ESAIM: PS 17 (2013) 740–766
DOI: 10.1051/ps/2012025

ESAIM: Probability and Statistics
www.esaim-ps.org

RISK BOUNDS FOR NEW M-ESTIMATION PROBLEMS *

Nabil Rachdi[1,2], Jean-Claude Fort[3] and Thierry Klein[4]

Abstract. In this paper, we consider a new framework where two types of data are available: experimental data $Y_1, \ldots, Y_n$ supposed to be i.i.d from Y and outputs from a simulated reduced model. We develop a procedure for parameter estimation to characterize a feature of the phenomenon Y. We prove a risk bound qualifying the proposed procedure in terms of the number of experimental data n, reduced model complexity and computing budget m. The method we present is general enough to cover a wide range of applications. To illustrate our procedure we provide a numerical example.

Mathematics Subject Classification. 65C60, 60F05, 62F12, 60G20, 65J22.

Received April 6, 2012. Revised August 14, 2012.

1. Introduction

In this paper we present some first results about the statistical study of a random variable Y in a new context: we have at disposal a sample of experimental data resulting from expensive real experiments or heavy computer code, hence we only have a few data. Besides these costly experiments or codes, various reduced models are available. Even if they still are complicated, one can use them to perform simulations in a reasonable computing time and obtain large samples from simulations. This situation is frequently encountered in various field of industry: meteorology, oil extraction, nuclear safety, aeronautics, mechanical engineering *etc*...

Our purpose is to use these two types of information (experimental data, reduced models) to obtain a good statistical description of a feature of the variable Y: it can be its mean, its median, its variance or even its probability density function (*p.d.f.*).

Keywords and phrases. M-estimation, inverse problems, empirical processes, oracle inequalities, model selection.

* *We owe thanks to Fabien Mangeant for advice and discussions, research engineer at EADS Innovation Works, Suresnes, France.*

[1] EADS Innovation Works, 12 rue Pasteur, 92152 Suresnes, France. nabil.rachdi@eads.net

[2] Institut de Mathématiques de Toulouse, 118 route de Narbonne, 31062 Toulouse, France.

[3] Université Paris Descartes, SPC, MAP5, 45 rue des Saints-Pères, 75006 Paris, France. jean-claude.fort@parisdescartes.fr

[4] Institut de Mathématiques de Toulouse, 118 route de Narbonne, 31062 Toulouse, France. thierry.klein@math.univ-toulouse.fr

Article published by EDP Sciences

The reduced models depend on unknown parameters which need to be estimated: examples are a simple physical equation, linear or non linear regression models among them neural networks, kriging approximations. They take the following form: $(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta \mapsto h(\mathbf{x}, \boldsymbol{\theta})$. Generally the variables $\mathbf{x}$ used to build the reduced models are not the same as the ones that have been measured (if they have) as experimental conditions during the experiments leading to the data $Y_1, \ldots, Y_n$. That is why in this work we do not suppose that the available data are couples of input/output variables $(\mathbf{x}_i, Y_i)$: the variables $\mathbf{x}_i$ are not available or are not the same as those used in the reduced models. Thus our only experimental data are the Y_i's, what differs from a classical regression framework.

Let us take an example of particular interest coming from EADS[5] Research department: the effect of an electromagnetic field on the behavior of an aircraft. When lightning or an electromagnetic field strikes an aircraft, sensors measure data corresponding to the intensity of such field in various parts 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 $Y_1, \ldots, Y_n$. On another side, we have at disposal several computer codes h modeling the electromagnetic field in function of the input variables $\mathbf{x}$ and the parameter $\boldsymbol{\theta}$ which can be tuned. The result of such a computer code is a function $h(\mathbf{x}, \boldsymbol{\theta})$. The variables $\mathbf{x}$ will be modeled by random variables, for instance it could be variables describing the atmospheric conditions, the angles of the lightning *w.r.t.* the aircraft, etc... The vector parameter $\boldsymbol{\theta}$ is part of the model and will be estimated. In this case the computer codes have various degrees of complexity. Actually, one has at disposal 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 only one model h.

In general these reduced models remain complex in the following meaning: the mean, the variance, a quantile or the *p.d.f.* of the output cannot be analytically computed. We will say that h is a *complex model* if the feature we are interested in is analytically *unreachable* as function of $\boldsymbol{\theta}$. *Complex models* can arise from several ways. For example, the function $h(\cdot, \boldsymbol{\theta})$ may have a complicated form due to the complexity of the modeling (non linear regression, neural networks), 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.

This aspect and the fact that we only have at disposal the experimental data $(Y_1, \ldots, Y_n)$ are the principal motivations of our work.

In this context, our goal is to construct a *Random Simulator*, $\mathbf{X} \mapsto h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ with $\mathbf{X}$ some random variable, predicting as well as possible a given feature of the distribution of the observed data $Y_1, \ldots, Y_n$. We present a general method based on a criterion to minimize which depends on both experimental and simulated data.

Our framework is not very far from the framework of Y. Auffray, P. Barbillon & J-M. Marin where they look for good metamodels of a time consuming black-box in order to evaluate the probability of rare events by simulation. Yet, in this paper we are not interested in building or analyzing metamodels, but we try to optimally use experimental data and simulated data of a given metamodel, which are not directly coupled. See also the work of P. Barbillon, G. Celeux, Grimaud, Lefebvre and De Rocquigny [1].

This paper is the theoretical part of a work on industrial applications in the field of "Uncertainty Management" [3]. The results we present are theoretical in that the estimation procedures we propose don't include practical implementations. The same is true for the modeling aspect: we deal with

[5]EADS: European Aeronautic Defense and Space Company

(input/output) models without specifying what can be done in practice. We do not deal with the pertinence of the possible reduced models (*metamodels*) (see [10, 18, 20, 25]). The impact of modeling technics will be treated in a forthcoming paper where we will apply some results obtained in this study in an industrial context [17].

The main tool behind the theoretical results we present is the empirical processes theory. This theory constitutes a mathematical toolbox of asymptotic statistics and more recently non-asymptotic statistics. It 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, Dudley [5], Pollard [16], Gaenssler [6], Shorack and Wellner [19] and others. More recently, many references give a general overview of this theory with its applications to statistics, for example [12, 22, 24].

Essential developments of non-asymptotic theory have been done in the last decade by the use of concentration inequalities to derive risk bounds [11, 15, 21] among others. Our work directly derives from these advances.

The starting point of our procedure of estimation is to minimize a contrast. Estimation based on minimizing a function was introduced by Huber in 1964 [8] where he proposed to generalize the maximum likelihood estimation. The resulting estimators are called *M-estimator* [9]. Asymptotic properties of these estimators were widely studied in a general context, and many authors like [22] or [23] used empirical processes theory, which turn out to be a very valuable tool.

This paper is organized as follows. In Section 2 we describe our general framework. In Section 3 we present our method of estimation. Our main result is presented in Section 4: we establish Theorem 4.4 providing a risk bound based on both experimental and simulated data. A numerical illustration is given in Section 5. In Section 6 we make some comments. Then we postponed in Annex A the discussion of the constants in Theorem 4.4. Finally Annex B is devoted to the proofs.

2. General setting

We first present the framework we use along this article.

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 be modeled by a random real valued variable $Y \in \mathcal{Y}$, with unknown distribution Q and f the associated *p.d.f.*. Let us assume that $\mathcal{Y} \subset [-M, M]$, $M > 0$.

Let us suppose that a n-sample $Y_1, \ldots, Y_n$ is available: we call it *experimental data*.
Next, we assume that this complex phenomenon can be approximately represented by the outputs $h(\mathbf{x}, \boldsymbol{\theta})$ of a *reduced model* 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 (*parameter space*).

We equip the input space $\mathcal{X}$ with a probability measure $P^{\mathbf{x}}$ and we get 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 assume that a sample drawn from this distribution is available.

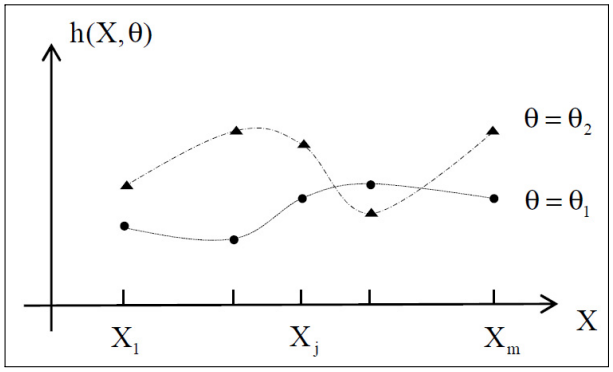

FIGURE 1. Example of model outputs with 2 different parameters.

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

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

which provides m outputs called *simulated data*

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

We emphasize that the $\mathbf{X}_j$'s are the variables used to produce the model outputs but are not the inputs that gave the Y_i's.

In practice, the data $\mathbf{X}_1, \ldots, \mathbf{X}_m$ either arise from simulations of the random variable $\mathbf{X}$ with known distribution $P^{\mathbf{x}}$ or from a large database.

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

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

In this paper, we develop a general method for estimating the parameter $\boldsymbol{\theta}$ based on the *training data* made of the experimental results $Y_1, \ldots, Y_n$ and the simulated inputs of the reduced model h, $\mathbf{X}_1, \ldots, \mathbf{X}_m$. The outputs of the model will depend on the parameter $\boldsymbol{\theta}$ to be estimated.

The method we propose is general enough to include some specific problems met in practice. Two kinds of statistical analysis involving inverse problems can be considered:

- On the one hand to estimate the "true" parameter $\boldsymbol{\theta}^\star$. It aims at estimating "physical" parameters having a real signification like dimensions or material properties for instance.
- On the other hand to estimate a parameter $\boldsymbol{\theta}^\star$ (not necessarily unique) in order to predict the random phenomenon Y. One hopes that $h(\mathbf{X}, \boldsymbol{\theta}^\star) \approx Y$, in the sense that its distribution shares some features with the distribution of Y: the same mean, variance, probability tail or the same $p.d.f$.
  Here, the parameter $\boldsymbol{\theta}^\star$ may have no real (physical) meaning. For instance it is the case when using reduced models given by a Multi-Layer Perceptron, where the parameter is the values of the connections.

2.2. Tools for evaluating the estimation performance

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

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

A *feature* of the distribution μ is a quantity of the form $\rho_{\mathcal{F}}(\mu) \in \mathcal{F}$ 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). The former case is part of the later one, identifying scalars to constant functions.

We equip the feature space $\mathcal{F}$ with the norm $\|\cdot\|_{\mathcal{F}}$ which can be a L_r-norm ($r \geq 1$) when $\mathcal{F}$ is a space of functions defined on $\mathcal{Y}$.

In all what follows, we denote by $\rho_h(\boldsymbol{\theta})$ a feature of the distribution of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$.

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

$$F_{h,\Theta} = \{\rho_h(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}. \tag{2.1}$$

Definition 2.1 (**Contrast and risk function).**
A **contrast function** (with value in $L_1(Q)$) is any function

$$\begin{aligned} \Psi \,:\, \mathcal{F} &\longrightarrow L_1(Q) \\ \rho &\longmapsto \Psi(\rho, \cdot) \,:\, y \in \mathcal{Y} \longmapsto \Psi(\rho, y), \end{aligned} \tag{2.2}$$

such that

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

is *unique*.

We call **risk function** the application

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

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

$$\mathcal{R}_\Psi(h, \boldsymbol{\theta}) := \mathbb{E}_Y\, \Psi\left(\rho_h(\boldsymbol{\theta}),\, Y\right), \tag{2.3}$$

where, for a random variable ξ, we used the notation $\mathbb{E}_\xi$ for the expectation *w.r.t.* the variable ξ.

Example 2.2 (Some classical features, associated contrasts and classical risk functions)**.**

– $\mathcal{F} = \mathbb{R}$ (constant functions): we may consider $\rho_h(\theta) = \mathbb{E}_X h(X, \theta)$ (mean), $\rho_h(\theta) = \mathbb{E}_X \mathbb{1}_{[s,+\infty[}(h(X, \theta))$(exceeding probability).

Mean-contrast

$$\Psi\left(\rho, y\right) = (y - \rho)^2, \quad \mathcal{R}_\Psi(h, \boldsymbol{\theta}) = \left(\mathbb{E}(Y) - \rho_h(\boldsymbol{\theta})\right)^2 + \operatorname{Var}(Y)$$

– $\mathcal{F} = \{\textit{set of density functions}\}$

log**-contrast**

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

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

L_2**-contrast**

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

In view of these examples, it makes sense to investigate models h or/and parameters $\boldsymbol{\theta}$ providing small risk values. Here we restrict our study to parameters of one model h.

3. The method.

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. Let us introduce our method:
We want to estimate a parameter $\boldsymbol{\theta}^\star$ minimizing the risk (2.3), *i.e.*

$$\boldsymbol{\theta}^\star \in \operatorname*{Argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_\Psi(h, \boldsymbol{\theta}). \tag{3.1}$$

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 $\mathcal{R}_\Psi(h, \boldsymbol{\theta}^\star)$, also called *ideal risk*.

However, the risk function $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$ is not computable (hence $\boldsymbol{\theta}^\star$) for two reasons: the measure Q is unknown, and we are dealing with complex models.
We aim at computing a parameter $\widehat{\boldsymbol{\theta}}$ that performs as well as $\boldsymbol{\theta}^\star$, that is

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

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

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

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 $Y_1, \ldots, Y_n$. 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 $h(\mathbf{X}_1, \boldsymbol{\theta}), \ldots, h(\mathbf{X}_m, \boldsymbol{\theta})$. We suppose that $\rho_h^m(\boldsymbol{\theta})$ takes the following form

$$\rho_h^m(\boldsymbol{\theta}) := \frac{1}{m} \sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) \tag{3.2}$$

where $\frac{1}{m}\widetilde{\rho} : \mathcal{Y} \to \mathcal{F}$ is a *weight function* depending on the contrast Ψ considered.

For sake of simplicity, we may also call $\widetilde{\rho}$ weight function.

Example 3.1 (Examples of weight functions)**.**

– *mean-contrast*

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

– log-*contrast or* L_2*-contrast density estimation*

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

where $K_b(\cdot - y) = \frac{1}{b}K(\frac{\cdot - y}{b})$ for a kernel K and a bandwidth b (See Figure 2 for an illustration). It means that the method of estimation relies on kernel density estimation.

Another choice is to use an expansion on a given (truncated) L_2-basis, $(\varphi_j, 0 \leq j \leq L)$, which leads to the weight function

$$\frac{1}{m}\widetilde{\rho}(y)(\cdot) = \frac{1}{m} \sum_{j=1}^{L} \varphi_j(\cdot)\varphi_j(y)$$

Remark 3.2. The weight function $\frac{1}{m}\widetilde{\rho}(y)$ evaluated at $y \in \mathcal{Y}$ can be either a scalar value ($\frac{1}{m}$ for the mean) or a function (for the density). So that without loss of generality, one can see the weight function $\frac{1}{m}\widetilde{\rho}(y)$ at a point $y \in \mathcal{Y}$ as a function,

$$\widetilde{\rho}(y) \ : \ \lambda \in \mathcal{Y} \longmapsto \widetilde{\rho}(y)(\lambda).$$

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

In the sequel, the examples of density estimation will be carried out from the kernel method. We chose this method because it is simple to write and so very popular in the uncertainty management in industrial context. Notice that we will assume some adaptivity of our kernel estimator by choosing a bandwidth $b = b_m$ that will depend on m.

Definition 3.3. We denote by $\sigma_h^m(\boldsymbol{\theta})$, called *simulation error*, the error committed while estimating the feature $\rho_h(\boldsymbol{\theta})$ by the estimator $\rho_h^m(\boldsymbol{\theta})$,

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

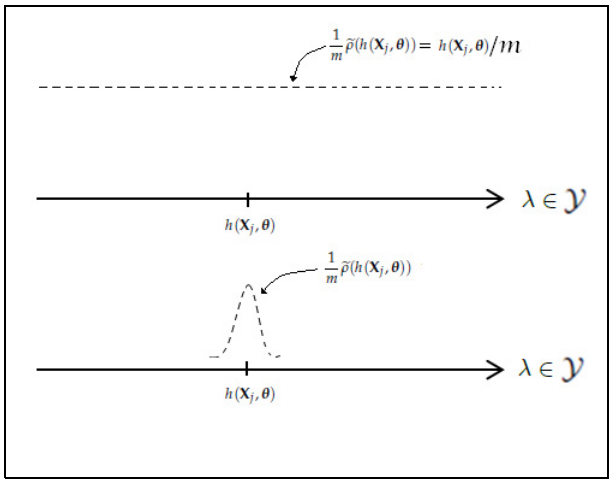

FIGURE 2. Example of weight function in the case of the mean (top) and in the case of the density (bottom).

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

$$\begin{aligned}
\sigma_h^m(\boldsymbol{\theta}) &= \|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
&= \|\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
&= \|\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) + \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
&\leq \|\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\|_{\mathcal{F}} + \|\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \\
&= \left\| \frac{1}{m}\sum_{j=1}^{m} [\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))] \right\|_{\mathcal{F}} + B_h^m(\boldsymbol{\theta})
\end{aligned} \tag{3.3}$$

with

$$B_h^m(\boldsymbol{\theta}) := \|\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})\|_{\mathcal{F}} \tag{3.4}$$

the *bias error*.

The first term in the right hand side of inequality (3.3) is a *variance* (random) term, and the second is a *bias* (deterministic) term.

Assumption 3.4. We assume that the plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ (3.2) has a uniformly bounded bias, *i.e.* it exists some constant $B_h(m)$ depending on h and m such that the bias error (3.4) satisfies

$$\sup_{\boldsymbol{\theta}\in\Theta} B_h^m(\boldsymbol{\theta}) < B_h(m) < \infty. \tag{3.5}$$

We give an example of such a constant $B_h(m)$ in Annex A.

Finally, the criterion we propose to minimize has the form

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

which provides the estimator

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta})),\, Y_i\right), \tag{3.6}$$

or equivalently

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

In the various cases we mentioned it gives $\widehat{\boldsymbol{\theta}}_M = \operatorname{Argmin}_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}(Y_i - h(\mathbf{X}_j,\boldsymbol{\theta}))\right)^2$ for the *mean-contrast*, $\widehat{\boldsymbol{\theta}}_{\log} = \operatorname{Argmin}_{\boldsymbol{\theta}\in\Theta} -\sum_{i=1}^{n}\log\left(\sum_{j=1}^{m}K_b(Y_i - h(\mathbf{X}_j,\boldsymbol{\theta}))\right)$ for the log-*contrast* and $\widehat{\boldsymbol{\theta}}_{L_2} = \operatorname{Argmin}_{\boldsymbol{\theta}\in\Theta}\left\{\left\|\sum_{j=1}^{m}K_b(\cdot - h(\mathbf{X}_j,\boldsymbol{\theta}))\right\|_2^2 - \frac{2\,m}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}K_b(Y_i - h(\mathbf{X}_j,\boldsymbol{\theta}))\right\}$ for the L_2-*contrast*.

Remark 3.5. The estimator $\widehat{\boldsymbol{\theta}}$ depends on the model h, the number of experimental data n and the number of simulation data m. The number of simulations m has to be thought greater than n. In our framework experimental data are difficult to obtain whereas simulated data are more reachable.

Now the issue is to find the statistical properties of this procedure taking into account the two kinds of data: experimental and simulated data.

Once we defined the procedure for computing $\widehat{\boldsymbol{\theta}}$, we have to qualify the *quality* of this procedure, which is 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

$$\mathcal{R}_\Psi(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h(\boldsymbol{\theta}), Y\right),$$

$$\boldsymbol{\theta}^\star \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}}\ \mathcal{R}_\Psi(h,\boldsymbol{\theta}),$$

and that $\widehat{\boldsymbol{\theta}}$ is defined by (3.6). Now, we give some definitions and notations useful to set Theorem 4.4. Denote by

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

and

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

TABLE 1. Example of classes of functions and constant A_Ψ (see Annex A).

	$\mathcal{W}_{(\widetilde{\rho},\Psi)}$	$\mathcal{P}_{(\widetilde{\rho},h)}$	A_Ψ
mean-contrast	$y \mapsto (y-\lambda)^2$, $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto h(\mathbf{x},\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$	$4\,M$
log-contrast	$y \mapsto -\log(K_b(y-\lambda))$, $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto K_b(\lambda - h(\mathbf{x},\boldsymbol{\theta}))$, $(\lambda,\boldsymbol{\theta}) \in \Theta \times \mathcal{Y}$	$\Vert f\Vert_2/\eta$
L_2-contrast	$y \mapsto \Vert K_b(\cdot - \lambda)\Vert_2 - 2\,K_b(y-\lambda)$, $\lambda \in \mathcal{Y}$	*idem*	$2\,(\Vert f\Vert_2 + B)$

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

Let's define the classes of functions

$$\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\widetilde{\rho}(\lambda),\, y),\, \lambda \in \mathcal{Y}\}, \tag{4.1}$$

$$\mathcal{P}_{(\widetilde{\rho},h)} = \{\mathbf{x} \in \mathcal{X} \mapsto \widetilde{\rho}(h(\mathbf{x},\boldsymbol{\theta}))(\lambda),\, (\boldsymbol{\theta},\lambda) \in \Theta \times \mathcal{Y}\}. \tag{4.2}$$

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

$$P\,g := \int g(u)P(\mathrm{d}u) \qquad g \in \mathcal{G}$$

and

$$\Vert P\Vert_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |P\,g|.$$

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

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

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

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

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

In our applications, the class of functions $\mathcal{G}_{\mathcal{Y}}$ is $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ and $\mathcal{G}_{\mathcal{X}}$ is $\mathcal{P}_{(\widetilde{\rho},h)}$, respectively defined in (4.1) and (4.2).

We make the following assumptions.

Assumption 4.2. Let $\mathcal{W}^m_{(\widetilde{\rho},\Psi)}$ be the class of functions

$$\mathcal{W}^m_{(\widetilde{\rho},\Psi)} = \left\{ y \in \mathcal{Y} \mapsto \Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(\lambda_j),\, y\right),\ (\lambda_j)_{1\le j\le m} \in \mathcal{Y}^m \right\}.$$

We assume that it exists some universal constant $\gamma > 0$ such that almost surely (a.s.)

$$\|\mathbb{G}_n\|_{\mathcal{W}^m_{(\widetilde{\rho},\Psi)}} \le \gamma\, \|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}},$$

where $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ is given in (4.1).

This assumption may be explained by the fact that the "complexity" of the class of functions $\mathcal{W}^m_{(\widetilde{\rho},\Psi)}$ is "close" to the complexity of $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ which can be viewed as $\mathcal{W}_{(\widetilde{\rho},\Psi)} = \mathcal{W}^{m=1}_{(\widetilde{\rho},\Psi)}$. In other words, we assume that the summation of functions $\widetilde{\rho}(\lambda_j)$ does not have a significant impact on the behavior of $\|\mathbb{G}_n\|_{\mathcal{W}^m_{(\widetilde{\rho},\Psi)}}$.

Assumption 4.3. Let us assume that the contrast Ψ satisfies

– for all $\rho_1, \rho_2 \in \mathcal{F}$

$$\mathbb{E}_Y\, |\Psi(\rho_1, Y) - \Psi(\rho_2, Y)| \le A_\Psi\, \|\rho_1 - \rho_2\|_{\mathcal{F}}$$

with a constant $A_\Psi < \infty$ independent of ρ_1, ρ_2.

In Annex A we compute the constant A_Ψ in various cases (see Tab. 1).

Then our main result follows:

Theorem 4.4 (Risk bound for Parameter Estimation)**.**

Under the assumptions (4.2), (4.3) *and* (3.4)*, suppose that the sequences of random variables* $\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$ *and* $\|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\widetilde{\rho},h)}}$ *are tight. Denote by* $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ *and* $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ *the associated constants, uniform (or decreasing) in* n *and* m*, respectively.*

Let the feature space $\mathcal{F}$ *be 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*

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}) \le \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h, \boldsymbol{\theta})\right) + \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\left(1 + \sqrt{\frac{n}{m}}\left(K^{\varepsilon}_{(\widetilde{\rho},h)} + B_m\right)\right)$$

where the constants $K^{\varepsilon}_{(\widetilde{\rho},\Psi)}$, $K^{\varepsilon}_{(\widetilde{\rho},h)}$ *depend on* $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$, $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$, A_Ψ, M *and* r*.* B_m *is a bias factor depending on* $B_h(m)$*.*

Remark 4.5. This result may yield consistency results if we assume that $B_h(m)$ tends to 0 when m tends to ∞. They would depend on the choice of the contrast.

5. Numerical illustration

Let us consider the following academic example. Suppose that the random phenomenon Y follows

$$Y = \sin(\xi) + 0.01\,\varepsilon, \tag{5.1}$$

where ξ and ε are two standard gaussian independent random variables.

Denote by $Y_1, \ldots, Y_n$ n i.i.d realizations of the random variable Y. We aim at estimating the *p.d.f.* of Y, that we denote by f, from the n-sample $Y_1, \ldots, Y_n$ and the following model h given by

$$h(\mathbf{X}, \boldsymbol{\theta}) = \theta_1 + \theta_2\,\mathbf{X} + \theta_3\,\mathbf{X}^3, \quad \mathbf{X} \sim \mathcal{N}(0,1),\ \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3). \tag{5.2}$$

In this example the Taylor expansion of the sin function takes place of a "metamodel".

Now our problem amounts to estimating the parameter $\boldsymbol{\theta}$ by an estimator $\widehat{\boldsymbol{\theta}}$, and then predict the *p.d.f.* of Y by the *p.d.f.* of the random variable $h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$.

In the previous setting, we built the estimator $\widehat{\boldsymbol{\theta}}$ as a M-estimator given by (3.6)

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

for some contrast $\Psi : \mathcal{F} \to L_1(Q)$ and m realizations $\mathbf{X}_1, \ldots, \mathbf{X}_m$ i.i.d from X. For this illustration, we propose to use the log-contrast

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

and the weight function

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

where $K_b(\cdot - y) = \frac{1}{b} K(\frac{\cdot - y}{b})$ with the gaussian kernel $K(u) = \frac{1}{\sqrt{2\,\pi}}\, e^{-u^2/2}$, and a bandwidth $b = b^m_{\boldsymbol{\theta}}$ computed from the sample $h(\mathbf{X}_j, \boldsymbol{\theta})$, $j = 1, \ldots, m$ for $\boldsymbol{\theta} \in \Theta$, by the Silverman's rule-of-thumb:

$$b^m_{\boldsymbol{\theta}} = 1.06\, m^{-1/5}\, \hat{\sigma}_{\boldsymbol{\theta}}. \tag{5.3}$$

The quantity $\hat{\sigma}_{\boldsymbol{\theta}}$ is the empirical standard deviation of the sample $h(\mathbf{X}_j, \boldsymbol{\theta})$, $j = 1, \ldots, m$

$$\hat{\sigma}_{\boldsymbol{\theta}} = \frac{1}{m} \sum_{j=1}^{m} \left(h(\mathbf{X}_j, \boldsymbol{\theta}) - \frac{1}{m} \sum_{j=1}^{m} h(\mathbf{X}_j, \boldsymbol{\theta}) \right)^2.$$

Finally, the estimator $\widehat{\boldsymbol{\theta}}$ becomes

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}} - \sum_{i=1}^{n} \log\left(\frac{1}{m} \sum_{j=1}^{m} K_{b^m_{\boldsymbol{\theta}}}(Y_i - h(\mathbf{X}_j, \boldsymbol{\theta})) \right). \tag{5.4}$$

Figure 3 shows the *p.d.f.* of the random variable $h(X, \widehat{\boldsymbol{\theta}})$, with $h(\cdot,\cdot)$ given in (5.2), for the three computed values $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3)$ in Table 2. The computation of the *p.d.f.* of $h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ is made through an intensive kernel smoothing.

We clearly see that the use of an approximate but reasonable "metamodel" greatly improves the estimation of f. We also notice, as expected, that increasing n (and m as a consequence) strongly impacts the quality of the M-estimator $\widehat{\boldsymbol{\theta}}$.

TABLE 2. M-estimator computations by varying n and m.

	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$KL(f,\widehat{f})$
$n=50$, $m=10^3$	$(3.9\pm 5).10^{-2}$	$(9.95\pm 0.5).10^{-1}$	$(-1.65\pm 0.4).10^{-1}$	$(4.9\pm 2).10^{-2}$
$n=500$, $m=5.10^3$	$(2.22\pm 0.5).10^{-2}$	$(9.56\pm 0.3).10^{-1}$	$(-1.36\pm 0.3).10^{-1}$	$(2.3\pm 0.4).10^{-2}$
$n=1000$, $m=10^4$	$(7\pm 3).10^{-3}$	$(9.52\pm 0.2).10^{-1}$	$(-1.31\pm 0.2).10^{-1}$	$(1.1\pm 0.2).10^{-2}$

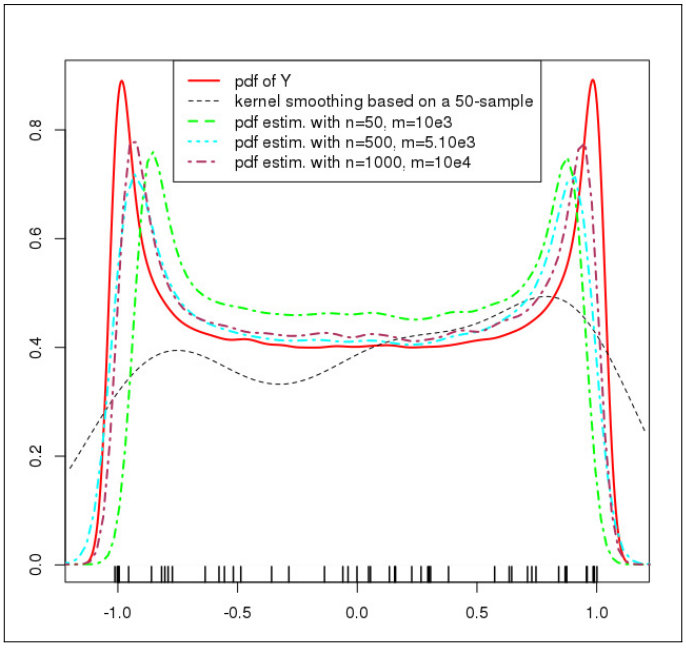

FIGURE 3. Comparison of probability density functions.

6. SOME COMMENTS

It is of interest to compare the methodology we are developing 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 (3.6) is classically

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

and we can derive immediately a risk bound.

Proposition 6.1 (Basic risk bound).
It holds that

$$\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}_n) \le \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2}{\sqrt{n}}\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi}, \tag{6.1}$$

where

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

Proof. The proof comes from a classical calculus in M-estimation, see for example [23] (p. 46). □

For most of statistical procedures, as likelihood, regression, classification etc ... risk bound like (6.1) can be found. Such procedures have been widely studied, with a large literature available. Recently, authors use the Empirical Processes theory (see [12, 22–24] among others) to derive limit theorems. Indeed, the asymptotic (and non-asymptotic) properties of the estimator $\widehat{\boldsymbol{\theta}}_n$ 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.* $\boldsymbol{\theta}^\star$ 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 [22].

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 (6.1), with high probability

$$\mathcal{R}_{\Psi}(h, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_{\Psi}(h, \boldsymbol{\theta})\right) + \frac{K}{\sqrt{n}}. \tag{6.2}$$

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, 14, 21, 24].
Inequality (6.1) can not be applied to our framework because the induced procedure $\widehat{\boldsymbol{\theta}}_n$ involves the quantity $\rho_h(\boldsymbol{\theta})$ which is untractable for *complex models*.

The result of Theorem 4.4 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 [2]). It derives from our assumptions, more precisely, when we supposed the tightness of the sequences of the random variables $\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$ ($Y_{1..n}$-dependent) and $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}}$ ($\mathbf{X}_{1..m}$-dependent). Moreover, we insist on the fact that the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ (that bounds $\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$) and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ (that bounds $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}}$) are uniform (or decreasing) in n and m, respectively. The advantage of this uniformity is the explicit expression of the *residual* term

$$\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\left(1 + \sqrt{\frac{n}{m}}\left(K^{\varepsilon}_{(\widetilde{\rho},h)} + B_m\right)\right) \tag{6.3}$$

depending on the data (n and m) on one hand, and on the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$, $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ 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 Annex A how to compute such constants using concentration inequalities. Let us assume for a moment the existence of these constants.
We showed that the estimation procedure $\widehat{\boldsymbol{\theta}}$ defined in (3.6) "mimics" the ideal risk $\mathcal{R}_{\Psi}(h, \boldsymbol{\theta}^\star) = \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_{\Psi}(h, \boldsymbol{\theta})\right)$ up to the residual term (6.3). Making $m \longrightarrow +\infty$, this residual becomes simply $\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$ which has the same form as the ones found in classical cases (6.2). We find the usual rate of convergence $\sqrt{n}$.

In our purpose, the factor

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

we call *simulation factor*, is due to the simulations used to estimate the feature $\rho_h(\boldsymbol{\theta})$ of the random output $h(\mathbf{X},\boldsymbol{\theta})$ by the plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ we defined in (3.2).

It appears that for a given n, one should have a number of simulation data m greater than n.
The *ideal risk* is $\inf_{\boldsymbol{\theta}\in\Theta}(\mathcal{R}_{\Psi}(h,\boldsymbol{\theta}))$. It can be understood as the "distance" between the *a priori* knowledge one has and the observed phenomenon, or a "distance" between the "best" a priori information available and the "target". Let's consider the case of density estimation. If the *ideal risk* is supposed equal to zero, it means that we believe that the *p.d.f.* 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):

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

However, such *a priori* has to be made with precautions. It is necessary to verify that the model h is able to reach a sufficiently large range of distributions, so that one believes that f belongs to the model.

Conclusion

We introduced a new framework to estimate some feature of a random variable Y, framework which corresponds to many situations encountered in practice. In this framework we proposed a procedure of estimation, that reveals to be satisfactory in a numerical illustration. For a more valuable application we refer to [17].

Our first theoretical result gives an upper bound for the risk of the estimator we proposed. It can be used to get consistency results under suitable assumptions.

This article is only a first step, for instance in the asymptotic ($n\to\infty, m\to\infty$) the next step would be to obtain a rate of convergence and a central limit theorem. To this end some technical difficulties related to the class of functions $\mathcal{W}^m_{(\widetilde{\rho},\Psi)}$ are to be overcome. We hope to make it in a forthcoming work.

ANNEXES

A. About the constants in Theorem 4.4

Constant A_{Ψ}

We will show how we obtain the constants A_{Ψ} in Table (1). Let us 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

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

This yields $A_{\Psi}=4M$.

– log-*contrast.*

Let $y\in\mathcal{Y}$, $\rho_1,\rho_2\in\mathcal{F}$, with $\mathcal{F}$ some set of *p.d.f.*

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 *p.d.f.* 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

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

This yields $A_\Psi = \frac{\|f\|_2}{\eta}$.

– *L_2-contrast.*

Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F}$, with $\mathcal{F}$ be some set of *p.d.f.*.

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

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

By triangular inequality

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

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

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

We get $A_\Psi = 2\,(B + \|f\|_2)$.

The two previous assumptions on the densities (uniformly lower bounded or upper bounded in L_2) are restrictive. Yet many densities with a fixed compact support belong to one or the other set, which allows to choose between the two contrasts. Of course it needs an *a priori* information, which is not always available or true.

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 (3.4) is zero for all $\boldsymbol{\theta} \in \Theta$ and all $m > 0$, hence $B_h(m) = 0$ too. This is not the case for the density estimation.

We study the example of the kernel estimator, *i.e.* when the weight function $\widetilde{\rho}$ is a function of the form

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

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

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

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

Using Theorem 24.1 in [23] (p. 345) we easily obtain:

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

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

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

We detail the arguments for computing the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$. Since these constants are tightness constants relative to some empirical processes (see the assumptions of Thm. 4.4), 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

$$\mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \leq K(\varepsilon)) \geq 1 - \varepsilon \quad \text{for small } \varepsilon > 0. \tag{6.4}$$

For this, we propose to use the work of T. Klein and E. Rio [11], in particular Theorem 1.1, which deals 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 $[-1, 1]$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \mathbb{W}_p(g) \geq \mathbb{E}(\sup_{g\in\mathcal{G}} \mathbb{W}_p(g)) + t \right) \leq \exp\left(-\frac{t^2}{2v + 3\,t/\sqrt{p}} \right), \tag{6.5}$$

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-sides control. Hence, according to the work of T. Klein and E. Rio [11], it exists some function $\varphi_{\mathcal{G}} : \mathbb{R}_+ \to [0, 1]$ decreasing to zero such that for all positive t

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{G}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + t\right) \leq \varphi_{\mathcal{G}}(t). \tag{6.6}$$

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 6.2. *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 C of $\mathbb{R}^l$, $l \geq 1$. Suppose that the application*

$$\lambda \in \mathcal{C} \longmapsto g_\lambda \in \mathcal{G} \subset L_2 \tag{6.7}$$

is continuous.

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

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{G}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + t\right) \leq \varphi_{\mathcal{G}}(t). \tag{6.8}$$

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

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

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

We define the sets $\mathcal{Y}^s = \{y_1^s, \ldots, y_{i_s}^s\}$ for $s \geq 1$ recursively. $\mathcal{Y}^1 = \{-M, 0, M\}$, assuming that the set $\mathcal{Y}^s = \{y_1^s, \ldots, y_{i_s}^s\}$ is built and reordering the elements in increasing order, we take the middle points $\tilde{y}_j^s = \frac{y_j^s + y_{j+1}^s}{2}$ and obtain $\widetilde{\mathcal{Y}}^s = \{\tilde{y}_j^s,\, i = 1, \ldots, i_{s-1} - 1\}$. Then we define

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

reordered to have increasing elements, and it holds $\mathrm{Card}(\mathcal{Y}^s) = 2^s + 1$.
Now, we define the classes of functions

$$\mathcal{W}^s_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\widetilde{\rho}(\lambda),\, y),\, \lambda \in \mathcal{Y}_s\}$$

noticing that for all $s \geq 1$,

$$\mathcal{W}^{s-1}_{(\widetilde{\rho},\Psi)} \subsetneq \mathcal{W}^s_{(\widetilde{\rho},\Psi)} \subsetneq \mathcal{W}_{(\widetilde{\rho},\Psi)}. \tag{6.9}$$

By this previous display and the fact that $\bigcup_{s\geq 1} \mathcal{Y}^s$ is dense in $[-M, M]$ and by the continuity Assumption (6.7), we have

$$\overline{\lim_{s\to\infty} \mathcal{W}^s_{(\widetilde{\rho},\Psi)}} = \overline{\bigcup_{s\geq 1} \mathcal{W}^s_{(\widetilde{\rho},\Psi)}} = \mathcal{W}_{(\widetilde{\rho},\Psi)}. \tag{6.10}$$

The classes of functions $\mathcal{W}^s_{(\widetilde{\rho},\Psi)}$, $s \geq 1$ are countable with values in $[-1,1]$ and we may apply the inequality (6.6) to the classes $\mathcal{W}^s_{(\widetilde{\rho},\Psi)}$. We get for all $t \geq 0$ and $s \geq 1$

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}}\right) + t\right) \leq \varphi_s(t). \tag{6.11}$$

We then prove that the two members of this inequality converge when $s \to \infty$.

Write the left member as follows

$$\begin{aligned}
&\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}} \geq \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}}\right) + t\right)\\
&= \mathbb{E}\left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}} \geq \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}}\right)+t}\right)\\
&= \mathbb{E}\left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}} - \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}}\right) \geq t}\right). \qquad (6.12)
\end{aligned}$$

The inclusions (6.9) yield

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

so the sequence $\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}}\right)_{s\geq 1}$ is increasing and bounded, thus it converges. By monotone convergence, we obtain that the sequence $\left(\mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}}\right)\right)_{s\geq 1}$ converges too provided that $\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\tilde{\rho},\Psi)}}) < \infty$. Thus, the sequence $\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}} - \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\tilde{\rho},\Psi)}}\right)\right)_{s\geq 1}$ converges, and by dominated convergence the quantity (6.12) converges to the wanted limit

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

For the right member of (6.11), by similar arguments, it can be shown that $\varphi_s(t) \to \varphi(t) = \varphi_{\mathcal{G}}(t)$.

That concludes the proof. □

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

$$\forall\, t \geq 0 \quad \kappa_{\mathcal{G}}^{-1}(t) = \varphi_{\mathcal{G}}(t). \qquad (6.13)$$

Then, we can write (6.8) as follows, for all $\varepsilon \in]0,1[$

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

or equivalently

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

Thus, one can take $K(\varepsilon)$ equal to $\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon)$ and $K(\varepsilon)$ satisfies (6.4), *i.e.*

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

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*. Although these methods are known to be not sharp, the bounds we will obtain are of interest for our purpose.

Before, let us recall some useful notations from [24] (p. 83–85).

Let $\mathcal{G}$ be a class of functions and W some probability measure. We denote $G : y \mapsto G(y)$ an *envelope function* of the class $\mathcal{G}$. The bracketing number is $N_{[]}(\epsilon, \mathcal{G}, L_2(W))$ and the entropy with bracketing is the logarithm of the bracketing number. Last, the bracketing integral is defined as

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

Now we apply Corollary 19.35 of [23] (p. 288), it holds that

$$\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) \leq a_{\mathcal{G}}\, J_{[]}(\|G\|_{2,W}, \mathcal{G}, L_2(W)), \tag{6.14}$$

where

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

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

Remark 6.3. The quantity $J_{[]}(\|G\|_{2,W}, \mathcal{G}, L_2(W))$ is computable if one has the bracketing numbers $N_{[]}(\epsilon, \mathcal{G}, L_2(W))$ $(\forall\, \epsilon > 0)$.

Finally, setting

$$K(\varepsilon) = a_{\mathcal{G}}\, J_{[]}\left(\|G\|_{2,Q}, \mathcal{G}_{(\widetilde{\rho},\Psi)}, L_2(W)\right) + \kappa_{\mathcal{G}}(\varepsilon) \tag{6.15}$$

provides the claimed constant. In particular, we should take $\mathcal{G} = \mathcal{W}_{(\widetilde{\rho},\Psi)}$ $(W = Q)$ and $\mathcal{G} = \mathcal{P}_{(\widetilde{\rho},h)}$ $(W = P^{\mathbf{x}})$ in order to compute $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$, respectively. We give explicit computations in the two following cases.

- $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ for the Mean-contrast.
  We recall that in this case
  $$\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{y \mapsto (y-\lambda)^2,\, \lambda \in \mathcal{Y}\}.$$
  This class is uniformly bounded by $4\,M^2$, we take the envelop function $G = 4\,M^2$. Then, we have
  $$|(y-\lambda_1)^2 - (y-\lambda_2)^2| \leq |\lambda_1 - \lambda_2|\, F(y),$$
  with $F(y) = |2y + 2\,M|$.
  Following the lines of [24] we obtain the constant:
  $$\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} = 8\, a_1 \sqrt{\pi}\, M^2 + \kappa_1(\varepsilon). \tag{6.16}$$
- $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ with the weight function $\widetilde{\rho}(y) = y$.
  In this case, the class of functions $\mathcal{P}_{(\widetilde{\rho},h)}$ is
  $$\mathcal{P}_{(\widetilde{\rho},h)} = \{\mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x}, \boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\} \quad (\mathcal{X} \subset \mathbb{R}^d).$$
  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 denoting by P an envelop of $\mathcal{P}_{(\widetilde{\rho},h)}$, we may take $P = M$.

Moreover, let us 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

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

where

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

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

$$D^\nu = \frac{\partial^{|\nu|}}{\partial \nu_1^{\nu_1} \ldots \partial \nu_d^{\nu_d}}, \quad \textit{with} \quad |\nu| = \sum_{i=1}^{d} \nu_i.$$

We aim at computing the entropy integral $J_{[]}\left(\|P\|_{2,Q}, \mathcal{P}_{(\tilde{\rho},h)}, L_2(Q)\right)$ by integrating the entropy $\log N_{[]}(\epsilon, \mathcal{P}_{(\tilde{\rho},h)}, L_2(Q))$.
Corollary 2.7.2 in [24] (p. 157) gives an entropy bound for the Hölder space $\mathbb{H}(\mathcal{X}, \alpha, 1)$:

$$\log N_{[]}(\epsilon, \mathbb{H}(\mathcal{X}, \alpha, 1), L_2(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{d/\alpha} \quad \forall \epsilon > 0, \tag{6.17}$$

where K depends on α, $\mathrm{diam}(\mathcal{X})$ and d.
Using (6.17) and the inequality

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

it holds for $d < 2\alpha$

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

hence

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

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

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

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 6.4. To compute the constants $\bar{K}^{\varepsilon}_{(\tilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\tilde{\rho},h)}$ is difficult but we have adopted a nonasymptotic point of view, so that these computations are necessary in order to get numerical values of the risk bounds.

B. Proofs

To prove the risk bound of Theorem (4.4), we need the following lemmas.

Preliminary lemmas

Lemma 6.5. *Consider the random functions*

$$y \mapsto \Psi(\rho_h^m(\boldsymbol{\theta}),\, y), \quad \boldsymbol{\theta} \in \Theta \quad \textit{with} \quad \rho_h^m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta}))$$

We have (a.s.)

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \le \gamma\, \|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}},$$

where $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ *is defined in* (4.1).

Proof. Conditionally to $\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_m = \mathbf{x}_m$, we have trivially

$$\left\{ \rho_h^m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{x}_j, \boldsymbol{\theta})),\, \boldsymbol{\theta} \in \Theta \right\} \subset \left\{ \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(\lambda_j),\, (\lambda_j)_{1\le j\le m} \in \mathcal{Y}^m \right\}.$$

Hence it yields that

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \le \|\mathbb{G}_n\|_{\mathcal{W}^m_{(\widetilde{\rho},\Psi)}},$$

with $\mathcal{W}^m_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\frac{1}{m}\sum_{j=1}^m \widetilde{\rho}(\lambda_j),\, y),\, (\lambda_j)_{1\le j\le m} \in \mathcal{Y}^m\}$. By Assumption 4.2 we obtain that

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \le \gamma\, \|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}},$$

for some universal constant $\gamma > 0$.

Finally, the right member does not depend on $\mathbf{x}_1, \ldots, \mathbf{x}_m$, and the result follows. □

Lemma 6.6. *Consider the* $P^{\mathbf{x}}$*-empirical process* $\mathbb{K}_m^{\mathbf{x}}$ *and let* $\|\cdot\|_{\mathcal{F}} = |\cdot|$ *or* $\|\cdot\|_r$ *and define*

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

We have

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

where $\mathcal{P}_{(\widetilde{\rho},h)}$ *is defined in* (4.2).

Proof. Let us notice that the quantity

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

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

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

Taking $\|\cdot\|_{\mathcal{F}} = |\cdot|$ the absolute value yields

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

Remark 6.7. In this case, $\widetilde{\rho}(y)(\lambda) = \widetilde{\rho}(y)$ for all y and λ in $\mathcal{Y}$.

– If, for all $y \in \mathcal{Y}$, $\widetilde{\rho}(y)$ is a real valued function defined on $\mathcal{Y}$.

We take $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_r$, $r \geq 1$, the L_r norm. By integration properties and the fact that $z \to z^r$ is increasing on $\mathbb{R}^+$ we have

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

Finally, notice that

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

and the result follows. □

Remark 6.8. In the case where the weight function is a kernel $K_b(\cdot - \cdot)$, the quantity

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

is treated as a sum of independent random functions in the recent work of Goldenshluger and 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].

Proof of Theorem (4.4)

Proof. We denote by

- $M(h,\boldsymbol{\theta}) = \mathcal{R}_\Psi(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h(\boldsymbol{\theta}), Y\right)$
- $M_n(h,\boldsymbol{\theta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\Psi\left(\rho_h(\boldsymbol{\theta}), Y_i\right)$
- $M_m(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h^m(\boldsymbol{\theta}), Y\right)$
- $M_{n,m}(h,\boldsymbol{\theta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\Psi\left(\rho_h^m(\boldsymbol{\theta}), Y_i\right)$
- $\mathbb{G}_n\Psi\left(\rho_h^m(\boldsymbol{\theta})\right) = \sqrt{n}\left(M_{n,m}(h,\boldsymbol{\theta}) - M_m(h,\boldsymbol{\theta})\right)$

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

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}}\; M_{n,m}(h,\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}}\; M(h,\boldsymbol{\theta}). \tag{6.18}$$

We have,

$$\begin{aligned}
&\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}})\\
&= M(h,\widehat{\boldsymbol{\theta}}) - M_m(h,\widehat{\boldsymbol{\theta}}) + M_m(h,\widehat{\boldsymbol{\theta}}) - M_{n,m}(h,\widehat{\boldsymbol{\theta}}) + M_{n,m}(h,\widehat{\boldsymbol{\theta}})\\
&= -\left(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\right) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_{n,m})\right) + \underbrace{M_{n,m}(h,\widehat{\boldsymbol{\theta}}) - M_{n,m}(h,\boldsymbol{\theta}^\star)}_{\leq 0\ (6.18)} + M_{n,m}(h,\boldsymbol{\theta}^\star)\\
&\leq -\left(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\right) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right) + M_{n,m}(h,\boldsymbol{\theta}^\star) - M_m(h,\boldsymbol{\theta}^\star) + M_m(h,\boldsymbol{\theta}^\star)\\
&\leq -\left(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\right) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right) + \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\boldsymbol{\theta}^\star)\right) + M_m(h,\boldsymbol{\theta}^\star)\\
&\leq -\left(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\right) + \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta}^\star)\right) - \Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right)\right)\\
&\quad + M_m(h,\boldsymbol{\theta}^\star) - M(h,\boldsymbol{\theta}^\star) + M(h,\boldsymbol{\theta}^\star)\\
&\leq \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta}^\star)\right) - \Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right)\right) + \left(M_m(h,\boldsymbol{\theta}^\star) - M(h,\boldsymbol{\theta}^\star)\right) - \left(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\right)\\
&\quad + M(h,\boldsymbol{\theta}^\star)\\
&\leq \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2}{\sqrt{n}}\sup_{\boldsymbol{\theta}\in\Theta}\left|\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)\right| + 2\sup_{\boldsymbol{\theta}\in\Theta}\left|M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})\right|
\end{aligned}$$

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

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

Second term. The Lemma 6.5 provides that (a.s.)

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

where $\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{\Psi(\widetilde{\rho}(\lambda),\,\cdot),\,\lambda\in\mathcal{Y}\}$. Thus the second term is bounded by $\frac{2\gamma}{\sqrt{n}}\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$.

Third term. We have

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

By Assumption 4.3, we obtain

$$|M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})| \leq A_\Psi \left\|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\right\|_{\mathcal{F}}, \tag{6.19}$$

for some positive constant A_Ψ.

Moreover, the inequality (3.3) yields

$$\left\|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\right\|_{\mathcal{F}} \leq \left\|\frac{1}{m}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]\right\|_{\mathcal{F}} + B_h^m(\boldsymbol{\theta}). \tag{6.20}$$

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

$$\left\|\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})\right\|_{\mathcal{F}} \leq \frac{1}{\sqrt{m}}\left\|\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta}))\right\|_{\mathcal{F}} + B_h^m(\boldsymbol{\theta}) \tag{6.21}$$

$$\leq \frac{1}{\sqrt{m}}\left(\left\|\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta}))\right\|_{\mathcal{F}} + \sqrt{m}\,B_h^m(\boldsymbol{\theta})\right). \tag{6.22}$$

Taking the *supremum* over Θ and combining the Lemma (6.6) and the Assumption 3.4 gives

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

Hence, in (6.19) we obtain

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

Finally, the following bound holds for the procedure risk

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

Now, let us notice that for any 3 events E_1, E_2, E_3 we have:

$$\mathbb{P}(E_1) \leq \mathbb{P}(E_1\cap E_2\cap E_3) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c). \tag{6.23}$$

We consider the following events

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

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

and

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

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

$$\mathbb{P}_{Y_{1\dots n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde\rho,\Psi)}} \le \bar K^{\varepsilon}_{(\tilde\rho,\Psi)}\right) \ge 1-\varepsilon$$

and

$$\mathbb{P}_{\mathbf{X}_{1\dots m}}\left(\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde\rho,h)}} \le \bar K^{\varepsilon}_{(\tilde\rho,h)}\right) \ge 1-\varepsilon$$

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

Using the inequality (6.23) with the fact that $\mathbb{P}(E_2) = \mathbb{P}_{Y_{1\dots n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde\rho,\Psi)}} \le \bar K^{\varepsilon}_{(\tilde\rho,\Psi)})$ and $\mathbb{P}(E_3) = \mathbb{P}_{\mathbf{X}_{1\dots m}}(\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde\rho,h)}} \le \bar K^{\varepsilon}_{(\tilde\rho,h)})$, we obtain

$$\mathbb{P}(E_1) \le \mathbb{P}_{Y_{1\dots n},\mathbf{X}_{1,\dots,m}}\left(\mathcal{R}_\Psi(h,\widehat{\boldsymbol\theta}) \le \inf_{\boldsymbol\theta\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol\theta)\right) + \frac{2\,\gamma}{\sqrt{n}}\,\bar K^{\varepsilon}_{(\tilde\rho,\Psi)} + 2\,\frac{A_\Psi}{\sqrt{m}}\left(c\,\bar K^{\varepsilon}_{(\tilde\rho,h)} + \sqrt{m}\,B_h(m)\right)\right) + 2\,\varepsilon.$$

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

$$\mathbb{P}_{Y_{1\dots n},\mathbf{X}_{1,\dots,m}}\left(\mathcal{R}_\Psi(h,\widehat{\boldsymbol\theta}) \le \inf_{\boldsymbol\theta\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol\theta)\right) + \frac{2\,\gamma}{\sqrt{n}}\,\bar K^{\varepsilon}_{(\tilde\rho,\Psi)} + 2\,\frac{A_\Psi}{\sqrt{m}}\left(c\,\bar K^{\varepsilon}_{(\tilde\rho,h)} + \sqrt{m}\,B_h(m)\right)\right) \ge 1-2\,\varepsilon.$$

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

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

where

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

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

and

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

That concludes the proof. □

References

[1] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefebvre and E.De Rocquigny, Nonlinear methods for inverse statistical problems. *Comput. Stat. Data Anal.* **55** (2011) 132–142.

[2] P. Billingsley, *Convergence of probability measures*. Wiley New York (1968).

[3] E.de Rocquigny, N. Devictor and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

[4] M.D. Donsker, Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annal. Math. Stat.* (1952) 277–281.

[5] R.M. Dudley, Weak convergence of measures on nonseparable metric spaces and empirical measures on euclidian spaces. *Illinois J. Math.* **11** (1966) 109–126.

[6] P. Gaenssler, Empirical Processes. *Instit. Math. Stat.*, Hayward, CA (1983).

[7] A. Goldenshluger and O. Lepski, Uniform bounds for norms of sums of independent random functions (2009) Preprint: `arXiv:0904.1950`.
[8] P.J. Huber, Robust estimation of a location parameter. *Annal. Math. Stat.* (1964) 73–101.
[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. *Ann. Prob.* **33** (2005) 1060–1077.
[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. *Annal. Stat.* **34** (2006) 2326–2366.
[16] D. Pollard, Empirical processes: theory and applications. *Regional Conference Series in Probability and Statistics Hayward* (1990).
[17] N. Rachdi, J.C. Fort and T. Klein, Stochastic inverse problem with noisy simulator- an application to aeronautic model. *Annal. Facult. Sci. Toulouse* **21**.
[18] T.J. Santner, B.J. Williams and W. Notz, *The design and analysis of computer experiments*. Springer Verlag (2003).
[19] G.R Shorack and J.A Wellner. Empirical processes with applications to statistics. Wiley Series in *Probability and Statistics* (1986).
[20] C. Soize and R. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput.* **26** (2004) 395–410.
[21] M. Talagrand, Sharper bounds for Gaussian and empirical processes. *Annal. Prob.* **22** (1994) 28–76.
[22] S. van de Geer, *Empirical processes in M-estimation*. Cambridge University Press (2000).
[23] A.W. van der Vaart, *Asymptotic statistics*. Cambridge University Press (2000).
[24] A.W. van der Vaart and J.A. Wellner, Weak Convergence and Empirical Processes. Springer Series in *Statistics* (1996).
[25] E. Vazquez. Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. Ph.D. thesis (2005).