Oracle inequalities for new M-estimation and model selection problems

Nabil Rachdi, Jean-Claude Fort, Thierry Klein

Oracle inequalities for new M-estimation and model selection problems

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

Abstract

In this paper, we develop new algorithms for parameter estimation and model selection in the case of models type Input/Output in order to represent and to characterize a phenomenon Y. From experimental data $Y_1, ..., Y_n$ supposed to be i.i.d from Y, we prove oracle inequalities 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.

Introduction

As in many statistical problems, we are interested to investigate the stochastic behaviour of a random variable Y. We have at disposal an i.i.d sample $Y_1, ..., Y_n$. 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, theses variables are not directly related to the "conditions" leading to the "experimental" data $Y_1, ..., Y_n$. Indeed, in our study, we don't suppose having the data $(\mathbf{x}_1, Y_1), ..., (\mathbf{x}_n, Y_n)$ which differs our framework from the classical regression one. That's why we assume that the available data reduced to $Y_1, ..., Y_n$: 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 [5] Research department: the effect of an electromagnetic field on the behaviour of an aircraft. When lightning or an electromagnetic field strike an aircraft, sensors measure data corresponding to the intensity of such

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

[2]We owe thanks to Fabien Mangeant for advice and discussions, research engineer at EADS Innovation Works, Suresnes France

[3]Université Paris Descartes, 45 rue des saints pères, 75006 Paris, France

[4]Institut de Mathématiques de Toulouse, 118 route de Narbonne F-31062 Toulouse

[5]EADS : European Aeronautic Defense and Space Company

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 $Y_1, ..., Y_n$. On another side, we dispose of several computer codes h modeling the lightning phenomenon 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 lightning properties (angles, atmospheric conditions etc...). The uncontrolled variables $\mathbf{x}$ will be modeled by a random variable $\mathbf{X}$.
In this case, 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.
So, shortly speaking, our goal is to construct a *Random Simulator*, $\mathbf{X} \mapsto \hat{h}(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ with $\mathbf{X}$ some random variable, predicting as well as possible the observed data $Y_1, ..., Y_n$. 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* $h(\mathbf{X}, \widehat{\boldsymbol{\theta}}) \sim \mathcal{U}([0,1])$ ($\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_1$ 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" [4]. 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 [14, 28, 20, 21]). 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 [6]. Along the years, the development of empirical processes theory increased successfully thanks to work of many contributors, R.M. Dudley [8], D. Pollard [7], P. Gaenssler [10], Galen R. Shorack and Jon A. Wellner [9] and others. More recently, many references give a general overview of this theory with its applications to statistics, for example [25, 23, 16]. Empirical processes give power tools for evaluating statistical estimation and inference problems. In particular, estimation based on minimizing a function was introduced by Huber in 1964 [12] where he proposed generalizing maximum likelihood estimation. The estimators resulting are called *M-estimator* ("M" for minimizing or maximizing) [13]. 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 [23] or [24] 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 an simulated data. This paper is divided into five parts. In Section 1 we describe our general framework. In Section 2 we establish Theorem 2.1 providing an oracle inequality for inverse problems based on both experimental and simulated data. In Section 3 we discuss about con-

stants in Theorem 2.1. In Section 4 we tackle model selection problem and we prove an oracle inequality. Section 5 is devoted to the proofs of the main results.

1 General setting

1.1 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 $\mathbb{Q}$. Denote by f the associated (Lebesgue) density function. Let assume that $\mathcal{Y} \subset [-M, M]$, $M > 0$.
Suppose that a n-sample $Y_1, ..., Y_n$ 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 $\mathcal{H}$

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

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 $\mathbb{P}^{\mathbf{x}}$ which forms a probability space $(\mathcal{X}, \mathcal{B}, \mathbb{P}^{\mathbf{x}})$. The probability measure $\mathbb{P}^{\mathbf{x}}$ is not supposed to be known, we will only dispose of a sample drawn from this distribution. In the case where $\mathbb{P}^{\mathbf{x}}$ 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

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

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

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

which provides m output *simulated data*

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

Remark 1.1. In practice, the data $\mathbf{X}_1, ..., \mathbf{X}_m$ may either arise from a data base (from experiments etc...) or simply arise from simulations of the random variable $\mathbf{X}$ with known distribution $\mathbb{P}^{\mathbf{x}}$.

In this paper, we develop a general method for estimating the parameter $\boldsymbol{\theta}$ and/or selecting a model h among the family $\mathcal{H}$ based on the *training data*

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

Owing to the *a priori* knowledge one can get, four scenarii are likely to appear, see Table 1, where

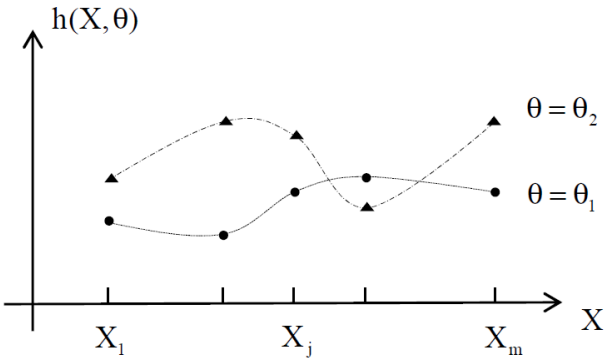

Figure 1: Example of model outputs with 2 different parameters.

- "$\mathcal{H}$" means that one dispose of more than one model
- "h" means that one dispose of a unique model only
- "Θ" means that at least one model is parametric
- "-" means non-parametric

	Model	Parameter
Scenario 1.	$\mathcal{H}$	Θ
Scenario 2.	h	Θ
Scenario 3.	$\mathcal{H}$	-
Scenario 4.	h	-

Table 1: Example of current scenarios.

In this general setting, we will keep the notations $\mathcal{H}$ and Θ even if $\mathcal{H} = \{h\}$ or $\Theta = \{\boldsymbol{\theta}_0\}$. 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 $\boldsymbol{\theta}^\star$ and/or the "true" model $h^\star$. It aims at estimating "physical" parameters having a real signification like dimensions or material properties, for instance.
The Scenario 1 and Scenario 2 (see Table 1) could lead to such problems.

- *Prediction.*
In prediction, one wants to estimate a parameter $\boldsymbol{\theta}^\star$ and/or a model $h^\star$, with $\boldsymbol{\theta}^\star$ and $h^\star$ not necessarily unique, in order to predict the random phenomenon Y. Informally, one hopes that

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

Here, the parameter $\boldsymbol{\theta}^\star$ and the model $h^\star$ may have no real signification. It is the case in models calibration for example.

It seems that prediction and identification provide common techniques but with different objectives. However, these two procedures can be in "conflict" in some cases, see [29].

Roughly, we can think that prediction procedures generalize identification ones, at the risk of restricting the parameter space and the family of models. By this way, through this paper, we will adopt a prediction strategy that includes both calibration and inverse problems.

1.2 Model performance

1.2.1 Motivation

In modeling, one can get a large degree of freedom concerning the models, the parameters, the *a priori* knowledge etc..., and the aimed objective. By objective, we mean a specific feature of the unknown random phenomenon Y: mean, exceeding probability, density function etc...
However, in many cases (EADS applications for example), only a few experimental data are available. But, if a phenomenon is modeled by full parameterized models (cf. Scenario 1 in Table 1 for example), each choice of models and/or parameters may lead to different results, see Figure (1). So, a reasonable strategy would be to find and to work with one model which "represents well" the phenomenon (for us Y) for the aimed objective.
We propose to use experimental data in order to adjust a simulating model: we called it *Random Simulator* in the introduction.

1.2.2 Tools for evaluating the model performance

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

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

Let a random variable W with probability distribution μ. We define a *feature* of the distribution μ as a quantity $\rho(\mu) \in \mathbb{F}$ (or simply ρ if there's no ambiguity), where $\mathbb{F}$ will be called the *feature space*.

Notice that the feature space $\mathbb{F}$ can be either a scalar space (mean, threshold probability, etc...) or a functional space (density distribution, cumulative distribution function).
We equip the feature space $\mathbb{F}$ with the norm $||\cdot||_{\mathbb{F}}$ which be either the absolute value norm $|\cdot|$ when $\mathbb{F} \subset \mathbb{R}$, or a L_r-norm ($r \geq 1$) when $\mathbb{F}$ is a functional space (with functions define on $\mathcal{Y}$).

In all what follows, we denote by

$$\rho(\mathbb{P}_{h,\boldsymbol{\theta}}) := \rho_h(\boldsymbol{\theta})$$

a feature of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$.

Definition 1.1. Contrast function.
We define any function

$$\begin{array}{rccc} \Psi : & \mathbb{F} \times \mathcal{Y} & \longrightarrow & \mathbb{R} \\ & (\rho, y) & \longmapsto & \Psi(\rho, y) \end{array} \tag{1}$$

as a *contrast* function.
For a random variable W, we use the notation $\mathbb{E}_W$ for the expectation under the variable W.
We make the assumption:

Assumption 1.1. 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 $\rho_1, \rho_2 \in \mathbb{F}$

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

with $L_\Psi : \mathcal{Y} \to \mathbb{R}$ 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.

Example 1.1. Some classical features and associated contrasts.

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

  ▷ Mean-Squared contrast

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

- $\mathbb{F} = \{set\ of\ density\ functions\}$

  ▷ log-contrast

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

  ▷ L_2-contrast

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

- etc...

See Table 2, page 12.
Now, we define the *risk function*.

Definition 1.2. Risk function.
Let $\rho_h(\boldsymbol{\theta}) \in \mathbb{F}$ a feature of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$, and let Ψ an associated contrast. The risk function (relative to Ψ) of the couple $(h, \boldsymbol{\theta})$ is defined as

$$\mathcal{R}_\Psi(h, \boldsymbol{\theta}) := \mathbb{E}_Y\, \Psi\left(\rho_h(\boldsymbol{\theta})\,,\, Y\right). \tag{2}$$

Example 1.2. Some classical risk functions.
By elementary calculus, we see that

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

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

- the log-contrast gives the Kullbach-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 $\rho_h(\boldsymbol{\theta}) = \int_{\mathcal{X}} h(x,\boldsymbol{\theta})\, \mathbb{P}^{\mathbf{x}}(dx)$, this integral is not necessarily tractable, even if the probability measure $\mathbb{P}^{\mathbf{x}}$ 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 an analytical form.
This situation is very common in engineering, where complex models exist and are only known through simulations

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

This aspect is the principal motivation of our work.

2 Inverse Problem.

Now, let us fix a model h in the set $\mathcal{H}$.
Our goal is to compute a parameter $\boldsymbol{\theta} \in \Theta$, depending on h, making the risk function $\mathcal{R}_\Psi(h,\boldsymbol{\theta})$ as small as possible.

- *Oracle.*

We want to estimate a parameter $\boldsymbol{\theta}^\star$ minimizing the risk (2), i.e

$$\boldsymbol{\theta}^\star \in \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \mathcal{R}_\Psi(h,\boldsymbol{\theta})\,. \tag{3}$$

In the literature, the parameter $\boldsymbol{\theta}^\star$ is also called the *oracle*. This term was introduce by Donoho and Johnstone [5].
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 uncomputable (hence $\boldsymbol{\theta}^\star$) for two reasons. First, the measure $\mathbb{Q}$ 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 $\boldsymbol{\theta}^\star$, that is

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

In the next we establish an *oracle inequality* of the form

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

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

As $\mathbb{Q}$ is unknown, we replace it by its empirical version

$$\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$$

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

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

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

$$\rho_h^m(\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{X}_j, \boldsymbol{\theta})) \tag{4}$$

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

Example 2.1. Examples of weight functions.

- *Mean-Squared contrast*

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

- log-*contrast or* L_2-*contrast*

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

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.

Remark 2.1. 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 (a kernel for the density), see Figure (2).
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.

For notation convenience, we may use the notation $W_{1..l}$ for a sample $W_1, ..., W_l$ of random variables, and $\mathbb{E}_{W_{1..l}}$ will be the expectation under the joint law of $(W_1, ..., W_l)$.

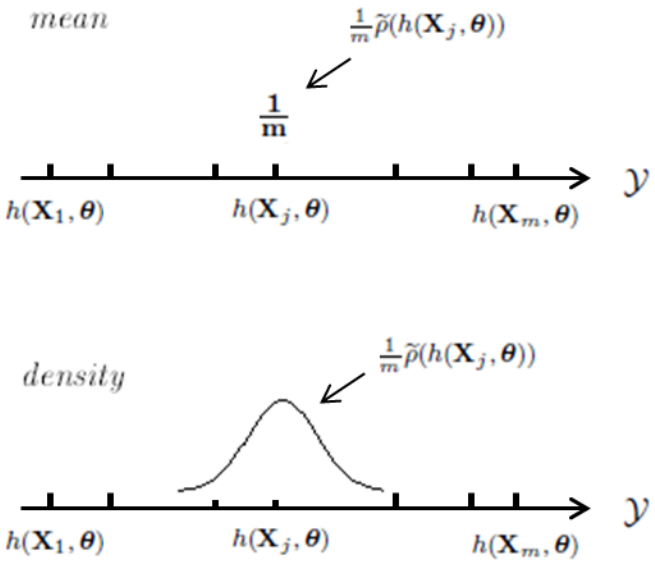

Figure 2: Example of weight function in the case of the mean (top) and the case of the density (bottom).

Definition 2.1. We denote by $\sigma_h^m(\boldsymbol{\theta})$, called *simulation error*, the error committed estimating the feature $\rho_h(\boldsymbol{\theta})$ by the estimator $\rho_h^m(\boldsymbol{\theta})$,

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

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

$$\begin{aligned}
\sigma_h^m(\boldsymbol{\theta}) &= ||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \\
&= ||\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \\
&= ||\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) + \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \\
&\leq ||\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))||_{\mathbb{F}} + ||\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \\
&\leq \left|\left|\frac{1}{m}\sum_{j=1}^{m}[\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))]\right|\right|_{\mathbb{F}} + b_h^m(\boldsymbol{\theta}) \qquad (5)
\end{aligned}$$

with

$$b_h^m(\boldsymbol{\theta}) := ||\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \qquad (6)$$

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

The first term in the right hand side of inequality (5) is a *variance* (random) term, and the second is a *bias* (deterministic) term.
For our statistical analysis, the variability term

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

will play a crucial rule whereas the bias term

$$||\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}}$$

behaves like a parasite term.

Assumption 2.1. We assume that the plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ (4) is uniformly asymptotically unbiased, i.e it exists some constant $b_h(m)$ depending on h and m such that the bias error (6) satisfies

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

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

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

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta}))\,,\,Y_i\right)\,, \tag{7}$$

or

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

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

Example 2.2. Examples of estimators.

- *Mean-Squared contrast*

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

- log-*contrast*

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

- L_2-*contrast*

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

Remark 2.2. 1. The estimator $\widehat{\boldsymbol{\theta}}$ 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 $\widehat{\boldsymbol{\theta}}$, we have to qualify the *quality* of this procedure. It's the topic of the following section.

2.1 Main Result

In this section, we aim at establishing an *oracle inequality* which provides a qualification of the estimation procedure previously defined.
We recall that

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

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

and

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

Now, we give some definitions and notations useful for setting the Theorem 2.1.

Denote by

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

and

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

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

$$\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\widetilde{\rho}(\lambda) , y) , \lambda \in \mathcal{Y}\} , \tag{8}$$

$$\mathcal{P}_{(\widetilde{\rho},h)} = \{\mathbf{x} \in \mathcal{X} \mapsto \widetilde{\rho}(h(\mathbf{x},\boldsymbol{\theta}))(\lambda) , (\boldsymbol{\theta},\lambda) \in \Theta \times \mathcal{Y}\} . \tag{9}$$

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

$$\mathbb{W}\, g := \int g(u)\mathbb{W}(du) \qquad g \in \mathcal{G}$$

and

$$\|\mathbb{W}\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |\mathbb{W}\, g| .$$

	$\mathcal{W}_{(\widetilde{\rho},\Psi)}$	$\mathcal{P}_{(\widetilde{\rho},h)}$	A_Ψ
M-S 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	$y \mapsto -\log\left(K_b(y-\lambda)\right)$, $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto K_b(\lambda - h(\mathbf{x},\boldsymbol{\theta}))$, $(\lambda,\boldsymbol{\theta}) \in \Theta \times \mathcal{Y}$	$\|f\|_2/\eta$
L_2-contrast	$y \mapsto \|\|K_b(\cdot\,-\lambda)\|\|_2 - 2\,K_b(y-\lambda)$, $\lambda \in \mathcal{Y}$	*idem*	$2\left(\|f\|_2 + B\right)$

Table 2: Example of classes of functions and constant A_Ψ (see section (3.1)).

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

$$\begin{aligned}\mathbb{G}_n\, g &= \int_\mathcal{Y} g(u)\mathbb{G}_n(du) \\ &= \sqrt{n}\int_\mathcal{Y} g(u)(\mathbb{Q}_n - \mathbb{Q})(du) \\ &= \frac{1}{\sqrt{n}}\sum_{i=0}^{n}\left(g(Y_i) - \mathbb{E}(g(Y))\right)\,.\end{aligned}$$

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

$$\mathbb{K}_m^{\mathbf{x}}\, g = \frac{1}{\sqrt{m}}\sum_{j=0}^{m}\left(g(\mathbf{X}_j) - \mathbb{E}(g(\mathbf{X}))\right)\,.$$

Remark 2.3. 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 $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ and $\mathcal{G}_\mathcal{X}$ is $\mathcal{P}_{(\widetilde{\rho},h)}$, respectively defined in (8) and (9).

Definition 2.2. Tightness.
Let $(W_l)_{l\geq 1}$ 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

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

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

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

Remark 2.4. The constant $\bar{K}^\varepsilon$ in this definition has to be uniform in $l \geq 1$. However, if $\bar{K}^\varepsilon$ is decreasing with l, we will prefer this constant (which improves the uniform one).

Theorem 2.1. ***Oracle Inequality for Parameter Estimation.***
Under the Assumptions (1.1) and (2.1), suppose that the sequences of random variables $||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$

and $||\mathbb{K}_m^{\mathbf{x}}||_{\mathcal{P}_{(\widetilde{\rho},h)}}$ *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 $\mathbb{F}$ *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*

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

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 2.5. For a fixed weight function $\widetilde{\rho}$, notice that the constant $K^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ depends on the regularity of the contrast function $y \mapsto \Psi(\widetilde{\rho}, y)$, and the constant $K^{\varepsilon}_{(\widetilde{\rho},h)}$ depends on the regularity of the maps $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta})$. Hence, we distinguish the effect of the contrast with those of the model regularity. In Section 3 we provide some examples of constants.

2.2 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 (7) is classically

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

and we can derive immediately an oracle inequality.

Proposition 2.1. ***Basic Oracle Inequality.***
It holds that

$$\mathcal{R}_{\Psi}(h, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_{\Psi}(h, \boldsymbol{\theta})\right) + \frac{2}{\sqrt{n}} \|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_{\Psi}}, \tag{10}$$

where

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

Proof. The proof comes from a classical calculus in M-estimation, see for example [24] (p. 46) . □

Most of statistical procedures, as likelihood, regression, classification etc... can be written like (10). Such procedures have been widely studied with a large literature available. Recently, authors use the Empirical Processes theory (see [23, 24, 25, 16] 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 behaviour 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 [23].
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 (10), with high probability

$$\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{K}{\sqrt{n}}. \tag{11}$$

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 [17, 22, 25, 19].

Inequality (10) can not be applied to our framework because the induced procedure $\widehat{\boldsymbol{\theta}}_n$ involves the quantity $\rho_h(\boldsymbol{\theta})$ intractable for *complex models*.
The result of Theorem (2.1) 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 [17], Billingsley [3]) presents in the 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

$$\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\left(1+\sqrt{\frac{n}{m}}(K^{\varepsilon}_{(\widetilde{\rho},h)}+B_m)\right) \tag{12}$$

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 Section 3.3 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 $\widehat{\boldsymbol{\theta}}$ defined in (7) "mimic" 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 (12). Making $m \longrightarrow +\infty$, this residual becomes simply $\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$ which has the same form as those found in classical cases (11). We find the usual rate of convergence $\sqrt{n}$.
In our purpose, the factor

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

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 (4).

Example 2.3. For unbiased plug-in estimator $\rho_h^m(\boldsymbol{\theta})$, $B_m = 0$, hence, the simulation factor is simply

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

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

$$m = n^\beta \quad or \quad n\left(\log(n)\right)^\beta,$$

we can make the simulation factor close to 1.

Remark 2.6. The term $\inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right)$ in Theorem (2.1) 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 (1.2) and (2.2), we can write the oracle inequality in Theorem (2.1) in specific cases as follows.

Example 2.4. Oracle Inequalities in specific cases.

- Mean-squared contrast

$$\left(\mathbb{E}(Y)-\rho_h(\widehat{\boldsymbol{\theta}}_{MS})\right)^2 \leq \inf_{\boldsymbol{\theta}\in\Theta}\left(\left(\mathbb{E}(Y)-\rho_h(\boldsymbol{\theta})\right)^2\right)+\frac{K^{\varepsilon}_{(\widetilde{\rho},MS)}}{\sqrt{n}}\left(1+\sqrt{\frac{n}{m}}(K^{\varepsilon}_{(\widetilde{\rho},h)}+B_m)\right)$$

  In practice, $B_m=0$.

- log-contrast

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

- L_2-contrast

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

The terms $\inf_{\boldsymbol{\theta}\in\Theta}\left(\left(\mathbb{E}(Y)-\rho_h(\boldsymbol{\theta})\right)^2\right)$, $\inf_{\boldsymbol{\theta}\in\Theta}\left(KL(\rho_h(\boldsymbol{\theta}),f)\right)$ and $\inf_{\boldsymbol{\theta}\in\Theta}\left(\|\rho_h(\boldsymbol{\theta})-f\|_2^2\right)$ are the *ideal risks* $\inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right)$ in different situations. These examples show clearly that these terms represent a "distance" between the "target" and the "best" information available, see Remark 2.6. These terms can be supposed equal to zero, in this case we obtain for example (L_2-contrast)

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

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

3 About the constants in Theorem (2.1)

3.1 Constant A_Ψ

We will show how we obtain the constants A_Ψ in Table (2). Let recall that $\mathcal{Y}\in[-M,M]$.

- *Mean-squared contrast.*
Let $y\in\mathcal{Y}$, $\rho_1,\rho_2\in\mathbb{F}\subset\mathcal{Y}$. We have

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

- log-*contrast.*
Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F}$, with $\mathcal{F} \subset \mathbb{F}$ and $\mathbb{F}$ 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 $\mathbb{Q}$ (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 $\mathcal{F} \subset \mathbb{F}$ and $\mathbb{F}$ 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

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

Taking the expectation under $\mathbb{Q}$ and by Cauchy-Schwarz inequality (as before) yields

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

3.2 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 (6) 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 $\widetilde{\rho}$ 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 a kernel $K()$ and a bandwidth b.
Consider that $\|\cdot\|_{\mathbb{F}} = \|\cdot\|_2$, for all $\boldsymbol{\theta} \in \Theta$ we have

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

Theorem (24.1) in [24] (p. 345) gives the following result.

Theorem 3.1. *Let $\xi_1, ..., \xi_m \in \mathcal{Y}$ an i.i.d sample drawn from a probability density function g and $K : \mathcal{Y} \to \mathbb{R}^+$ some function (kernel). Denote by*

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

If the following assumptions are valid

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

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

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

In particular, the bias term $\|\mathbb{E}_{\xi_{1..m}} \widehat{g} - g\|_2$ 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 $b_h(m) = \sup_{\boldsymbol{\theta} \in \Theta} b_h^m(\boldsymbol{\theta})$.

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

We detail the arguments for computing the constant $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$. The constant $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ will be obtained in the same way.
As defined in Theorem (2.1), these constants are related to the *supremum* of empirical processes.
We will use special cases of Theorem (1.1) in [15].

- *Constant* $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$.

We need to extend the Theorem (1.1) in [15] which deals with countable classes of functions. So, we prove the following proposition.

Proposition 3.1. *Let the empirical process* $\mathbb{G}_n$ *indexed by the class of functions* $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ *(defined in (8)). Supposed that* $\mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) < \infty$ *, it holds that for all* $t \geq 0$

$$\mathbb{P}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}_{Y_{1...n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) + t\right) \quad \leq \quad exp\left(-\frac{t^2}{2\,v + 3\,M\,t/\sqrt{n}}\right) \tag{13}$$

with $v = \sup_{w\in\mathcal{W}_{(\widetilde{\rho},\Psi)}} \mathrm{Var}(w(Y))$.

Proof. Recall that

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

and consider that $\mathcal{Y} = [-M, M]$.
We define the sets $\mathcal{Y}^s = \{y_1^s, ..., y_{i_s}^s\}$ for $s \geq 1$ recursively, as follows:

- $\mathcal{Y}^1 = \{-M, 0, M\}$.
- Assume that the set $\mathcal{Y}^s = \{y_1^s, ..., y_{i_s}^s\}$ is construct, with increasing elements, i.e $y_1^s < ... < y_{i_s}^s$.
  For $j = 1, ..., i_s - 1$, let
  $$\tilde{y}_j^s = \frac{y_j^s + y_{j+1}^s}{2}$$
  and
  $$\widetilde{\mathcal{Y}}^s = \{\tilde{y}_j^s,\, i = 1, ..., i_{s-1} - 1\}\,.$$
- Define
  $$\mathcal{Y}^{s+1} = \mathcal{Y}^s \cup \widetilde{\mathcal{Y}}^s$$
  with increasing elements.

Remark 3.1. One can verify that

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

Now, define the classes of functions

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

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

$$\mathcal{W}^{s-1}_{(\widetilde{\rho},\Psi)} \subsetneq \mathcal{W}^{s}_{(\widetilde{\rho},\Psi)} \subsetneq \mathcal{W}_{(\widetilde{\rho},\Psi)}\,. \tag{14}$$

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

$$\overline{\lim_{s\to\infty} \mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} = \overline{\bigcup_{s\geq 1} \mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} = \mathcal{W}_{(\widetilde{\rho},\Psi)}. \tag{15}$$

The classes of functions $\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}$, $s \geq 1$, are countable ($2^s + 1$ elements) with values in $[-M, M]$. Finally, we apply Theorem (1.1) in [15] to the classes $\frac{1}{M}\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}$, we get for all $t \geq 0$ and $s \geq 1$

$$\mathbb{P}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}) + t\right) \;\leq\; exp\left(-\frac{t^2}{2\,v_s + 3\,M\,t/\sqrt{n}}\right) \tag{16}$$

where $v_s = \sup_{w\in\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} \mathrm{Var}(w(Y))\,.$
We wish to prove that the left and right member of this last inequality converge when $s \to \infty$. Write the left member as follows

$$\begin{aligned}
&\mathbb{P}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}) + t\right)\\
=\;& \mathbb{E}_{Y_{1..n}}\left(\mathbb{1}_{\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}) + t}\right)\\
=\;& \mathbb{E}_{Y_{1..n}}\left(\mathbb{1}_{\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} - \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}) \geq t}\right)
\end{aligned} \tag{17}$$

The inclusions (14) yields

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

so the sequence $\left(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}\right)_{s\geq 1}$ is increasing and bounded, thus it converges. By monotone convergence, we obtain that the sequence $\left(\mathbb{E}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}\right)\right)_{s\geq 1}$ converges too provided that $\mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) < \infty$. Thus, the sequence $\left(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}} - \mathbb{E}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}^{s}_{(\widetilde{\rho},\Psi)}}\right)\right)_{s\geq 1}$ converges too, and by dominated convergence the quantity (17) converges to the wanted limit

$$\mathbb{E}_{Y_{1..n}}\left(\mathbb{1}_{\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} - \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) \geq t}\right) = \mathbb{P}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) + t\right).$$

For the right member of (16). By similar arguments, one can check that the sequence $(v_s)_{s\geq 1}$ is increasing and bounded, thus it converges. The limit is $v = \sup_{w\in\mathcal{W}_{(\widetilde{\rho},\Psi)}} \mathrm{Var}(w(Y))\,.$ That concludes the proof.

□

The function $t \mapsto exp\left(-\frac{t^2}{2\,v + 3\,M\,t/\sqrt{n}}\right)$ is decreasing from $\mathbb{R}^*_+$ into $]0,1[$, hence it exists a unique function $\kappa :]0,1[\to \mathbb{R}^*_+$ such that

$$\forall\, t \geq 0 \quad \kappa^{-1}(t) = exp\left(-\frac{t^2}{2\,v + 3\,M\,t/\sqrt{n}}\right)\,. \tag{18}$$

(Note that for all $\epsilon > 0$, $\kappa(\epsilon)$ decreases with n).
Then, we can write (13) as follows, for all $\varepsilon \in]0,1[$

$$\mathbb{P}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) + \kappa(\varepsilon)\right) \leq \varepsilon$$

or equivalently

$$\mathbb{P}_{Y_{1..n}}\left(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \leq \mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) + \kappa(\varepsilon)\right) \geq 1 - \varepsilon\,.$$

Thus, one could take the constant $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ equal to

$$\mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) + \kappa(\varepsilon)\,,$$

but the quantity $\mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}})$ seems to be 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 μ 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

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

The three following definitions are from [25] (p. 83-85).

Definition 3.1. $L_2(\mu)$ **Covering numbers and Entropy.**
The covering number $N(\varepsilon,\, \mathcal{G},\, L_2(\mu))$ is the minimal number of balls $\{j\,,\, \|j - g\|_{2,\mu} < \epsilon\}$ 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 3.2. $L_2(\mu)$ **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 $||u - l||_{2,\mu} < \epsilon$. The bracketing number $N_{[]}(\epsilon,\, \mathcal{G},\, L_2(\mu))$ 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 3.3. $L_2(\mu)$ **Bracketing integral.**
The bracketing integral is defined as

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

Now we apply Theorem (19.35) of [24] (p. 288), it holds that

$$\mathbb{E}_{Y_{1..n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) \quad \leq \quad a_1\, J_{[]}\left(\|W\|_{2,\mathbb{Q}},\, \mathcal{W}_{(\widetilde{\rho},\Psi)},\, L_2(\mathbb{Q})\right)\,, \tag{19}$$

where

- a_1 is some universal constant
- $W : \mathcal{Y} \to \mathbb{R}$ is an envelop function of $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ and

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

Remark 3.2. The quantity $J_{[]}\left(\|W\|_{2,\mathbb{Q}},\, \mathcal{W}_{(\widetilde{\rho},\Psi)},\, L_2(\mathbb{Q})\right)$ is computable if one has the bracketing numbers $N_{[]}(\epsilon,\, \mathcal{W}_{(\widetilde{\rho},\Psi)},\, L_2(\mathbb{Q}))$ $(\forall\, \epsilon > 0)$, see examples in Section 3.4 below.

Finally, setting

$$\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} = a_1\, J_{[]}\left(\|W\|_{2,\mathbb{Q}},\, \mathcal{W}_{(\widetilde{\rho},\Psi)},\, L_2(\mathbb{Q})\right) + \kappa(\varepsilon) \tag{20}$$

provides the claimed constant.

Remark 3.3. Since the quantity $\kappa(\varepsilon)$ decreases with n, $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ too, and the assumptions of Theorem (2.1) are satisfied.

- *Constant* $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$.

By the same argument, let consider the empirical process $\mathbb{K}^{\mathbf{x}}_m$ and the class of functions $\mathcal{P}_{(\widetilde{\rho},h)}$ (defined in (9)) with an envelop noted $P : \mathcal{X} \to \mathbb{R}$.
We obtain

$$\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)} = a_2\, J_{[]}\left(\|P\|_{2,\mathbb{Q}},\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(\mathbb{Q})\right) + \kappa(\varepsilon)\,, \tag{21}$$

where a_2 is some universal constant and κ is defined in (18).

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

3.4.1 $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ for the Mean-squared contrast

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 $W = 4\,M^2$. Then, we have

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

with $F(y) = |2y + 2\,M|$, and by Theorem (2.7.11) in [25] (p. 164) it holds that

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

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

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

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

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

Now, we compute the bracketing integral

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

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

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

Finally, we obtain the following constant

$$\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} = 8\,a_1\,\sqrt{\pi}\,M^2 + \kappa(\varepsilon)\,. \tag{22}$$

3.4.2 $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ 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 denoted by P an envelop of $\mathcal{P}_{(\widetilde{\rho},h)}$, 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

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

where

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

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

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

We aim at computing the entropy integral $J_{[]}\left(\|P\|_{2,\mathbb{Q}},\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(\mathbb{Q})\right)$ by integrating the entropy $\log N_{[]}(\epsilon,\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(\mathbb{Q}))$.
Corollary (2.7.2) in [25] (p. 157) gives an entropy bound for the Hölder space $\mathbb{H}(\mathcal{X}, \alpha, 1)$:

$$\log N_{[]}(\epsilon,\, \mathbb{H}(\mathcal{X}, \alpha, 1),\, L_2(\mathbb{Q})) \leq K \left(\frac{1}{\epsilon}\right)^{d/\alpha} \qquad \forall\, \epsilon > 0\,, \tag{23}$$

where K depends on α, $\mathrm{diam}(\mathcal{X})$ and d.
We supposed that $\mathcal{P}_{(\widetilde{\rho},h)} \subset \mathbb{H}(\mathcal{X}, \alpha, L)$, and one can easily check that $\mathbb{H}(\mathcal{X}, \alpha, L) = L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1)$, where

$$L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1) = \{L\, g : g \in \mathbb{H}(\mathcal{X}, \alpha, 1)\}\,. \tag{24}$$

Remark 3.4. If $\mathcal{P}_{(\widetilde{\rho},h)} \subset \mathbb{H}(\mathcal{X}, \alpha, L)$, 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 3.1.

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

Proof. The first equality is clear by (24). Let $([l_i, u_i])_{i=1\ldots N}$ be a set of ϵ-brackets covering $\mathbb{H}(\mathcal{X}, \alpha, 1)$. Then the brackets $([L\, l_i, L\, u_i])_{i=1\ldots N}$ cover $L \cdot \mathbb{H}(\mathcal{X}, \alpha, 1)$ since for $g \in \mathbb{H}(\mathcal{X}, \alpha, 1)$

$$l \leq g \leq u \Longrightarrow L\, l \leq L\, g \leq L\, u\,.$$

Finally, the brackets $[L\, l_i, L\, u_i]$ are of size $L\, \epsilon$, and the result follows. □

Using (23), Lemma 3.1 and the inequality

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

it holds for $d < 2\alpha$

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

hence

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

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

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

Remark 3.5. 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 3.6. The computation of the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ 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.

4 Model selection

4.1 Results

In the previous section, for each fixed model h in $\mathcal{H}$, we computed a parameter $\widehat{\boldsymbol{\theta}}_h := \widehat{\boldsymbol{\theta}}$ depending on the model h. Thus we have constructed a family of models

$$\{h(\,\cdot\,,\widehat{\boldsymbol{\theta}}_h)\,,\, h\in\mathcal{H}\}\,,$$

with

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

where we recall that $\mathcal{H}$ is a set of functions $\mathcal{Z}\longrightarrow\mathcal{Y}$, with $\mathcal{Z}\subset\mathbb{R}^{d+k}$.

In this section, we establish an oracle inequality qualifying the model selection procedure we propose. Some results of this section are inspired from the previous section.
Recall that the risk function is defined as it follows

$$\mathcal{R}_\Psi(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h(\boldsymbol{\theta}),Y\right) \quad \textit{for all} \quad (h,\boldsymbol{\theta})\in\mathcal{H}\times\Theta\,.$$

Here, $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_h$, and the risk is simply function of h: $\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}_h)$.

An ideal selection procedure would be to choose the *oracle*

$$h^\star = \underset{h\in\mathcal{H}}{\mathrm{Argmin}}\,\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}_h)\,.$$

Recall that

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

- *First approach.*
In the same spirit of the section 2, it seems natural to propose the following selection procedure inspired from (7)

$$\widehat{h} = \underset{h\in\mathcal{H}}{\mathrm{Argmin}}\,\frac{1}{n}\sum_{i=1}^{n}\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_h)\,,\,Y_i\right)\,. \tag{25}$$

Let introduce the quantities

$$D(h) = 2\,A_\Psi\,c\left(\sup_{\lambda\in\mathcal{Y}}\left|\frac{1}{m}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j,\widehat{\boldsymbol{\theta}}_h))(\lambda)-\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\widehat{\boldsymbol{\theta}}_h))(\lambda)\right]\right| + \frac{b_h(m)}{c}\right) \tag{26}$$

and

$$D(\mathcal{H}) = \sup_{h\in\mathcal{H}}\,D(h)\,. \tag{27}$$

Lemma 4.1. *Under the Assumptions (1.1) and (2.1), suppose that the sequence of random variables* $\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$ *is tight and let the constant* $K^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ *defined in Theorem 2.1.*
Let the feature space $\mathbb{F}$ *equipped with either the absolute value norm or some* L_r *norm.*
Then, with probability at least $1-\varepsilon$

$$\mathcal{R}_\Psi(\widehat{h},\widehat{\boldsymbol{\theta}}_{\widehat{h}}) \leq \inf_{h\in\mathcal{H}}\Big(\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}_h)\Big) + D(\mathcal{H}) + \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\,,$$

where the constants A_Ψ *and* c *are those of Theorem (2.1).*

Remark 4.1. 1. The term $\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$ depends only on the number of *experimental data* $Y_1,...,Y_n$ (through n), on the contrast Ψ and the weight function $\widetilde{\rho}$ considered (through $K^{\varepsilon}_{(\widetilde{\rho},\Psi)}$). Thus, it doesn't depend on the models $h\in\mathcal{H}$, and roughly speaking, it appears that it can't be reduced.
2. The term $D(\mathcal{H})$ depends on the "richness" of the class $\mathcal{H}$. This term acts as a *penalization* term.

Finally, the selection procedure (25) provides an oracle inequality with a residual term

$$D(\mathcal{H}) + \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\,,$$

where $D(\mathcal{H})$ can be non-negligible.

Can we perform the selection procedure (25) ?

The problem we meet here can be understood following the so-called *structural risk minimization*, see [27] or [26] for instance. Indeed, the quantity

$$\frac{1}{n}\sum_{i=1}^{n}\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_h)\,,\,Y_i\right)$$

estimates the risk $\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}_h)$ up to a *complexity term*. We will see that if we take into account the *complexity* of the models $h\in\mathcal{H}$ in the procedure (25), we can improve it.

- *Penalized selection.*
The penalization in model selection has been first proposed by Akaike [1] and Mallows [18] in specific cases. Since, many authors have contributed to remarkable developments in the area of *Model selection and penalization*, we cite [19] with the associated references. We recall here that our goal is not to improve some well established results in model selection, but to link the model selection theory to our framework. For this, we adopt the philosophy of [19] without getting into deep details, which would lead to considerations far from our topic.

Let some *penalty function* $\mathrm{pen}:\mathcal{H}\longrightarrow\mathbb{R}^+$ and consider the following selection procedure

$$\widehat{h} = \operatorname*{Argmin}_{h\in\mathcal{H}}\ \frac{1}{n}\sum_{i=1}^{n}\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_h)\,,\,Y_i\right) + \mathrm{pen}(h)\,, \tag{28}$$

where we simply add the penalty function pen to the procedure (25).

Theorem 4.1. ***Oracle Inequality for Model Selection.***
Let the penalty function

$$\widetilde{\text{pen}}(h) = \mathbb{E}_Y\left(\Psi\left(\rho_h(\widehat{\boldsymbol{\theta}}_h), Y\right) - \Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_h), Y\right)\right) \quad \textit{for all} \quad h \in \mathcal{H}\,,$$

and assume that

$$\forall\, h \in \mathcal{H} \quad \text{pen}(h) \geq \widetilde{\text{pen}}(h)\,.$$

Moreover, we suppose that the sequence of random variables $||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$ *is tight. Let the constant* $K^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ *defined in Theorem 2.1.*
Then, it holds with probability at least $1-\varepsilon$

$$\mathcal{R}_\Psi(\widehat{h}, \widehat{\boldsymbol{\theta}}_{\widehat{h}}) \leq \inf_{h\in\mathcal{H}}\left(\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}_h) + (\text{pen} - \widetilde{\text{pen}})(h)\right) + \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\,.$$

Remark 4.2. 1. Suppose that $\text{pen} \simeq \widetilde{\text{pen}}$, then the penalized selection would have good performance since the residual term would be close to $\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$. This is better than $D(\mathcal{H}) + \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$ obtained using the (unpenalized) procedure (25).
2. For given experimental data $Y_1, ..., Y_n$, the term $\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$ can't be reduced (see Remark (4.1)), hence if the penalty function pen is well chosen, the procedure (28) certainly improves (25).
3. It could seem surprising that the Theorem (4.1) doesn't need Assumptions (1.1) and (2.1) like the Lemma (4.1) previously set. In fact, the assumption on the penalty function pen in this proposition is strong enough. The whole problem will be to evaluate the penalty function $\widetilde{\text{pen}}$ which depends on the probability measure $\mathbb{Q}$ unknown, and on the feature $\rho_h(\widehat{\boldsymbol{\theta}}_h)$ uncomputable.

The main difficulty is to find a "good" penalty function pen, i.e satisfying

- $\forall\, h \in \mathcal{H} \quad \text{pen}(h) \geq \widetilde{\text{pen}}(h)$
- $\text{pen} \simeq \widetilde{\text{pen}}\,.$

These conditions on the penalty function are roughly those in [19]. Recent developments deal with the penalization choice (data-driven construction), in the case of least-squares regression, [2]. We won't carry on the penalization calibration aspect here, it will be considered in a further work.
Using Assumptions (1.1) and (2.1), one can easily check that,

$$\widetilde{\text{pen}}(h) \leq \frac{D(h)}{2}\,, \quad \textit{for all} \quad h \in \mathcal{H}\,, \tag{29}$$

where $D(h)$ is defined in (26).
Notice that $\sup_{h\in\mathcal{H}} D(h) = D(\mathcal{H})$ where $D(\mathcal{H})$ is defined in Lemma (4.1).
Then, a possible candidate for pen would be

$$\text{pen}(h) = \frac{D(h)}{2}\,,$$

which satisfies $\text{pen}(h) \geq \widetilde{\text{pen}}(h)$ according (29).
Finally, we propose the following selection procedure

$$\widehat{h} = \underset{h\in\mathcal{H}}{\text{Argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_h)\,,\, Y_i\right) + \frac{D(h)}{2}\,. \tag{30}$$

Example 4.1. For the weight $\widetilde{\rho}(y) = y$ in the Mean Squared framework.
In this case, $A_\Psi = 4\,M$, $c = 1$ and $b_h(m) = 0$. We obtain the following penalty function

$$\begin{aligned} \text{pen}(h) &= 4\,M \left| \frac{1}{m} \sum_{j=1}^{m} \Big(h(\mathbf{X}_j, \widehat{\boldsymbol{\theta}}_h) - \mathbb{E}_{\mathbf{X}}\, h(\mathbf{X}, \widehat{\boldsymbol{\theta}}_h) \Big) \right| \\ &= 4\,M\, |(\mathbb{P}_m^{\mathbf{x}} - \mathbb{P}^{\mathbf{x}})(h_{\mathcal{X}})| \,, \end{aligned}$$

where $h_{\mathcal{X}}(\mathbf{x}) = h(\mathbf{x}, \widehat{\boldsymbol{\theta}}_h)$.

5 Proofs

In order to to prove the *oracle inequality* of Theorem (2.1), we need the following lemmas.

5.1 Preliminary lemmas

Lemma 5.1. *Consider the random functions*

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

We have (a.s.)

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

where $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ is defined in (8).

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 $\mathbf{X} = \mathbf{x}_0$

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

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

Remark 5.1. The left member of the inequality in the lemma (5.1) 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 5.2. *Consider the $\mathbb{P}^{\mathbf{x}}$-empirical process $\mathbb{K}_m^{\mathbf{x}}$ and let $\|\cdot\|_{\mathbb{F}} = |\cdot|$ or $\|\cdot\|_r$ and define*

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

We have

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

where $\mathcal{P}_{(\widetilde{\rho},h)}$ is defined in (9).

Proof. Let notice that the quantity

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

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

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

Remark 5.2. 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 $\mathcal{Y}$.

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

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

we have

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

Finally, notice that

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

and the result follows. □

Remark 5.3. In the case where the weight function is a kernel $K_b(\cdot - \cdot)$, the quantity

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

is treated as a sum of independent random functions in the recent work of A. Goldenshluger and O. Lepski [11]. 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 [11].

5.2 Proof of Theorem (2.1)

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 recall that

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\; M_{n,m}(h,\boldsymbol{\theta}) \tag{31}$$

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\,(31)} + 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) \\
& + 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) \\
& + 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. Since $\rho_h^m(\boldsymbol{\theta}) = \dfrac{1}{m}\sum_{j=1}^m \widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta}))$ and $\rho \longmapsto \Psi(\rho,y)$ is convex by Assumption (1.1), we have the inequality
for all $y \in \mathcal{Y}$,

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

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

$$\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right) \leq \frac{1}{m}\sum_{j=1}^m \mathbb{G}_n\Psi\left(\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta}))\right)\,. \tag{32}$$

By Lemma (5.1) we have (a.s)

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

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

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

Third term. We have

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

By Assumption (1.1)

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

then

$$\left|M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})\right| \leq ||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}}\,\mathbb{E}_Y L_\Psi(Y)\,. \tag{33}$$

Let $A_\Psi = \mathbb{E}_Y L_\Psi(Y)$.
Moreover, the inequality (5) yields

$$||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \leq \left|\left|\frac{1}{m}\sum_{j=1}^m\left[\widetilde{\rho}(h(\mathbf{X}_j,\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]\right|\right|_{\mathbb{F}} + b_h^m(\boldsymbol{\theta})\,. \tag{34}$$

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

$$||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathbb{F}} \leq \frac{1}{\sqrt{m}} ||\mathbb{K}_m^{\mathbf{x}} \widetilde{\rho}(h(\cdot, \boldsymbol{\theta}))||_{\mathbb{F}} + b_h^m(\boldsymbol{\theta}) \tag{35}$$

$$\leq \frac{1}{\sqrt{m}} \left(||\mathbb{K}_m^{\mathbf{x}} \widetilde{\rho}(h(\cdot, \boldsymbol{\theta}))||_{\mathbb{F}} + \sqrt{m}\, b_h^m(\boldsymbol{\theta}) \right) . \tag{36}$$

Taking the *supremum* over Θ and combining the Lemma (5.2) and the Assumption (2.1) gives

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

Hence, in (33) we obtain

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

Finally, the following bound holds for the procedure risk

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

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

$$\mathbb{P}(E_1) \leq \mathbb{P}(E_1 \cap E_2 \cap E_3) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c) . \tag{37}$$

Take the following events

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

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

and

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

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

$$\mathbb{P}_{Y_{1\ldots n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}) \geq 1 - \varepsilon$$

and

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

respectively (for all $\varepsilon > 0$).
Using the inequality (37) with the fact that $\mathbb{P}(E_2) = \mathbb{P}_{Y_{1\ldots n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)})$ and $\mathbb{P}(E_3) = \mathbb{P}_{\mathbf{X}_{1\ldots m}}(\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},h)})$, we obtain

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

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

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

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

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

where

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

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

and

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

That concludes the proof. □

5.3 Proofs of Theorem 4.1

Proof. (Lemma (4.1)). For this proof, consider the notations in the proof of Theorem (4.1) and develop the same calculus than in the proof of Theorem (2.1). □

Proof. (Theorem (4.1)). This proof is similar to the beginning of the proof of Theorem 2.1.

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

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

$$\widehat{h} = \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} M_{n,m}(h) + \mathrm{pen}(h) \tag{38}$$

and

$$h^* = \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} M(h)\,. \tag{39}$$

We have,

$$
\begin{aligned}
& \mathcal{R}_\Psi(\widehat{h}, \widehat{\boldsymbol{\theta}}_h) \\
= \; & M(\widehat{h}) - M_m(\widehat{h}) + M_m(\widehat{h}) - M_{n,m}(\widehat{h}) + M_{n,m}(\widehat{h}) \\
= \; & -\left(M_m(\widehat{h}) - M(\widehat{h})\right) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho^m_{\widehat{h}}(\widehat{\boldsymbol{\theta}}_h)\right) + \underbrace{M_{n,m}(\widehat{h}) + \mathrm{pen}(\widehat{h}) - M_{n,m}(h^\star) - \mathrm{pen}(h^\star)}_{\leq 0 \,(38)} \\
& + \mathrm{pen}(h^\star) - \mathrm{pen}(\widehat{h}) + M_{n,m}(h^\star) \\
\leq \; & \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\Psi\left(\rho^m_{h^\star}(\widehat{\boldsymbol{\theta}}_h)\right) - \Psi\left(\rho^m_{\widehat{h}}(\widehat{\boldsymbol{\theta}}_h)\right)\right) - \left(\mathrm{pen}(\widehat{h}) - \left(M(\widehat{h}) - M_m(\widehat{h})\right)\right) \\
& + \left(\mathrm{pen}(h^\star) - (M(h^\star) - M_m(h^\star))\right) + M(h^\star)
\end{aligned}
$$

Let $\widetilde{\mathrm{pen}}(h) = M(h) - M_m(h) = \mathbb{E}_Y\left(\Psi\left(\rho_h(\widehat{\boldsymbol{\theta}}_h), Y\right) - \Psi\left(\rho^m_h(\widehat{\boldsymbol{\theta}}_h), Y\right)\right)$, by the assumption

$$
\forall\, h \in \mathcal{H} \quad \mathrm{pen}(h) \geq \widetilde{\mathrm{pen}}(h)\,.
$$

We have

$$
-\left(\mathrm{pen}(\widehat{h}) - \left(M(\widehat{h}) - M_m(\widehat{h})\right)\right) \leq 0\,.
$$

Thus, writing

$$
\left(\mathrm{pen}(h^\star) - (M(h^\star) - M_m(h^\star))\right) + M(h^\star) = \inf_{h\in\mathcal{H}}\left(\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}_h) + (\mathrm{pen} - \widetilde{\mathrm{pen}})(h)\right)\,,
$$

we obtain

$$
\mathcal{R}_\Psi(\widehat{h}, \widehat{\boldsymbol{\theta}}_{\widehat{h}}) \leq \inf_{h\in\mathcal{H}}\left(\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}_h) + (\mathrm{pen} - \widetilde{\mathrm{pen}})(h)\right) + \frac{2}{\sqrt{n}}\,||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}\,,
$$

and with probability at least $1-\varepsilon$

$$
\mathcal{R}_\Psi(\widehat{h}, \widehat{\boldsymbol{\theta}}_{\widehat{h}}) \leq \inf_{h\in\mathcal{H}}\left(\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}_h) + (\mathrm{pen} - \widetilde{\mathrm{pen}})(h)\right) + \frac{K^\varepsilon_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\,,
$$

that concludes the proof. □

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory*, pages 267–281, 1973.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10:245–279, 2009.

[3] P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.

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

[5] D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[6] 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.

[7] D.Pollard. Empirical processes: theory and applications. *Regional Conference Series in Probability and Statistics Hayward*, 1990.

[8] 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.

[9] J. A. Wellner G. R. Shorack. *Empirical processes with applications to statistics.* Wiley Series in Probability and Statistics, 1986.

[10] P. Gaenssler. *Empirical Processes.* Institute of Mathematical Statistics, Hayward, CA, 1983.

[11] A. Goldenshluger and O. Lepski. Uniform bounds for norms of sums of independent random functions. *Arxiv preprint arXiv:0904.1950*, 2009.

[12] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[13] P.J. Huber. *Robust statistics.* Wiley-Interscience, 1981.

[14] J.P.C. Kleijnen. *Design and analysis of simulation experiments.* Springer Verlag, 2007.

[15] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of probability*, 33(3):1060–1077, 2005.

[16] M. R. Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer series in statistics, 2008.

[17] M. Ledoux. *The concentration of measure phenomenon.* AMS, 2001.

[18] CL Mallows. Some comments on CP. *Technometrics*, (15):661–675, 1973.

[19] P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003.* Springer Verlag, 2007.

[20] T.J. Santner, B.J. Williams, and W. Notz. *The design and analysis of computer experiments.* Springer Verlag, 2003.

[21] 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.

[22] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

[23] S. Van De Geer. *Empirical processes in M-estimation.* Cambridge University Press, 2000.

[24] AW. Van der Vaart. *Asymptotic statistics.* Cambridge University Press, 2000.

[25] AW. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer Series in Statistics, 1996.

[26] V. Vapnik. *Statistical learning theory.* Wiley New York, 1998.

[27] V. Vapnik and A. Chervonenkis. *Theory of pattern recognition.* Nauka, Moscow, 1974.

[28] E. Vazquez. (PhD thesis) Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. 2005.

[29] Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.