epsilon, \mathcal{W}_{(\tilde{\rho}, \Psi)}, L_2(Q)) \leq \frac{16 M^2}{\epsilon}.$$

Now, we compute the bracketing integral

$$\begin{aligned} J_{[]}(\|G\|_{2,Q}, \mathcal{W}_{(\tilde{\rho}, \Psi)}, L_2(Q)) &= \int_0^{\|G\|_{2,Q}} \sqrt{\log(N_{[]}(\epsilon, \mathcal{W}_{(\tilde{\rho}, \Psi)}, L_2(Q)))} d\epsilon \\ &\leq \int_0^{4M^2} \sqrt{\log\left(\frac{16 M^2}{\epsilon}\right)} d\epsilon, \end{aligned}$$

and with the variable substitution $u = 2 \log(16 M^2/\epsilon)$, this integral becomes

$$4\sqrt{2} M^2 \int_{\log(16)}^{+\infty} \sqrt{u} e^{-u/2} du.$$

Moreover, since $\int_0^{+\infty} \sqrt{u} e^{-u/2} du = \sqrt{2\pi}$, the bracketing integral is bounded by

$$J_{[]}(\|G\|_{2,Q}, \mathcal{W}_{(\tilde{\rho}, \Psi)}, L_2(Q)) \leq 8\sqrt{\pi} M^2.$$

Finally, we obtain the following constant

$$(26) \quad \bar{K}_{(\tilde{\rho}, \Psi)}^\epsilon = 8 a_1 \sqrt{\pi} M^2 + \kappa_1(\epsilon).$$

7.2 $\bar{K}_{(\tilde{\rho}, h)}^\epsilon$ with the weight function $\tilde{\rho}(y) = y$

In this case, the class of functions $\mathcal{P}_{(\tilde{\rho}, h)}$ is

$$\mathcal{P}_{(\tilde{\rho}, h)} = \{\mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\} \quad (\mathcal{X} \subset \mathbb{R}^d).$$

We assumed in the introduction that the models $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$ are uniformly bounded by M , thus denote by P an envelop of $\mathcal{P}_{(\tilde{\rho}, h)}$, take $P = M$.

Moreover, let suppose that the models $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$ belong to the Hölder space $\mathbb{H}(\mathcal{X}, \alpha, L)$ ($\alpha, L > 0$) defined as

$$\mathbb{H}(\mathcal{X}, \alpha, L) = \{g : \mathcal{X} \rightarrow \mathbb{R} \text{ continuous}, \|g\|_\alpha \leq L\}$$

where

$$\|g\|_\alpha = \max_{|\nu| \leq [\alpha]} \sup_{x \in \mathcal{X}} |D^\nu g(x)| + \max_{\nu: |\nu| = [\alpha]} \sup_{x, x' \in \mathcal{X}} \frac{|D^\nu g(x) - D^\nu g(x')|}{\|x - x'\|^{\alpha - [\alpha]}}$$

with $[\alpha]$ the largest integer smaller than α , and the differential operator D^ν is defined as, for $\nu = (\nu_1, \dots, \nu_d) \in \mathbb{N}^d$

$$D^\nu = \frac{\partial^{|\nu|}}{\partial \nu_1^{\nu_1} \dots \partial \nu_d^{\nu_d}}, \quad \text{and} \quad |\nu| = \sum_{i=1}^d \nu_i.$$

We aim at computing the entropy integral $J_{[]}(\|P\|_{2,Q}, \mathcal{P}_{(\tilde{\rho}, h)}, L_2(Q))$ by integrating the entropy $\log N_{[]}(\epsilon, \mathcal{P}_{(\tilde{\rho}, h)}, L_2(Q))$.

Corollary 2.7.2 in [23] (p. 157) gives an entropy bound for the Hölder space $\mathbb{H}(\mathcal{X}, \alpha, 1)$:

$$(27) \quad \log N_{[]}(\epsilon, \mathbb{H}(\mathcal{X}, \alpha, 1), L_2(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{d/\alpha} \quad \forall \epsilon > 0,$$

where K depends on α , $\text{diam}(\mathcal{X})$ and d .

We supposed that $\mathcal{P}_{(\tilde{\rho}, h)} \subset \mathbb{H}(\mathcal{X}, \alpha, L)$, and one can easily check that

$$(28) \quad \mathbb{H}(\mathcal{X}, \alpha, L) = L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1).$$

where $L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1) = \{Lg : g \in \mathbb{H}(\mathcal{X}, \alpha, 1)\}$.

Remark 7.1. If $\mathcal{P}_{(\tilde{\rho}, h)} \subset \mathbb{H}(\mathcal{X}, \alpha, L)$, then necessarily $L \geq M$. It comes from the fact that $\|g\|_\alpha \geq \|g\|_\infty$ for all $\alpha > 0$.

Next, we will use the following lemma.

Lemma 7.1.

$$\begin{aligned} N_{[\cdot]}(\epsilon, \mathbb{H}(\mathcal{X}, \alpha, L), L_2(Q)) &= N_{[\cdot]}(\epsilon, L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1), L_2(Q)) \\ &= N_{[\cdot]}(\epsilon/L, \mathbb{H}(\mathcal{X}, \alpha, 1), L_2(Q)). \end{aligned}$$

Proof. The first equality is clear by (28). Let $([l_i, u_i])_{i=1 \dots N}$ be a set of ϵ -brackets covering $\mathbb{H}(\mathcal{X}, \alpha, 1)$. Then the brackets $([Ll_i, Lu_i])_{i=1 \dots N}$ cover $L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1)$ since for $g \in \mathbb{H}(\mathcal{X}, \alpha, 1)$

$$l \leq g \leq u \implies Ll \leq Lg \leq Lu.$$

Finally, the brackets $[Ll_i, Lu_i]$ are of size $L\epsilon$, and the result follows. \square

Using (27), Lemma 7.1 and the inequality

$$J_{[\cdot]}(\|P\|_{2,Q}, \mathcal{P}_{(\tilde{\rho}, h)}, L_2(Q)) \leq J_{[\cdot]}(\|P\|_{2,Q}, \mathbb{H}(\mathcal{X}, \alpha, L), L_2(Q)),$$

it holds for $d < 2\alpha$

$$J_{[\cdot]}(\|P\|_{2,Q}, \mathcal{P}_{(\tilde{\rho}, h)}, L_2(Q)) \leq \sqrt{K} \int_0^M \left(\frac{L}{\epsilon}\right)^{d/2\alpha} d\epsilon,$$

hence

$$J_{[\cdot]}(\|P\|_{2,Q}, \mathcal{P}_{(\tilde{\rho}, h)}, L_2(Q)) \leq M \sqrt{K} \left(\frac{L}{M}\right)^{d/2\alpha} \frac{1}{1 - d/2\alpha}.$$

Finally, under the condition $d < 2\alpha$, we get the constant

$$\bar{K}_{(\tilde{\rho}, h)}^\epsilon = a_2 M \sqrt{K} \left(\frac{L}{M}\right)^{d/2\alpha} \frac{1}{1 - d/2\alpha} + \kappa_2(\epsilon).$$

Remark 7.2. The condition $d < 2\alpha$ above, means that the dimension of the random input \mathbf{X} (equal to d) is limited by the "smoothness" of the models $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. The smoother the models are (i.e α large), the larger the dimension d can be.

Remark 7.3. The computation of the constants $\bar{K}_{(\tilde{\rho}, \Psi)}^\epsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\epsilon$ are difficult enough to obtain, as we saw. However, we adopt a nonasymptotic point of view and so such computations are crucial in order to give sense to the risk bounds.

8 Proofs

In order to prove the risk bound of Theorem (4.1), we need the following lemmas.

8.1 Preliminary lemmas

Lemma 8.1. *Consider the random functions*

$$y \mapsto \Psi(\tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})), y), \quad \boldsymbol{\theta} \in \Theta.$$

We have (a.s.)

$$\sup_{\boldsymbol{\theta} \in \Theta} |\mathbb{G}_n(\Psi(\tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))))| \leq \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}},$$

where $\mathcal{W}_{(\tilde{\rho}, \Psi)}$ is defined in (9).

Proof. The key ingredient is *re-parametrization*.

Since for all $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$, $h(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{Y}$, conditionally to $\mathbf{X} = \mathbf{x}_0$

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta} |\mathbb{G}_n(\Psi(\tilde{\rho}(h(\mathbf{x}_0, \boldsymbol{\theta}))))| &\leq \sup_{\lambda \in \mathcal{Y}} |\mathbb{G}_n(\Psi(\tilde{\rho}(\lambda)))| \\ &= \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}}. \end{aligned}$$

The right member does not depend on \mathbf{x}_0 , and the result follows. \square

Remark 8.1. The left member of the inequality in the lemma (8.1) depends on the model h , contrary to the right member. Indeed, this last term depends only on the weight function with the associated contrast, and on n .

Lemma 8.2. *Consider the $P^{\mathbf{x}}$ -empirical process $\mathbb{K}_m^{\mathbf{x}}$ and let $\|\cdot\|_{\mathcal{F}} = |\cdot|$ or $\|\cdot\|_r$ and define*

$$c = \begin{cases} 1 & \text{if } \tilde{\rho}(y) \text{ is constant, } \forall y \in \mathcal{Y}, \\ (2M)^{1/r} & \text{else} \end{cases}.$$

We have

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta}))\|_{\mathcal{F}} \leq c \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}},$$

where $\mathcal{P}_{(\tilde{\rho}, h)}$ is defined in (10).

Proof. Let notice that the quantity

$$\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta})) = \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))]$$

can be (up to a factor) either a sum of independent random real variables or a sum of independent random functions.

- If $\tilde{\rho}(y) \in \mathbb{R}$ for all $y \in \mathcal{Y}$ (we have a sum of random variables).

Taking $\|\cdot\|_{\mathcal{F}} = |\cdot|$ the absolute value norm, it comes directly that

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta} \|\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta}))\|_{\mathcal{F}} &= \sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))] \right| \\ &= \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} \end{aligned}$$

Remark 8.2. In this case, $\tilde{\rho}(y)(\lambda) = \tilde{\rho}(y)$ for all y and λ in \mathcal{Y} .
- If, for all $y \in \mathcal{Y}$, $\tilde{\rho}(y)$ is a real valued function defined on \mathcal{Y} .

Take $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_r$, $r \geq 1$, the L_r norm. By integration properties and the fact that

$$\sup_{z \geq 0} z^r = (\sup_{z \geq 0} z)^r,$$

we have

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta} \|\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta}))\|_r &= \sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))] \right\|_r \\ &= \sup_{\boldsymbol{\theta} \in \Theta} \left(\int_{\mathcal{Y}} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))(\lambda))] \right|^r d\lambda \right)^{1/r} \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta} \left(\int_{\mathcal{Y}} \left(\sup_{\lambda \in \mathcal{Y}} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))(\lambda))] \right| \right)^r d\lambda \right)^{1/r} \\ &= \sup_{\boldsymbol{\theta} \in \Theta} \sup_{\lambda \in \mathcal{Y}} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))(\lambda))] \right| \left(\int_{\mathcal{Y}} d\lambda \right)^{1/r} \\ &= (2M)^{1/r} \sup_{(\boldsymbol{\theta}, \lambda) \in \Theta \times \mathcal{Y}} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))(\lambda))] \right|. \end{aligned}$$

Finally, notice that

$$\sup_{(\boldsymbol{\theta}, y) \in \Theta \times \mathcal{Y}} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))(y) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))(y)] \right| = \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}(\tilde{\rho}, h)}$$

and the result follows. \square

Remark 8.3. In the case where the weight function is a kernel $K_b(\cdot - \cdot)$, the quantity

$$\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta})) = \frac{1}{\sqrt{m}} \sum_{j=1}^m [K_b(\cdot - h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} K_b(\cdot - h(\mathbf{X}, \boldsymbol{\theta}))]$$

is treated as a sum of independent random functions in the recent work of A. Goldenshluger and O. Lepski [7]. Here we have made the restrictive assumption that $\mathcal{Y} \subset [-M, M]$. A valuable challenge would be to extend our results to the unbounded case using [7].

8.2 Proof of Theorem (4.1)

Proof. We denote by

$$- M(h, \boldsymbol{\theta}) = \mathcal{R}_{\Psi}(h, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{Y}} \Psi(\rho_h(\boldsymbol{\theta}), Y)$$

$$- M_n(h, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \Psi(\rho_h(\boldsymbol{\theta}), Y_i)$$

- $M_{,m}(h, \boldsymbol{\theta}) = \mathbb{E}_Y \Psi(\rho_h^m(\boldsymbol{\theta}), Y)$
- $M_{n,m}(h, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \Psi(\rho_h^m(\boldsymbol{\theta}), Y_i)$
- $\mathbb{G}_n \Psi(\rho_h^m(\boldsymbol{\theta})) = \sqrt{n} (M_{n,m}(h, \boldsymbol{\theta}) - M_{,m}(h, \boldsymbol{\theta}))$

where $\rho_h^m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))$ and recall that

$$(29) \quad \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} M_{n,m}(h, \boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} M(h, \boldsymbol{\theta}).$$

We have,

$$\begin{aligned} & \mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \\ = & M(h, \hat{\boldsymbol{\theta}}) - M_m(h, \hat{\boldsymbol{\theta}}) + M_m(h, \hat{\boldsymbol{\theta}}) - M_{n,m}(h, \hat{\boldsymbol{\theta}}) + M_{n,m}(h, \hat{\boldsymbol{\theta}}) \\ = & - \left(M_m(h, \hat{\boldsymbol{\theta}}) - M(h, \hat{\boldsymbol{\theta}}) \right) - \frac{1}{\sqrt{n}} \mathbb{G}_n \Psi \left(\rho_h^m(\hat{\boldsymbol{\theta}}_{n,m}) \right) + \underbrace{M_{n,m}(h, \hat{\boldsymbol{\theta}}) - M_{n,m}(h, \boldsymbol{\theta}^*)}_{\leq 0 (29)} + M_{n,m}(h, \boldsymbol{\theta}^*) \\ \leq & - \left(M_m(h, \hat{\boldsymbol{\theta}}) - M(h, \hat{\boldsymbol{\theta}}) \right) - \frac{1}{\sqrt{n}} \mathbb{G}_n \Psi \left(\rho_h^m(\hat{\boldsymbol{\theta}}) \right) + M_{n,m}(h, \boldsymbol{\theta}^*) - M_m(h, \boldsymbol{\theta}^*) + M_m(h, \boldsymbol{\theta}^*) \\ \leq & - \left(M_m(h, \hat{\boldsymbol{\theta}}) - M(h, \hat{\boldsymbol{\theta}}) \right) - \frac{1}{\sqrt{n}} \mathbb{G}_n \Psi \left(\rho_h^m(\hat{\boldsymbol{\theta}}) \right) + \frac{1}{\sqrt{n}} \mathbb{G}_n \Psi \left(\rho_h^m(\boldsymbol{\theta}^*) \right) + M_m(h, \boldsymbol{\theta}^*) \\ \leq & - \left(M_m(h, \hat{\boldsymbol{\theta}}) - M(h, \hat{\boldsymbol{\theta}}) \right) + \frac{1}{\sqrt{n}} \mathbb{G}_n \left(\Psi \left(\rho_h^m(\boldsymbol{\theta}^*) \right) - \Psi \left(\rho_h^m(\hat{\boldsymbol{\theta}}) \right) \right) \\ & + M_m(h, \boldsymbol{\theta}^*) - M(h, \boldsymbol{\theta}^*) + M(h, \boldsymbol{\theta}^*) \\ \leq & \frac{1}{\sqrt{n}} \mathbb{G}_n \left(\Psi \left(\rho_h^m(\boldsymbol{\theta}^*) \right) - \Psi \left(\rho_h^m(\hat{\boldsymbol{\theta}}) \right) \right) + (M_m(h, \boldsymbol{\theta}^*) - M(h, \boldsymbol{\theta}^*)) - \left(M_m(h, \hat{\boldsymbol{\theta}}) - M(h, \hat{\boldsymbol{\theta}}) \right) \\ & + M(h, \boldsymbol{\theta}^*) \\ \leq & \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \sup_{\boldsymbol{\theta} \in \Theta} |\mathbb{G}_n (\Psi(\rho_h^m(\boldsymbol{\theta})))| + 2 \sup_{\boldsymbol{\theta} \in \Theta} |M_m(h, \boldsymbol{\theta}) - M(h, \boldsymbol{\theta})| \end{aligned}$$

since $M(h, \boldsymbol{\theta}^*) = \mathcal{R}_\Psi(h, \boldsymbol{\theta}^*) = \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta}))$.

Now, we want to bound the second and third terms in the right member of the last inequality.

Second term. Since $\rho_h^m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))$ and $\rho \mapsto \Psi(\rho, y)$ is convex by Assumption (4.1), we have the inequality for all $y \in \mathcal{Y}$,

$$\begin{aligned} \Psi(\rho_h^m(\boldsymbol{\theta}), y) &= \Psi \left(\frac{1}{m} \sum_{j=1}^m \tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})), y \right) \\ &\leq \frac{1}{m} \sum_{j=1}^m \Psi(\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})), y). \end{aligned}$$

Then, by the linearity of the measure \mathbb{G}_n , it yields

$$(30) \quad \mathbb{G}_n (\Psi(\rho_h^m(\boldsymbol{\theta}))) \leq \frac{1}{m} \sum_{j=1}^m \mathbb{G}_n \Psi(\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))).$$

By Lemma 8.1 we have (a.s)

$$\sup_{\boldsymbol{\theta} \in \Theta} |\mathbb{G}_n(\Psi(\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))))| \leq \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}}$$

where $\mathcal{W}_{(\tilde{\rho}, \Psi)} = \{\Psi(\tilde{\rho}(\lambda), \cdot), \lambda \in \mathcal{Y}\}$, then (a.s)

$$\sup_{\boldsymbol{\theta} \in \Theta} |\mathbb{G}_n(\Psi(\rho_h^m(\boldsymbol{\theta})))| \leq \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}}.$$

Third term. We have

$$\begin{aligned} |M_m(h, \boldsymbol{\theta}) - M(h, \boldsymbol{\theta})| &= |\mathbb{E}_Y(\Psi(\rho_h^m(\boldsymbol{\theta}), Y) - \Psi(\rho_h(\boldsymbol{\theta}), Y))| \\ &\leq \mathbb{E}_Y |\Psi(\rho_h^m(\boldsymbol{\theta}), Y) - \Psi(\rho_h(\boldsymbol{\theta}), Y)|. \end{aligned}$$

By Assumption (4.1)

$$|\Psi(\rho_h^m(\boldsymbol{\theta}), Y) - \Psi(\rho_h(\boldsymbol{\theta}), Y)| \leq L_\Psi(Y) \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}},$$

then

$$(31) \quad |M_m(h, \boldsymbol{\theta}) - M(h, \boldsymbol{\theta})| \leq \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \mathbb{E}_Y L_\Psi(Y).$$

Let $A_\Psi = \mathbb{E}_Y L_\Psi(Y)$.

Moreover, the inequality (6) yields

$$(32) \quad \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \leq \left\| \frac{1}{m} \sum_{j=1}^m [\tilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}} \tilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta}))] \right\|_{\mathcal{F}} + b_h^m(\boldsymbol{\theta}).$$

Equivalently, by considering the empirical process $\mathbb{K}_m^{\mathbf{x}} = \sqrt{m}(\mathbb{P}_m^{\mathbf{x}} - P^{\mathbf{x}})$, we obtain

$$(33) \quad \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \leq \frac{1}{\sqrt{m}} \|\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta}))\|_{\mathcal{F}} + b_h^m(\boldsymbol{\theta})$$

$$(34) \quad \leq \frac{1}{\sqrt{m}} (\|\mathbb{K}_m^{\mathbf{x}} \tilde{\rho}(h(\cdot, \boldsymbol{\theta}))\|_{\mathcal{F}} + \sqrt{m} b_h^m(\boldsymbol{\theta})).$$

Taking the *supremum* over Θ and combining the Lemma (8.2) and the Assumption (3.1) gives

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \leq \frac{1}{\sqrt{m}} \left(c \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} + \sqrt{m} b_h(m) \right).$$

Hence, in (31) we obtain

$$\sup_{\boldsymbol{\theta} \in \Theta} |M_m(h, \boldsymbol{\theta}) - M(h, \boldsymbol{\theta})| \leq \frac{A_\Psi}{\sqrt{m}} \left(c \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} + \sqrt{m} b_h(m) \right).$$

Finally, the following bound holds for the procedure risk

$$\mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}} + 2 \frac{A_\Psi}{\sqrt{m}} \left(c \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} + \sqrt{m} b_h(m) \right).$$

Now, let notice that for any 3 events E_1, E_2, E_3 we have by elementary probability calculus

$$(35) \quad \mathbb{P}(E_1) \leq \mathbb{P}(E_1 \cap E_2 \cap E_3) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c).$$

Take the following events

$$E_1 = \left\{ \mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}} + 2 \frac{A_\Psi}{\sqrt{m}} \left(c \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} + \sqrt{m} b_h(m) \right) \right\}$$

$$E_2 = \left\{ \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}} \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon \right\}$$

and

$$E_3 = \left\{ 2 \frac{A_\Psi}{\sqrt{m}} \left(c \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} + \sqrt{m} b_h(m) \right) \leq 2 \frac{A_\Psi}{\sqrt{m}} \left(c \bar{K}_{(\tilde{\rho}, h)}^\varepsilon + \sqrt{m} b_h(m) \right) \right\},$$

where $\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon$ and $\bar{K}_{(\tilde{\rho}, h)}^\varepsilon$ are such that

$$\mathbb{P}_{Y_1 \dots Y_n} (\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}} \leq \bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon) \geq 1 - \varepsilon$$

and

$$\mathbb{P}_{\mathbf{X}_1 \dots \mathbf{X}_m} (\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} \leq \bar{K}_{(\tilde{\rho}, h)}^\varepsilon) \geq 1 - \varepsilon$$

respectively (for all $\varepsilon > 0$).

Using the inequality (35) with the fact that $\mathbb{P}(E_2) = \mathbb{P}_{Y_1 \dots Y_n} (\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho}, \Psi)}} \leq \bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon)$ and $\mathbb{P}(E_3) = \mathbb{P}_{\mathbf{X}_1 \dots \mathbf{X}_m} (\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho}, h)}} \leq \bar{K}_{(\tilde{\rho}, h)}^\varepsilon)$, we obtain

$$\mathbb{P}(E_1) \leq \mathbb{P}_{Y_1 \dots Y_n, \mathbf{X}_1, \dots, \mathbf{X}_m} \left(\mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon + 2 \frac{A_\Psi}{\sqrt{m}} \left(c \bar{K}_{(\tilde{\rho}, h)}^\varepsilon + \sqrt{m} b_h(m) \right) \right) + 2\varepsilon.$$

But note that $\mathbb{P}(E_1) = 1$, so

$$\mathbb{P}_{Y_1 \dots Y_n, \mathbf{X}_1, \dots, \mathbf{X}_m} \left(\mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{2}{\sqrt{n}} \bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon + 2 \frac{A_\Psi}{\sqrt{m}} \left(c \bar{K}_{(\tilde{\rho}, h)}^\varepsilon + \sqrt{m} b_h(m) \right) \right) \geq 1 - 2\varepsilon.$$

Equivalently, we have with probability at least $1 - 2\varepsilon$

$$\mathcal{R}_\Psi(h, \hat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta} \in \Theta} (\mathcal{R}_\Psi(h, \boldsymbol{\theta})) + \frac{K_{(\tilde{\rho}, \Psi)}^\varepsilon}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K_{(\tilde{\rho}, h)}^\varepsilon + B_m) \right)$$

where

$$K_{(\tilde{\rho}, \Psi)}^\varepsilon = 2 \bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon,$$

$$K_{(\tilde{\rho}, h)}^\varepsilon = A_\Psi c \frac{\bar{K}_{(\tilde{\rho}, h)}^\varepsilon}{\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon}$$

and

$$B_m = \sqrt{m} \frac{A_\Psi}{\bar{K}_{(\tilde{\rho}, \Psi)}^\varepsilon} b_h(m).$$

That concludes the proof. \square

References

- [1] P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.
- [2] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

- [3] D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.
- [4] M.D. Donsker. Justification and extension of Doob’s heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of mathematical statistics*, pages 277–281, 1952.
- [5] R.M. Dudley. Weak convergence of measures on nonseparable metric spaces and empirical measures on euclidian spaces. *Illinois Journal of Mathematics*, 11:109–126, 1966.
- [6] P. Gaenssler. *Empirical Processes*. Institute of Mathematical Statistics, Hayward, CA, 1983.
- [7] A. Goldenshluger and O. Lepski. Uniform bounds for norms of sums of independent random functions. *Arxiv preprint arXiv:0904.1950*, 2009.
- [8] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.
- [9] P.J. Huber. *Robust statistics*. Wiley-Interscience, 1981.
- [10] J.P.C. Kleijnen. *Design and analysis of simulation experiments*. Springer Verlag, 2007.
- [11] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of probability*, 33(3):1060–1077, 2005.
- [12] M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer series in statistics, 2008.
- [13] M. Ledoux. *The concentration of measure phenomenon*. AMS, 2001.
- [14] P. Massart. *Concentration inequalities and model selection: Ecole d’Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer Verlag, 2007.
- [15] P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.
- [16] D. Pollard. Empirical processes: theory and applications. *Regional Conference Series in Probability and Statistics Hayward*, 1990.
- [17] T.J. Santner, B.J. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer Verlag, 2003.
- [18] G.R. Shorack and J.A. Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Statistics, 1986.
- [19] C. Soize and R. Ghanem. Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM Journal on Scientific Computing*, 26:395–410, 2004.
- [20] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.
- [21] S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.
- [22] A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

- [23] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.
- [24] E. Vazquez. (PhD thesis) Modélisation comportementale de systèmes non-linéaires multi-variables par méthodes à noyaux et applications. 2005.