Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ann. I. H. Poincaré – PR 42 (2006) 273–325
ANNALES DE L'INSTITUT HENRI POINCARÉ PROBABILITÉS ET STATISTIQUES
www.elsevier.com/locate/anihpb

Model selection via testing: an alternative to (penalized) maximum likelihood estimators

Lucien Birgé

UMR 7599 "Probabilités et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Paris VI, 4, place Jussieu, 75252 Paris cedex 05, France

Received 9 July 2003; received in revised form 28 February 2005; accepted 12 April 2005

Available online 18 November 2005

Abstract

This paper is devoted to the definition and study of a family of model selection oriented estimators that we shall call T-estimators ("T" for tests). Their construction is based on former ideas about deriving estimators from some families of tests due to Le Cam [L.M. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 1 (1973) 38–53 and L.M. Le Cam, On local and global properties in the theory of asymptotic normality of experiments, in: M. Puri (Ed.), Stochastic Processes and Related Topics, vol. 1, Academic Press, New York, 1975, pp. 13–54] and Birgé [L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 65 (1983) 181–237, L. Birgé, Sur un théorème de minimax et son application aux tests, Probab. Math. Statist. 3 (1984) 259–282 and L. Birgé, Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées, Ann. Inst. H. Poincaré Sect. B 20 (1984) 201–223] and about complexity based model selection from Barron and Cover [A.R. Barron, T.M. Cover, Minimum complexity density estimation, IEEE Trans. Inform. Theory 37 (1991) 1034–1054].

It is well-known that maximum likelihood estimators and, more generally, minimum contrast estimators do suffer from various weaknesses, and their penalized versions as well. In particular they are not robust and they require restrictive assumptions on both the models and the underlying parameter set to work correctly. We propose an alternative construction, which derives an estimator from many simultaneous tests between some probability balls in a suitable metric space. In many cases, although not in all, it results in a penalized M-estimator restricted to a suitable countable set of parameters.

On the one hand, this construction should be considered as a theoretical rather than a practical tool because of its high computational complexity. On the other hand, it solves many of the previously mentioned difficulties provided that the tests involved in our construction exist, which is the case for various statistical frameworks including density estimation from i.i.d. variables or estimating the mean of a Gaussian sequence with a known variance. For all such frameworks, the robustness properties of our estimators allow to deal with minimax estimation and model selection in a unified way, since bounding the minimax risk amounts to performing our method with a single, well-chosen, model. This results, for those frameworks, in simple bounds for the minimax risk solely based on some metric properties of the parameter space. Moreover the method applies to various statistical frameworks and can handle essentially all types of models, linear or not, parametric and non-parametric, simultaneously. It also provides a simple way of aggregating preliminary estimators.

From these viewpoints, it is much more flexible than traditional methods and allows to derive some results that do not presently seem to be accessible to them.

E-mail address: lb@ccr.jussieu.fr (L. Birgé).

0246-0203/$ – see front matter
doi:10.1016/j.anihpb.2005.04.004

Résumé

Cet article est consacré à la défintion et à l'étude d'une classe d'estimateurs, que nous appellerons T-estimateurs ("T" pour test), destinés à faire de la sélection de modèle. Leur construction se fonde sur d'anciennes méthodes de fabrication d'estimateurs à partir de tests dues à Le Cam [L.M. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 1 (1973) 38–53 et L.M. Le Cam, On local and global properties in the theory of asymptotic normality of experiments, in: M. Puri (Ed.), Stochastic Processes and Related Topics, vol. 1, Academic Press, New York, 1975, pp. 13–54] et Birgé [L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 65 (1983) 181–237, L. Birgé, Sur un théorème de minimax et son application aux tests, Probab. Math. Statist. 3 (1984) 259–282 et L. Birgé, Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées, Ann. Inst. H. Poincaré Sect. B 20 (1984) 201–223] et sur des idées de Barron et Cover [A.R. Barron, T.M. Cover, Minimum complexity density estimation, IEEE Trans. Inform. Theory 37 (1991) 1034–1054] à propos de l'utilisation de notions de complexité pour faire de la sélection de modèle.

Il est bien connu que les estimateurs du maximum de vraisemblance et, plus généralement, les estimateurs par minimum de contraste, souffrent de diverses limitations de même que leurs versions pénalisées. Parmi celles-ci, on peut noter qu'ils ne sont généralement pas robustes et ne donnent de bons résultats que moyennant des hypothèses restrictives portant à la fois sur les modèles et sur l'ensemble des paramètres. Nous proposons ici une construction alternative à partir d'une famille de tests entre les boules de l'espace des probabilités muni d'une métrique convenable. Dans un certain nombre de situations, l'estimateur obtenu n'est autre qu'un M-estimateur pénalisé défini sur un certain ensemble dénombrable de paramètres.

Cette construction doit être considérée davantage comme un outil théorique que pratique, compte-tenu de sa complexité numérique, mais elle permet de régler la plupart des problèmes précités dès que les tests robustes requis existent, ce qui est le cas dans divers problèmes statistiques tels que l'estimation d'une densité à partir d'un échantillon ou l'estimation de la moyenne d'une suite de variables gaussiennes indépendantes de même variance connue. Dans de telles situations, les propriétés de robustesse de nos estimateurs permettent de traiter simultanément les problèmes de minimax et de sélection de modèle dans la mesure où l'évaluation du risque minimax revient à utiliser notre méthode sur un modèle unique, convenablement choisi. Nous obtenons alors des bornes du risque minimax qui ne dépendent que de la structure métrique de l'espace des paramètres. Cette construction s'applique à des problèmes statistiques variés et permet de considérer divers types de modèles, linéaires ou non, paramétriques ou non, simultanément. La même construction permet également de sélectionner ou combiner divers estimateurs préliminaires.

Pour toutes ces raisons, notre méthode est bien plus flexible que les méthodes traditionnelles et permet en particulier d'obtenir certains résultats qui ne semblent pas leur être actuellement accessibles.

MSC: primary 62F35; secondary 62G05

Keywords: Maximum likelihood; Robustness; Robust tests; Metric dimension; Minimax risk; Model selection; Aggregation of estimators

1. Introduction

1.1. Some motivations

The starting point for this paper has been the well-known fact that the celebrated maximum likelihood estimator (m.l.e. for short) and more generally minimum contrast estimators like least squares or projection estimators as well as their penalized versions do share good properties under suitable but somewhat restrictive assumptions, but otherwise may not have the right rate of convergence or even be inconsistent. This fact has been recognized for a long time; examples about the m.l.e. and further references can be found in Le Cam [46].

Another serious deficiency of maximum likelihood (or similar) estimators is their lack of "robustness". By this, we mean the property that an estimator still behaves well (its risk does not change too much) if the true underlying distribution of the observations does not belong to the parameter set but remains close to it. Unfortunately, the performances of the m.l.e. can deteriorate considerably owing to some small departures from the assumptions, as shown by the simple illustration given below in Section 2.3.

After some years of study of minimum contrast estimators, we became convinced of the need for a more flexible and less demanding alternative method for estimation and model selection. We looked for a method that would avoid many difficulties connected with the study of penalized minimum contrast estimators: for instance the systematic use of delicate empirical processes, chaining, or concentration of measures arguments which typically require restrictive assumptions. We wanted to get rid of entropy with bracketting assumptions and Kullback-Leibler information numbers in connection with the m.l.e. and to avoid the various boundedness restrictions that often mar the proofs about

penalized least squares and projection density estimators. Some illustrations of these difficulties can be found in most papers and books on the subject, among which (small sample) [58,59,14,5,21,22] or [63]. Further limitations of the classical methods for model selection are connected with the choice of the models which have to share some special properties: they should, for instance, be finite dimensional linear spaces generated by special bases, as in Baraud [3] or be uniformly bounded as in Yang ([66,67] and [70]).

1.2. About T-estimators

In this paper, we present and study an alternative estimation method which is based on two ingredients: one or several discrete models and a family of tests between the points of the models. By *model*, we mean an approximating set for the true unknown distribution of the observation(s). As to the tests, they are tests between balls in suitable metric spaces of probability measures and therefore enjoy some nice robustness properties. The existence of such tests is granted for various stochastic frameworks, among which those corresponding to i.i.d. observations, homoscedastic Gaussian sequences and bounded regression that we shall consider in this paper and to Gaussian regression with random design which has been studied in Birgé [12].

The resulting estimators, which we call T-estimators ("T" for *tests*) posess a number of interesting properties:

(i) The maximal risk over some parameter set $\mathcal{S}$ of a suitable T-estimator $\hat{s}$ (depending on $\mathcal{S}$) can be bounded in terms of simple metric properties of $\mathcal{S}$. This implies that one can derive upper bounds for the minimax risk over $\mathcal{S}$ in terms of those metric properties.

(ii) T-estimators inherit the robustness properties of the tests they are built from, a quality which is definitely not shared by maximum likelihood estimators. More precisely, if we use as our loss function some suitable distance d, the increase of risk incurred when the true parameter s does not belong to $\mathcal{S}$ (as compared to the risk when it does belong to $\mathcal{S}$) is bounded by $Cd(s,\mathcal{S})$, for some constant C independent of s.

(iii) If the T-estimator derives from a family of models, it automatically provides a model selection procedure, tending to choose the best model (in a suitable sense) among the family. In particular, good choices of the families of models result in adaptive estimators. From this point of view, one important property of T-estimators is the fact that they can cope with fairly arbitrary countable families of models, possibly non-linear or infinite dimensional. In particular, one can mix conventional parametric models with those used for non-parametric estimation.

The main advantage of this flexibility with respect to the structure of the models is to provide a complete decoupling between the choice of the models and the analysis of T-estimators. The existence of T-estimators depends on the existence of suitable robust tests which is only connected to the stochastic framework we consider together with the choice of a proper distance. As to the models' choice, it should be motivated only by the ideas we have about the true unknown parameter or the assumptions we make about it. Therefore models will be provided by approximation theory or our prior information or belief. Moreover, the same families of models may be used for different stochastic frameworks, leading to similar results. We shall in particular emphasize here the complete parallelism between model selection using T-estimators within the "white noise framework" and the i.i.d. framework (density estimation) with Hellinger loss.

There is a counterpart to these nice properties: our construction is often complicated. As a consequence, although our estimators could be implemented in some favourable cases, their complexity will often be too large so that they can actually be computed. They should be considered as "abstract" estimators providing a good indication of what is "theoretically" feasible to solve a given estimation problem.

Another price to pay for this level of generality is that our risk bounds will be given up to universal constants that may be large. We actually decided to sacrifice to simplicity and made no serious effort to optimize the constants. This would have been at the price of an increased complexity of both the assumptions and the proofs: bounding the constants efficiently requires to take advantage of the specificity of each particular situation, which is just the opposite to the philosophy of this paper. The concerned reader could adapt the method to any specific problem he considers in order to improve the constants.

1.3. Some historical remarks

It has been known for a long time that one could build confidence intervals from suitable families of tests, but, as far as we know, the idea of using tests between probability balls to build estimators is due to Le Cam who was

looking for a "universal" $\sqrt{n}$-consistent preliminary estimator for parametric models to replace the m.l.e. In [43] and [44], Le Cam described the construction of estimators from families of tests and analyzed their performances in terms of the "dimension" (in a suitable metric sense) of the set of parameters. In [8–10] and [11], using an alternative construction, still based on testing, we extended Le Cam's results with an emphasis on the minimax risk for non-parametric estimation, robustness and the treatment of some cases of dependent variables. Although a more recent summary of Le Cam's point of view on this subject appeared in [47], these ideas remained widely unknown since then (with the exception of Groeneboom [32]) and, to our knowledge, nobody (including the present author) tried to apply the method to other stochastic frameworks like the Gaussian one for instance, that we consider here, or to extend it. Related points of view about the relationships between the minimax risk and the metric structure of the parameter space are to be found in Yatracos [73], Yang and Barron [72], Devroye and Lugosi [28] and Yang [69].

Some fundamental ideas for complexity-based model selection, which are also somewhat related to testing, appeared later in Barron and Cover [6] and Barron [4]. They gave birth to a considerable amount of litterature on model selection based on penalized minimum contrast estimators and empirical processes techniques, which, unavoidably, suffer from the same defects as ordinary minimum contrast estimators. Mixing the old idea of building estimators from tests together with some newer ones about penalization borrowed from [6] and subsequent works will allow us to substantially improve and generalize the constructions of [8] and [10] in particular towards model selection and adaptive estimation.

An alternative trend of methods for model selection and adaptation that received a great attention in the recent years is based on selection or mixing of procedures. These methods, which have more practical relevance, share many of the advantages of our approach, in particular its flexibility and adaptation properties, but there are some noticeable differences. We defer a comparison of the two points of view to Section 9. Let us just mention here a few key references on the subject of aggregation like Juditsky and Nemirovski [37], Nemirovski [50], Yang ([66–68] and [70]), Catoni ([23] and [24]), Tsybakov [56] and Wegkamp [63].

Although we shall study at length the performances of T-estimators, we shall not discuss their optimality properties here. This would involve the comparison of our upper bounds with lower bounds based on dimensional arguments, as in [8,10,11,72,69] or [26]. Part of this task has already been achieved there and many other lower bounds results are known for various special situations. It suffices, in many cases, to compare those known lower bounds with our upper bounds to check that properly constructed T-estimators are often (approximately) minimax.

1.4. About the content of this paper

We begin our analysis by two introductory sections which give both motivations and heuristics for our construction. Although they provide some useful hints for the understanding of our somewhat abstract developments, they are not, strictly speaking, technically mandatory for reading the sequel and the impatient reader could jump directly to Section 4. We first illustrate, via three examples, some weaknesses of the maximum likelihood method: it does not work at all when the likelihood process behaves in an erratic way, it is not robust and it can be fooled by the "massiveness" of the parameter set, even if we merely want to estimate the mean of a Gaussian vector with identity covariance matrix under the assumption that this mean belongs to some convex, compact subset of some (high-dimensional) Euclidean space. A careful analysis of the performances of the m.l.e. on a finite set then provides some hints about a possible solution to the above mentioned problems.

Section 4 describes the abstract stochastic framework we shall work with all along the paper and explains the construction of T-estimators based on a discrete set S and a family of tests between the points of this set. In Section 5, we state the assumptions that should be satisfied by S and the tests when S can be viewed as a single model for the unknown parameter to be estimated. Then we give the resulting risk bounds for T-estimators and show that the required assumptions are satisfied for the frameworks we consider here: independent variables, Gaussian sequences and bounded regression. In the next section, we show how to build discrete models and the corresponding T-estimators in order to bound the minimax risk over some given parameter set $\mathcal{S}$ by a function depending on its metric properties only and which we call its *metric dimension*. Section 7 explains how to extend the previous construction to the case when we want to use several (possibly many) models simultaneously. The resulting T-estimators have a risk which is roughly bounded by the smallest among all risk bounds for the T-estimators derived from one model in the family, plus (possibly) an additional term due to the complexity of the family of models.

Section 8 is devoted to various applications. In particular, we show how to mix models for parametric and non-parametric estimation. In the Gaussian sequence framework, we show that T-estimators not only allow to recover all the results of Birgé and Massart [17] since they can handle in the same way arbitrary families of linear models, but also allow to mix other sorts of models with the previous ones, possibly infinite-dimensional like ellipsoids or finite-dimensional but non-linear like classical parametric models. As to density estimation with Hellinger loss, our analysis demonstrates that any result about T-estimators we can prove in the white noise framework has a parallel (modulo a simple translation) for density estimation, which is far from being true with minimum contrast estimators. In particular we consider here the problem of adaptive estimation over general Besov balls with Hellinger loss, but all the other results of [17] about the white noise framework could be transfered to density estimation with Hellinger loss in the same way.

As previously mentioned, one can distinguish between two types of T-estimators: the simplest ones are based on a single model while a more sophisticated construction can handle many models simultaneously, for instance in order to derive adaptive estimators. In the case of i.i.d. observations, an alternative approach based on procedure selection works as follows. First build an estimator on each model, possibly a T-estimator or any other one likes, and select one of them to get the final estimator. This is a particular case of aggregation of estimators. We consider the problem of aggregation using T-estimators in Section 9. In particular we show that this two-steps procedure, based on initial T-estimators for each model, is essentially equivalent to the general procedure when the sample can be split into two ones with the same distribution. We apply the method to selecting a partition for histogram estimation or a particular linear approximating space among a family for estimating a bounded regression function, among other examples. We also investigate the similarities and differences between aggregation procedures and T-estimators.

1.5. Three illustrations with independent variables

Let us conclude this section with three specific applications, as an appetizer for the reader.

Estimating a seemingly uniform distribution. Our first illustration deals with the problem of robust estimation within the model of uniform distributions on $[0,\theta]$, $\theta > 0$. The difficulty here comes from the fact that our observations $X_1,\dots,X_n$, although independent, do not necessarily follow the assumed model.

Proposition 1. *Let $X_1,\dots,X_n$ be independent random variables with arbitrary unknown distributions $\overline{P}_i$, $1 \leqslant i \leqslant n$, on $\mathbb{R}_+$. Let $\mathcal{U}_\theta$ denote the uniform distribution on $[0,\theta]$, $\theta > 0$ and h the Hellinger distance between probabilities. There exists an estimator $\hat{\theta}(X_1,\dots,X_n)$ such that, whatever the distributions $\overline{P}_i$,*

$$\mathbb{E}\Big[\sup_{1\leqslant i\leqslant n} h^2(\overline{P}_i,\mathcal{U}_{\hat{\theta}})\Big] \leqslant C \inf_{\theta>0}\left\{\sup_{1\leqslant i\leqslant n} h^2(\overline{P}_i,\mathcal{U}_\theta) + \frac{\max\{\log(\Gamma_n^{-1}|\log\theta|);1\}}{n}\right\}$$

where C denotes a universal constant and

$$\Gamma_n = 33.6\times 10^5 n^{-1}\big(4.5\exp\big[\max\big\{(n/84);2\big\}\big]-1\big). \tag{1.1}$$

These performances should be compared with those of the maximum likelihood estimator, which is the largest observation $X_{(n)}$. If the model is true, i.e. $X_1,\dots,X_n$ are i.i.d. $\mathcal{U}_{\theta_0}$, then the risk of the m.l.e. is $(2n+1)^{-1}$. For our estimator the risk is of the right order n^{-1} apart from the factor $\max\{\log(|\log\theta|/\Gamma_n);1\}$ (which equals 1 unless $\log(|\log\theta_0|)$ is really huge) and the (unfortunately) large constant C, which is the price to pay for robustness. On the other hand, if the model is not correct because $X_1,\dots,X_n$ are not i.i.d. $\mathcal{U}_{\theta_0}$ but it is only slightly wrong in the sense that $\sup_{1\leqslant i\leqslant n} h^2(\overline{P}_i,\mathcal{U}_{\theta_0}) \leqslant 5/(4n)$ for some $\theta_0 > 0$, the risk of our estimator remains of order $n^{-1}\max\{\log(\Gamma_n^{-1}|\log\theta|);1\}$ while the risk of the m.l.e. may become larger than 0.38 as shown in Section 2.3.

Adaptive estimation in Besov spaces with $\mathbb{L}_1$-loss. Our second example deals with adaptive density estimation for general Besov balls when the loss is the $\mathbb{L}_1$-distance between densities.

Theorem 1. *Let $X_1, \dots, X_n$ be an n-sample from some distribution $\overline{P_s}$ with density s with respect to Lebesgue measure on $[0,1]^k$. One can build a T-estimator $\hat{s}(X_1, \dots, X_n)$ for s such that, if the Besov semi-norm of s satisfies $|s|_{B^\alpha_{p,\infty}} \leqslant R$ for some $p > 0$, $\alpha > k(1/p-1)_+$ and $R \geqslant 1/\sqrt{n}$, then, for $1 \leqslant q \leqslant 79$,*

$$\mathbb{E}_s\big[\|s-\hat{s}\|_1^q\big] \leqslant C(\alpha, p, q, k) R^{kq/(2\alpha+k)} n^{-q\alpha/(2\alpha+k)}.$$

As far as we know, all results about this density estimation problem (without additional boundedness assumptions), even those dealing with the minimax risk for known α and p, are limited to a range of the form $r > \alpha > k/p$, with r some positive integer as in Donoho et al. [29]. More recent improved results of Kerkyacharian and Picard [38] extend this range but they use projection estimators over some wavelet basis and assume some large deviation inequalities for the empirical coefficients that we are unable to check without additional assumptions on s. By nature, the procedure is also limited to $\alpha < r$ for some given r depending on the choice of the basis. Our method allows to handle the larger scale of Besov spaces given by $\alpha > (k/p-k)_+$.

Model selection for bounded regression with random design. In this case, we observe an even number n of i.i.d. pairs of variables (X_i, Y_i), $1 \leqslant i \leqslant n$, with $X_i \in \mathcal{X}$ and $Y_i \in [0,1]$. The distribution μ of X_i on $\mathcal{X}$ is unknown, $\|\cdot\|_2$ denotes the norm in $\mathbb{L}_2(\mu)$ and we assume that Y_i and X_i are connected in the following way: $Y_i = s(X_i) + \xi_i$ for some function s from $\mathcal{X}$ to $[0,1]$ and $\mathbb{E}[\xi_i | X_i] = 0$. To derive an estimator of the unknown parameter s we can, for instance, use a countable family of linear spaces of bounded functions on $\mathcal{X}$ and get the following result.

Theorem 2. *Given the observations (X_i, Y_i), $1 \leqslant i \leqslant n$, a countable family $\{T_m,\ m \in \mathcal{M}\}$ of finite dimensional linear subspaces of bounded functions on $\mathcal{X}$ with respective dimensions D_m and a family $\{\Delta_m,\ m \in \mathcal{M}\}$ of positive weights with $\Delta_m \geqslant 1$ and $\sum_{m\in\mathcal{M}} \exp[-\Delta_m] \leqslant \mathrm{e}$, one can construct an estimator $\hat{s}$ which is a function from $\mathcal{X}$ to $[0,1]$ satisfying, for all s,*

$$\mathbb{E}\big[\|\hat{s}-s\|_2^2\big] \leqslant C \inf_{m\in\mathcal{M}} \Big\{ \inf_{t\in T_m} \|t-s\|_2^2 + n^{-1} \max\big\{(D_m \log n), \Delta_m\big\}\Big\},$$

where C denotes a universal constant.

2. The difficulties connected with maximum likelihood estimation

The m.l.e. is known to behave in an optimal way for parametric estimation under suitable regularity assumptions (see, for instance, Le Cam [41] or the book [60] by van der Vaart) and to have the right rate of convergence in non-parametric situations under specific entropy assumptions ([57–59,13,54] and [64]). It has nevertheless been recognized for a long time that it can also behave quite poorly when such assumptions are not satisfied. Many counterexamples to consistency or optimality of the m.l.e. have been found in the past and the interested reader should look at those given by Le Cam in [46] which is a real advertisement against the systematic use of the m.l.e. without caution. As Le Cam said in the introduction of this paper, "one of the most widely used methods of statistical estimation is that of maximum likelihood .... Qualms about the general validity of optimality properties (of maximum likelihood estimators) have been expressed occasionally." Then a long list of examples follows, showing that the m.l.e. may behave in a terrible way. Further ones are to be found in [13], Section 4 and [28], Section 6.4. We shall add three more below. All these examples emphasize the fact that the m.l.e. is in no way a universal estimator. Indeed, all positive results about the m.l.e. involve much stronger assumptions (like L.A.N. in the parametric case, or entropy with bracketing conditions as in [58] and [59]) than those we want to use here. Even if the parameter set is compact, which prevents the m.l.e. to go to infinity, one can get into troubles for two reasons: either the likelihood process does not behave in a smooth way locally or the space is so "massive" (in an informal sense, see an example below) that it is not possible to get a local control of the supremum of the likelihood process.

2.1. Erratic behaviour of the likelihood process

The difficulties caused by irregularity of the likelihood function for the i.i.d. setting, even in the simplest parametric case of a translation family, are easy to demonstrate. Consider some density f with respect to Lebesgue measure on the line satisfying $f(x) > 0$ for all $x \in \mathbb{R}$ and $\lim_{x\to 0} f(x) = +\infty$. If we observe a sample $X_1, \dots, X_n$ of some translate

of the density $f_s(x) = f(x-s)$ with $s \in \mathbb{R}$, the maximum likelihood estimator does not exist since the likelihood is infinite at every observation. This phenomenon is neither due to the non-compactness of the parameter space (it remains true if we restrict s to some compact interval) nor to the massiveness of the parameter space, but rather to the erratic behaviour of the likelihood function. Nevertheless, setting $p = \int_{-\infty}^{0} f(t)\,\mathrm{d}t$, the corresponding empirical p-quantile provides quite a good estimator of s, which means that the statistical problem to be solved is not a difficult one at all.

2.2. Some difficulties encountered with high-dimensional parameter sets

More subtle than the effects of the lack of smoothness of the likelihood function are the difficulties due to the "massiveness" of the parameter space. Some asymptotic results in this direction have been given in Section 4 of [13] relying on the construction of rather complicated infinite dimensional parameter sets. A much simpler and non-asymptotic illustration of the suboptimality of the m.l.e. when the parameter set is too "massive", although convex and compact, is as follows.

Let $\boldsymbol{X} = (X_0, \ldots, X_k)$ be a $(k+1)$-dimensional Gaussian vector with distribution $\mathcal{N}(s, I_{k+1})$, where I_{k+1} denotes the identity matrix of dimension $k+1$. For any vector $s = (s_0, \ldots, s_k)$ in $\mathbb{R}^{k+1}$, we denote by s' its projection onto the k-dimensional linear space spanned by the k last coordinates and by $\|s\|$ its Euclidean norm.

Proposition 2. *Let the integer k be not smaller than* 128 *and*

$$\mathcal{S} = \bigl\{ s \in \mathbb{R}^{k+1} \mid |s_0| \leqslant k^{1/4} \text{ and } \|s'\| \leqslant 2\bigl(1 - k^{-1/4}|s_0|\bigr) \bigr\}.$$

The quadratic risk of the maximum likelihood estimator $\hat{s}$ on $\mathcal{S}$ and the minimax risk satisfy respectively

$$\sup_{s\in\mathcal{S}} \mathbb{E}_s\bigl[\|s-\hat{s}\|^2\bigr] \geqslant (3/4)\sqrt{k} + 3 \quad \text{and} \quad \inf_{\tilde{s}} \sup_{s\in\mathcal{S}} \mathbb{E}_s\bigl[\|s-\tilde{s}\|^2\bigr] \leqslant 5.$$

This demonstrates that the maximal risk of the m.l.e. may be much larger than the minimax risk when k is large. The proof is given in Appendix A.

2.3. Lack of robustness of the parametric m.l.e.

We shall conclude this study by showing that the m.l.e. is definitely not a robust estimator in the sense that its risk can increase dramatically if the parametric assumption is only slightly violated. Let us assume that we observe an i.i.d. sample of size $n \geqslant 4$ from some unknown distribution $\overline{P}$ on $[0,1]$ and we use for our statistical model the parametric family $\mathcal{S}$ of all uniform distributions $\mathcal{U}_\theta$ on $[0,\theta]$ with $0 < \theta \leqslant 1$. Since $\overline{P}$ may not belong to this family, we cannot use the square of the distance between parameters as our loss function as one would usually do. We have to introduce a loss function which makes sense when $\overline{P} \notin \mathcal{S}$ and replace the distance between parameters by a distance between distributions. We choose, for reasons that will become clearer later on, the Hellinger distance. Let us recall that the Hellinger distance h between two probabilities P and Q defined on the same space and their Hellinger affinity ρ are given respectively by

$$h^2(P,Q) = \frac{1}{2}\int \bigl(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\bigr)^2, \qquad \rho(P,Q) = \int \sqrt{\mathrm{d}P\,\mathrm{d}Q} = 1 - h^2(P,Q), \tag{2.1}$$

where $\mathrm{d}P$ and $\mathrm{d}Q$ denote the densities of P and Q with respect to any dominating measure (the result being independent of the choice of such a measure). One can check that $\rho(\mathcal{U}_\theta, \mathcal{U}_{\theta'}) = \sqrt{\theta/\theta'}$ if $\theta < \theta'$. It follows that, if the parametric model is true (i.e. $\overline{P} = \mathcal{U}_\theta$ for some $\theta \in (0,1]$), the risk of the maximum likelihood estimator of θ, which is the largest observation $X_{(n)}$, is given by $\mathbb{E}_\theta[h^2(\mathcal{U}_\theta, \mathcal{U}_{X_{(n)}})] = 1/(2n+1)$. Let us now suppose that $\overline{P}$ does not belong to $\mathcal{S}$ but has the density

$$10\bigl[(1-2n^{-1})\mathbb{1}_{[0,1/10]} + 2n^{-1}\mathbb{1}_{[9/10,1]}\bigr]$$

with respect to Lebesgue measure. Since $\rho(\overline{P}, \mathcal{U}_{1/10}) = (1-2n^{-1})^{1/2}$, $h^2(\overline{P}, \mathcal{U}_{1/10}) < 5/(4n)$ for $n \geqslant 4$ and one would expect the increase of risk due to this small deviation from the parametric assumption to be $\mathrm{O}(1/n)$ if the

m.l.e. were robust. This is not the case: with probability $1-(1-2/n)^n > 1-\mathrm{e}^{-2}$, $X_{(n)} \geqslant 9/10$ and therefore $\rho(\overline{P}, \mathcal{U}_{X_{(n)}}) < (1/3) + (1/\sqrt{20})$. It follows that

$$\mathbb{E}_{\overline{P}}\bigl[h^2(\overline{P}, \mathcal{U}_{X_{(n)}})\bigr] > \left[\frac{2}{3} - \frac{1}{\sqrt{20}}\right](1-\mathrm{e}^{-2}) > 0.38.$$

3. How to rescue the m.l.e., some heuristics

In order to explain our point of view about maximum likelihood estimation, it will be convenient to work within a specific statistical framework and we shall assume here that our observation is an n-sample $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ from some unknown distribution $\overline{P}_s$ on the measurable space $(\mathcal{X}, \mathcal{W})$, where s belongs to some parameter set $\mathcal{S}$. We shall denote the corresponding probabilities by $\mathbb{P}_s$. Assuming that our parametrization is one-to-one, we can turn $\mathcal{S}$ into a metric space with metric h, setting $h(s,t) = h(\overline{P}_s, \overline{P}_t)$ where h denotes the Hellinger distance given by (2.1). We shall also assume that $\mathcal{S}$ is compact, which implies that our family of probabilities $\{\overline{P}_s,\ s \in \mathcal{S}\}$ is dominated with respective densities $\mathrm{d}\overline{P}_s$ and, to be consistent with the M-estimators approach, we shall denote by $\Lambda_n(t, \boldsymbol{X}) = -\sum_{i=1}^n \log(\mathrm{d}\overline{P}_t(X_i))$ minus the log-likelihood at t. Thus the m.l.e. with respect to some set S is the minimizer of $\Lambda_n(t, \boldsymbol{X})$ for $t \in S$.

3.1. About the m.l.e. on finite sets

As we have seen, the maximum likelihood estimator on $\mathcal{S}$ may behave poorly either because the likelihood process behaves in an erratic way on $\mathcal{S}$ or because $\mathcal{S}$ is too "massive". A natural idea to build an alternative estimator is to approximate the compact set $\mathcal{S}$ by a finite subset S such that for $s \in \mathcal{S}$ one can find $t \in S$ with $h(s,t) \leqslant \eta$ and restrict the maximization of the likelihood to S. Since S is finite, there is no problem with the local behaviour of the likelihood and the amount of discretization (the size of η) will allow to control the massiveness of S. For simplicity, let us assume that

$$\mathbb{P}_s\bigl[\Lambda_n(u, \boldsymbol{X}) = \Lambda_n(t, \boldsymbol{X})\bigr] = 0 \quad \text{for all } s \in \mathcal{S} \text{ and } t, u \in S,\ t \neq u. \tag{3.1}$$

Then the maximum likelihood estimator $\hat{s}$ on S exists and is unique $\mathbb{P}_s$-a.s.

If $s \in S$, one can bound the deviations of $\hat{s}$ from s by a simple argument which goes back to Wald [62]. The first step is to observe that, for all t and u, the errors of likelihood ratio tests between $\overline{P}_t$ and $\overline{P}_u$ are bounded by

$$\mathbb{P}_t\bigl[\Lambda_n(u, \boldsymbol{X}) \leqslant \Lambda_n(t, \boldsymbol{X})\bigr] \leqslant \exp\bigl[n\log\bigl(\rho(\overline{P}_u, \overline{P}_t)\bigr)\bigr] \leqslant \exp\bigl[-nh^2(u,t)\bigr], \tag{3.2}$$

which follows from (A.5) in Appendix A and (2.1). More precise results in this direction can be found in Chernoff [25].

Now, given $\eta > 0$, $K \geqslant 1$ and $s \in S$, we want to bound $\mathbb{P}_s[h(s,\hat{s}) \geqslant K\eta]$. For $k \geqslant 0$, we set $S_k = \{u \in S \mid 2^{k/2}K\eta \leqslant h(s,u) < 2^{(k+1)/2}K\eta\}$ and denote by $|S_k|$ the cardinality of S_k. We derive from (3.2) that

$$\begin{aligned}
\mathbb{P}_s\bigl[h(s,\hat{s}) \geqslant K\eta\bigr] &\leqslant \mathbb{P}_s\bigl[\exists u \in S \text{ with } h(s,u) \geqslant K\eta \text{ and } \Lambda_n(u, \boldsymbol{X}) \leqslant \Lambda_n(s, \boldsymbol{X})\bigr] \qquad (3.3)\\
&\leqslant \sum_{k=0}^{+\infty} \mathbb{P}_s\bigl[\exists u \in S_k \text{ with } \Lambda_n(u, \boldsymbol{X}) \leqslant \Lambda_n(s, \boldsymbol{X})\bigr]\\
&\leqslant \sum_{k=0}^{+\infty} |S_k| \sup_{u \in S_k} \mathbb{P}_s\bigl[\Lambda_n(u, \boldsymbol{X}) \leqslant \Lambda_n(s, \boldsymbol{X})\bigr] \leqslant \sum_{k=0}^{+\infty} |S_k| \exp[-2^k n K^2 \eta^2]. \qquad (3.4)
\end{aligned}$$

In order to get a small bound for the right-hand side of (3.4) for $K \geqslant 1$, one should first require that, when $K = 1$, the first term of the series, $|S_0|\exp[-n\eta^2]$, be small, which will determine the choice of η. In particular, one should require that $n\eta^2 \geqslant 1$. Then one should put a suitable assumption about the massiveness of S implying that $|S_k|$ does not grow too fast with k so that the sum of the whole series is not much larger than its first term. For this, something akin to $|S_k| \leqslant |S_0|\exp(2^{k-1})$ would do.

3.2. An alternative point of view on the previous analysis

If $s \in \mathcal{S} \setminus S$, one can find $s' \in S$ with $h(s,s') \leqslant \eta$, hence $\mathbb{P}_s[h(s,\hat{s}) \geqslant (K+1)\eta] \leqslant \mathbb{P}_s[h(s',\hat{s}) \geqslant K\eta]$ and the previous arguments could be extended straightforwardly, at least for K large enough, if we could bound $\mathbb{P}_s[\Lambda_n(u,\boldsymbol{X}) \leqslant \Lambda_n(t,\boldsymbol{X})]$ by an analogue of (3.2) when $h(s,t) \leqslant \eta$ and $h(u,t)$ is large enough. This would mean that the likelihood ratio tests between two points t and u of S do have small errors, when n is large, for testing the Hellinger balls of radius η and respective centers t and u, provided that $h(u,t)$ is large, which is a robustness property. Unfortunately, unless one puts some additional assumptions on the likelihood ratios, such a property does not hold in general, as shown by the following counter-example.

Let μ denote the Lebesgue measure on $[0,1]$, $\overline{P}_w = w \cdot \mu$ for any density w with respect to μ, $s = \mathbb{1}_{[0,1]}$, $\lambda = 1-(2n)^{-1}$, $t = \lambda^{-1}\mathbb{1}_{[0,\lambda]}$ and let u be any density such that $\sup_{x\in[0,1]} |\log u(x)| < +\infty$. Then $\rho^2(\overline{P}_s, \overline{P}_t) = \lambda$ and it follows from [43] that the sum of the errors of any test between s and t based on a sample of size $n \geqslant 12$ is at least $\lambda/2 = (1/2) - (4n)^{-1}$. This means that we cannot test well whether s or t is true. Nevertheless, even if $h(t,u)$ is close to one, which means that u is far away from t,

$$\mathbb{P}_s\big[\Lambda_n(u,\boldsymbol{X}) \leqslant \Lambda_n(t,\boldsymbol{X})\big] \geqslant \mathbb{P}_s\Big[\sup_{1\leqslant i\leqslant n} X_i \geqslant \lambda\Big] = 1 - \left(1 - \frac{1}{2n}\right)^n > 1 - \mathrm{e}^{-1/2}.$$

In order to see how we can fix the problem, let us carefully review the previous analysis of the performances of the m.l.e. on a finite set. The key point is to notice that, by (3.1), the m.l.e. $\hat{s}$ is the unique point in S such that all likelihood ratio tests between $\hat{s}$ and any other point accept $\hat{s}$. An equivalent way of stating this fact is to set, for any $t \in S$, $\mathcal{R}_t = \{u \in S \mid \Lambda_n(u,\boldsymbol{X}) < \Lambda_n(t,\boldsymbol{X})\}$ and $\mathcal{D}_{\boldsymbol{X}}(t) = \sup_{u\in\mathcal{R}_t} h(t,u)$, (with the convention $\sup_{u\in\emptyset} h(t,u) = 0$), then define $\hat{s}$ as $\operatorname{argmin}_{t\in S} \mathcal{D}_{\boldsymbol{X}}(t)$ since $\hat{s} = \operatorname{argmin}_{t\in S} \Lambda_n(t,\boldsymbol{X})$ is equivalent to $\mathcal{D}_{\boldsymbol{X}}(\hat{s}) = 0$. It follows from the definition of $\mathcal{D}_{\boldsymbol{X}}$ that, for $t,u \in S$, $\mathcal{D}_{\boldsymbol{X}}(t) \vee \mathcal{D}_{\boldsymbol{X}}(u) \geqslant h(t,u)$ and therefore that $h(t,\hat{s}) \leqslant \mathcal{D}_{\boldsymbol{X}}(t) \vee \mathcal{D}_{\boldsymbol{X}}(\hat{s}) \leqslant \mathcal{D}_{\boldsymbol{X}}(t)$. Finally, if $h(s,s') \leqslant \eta$,

$$\mathbb{P}_s\big[h(s,\hat{s}) \geqslant (K+1)\eta\big] \leqslant \mathbb{P}_s\big[h(s',\hat{s}) \geqslant K\eta\big] \leqslant \mathbb{P}_s\big[\mathcal{D}_{\boldsymbol{X}}(s') \geqslant K\eta\big] = \mathbb{P}_s\big[\exists u \in \mathcal{R}_{s'} \text{ with } h(s',u) \geqslant K\eta\big],$$

since S is finite. This is equivalent to (3.3) but the proof of (3.4) cannot proceed as before because, as we have just seen, likelihood ratio tests are not robust. Now suppose we can replace the likelihood ratio tests between t and u by some alternative ones which are robust and redefine $\mathcal{R}_t$ accordingly: $\mathcal{R}_t$ is the set of points $u \in S$ such that the test between t and u decides u. This still makes sense and the definitions of the function $\mathcal{D}_{\boldsymbol{X}}$ and of $\hat{s} = \operatorname{argmin}_{t\in S} \mathcal{D}_{\boldsymbol{X}}(t)$ as well. Of course, since we started from some arbitrary family of robust tests, there is no reason anymore that one could still express $\hat{s}$ as $\operatorname{argmin}_{t\in S} \gamma_n(t,\boldsymbol{X})$ for some function γ_n. Nevertheless, the bound

$$\mathbb{P}_s\big[h(s,\hat{s}) \geqslant (K+1)\eta\big] \leqslant \mathbb{P}_s\big[\exists u \in \mathcal{R}_{s'} \text{ with } h(s',u) \geqslant K\eta\big] \tag{3.5}$$

still holds and, if we could bound the errors of the new robust tests by some suitable analogue of (3.2), we could proceed as before from (3.5) to some analogue of (3.4). Typically, we shall require that the robust tests we use satisfy the following error bound for some constants c and κ:

$$\mathbb{P}_s[\text{the test between } t \text{ and } u \text{ decides } u] \leqslant \exp\big[-cnh^2(t,u)\big] \quad \text{if } h(t,u) \geqslant \kappa h(s,t). \tag{3.6}$$

Deriving an estimator of $s \in \mathcal{S}$ from a family of tests satisfying (3.6) by setting $\hat{s} = \operatorname{argmin}_{t\in S} \mathcal{D}_{\boldsymbol{X}}(t)$ is actually very natural: if the true parameter s is close to $s' \in S$, all the tests between s' and the points $u \in S$ far enough from s' will accept t with large probability and $\mathcal{D}_{\boldsymbol{X}}(s')$ should therefore not be large. On the other hand, if $u \in S$ is far enough from s, the test between s' and u will accept s' which will result in a large value of $\mathcal{D}_{\boldsymbol{X}}(u)$.

Obviously, the previous reasoning essentially relies on the existence of robust tests satisfying an analogue of (3.2) like (3.6), but it has been known for a long time that such tests do exist, as shown by Le Cam [44] and Birgé [9]. They actually also exist in other stochastic frameworks, not only for i.i.d. samples, which accounts for the introduction of the general setting that follows. Note also that the interpretation of estimators in terms of testing was absolutely essential for our construction. It is indeed, together with the elementary arguments used for deriving (3.4) (counting the number of points of S contained in balls and bounding the errors of tests), at the chore of our method.

4. A robust substitute for the (penalized) m.l.e.

4.1. A general statistical framework

We observe some random element $\boldsymbol{X}$ from $(\Omega, \mathcal{A})$ to $(\Xi, \mathcal{Z})$ with distribution $P_{\boldsymbol{X}}$ belonging to some set $\mathcal{P}$ of possible distributions on $(\Xi, \mathcal{Z})$ and we have at hand, to serve as parameter set, a semi-metric space (M, d), which means that the function d is a semi-distance on M, i.e. it satisfies

$$d(t,t) = 0 \quad \text{and} \quad d(t,u) = d(u,t) \geqslant 0 \quad \text{for all } t, u \in M, \tag{4.1}$$

but not necessarily some version of the triangular inequality:

$$d(t,u) \leqslant A\big[d(t,r) + d(r,u)\big] \quad \text{for all } r, t, u \in M \text{ and some } A \geqslant 1. \tag{4.2}$$

For the applications we develop in this paper, d is a genuine distance satisfying (4.2) with $A = 1$. Nevertheless, part of the results we shall prove here only require that (4.1) holds and since those particular results will be useful for further applications (to be given in subsequent papers) which do involve semi-distances, we shall distinguish hereafter between the results that assume that d is a genuine distance from the others. One could actually only assume that (4.2) holds with $A > 1$. This would only affect the value of the constants in all our results. For simplicity, we only consider here the case $A = 1$, the extension to the case of $A > 1$ being straightforward. Even if d is not a distance, we shall use the following notations for open and closed balls in M with center t and radius $r \geqslant 0$:

$$\mathcal{B}_d(t,r) = \big\{u \in M \mid d(u,t) < r\big\} \quad \text{and} \quad \overline{\mathcal{B}}_d(t,r) = \big\{u \in M \mid d(u,t) \leqslant r\big\}, \tag{4.3}$$

possibly omitting the subscript d when no confusion is possible. The (semi-)distance $d(t, S)$ from some point $t \in M$ to some subset S of M is defined as $d(t, S) = \inf_{u \in S} d(t, u)$.

Our purpose in this paper is to design estimators of the unknown parameter $s = F(P_{\boldsymbol{X}})$ where F is some mapping from $\mathcal{P}$ to M. In most examples, the application F is one-to-one and F^{-1} is merely a parametrization of $\mathcal{P}$ by M, in which case we shall set $P_t = F^{-1}(t)$, $\mathcal{P} = \{P_t,\ t \in M\}$ and systematically identify t with P_t, M with $\mathcal{P}$ and write indifferently $d(P_t, P_u)$ or $d(t, u)$. We denote by $\mathbb{P}_s$ the probability on $(\Omega, \mathcal{A})$ that gives $\boldsymbol{X}$ its true distribution $P_{\boldsymbol{X}}$, by $\mathbb{E}_s$ the corresponding expectation operator and we set $P_{\boldsymbol{X}} = P_s$ when F is one-to-one. To any measurable map $\hat{s}$ from Ξ to M corresponds the estimator $\hat{s}(\boldsymbol{X})$. By a *model* for s we mean any subset (often denoted by S, S', $\overline{S}$ or $\mathcal{S}$) of M, which may or may not contain s. When speaking of a *countable* set, we always mean a finite or countable set. Constants will be denoted by $C, C', c_1, \ldots$ or by $C(x, y, \ldots)$ to emphasize their dependence on some input parameters $x, y, \ldots$. For simplicity, the same notation may be used to denote different constants. We shall systematically use $x \vee y$ and $x \wedge y$ for $\max\{x, y\}$ and $\min\{x, y\}$ respectively, we shall denote by $|S|$ the cardinality of the set S, by $\mathbb{N}^\star = \mathbb{N} \setminus \{0\}$ the set of positive integers and set

$$\lceil x \rceil = \inf\{n \in \mathbb{N} \mid n \geqslant x\}; \quad \lfloor x \rfloor = \sup\{n \in \mathbb{N} \mid n \leqslant x\} \quad \text{for } x \in \mathbb{R}_+. \tag{4.4}$$

4.2. The construction of T-estimators

4.2.1. Defining T-estimators

The construction of what we shall call a *T-estimator* ("T" for "test") requires

(i) a countable subset S of M which plays the role of an approximating set for the true unknown s;
(ii) a non-negative number ε and a positive *weight function* η from S to $\mathbb{R}_+$;
(iii) a family of tests between the points of S.

At this stage, we have to make quite precise what we actually mean by *a test between t and u* since, in our approach, there is no *hypothesis* or *alternative* or rather we ignore which of the two points will play each role.

Definition 1. Given a random element $\boldsymbol{X}$ with values in Ξ and two distinct points t and $u \in M$, a (non-randomized) test between t and u is a pair of measurable functions $\psi(t, u, \boldsymbol{X}) = 1 - \psi(u, t, \boldsymbol{X})$ with values in $\{0; 1\}$, our convention being that $\psi(t, u, \boldsymbol{X}) = 0$ means accepting t while $\psi(t, u, \boldsymbol{X}) = 1$ (or equivalently $\psi(u, t, \boldsymbol{X}) = 0$) means accepting u.

We shall stick to this convention throughout the paper. We can now define a T-estimator in the following way.

Definition 2. Let S be a countable subset of M, η be a non-negative function on S and $\varepsilon \geqslant 0$. Let $\{\psi(t,u,\boldsymbol{X})\}$ be a family of tests indexed by the pairs $(t,u) \in S^2$ with $t \neq u$ and satisfying the coherence relationship $\psi(u,t,\boldsymbol{X}) = 1-\psi(t,u,\boldsymbol{X})$. Setting $\mathcal{R}_t = \{u \in S, u \neq t \mid \psi(t,u,\boldsymbol{X}) = 1\}$, we define the random function $\mathcal{D}_{\boldsymbol{X}}$ on S by

$$\mathcal{D}_{\boldsymbol{X}}(t) = \begin{cases} \sup\limits_{u\in\mathcal{R}_t} \{d(t,u)\} & \text{if } \mathcal{R}_t \neq \emptyset; \\ 0 & \text{if } \mathcal{R}_t = \emptyset. \end{cases} \tag{4.5}$$

We call T-*estimator* (T_ε-*estimator* when we want to emphasize the value of ε) derived from S, η, ε and the family of tests $\{\psi(t,u,\boldsymbol{X})\}$ any measurable application $\hat{s}(\boldsymbol{X})$ with values in S satisfying

$$\big[\mathcal{D}_{\boldsymbol{X}}\big(\hat{s}(\boldsymbol{X})\big)\big] \vee \big[\varepsilon\eta\big(\hat{s}(\boldsymbol{X})\big)\big] = \inf_{t\in S}\big[\mathcal{D}_{\boldsymbol{X}}(t) \vee \varepsilon\eta(t)\big]. \tag{4.6}$$

Obviously, T-estimators need neither exist nor be unique in general, but we shall work under assumptions that, at least, ensure their existence.

4.2.2. A special case: M-estimators

As a special case, we find tests of likelihood ratio type, i.e. tests that derive from the comparison between the values at t and u of some random function $\gamma(\cdot,\boldsymbol{X})$ from M to $[-\infty,+\infty]$. Typically, $\gamma(\cdot,\boldsymbol{X})$ is a (possibly penalized) *empirical contrast function* as defined in Birgé and Massart [13] and Barron et al. [5], the case of likelihood ratio tests of Section 3.1 corresponding to $\gamma(t,\boldsymbol{X}) = \Lambda_n(t,\boldsymbol{X})$. In this case, computing a T-estimator derived from S almost amounts to finding a point minimizing the function $t \mapsto \gamma(t,\boldsymbol{X})$ for $t \in S$. Estimators based on the minimization on a criterion like $\gamma(t,\boldsymbol{X})$ are usually called M-estimators, as in Chapter 5 of van der Vaart [60], and we shall precisely define them in the following way.

Definition 3. Let $\gamma'(\cdot,\boldsymbol{X})$ be a random function from S to $[-\infty,+\infty]$, η a weight function from S to $\mathbb{R}_+$ and τ a non-negative number. Set

$$\gamma(t,\boldsymbol{X}) = \gamma'(t,\boldsymbol{X}) + \tau\eta^2(t) \quad \text{for all } t \in S. \tag{4.7}$$

A family of tests $\psi(t,u,\boldsymbol{X}) = 1-\psi(u,t,\boldsymbol{X})$ between t and u with $t,u \in S$ will be called a family of M-tests derived from the function $\gamma'(\cdot,\boldsymbol{X})$ with penalty $\tau\eta^2$ if

$$\psi(t,u,\boldsymbol{X}) = \begin{cases} 0 & \text{if } \gamma(t,\boldsymbol{X}) < \gamma(u,\boldsymbol{X}); \\ 1 & \text{if } \gamma(t,\boldsymbol{X}) > \gamma(u,\boldsymbol{X}); \end{cases} \tag{4.8}$$

for all $t,u \in S$, $t \neq u$, the value of ψ being arbitrary when $\gamma(t,\boldsymbol{X}) = \gamma(u,\boldsymbol{X})$. Any minimizer of $\gamma'(t,\boldsymbol{X}) + \tau\eta^2(t)$ over S will then be called an M-estimator.

Note that the difference between M-estimators and the T_0-estimators derived from the tests defined by (4.8) is rather subtle and only due to the possible differences in case of equality in (4.8). If, for all $t,u \in S$ with $t \neq u$, $\mathbb{P}_s[\gamma(t,\boldsymbol{X}) = \gamma(u,\boldsymbol{X})] = 0$, since S is countable, either there exists a minimizer of $\gamma' + \tau\eta^2$ over S and it is the unique T_0- and M-estimator, or there does not exist any M-estimator. More generally, when there exists a unique minimizer $\hat{s}(\boldsymbol{X})$ of $\gamma' + \tau\eta^2$, then $\mathcal{D}_{\boldsymbol{X}}(\hat{s}(\boldsymbol{X})) = 0$ and $\hat{s}$ is also the unique T_0-estimator. Apart from this situation, the relationship between T-estimators and M-estimators is not clear in general, although we shall see later that their properties are quite similar.

4.2.3. Elementary properties

The definition of $\mathcal{D}_{\boldsymbol{X}}$ implies that

$$d(t,u) \leqslant \mathcal{D}_{\boldsymbol{X}}(t) \vee \mathcal{D}_{\boldsymbol{X}}(u) \quad \text{for all } (t,u) \in S^2. \tag{4.9}$$

Consequently, any T_ε-estimator $\hat{s}(\boldsymbol{X})$ satisfies, by (4.6),

$$d\big(t,\hat{s}(\boldsymbol{X})\big) \leqslant \big[\mathcal{D}_{\boldsymbol{X}}(t) \vee \varepsilon\eta(t)\big] \quad \text{for all } t \in S. \tag{4.10}$$

It follows from the definition of $\mathcal{R}_t$ that $\mathcal{D}_{\boldsymbol{X}}(t)$ should be viewed as a *plausibility index* playing the role of minus the (penalized) likelihood at t: if it is large, one should believe that the true s is far from t. If d is a genuine distance and $\varepsilon = 0$, then (4.10) implies that $d(s,\hat{s}) \leqslant \inf_{t\in S}[d(s,t) + \mathcal{D}_{\boldsymbol{X}}(t)]$ which means that a T_0-estimator makes the best compromise among the points in S between the distance from t to the true s and its plausibility.

To deal with M-estimators, it will be convenient to introduce

$$\mathcal{R}'_t = \big\{u \in S \mid \gamma(u,\boldsymbol{X}) \leqslant \gamma(t,\boldsymbol{X})\big\} \quad \text{and} \quad \mathcal{D}'_{\boldsymbol{X}}(t) = \sup_{u\in\mathcal{R}'_t} \big\{d(t,u)\big\}, \tag{4.11}$$

so that

$$d(t,u) \leqslant \mathcal{D}'_{\boldsymbol{X}}(t) \vee \mathcal{D}'_{\boldsymbol{X}}(u) \quad \text{for all } (t,u) \in S^2 \tag{4.12}$$

and, if $\hat{s}$ is any minimizer of $\gamma(\cdot,\boldsymbol{X})$ over S, $\hat{s} \in \mathcal{R}'_t$ for all $t \in S$, hence

$$d\big(t,\hat{s}(\boldsymbol{X})\big) \leqslant \mathcal{D}'_{\boldsymbol{X}}(t) \quad \text{for all } t \in S. \tag{4.13}$$

We shall also use the fact, which follows from (4.8), that

$$\big\{u \in S \mid \gamma(u,\boldsymbol{X}) < \gamma(t,\boldsymbol{X})\big\} \subset \overline{\mathcal{B}}_d\big(t,\mathcal{D}_{\boldsymbol{X}}(t)\big). \tag{4.14}$$

4.3. Some basic assumptions

In order to ensure the existence of T- or M-estimators and show that they enjoy nice properties we have to choose both the set S and the family of tests ψ in a proper way and require that they satisfy some suitable assumptions. Since we want to mimic the "proof" we gave in Section 3, our tests should satisfy a suitable analogue of (3.6) and we should have some control on the "massiveness" of S.

4.3.1. Assumptions about our family of tests

In order to be sure that suitable tests exist that warrant the existence of T-estimators we shall always work under the following assumption and choose S as a subset of M_T.

Assumption 1. There exists a subset M_T of M, a function δ from $M \times M_T$ to $[0,+\infty]$ and two constants $a, B > 0$ such that, for any pair $(t,u) \in M_T^2$ with $t \neq u$ and any $x \in \mathbb{R}$, one can find a test $\psi(t,u,\boldsymbol{X})$ satisfying

$$\sup_{\{s\in M\mid\delta(s,t)\leqslant d(t,u)\}} \mathbb{P}_s\big[\psi(t,u,\boldsymbol{X}) = 1\big] \leqslant B\exp\big[-a\big(d^2(t,u)+x\big)\big];$$

$$\sup_{\{s\in M\mid\delta(s,u)\leqslant d(t,u)\}} \mathbb{P}_s\big[\psi(u,t,\boldsymbol{X}) = 1\big] \leqslant B\exp\big[-a\big(d^2(t,u)-x\big)\big].$$

For M-tests, as given by Definition 3, Assumption 1 derives from part (A) of the more specific one that follows, setting $\psi(t,u,\boldsymbol{X}) = 1$ if $\gamma'(t,\boldsymbol{X}) - \gamma'(u,\boldsymbol{X}) > \tau x$ and $\psi(t,u,\boldsymbol{X}) = 0$ if $\gamma'(t,\boldsymbol{X}) - \gamma'(u,\boldsymbol{X}) < \tau x$.

Assumption 2. (A) There exists a subset M_T of M, a random function $\gamma'(\cdot,\boldsymbol{X})$ on M_T, a function δ from $M \times M_T$ to $[0,+\infty]$ and three constants $\tau, a, B > 0$ such that, for all $x \in \mathbb{R}$ and all pairs $(t,u) \in M_T^2$ with $t \neq u$,

$$\sup_{\{s\in M\mid\delta(s,t)\leqslant d(t,u)\}} \mathbb{P}_s\big[\gamma'(t,\boldsymbol{X}) - \gamma'(u,\boldsymbol{X}) \geqslant \tau x\big] \leqslant B\exp\big[-a\big(d^2(t,u)+x\big)\big].$$

(B) There exists a constant $\kappa' > 0$ such that, for all $x \in \mathbb{R}$, all $s \in M$ and all pairs $(t,u) \in M_T^2$ with $t \neq u$,

$$\mathbb{P}_s\big[\gamma'(t,\boldsymbol{X}) - \gamma'(u,\boldsymbol{X}) \geqslant \tau x\big] \leqslant B\exp\big[a\big(\kappa' d^2(s,t) - x\big)\big].$$

Under Assumption 1 (or 2(A)), we have to choose suitable values of x in order to get a well-defined family of tests. Given the weight function η on $S \subset M_T$, we shall always base our construction of T-estimators (or M-estimators), as explained in the previous sections, on the tests provided by these assumptions with $x = \eta^2(u) - \eta^2(t)$. It then follows that, for all $s \in M$ and $t,u \in S$ with $t \neq u$,

$$\sup_{\{s\in M\mid\delta(s,t)\leqslant d(t,u)\}} \mathbb{P}_s\big[\psi(t,u,\boldsymbol{X}) = 1\big] \leqslant B\exp\big[-a\big(d^2(t,u) - \eta^2(t) + \eta^2(u)\big)\big]. \tag{4.15}$$

Under Assumption 2(A), we get, for each pair $(t,u)\in S^2$, $t\neq u$, and γ given by (4.7),

$$\sup_{\{s\in M|\delta(s,t)\leqslant d(t,u)\}}\mathbb{P}_s\big[\gamma(t,\boldsymbol{X})\geqslant\gamma(u,\boldsymbol{X})\big]\leqslant B\exp\big[-a\big(d^2(t,u)-\eta^2(t)+\eta^2(u)\big)\big]. \tag{4.16}$$

Therefore the M-tests derived from γ' according to (4.7) and (4.8) also satisfy (4.15). Note that, in this case, the function $\tau\eta^2$ plays the role of the penalty for penalized maximum likelihood estimators or penalized least squares estimators. If, moreover, Assumption 2(B) holds, then

$$\mathbb{P}_s\big[\gamma(t,\boldsymbol{X})\geqslant\gamma(u,\boldsymbol{X})\big]\leqslant B\exp\big[a\big(\kappa' d^2(s,t)+\eta^2(t)-\eta^2(u)\big)\big]\quad\text{for all } s\in M. \tag{4.17}$$

One should view δ as a function measuring the robustness of the tests $\psi(t,u,\boldsymbol{X})$ with respect to deviations from the assumption that t obtains. If $\delta(s,t)=0$ the probability of rejecting t when s obtains is bounded by the right-hand side of (4.15) for all $u\neq t$ and this remains true as long as s remains "close enough" to t in the sense that $\delta(s,t)\leqslant d(t,u)$. If $\delta(s,t)$ is large, one can test t efficiently only against points u which are far away. In the simplest cases, and in particular those we consider in this paper, $\delta=\kappa d$ for some $\kappa>0$, but the introduction of a general δ (which, in particular, may take the value $+\infty$) proves useful in some special situations and does not involve any additional complication. Note also that not all (semi-)distances do suit our needs: the construction of tests that satisfy the previous assumption is only possible for some very special (semi-)distances.

4.3.2. Definition and elementary properties of D-models

In order to measure the massiveness of S and, more precisely, to bound the number of points of S that are contained in balls, we shall introduce the following notion of a *D-model* ("D" for discrete and dimension).

Definition 4. Let η, D and B' be positive numbers and S' be a subset of the semi-metric space (M,d). It will be called *a D-model with parameters η, D and B'* if

$$\big|S'\cap\mathcal{B}_d(t,x\eta)\big|\leqslant B'\exp[Dx^2]\quad\text{for all } x\geqslant 2 \text{ and } t\in M, \tag{4.18}$$

or equivalently

$$\big|S'\cap\mathcal{B}_d(t,r)\big|\leqslant B'\exp\big[D\big[(r/\eta)\vee 2\big]^2\big]\quad\text{for all } r>0 \text{ and } t\in M.$$

The number 2 has no magic meaning here and has been chosen for convenience. Other numbers would do and we could even parametrize this constant but this would lead to more complicated proofs and results without any substantial benefit. Finite sets do satisfy this assumption for suitable values of the parameters η, D and B' and lattices in Euclidean spaces as well. Further examples will be given in Section 6. Note that when the distance d is bounded, as is the case for Hellinger and variation distances, D-models are necessarily finite sets.

Some straightforward consequences of this definition to be used in the sequel, are as follows.

Lemma 1. *If S' is a D-model with parameters η, D and B', then it is at most countable and it is also a D-model with parameters η', D' and B' for all $\eta'>0$ and $D'=D[(\eta'/\eta)^2\vee 1]$. If, moreover, d is a distance and δ some function from $M\times S'$ to $[0,+\infty]$ such that $\delta(s,t)\geqslant\kappa d(s,t)$ for some positive κ, there exists a well-defined* minimum distance operator *π' from M to S' satisfying $\delta(s,\pi'(s))=\delta(s,S')=\inf_{t\in S'}\delta(s,t)$. In particular, one can define a* minimum distance operator *π from M to S' satisfying $d(s,\pi(s))=d(s,S')=\inf_{t\in S'}d(s,t)$.*

In order to check that S' is a D-model, the following result will sometimes be useful:

Lemma 2. *If d is a distance and*

$$\big|S'\cap\mathcal{B}_d(t,x\eta)\big|\leqslant B'\exp[Dx^2/4]\quad\textit{for all } x\geqslant 2 \textit{ and } t\in S', \tag{4.19}$$

then S' is a D-model with parameters η, D and B'.

Proof. If d is a distance and $S'\cap\mathcal{B}_d(t,x\eta)$ is not empty, it contains at least one point u and is therefore included in $S'\cap\mathcal{B}_d(u,2x\eta)$ with $u\in S'$. Hence (4.18) follows from (4.19). □

5. T-estimators based on a single D-model

In this section, we consider the simplest case of T-estimators, as defined in Section 4.2.1, i.e. those based on a single D-model S. This is a natural generalization of (non-penalized) m.l.e. estimators, as explained in Section 3.2, corresponding to the choice $\eta \equiv 0$. It then follows from (4.6) that the value of ε is irrelevant so that we can restrict our study to T_0-estimators. Then, (4.15) becomes

$$\sup_{\{s\in M|\delta(s,t)\leqslant d(t,u)\}} \mathbb{P}_s\big[\psi(t,u,\boldsymbol{X})=1\big] \leqslant B\exp\big[-ad^2(t,u)\big], \tag{5.1}$$

for all pairs (t,u), $t\neq u$ in S^2 and, in the case of M-tests, $\gamma=\gamma'$ and (4.16) becomes

$$\sup_{\{s\in M|\delta(s,t)\leqslant d(t,u)\}} \mathbb{P}_s\big[\gamma(t,\boldsymbol{X})\geqslant\gamma(u,\boldsymbol{X})\big] \leqslant B\exp\big[-ad^2(t,u)\big]. \tag{5.2}$$

5.1. Working within the general framework

Our aim is to prove some large deviation results for the minimizer(s) of $\mathcal{D}_{\boldsymbol{X}}$ or $\mathcal{D}'_{\boldsymbol{X}}$ which allow us, via (4.9) or (4.12), to derive the existence of T_0- or M-estimators and bound their risk. To this end, we have to introduce, for each integer $q\geqslant 1$ the function ζ_q defined in the following proposition to be proved in Appendix A.

Proposition 3. *Let Y be a non-negative random variable such that*

$$\mathbb{P}[Y>y]\leqslant\alpha\exp(-\beta y^2) \quad \textit{for } y\geqslant\bar{y}, \tag{5.3}$$

where α, β and $\bar{y}$ denote some positive constants. Then, for all $w\geqslant 0$ and $q\geqslant 1$,

$$\mathbb{E}\big[(Y+w)^q\big]\leqslant\big[1+\alpha\zeta_q(\beta\bar{y}^2)\big](\bar{y}+w)^q, \tag{5.4}$$

where ζ_q is the decreasing function defined on $(0,+\infty)$ by

$$\zeta_q(x)=\sqrt{\frac{\pi\,\mathrm{e}q}{2}}\left[\frac{q}{2\,\mathrm{e}x}\right]^{q/2}\mathbb{1}_{(0,cq)}(x)+\frac{q}{2}\,\mathrm{e}^{-x}\mathbb{1}_{[cq,+\infty)}(x); \quad c=\begin{cases}1/2 & \textit{if } q\leqslant 2\pi\mathrm{e};\\ 0.612 & \textit{if } q>2\pi\mathrm{e}.\end{cases} \tag{5.5}$$

We now have at hand all the required tools to state the main result of this section.

Theorem 3. *Let (M,d) be a semi-metric space for which Assumption 1 holds, let $S\subset M_T$ be a D-model with parameters η, D and B', $D\geqslant 1/2$, and let $\{\psi(t,u,\boldsymbol{X}),\ (t,u)\in S^2,\ t\neq u\}$ be a family of tests satisfying (5.1) with $2a\eta^2\geqslant 3D$. Then, for all $s\in M$ such that $\delta(s,S)<+\infty$, $\mathbb{P}_s$-a.s. there exists T_0-estimators $\hat{s}(\boldsymbol{X})$ derived from these tests and any of them satisfies, for any $s'\in S$,*

$$\mathbb{P}_s\big[d(s',\hat{s})>y\big]<2.2BB'\exp[-ay^2/6] \quad \textit{for } y\geqslant\big[\delta(s,s')\big]\vee(4\eta). \tag{5.6}$$

If d is a distance and $\delta=\kappa d$ for some $\kappa>0$, then

$$\mathbb{E}_s\big[d^q(s,\hat{s})\big]\leqslant(\kappa+1)^q\left[d(s,S)\vee\frac{4\eta}{\kappa}\right]^q\left[1+2.2BB'\zeta_q\left(\frac{8a\eta^2}{3}\right)\right] \textit{ for } q\geqslant 1, \tag{5.7}$$

with ζ_q given by (5.5). We get in particular, for all $s\in M$,

$$\mathbb{E}_s\big[d^q(s,\hat{s})\big]\leqslant(1+0.15qBB')(\kappa+1)^q\left[d(s,S)\vee\frac{4\eta}{\kappa}\right]^q \quad \textit{if } 1\leqslant q\leqslant\frac{16a\eta^2}{3}\wedge 17. \tag{5.8}$$

If Assumption 2(A) holds and the function γ satisfies (5.2), under the previous assumptions, there exists at least one minimizer of the function $\gamma(t,\boldsymbol{X})$ with respect to $t\in S$ and any such M-estimator $\hat{s}(\boldsymbol{X})$ satisfies (5.6), (5.7) and (5.8).

Remarks.

(i) The term involving BB' in (5.7) should be considered as a "remainder" term which is small for large values of $a\eta^2$.

(ii) Since $D \geqslant 1/2$, $16a\eta^2/3 \geqslant 4$ and (5.8) covers at least the case $1 \leqslant q \leqslant 4$.

(iii) The bound (5.7) reveals the relevance of the parameter B' which controls the remainder term. Indeed, it is easy to see that (4.18) is over parametrized since we could always normalize B' to one. This is clear if $B' < 1$ and, when $B' > 1$, we could merely change D to $D + (\log B')/4$. If B' is smaller than one, this would be a net loss for the risk bounds. In the opposite case, the modification improves the remainder term but deteriorates the main one because of the requirement $2a\eta^2 \geqslant 3D$ which typically forces us to enlarge η if we enlarge D. In situations where D is large, B' is compensated by $\zeta_q(8a\eta^2/3) \leqslant \zeta_q(4D)$ while enlarging η is not a good strategy. This accounts a posteriori for the introduction of B' in Definition 4, even if we shall often normalize it to one in the examples.

Proof of Theorem 3. We first want to show that (5.1) or (5.2) (in the case of M-tests) imply that

$$\left.\begin{array}{l}\mathbb{P}_s\big[\mathcal{D}_{\boldsymbol{X}}(s') > y\big] \\ \text{or} \\ \mathbb{P}_s\big[\mathcal{D}'_{\boldsymbol{X}}(s') > y\big]\end{array}\right\} < 2.2BB'\exp\left[-\frac{ay^2}{6}\right] \quad \text{for } y \geqslant y_0 = \delta(s,s') \vee (4\eta). \tag{5.9}$$

To prove this, we set $\theta = 5/4$ and $S_k = \{t \in S \mid \theta^{k/2}y \leqslant d(s',t) < \theta^{(k+1)/2}y\}$. Then

$$\begin{aligned}\mathbb{P}_s\big[\mathcal{D}_{\boldsymbol{X}}(s') > y\big] &= \mathbb{P}_s\big[\exists t \in S \text{ with } d(t,s') > y \text{ and } \psi(s',t,\boldsymbol{X}) = 1\big] \\ &\leqslant \sum_{k=0}^{+\infty} \mathbb{P}_s\big[\exists t \in S_k \text{ with } \psi(s',t,\boldsymbol{X}) = 1\big] \leqslant \sum_{k=0}^{+\infty} |S_k| \sup_{t\in S_k} \mathbb{P}_s\big[\psi(s',t,\boldsymbol{X}) = 1\big].\end{aligned} \tag{5.10}$$

Since $S_k \subset \mathcal{B}_d\big(s, \theta^{(k+1)/2}y\big) \cap S$ and $\theta^{(k+1)/2}y \geqslant \sqrt{\theta}y \geqslant \sqrt{\theta}y_0 > 2\eta$, it follows from (4.18) that $|S_k| \leqslant B'\exp[\theta^{k+1}(y/\eta)^2 D]$. If $t \in S_k$, then $d(s',t) \geqslant \theta^{k/2}y \geqslant y_0 \geqslant \delta(s,s')$ and we can use (5.1) to derive from (5.10) and $D \leqslant 2a\eta^2/3$ that $\mathbb{P}_s[\mathcal{D}_{\boldsymbol{X}}(s') > y] \leqslant G(y)$ with

$$G(y) = BB'\sum_{k=0}^{+\infty}\exp\left[\theta^{k+1}\frac{y^2}{\eta^2}D - a\theta^k y^2\right] \leqslant BB'\sum_{k=0}^{+\infty}\exp\left[-\frac{a\theta^k y^2}{3}(3-2\theta)\right].$$

Since $ay^2 \geqslant ay_0^2 \geqslant 16a\eta^2 \geqslant 24D \geqslant 12$, we finally get

$$\begin{aligned}G(y) &\leqslant BB'\exp\left[-\frac{ay^2}{6}\right]\sum_{k=0}^{+\infty}\exp\left[\frac{ay^2}{6}(1-\theta^k)\right] \leqslant BB'\exp\left[-\frac{ay^2}{6}\right]\sum_{k=0}^{+\infty}\exp\left[-2\left(\left(\frac{5}{4}\right)^k - 1\right)\right] \\ &< 2.2BB'\exp\left[-\frac{ay^2}{6}\right],\end{aligned}$$

which proves (5.9) for $\mathcal{D}_{\boldsymbol{X}}(s')$. If (5.2) holds, we replace $\psi(s',t,\boldsymbol{X}) = 1$ by $\gamma(t,\boldsymbol{X}) \leqslant \gamma(s',\boldsymbol{X})$ and show by the same arguments that $\mathbb{P}_s[\mathcal{D}'_{\boldsymbol{X}}(s') > y] \leqslant G(y)$.

Since $\mathcal{D}_{\boldsymbol{X}}(s')$ is a.s. finite and, by (4.9), the set of points $t \in S$ such that $\mathcal{D}_{\boldsymbol{X}}(t) \leqslant \mathcal{D}_{\boldsymbol{X}}(s')$ is a subset of $S \cap \overline{\mathcal{B}}_d(s', \mathcal{D}_{\boldsymbol{X}}(s'))$ which is finite by (4.18), it follows that $\{u \in S \mid \mathcal{D}_{\boldsymbol{X}}(u) = \inf_{t\in S}\mathcal{D}_{\boldsymbol{X}}(t)\}$ is non-empty and finite. Since S is countable, it is possible to select an element $\hat{s}$ of this set in a measurable way, which provides the required T_0-estimator $\hat{s}$. Then (5.6) derives from (5.9) and (4.10). If d is a distance we choose $s' = \pi(s)$ where π is the minimum distance operator provided by Lemma 1. Hence $d(s,s') = d(s,S)$ and (4.10) implies that $d(s,\hat{s}) \leqslant d(s,S) + \mathcal{D}_{\boldsymbol{X}}(s')$. We may then bound $\mathbb{E}_s[d^q(s,\hat{s})]$ from (5.9) via Proposition 3 with $Y = \mathcal{D}_{\boldsymbol{X}}(s')$, $\alpha = 2.2BB'$, $\beta = a/6$, $\bar{y} = y_0 = \kappa d(s,S) \vee 4\eta$ and $w = d(s,S)$, taking into account that

$$\bar{y} + w = \kappa\big[d(s,S) \vee (4\eta/\kappa)\big] + d(s,S) \leqslant (\kappa+1)\big[d(s,S) \vee (4\eta/\kappa)\big]. \tag{5.11}$$

Then (5.7) follows from (5.4) and (5.8) from the fact that $8a\eta^2/3 \geqslant 2$ and $2\pi\mathrm{e} > 17$. For M-tests satisfying (5.2), we proceed exactly in the same way, (4.14) implying that $\{t \in S \mid \gamma(t,\boldsymbol{X}) < \gamma(s',\boldsymbol{X})\}$ is finite, which proves the existence of a minimizer of $\gamma(\cdot,\boldsymbol{X})$. We then conclude as before, replacing (4.9) by (4.12) and (4.10) by (4.13). □

As an immediate consequence of (5.8), we can derive upper bounds for the minimax risk over subsets $\mathcal{S}$ of M. Let us denote the maximal risk of an estimator $\hat{s}(\boldsymbol{X})$ and the minimax risk over $\mathcal{S}$ respectively by

$$R(\hat{s},\mathcal{S},q)=\sup_{s\in\mathcal{S}}\mathbb{E}_s\bigl[d^q(s,\hat{s})\bigr]\quad\text{and}\quad R(\mathcal{S},q)=\inf_{\hat{s}}R(\hat{s},\mathcal{S},q), \tag{5.12}$$

where the infimum is over all possible estimators $\hat{s}$. Then

Corollary 1. *Let (M,d) be a metric space, S be a D-model with parameters η, D and B', $D\geqslant 1/2$, $\{\psi(t,u,\boldsymbol{X}),(t,u)\in S^2,\ t\neq u\}$ be a family of tests satisfying* (5.1) *with $\delta=\kappa d,\kappa>0$ and $2a\eta^2\geqslant 3D$ and $\hat{s}$ be a* T_0*-estimator* (*which exists a.s.*) *derived from these tests, then, for $1\leqslant q\leqslant(16a\eta^2/3)\wedge 17$,*

$$R(\mathcal{S},q)\leqslant R(\hat{s},\mathcal{S},q)\leqslant(1+0.15qBB')(\kappa+1)^q\left[\Bigl(\sup_{s\in\mathcal{S}}d(s,S)\Bigr)\vee\frac{4\eta}{\kappa}\right]^q. \tag{5.13}$$

5.2. Some particular statistical frameworks

To apply the previous results under Assumption 1, we just have to find a suitable D-model $S\subset M_T$ with parameters η and D satisfying $2a\eta^2/3\geqslant D\geqslant 1/2$. Assumption 1 actually holds for various statistical frameworks. To keep this paper to an acceptable size, we shall only consider three simple illustrations here, namely independent observations, Gaussian sequences and bounded regression. The case of Gaussian regression with random design has been considered in [12]. Further examples will be given in subsequent papers.

5.2.1. Independent observations with an unknown distribution

The independent setting. Here, we observe a set $\boldsymbol{X}=(X_1,\dots,X_n)$ of n independent random variables X_i with values in $(\mathcal{X},\mathcal{W})$. We denote by $\overline{M}$ the set of all distributions on $(\mathcal{X},\mathcal{W})$, set $\Xi=\mathcal{X}^n$, $\mathcal{Z}=\mathcal{W}^{\otimes n}$ and take for M the set of all product distributions on $(\mathcal{X}^n,\mathcal{W}^{\otimes n})$: $M=\{P_t=\bigotimes_{i=1}^n\overline{P}_i,\ \overline{P}_i\in\overline{M}\ \text{for } 1\leqslant i\leqslant n\}$. As indicated before, we identify t and P_t and M with the set of parameters t, denoting by P_s the true distribution of $\boldsymbol{X}$, which is assumed to belong to M.

If $P_t\in M$ is the distribution of an i.i.d. sample, we denote by $\overline{P}_t$ the common distribution of the X_i and, for simplicity, since there will be no ambiguity in the sequel, we shall also denote by $\overline{M}$ the subset of M consisting of those power distributions $P_t=\overline{P}_t^{\otimes n}$ so that the distributions P_t with $t\in\overline{M}$ are those for i.i.d. samples $(X_1,\dots,X_n)$ with marginal distributions $\overline{P}_t$ on $\mathcal{X}$. In this paper, we shall systematically restrict ourselves to considering models $S\subset\overline{M}$, which corresponds to the situation where we assume that our observations are independent and believe that they are close to i.i.d. but allow some departures from equidistribution.

To turn M into a metric space, we use either the sup-Hellinger distance $\bar{h}$ of the coordinates or the sup-variation distance $\bar{v}$. We recall that the Hellinger distance h is given by (2.1) and the variation distance v between two probabilities P and Q is defined by

$$v(P,Q)=\frac{1}{2}\int|\mathrm{d}P-\mathrm{d}Q|=\sup_A\bigl|P(A)-Q(A)\bigr|, \tag{5.14}$$

where the supremum is over all measurable sets. It is well-known from [43] that the two distances satisfy the inequalities

$$h^2(P,Q)\leqslant v(P,Q)\leqslant h(P,Q)\sqrt{2-h^2(P,Q)}. \tag{5.15}$$

If $P_t=\bigotimes_{i=1}^n\overline{P}_i$ and $P_u=\bigotimes_{i=1}^n\overline{Q}_i$, we define $\bar{h}$ and $\bar{v}$ on M by

$$\bar{h}(t,u)=\sup_{1\leqslant i\leqslant n}h(\overline{P}_i,\overline{Q}_i)\quad\text{and}\quad\bar{v}(t,u)=\sup_{1\leqslant i\leqslant n}v(\overline{P}_i,\overline{Q}_i).$$

In particular, if $u\in\overline{M}$, i.e. $P_u=\overline{P}_u^{\otimes n}$, then $\bar{h}(t,u)=\sup_{1\leqslant i\leqslant n}h(\overline{P}_i,\overline{P}_u)$ and $\bar{v}(t,u)=\sup_{1\leqslant i\leqslant n}v(\overline{P}_i,\overline{P}_u)$. If both t and $u\in\overline{M}$, then

$$\bar{h}(t,u)=h(\overline{P}_t,\overline{P}_u)\quad\text{and}\quad\bar{v}(t,u)=v(\overline{P}_t,\overline{P}_u), \tag{5.16}$$

which allows us to identify $\bar{h}$ with h and $\bar{v}$ with v on $\overline{M}$ and turn it into a metric space with distance either h or v.

The i.i.d. setting. It corresponds to the particular case where we only consider distributions for i.i.d. samples or equivalently restrict ourselves to $t \in \overline{M}$, as defined for the independent setting, and choose for M either $\overline{M}$ itself or some subset of it, with the metric given either by h or v. For instance, we may take for M the set of all probability densities t with respect to some measure μ on $\mathcal{X}$ and set $\mathrm{d}\overline{P}_t/\mathrm{d}\mu = t$, hence $P_t = (t \cdot \mu)^{\otimes n}$.

5.2.2. The Gaussian setting

The Gaussian setting corresponds to the so-called "Gaussian sequence" framework in which we observe a sequence $\boldsymbol{X} = (X_i)_{i\geqslant 1}$ of independent Gaussian variables with known variance σ^2 and respective means s_i. Then $\Xi = \mathbb{R}^{\mathbb{N}^\star}$, $X_i \sim \mathcal{N}(s_i, \sigma^2)$ for each i and $s = (s_i)_{i\geqslant 1} \in M = \boldsymbol{l}_2(\mathbb{N}^\star)$. We denote by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ respectively the scalar product and the norm in $\boldsymbol{l}_2(\mathbb{N}^\star)$, by d_2 the corresponding distance ($d_2(s,t) = \|s-t\|$) and P_s the true distribution of $\boldsymbol{X}$. All possible distributions P_t for $\boldsymbol{X}$, with $t \in \boldsymbol{l}_2(\mathbb{N}^\star)$, being mutually absolutely continuous, we can choose the centered distribution $P_0 = \overline{P}_0^{\otimes\mathbb{N}^\star}$ with $\overline{P}_0 = \mathcal{N}(0,\sigma^2)$ for reference measure, getting

$$\frac{\mathrm{d}P_t}{\mathrm{d}P_0}(\boldsymbol{X}) = \exp\left[\frac{1}{\sigma^2}\left(\langle t, \boldsymbol{X}\rangle - \frac{\|t\|^2}{2}\right)\right]. \tag{5.17}$$

Although the case of $X_i \sim \mathcal{N}(s_i,\sigma^2)$ with a known value of σ can be reduced to the case of $X_i/\sigma \sim \mathcal{N}(s_i/\sigma, 1)$, it will be more instructive to give our results within the original framework in order to emphasize the influence of σ.

The Gaussian setting is merely an infinite-dimensional extension of the classical problem of estimating the mean s of a Gaussian vector with known covariance matrix in $\mathbb{R}^n$ which can be viewed as a particular case of the Gaussian setting with $s_i = 0$ for $i > n$. We recover the classical Gaussian linear regression framework if we assume that s belongs to some given linear subspace of $\mathbb{R}^n$.

Alternatively, the Gaussian setting can be identified with the classical "white noise framework" which corresponds to the observation of the process

$$Y(z) = \int_0^z s(x)\,\mathrm{d}x + \sigma W(z), \quad 0 \leqslant z \leqslant 1, \tag{5.18}$$

where s is an unknown function in $\mathbb{L}_2([0,1],\mathrm{d}x)$ and W is a Wiener process with $W(0) = 0$. Choosing some orthonormal basis $\{\varphi_i,\ i \geqslant 1\}$ of $\mathbb{L}_2([0,1],\mathrm{d}x)$ and defining $s_i = \int_0^1 s(x)\varphi_i(x)\,\mathrm{d}x$, $X_i = \int_0^1 \varphi_i(x)\,\mathrm{d}Y(x)$ leads to the Gaussian setting. The function s in (5.18) can be identified with the sequence $(s_i)_{i\geqslant 1}$ of its Fourier coefficients with respect to the basis $\{\varphi_i,\ i \geqslant 1\}$ via Plancherel's formula. Since this correspondence is an isometry, it allows us to view the white noise framework (5.18) as an alternative representation of the Gaussian setting with parameter space $M = \mathbb{L}_2([0,1],\mathrm{d}x)$ and distance d corresponding to the $\mathbb{L}_2$-norm. Much more on this is to be found in Sections 1 and 6 of [17].

5.2.3. The bounded regression setting

This statistical problem has recently received much attention in view of his connections with the fashionable domain of Statistical Learning. A major reference is the book [33] by Györfi, Kohler, Kryżak and Walk and a very recent one including many additional references is [26] by DeVore, Kerkyacharian, Picard and Temlyakov.

As in the case of independent variables, we can distinguish between two situations.

Bounded regression with random design. We observe an n-sample $\{(X_i,Y_i),\ 1 \leqslant i \leqslant n\}$, $X_i \in \mathcal{X}$, $Y_i \in I$ from some unknown distribution on $\mathcal{X}\times I$ where I is a *known* compact interval of $\mathbb{R}$. Performing if necessary an affine transform of the Y_i, we may assume without loss of generality that $I = [0,1]$, hence $\Xi = (\mathcal{X}\times[0,1])^n$, which we shall do throughout this paper. The problem would perfectly fit into our i.i.d. setting if the unknown parameter to be estimated were the joint distribution of X and Y, but here we focus on the estimation of the conditional mean s of Y given X, i.e. $s(x) = \mathbb{E}_s[Y|X=x] \in [0,1]$, denoting by μ the *unknown* marginal distribution of X on $\mathcal{X}$. We may therefore rewrite this statistical framework in regression form as

$$Y = s(X) + \xi \quad \text{with } X \sim \mu, \qquad Y, s(X) \in [0,1] \quad \text{and} \quad \mathbb{E}_s[\xi|X] = 0. \tag{5.19}$$

It then corresponds to random design regression with bounded observations, which is a classical framework used in Statistical Learning. Here s is the only parameter to be estimated but μ and the conditional distribution of ξ given X

are unknown nuisance parameters. This is the only case we shall consider in this paper for which the mapping F such that $s = F(P_{\boldsymbol{X}})$ is not one-to-one so that s does not determine $P_{\boldsymbol{X}}$.

Bounded regression with fixed design. It is the same framework as before apart from the fact that, instead of being i.i.d., the variables $X_1, \dots, X_n$ are now fixed (deterministic), equal to $x_1, \dots, x_n$, so that we have independent observations $Y_1, \dots, Y_n$ satisfying

$$Y_i = s(x_i) + \xi_i \quad \text{with } Y_i,\ s(x_i) \in [0,1] \quad \text{and} \quad \mathbb{E}_s[\xi_i] = 0. \tag{5.20}$$

This situation occurs in particular when we analyze the random design problem conditionally to the values of the X_i. We do not assume here that all values x_i are distinct so that the cardinality of the set $\mathcal{X} = \{x_1, \dots, x_n\}$ may be smaller than n.

A unified framework. In order to save space and avoid a lot of redundancies, we shall treat both regression cases (random and fixed design) simultaneously, using the following conventions: In the random design case, M is the set of measurable functions from $\mathcal{X}$ to $[0,1]$ with $\mathbb{L}_2(\mu)$-norm $\|\cdot\| = \|\cdot\|_2$ and the corresponding distance $d = d_2$. We also set $\boldsymbol{X} = \{(X_i, Y_i),\ 1 \leqslant i \leqslant n\}$ and $\gamma'(t, \boldsymbol{X}) = \sum_{i=1}^n [Y_i - t(X_i)]^2$. In the fixed design case, M is the set of functions from $\mathcal{X} = \{x_1, \dots, x_n\}$ to $[0,1]$, which can be identified to the metric space $[0,1]^{|\mathcal{X}|}$ with the distance $d = d_n$ defined by $d_n^2(t,u) = n^{-1} \sum_{i=1}^n [t(x_i) - u(x_i)]^2$ and the corresponding norm $\|\cdot\| = \|\cdot\|_n$ with $\|t-u\|_n = d_n(t,u)$. Then $\boldsymbol{X} = \{(x_i, Y_i),\ 1 \leqslant i \leqslant n\}$ and $\gamma'(t, \boldsymbol{X}) = \sum_{i=1}^n [Y_i - t(x_i)]^2$.

5.3. Application of the general theory to the particular frameworks

5.3.1. The existence of robust tests

Gaussian sequences. It is not difficult to check that Assumption 2 holds in the Gaussian setting since then likelihood ratio tests are naturally robust as shown by the following proposition.

Proposition 4. *Let $\boldsymbol{X} = (X_i)_{i \geqslant 1} \in \mathbb{R}^{\mathbb{N}^\star}$ be a random sequence with independent Gaussian coordinates of variance σ^2 and mean vector belonging to $\boldsymbol{l}_2(\mathbb{N}^\star)$. Let P_t denote the distribution of $\boldsymbol{X}$ when the mean vector is t. Then, for all $s, t, u \in \boldsymbol{l}_2(\mathbb{N}^\star)$ and $z \in \mathbb{R}$,*

$$\mathbb{P}_s\left[\log\left(\frac{\mathrm{d}P_u}{\mathrm{d}P_t}\right)(\boldsymbol{X}) \geqslant z\right] \leqslant \exp\left[-\frac{z}{2} - \frac{\|t-u\|(\|t-u\| - 4\|s-t\|)}{8\sigma^2}\right].$$

In particular, for all $x \in \mathbb{R}$,

$$\sup_{\{s \in \boldsymbol{l}_2(\mathbb{N}^\star) \,|\, \|s-t\| \leqslant \|t-u\|/6\}} \mathbb{P}_s\left[\log\left(\frac{\mathrm{d}P_u}{\mathrm{d}P_t}\right)(\boldsymbol{X}) \geqslant \frac{x}{12\sigma^2}\right] \leqslant \exp\left[-\frac{\|t-u\|^2 + x}{24\sigma^2}\right]$$

and, for all $s \in \boldsymbol{l}_2(\mathbb{N}^\star)$ and $x \in \mathbb{R}$,

$$\mathbb{P}_s\left[\log\left(\frac{\mathrm{d}P_u}{\mathrm{d}P_t}\right)(\boldsymbol{X}) \geqslant \frac{x}{12\sigma^2}\right] \leqslant \exp\left[\frac{\|t-s\|^2}{2\sigma^2} - \frac{x}{24\sigma^2}\right].$$

Bounded regression. The case of bounded regression is only slightly more complicated. An application of Bernstein's Inequality actually leads to the following:

Proposition 5. *Let $\boldsymbol{X}$, M, $\|\cdot\|$ and $\gamma'(\cdot, \boldsymbol{X})$ be defined according to the conventions of Section* 5.2.3 *for bounded regression with either random or fixed design. For all $s, t, u \in M$ and $z \in \mathbb{R}$, if $y = 4\|s-t\|^2 - \|t-u\|^2/4$, then*

$$\mathbb{P}_s\big[\gamma'(t, \boldsymbol{X}) - \gamma'(u, \boldsymbol{X}) \geqslant nz\big] \leqslant \exp\left[\frac{-3n}{100}\left(\|t-u\|^2 + \frac{98(z-y)}{25}\right)\right]. \tag{5.21}$$

In particular, for all $x \in \mathbb{R}$,

$$\sup_{\{s \in M \,|\, \|s-t\| \leqslant \|t-u\|/4\}} \mathbb{P}_s\left[\gamma'(t, \boldsymbol{X}) - \gamma'(u, \boldsymbol{X}) \geqslant \frac{25nx}{98}\right] \leqslant \exp\left[\frac{-3n}{100}\big(\|t-u\|^2 + x\big)\right], \tag{5.22}$$

and, for all $s \in M$ and $x \in \mathbb{R}$,

$$\mathbb{P}_s\left[\gamma'(t,\boldsymbol{X}) - \gamma'(u,\boldsymbol{X}) \geqslant \frac{25nx}{98}\right] \leqslant \exp\left[\frac{3n}{100}\left(\frac{392\|s-t\|^2}{25} - x\right)\right]. \tag{5.23}$$

Remark. It is easily seen, either from the proof or from a scaling argument, that, if the Y_i take their values in $[0, A]$ instead of $[0, 1]$ and M is defined accordingly as a set of functions with values in $[0, A]$, all the previous bounds hold with $3n/100$ replaced by $3n/(100A^2)$. This means that, in all subsequent results dealing with bounded regression, the value of a should be changed from $3n/100$ to $3n/(100A^2)$. In order to save space, we shall not pursue into this direction leaving such extensions to the reader.

Independent observations. In the general independent setting, likelihood ratio tests are not robust, as we have seen, and one has to introduce special tests for our purposes. They have been constructed by Huber in [34] (see also [35], Section 10.3) for the variation distance and by Le Cam in [44] and Birgé in [9] for the Hellinger distance.

Proposition 6. *Let $\overline{P}_t$, $\overline{P}_u$ be two different distributions on some measurable space $\mathcal{X}$ and $x \in \mathbb{R}$, d be either the Hellinger or the variation distance between probabilities on $\mathcal{X}$ and $\alpha = 1$ if $d = h$ or $\alpha = 2$ if $d = v$. One can find a test function ψ (depending on α, t, u and x) defined on $\mathcal{X}^n$, with $\psi(t,u,\boldsymbol{X}) = 1 - \psi(u,t,\boldsymbol{X})$, such that, if $\boldsymbol{X} = (X_1,\dots,X_n)$ is a set of independent random variables with distribution $P_s = \bigotimes_{i=1}^n \overline{P}_i$, then*

$$\mathbb{P}_s\big[\psi(t,u,\boldsymbol{X}) = 1\big] \leqslant \exp\big[-n\big(d^2(t,u) + x\big)/(4\alpha)\big] \quad \text{if} \sup_{1\leqslant i\leqslant n} d(\overline{P}_i, \overline{P}_t) \leqslant d(t,u)/4$$

and

$$\mathbb{P}_s\big[\psi(u,t,\boldsymbol{X}) = 1\big] \leqslant \exp\big[-n\big(d^2(t,u) - x\big)/(4\alpha)\big] \quad \text{if} \sup_{1\leqslant i\leqslant n} d(\overline{P}_i, \overline{P}_u) \leqslant d(t,u)/4.$$

The proofs of the previous propositions are given in Appendix A.

5.3.2. The resulting risk bounds for T-estimators

It follows from the previous section that, in all the frameworks we considered, it is possible to construct families of tests satisfying Assumption 1 (or 2 for the Gaussian and bounded regression settings) with $B = 1$ for suitable values of a and $\delta = \kappa d$. We can take $M_T = M$ in all cases, except for the independent non i.i.d. one for which $M_T = \overline{M}$. Applying Theorem 3 in each case leads to the following corollary which covers all the specific frameworks we consider in this paper.

Corollary 2. *Suppose that we want to estimate an unknown element $s \in M$ in the independent, the Gaussian or the bounded regression setting and that $S \subset M_T$ is a D-model with parameters η, D and B'.*

(i) *For independent observations, $M_T = \overline{M}$ and one can build tests that satisfy Assumption 1 with $B = 1$, $a = n/4$ and $\delta = 4\bar{h}$ in the Hellinger case or $a = n/8$ and $\delta = 4\bar{v}$ in the variation case.*

(ii) *In the Gaussian setting, $M_T = M$ and Assumption 2 holds with $B = 1$,*

$$\gamma'(t,\boldsymbol{X}) = \langle t,\boldsymbol{X}\rangle - \frac{\|t\|^2}{2}, \quad \tau = \frac{1}{12\sigma^2}, \quad a = \frac{1}{24\sigma^2}, \quad \delta = 6d_2 \quad \text{and} \quad \kappa' = 12.$$

(iii) *For bounded regression with either random or fixed design, $M_T = M$ and Assumption 2 holds with $B = 1$, $\tau = 25n/98$, $\kappa' = 15.68$,*

$$\gamma'(t,\boldsymbol{X}) = \sum_{i=1}^n \big[Y_i - t(X_i)\big]^2 \quad \text{or} \quad \sum_{i=1}^n \big[Y_i - t(x_i)\big]^2, \quad a = \frac{3n}{100} \quad \text{and} \quad \delta = 4d_2 \text{ or } 4d_n.$$

Therefore, if $2a\eta^2/3 \geqslant D \geqslant 1/2$, we can, in each case, build T_0-estimators $\hat{s}$ that satisfy (5.6), (5.7) *and* (5.8) *for all $s \in M$ with $B = 1$ and the relevant values of d, a and κ. For the Gaussian and bounded regression settings, the corresponding M-estimators satisfy the same results.*

Remark. In the independent and bounded regression settings, the distance d is bounded by one and the risk of any estimator with values in M as well. But it may well happen that the upper bounds in (5.7) and (5.8) be larger than one, especially for small values of n, because our method is far from optimal at the level of constants. Therefore these upper bounds and the ones which we shall get later should systematically be truncated to one. In order to simplify the presentation of our results we shall, most of the time, omit to do this explicitly.

5.3.3. An application to the model of uniform distributions

In order to illustrate the relationship between the classical approach and ours, let us go back to the problem we considered in Section 2.3 and suppose that we want to estimate some distribution on $\mathbb{R}_+$ from n independent observations $X_1, \ldots, X_n$ via the model of uniform distributions $\mathcal{U}_\theta$ on $[0,\theta]$ with $\theta > 0$. When we say that we use this model, this means that we believe that the true distribution $\overline{P}_i$ of X_i is close to some $\mathcal{U}_\theta$ (independent of i) but we do not assume that $\overline{P}_i$ belongs to the model. It will be convenient here to reparametrize the uniform distributions, denoting by $\overline{P}_t$, $t \in \mathbb{R}$ the uniform distribution on $[0, \mathrm{e}^t]$, since then

$$h^2(t,u) = h^2(\overline{P}_t, \overline{P}_u) = 1 - \exp\bigl(-|t-u|/2\bigr) \leqslant |t-u|/2. \tag{5.24}$$

Given some $\alpha \in \mathbb{R}$, we shall set, for $D \geqslant 1/2$,

$$\eta^2 = 16.8D/n; \qquad J = \sup\bigl\{ j \in \mathbb{N} \mid j \leqslant 4.5\exp\bigl[(4D) \vee (n/84)\bigr]\bigr\}; \tag{5.25}$$

$$I = [\alpha, \alpha + 4J\eta^2] \quad \text{and} \quad S = \bigl\{\alpha + 2\eta^2(1+2j),\ j \in \mathbb{N},\ j \leqslant J-1\bigr\}. \tag{5.26}$$

It follows from (5.24) that $\inf_{t\in S} h(\overline{P}_t, \overline{P}_u) \leqslant \eta$ for all $u \in I$ and consequently that, whatever the distribution $P_s = \bigotimes_{i=1}^n \overline{P}_i$ of $\boldsymbol{X}$,

$$\inf_{t\in S} \bar{h}^2(P_s, P_t) = \inf_{t\in S} \sup_{1\leqslant i\leqslant n} h^2(\overline{P}_i, \overline{P}_t) \leqslant 2\Bigl[\eta^2 + \inf_{t\in I} \sup_{1\leqslant i\leqslant n} h^2(\overline{P}_i, \overline{P}_t)\Bigr]. \tag{5.27}$$

In order to check that S is a D-model, we shall apply the following lemma. Its conclusions go beyond what we need here but they will prove useful later.

Lemma 3. *Let η and S be defined by* (5.25) *and* (5.26)*. Then, whatever the probability $P = \bigotimes_{i=1}^n \overline{Q}_i \in M$,*

$$\bigl|S \cap \mathcal{B}_{\bar{h}}(P, r)\bigr| \leqslant 4.5\exp\bigl[D\bigl[(r/\eta) \vee 2\bigr]^2\bigr] \quad \textit{for all } r > 0. \tag{5.28}$$

Proof. We shall distinguish between two situations. When $r^2 \geqslant 1/5$, then $(r/\eta)^2 \geqslant n/(84D)$ and (5.28) follows from (5.25) since $|S| = J$. If $r^2 < 1/5$, there is obviously nothing to prove if $r \leqslant \inf_{t\in S} \bar{h}(P, P_t)$ and we can therefore assume that there exists some $t \in S$ such that $\mathcal{B}_{\bar{h}}(P,r) \subset \mathcal{B}_{\bar{h}}(P_t, 2r)$. If P_u belongs to $\mathcal{B}_{\bar{h}}(P_t, 2r)$, it follows from (5.24) that $1 - \exp(-|t-u|/2) < 4r^2$, hence $(4\eta^2)^{-1}|t-u| < -(2\eta^2)^{-1}\log(1-4r^2)$ and the definition of S implies that

$$\bigl|S \cap \mathcal{B}_{\bar{h}}(P,r)\bigr| \leqslant \bigl|S \cap \mathcal{B}_{\bar{h}}(P_t, 2r)\bigr| < -\eta^{-2}\log(1-4r^2) + 1 < (5\log 5)(r/\eta)^2 + 1$$

since $-\log(1-4r^2) < (5\log 5)r^2$ for $r^2 < 1/5$. Finally (5.28) follows from the lower bound $D \geqslant 1/2$. □

The lemma implies that S is a D-model with parameters η, D and 4.5. We can therefore apply Corollary 2 to S with $D = 1/2$, $\eta^2 = 8.4/n$, $B' = 4.5$ and get, in view of (5.27), the risk bound

$$\mathbb{E}_s\bigl[\bar{h}^2(s,\hat{s})\bigr] \leqslant C\Bigl[\inf_{\theta\in\Theta} \sup_{1\leqslant i\leqslant n} h^2(\overline{P}_i, \mathcal{U}_\theta) + n^{-1}\Bigr],$$

where Θ denotes the interval $[\exp(\alpha), \exp(\alpha + 4J\eta^2)]$. One should note here that it is necessary to put some restriction on the length of Θ: if it were infinite, the set S would also be infinite and Assumption 2 could not hold since any Hellinger ball of radius one contains S.

6. Metric dimension and minimax risk

If we consider a stochastic framework for which Assumption 1 holds with $M_T = M$, we can bound the minimax risk over subsets $\mathcal{S}$ of M via Corollary 1. Optimizing the upper bound in (5.13) for given values of q, κ, B and B' amounts

to minimize $[(\kappa/4)\sup_{s\in\mathcal{S}} d(s,S)]\vee\eta$ with respect to those S and η such that S is a D-model with parameters η, D and B' and $\eta^2\geqslant 3D/(2a)$. Since, by Lemma 1, for any $\eta'>\eta$, one can always replace the pair (η,D) by (η',D') satisfying the same relationship, one can assume that $\eta\geqslant(\kappa/4)\sup_{s\in\mathcal{S}} d(s,S)$. Therefore, we have to look for the minimal value of η such that there exists some D-model S with parameters η, D and B' satisfying $\eta^2\geqslant 3D/(2a)$ and $\eta\geqslant(\kappa/4)d(s,S)$ for all $s\in\mathcal{S}$.

It is one purpose of Approximation Theory to find sets with prescribed approximation properties and controlled massiveness. For instance the entropy numbers of the compact set $\mathcal{S}$ can be used to build suitable D-models, although entropy is not the most adequate tool in our case, as we shall see. Often Approximation Theory provides simple sets (like finite dimensional linear spaces) S' which can be used to approximate the elements of $\mathcal{S}$ with prescribed accuracy: $\sup_{s\in\mathcal{S}} d(s,S')\leqslant\varepsilon$, but are not D-models and cannot therefore be used directly for our construction. Some additional step is needed in order to apply the classical results of Approximation Theory to the construction of T-estimators. Similar arguments are required to discretize in a suitable way the sets of densities that are used in parametric estimation as illustrated in Section 5.3.3. It is therefore important to understand how one can derive T-estimators from "natural" approximation spaces or simple parametric families.

6.1. Introducing metric dimensions

The previous reasoning assumed a fixed value of B' and, as we noticed in Section 5.1, there are some possible balances in (4.18) between B' and D, hence η. Playing with all three parameters would make everything more complicated in what follows and we shall set $B'=1$ for the remainder of Section 6. If one wants to use the influence of B' efficiently, it is better to go back to the general point of view we took in Section 5.

Let us now recall the following definitions from Approximation Theory.

Definition 5. Let (M,d) be a metric space. A subset S of M is η-separated if $d(t,u)>\eta$ for all pairs $(t,u)\in S^2$ with $t\neq u$; it is called an η-net for $\mathcal{S}\subset M$ if, for all $s\in\mathcal{S}$, one can find $t\in S$ such that $d(s,t)\leqslant\eta$. An η-separated subset S of $\mathcal{S}\subset M$ is said to be maximal (in $\mathcal{S}$) if any S' with $S\subsetneq S'\subset\mathcal{S}$ is not η-separated.

Observe that an η-net for $\mathcal{S}$ needs not be a subset of $\mathcal{S}$ and that a maximal η-separated subset of $\mathcal{S}$ is an η-net for $\mathcal{S}$.

Given $\mathcal{S}\subset M$, the values of η such that there exists an η-net S_η for $\mathcal{S}$ which is a D-model with parameters η, D, 1 and $1/2\leqslant D\leqslant 2a\eta^2/3$ only depend on some metric properties of $\mathcal{S}$ that describe its "massiveness". In view of Definition 4, it may seem natural to characterize this massiveness by the function D' defined on $(0,+\infty)$ by

$$D'(\eta)=\eta^2\inf_{S_\eta}\sup_{t\in M;\ r\geqslant 2\eta} r^{-2}\log\bigl(\bigl|S_\eta\cap\mathcal{B}(t,r)\bigr|\bigr),$$

where the infimum is over all η-nets S_η for $\mathcal{S}$. This function can be degenerate ($D'(\eta)=+\infty$ for all $\eta>0$), for instance when $\mathcal{S}=\boldsymbol{l}_2(\mathbb{N}^\star)$, in which case it is of no use. Moreover, it can behave in a rather erratic way: when $\mathcal{S}=\mathbb{Z}\subset M=\mathbb{R}$, one can show that $D'(\eta)\leqslant\eta^2\log 3$ for $\eta<1/2$ and $D'(1/2)\geqslant(\log 2)/4$. This example also illustrates the difficulty to compute the function D' even in the simplest situations. Apart from some quite exceptional cases (if $\mathcal{S}=\mathbb{R}$, then $D'(\eta)=(\log 3)/4$ for all $\eta>0$), it is impossible to compute it precisely and the best one can do is to bound it from above and below. It will therefore be more convenient here to work with some upper bound $\widetilde{D}$ for D' that we call *a bound for the metric dimension* of $\mathcal{S}$.

Definition 6. Let $\mathcal{S}$ be a subset of some metric space (M,d) and $\widetilde{D}$ be a right-continuous function from $(0,+\infty)$ to $[1/2,+\infty]$, such that $\widetilde{D}(x)=o(x^2)$ when $x\to+\infty$. We say that $\mathcal{S}$ has *a metric dimension bounded by* $\widetilde{D}$ if, for every $\eta>0$, there exists an η-net S_η for $\mathcal{S}$ which is a D-model with parameters η, $\widetilde{D}(\eta)$ and 1. If one can choose $\widetilde{D}(\eta)=\overline{D}\in[1/2,+\infty)$ for all $\eta>0$, we say that $\mathcal{S}$ has *a finite metric dimension bounded by* $\overline{D}$. We shall speak of *inner metric dimension bounded by* $\widetilde{D}$ (or by $\overline{D}$) if $S_\eta\subset\mathcal{S}$ for all $\eta>0$.

The restrictions that $\widetilde{D}$ be right-continuous and $\geqslant 1/2$ will avoid to introduce additional assumptions in the sequel and we can always enlarge $\widetilde{D}$ to get them. It is easily seen that $\widetilde{D}(\lambda x)\leqslant\lambda^2\widetilde{D}(x)$ for $\lambda>1$ and we do not know of any example where the condition $\widetilde{D}(x)=o(x^2)$ when $x\to+\infty$ is not satisfied. The introduction of the inner metric dimension will simplify our further analysis. The following properties are straightforward.

Proposition 7. *If $\mathcal{S}$ has a metric dimension bounded by $\widetilde{D}$, this remains true for any subset of $\mathcal{S}$ and $\mathcal{S}$ has an inner metric dimension $\widetilde{D}'(\eta)$ bounded by $(25/4)\widetilde{D}(\eta/2)$.*

Proof. Let S'_η be a maximal η-separated subset of $\mathcal{S}$ (hence an η-net for $\mathcal{S}$) and S be an $\eta/2$-net for $\mathcal{S}$ which is a D-model with parameters $\eta/2$, $\widetilde{D}(\eta/2)$ and 1. Then S is an $\eta/2$-net for S'_η so that $|S'_\eta \cap \mathcal{B}(t, x\eta)| \leqslant |S \cap \mathcal{B}(t, (x+1/2)\eta)|$. The result follows since $x \geqslant 2$ in the definition of the metric dimension. □

6.2. Some historical remarks

The fact that the minimax risk over $\mathcal{S}$ can be bounded using its metric properties has already been recognized in the early seventies by Le Cam ([43] and [44]) who introduced the following notion of *metric dimension* to measure the massiveness of a set: he defined the $D(\eta)$-metric dimension of $\mathcal{S}$ as the smallest number z such that any set in $\mathcal{S}$ with diameter $2x \geqslant 2\eta$ can be covered by no more than 2^z sets of diameter not larger than x. In [8] (Assumption H1, p. 186) we introduced a slightly different notion of metric dimension, bounding the maximal number of points of some η-net contained in an arbitrary ball of radius $2^j\eta$ for $j \geqslant j_0 \geqslant 1$ by $2^{jD(\eta)}$. This is actually very similar to (4.18), apart from the fact that in our assumption we replaced x^D by the less restrictive $\exp(x^2 D)$. The initial definitions may seem more natural because both were inspired by the example of k-dimensional Euclidean spaces. If $\mathcal{S}$ is such a space, any ball of radius 2η can be covered by $2^{c_1 k}$ balls of radius η and there exists an η-net S_η for $\mathcal{S}$ such that $|S_\eta \cap \mathcal{B}(t,r)| \leqslant (r/\eta)^{c_2 k}$ for all $r \geqslant 2\eta$ and $t \in \mathcal{S}$. Apart from the constants c_1, c_2, these bounds are optimal. Changing these definitions to Assumption 2 gives slightly more flexibility and simplifies the proofs. There is a little cost for that at the level of constants but this is a minor point which does not change anything to the philosophy of our approach.

6.3. Bounds for the risk based on metric dimensions

The importance of metric dimensions follows from the fact that if $\mathcal{S}$ has a metric dimension bounded by $\widetilde{D}$ one can find, for any $\eta > 0$ such that $\widetilde{D}(\eta) < +\infty$, an η-net for $\mathcal{S}$ which is a D-model. As a consequence, one can bound the minimax risk on $\mathcal{S}$ as soon as one can get a bound for its metric dimension.

Theorem 4. *Assume that (M,d) is a metric space such that there exists, for each pair $(t,u) \in M^2$, $t \neq u$, a test $\psi(t,u,\boldsymbol{X})$ between t and u satisfying the error bound*

$$\mathbb{P}_s\big[\psi(t,u,\boldsymbol{X}) = 1\big] \leqslant \exp\big[-ad^2(t,u)\big] \quad \text{if } d(s,t) \leqslant \kappa^{-1} d(t,u),$$

for some $a > 0$ and $\kappa \geqslant 4$, independent of s,t,u. Let $\mathcal{S}$ be some subset of M with a metric dimension bounded by $\widetilde{D}$ and set $\tilde{\eta} = \inf\{x > 0 \mid 2ax^2 \geqslant 3\widetilde{D}(x)\}$. There exists a T-estimator $\hat{s}$ such that, for all $s \in \mathcal{S}$,

$$\mathbb{P}_s\big[d(s,\hat{s}) > (x+1)\tilde{\eta}\big] < 2.2\exp(-ax^2\tilde{\eta}^2/6) \quad \text{for } x \geqslant \kappa \tag{6.1}$$

and, for all $s \in M$,

$$\mathbb{E}_s\big[d^q(s,\hat{s})\big] \leqslant C(q)(\kappa+1)^q\big[d(s,\mathcal{S}) + \tilde{\eta}\big]^q \quad \text{for } q \geqslant 1.$$

In particular, the minimax risk $R(\mathcal{S},q)$ over $\mathcal{S}$ is bounded by $C(q)(\kappa+1)^q\tilde{\eta}^q$.

Proof. The behaviour of $\widetilde{D}$ when $x \to +\infty$ implies that $\tilde{\eta}$ is finite and its right-continuity that one can find a D-model S with parameters $\tilde{\eta}$, $D = \widetilde{D}(\tilde{\eta}) \leqslant 2a\tilde{\eta}^2/3$ and 1 which is an $\tilde{\eta}$-net for $\mathcal{S}$. We can therefore apply Theorem 3 with $\eta = \tilde{\eta}$ and $d(s,s') \leqslant \tilde{\eta}$. Since $B = 1$ and $\delta(s,s') \vee (4\tilde{\eta}) \leqslant \kappa\tilde{\eta}$, (6.1) follows from (5.6). The risk bound then follows from (5.7) since $\zeta_q\big(8a\tilde{\eta}^2/3\big) \leqslant \zeta_q(2)$. □

Such a theorem actually emphasizes a robustness property of T-estimators that is definitely not shared by the classical m.l.e. as we demonstrated in Section 2.3: even if s does not belong to $\mathcal{S}$, the T-estimators constructed in view of bounding the minimax risk over $\mathcal{S}$ still have a risk which remains under control. Up to constants depending on q only, we get the bound for the minimax risk plus an extra $d^q(s,\mathcal{S})$ term corresponding to a misspecification in the statistical model. The translation of the results of Theorem 4 to our particular frameworks leads to the following corollary.

Corollary 3. *Let us consider, in the i.i.d., the Gaussian or the bounded regression setting a subset $\mathcal{S}$ of M with metric dimension bounded by $\widetilde{D}$ with respect to the relevant distance. Let $\eta_n = \inf\{x > 0 \mid nx^2 \geqslant 6\alpha\widetilde{D}(x)\}$ (with $\alpha = 1$ in the Hellinger case and $\alpha = 2$ in the variation case) for the i.i.d. setting, $\eta_\sigma = \inf\{x > 0 \mid x^2 \geqslant 36\sigma^2\widetilde{D}(x)\}$ for the Gaussian setting and, for the bounded regression setting, $\eta_n = \inf\{x > 0 \mid nx^2 \geqslant 50\widetilde{D}(x)\}$. Then one can build in each case a suitable T-estimator (or an M-estimator for the two last cases) $\hat{s}$ which satisfies, for all $s \in M$ and $q \geqslant 1$, the following risk bounds.*

(i) *In the i.i.d. setting*

$$\mathbb{E}_s\big[h^q(s,\hat{s})\big] \leqslant C(q)\big[\eta_n + h(s,\mathcal{S})\big]^q \quad \text{or} \quad \mathbb{E}_s\big[v^q(s,\hat{s})\big] \leqslant C(q)\big[\eta_n + v(s,\mathcal{S})\big]^q.$$

(ii) *In the Gaussian setting*

$$\mathbb{E}_s\big[\|s-\hat{s}\|^q\big] \leqslant C(q)\Big[\eta_\sigma + \inf_{t\in\mathcal{S}}\|s-t\|\Big]^q.$$

(iii) *In the bounded regression setting*

$$\mathbb{E}_s\big[\|s-\hat{s}\|^q\big] \leqslant C(q)\Big[\eta_n + \inf_{t\in\mathcal{S}}\|s-t\|\Big]^q.$$

In particular, the minimax risk $R(\mathcal{S},q)$ over $\mathcal{S}$ is bounded by $C(q)\eta_n^q$ or $C(q)\eta_\sigma^q$.

For the i.i.d. setting, we essentially recover (in a slightly more general form), the results of Birgé [8] (see his Proposition 3.1 and Corollary 2.6). For the Gaussian setting, this corollary applies in particular to finite dimensional linear subspaces of $\boldsymbol{l}_2(\mathbb{N}^\star)$ and compact finite dimensional non-linear manifolds. As far as we are aware, it has never been stated before, although it could have been deduced from [8] via our Proposition 4.

6.4. A few typical illustrations

6.4.1. Finite dimensional normed linear spaces

One purpose of Approximation Theory (see, for instance Pinkus [52]) is, given some function space $\mathcal{S}$, to provide k-dimensional linear spaces S_k such that $\sup_{s\in\mathcal{S}} d(s,S_k) \leqslant \varepsilon$ where ε is a known function of $\mathcal{S}$ and k. Then, if S is an ε-net for S_k, it is a 2ε-net for $\mathcal{S}$. Finding such nets S amounts to get bounds on the metric dimension of k-dimensional linear subspaces S_k of normed linear spaces. The following proposition implies together with Theorem 4 that if $\mathcal{S}$ is any subset of a k-dimensional normed linear space, the minimax quadratic risk $R(\mathcal{S},2)$ is bounded by $Ca^{-1}k$ (Ck/n in the i.i.d. and bounded regression settings and $C\sigma^2 k$ in the Gaussian setting, as expected).

Proposition 8. *Let M' be a normed linear space, d be the distance derived from the norm and V_k be some k-dimensional linear subspace of M'. If $S' \subset V_k$, it has a finite inner metric dimension bounded by $k\log 5 < 5k/3$. If M' is a Hilbert space and $S' \subset V_k$ is convex, the dimension bound can be improved to $0.403k \vee 1/2$.*

Proof. There exists a maximal η-separated subset S of S', which is therefore an η-net for S'. By the classical Lemma 4 below (see the proof of Lemma 2.5 page 20 of [59]) S satisfies, for all $t \in V_k$,

$$\big|\{u \in S \mid \|t-u\| < x\eta\}\big| \leqslant (2x+1)^k \leqslant \exp\big[kx^2(\log 5)/4\big] \quad \text{for } x \geqslant 2. \tag{6.2}$$

We then conclude from Lemma 2. If M' is a Hilbert space, S' is convex in V_k and $t \in M'$, there exists $t' \in \overline{S'}$ such that $d(t,u) \geqslant d(t',u)$ for all $u \in S'$ so that (6.2) still holds for $t \in M'$ which gives the dimension bound $k(\log 5)/4 < 0.403k$. □

Lemma 4. *If S is an η-separated subset of some k-dimensional normed linear space V_k, then $|\{u \in S \mid \|t-u\| \leqslant x\eta\}| \leqslant (2x+1)^k$ for all $x > 0$ and all $t \in V_k$.*

Note that our bound (6.2) has the right order of magnitude with respect to k: only the constant $(\log 5)/4$ is too pessimistic. Indeed, a lower-bound argument shows that, if S is an arbitrary η-net for $\mathbb{R}^k$, then $|S \cap \mathcal{B}(t,3\eta)| \geqslant 2^k$.

This is a rather poor but straightforward lower bound which shows that the metric dimension is bounded from below by $k(\log 2)/9$. When M' is a Hilbert space, one can also build η-nets for S_k, $k \geqslant 2$ (the case of $k = 1$ being trivial) in a constructive way but this results in the suboptimal bound $0.458k$ for the inner metric dimension.

Proposition 9. *Let V_k be a k-dimensional linear subspace (identified to $\mathbb{R}^k$) of some Hilbert space M' and η be a positive number. If $S \subset V_k$ is the lattice $[(2\eta/\sqrt{k})\mathbb{Z}]^k$, then it is an η-net for V_k and for all $t \in M'$,*

$$\bigl|S \cap \mathcal{B}(t,r)\bigr| < (\pi k)^{-1/2} \exp\bigl[0.458k(r/\eta)^2\bigr] \quad \text{for } r \geqslant 2\eta.$$

The result follows immediately from the next lemma, which can be proved as Lemma 2 from [15].

Lemma 5. *For all positive integers k, $t \in \mathbb{R}^k$, $\lambda > 0$ and $r > 0$,*

$$\bigl|(\lambda\mathbb{Z})^k \cap \mathcal{B}(t,r)\bigr| \leqslant \frac{(\pi e/2)^{k/2}}{\sqrt{\pi k}} \left(\frac{2r}{\lambda\sqrt{k}} + 1\right)^k < \frac{1}{\sqrt{\pi k}} \exp\left[k\left(0.73 + \log\left(\frac{2r}{\lambda\sqrt{k}} + 1\right)\right)\right].$$

6.4.2. Parametric models

The initial construction of Le Cam in [43] was directed towards parametric models, i.e. sets $\mathcal{S}$ of distributions that are smooth images of subsets of $\mathbb{R}^k$. For many such models, or at least compact subsets of them, there is a smooth relationship of the form (6.3) below between the Euclidean distance on the parameters and the distance d on M (see Lemma 7 of [43]). In this case, one can mimic the proof of Proposition 8 to get

Proposition 10. *Assume that there exists a one-to-one parametrization $\theta \mapsto t(\theta) \in \mathcal{S}$ of $\mathcal{S}$ by a subset Θ of $\mathbb{R}^k$, a norm $\|\cdot\|$ on $\mathbb{R}^k$, three positive constants $b \leqslant b'$ and β and an increasing function ξ satisfying $\xi(x\lambda) \leqslant x^\beta \xi(\lambda)$ for all $x \geqslant 2$, $\lambda > 0$, such that*

$$b\|\theta - \theta'\| \leqslant \xi\bigl(d\bigl(t(\theta), t(\theta')\bigr)\bigr) \leqslant b'\|\theta - \theta'\| \quad \text{for all } \theta, \theta' \in \Theta. \tag{6.3}$$

Then the subset $\mathcal{S}$ of (M,d) has an inner metric dimension bounded by $k[\log(b'/b) + \beta \log 5] \vee (1/2)$.

Note that, for the i.i.d. setting, (6.3) can only hold for bounded sets Θ since d being either h or v, it is bounded.

An alternative way to deal with a parametric model, when the parametrization t is a smooth mapping from $\mathbb{R}^k$ to some Hilbert space, which may happen in the Gaussian setting, is to consider it as a manifold. It is then useful to introduce the following

Property 1. *A subset $\mathcal{V}$ of $\mathcal{S}$ enjoys Property 1 if there exists $D > 0$ such that, for all $\eta, r > 0$, $t \in M$ and any η-separated subset S_η of $\mathcal{S}$, $|\mathcal{V} \cap S_\eta \cap \mathcal{B}(t,r)| \leqslant \exp[D[(r/\eta) \vee 1]^2]$.*

If $\mathcal{S}$ is a k-dimensional smooth manifold, for any $s \in \mathcal{S}$, one can find a vicinity of s in $\mathcal{S}$ for which the projection onto the tangent space at s is almost isometric. It follows from the arguments used to prove Proposition 8 that there exists a vicinity $\mathcal{V}_s$ of s which enjoys Property 1 with $D = 3k/2$. If $\mathcal{S}$ is compact, one can even assume that $\mathcal{V}_s = \mathcal{B}(s, \bar{r})$ for some $\bar{r} > 0$ independent of s. Indeed, if this were not true, by compacity, one could find a sequence $(s_n)_{n\geqslant 1}$ in $\mathcal{S}$ converging to s_0 and a sequence $(r_n)_{n\geqslant 1}$ converging to 0 such that $\mathcal{S} \cap \mathcal{B}(s_n, r_n)$ does not enjoy Property 1. This would contradict the fact that $\mathcal{V}_{s_0}$ enjoys Property 1 since $\mathcal{B}(s_n, r_n) \subset \mathcal{V}_{s_0}$ for n large enough. In such a case, we can bound the metric dimension of $\mathcal{S}$ in the following way.

Proposition 11. *If $\mathcal{S}$ is compact, $\bar{r} > 0$ and, for all $s \in \mathcal{S}$, $\mathcal{B}(s, \bar{r}) \cap \mathcal{S}$ enjoys Property 1 with the same value of D, the metric dimension of $\mathcal{S}$ is bounded by*

$$\widetilde{D}(\eta) = \frac{D(\bar{r}/\eta)^2 + \log K}{4 \vee [\bar{r}/(2\eta)]^2} \vee 1/2, \tag{6.4}$$

where K denotes the minimal cardinality of a covering of $\mathcal{S}$ by balls of radius $\bar{r}$.

Proof. If $\eta \geqslant \bar{r}$, the K centers of the balls of radius $\bar{r}$ which cover $\mathcal{S}$ provide an η-net S for $\mathcal{S}$ with $\log|S| = \log K$ and (6.4) holds. If $\eta < \bar{r}$, we take for S a maximal η-separated subset of $\mathcal{S}$ and it follows from Property 1 that $|S \cap \mathcal{B}(s,\bar{r})| \leqslant \exp[D(\bar{r}/\eta)^2]$ for all $s \in \mathcal{S}$, which implies that $|S| \leqslant K\exp[D(\bar{r}/\eta)^2]$. Let t be an arbitrary point in M. We distinguish between two cases. If $2\eta \leqslant r < \bar{r}/2$, then either $\mathcal{B}(t,r) \cap S = \emptyset$ and there is nothing to prove, or $\mathcal{B}(t,r) \subset \mathcal{B}(s,2r) \subset \mathcal{B}(s,\bar{r})$ for some $s \in S$. Then $|S \cap \mathcal{B}(s,2r)| \leqslant \exp[4D(r/\eta)^2]$ by Property 1 and (6.4) holds since $\bar{r} > 4\eta$. If $r \geqslant (\bar{r}/2) \vee 2\eta$, then

$$\log\bigl(\bigl|S \cap \mathcal{B}(t,r)\bigr|\bigr) \leqslant \log\bigl(|S|\bigr) \leqslant D(\bar{r}/\eta)^2 + \log K \leqslant \frac{D(\bar{r}/\eta)^2 + \log K}{4 \vee [\bar{r}/(2\eta)]^2}(r/\eta)^2$$

and (6.4) holds again. □

6.4.3. Totally bounded sets and entropy numbers

For totally bounded sets, a classical way of measuring massiveness is via *entropy numbers*. Let us recall their definition.

Definition 7. If $\mathcal{S}$ is totally bounded, its η-covering number $\mathcal{N}(\eta)$ is the smallest number of closed balls of radius η that are needed to cover it and its η-entropy is $\mathcal{H}(\eta) = \log_2[\mathcal{N}(\eta)]$.

The η-entropy is non-increasing with respect to η, $\mathcal{H}(\eta) = 0$ for η large enough and the metric dimension of $\mathcal{S}$ is bounded by $[\mathcal{H}(\eta)\log 2]/4$. One way of bounding $\mathcal{H}(\eta)$ is to find an upper bound for the cardinality of some maximal η-separated set in $\mathcal{S}$. Much more on the subject, in particular examples of evaluations of $\mathcal{H}$ for various sets and distances, can be found in [39,48,18] and [49] among other references. Nevertheless, the approach based on entropy is not always adequate for our purpose, even for compact sets. For instance, we have seen that Euclidean balls in $M = \mathbb{R}^k$ with $k > 1$ have a metric dimension bounded by $0.403k$ independently of their radius $\bar{r}$. But their η-entropy $\mathcal{H}(\eta)$ is bounded from below by $k\log_2(\bar{r}/\eta)$. Using $\mathcal{H}(\eta)(\log 2)/4$, which is at least $[k(\log 2)/4]\log_2(\bar{r}/\eta)$, as an upper bound for the metric dimension of those balls would not lead to the right bound when $\bar{r}$ is large.

6.4.4. Building D-models for bounded regression with random design

There are special difficulties to derive D-models in this case when the distribution μ of the design is unknown because, even if we know a set S' with finite metric dimension, for instance a finite-dimensional linear space, we do not know how to discretize it since the distance d_2 is unknown to the statistician. A possible way to bypass this problem is to start from a finite dimensional linear subspace of the space $\mathcal{L}_\infty(\mathcal{X})$ of bounded functions on $\mathcal{X}$ endowed with the uniform distance d_∞ given by $d_\infty(t,u) = \sup_{x\in\mathcal{X}}|t(x) - u(x)|$. If T' is a k-dimensional linear subspace of $\mathcal{L}_\infty(\mathcal{X})$ which contains the constant functions and T a maximal η-separated subset of T' which contains the constant function $t_0 \equiv 1/2$, T is an η-net for T' with respect to d_∞ and, if $S' = T \cap \mathcal{B}_{d_\infty}(t_0,1)$, by Lemma 4, $|S'| \leqslant (2\eta^{-1}+1)^k$. Introducing the mapping $\bar{\pi}$ from $\mathcal{L}_\infty(\mathcal{X})$ to M given by

$$\bar{\pi}(t) = (t \wedge 1) \vee 0 \quad \text{for } t \in \mathcal{L}_\infty(\mathcal{X}), \tag{6.5}$$

we set $S = \bar{\pi}(S')$. Then S is a D-model for the distance d_2 with parameters η, $(k/4)\log\bigl(2\eta^{-1}+1\bigr)$ and 1. Note here the presence of the unexpected logarithmic factor due to the use of entropy instead of metric dimension for deriving the parameters of D-models. Moreover, if $s \in M$, $d_\infty(s,T) \leqslant d_\infty(s,t_0) \leqslant 1/2$, hence

$$d_2(s,S) \leqslant d_\infty(s,S) \leqslant d_\infty(s,S') = d_\infty(s,T) \leqslant d_\infty(s,T') + \eta.$$

6.4.5. Ellipsoids

Let us now consider a situation where the function $\widetilde{D}(\eta)$ converges to infinity when η goes to 0. Given a non-increasing sequence $\boldsymbol{a} = (a_i)_{i\geqslant 1}$ in $[0,+\infty]$ with $a_1 > 0$ and $\lim_i a_i = 0$ we define the ellipsoid $\mathcal{E}(\boldsymbol{a}) \subset \boldsymbol{l}_2(\mathbb{N}^\star)$ as

$$\mathcal{E}(\boldsymbol{a}) = \left\{ s = (s_i)_{i\geqslant 1} \;\Big|\; \sum_{i=1}^{+\infty}\left(\frac{s_i}{a_i}\right)^2 \leqslant 1 \right\}, \tag{6.6}$$

with the convention that if $a_i = 0$ then $s_i = 0$ and if $a_i = +\infty$ then s_i is arbitrary. To bound the metric dimension of $\mathcal{E}(\boldsymbol{a})$, we observe that $\sum_{i=k+1}^{+\infty} s_i^2 \leqslant a_{k+1}^2$ for $k \geqslant 0$ and $s \in \mathcal{E}(\boldsymbol{a})$. Applying this with $k = 0$ and $S_0 = \{0\}$, we

can set $\widetilde{D}(\eta)$ to any number $\geqslant 1/2$ for $\eta \geqslant a_1$. For $\eta \geqslant \sqrt{2}a_{k+1}$ with $k \geqslant 1$, we set $\lambda = \eta\sqrt{2/k} \geqslant 2a_{k+1}k^{-1/2}$ and $S_k = (\lambda\mathbb{Z})^k \subset \boldsymbol{l}_2(\mathbb{N}^\star)$ ($t_i = 0$ for $i > k$ if $t \in S_k$). Therefore, if $s \in \mathcal{E}(\boldsymbol{a})$, one can find some $t \in S_k$ with

$$d^2(s,t) \leqslant a_{k+1}^2 + k\lambda^2/4 \leqslant k\lambda^2/2 = \eta^2$$

and S_k is an η-net for $\mathcal{E}(\boldsymbol{a})$. If $t \in \boldsymbol{l}_2(\mathbb{N}^\star)$ and $r \geqslant 2\eta$, it follows from Lemma 5 that

$$k^{-1}\log\bigl(\bigl|S_k \cap \mathcal{B}(t,r)\bigr|\bigr) < 0.73 + \log\bigl(1+\sqrt{2}(r/\eta)\bigr) < 0.52(r/\eta)^2.$$

This proves that the metric dimension of $\mathcal{E}(\boldsymbol{a})$ is bounded by $0.52k$ when $\eta \geqslant \sqrt{2}a_{k+1}$ and we can finally choose for $\widetilde{D}$ the function given by

$$\widetilde{D}(\eta) = 0.52\inf\bigl\{k \in \mathbb{N}^\star \mid \sqrt{2}a_{k+1} \leqslant \eta\bigr\}. \tag{6.7}$$

We can then derive from Corollary 3 and Lemma 6 below the following upper bound for the minimax risk over $\mathcal{E}(\boldsymbol{a})$:

$$R\bigl(\mathcal{E}(\boldsymbol{a}), q\bigr) \leqslant C(q)\Bigl(\inf_{k\geqslant 1}\bigl\{a_{k+1} \vee \sigma\sqrt{k}\bigr\}\Bigr)^q \quad \text{for } q \geqslant 1. \tag{6.8}$$

This bound is similar to the one we got for the i.i.d. setting in [8], Section 4. Such a result is not new and just given as an illustration of the way our method works. It can, for instance, be deduced from Donoho et al. [30] (see also [17], Section 6.2). An exact asymptotic evaluation of the minimax risk over ellipsoids has been given by Pinsker in [53].

Lemma 6. *Let $(b_k)_{k\geqslant 1}$ be some non-increasing sequence with values in $[0,+\infty]$ such that $\lim_{k\to+\infty} b_k = 0$ and let the function G be defined on $(0,+\infty)$ by $G(x) = \inf\{k \geqslant 1 \mid b_k \leqslant x\}$. Then, for all $t > 0$,*

$$\inf\bigl\{x \mid x^2 \geqslant tG(x)\bigr\} = \inf_{k\geqslant 1}\bigl(b_k \vee \sqrt{tk}\bigr).$$

The elementary proof will be omitted.

The problem of estimating some $s \in \mathcal{E}(\boldsymbol{a})$ typically occurs when it comes from the filtering of the white noise framework by the trigonometric basis as explained in Section 5.2.2. It is then of common practice to put some Sobolev-type restriction on the unknown s, of the form $\|s^{(\alpha)}\| \leqslant R$. This amounts to assume that s belongs to the ellipsoid $\mathcal{E}(\boldsymbol{a}) = \mathcal{E}'(\alpha, R)$ defined by (6.6) with coefficients

$$a_1 = +\infty \quad \text{and} \quad a_{2j} = a_{2j+1} = R(2\pi j)^{-\alpha} \quad \text{for } j \geqslant 1. \tag{6.9}$$

We refer to Section 1.1.4 of [17] for additional details. This and (6.7) lead to the following upper bound $\widetilde{D}_{\alpha,R}$ for the metric dimension of $\mathcal{E}'(\alpha, R)$:

$$\frac{\widetilde{D}_{\alpha,R}(\eta)}{0.52} = \begin{cases} 1 & \text{if } \eta \geqslant \sqrt{2}R(2\pi)^{-\alpha}; \\ 2j+1 & \text{if } \sqrt{2}R(2\pi j)^{-\alpha} > \eta \geqslant \sqrt{2}R\bigl[2\pi(j+1)\bigr]^{-\alpha},\ j \geqslant 1. \end{cases}$$

Equivalently,

$$\frac{\widetilde{D}_{\alpha,R}(\eta)}{0.52} = 2\left\lceil \frac{1}{2\pi}\left(\frac{\sqrt{2}R}{\eta}\right)^{1/\alpha}\right\rceil - 1, \tag{6.10}$$

with $\lceil\cdot\rceil$ given by (4.4). The resulting upper bound for the minimax risk then derives from (6.8):

$$R\bigl(\mathcal{E}'(\alpha,R), q\bigr) \leqslant C(q)\sigma^q\left[\left(\frac{R}{\sigma\pi^\alpha}\right)^{1/(2\alpha+1)} \vee 1\right]^q.$$

This upper bound is known to be sharp, up to the constant $C(q)$.

7. Working with several models

The limitations of our previous approach which amounts to basing our estimation procedure on a single well-chosen discrete model S appear clearly in the applications. When we estimate a uniform distribution on $[0, \mathrm{e}^t]$ in

Section 5.3.3, we have to restrict to a compact set of values for t and when we design a minimax (up to constants) estimator for the ellipsoid $\mathcal{E}'(\alpha, R)$ in Section 6.4.5, we have to know the values of α and R in order to construct the estimator. It would clearly be more satisfactory to be able to use several models simultaneously in the construction of our estimators. For instance, with a countable number of them, one could approximate properly the whole parameter space in the first problem and if we had at hand one approximating space for each pair (α, R) and could use all of them together in the second problem, we would not have to know α and R and therefore get an adaptive estimator. Further motivations for the introduction of several models can be found in [5] and [17].

There are various ways to handle several models simultaneously. One possible way that we shall consider in Section 9 is to build an estimator on each model and then aggregate them. An alternative solution is to use a penalized M-estimator, for instance the penalized m.l.e., but this estimator suffers from the same defects as the ordinary m.l.e. and proving results for the penalized m.l.e. leads to similar technical difficulties, as can be seen from [5,21,22,59] or [31]. Just as the construction of T-estimators provided an alternative to the ordinary m.l.e., the construction that follows offers an alternative to the penalized m.l.e.

7.1. T-estimators based on several models

In this set up, S is a union of several D-models which satisfy the following assumption.

Assumption 3. The set $S = \bigcup_{m\in\mathcal{M}} S_m$ is a finite or countable union of D-models S_m with respective parameters η_m, D_m and B' and $D_m \geqslant 1/2$ for all $m \in \mathcal{M}$.

This implies in particular that S is countable and that Lemma 1 applies to each S_m.

In the case of several models, we have to specify a suitable function η on S. We would like that $\eta(t) = \eta_m$ when t belongs to S_m but this is not a proper definition because t may belong to several S_m simultaneously. We therefore set

$$\eta(t) = \inf\{\eta_m \mid m \in \mathcal{M} \text{ and } t \in S_m\}. \tag{7.1}$$

Note that, if we work with a single model ($|\mathcal{M}| = 1$), η is a constant function so that (4.15) reduces to (5.1) and (4.16) to (5.2) and the case of a single model appears to be a special case of the general one but its treatment is much simpler.

We can now prove a general result about the existence and performances of T-estimators which should be viewed as the extension of Theorem 3 to the case of several models. Not only the relationship between D and $a\eta^2$ should now hold for each model S_m, uniformly in m, but we also require the following additional assumption

$$\sum_{m\in\mathcal{M}} \exp[-a\eta_m^2/21] = \Sigma < +\infty, \tag{7.2}$$

which bounds the "complexity" of our family of models in the sense that it controls the growth of the sequence of numbers $|\{m \in \mathcal{M} \mid j - 1 < \eta_m \leqslant j\}|$. This condition should be viewed as an analogue of (3.1) in [6], (19) in [14], (2.2) in [5] or (3.3) in [17]. It implies in particular that

$$\big|\{m \in \mathcal{M} \mid \eta_m \leqslant z\}\big| < +\infty \quad \text{for all } z > 0. \tag{7.3}$$

Theorem 5. *Let (M, d) be a semi-metric space and Assumptions* 1 *and* 3 *hold with $S \subset M_T$,* (7.2) *and*

$$a\eta_m^2 \geqslant 21D_m/5 \quad \textit{for all } m \in \mathcal{M}. \tag{7.4}$$

If the tests ψ satisfy (4.15) *with η given by* (7.1)*, for all $s \in M$ such that $\delta(s, S) < +\infty$, $\mathbb{P}_s$-a.s. there exist T_ε-estimators $\hat{s}(\boldsymbol{X})$ if $\varepsilon > 0$ or if $\mathcal{M}$ is finite. If $0 \leqslant \varepsilon \leqslant 4$, any of them satisfies, for all $s' \in S$ such that $\delta(s, s') < +\infty$,*

$$\mathbb{P}_s\big[d(s', \hat{s}) > y\big] < (BB'\Sigma/7)\exp\big[-(2/3)ay^2\big] \quad \textit{for } y \geqslant \delta(s, s') \vee \big[4\eta(s')\big]. \tag{7.5}$$

If, moreover, d is a distance and $\delta = \kappa d$, then, for all $s \in M$,

$$\mathbb{E}_s\big[d^q(s, \hat{s})\big] \leqslant \big[1 + (BB'\Sigma/7)\zeta_q(22.4)\big](\kappa + 1)^q \inf_{m\in\mathcal{M}}\big\{d(s, S_m) \vee (4\eta_m/\kappa)\big\}^q. \tag{7.6}$$

In particular, for $1 \leqslant q \leqslant 79$,

$$\mathbb{E}_s\big[d^q(s, \hat{s})\big] \leqslant [1 + 10^{-7}BB'\Sigma](\kappa + 1)^q \inf_{m\in\mathcal{M}}\big\{d(s, S_m) \vee (4\eta_m/\kappa)\big\}^q. \tag{7.7}$$

If Assumption 2 *replaces Assumption* 1, $\mathbb{P}_s$*-a.s., there exists at least one M-estimator* $\hat{s} = \hat{s}(\boldsymbol{X}) \in S$ *such that* $\gamma'(\hat{s}, \boldsymbol{X}) + \tau\eta^2(\hat{s}) = \inf_{t\in S}\{\gamma'(t, \boldsymbol{X}) + \tau\eta^2(t)\}$ *and any such M-estimator satisfies* (7.5), (7.6) *and* (7.7) *under the same conditions.*

Remark. Since $D_m \geqslant 1/2$, it follows from (7.4) and (7.5) that

$$a\eta_m^2 \geqslant 2.1 \quad \text{for all } m \in \mathcal{M} \quad \text{and} \quad ay^2 \geqslant 16a\eta^2(s') \geqslant 33.6, \tag{7.8}$$

so that the right-hand side of (7.5) is bounded by $2.7 \times 10^{-11} B B' \Sigma$. Therefore (7.5) implies that $\mathbb{P}_s[d(s', \hat{s}) \leqslant \delta(s, s') \vee [4\eta(s')]] \simeq 1$ unless $BB'\Sigma$ is very large.

As a consequence of Theorem 5, we can build T-estimators from a suitable discretization of some collection of approximating models such as those provided by Approximation Theory. The following result is easily comparable to more classical ones about the performances of penalized estimators as in [5] or [17].

Corollary 4. *Let Assumption* 1 *or* 2 *hold with* $M_T = M$, $\delta = \kappa d$, $\kappa \geqslant 4$ *and let* $\{\overline{S}_m\}_{m\in\mathcal{M}}$ *be a finite or countable family of subsets of the metric space* (M, d) *with respective finite metric dimensions bounded by* $\overline{D}_m$. *Let* $\{\Delta_m\}_{m\in\mathcal{M}}$ *be a family of non-negative weights such that*

$$\sum_{m\in\mathcal{M}} \exp[-\Delta_m/5] = \Sigma. \tag{7.9}$$

There exists a T-estimator (*or an M-estimator under Assumption* 2) $\hat{s}$ *satisfying, for all* $s \in M$

$$\mathbb{P}_s\big[d(s, \hat{s}) > y\big] < (B\Sigma/7)\exp\big[-(32/75)ay^2\big] \quad \text{for } y \geqslant (\kappa+1)\bar{y}, \tag{7.10}$$

$$\bar{y} = \inf_{m\in\mathcal{M}}\left\{d(s, S_m) \vee \sqrt{21(\overline{D}_m \vee \Delta_m)/(5a)}\right\}.$$

For $1 \leqslant q \leqslant 79$, *we get the following risk bound*:

$$\mathbb{E}_s\big[d^q(s, \hat{s})\big] \leqslant \left[1 + \frac{B\Sigma}{10^7}\right](\kappa+1)^q \inf_{m\in\mathcal{M}}\left\{d\big(s, \overline{S}_m\big) + \sqrt{\frac{21}{5a}(\overline{D}_m \vee \Delta_m)}\right\}^q. \tag{7.11}$$

Proof. For each m, set $\eta_m^2 = 21(\overline{D}_m \vee \Delta_m)/(5a)$ and $D_m = \overline{D}_m \geqslant 1/2$ by definition. Then (7.2) and (7.4) hold. Moreover, the definition of the metric dimension implies that, for each $m \in \mathcal{M}$, there exists $S_m \subset M = M_T$ which is an η_m-net for $\overline{S}_m$ and a D-model with parameters η_m, D_m and 1. Assumption 3 then holds with $S = \bigcup_{m\in\mathcal{M}} S_m$ and $B' = 1$. We may apply Theorem 5 which implies the existence of T-estimators satisfying (7.5) and (7.7) from which (7.11) follows since $d(s, S_m) \leqslant d(s, \overline{S}_m) + \eta_m$. It follows from (7.3) that one can find some $m \in \mathcal{M}$ and some $s' \in S_m$ such that $d(s, s') \vee \eta_m = \bar{y}$. Since $\delta(s, s') \vee (4\eta_m) \leqslant \kappa\bar{y}$, it follows from (7.5) that

$$\mathbb{P}_s\big[d(s, \hat{s}) > z + \bar{y}\big] < (B\Sigma/7)\exp[-2az^2/3] \quad \text{for } z \geqslant \kappa\bar{y},$$

from which we derive (7.10). □

Remark. We could, alternatively, adopt a Bayesian point of view. Choosing some prior distribution ν on $\mathcal{M}$ with $\nu_m = -\log(\nu(\{m\})) > 0$ for each m and setting $\Delta_m = 5(\nu_m - 14)$ leads to (7.2) with $\Sigma = \mathrm{e}^{14}$ so that (7.11) becomes

$$\mathbb{E}_s\big[d^q(s, \hat{s})\big] \leqslant \left[1 + \frac{B}{8}\right](\kappa+1)^q \inf_{m\in\mathcal{M}}\left\{d(s, \overline{S}_m) + \sqrt{\frac{21}{a}\left[\frac{\overline{D}_m}{5} \vee (\nu_m - 14)\right]}\right\}^q.$$

A very small prior probability $\nu(\{m\})$ for the model $\overline{S}_m$, implies a large value of η_m^2, which, as we already mentioned, can be viewed as a penalty for model S_m. This Bayesian viewpoint should be compared with the one given in Section 3.4 of [17] for penalized least squares.

7.2. Proof of Theorem 5

As in Section 5.3, we start by proving some deviation bound for $\mathcal{D}_{\boldsymbol{X}}$ and $\mathcal{D}'_{\boldsymbol{X}}$:

$$\left.\begin{array}{l}\mathbb{P}_s\big[\mathcal{D}_{\boldsymbol{X}}(s') > y\big] \\ \text{or} \\ \mathbb{P}_s\big[\mathcal{D}'_{\boldsymbol{X}}(s') > y\big]\end{array}\right\} \leqslant \frac{BB'\Sigma}{7}\exp\left[-\frac{2ay^2}{3}\right] \quad \text{for } y \geqslant y_0 = \delta(s,s') \vee \big[4\eta(s')\big]. \tag{7.12}$$

It follows from (4.5) and (4.15) that

$$\begin{aligned}\mathbb{P}_s\big[\mathcal{D}_{\boldsymbol{X}}(s') > y\big] &= \mathbb{P}_s\big[\exists t \in S \text{ with } d(t,s') > y \text{ and } \psi(s',t,\boldsymbol{X}) = 1\big] \\ &\leqslant \sum_{\substack{t\in S \\ d(t,s')>y}} \mathbb{P}_s\big[\psi(s',t,\boldsymbol{X}) = 1\big] \leqslant B \sum_{\substack{t\in S \\ d(t,s')>y}} \exp\big[-a\big(d^2(t,s') - \eta^2(s') + \eta^2(t)\big)\big],\end{aligned} \tag{7.13}$$

since $d(t,s') > y \geqslant \delta(s,s')$. Under Assumption 2(A), the same bound holds with $\mathcal{D}'_{\boldsymbol{X}}(s')$ replacing $\mathcal{D}_{\boldsymbol{X}}(s')$ and $\{\gamma(t,\boldsymbol{X}) \leqslant \gamma(s',\boldsymbol{X})\}$ replacing $\{\psi(s',t,\boldsymbol{X}) = 1\}$, by (4.16). Let us now bound the right-hand side of (7.13). Using (7.3) we can index the set $\mathcal{M}$ as $\{m_j,\ j \in \mathbb{N},\ j < |\mathcal{M}|\}$ in non-increasing order of the η_m, so that $i < j$ implies $\eta_{m_i} \leqslant \eta_{m_j}$. Then we derive from the S_m a partition $\{S'_m, m \in \mathcal{M}\}$ of S by setting, using this indexing of $\mathcal{M}$,

$$S'_{m_k} = S_{m_k} \cap \left(\bigcup_{j<k} S_{m_j}\right)^c \quad \text{for all } k \in \mathbb{N}, k < |\mathcal{M}|. \tag{7.14}$$

Since $S'_m \subset S_m$, the S'_m still satisfy Assumption 3 with the same constants η_m, D_m and B'. We first want to show that

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geqslant y}} \exp\big[-a\big(d^2(t,s') + \eta_p^2\big)\big] \leqslant \frac{B'}{7}\exp\left[-\frac{35ay^2}{48} - \frac{a\eta_p^2}{21}\right]. \tag{7.15}$$

To prove (7.15), we consider two cases.

Case 1: $y \geqslant 2\eta_p$. Let us observe that, if $z \geqslant 2\eta_p$, Assumption 3 and (7.4) imply that

$$\big|\big\{t \in S'_p \mid d(t,s') < z\big\}\big| \leqslant B'\exp\big[(z/\eta_p)^2 D_p\big] \leqslant B'\exp[5az^2/21].$$

Setting $\theta = 91/80$, we derive from this bound and (7.8) that

$$\begin{aligned}\frac{1}{B'}\sum_{\substack{t\in S'_p \\ d(t,s')\geqslant y}} \mathrm{e}^{-ad^2(t,s')} &= \frac{1}{B'}\sum_{j\geqslant 0} \sum_{\theta^{j/2}y \leqslant d(t,s') < \theta^{(j+1)/2}y} \mathrm{e}^{-ad^2(t,s')} \\ &\leqslant \sum_{j\geqslant 0}\exp[5\theta^{j+1}ay^2/21 - a\theta^j y^2] \leqslant \sum_{j\geqslant 0}\exp[-35ay^2\theta^j/48] \\ &\leqslant \exp[-35ay^2/48]\sum_{j\geqslant 0}\exp\big[-(49/2)\big[(91/80)^j - 1\big]\big] \\ &< 1.036\exp[-35ay^2/48].\end{aligned} \tag{7.16}$$

Case 2: $y < 2\eta_p$. In this case we split the sum in (7.15) into two parts. For $d(t,s') \geqslant 2\eta_p$, we can apply the results of Case 1, i.e. (7.16) with y replaced by $2\eta_p$ and then the assumption $y < 2\eta_p$, to get

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geqslant 2\eta_p}} \exp\big[-a\big(d^2(t,s') + \eta_p^2\big)\big] < 1.036B'\exp\big[-a\big(35ay^2/48 + \eta_p^2\big)\big].$$

For $y \leqslant d(t,s') < 2\eta_p$, we use (7.4) again to derive

$$\sum_{\substack{t\in S'_p\\ y\leqslant d(t,s')<2\eta_p}} \exp\bigl[-a\bigl(d^2(t,s')+\eta_p^2\bigr)\bigr]\leqslant B'\exp[4D_p-ay^2-a\eta_p^2]\leqslant B'\exp\bigl[-a(y^2+\eta_p^2/21)\bigr].$$

Adding the sums for $d(t,s')\geqslant 2\eta_p$ and $d(t,s')<2\eta_p$ shows that the resulting bound for Case 2 is larger than the one we derived for Case 1, so that, for all $y\geqslant y_0$,

$$\sum_{\substack{t\in S'_p\\ d(t,s')\geqslant y}} \exp\bigl[-a\bigl(d^2(t,s')+\eta_p^2\bigr)\bigr]\leqslant B'\exp\biggl[-\frac{35ay^2}{48}-\frac{a\eta_p^2}{21}\biggr]\biggl(1.036\exp\biggl[-\frac{20a\eta_p^2}{21}\biggr]+\exp\biggl[-\frac{13ay^2}{48}\biggr]\biggr)$$

and (7.15) follows from (7.8). Using the fact that our ordering of $\mathcal{M}$ has been chosen in such a way that for all $t\in S$, $\eta(t)=\eta_p$ if $t\in S'_p$ and summing (7.15) with respect to $p\in\mathcal{M}$, we deduce from (7.2) that

$$\sum_{\substack{t\in S\\ d(t,s')\geqslant y}} \exp\bigl[-a\bigl(d^2(t,s')+\eta^2(t)\bigr)\bigr]\leqslant\frac{B'\Sigma}{7}\exp\biggl[-\frac{35ay^2}{48}\biggr]. \tag{7.17}$$

Finally (7.12) follows from (7.13) and (7.17) since $a\eta^2(s')\leqslant ay^2/16$.

Since (7.12) holds for $y\geqslant y_0$, $\mathcal{D}_{\boldsymbol{X}}(s')<+\infty$, $\mathbb{P}_s$-a.s. for any $s'\in S$ such that $\delta(s,s')<+\infty$. Moreover, $S=\bigcup_{m\in\mathcal{M}}S'_m$ and, by (4.9),

$$T_m=\bigl\{t\in S'_m\mid \mathcal{D}_{\boldsymbol{X}}(t)\vee\varepsilon\eta(t)\leqslant\mathcal{D}_{\boldsymbol{X}}(s')\vee\varepsilon\eta(s')\bigr\}\subset S'_m\cap\overline{\mathcal{B}}\bigl(s',\mathcal{D}_{\boldsymbol{X}}(s')\vee\varepsilon\eta(s')\bigr).$$

It follows that T_m is finite for each m. Then $\bigcup_{m\in\mathcal{M}}T_m$ is finite when $\mathcal{M}$ is finite. If $\varepsilon>0$ and $\eta_m>[\varepsilon^{-1}\mathcal{D}_{\boldsymbol{X}}(s')]\vee\eta(s')$, then $\varepsilon\eta(t)>\mathcal{D}_{\boldsymbol{X}}(s')\vee\varepsilon\eta(s')$ for all $t\in S'_m$ and T_m is empty. Therefore $\bigcup_{m\in\mathcal{M}}T_m$ is again finite by (7.3). In both cases there exists at least one T-estimator and (7.5) follows from (7.12) and (4.10).

Let us now fix some $m\in\mathcal{M}$ and set $s'=\pi_m(s)$ where π_m is a minimum distance operator from M to S_m provided by Lemma 1 via Assumption 3. Then $\eta(s')\leqslant\eta_m$, $d(s,s')=d(s,S_m)$, $\bar{y}=\kappa d(s,S_m)\vee 4\eta_m\geqslant y_0$ and, since $\varepsilon\leqslant 4$, (4.10) implies that

$$d\bigl(s,\hat{s}(\boldsymbol{X})\bigr)\leqslant d(s,s')+\mathcal{D}_{\boldsymbol{X}}(s')\vee\varepsilon\eta(s')\leqslant\mathcal{D}_{\boldsymbol{X}}(s')\vee 4\eta_m+d(s,S_m).$$

It follows from (7.12) that

$$\mathbb{P}_s\bigl[\mathcal{D}_{\boldsymbol{X}}(s')\vee 4\eta_m>y\bigr]\leqslant(BB'\Sigma/7)\exp[-2ay^2/3]\quad\text{for } y\geqslant\bar{y}$$

and we may therefore apply Lemma 3 with $Y=\mathcal{D}_{\boldsymbol{X}}(s')\vee 4\eta_m$, $\alpha=BB'\Sigma/7$, $\beta=2a/3$ and $w=d(s,S_m)$, hence $\beta\bar{y}^2\geqslant 32a\eta_m^2/3\geqslant 22.4$ by (7.8). Arguing as in the proof of Theorem 3 with S replaced by S_m in (5.11), we get

$$\mathbb{E}_s\bigl[d^q(s,\hat{s})\bigr]\leqslant(\kappa+1)^q\bigl[d(s,S_m)\vee(4\eta_m/\kappa)\bigr]^q\bigl[1+(BB'\Sigma/7)\zeta_q(22.4)\bigr].$$

An optimization with respect to m, which is arbitrary in $\mathcal{M}$, then leads to (7.6). Since, by (5.5), $\zeta_q(22.4)/7\leqslant 10^{-7}$ if $q\leqslant 79$, we derive (7.7).

Let us now assume that Assumption 2 holds and fix $s'\in S$. We want to show that

$$\theta(y)=\mathbb{P}_s\bigl[\exists t\in S\text{ with }\gamma(t,\boldsymbol{X})\leqslant\gamma(s',\boldsymbol{X})\text{ and }\eta(t)\geqslant y\bigr]\underset{y\to+\infty}{\longrightarrow}0. \tag{7.18}$$

We may therefore assume that $y\geqslant y_0/2$. Setting $\mathcal{M}_y=\{p\in\mathcal{M}\mid\eta_p\geqslant y\}$, we get, since $\eta(t)=\eta_p$ if $t\in S'_p$,

$$\theta(y)\leqslant\sum_{p\in\mathcal{M}_y}\Biggl[\sum_{\substack{t\in S'_p\\ d(t,s')\geqslant 2\eta_p}}\mathbb{P}_s\bigl[\gamma(t,\boldsymbol{X})\leqslant\gamma(s',\boldsymbol{X})\bigr]+\sum_{\substack{t\in S'_p\\ d(t,s')<2\eta_p}}\mathbb{P}_s\bigl[\gamma(t,\boldsymbol{X})\leqslant\gamma(s',\boldsymbol{X})\bigr]\Biggr].$$

Since $2\eta_p\geqslant 2y\geqslant y_0\geqslant 4\eta(s')$, we can bound the first term using (4.16) and (7.15):

$$\sum_{\substack{t\in S'_p\\ d(t,s')\geqslant 2\eta_p}}\mathbb{P}_s\bigl[\gamma(t,\boldsymbol{X})\leqslant\gamma(s',\boldsymbol{X})\bigr]\leqslant\frac{BB'}{7}\exp\biggl[-\frac{35ay^2}{12}-\frac{a\eta_p^2}{21}+a\eta^2(s')\biggr],$$

hence by (7.2), since $\eta(s') \leqslant y/2$,

$$\sum_{p\in\mathcal{M}_y} \sum_{\substack{t\in S'_p \\ d(t,s')\geqslant 2\eta_p}} \mathbb{P}_s\big[\gamma(t,\boldsymbol{X}) \leqslant \gamma(s',\boldsymbol{X})\big] \leqslant \frac{\Sigma BB'}{7} \exp\left[-\frac{8ay^2}{3}\right] \underset{y\to+\infty}{\longrightarrow} 0. \tag{7.19}$$

Then, we use (4.17), (7.4), $d(s,s') \leqslant \kappa^{-1}\delta(s,s') \leqslant y_0/4$ and $y_0 \geqslant 4\eta(s')$ to get

$$\begin{aligned}\sum_{p\in\mathcal{M}_y} \sum_{\substack{t\in S'_p \\ d(t,s')<2\eta_p}} \mathbb{P}_s\big[\gamma(t,\boldsymbol{X}) \leqslant \gamma(s',\boldsymbol{X})\big] &\leqslant \sum_{p\in\mathcal{M}_y} BB' \exp\big[4D_p + a\kappa' d^2(s,s') + a\eta(s')^2 - a\eta_p^2\big] \\ &\leqslant BB' \exp\big[(ay_0^2/16)(\kappa'+1)\big] \sum_{p\in\mathcal{M}_y} \exp\big[-a\eta_p^2/21\big],\end{aligned}$$

which goes to zero when $y \to +\infty$, since, by (7.2), $\sum_{p\in\mathcal{M}_y} \exp\big[-a\eta_p^2/21\big] \underset{y\to+\infty}{\longrightarrow} 0$. Together with (7.19) this shows that (7.18) holds. Now, by (4.14),

$$\big\{t \in S \mid \gamma(t,\boldsymbol{X}) < \gamma(s',\boldsymbol{X}) \text{ and } \eta(t) < y\big\} \subset \bigcup_{p\in\mathcal{M}\setminus\mathcal{M}_y} \big[S_p \cap \overline{\mathcal{B}}\big(s', \mathcal{D}_{\boldsymbol{X}}(s')\big)\big].$$

The set $S_p \cap \overline{\mathcal{B}}(s', \mathcal{D}_{\boldsymbol{X}}(s'))$ is finite a.s. for each p and $\mathcal{M}\setminus\mathcal{M}_y$ as well for any $y > 0$ by (7.3). This implies that, with a probability at least $1-\theta(y)$, $\{t \in S \mid \gamma(t,\boldsymbol{X}) < \gamma(s',\boldsymbol{X})\}$ is a finite set and there exists some minimizer $\hat{s}(\boldsymbol{X})$ of $\gamma(\cdot,\boldsymbol{X})$. Letting y go to infinity, we conclude from (7.18) that such a minimizer $\hat{s}(\boldsymbol{X})$ exists with probability one. Moreover, by (4.13), any such M-estimator satisfies, for each $m \in \mathcal{M}$, $d(s,\hat{s}) \leqslant d(s,S_m) + \mathcal{D}'_{\boldsymbol{X}}(\pi_m(s))$ and we conclude as before from (7.12), replacing $\mathcal{D}_{\boldsymbol{X}}$ by $\mathcal{D}'_{\boldsymbol{X}}$ and ε by 0.

8. Some applications

8.1. Application to the Gaussian setting

As follows from Proposition 4, for the Gaussian setting, Assumption 2 and therefore (4.16) and (4.17) hold with $\gamma'(t,\boldsymbol{X}) = -\log(\mathrm{d}P_t/\mathrm{d}P_0)(\boldsymbol{X})$,

$$\tau = (12\sigma^2)^{-2}, \quad d = d_2, \quad B = 1, \quad a = (24\sigma^2)^{-1}, \quad \kappa = 6 \quad \text{and} \quad \kappa' = 12.$$

By (5.17) and (4.7), given η, the corresponding function $\gamma(t,\boldsymbol{X})$ can be written as

$$\gamma(t,\boldsymbol{X}) = \frac{1}{\sigma^2}\left[\frac{\|t\|^2}{2} - \langle t,\boldsymbol{X}\rangle + \frac{\eta^2(t)}{12}\right] \quad \text{for all } t \in S, \tag{8.1}$$

and (4.8) then implies that the relevant tests ψ are the likelihood ratio tests defined by

$$\psi(t,u,\boldsymbol{X}) = \begin{cases} 0 & \text{if } \log\left(\dfrac{\mathrm{d}P_u}{\mathrm{d}P_t}\right)(\boldsymbol{X}) + \dfrac{\eta^2(t)-\eta^2(u)}{12\sigma^2} < 0; \\ 1 & \text{if } \log\left(\dfrac{\mathrm{d}P_u}{\mathrm{d}P_t}\right)(\boldsymbol{X}) + \dfrac{\eta^2(t)-\eta^2(u)}{12\sigma^2} > 0. \end{cases} \tag{8.2}$$

This means that Theorem 5 applies to this setting provided that a countable subset S of $M = \boldsymbol{l}_2(\mathbb{N}^\star)$ has been chosen which satisfies Assumption 3 with (7.2) and (7.4). In such a case, there exists a minimizer over S of the function γ given by (8.1) with η defined by (7.1) and, since $\mathbb{P}_s[\gamma(t,\boldsymbol{X}) = \gamma(u,\boldsymbol{X})] = 0$ for $t \neq u$ and S is countable, such a minimizer $\hat{s}$ is a.s. unique, hence is also the unique T_0-estimator, as explained in Section 4.2.2. Since it is a minimizer of $\|t\|^2 - 2\langle t,\boldsymbol{X}\rangle + \eta^2(t)/6$, it is merely a penalized least squares estimator on S, as considered in [17], with penalty $\eta^2(t)/6$.

Starting with a family of general models with controlled metric dimensions instead of D-models, we can derive from Theorem 5 the following result, the proof of which is analogue to the one of Corollary 4.

Corollary 5. *Assume that we have at disposal a finite or countable family of subsets* $\{\overline{S}_m\}_{m\in\mathcal{M}}$ *of* $\boldsymbol{l}_2(\mathbb{N}^\star)$ *with respective metric dimensions bounded by functions* $\widetilde{D}_m$. *Assume moreover that the numbers* η_m *satisfy the inequalities*

$$\eta_m^2 \geqslant 100.8\sigma^2\widetilde{D}_m(\eta_m) \quad \text{for all } m\in\mathcal{M} \quad \text{and} \quad \sum_{m\in\mathcal{M}} \exp\left[-\frac{\eta_m^2}{504\sigma^2}\right] = \Sigma < +\infty. \tag{8.3}$$

Then one can build a T_0*-estimator* $\hat{s}$ *which is the unique penalized least squares estimator on some suitable countable subset* S *of* $\boldsymbol{l}_2(\mathbb{N}^\star)$ *and it satisfies, for all* $s\in M$,

$$\mathbb{E}_s\left[\|s-\hat{s}\|^q\right] \leqslant [1+10^{-7}\Sigma]7^q \inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\|s-t\|+\eta_m\right\}^q \quad \text{for } 1\leqslant q\leqslant 79.$$

If, in particular, the sets $\overline{S}_m$ *have respective finite metric dimensions bounded by* $\overline{D}_m$ *and* (7.9) *holds, one gets, for* $1\leqslant q\leqslant 79$,

$$\mathbb{E}_s\left[\|s-\hat{s}\|^q\right] \leqslant [1+10^{-7}\Sigma]7^q \inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\|s-t\|+10.04\sigma\sqrt{\overline{D}_m\vee\Delta_m}\right\}^q. \tag{8.4}$$

Let us observe that the result given in the second part of this corollary completely parallels, and actually generalizes, since we are not restricted to considering linear models $\overline{S}_m$, the results of Theorem 2 of [17] for penalized projection estimators on linear models. Indeed, this theorem involves a family of linear subspaces of $\boldsymbol{l}_2(\mathbb{N}^\star)$ with respective linear dimensions D_m satisfying $\sum_{m\in\mathcal{M}}\exp(-L_mD_m)=\Sigma'<+\infty$. If we consider such a family of spaces $\{\overline{S}_m\}_{m\in\mathcal{M}}$ in Corollary 5 and set $\overline{D}_m = D_m/2$ and $\Delta_m = 5[L_mD_m+\log(\Sigma'/\Sigma)]$, then (7.9) is satisfied, $\overline{S}_m$ has a finite metric dimension bounded by $\overline{D}_m$ and (8.4) holds. The resulting risk bound is (apart from the constants) the exact analogue of (3.5) in [17]. As an immediate consequence, all the strategies of model selection and all corresponding adaptation results (for ellipsoids, $\boldsymbol{l}_p$-balls and Besov bodies) that have been considered for penalized projection estimators in Sections 5 and 6 of [17] remain valid for T-estimators. The novelty is that Corollary 5 allows us to consider more sophisticated strategies that possibly mix linear and non-linear models and even allows to consider models which are not of finite metric dimension. Let us now illustrate these new possibilities by two examples.

8.1.1. Handling a parametric model

Let us consider in $M=\boldsymbol{l}_2(\mathbb{N}^\star)$, the parametric family $\overline{S}=\{t(\theta),\ \theta>0\}$ with $t_i(\theta)=\exp(-i\theta)$ for $i\geqslant 1$. If we suspect that the true s may belong to $\overline{S}$, it seems reasonable to include $\overline{S}$ into a list of other models which should take care of the case when $s\notin\overline{S}$. For instance we can use $\overline{S}$ with a family of linear models such as those studied in [17] or the family of ellipsoids we shall consider in the next section. We shall prove below that $\overline{S}$ has a finite metric dimension bounded by 4.5. Adding a finite-dimensional model with dimension bounded by $\overline{D}$ to a given list $\{\overline{S}_m,\ m\in\mathcal{M}\}$ has little cost. Starting with weights Δ_m satisfying (7.9), we merely set $\Delta=\overline{D}$ for the new model. This leads only to a negligeable increase in the risk if $s\notin\overline{S}$, due to the increase of Σ by $\exp[-\Delta/5]<1$, but if s truly belongs to $\overline{S}$, we recover the classical parametric risk of order $\left(\sigma\sqrt{\overline{D}}\right)^q$. For a deeper analysis of the strategies to use to mix families of models, we refer to Section 4.1 of [17].

For simplicity, we shall identify $t(\theta)$ and θ, setting $d_2(\theta,\theta')=d_2(t(\theta),t(\theta'))$. Then, for $\theta<\theta'$,

$$d_2^2(\theta,\theta') = \sum_{i\geqslant 1}\left[\frac{1}{\mathrm{e}^{i\theta}}-\frac{1}{\mathrm{e}^{i\theta'}}\right]^2 = \frac{1}{\mathrm{e}^{2\theta}-1}+\frac{1}{\mathrm{e}^{2\theta'}-1}-\frac{2}{\mathrm{e}^{\theta+\theta'}-1} < \frac{1}{\mathrm{e}^{2\theta}-1}. \tag{8.5}$$

Let us set $g(x)=(\mathrm{e}^x-1)^{-1}$ for $x>0$. It follows from Taylor's formula that

$$d_2^2(\theta,\theta') = g(2\theta)+g(2\theta')-2g(\theta+\theta') = (\theta'-\theta)^2g''(2\theta^*) \quad \text{with } \theta\leqslant\theta^*\leqslant\theta'.$$

Since $g''(x)=\mathrm{e}^x(\mathrm{e}^x+1)(\mathrm{e}^x-1)^{-3}$ is a decreasing function of x, we may conclude that

$$(\theta'-\theta)^2g''(2\theta') \leqslant d_2^2(\theta,\theta') \leqslant (\theta'-\theta)^2g''(2\theta) \quad \text{for all } \theta<\theta'.$$

Since it is rather hard to invert g'', it will prove convenient to replace it by a simpler function. One can indeed check that

$$(8/5)x^{-3}\leqslant g''(x)\leqslant 2x^{-3} \quad \text{for } x\leqslant 3 \quad \text{and} \quad \mathrm{e}^{-x}\leqslant g''(x)\leqslant (5/4)\,\mathrm{e}^{-x} \quad \text{for } x\geqslant 3.$$

It finally follows that

$$(\theta'-\theta)^2 f(\theta') \leqslant d_2^2(\theta,\theta') \leqslant (5/4)(\theta'-\theta)^2 f(\theta) \quad \text{for all } \theta<\theta', \tag{8.6}$$

for a decreasing function f given by

$$f(x)=\begin{cases} x^{-3}/5 & \text{for } x<3/2;\\ \mathrm{e}^{-2x} & \text{for } x\geqslant 3/2.\end{cases}$$

Let now $\eta>0$ be given. In order to build an η-net for $\overline{S}$ we shall define suitable numbers $\theta_{k,j}$ for $k\in\mathbb{Z}$, $j\in\mathbb{N}$. We first define $\theta_{k,0}=\theta_k$ by $f(\theta_k)=\exp(-3-k)$ so that

$$\theta_k=\begin{cases} (3+k)/2\geqslant 3/2 & \text{for } k\geqslant 0;\\ 5^{-1/3}\exp(1+k/3)<3/2 & \text{for } k<0.\end{cases}$$

Then we set

$$\theta_{k,j}=\theta_k+8j\eta\,\mathrm{e}^{k/2}, \quad J_k=\sup\{j\in\mathbb{N}\mid \theta_{k,j}<\theta_{k+1}\} \quad \text{and} \quad I_k=[\theta_k,\theta_{k+1}[.$$

It follows from these definitions that, for any $\theta\in I_k$ one can find some $\theta'\in\{\theta_{k,j},\ 0\leqslant j\leqslant J_k\}\cup\{\theta_{k+1}\}$ such that $|\theta-\theta'|\leqslant 4\eta\,\mathrm{e}^{k/2}$, hence by (8.6), $d_2(\theta,\theta')<\eta$ and

$$\sup_{\theta\in I_k}\Big[\Big(\inf_{0\leqslant j\leqslant J_k} d_2(\theta,\theta_{k,j})\Big)\wedge d_2(\theta,\theta_{k+1})\Big]<\eta. \tag{8.7}$$

Moreover, the definition of θ_k implies that

$$\theta_{k+1}-\theta_k=\begin{cases} 1/2 & \text{for } k\geqslant 0;\\ 3/2-5^{-1/3}\,\mathrm{e}^{2/3}\approx 0.361>(1/2)\,\mathrm{e}^{-1/3} & \text{for } k=-1;\\ 5^{-1/3}(\mathrm{e}^{4/3}-\mathrm{e})\,\mathrm{e}^{k/3}\approx 0.629\,\mathrm{e}^{k/3} & \text{for } k\leqslant -2.\end{cases} \tag{8.8}$$

Since $J_k<(8\eta)^{-1}\,\mathrm{e}^{-k/2}(\theta_{k+1}-\theta_k)$, we get

$$J_k<G(k)/\eta \quad \text{with } G(k)=\begin{cases} (1/16)\exp(-k/2) & \text{for } k\geqslant 0;\\ 0.079\exp(-k/6) & \text{for } k<0.\end{cases} \tag{8.9}$$

Set $K=\inf\{k\in\mathbb{Z}\mid G(k)<\eta\}$. If $K+12<0$, then $\eta>0.079\exp(-K/6)$ and

$$\mathrm{e}^{2\theta_{K+12}}-1>2\theta_{K+12}=2\times 5^{-1/3}\exp(5+K/3)>\eta^{-2}.$$

One can check in the same way that this inequality remains true if $K<0$ and $K+12\geqslant 0$ or if $K\geqslant 0$. Then, by (8.5), $d(\theta_{K+12},\theta)<\eta$ for $\theta>\theta_{K+12}$ and it follows from the previous arguments that the set $S=\{\theta_{k,j},\ k<K,\ 0\leqslant j\leqslant J_k\}\cup\{\theta_K,\dots,\theta_{K+12}\}$ is an η-net for $\overline{S}$. Since, for $k<K$, $|S\cap I_k|=J_k+1\leqslant 2G(k)/\eta$ and for $k\geqslant K$, $|S\cap I_k|=1$, we get, for $k\geqslant 0$,

$$\big|S\cap[\theta_k,+\infty)\big|\leqslant\frac{1}{8\eta}\sum_{j\geqslant k}\exp\Big(\frac{-j}{2}\Big)+13<\frac{0.318}{\eta}\exp\Big(\frac{-k}{2}\Big)+13 \tag{8.10}$$

and, for $k<0$,

$$\begin{aligned}\big|S\cap[\theta_k,+\infty)\big|&\leqslant\frac{0.158}{\eta}\sum_{j=k}^{-1}\exp\Big(\frac{-j}{6}\Big)+\frac{1}{8\eta}\sum_{j\geqslant 0}\exp\Big(\frac{-j}{2}\Big)+13\\ &\leqslant\frac{0.158}{\eta}\sum_{j\geqslant k}\exp\Big(\frac{-j}{6}\Big)+13<\frac{1.03}{\eta}\exp\Big(\frac{-k}{6}\Big)+13.\end{aligned} \tag{8.11}$$

We are now in a position to bound the cardinality of $S\cap\mathcal{B}(\theta',r)$ for $\theta'\in\overline{S}$ and $r\geqslant 2\eta$. Note that the part of the ball that matters is merely the interval $(\underline{\theta},\bar{\theta})$ with $\underline{\theta}<\theta'<\bar{\theta}$ and $d_2(\theta',\underline{\theta})=d_2(\theta',\bar{\theta})=r$. If the whole interval is contained in some I_k, then by (8.6), $\bar{\theta}-\underline{\theta}\leqslant 2r[f(\theta_{k+1})]^{-1/2}=2r\,\mathrm{e}^{2+k/2}$, hence $|(\underline{\theta},\bar{\theta})\cap S|=|(\underline{\theta},\bar{\theta})\cap I_k|\leqslant \mathrm{e}^2 r/(4\eta)+1$. If

$\underline{\theta} \in I_{k-1}$ and $\bar{\theta} \in I_k$, using the previous argument twice, we get $|(\underline{\theta}, \bar{\theta}) \cap S| \leqslant \mathrm{e}^2 r/(2\eta) + 2$. Finally, if $\underline{\theta} \in I_{k-1}$ and $\bar{\bar{\theta}} \geqslant \theta_{k+1}$, then

$$\bigl|(\underline{\theta}, \bar{\theta}) \cap S\bigr| \leqslant \bigl|(\underline{\theta}, \bar{\theta}) \cap I_{k-1}\bigr| + \bigl|S \cap [\theta_k, +\infty)\bigr| \leqslant \mathrm{e}^2 r/(4\eta) + 1 + \bigl|S \cap [\theta_k, +\infty)\bigr|$$

and by (8.6), $2r \geqslant d_2(\theta_k, \theta_{k+1}) \geqslant \mathrm{e}^{-2-k/2}(\theta_{k+1} - \theta_k)$. If $k < 0$, it follows from (8.8) that $\theta_{k+1} - \theta_k > (1/2)\,\mathrm{e}^{k/3}$, hence $\mathrm{e}^{-k/6} < 4\,\mathrm{e}^2 r$ and by (8.11),

$$\bigl|(\underline{\theta}, \bar{\theta}) \cap S\bigr| \leqslant \mathrm{e}^2 r/(4\eta) + 14 + 30.5 r/\eta < \exp\bigl[1.125(r/\eta)^2\bigr] \quad \text{for } r \geqslant 2\eta.$$

One can check that the same bound holds for $k \geqslant 0$ as well as in the cases we started with. It finally follows from Lemma 1 that the metric dimension of $\overline{S}$ is bounded by 4.5.

8.1.2. An example of roughness penalization

It is of common practice, for estimating an unknown function s belonging to some non-compact class of functions, like a Sobolev space W_2^α, to use a penalized maximum likelihood or least squares method with a roughness penalty. In the previous Sobolev case, it is often recommended to use a penalty proportional to $\|s^{(\alpha)}\|^2$. Many examples and further references can be found in Silverman [55], Wahba [61], Eggermont and LaRiccia [31] and Györfi et al. [33].

When the original statistical framework is the white noise framework and we filter it via the trigonometric basis, the initial estimation problem becomes, as explained in Section 6.4.5, estimate s in the Gaussian setting under the assumption that it belongs to the ellipsoid $\mathcal{E}(\boldsymbol{a}) = \mathcal{E}'(\alpha, R)$ with coefficients a_i given by (6.9) for a known value of α but an unknown value of $R = \|s^{(\alpha)}\|$. In order to estimate s, we may consider the family of models $\overline{S}_m = \mathcal{E}'(\alpha, 2^m\sigma)$ for $m \in \mathcal{M} = \mathbb{N}$ and apply Corollary 5 with $\eta_m^2 = 53\,\sigma^2\,2^{2m/(2\alpha+1)}$. It follows from (6.10) with $R = 2^m\sigma$ that

$$\frac{\widetilde{D}_m(\eta_m)}{0.52} = 2\left\lceil \frac{2^{2m/(2\alpha+1)}}{2\pi(53/2)^{1/(2\alpha)}} \right\rceil - 1 \leqslant 2^{2m/(2\alpha+1)} = \frac{\eta_m^2}{53\sigma^2},$$

for all $m \in \mathcal{M}$. Then (8.3) holds with

$$\Sigma = \Sigma(\alpha) < \sum_{m \geqslant 0} \exp\bigl[-2^{2m/(2\alpha+1)}/9.51\bigr]$$

and one can conclude that the corresponding T-estimator $\hat{s}$ satisfies

$$\mathbb{E}_s\bigl[\|s - \hat{s}\|^q\bigr] \leqslant \bigl[1 + 10^{-7}\Sigma(\alpha)\bigr]7^q \inf_{m \in \mathcal{M}} \bigl\{d_2(s, \overline{S}_m) + \eta_m\bigr\}^q \quad \text{for } 1 \leqslant q \leqslant 79.$$

If we choose $m = \inf\{j \in \mathbb{N} \mid \|s^{(\alpha)}\| \leqslant 2^j\sigma\}$, then $s \in \overline{S}_m$, $2^m \leqslant (2\sigma^{-1}\|s^{(\alpha)}\|) \vee 1$ and we conclude that

$$\mathbb{E}_s\bigl[\|s - \hat{s}\|^q\bigr] \leqslant C(q)\bigl[1 + 10^{-7}\Sigma(\alpha)\bigr]\sigma^q\bigl[\bigl(\sigma^{-1}\|s^{(\alpha)}\|\bigr) \vee 1\bigr]^{q/(2\alpha+1)},$$

although $\|s^{(\alpha)}\|$ is unknown. This is the best one can do from the minimax point of view even when $\|s^{(\alpha)}\|$ is known. We recall that the resulting estimator is a penalized least squares estimator on some discrete set with penalty roughly proportional to $\|s^{(\alpha)}\|^{2/(2\alpha+1)}$ (the penalty for the set S_m being proportional to η_m^2), which is different from the classical one. Using a penalty proportional to $\|s^{(\alpha)}\|^2$ would not lead to the right bound for the risk (see also Section 21 of [33] for similar results with random design regression). The choice of the classical penalty is actually motivated by computational reasons and solutions of the minimization problem while our penalty is only motivated by dimensional arguments.

Note that adaptation over all ellipsoids can be obtained in a much simpler way, as explained in Section 6.2 of [17]. The previous construction was merely an illustration of the fact that one can use models of unbounded dimension and the connection with classical roughness penalties.

8.2. Application to the independent and i.i.d. settings

In the independent or i.i.d. settings, the construction of the estimator is more involved since the tests $\psi(t, u, \boldsymbol{X})$ satisfying Assumption 1 are not likelihood ratio tests between t and u and the resulting T-estimators are not penalized maximum likelihood estimators over some discrete set S. Nevertheless Theorem 5 easily translates to those settings.

Corollary 6. *Assume that we observe n independent random variables with unknown joint distribution P_s, $s \in (M, d)$, where d is either the sup-variation $\bar{v}$ or the sup-Hellinger distance $\bar{h}$, and that we have at disposal a finite or countable family of discrete subsets $\{S_m\}_{m\in\mathcal{M}}$ of the set $\overline{M}$ of all distributions for i.i.d. variables. Let those sets S_m satisfy Assumption 3 with*

$$\eta_m^2 \geqslant 16.8\alpha D_m/n \quad \text{for all } m \in \mathcal{M}, \qquad \sum_{m\in\mathcal{M}} \exp\bigl[-n\eta_m^2/(84\alpha)\bigr] = \Sigma < +\infty, \tag{8.12}$$

and $\alpha = 2$ if $d = \bar{v}$, $\alpha = 1$ if $d = \bar{h}$. Then one can build in each case a T-estimator $\hat{s}$ such that, for all $s \in M$,

$$\mathbb{E}_s\bigl[d^q(s,\hat{s})\bigr] \leqslant [1 + 10^{-7}B'\Sigma]5^q \inf_{m\in\mathcal{M}} \bigl\{d(s,S_m) \vee \eta_m\bigr\}^q \quad \text{for } 1 \leqslant q \leqslant 79,$$

with either $d = \bar{v}$ or $d = \bar{h}$ according to the metric used.

Proof. It follows from Proposition 6 that in the independent setting, Assumption 1 holds with $M_T = \overline{M}$, $B = 1$, $\delta = 4d$ and $a = n/8$ when $d = \bar{v}$, $a = n/4$ when $d = \bar{h}$. In view of these values, our assumptions on η_m and D_m imply (7.2) and (7.4) and the conclusion follows from (7.7) of Theorem 5. □

Models based on uniform distributions. As we noticed in Section 5.3.3, it is not possible, whatever $D > 0$, to find a single D-model which is an η-net for the whole set $\overline{S} = \{\overline{P}_t^{\otimes n},\ t \in \mathbb{R}\}$, where $\overline{P}_t = \mathcal{U}_{\mathrm{e}^t}$ denotes the uniform distribution on $[0, \mathrm{e}^t]$ with $t \in \mathbb{R}$. In order to solve this problem, we shall use a device that we call *stratification*, consisting in splitting a large model, like $\overline{S}$, which does not have a finite metric dimension into a countable number of pieces, each one with a finite metric dimension (but not necessarily the same). This method has already been used by Yang and Barron in [71] and Birgé in [12]. Here, we replace $\overline{S}$ (identified to $\mathbb{R}$) by the union of submodels $\{\overline{S}_m,\ m \in \mathbb{Z}\}$ with $\overline{S}_m = [10^{-5}m\Gamma_n, 10^{-5}(m+1)\Gamma_n)$ and Γ_n given by (1.1).

Proof of Proposition 1. In order to apply Corollary 6 with $d = \bar{h}$, we first apply the construction of Section 5.3.3 to each interval $\overline{S}_m$ with $D_m = [10\log(10^{-5}|m|)] \vee (1/2)$ and $\eta_m^2 = 16.8D_m/n$. Then the resulting value of J, as defined by (5.25) with $D = D_m$, satisfies $J > 4.5\exp[(n/84) \vee 2] - 1$ and the corresponding interval I, with $\alpha = 10^{-5}m\Gamma_n$, has a length

$$4J\eta_m^2 > 33.6n^{-1}\bigl(4.5\exp\bigl[(n/84) \vee 2\bigr] - 1\bigr) = 10^{-5}\Gamma_n.$$

It follows that $\overline{S}_m \subset I$ and the set S_m provided by (5.26) (with $\eta = \eta_m$) is an η_m-net for $\overline{S}_m$. Moreover, by Lemma 3, the family of sets $\{S_m,\ m \in \mathbb{Z}\}$ satisfies Assumption 3 with $B' = 4.5$. Our choices for D_m and η_m imply that (8.12) is satisfied and, since $D_m > 1/2$ is equivalent to $|m| \geqslant K' = 105128$,

$$\Sigma = \sum_{m\in\mathbb{Z}} \exp[-D_m/5] = \mathrm{e}^{-1/10}(2K' - 1) + 2 \times 10^{10} \sum_{i \geqslant K'} i^{-2} < 4 \times 10^5.$$

Finally we derive from Corollary 6 that, whatever the true probability P_s,

$$\mathbb{E}_s\bigl[\bar{h}^2(s,\hat{s})\bigr] < 30 \inf_{m\in\mathcal{M}} \bigl\{\bar{h}(s,S_m) \vee \eta_m\bigr\}^2 \leqslant C \inf_{m\in\mathcal{M}} \bigl\{\bar{h}^2(s,\overline{S}_m) + D_m/n\bigr\}.$$

If the original parameter $\theta = \mathrm{e}^t$ satisfies $\log\theta \in \overline{S}_m$, then $10^5\Gamma_n^{-1}\log\theta \in [m, m+1)$, hence $10^{-5}|m| \leqslant \Gamma_n^{-1}|\log\theta| + 10^{-5}\mathbb{1}_{m<0}$ and $D_m \leqslant [11\log(\Gamma_n^{-1}|\log\theta|)] \vee (1/2)$, which concludes the proof with $\hat{\theta} = \exp(\hat{s})$. □

8.3. Density estimation

8.3.1. From Approximation Theory to discrete models

Density estimation in the i.i.d. setting has been the subject of hundreds of papers during the last decades. Modern results tend to put as few assumptions as possible on the underlying densities and insist on adaptive procedures. In particular, they rely quite heavily on results from Approximation Theory. Unfortunately, Approximation Theory deals with approximation of functions, not of densities, and provides approximation spaces, in particular finite dimensional linear spaces, which are not sets of densities. In this section, we wish to explain how, starting from a D-model with

some approximation properties, but which is a set of functions, we can replace it by a D-model which is a set of densities (and can therefore be used for our estimation purposes) and enjoys similar approximation properties. The following proposition actually covers more general situations. In the case of density estimation, one should understand M' as some function space and M_0 as the subset of all density functions in M' (with respect to some given reference measure).

Proposition 12. *Let M_0 and T be two non-empty subsets of some metric space (M', d) with $|T \cap \mathcal{B}(t,r)| < +\infty$ for all $t \in M'$ and $r > 0$. Let $\bar\pi$ be a mapping from T to M_0 such that one of the two following conditions* (8.13) *or* (8.14) *is satisfied for some $\lambda \geqslant 1$:*

$$d\big(t, \bar\pi(t)\big) \leqslant \lambda d(t, M_0) \quad \text{for all } t \in T; \tag{8.13}$$

$$d\big(u, \bar\pi(t)\big) \leqslant \lambda d(u,t) \quad \text{for all } t \in T \text{ and } u \in M_0. \tag{8.14}$$

Then, for any positive ε and η, one can build a subset S' of $\bar\pi(T)$ such that

$$\big|S' \cap \mathcal{B}(t,r)\big| \leqslant \big|T \cap \mathcal{B}(t,3r)\big| \vee 1 \quad \text{for all } t \in M' \text{ and } r \geqslant \eta/2 \tag{8.15}$$

and

$$d(u, S') \leqslant (2\lambda + 1 + \varepsilon) d(u, T) \quad \text{for all } u \in M_0. \tag{8.16}$$

Proof. Once again, our main tool will be a stratification procedure. We fix $\theta = 1 + \varepsilon/(1+\lambda)$ and introduce the increasing sequence of numbers η_j, $j \in \mathbb{N}$ given by $\eta_0 = 0$ and $\eta_j = \theta^{j-1}\eta$ for $j \geqslant 1$. We then set $T_j = \{t \in T \mid \eta_{j-1} \leqslant d(t, \bar\pi(t)) < \eta_j\}$ for $j \geqslant 1$ so that $T = \bigcup_{j \geqslant 1} T_j$. Then we define the sets S_j inductively starting from $S_1 = \bar\pi(T_1)$ and, for $j > 1$, choosing for S_j a maximal η_j-separated subset, therefore an η_j-net, of

$$T_j' = \left\{ t' \in \bar\pi(T_j) \;\middle|\; d\left(t', \bigcup_{1 \leqslant k < j} S_k\right) > \eta_j \right\}.$$

We finally set $S' = \bigcup_{j \geqslant 1} S_j$. It follows from this construction that

$$S_k \cap S_j = \emptyset \quad \text{for } k \neq j \quad \text{and} \quad S' \cap \overline{\mathcal{B}}(u, \eta_j) = \{u\} \quad \text{if } u \in S_j,\ j > 1. \tag{8.17}$$

Moreover, $d(\bar\pi(t), S') \leqslant \theta d(t, \bar\pi(t))$ for all $t \in T$. Indeed this is true when $t \in T_1$ since then $\bar\pi(t) \in S_1 \subset S'$ and if $t \in T_j$ with $j > 1$, $d(\bar\pi(t), S') \leqslant \eta_j = \theta\eta_{j-1} \leqslant \theta d(t, \bar\pi(t))$. Therefore, if (8.13) holds, we can write for any $t \in T$ and $u \in M_0$,

$$\begin{aligned} d(u, S') &\leqslant d(u,t) + d\big(t, \bar\pi(t)\big) + d\big(\bar\pi(t), S'\big) \leqslant d(u,t) + (1+\theta) d\big(t, \bar\pi(t)\big) \\ &\leqslant d(u,t) + \lambda(1+\theta) d(t, M_0) \leqslant \big[1 + \lambda(1+\theta)\big] d(u,t) \end{aligned}$$

and (8.16) follows from our choice of θ and a minimization over $t \in T$. If (8.14) holds, we get

$$\begin{aligned} d(u, S') &\leqslant d\big(u, \bar\pi(t)\big) + d\big(\bar\pi(t), S'\big) \leqslant \lambda d(u,t) + \theta d\big(t, \bar\pi(t)\big) \\ &\leqslant \lambda d(u,t) + \theta\big[d(t,u) + d\big(u, \bar\pi(t)\big)\big] \leqslant \big[\theta + \lambda(1+\theta)\big] d(u,t) \end{aligned}$$

and we conclude as before.

Let us now turn to the proof of (8.15). Let $t \in M'$, $r \geqslant \eta/2$ be given and J be defined by $\eta_J \leqslant 2r < \eta_{J+1}$. Then $J \geqslant 1$ and $d(u,u') < 2r < \eta_{J+1}$ for all u and $u' \in \mathcal{B}(t,r)$. Setting $S_J' = \bigcup_{1 \leqslant j \leqslant J} S_j$, we derive from (8.17) that either

$$\big|(S' \setminus S_J') \cap \mathcal{B}(t,r)\big| = 1 \quad \text{hence} \quad S_J' \cap \mathcal{B}(t,r) = \emptyset \quad \text{or} \quad (S' \setminus S_J') \cap \mathcal{B}(t,r) = \emptyset.$$

In the first case, $|S' \cap \mathcal{B}(t,r)| = 1$. In the second case, $|S' \cap \mathcal{B}(t,r)| = |S_J' \cap \mathcal{B}(t,r)|$. If $u' \in S_J' \cap \mathcal{B}(t,r)$, then $u' = \bar\pi(u)$ for some $u \in T_j$ with $j \leqslant J$, hence $d(u,u') < \eta_J \leqslant 2r$ and $d(u,t) < 3r$ which proves (8.15). □

Remark. If M_0 is a linear subspace of a Hilbert space (M', d), then the projection operator $\bar\pi$ onto M_0 satisfies both (8.13) and (8.14) with $\lambda = 1$. In any case, if $\lambda > 1$, one can always find an approximate minimum distance operator with respect to M_0 which satisfies (8.13).

8.3.2. Density estimation with Hellinger loss

Let us now show how, given some reference measure μ and a general family of models in the space $\mathbb{L}_2(\mu)$, we can derive an estimator of a density s with respect to μ from an i.i.d. sample and bound its Hellinger risk.

Theorem 6. *Let μ be some positive measure on $\mathcal{X}$, M be the set of all probability densities with respect to μ and $\|\cdot\|_2$ be the norm in $\mathbb{L}_2(\mu)$. Let $\{\overline{S}_m\}_{m\in\mathcal{M}}$ be a finite or countable family of subsets of the metric space $\mathbb{L}_2(\mu)$ with respective finite metric dimensions bounded by $\overline{D}_m$ and let $\{\Delta_m\}_{m\in\mathcal{M}}$ be a family of non-negative weights satisfying* (7.9)*. Let $X_1,\dots,X_n$ be an i.i.d. sample from some distribution P_s with density s with respect to μ. One can build a T-estimator $\hat{s}(X_1,\dots,X_n)$ satisfying, for all $s\in M$ and $1\leqslant q\leqslant 79$,*

$$\mathbb{E}_s\bigl[h^q(s,\hat{s})\bigr]\leqslant C(q)\left[1+\frac{\Sigma}{10^7}\right]\inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\|\sqrt{s}-t\|_2+\sqrt{\frac{\overline{D}_m\vee\Delta_m}{n}}\right\}^q. \tag{8.18}$$

Proof. In order to derive suitable D-models S_m from the sets $\overline{S}_m$, we have to apply Proposition 12. Let d_2 denote the distance in $\mathbb{L}_2(\mu)$, g the mapping from M to $\mathbb{L}_2(\mu)$ given by $g(w)=\sqrt{w}$ and let $\bar{\pi}$ be any mapping from $\mathbb{L}_2(\mu)$ onto $g(M)$ such that $d_2(t,\bar{\pi}(t))\leqslant 1.1d_2(t,g(M))$ for all $t\in\mathbb{L}_2(\mu)$. For each m, set $D_m=9\overline{D}_m$ and $\eta_m^2=(84/5)n^{-1}[D_m\vee\Delta_m]$. It follows that (8.12) holds. Let $\eta'_m=\eta_m\sqrt{2}$ and T_m be an η'_m-net for $\overline{S}_m$ with respect to the distance d_2 satisfying

$$\bigl|T_m\cap\mathcal{B}_{d_2}(t,r')\bigr|\leqslant\exp\bigl[\overline{D}_m(r'/\eta'_m)^2\bigr]\quad\text{for all } t\in\mathbb{L}_2(\mu)\text{ and } r'\geqslant 2\eta'_m.$$

Applying Proposition 12 (with $M'=\mathbb{L}_2(\mu)$, $M_0=g(M)$, $d=d_2$, $\lambda=1.1$, $\varepsilon=0.1$ and $\eta=\eta'_m$) to T_m, we get a subset S'_m of $g(M)$ with the following properties:

$$\bigl|S'_m\cap\mathcal{B}_{d_2}(t,r')\bigr|\leqslant\exp\bigl[9(r'/\eta'_m)^2\overline{D}_m\bigr]\quad\text{for all } t\in\mathbb{L}_2(\mu)\text{ and } r'\geqslant 2\eta'_m; \tag{8.19}$$

$$d_2(u,S'_m)\leqslant 3.3d_2(u,T_m)\leqslant 3.3[d_2(u,\overline{S}_m)+\eta'_m]\quad\text{for all } u\in g(M). \tag{8.20}$$

Let us set $S_m=g^{-1}(S'_m)$. Since g is an isometry between (M,d_2) and $(g(M),\sqrt{2}h)$, it follows from (8.19), with $t=g(w)$ and $r'=\sqrt{2}r$, that

$$\bigl|S_m\cap\mathcal{B}_h(w,r)\bigr|\leqslant\exp\bigl[9(r/\eta_m)^2\overline{D}_m\bigr]\quad\text{for all } w\in M\text{ and } r\geqslant 2\eta_m,$$

which implies that Assumption 3 holds with $B'=1$. Since, by (8.20),

$$h(w,S_m)\leqslant 3.3\bigl[d_2\bigl(g(w),\overline{S}_m\bigr)/\sqrt{2}+\eta_m\bigr]\quad\text{for all } w\in M,$$

(8.18) follows from Corollary 6. □

The interest of such a result is that it requires absolutely no assumption on s and on the approximating sets $\overline{S}_m$ apart from the fact that they have a finite metric dimension in $\mathbb{L}_2(\mu)$. In particular finite dimensional linear spaces will do. We therefore completely avoid the usual restrictions connected with maximum likelihood estimation like entropy with bracketing, $\mathbb{L}_\infty$-bounds on s or the introduction of Kullback–Leibler divergences (compare with [71], Theorem 1, [5] Theorem 2, [21,22] or [59]).

Remark. If we assume that we *know* an a priori bound $\overline{R}$ for the $\mathbb{L}_\infty$-norm of the unknown density s, we can immediately derive from Theorem 6 a bound for the $\mathbb{L}_2$-risk. For this we replace the estimator $\hat{s}$ by $\hat{s}_{\overline{R}}=\hat{s}\wedge\overline{R}$. Then

$$\|s-\hat{s}_{\overline{R}}\|_2^2=\int\left(\sqrt{s}+\sqrt{\hat{s}_{\overline{R}}}\right)^2\left(\sqrt{s}-\sqrt{\hat{s}_{\overline{R}}}\right)^2\mathrm{d}\mu\leqslant 4\overline{R}\int\left(\sqrt{s}-\sqrt{\hat{s}_{\overline{R}}}\right)^2\leqslant 4\overline{R}\int\left(\sqrt{s}-\sqrt{\hat{s}}\right)^2=8\overline{R}h^2(s,\hat{s}),$$

and a bound for $\mathbb{E}_s[\|s-\hat{s}_{\overline{R}}\|_2^2]$ could be derived from (8.18). The resulting estimator $\hat{s}_{\overline{R}}$ is not necessarily a true density but this could be fixed. Of course, assuming that we know a bound on the $\mathbb{L}_\infty$-norm of s is rather unrealistic, although such an assumption has often been used in papers dealing with model selection for density estimation or random design regression with $\mathbb{L}_2$-loss. A general treatment of density estimation using $\mathbb{L}_2$-loss requires more technicalities and will therefore be given elsewhere.

To illustrate the power of Theorem 6, we give one application relying on the following proposition which partly summarizes the results of Birgé and Massart [16]. We refer to this paper and the book [27] by DeVore and Lorentz for details on Besov spaces. In what follows, all constants depend on k, but since k is fixed, we omit to emphasize this dependence. As to the linear spaces provided by the proposition they are typically linear spans of finite subsets of some given wavelet basis or spaces of piecewise polynomials.

Proposition 13. *Given a positive integer r, one can find for each $j \geqslant 0$ a family $\{\overline{S}_m\}_{m\in\mathcal{M}_j(r)}$ of D_j-dimensional linear spaces of functions on $\mathbb{R}^k$ with the following properties*:

(i) *The integers D_j and $|\mathcal{M}_j(r)|$ satisfy*

$$D_j \leqslant c_1(r) + c_2(r)2^{jk} \quad \text{and} \quad \log\bigl|\mathcal{M}_j(r)\bigr| \leqslant c_3(r)2^{jk}, \tag{8.21}$$

where the constants $c_i \geqslant 1$ only depend on r and k;

(ii) *for any $p > 0$, $q \geqslant 1$ and α with $r > \alpha > (k/p - k/q)_+$ and any function t belonging to the Besov space $B^\alpha_{p,\infty}([0,1]^k)$ with Besov semi-norm $|t|_{B^\alpha_{p,\infty}}$, one can find some $t' \in \bigcup_{m\in\mathcal{M}_j} \overline{S}_m$ such that*

$$\|t - t'\|_q \leqslant C(r,k,\alpha,p,q)|t|_{B^\alpha_{p,\infty}} 2^{-j\alpha}, \tag{8.22}$$

where $\|\cdot\|_q$ denotes the $\mathbb{L}_q(dx)$-norm on $[0,1]^k$.

Restricting ourselves to the case $k = 1$ for simplicity, we can apply Theorem 6 to the family of models $\{\overline{S}_m\}_{m\in\mathcal{M}}$ with $\mathcal{M} = \bigcup_{i\geqslant 1}\bigcup_{j\geqslant 0}\mathcal{M}_j(2^i)$ provided by the previous proposition. If $m \in \mathcal{M}_j(2^i)$, we choose

$$\Delta_m = 5\bigl[c_3(2^i)2^j + i + j - 14\bigr] \quad \text{and} \quad \overline{D}_m = \bigl(c_1(2^i) + c_2(2^i)2^j\bigr)/2,$$

according to Proposition 8. Then (7.9) holds with $\Sigma < 1.11 \times 10^6$. Applying Proposition 13 with $t = \sqrt{s}$, $r = 2^i > \alpha \geqslant 2^{i-1}$ and $q = 2$, we derive from Theorem 6 that, if $|\sqrt{s}|_{B^\alpha_{p,\infty}} \leqslant R$ with $\alpha > (1/p - 1/2)_+$,

$$\mathbb{E}_s\bigl[h^2(s,\hat{s})\bigr] \leqslant C_1 \inf_{j\geqslant 0}\bigl\{C(\alpha,p)R^2 2^{-2j\alpha} + c_4(\alpha)n^{-1}2^j\bigr\}.$$

An optimization with respect to j leads to the following result.

Theorem 7. *Let $X_1,\dots,X_n$ be an n-sample from some distribution $\overline{P}_s$ with density s with respect to Lebesgue measure on $[0,1]$. One can build a T-estimator $\hat{s}(X_1,\dots,X_n)$ such that, if the Besov semi-norm of $\sqrt{s}$ satisfies $|\sqrt{s}|_{B^\alpha_{p,\infty}} \leqslant R$ for some $p > 0$, $\alpha > (1/p - 1/2)_+$ and $R \geqslant 1/\sqrt{n}$, then*

$$\mathbb{E}_s\bigl[h^2(s,\hat{s})\bigr] \leqslant C(\alpha,p)R^{2/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}. \tag{8.23}$$

Note that the use of Hellinger distance allows to get adaptation for the whole domain $\alpha > (1/p - 1/2)_+$ which is, to our knowledge, new for density estimation, the usual results being restricted to an interval of the form $(1/p, r)$. We could prove in the same way a multidimensional analogue.

One can also apply to the i.i.d. setting the results of [17], Section 4.1 to mix several families of approximating spaces. In particular, one could design a T-estimator that satisfies simultaneously the conclusions of Proposition 1 and Theorem 7. Therefore, if s is the uniform density on $[0,\theta]$ for some $\theta > 0$, we get the usual parametric rate n^{-1} for the quadratic Hellinger risk while (8.23) applies when $\sqrt{s}$ is a density belonging to some Besov ball. This is an illustration of the fact that T-estimators allow to cope simultaneously with parametric and non-parametric models, getting the parametric rate if the true distribution is in the parametric model and the non-parametric one otherwise.

8.3.3. A parallel with the white noise framework

One should stress the fact that the results of Theorem 6 about the i.i.d. setting completely parallel those which hold in the white noise framework. To be more precise, let us observe that Corollary 5 also applies to the white noise framework via the identification mentioned in Section 5.2.2.

Corollary 7. *Assume we are in the white noise framework, observing the process Y given by* (5.18) *with unknown parameter s and that we have at disposal a finite or countable family of subsets $\{\overline{S}_m\}_{m\in\mathcal{M}}$ of $\mathbb{L}_2([0,1],\mathrm{d}x)$ with respective finite metric dimensions bounded by $\overline{D}_m$. If the family of non-negative weights $\{\Delta_m\}_{m\in\mathcal{M}}$ satisfies* (7.9)*, one can build a T-estimator $\hat{s}(Y)$ satisfying, for all $s\in\mathbb{L}_2([0,1],\mathrm{d}x)$ and $1\leqslant q\leqslant 79$,*

$$\mathbb{E}_s\big[\|s-\hat{s}\|_2^q\big]\leqslant C(q)\left[1+\frac{\Sigma}{10^7}\right]\inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\|s-t\|_2+\sqrt{\frac{\overline{D}_m\vee\Delta_m}{n}}\right\}^q. \tag{8.24}$$

Clearly, (8.18) is the exact analogue of (8.24). This means that *any model selection procedure for the white noise framework based on a T-estimator has an analogue for density estimation with Hellinger loss in the i.i.d. setting which has exactly the same performances without any additional restriction, modulo the replacement of s in the white noise framework by $\sqrt{s}$ in the i.i.d. setting*. In particular, all the results obtained in Section 6 of [17], which are based on approximation by finite dimensional linear subspaces of $\mathbb{L}_2$, can immediately be translated into parallel results for the i.i.d. setting with Hellinger distance, provided that the assumptions are now put on $\sqrt{s}$. Our Theorem 7 is just one possible illustration of this fact among many others.

Of course, the previous remark is not a result of asymptotic equivalence of experiments, as defined by Le Cam ([42] and [45]) and illustrated, for instance, by Brown and Low [19] or Nussbaum [51], among other examples. Our parallelism has some limitations: it holds up to constants, it is restricted to loss functions of the form $\|s-\hat{s}\|_2^q$ and $h^q(s,\hat{s})$, although this could be generalized via (7.10), and to specific estimators, namely T-estimators. On the other hand, it has also some advantages: it is non-asymptotic, the parallelism is explicit and it works for classes of functions for which no equivalence of experiments result exists, as far as we know (for instance the class of densities s on $[0,1]$ such that $\sqrt{s}$ is 1/4-Hölderian).

8.3.4. Density estimation with $\mathbb{L}_1$-loss

If we want to use the $\mathbb{L}_1$-loss for density estimation, it is enough to get the result for the loss based on the variation distance since for two probabilities P and Q with respective densities f and g, $\|f-g\|_1=2v(P,Q)$. Combining Corollary 6 and Proposition 12 we get the following analogue of Theorem 6 for the independent setting. The proof being quite similar, it will be omitted.

Theorem 8. *Let μ be some positive measure on $\mathcal{X}$, $\overline{M}_\mu$ be the set of all probability densities with respect to μ and $\|\cdot\|_1$ be the norm in $\mathbb{L}_1(\mu)$. Let $\{\overline{S}_m\}_{m\in\mathcal{M}}$ be a finite or countable family of subsets of the metric space $\mathbb{L}_1(\mu)$ with respective finite metric dimensions bounded by $\overline{D}_m$ and $\{\Delta_m\}_{m\in\mathcal{M}}$ be a family of non-negative weights satisfying* (7.9)*. Let $X_1,\dots,X_n$ be n independent random variables on $\mathcal{X}$ with joint distribution $P_s=\bigotimes_{i=1}^n\overline{P}_i$ on $\mathcal{X}^n$ and M be the set of all such product distributions. One can build a T-estimator $\hat{s}(X_1,\dots,X_n)$ with values in $\overline{M}_\mu$ satisfying, for all $s\in M$, $1\leqslant q\leqslant 79$, and $\bar{v}(s,t)=\sup_{1\leqslant i\leqslant n}v(\overline{P}_i,t\cdot\mu)$ for $t\in\overline{M}_\mu$,*

$$\mathbb{E}_s\big[\bar{v}^q(s,\hat{s})\big]\leqslant C(q)[1+10^{-7}\Sigma]\inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\bar{v}(s,t)+\sqrt{n^{-1}[\overline{D}_m\vee\Delta_m]}\right\}^q.$$

Theorem 1 immediately follows from this last result and Proposition 13 by a proof which is completely similar to the proof of Theorem 7 and will therefore be omitted. Note here that one can bound the risk of the estimator $\hat{s}$ of Theorem 1 even if $\overline{P}_s$ is not absolutely continuous with respect to Lebesgue measure μ or if its density s does not belong to such Besov spaces. If s' is any density satisfying $|s'|_{B^\alpha_{p,\infty}}\leqslant R$ we get, when $k=1$,

$$\mathbb{E}_s\big[v^q(\overline{P}_s,\hat{s}\cdot\mu)\big]\leqslant C(\alpha,p,q)R^{q/(2\alpha+1)}n^{-q\alpha/(2\alpha+1)}+C'(q)v^q(\overline{P}_s,s'\cdot\mu).$$

8.4. Application to bounded regression

In the context of bounded regression, Proposition 5 implies that Assumption 2 holds with $\gamma'(t,\boldsymbol{X})$ being either $\sum_{i=1}^n[Y_i-t(X_i)]^2$ (random design) or $\sum_{i=1}^n[Y_i-t(x_i)]^2$ (fixed design),

$$\tau=25n/98,\quad d=d_2\text{ or }d_n,\quad B=1,\quad a=3n/100,\quad \kappa=4\quad\text{and}\quad \kappa'=15.68.$$

Arguing as for the Gaussian setting, we see that, given a subset S of M which satisfies Assumption 3, (7.2) and (7.4), we can apply Theorem 5 and prove the existence of a minimizer $\hat{s}$ (which is not necessarily unique here) of the penalized least squares criterion $\gamma'(t, \boldsymbol{X}) + 25n\eta^2(t)/98$ with respect to $t \in S$. More formally, we get:

Corollary 8. *In the bounded regression setting* (*with either random or fixed design*) *as described in Section* 5.2.3*, let* $S \subset M$ *satisfy Assumption* 3 *with*

$$\eta_m^2 \geqslant 140 D_m/n \quad \textit{for all } m \in \mathcal{M} \quad \textit{and} \quad \sum_{m\in\mathcal{M}} \exp[-n\eta_m^2/700] = \Sigma < +\infty. \tag{8.25}$$

Then there exists a.s. at least one minimizer over S *of the penalized least squares criterion*

$$\sum_{i=1}^{n}\bigl[Y_i - t(X_i)\bigr]^2 + 25n\eta^2(t)/98 \quad \textit{or} \quad \sum_{i=1}^{n}\bigl[Y_i - t(x_i)\bigr]^2 + 25n\eta^2(t)/98,$$

with $\eta(t) = \inf_{\{m\in\mathcal{M}|t\in S_m\}} \eta_m$*, and any such minimizer* $\hat{s}$ *satisfies, for all* $s \in M$*,*

$$\mathbb{E}_s\bigl[\|s-\hat{s}\|^q\bigr] \leqslant [1 + 10^{-7}B'\Sigma]5^q \inf_{m\in\mathcal{M}}\left\{\Bigl(\inf_{t\in S_m}\|s-t\|\Bigr) \vee \eta_m\right\}^q \quad \textit{for } 1 \leqslant q \leqslant 79.$$

8.4.1. The fixed design case

It is not difficult to derive D-models in the context of bounded regression with fixed design and $|\{x_1, \ldots, x_n\}| = n'$. To get a D-model, it suffices to start with a subset T' of the metric space $(\mathbb{R}^{n'}, d_n)$ with a metric dimension bounded by D. For instance a 2D-dimensional linear subspace of $\mathbb{R}^{n'}$ would do. We can then construct an η-net T for T' satisfying (4.18) with $S' = T$ and $B' = 1$. With $\bar{\pi}$ given by (6.5), we get $d(u, \bar{\pi}(t)) \leqslant d(u, t)$ for all $u \in M$ and $t \in \mathbb{R}^{n'}$. It follows that (8.14) holds with $\lambda = 1$, $M' = \mathbb{R}^{n'}$ and $M_0 = M$. We may therefore apply Proposition 12 (with $\varepsilon = 0.1$ for instance) to T, getting a subset S' of M which is a D-model with parameters η, $9D$ and 1, and, by (8.16), satisfies $d_n(s, S') \leqslant 3.1[d_n(s, T') + \eta]$ for all $s \in M$. Starting from a family $\{T'_m, m \in \mathcal{M}\}$ of subsets of $\mathbb{R}^{n'}$ with respective metric dimensions $D_m/9$ and choosing η_m to satisfy (8.25), we can use the previous construction to derive a family of D-models S_m with parameters η_m, D_m and 1 and apply Corollary 8.

8.4.2. The random design case

The previous construction also applies to the random design case if the distribution μ of the design is known. Otherwise it cannot be performed by the statistician since it involves the unknown distance d_2. We may alternatively build families of T-models via the uniform distance as explained in Section 6.4.4. Let us give here a simple illustration of this fact, assuming that $n \geqslant 30$.

For $m = (j, k) \in \mathbb{N}^2$, let us set $|m| = j(k+1)$ and define

$$\mathcal{M} = \bigl\{m = (j,k) \in \mathbb{N}^2 \mid 3 \leqslant |m| \leqslant n/10\bigr\}.$$

For $j \in \mathbb{N}^\star$ we denote by $\mathcal{I}_j$ a partition of $[0, 1]$ into j intervals of equal length j^{-1} and for each $m \in \mathcal{M}$, we consider the $|m|$-dimensional linear space $\mathcal{P}_m$ of piecewise polynomials on $[0, 1]$ of degree not larger than k based on the partition $\mathcal{I}_j$. This means that the restriction of any element of $\mathcal{P}_m$ to any interval of $\mathcal{I}_j$ is a polynomial of degree not larger than k. Setting $\eta_m = [18|m|n^{-1}\log(n/|m|)]^{1/2}$ (hence $\eta_m \leqslant 3\sqrt{(\log 10)/5}$), we use the construction of Section 6.4.4 to deduce from $\mathcal{P}_m$ a D-model S_m with parameters η_m, $D_m = (|m|/4)\log(2\eta_m^{-1} + 1) > 1/2$ and $B' = 1$ such that

$$d_2(s, S_m) \leqslant d_\infty(s, \mathcal{P}_m) + \eta_m \quad \text{for all } s \in M.$$

One can then check that (8.25) holds with $\Sigma < 100$, so that Corollary 8 leads to an M-estimator $\hat{s}$ satisfying, for all $s \in M$ and $1 \leqslant q \leqslant 79$,

$$\mathbb{E}_s\bigl[\|s-\hat{s}\|^q\bigr] \leqslant [1 + 10^{-5}]5^q \inf_{m\in\mathcal{M}}\left\{d_\infty(s, \mathcal{P}_m) + \sqrt{18|m|n^{-1}\log\bigl(n/|m|\bigr)}\right\}^q. \tag{8.26}$$

As noticed in Section 6.4.4, the use of approximation in the uniform distance results in the unexpected $\log(n/|m|)$ factor in the risk, as compared, for instance, to what we would get for density estimation. The following example shows that this entails unusual logarithmic factors in our risk bounds for estimating Hölderian functions.

Let s have a derivative of order l satisfying the following Hölderian condition for some $\alpha \in (0,1]$ and $R \geqslant (n/\log n)^{-1/2}$:

$$\left|s^{(l)}(x) - s^{(l)}(x)\right| \leqslant R|x-y|^{\alpha} \quad \text{for all } x, y \in [0,1].$$

It is well-known from Approximation Theory that, if $m = (j,l)$ and $\beta = l + \alpha$, then $d_{\infty}(s, \mathcal{P}_m) \leqslant C(l)Rj^{-\beta}$. Putting this bound into (8.26) and optimizing with respect to those $m \in \mathcal{M}$ with $k = l$, we get

$$\mathbb{E}_s\left[\|s - \hat{s}\|^2\right] \leqslant \left[C(l)R^{2/(2\beta+1)}(n/\log n)^{-2\beta/(2\beta+1)}\right] \wedge 1.$$

9. Aggregation of estimators

As shown by the results of the previous section, T-estimators allow to select among many models of different types and therefore provide adaptive estimators. An alternative way to reach the same objective is to start from a large family of estimation procedures and use an aggregation method to combine them, either selecting one in the family, or mixing them together. One method, based on some progressive mixing of estimators and information-theoretically oriented appears in Yang [65] and Catoni [23]. Further developments and adaptation results are to be found in [66,67] and [24]. An alternative stochastically oriented approach using one half of the sample to design estimators and the other half to choose between them or mix them together appears in Juditsky and Nemirovski [37] and Nemirovski [50] or Yang ([68] and [70]). Optimal rates for the various types of aggregation have been given in [37,56] and [70]. Additional results are to be found in [63,2] and [20].

The fact that, in the case of i.i.d. observations, our method can also be used to aggregate preliminary estimators has been suggested to us by Yannick Baraud and Sacha Tsybakov. Proceeding as in [37], we split the sample in two parts, use one part to build a countable family of estimators and the second part either to select one estimator in the family or to combine them together. Since the estimators have been built from an independent sample, they can be viewed as deterministic points in M for this step as in [56] and [20]. Consequently, we can reduce the analysis of this second step to the problem of aggregating a countable family of points in M, which is actually a particular case of the general framework of Section 7.

Although there are similar purposes and results in aggregation methods used by the previously mentioned authors and model selection based on T-estimators, like trying to get the best of each estimator or model or getting adaptive results over large classes of functions, there are also some differences. Aggregation methods work with all types of preliminary estimators and can even mix different ways of combining them (selection, convex or linear aggregation as in [70] or [20]). Moreover they are generally based on effective algorithms and result in risk bounds involving much better constants than ours, but they do not provide the initial estimators, while T-estimators can be used to build initial model-based estimators and possibly mix them with others. Most aggregation results also require some boundedness assumptions as in [37,66,68,70] or [63]. They also only deal with i.i.d. observations (density estimation or regression with random design) with one exception for Gaussian sequences with known variance, as explained by Nemirovski in [50], Chapter 6. This is due to the fact that, by a randomization procedure, one can derive from a variable $X \sim \mathcal{N}(\mu, \sigma^2)$ a pair of independent variables with distribution $\mathcal{N}(\mu, 2\sigma^2)$. This allows to duplicate a Gaussian sequence with only a small increase of the variance and get one sequence to build the initial estimators and one for the aggregation step. T-estimators can be used for other frameworks like bounded regression with fixed design where no sample splitting is possible.

9.1. Selecting a point from a countable family

Suppose we are in a statistical framework for which Assumption 1 (or 2) is satisfied and we are given a countable subset $S = \{t_m,\ m \in \mathcal{M}\}$ of M_T (possibly preliminary estimators built on an independent sample) and a family of positive weights $\{\Delta_m,\ m \in \mathcal{M}\}$. In order to select one point $t_{\widehat{m}}$ in the family from the observations via a selection procedure $\widehat{m}(\boldsymbol{X})$, we consider the family of models $S_m = \{t_m\}$ and define a function η on S by $\eta(t_m) = \sqrt{a^{-1}\Delta_m}$. From the tests provided by Assumption 1 or 2, we can build T_{ε} or M-estimators $t_{\widehat{m}}$. Their performances are given by the following simplified version of Theorem 5 especially tailored for this case.

Theorem 9. *Let (M,d) be a metric space and Assumption 1 or 2 hold with $\delta=\kappa d$, $\kappa>0$. Let $S=\{t_m,\ m\in\mathcal{M}\}$ be a countable subset of M_T and $\{\Delta_m,\ m\in\mathcal{M}\}$ be a set of weights satisfying*

$$\sum_{m\in\mathcal{M}}\exp[-\Delta_m]=\Sigma<+\infty\quad\text{and}\quad\lambda=\inf_{m\in\mathcal{M}}\Delta_m>0. \tag{9.1}$$

Let $0<\varepsilon\leqslant\sqrt{2}$. Then, under Assumption 1, T_ε-estimators exist $\mathbb{P}_s$-a.s. for all $s\in M$ (and also T_0-estimators if $\mathcal{M}$ is finite) and M-estimators as well under Assumption 2. Moreover, if $t_{\widehat{m}}$ is any of them, it satisfies, for all $q\geqslant 1$,

$$\mathbb{E}_s\bigl[d^q(s,t_{\widehat{m}})\bigr]\leqslant\bigl[1+B\Sigma\zeta_q(\lambda)\bigr](\kappa+1)^q\inf_{m\in\mathcal{M}}\left\{d(s,t_m)\vee\left[\kappa^{-1}\sqrt{2a^{-1}\Delta_m}\right]\right\}^q. \tag{9.2}$$

In particular, if $\lambda\geqslant 1/2$ and $1\leqslant q\leqslant(2\lambda)\wedge 17$,

$$\mathbb{E}_s\bigl[d^q(s,t_{\widehat{m}})\bigr]\leqslant\left(1+\frac{B\Sigma q}{2\,\mathrm{e}^\lambda}\right)(\kappa+1)^q\inf_{m\in\mathcal{M}}\left\{d(s,t_m)\vee\left[\frac{\sqrt{2a^{-1}\Delta_m}}{\kappa}\right]\right\}^q. \tag{9.3}$$

Proof. We use here a simplified version of the arguments leading to Theorem 5. Since, for $z\in\mathbb{R}_+$, $\{t\in S\mid\eta(t)\leqslant z\}$ is finite by (9.1), there exists at least one $s'\in S$ such that $[\kappa d(s,s')]\vee(\sqrt{2}\eta(s'))=\inf_{t\in S}\{[\kappa d(s,t)]\vee[\sqrt{2}\eta(t)]\}=y_0$ and it satisfies, for $y\geqslant y_0$,

$$\begin{aligned}\mathbb{P}_s\bigl[\mathcal{D}_{\boldsymbol{X}}(s')>y\bigr]&\leqslant\sum_{\substack{t\in S\\ d(t,s')>y}}\mathbb{P}_s\bigl[\psi(s',t,\boldsymbol{X})=1\bigr]\leqslant B\sum_{\substack{t\in S\\ d(t,s')>y}}\exp\bigl[-a\bigl(d^2(t,s')-\eta(s')^2+\eta(t)^2\bigr)\bigr]\\&\leqslant B\Sigma\exp\bigl[-a\bigl(y^2-\eta(s')^2\bigr)\bigr]\leqslant B\Sigma\exp[-ay^2/2].\end{aligned}$$

Since $\varepsilon\eta(s')\leqslant y_0$, this bound and (4.10) imply that $\mathbb{P}_s[d(s',\hat{s})>y]<B\Sigma\exp[-ay^2/2]$ for $y\geqslant y_0$. It then follows from Proposition 3, $ay_0^2\geqslant 2\lambda$ and $d(s,s')\leqslant\kappa^{-1}y_0$ that

$$\mathbb{E}_s\bigl[d^q(s,\hat{s})\bigr]\leqslant\bigl[1+B\Sigma\zeta_q(ay_0^2/2)\bigr](y_0+\kappa^{-1}y_0)^q\leqslant\bigl[1+B\Sigma\zeta_q(\lambda)\bigr](\kappa+1)^q(y_0/\kappa)^q,$$

which proves (9.2). Then (9.3) follows from (5.5). The case of Assumption 2 is quite similar, with the analogue of (7.18). We omit the details. □

This theorem applies in particular to the three settings considered in this paper with $B=1$ and suitable values of a and κ, leading to the following corollary which could be used, for instance, for bandwith selection in the i.i.d. setting.

Corollary 9. *Given a countable subset $S=\{t_m,\ m\in\mathcal{M}\}$ of $\overline{M}$ in the independent setting or of M in the Gaussian or bounded regression settings together with a family of weights $\{\Delta_m,\ m\in\mathcal{M}\}$ satisfying (9.1), one can build in all cases a selection procedure $\widehat{m}(\boldsymbol{X})$ with values in $\mathcal{M}$ and the following properties: for all $s\in M$ and $q\geqslant 1$, we get, in the independent setting,*

$$\mathbb{E}_s\bigl[d^q(s,t_{\widehat{m}})\bigr]\leqslant\bigl[1+\Sigma\zeta_q(\lambda)\bigr]5^q\inf_{m\in\mathcal{M}}\left\{d(s,t_m)\vee\sqrt{(\alpha/2)n^{-1}\Delta_m}\right\}^q,$$

with $\alpha=1$ if $d=\bar{h}$, $\alpha=2$ if $d=\bar{v}$, in the bounded regression setting,

$$\mathbb{E}_s\bigl[d^q(s,t_{\widehat{m}})\bigr]\leqslant\bigl[1+\Sigma\zeta_q(\lambda)\bigr]5^q\inf_{m\in\mathcal{M}}\left\{d(s,t_m)\vee\sqrt{(25/6)n^{-1}\Delta_m}\right\}^q,$$

and in the Gaussian setting

$$\mathbb{E}_s\bigl[d^q(s,t_{\widehat{m}})\bigr]\leqslant\bigl[1+\Sigma\zeta_q(\lambda)\bigr]7^q\inf_{m\in\mathcal{M}}\left\{d(s,t_m)\vee\left(2\sigma\sqrt{\Delta_m/3}\right)\right\}^q.$$

9.1.1. Aggregating T-estimators based on a single model

In all situations where we can split our sample in two parts, use the first one to build a family of estimators and the second one to select one estimator in the family, one could, instead of using the construction of Section 7.1 with a family $\{S_m,\ m\in\mathcal{M}\}$ of D-models, use instead the following procedure: build a T-estimator $\hat{s}_m$ on each D-model S_m

using the first part of the sample, which results in a family of T-estimators $\{\hat{s}_m,\ m\in\mathcal{M}\}$, then use the second part of the sample to select one estimator $\tilde{s}=\hat{s}_{\widehat{m}}$ according to the recipe of Theorem 9. The method also applies to Gaussian sequences if we duplicate our observation according to the recipe of Chapter 6 of Nemirovski [50].

The performances of this new two-steps procedure, which results from the successive applications of Theorems 3 and 9, are comparable to those of the one-step procedure of Section 7.1 with risk bounds similar to those derived from Theorem 5. For the sake of comparison with Corollary 6, we provide below the risk bounds corresponding to the i.i.d. setting for this two-steps procedure.

Proposition 14. *Assume that we observe an even number n of i.i.d. random variables with unknown distribution $\overline{P}_s$, $s\in(\overline{M},d)$, where d is either the variation or the Hellinger distance, and that we have at disposal a finite or countable family of D-models $\{S_m\}_{m\in\mathcal{M}}$ subsets of the set $\overline{M}$ of all distributions for i.i.d. variables and with respective parameters η_m, D_m and B', $D_m\geqslant 1/2$. Assume moreover that*

$$\eta_m^2\geqslant 12\alpha D_m/n\quad \textit{for all } m\in\mathcal{M};\qquad \sum_{m\in\mathcal{M}}\exp\bigl[-n\eta_m^2/(8\alpha)\bigr]=\Sigma<+\infty;$$

with $\alpha=2$ if $d=v$, $\alpha=1$ if $d=h$. Then one can use the procedure described just before to build an estimator $\tilde{s}$ such that, for all $s\in\overline{M}$ and $q\geqslant 1$,

$$\mathbb{E}_s\bigl[d^q(s,\tilde{s})\bigr]\leqslant C(q,B',\Sigma)\inf_{m\in\mathcal{M}}\bigl\{d(s,S_m)+\eta_m\bigr\}^q,$$

with either $d=v$ or $d=h$ according to the metric used.

Proof. Since $n=2p$, we use $X_1,\dots,X_p$ to build on each D-model S_m a T-estimator $\hat{s}_m=\hat{s}_m(X_1,\dots,X_p)$ and apply Corollary 2 to each of them. Since, in the i.i.d. setting, $\kappa=4$ and $a=p/(4\alpha)=n/(8\alpha)$, hence $2a\eta_m^2\geqslant 3D_m$ and $8a\eta_m^2/3\geqslant 2$, (5.7) becomes

$$\mathbb{E}_s\bigl[d^q\bigl(s,\hat{s}_m(X_1,\dots,X_p)\bigr)\bigr]\leqslant 5^q\bigl[1+2.2B'\zeta_q(2)\bigr]\bigl[d(s,S_m)\vee\eta_m\bigr]^q\quad\text{for all } m\in\mathcal{M},\ q\geqslant 1.$$

Working conditionally to $X_1,\dots,X_p$, we then apply Theorem 9 to the family $S=\{\hat{s}_m,\ m\in\mathcal{M}\}$ using the variables $X_{p+1},\dots,X_n$. This results in the estimator $\tilde{s}=\hat{s}_{\widehat{m}}$. Since here $\lambda\geqslant 3/4$, (9.2) becomes

$$\mathbb{E}_s\bigl[d^q(s,\tilde{s})\mid X_1,\dots,X_p\bigr]\leqslant 5^q\bigl[1+\Sigma\zeta_q(3/4)\bigr]\inf_{m\in\mathcal{M}}\bigl\{d\bigl(s,\hat{s}_m(X_1,\dots,X_p)\bigr)\vee[\sqrt{2}\eta_m/4]\bigr\}^q$$

and an integration with respect to $X_1,\dots,X_p$ allows us to conclude. □

The advantage of the two-steps procedure is that it allows to mix estimators which are not T-estimators with those derived from the D-models.

9.1.2. Aggregating preliminary estimators

Let us here illustrate aggregation of preliminary estimators for bounded regression with random design. Suppose now that we have at hand an n-sample $Z_1,\dots,Z_n$, $Z_i=(X_i,Y_i)$, with an even number of observations $n=2p$ and a countable number of procedures $\{\hat{s}_m,\ m\in\mathcal{M}\}$ with values in M to estimate s together with a family of weights $\Delta_m\geqslant 1$ satisfying $\sum_{m\in\mathcal{M}}\exp[-\Delta_m]\leqslant \mathrm{e}$ (hence $\lambda\geqslant 1$ and $\Sigma\leqslant\mathrm{e}$ in (9.1)). We first evaluate $\hat{s}_m(Z_1,\dots,Z_p)$ for each m and then apply Corollary 9 to the second half of the sample, $Z_{p+1},\dots,Z_n$, conditionally to the first half, with $S=\{\hat{s}_m(Z_1,\dots,Z_p),\ m\in\mathcal{M}\}$. If $\hat{s}$ is the resulting estimator we get, for $q=2$,

$$\mathbb{E}_s\bigl[\|\hat{s}-s\|_2^2\mid Z_1,\dots,Z_p\bigr]\leqslant 50\inf_{m\in\mathcal{M}}\bigl\{\|\hat{s}_m-s\|_2^2+\bigl[25\Delta_m/(6p)\bigr]\bigr\}.$$

Integrating with respect to $Z_1,\dots,Z_p$ leads to

$$\mathbb{E}_s\bigl[\|\hat{s}-s\|_2^2\bigr]\leqslant 50\inf_{m\in\mathcal{M}}\bigl\{\mathbb{E}_s\bigl[\bigl\|\hat{s}_m(Z_1,\dots,Z_p)-s\bigr\|_2^2\bigr]+\bigl[25\Delta_m/(3n)\bigr]\bigr\}.\tag{9.4}$$

Given a D_m-dimensional linear space T_m of bounded functions on $\mathcal{X}$, we consider the least squares estimator $\hat{s}_m$ over T_m, i.e. the minimizer, with respect to $t\in T_m$, of $\sum_{i=1}^n[Y_i-t(X_i)]$. Then $\hat{t}_m=\bar{\pi}(\hat{s}_m)$, with $\bar{\pi}$ given by (6.5), is an estimator with values in M which, according to Theorem 11.3 of [33], satisfies

$$\mathbb{E}_s\bigl[\bigl\|\hat{s}_m(Z_1,\dots,Z_p)-s\bigr\|_2^2\bigr]\leqslant C\bigl[d_2^2(s,T_m)+D_m p^{-1}\log p\bigr].\tag{9.5}$$

Together with (9.4), this proves our Theorem 2. We could, for instance, choose for the spaces T_m spaces of piecewise polynomials of varying degrees over different partitions. Rather than pursuing into this direction, let us analyze the simplest case of histograms, i.e. polynomials of degree 0. At the price of some extra logarithmic factors due to the use of (9.5), one could extend the results we get for histograms to truncated least squares estimators over families of piecewise polynomials with the same partitions by a straightforward application of Theorem 2.

9.2. Partition selection for histograms

We wish to explain here how to select a histogram estimator among a family of them based on different partitions of the underlying space $\mathcal{X}$, dealing simultaneously with the i.i.d. and the bounded regression with random design settings. In both cases we assume that we have chosen a countable family $\mathcal{M}$ of finite partitions $m = \{I_1, \ldots, I_{|m|}\}$ of $\mathcal{X}$ and a family of weights $\Delta_m \geqslant 1$ on $\mathcal{M}$ such that $\Sigma = \sum_{m\in\mathcal{M}} \exp[-\Delta_m] \leqslant 2\,\mathrm{e}$.

9.2.1. The i.i.d. setting

Suppose we observe i.i.d. random variables $X_1, \ldots, X_n$ on some measurable space $\mathcal{X}$ with unknown distribution $\overline{P}_s$. Given a finite reference measure μ on $\mathcal{X}$ and a finite partition m of $\mathcal{X}$ such that $\mu(I_j) > 0$ for $1 \leqslant j \leqslant |m|$, the histogram estimator of $\overline{P}_s$ based on the partition m has a density $\hat{s}_m$ with respect to μ given by

$$\hat{s}_m(X_1, \ldots, X_n) = \frac{1}{n}\sum_{j=1}^{|m|} \frac{N_j}{\mu(I_j)}\mathbb{1}_{I_j}, \quad \text{with } N_j = \sum_{i=1}^{n} \mathbb{1}_{I_j}(X_i). \tag{9.6}$$

One easily shows that

$$\mathbb{E}_s\big[v(\overline{P}_s, \hat{s}_m\cdot\mu)\big] \leqslant v(\overline{P}_s, \bar{s}_m\cdot\mu) + \frac{1}{2}\sqrt{\frac{|m|-1}{n}} \quad \text{with } \bar{s}_m = \sum_{j=1}^{|m|} \frac{\overline{P}_s(I_j)}{\mu(I_j)}\mathbb{1}_{I_j}. \tag{9.7}$$

Given $2n$ i.i.d. observations with distribution $\overline{P}_s$ and the family $\mathcal{M}$, we can first build all the histograms $\hat{s}_m(X_1, \ldots, X_n)$ based on the first n observations as in (9.6) and then select a partition, $\widehat{m}(X_{n+1}, \ldots, X_{2n}) \in \mathcal{M}$ using the last n observations which results in a density estimator $\tilde{s} = \hat{s}_{\widehat{m}}$. Applying Corollary 9 with $q = 1$ and $\lambda \geqslant 1$ conditionally on $X_1, \ldots, X_n$ gives

$$\mathbb{E}_s\big[v(\overline{P}_s, \tilde{s}\cdot\mu) \mid X_1, \ldots, X_n\big] \leqslant 10 \inf_{m\in\mathcal{M}}\Big\{v(\overline{P}_s, \hat{s}_m\cdot\mu) + \sqrt{n^{-1}\Delta_m}\Big\}.$$

Integrating with respect to $X_1, \ldots, X_n$ and applying (9.7) leads to

$$\mathbb{E}_s\big[v\big(\overline{P}_s, \tilde{s}\cdot\mu\big)\big] \leqslant 10 \inf_{m\in\mathcal{M}}\left\{v(\overline{P}_s, \bar{s}_m\cdot\mu) + \frac{(1/2)\sqrt{|m|-1} + \sqrt{\Delta_m}}{\sqrt{n}}\right\}. \tag{9.8}$$

This result, which allows arbitrary families of partitions and requires no assumption at all on $\overline{P}_s$ should be compared, for instance, with [21,22] and [28]. Note that, if $\Delta_m \leqslant c(|m|-1)$ for all $m \in \mathcal{M}$ and some $c > 0$, this bound corresponds, up to some multiplicative constant depending on c, to the risk bound (9.7) for the best histogram among the family.

9.2.2. Bounded regression with random design

In the bounded regression with random design setting with observations $Z_1, \ldots, Z_n$, $Z_i = (X_i, Y_i)$, we may analogously define the histogram (or partitioning) estimator based on the partition m of $\mathcal{X}$, using the convention $0/0 = 0$, by

$$\hat{s}_m(Z_1, \ldots, Z_n) = \sum_{j=1}^{|m|} \frac{\sum_{i=1}^{n} Y_i\mathbb{1}_{I_j}(X_i)}{N_j}\mathbb{1}_{I_j}, \quad \text{with } N_j = \sum_{i=1}^{n}\mathbb{1}_{I_j}(X_i).$$

It follows from (6) page 98 of Beirlant and Györfi [7] and some easy computations that

$$\mathbb{E}_s\big[\|\hat{s}_m - s\|_2^2\big] \leqslant \|\bar{s}_m - s\|_2^2 + |m|/n \quad \text{with } \bar{s}_m = \sum_{j=1}^{|m|} \frac{\int_{I_j} s(x)\,\mathrm{d}\mu(x)}{\mu(I_j)}\mathbb{1}_{I_j}. \tag{9.9}$$

Starting with $2n$ observations, the family $\mathcal{M}$ and proceeding as in the i.i.d. setting, we derive again from Corollary 9 that the aggregated estimator $\hat{s}$ satisfies

$$\mathbb{E}_s\big[\|\hat{s}-s\|_2^2\big] \leqslant C \inf_{m\in\mathcal{M}}\big\{\|\bar{s}_m - s\|_2^2 + n^{-1}\big(|m|+\Delta_m\big)\big\}, \tag{9.10}$$

with $\bar{s}_m$ given by (9.9).

9.2.3. Building special partitions on $[0,1]$

To illustrate the previous results, let us assume that $\mathcal{X}=[0,1]$ and consider a family $\mathcal{M}=\mathcal{M}'_1\cup\mathcal{M}'_2\cup\mathcal{M}'_3$ of partitions of $\mathcal{X}$ defined in the following way. The family $\mathcal{M}'_1$ is a set of so-called regular partitions of the form $m=\{I_1,\dots,I_{|m|}\}$ with $I_j=[(j-1)/|m|, j/|m|)$ if $j<|m|$ and $I_{|m|}=[(|m|-1)/|m|,1]$ with $1\leqslant|m|\leqslant n$.

To define $\mathcal{M}'_2$ we apply Proposition 13 with $r=1$. In this case, if $m\in\mathcal{M}_j(1)$, we can take for each D_j-dimensional space $\overline{S}_m$ provided by the proposition the linear span of some subset of cardinality D_j of the Haar basis including the function $\mathbb{1}_{[0,1]}$, as follows from [16]. In view of the form of the Haar basis the space $\overline{S}_m$ is a subspace of the linear space of piecewise constant functions based on some partition of $\mathcal{X}$ with a cardinality bounded by $3D_j$ and we shall also denote by m this partition, taking for $\mathcal{M}'_2$ the set of all such partitions when j varies.

To each increasing sequence $\mathcal{J}=\{0=x_0<x_1<\cdots<x_D=1\}$ we associate the partition $m_{\mathcal{J}}=\{I_1,\dots,I_D\}$ with $I_j=[x_{j-1},x_j)$ for $j<D$ and $I_D=[x_{D-1},x_D]$. For each positive integer k, we set $\mathcal{J}_k=\{j2^{-k},\ j=0,\dots,2^k\}$ and, if $2\leqslant D\leqslant 2^k$, we introduce the set $\mathcal{M}_{k,D}$ of all partitions $m_{\mathcal{J}}$ with $|m_{\mathcal{J}}|=D$ and k is the smallest integer such that $\mathcal{J}\subset\mathcal{J}_k$. Finally we set $\mathcal{M}'_3=\bigcup_{k\geqslant1}(\bigcup_{D=2}^{2^k}\mathcal{M}_{k,D})$.

If we choose $\Delta_m=|m|$ for $m\in\mathcal{M}'_1$, then $\sum_{m\in\mathcal{M}'_1}\exp[-\Delta_m]<(\mathrm{e}-1)^{-1}$. If $m\in\mathcal{M}'_2$, then $m\in\mathcal{M}_j(1)$ for some $j\geqslant0$ and we set $\Delta_m=c_3(1)2^j+j$. It follows from (8.21) that $\sum_{m\in\mathcal{M}'_2}\exp[-\Delta_m]<\mathrm{e}/(\mathrm{e}-1)$. For $m\in\mathcal{M}_{k,D}\subset\mathcal{M}'_3$, we choose $\Delta_m=Dk$. Since $|\mathcal{M}_{k,D}|\leqslant\binom{2^k-1}{D-1}<2^{k(D-1)}$,

$$\sum_{m\in\mathcal{M}'_3}\exp[-\Delta_m]=\sum_{k\geqslant1}\sum_{D=2}^{2^k}|\mathcal{M}_{k,D}|\,\mathrm{e}^{-Dk}<\sum_{k\geqslant1}2^{-k}\sum_{D\geqslant2}\exp\big[-Dk(1-\log2)\big]<5/4.$$

Putting those three bounds together, we conclude that $\Sigma<2\mathrm{e}$, as required.

9.2.4. Approximation properties of the partitions and risk bounds

An immediate application of (9.8) and (9.10) shows that there exist partition selection procedures $\widehat{m}$ such that, for all $\overline{P}_s$ in the i.i.d. setting,

$$\mathbb{E}_s\big[v(\overline{P}_s,\hat{s}_{\widehat{m}}\cdot\mu)\big]\leqslant C\inf_{k\geqslant1}\inf_{2\leqslant D\leqslant2^k}\inf_{m\in\mathcal{M}_{k,D}}\big\{v(\overline{P}_s,\bar{s}_m\cdot\mu)+\sqrt{Dk/n}\big\},$$

and for all $s\in M$ in the bounded regression setting,

$$\mathbb{E}_s\big[\|\hat{s}-s\|_2^2\big]\leqslant C\inf_{k\geqslant1}\inf_{2\leqslant D\leqslant2^k}\inf_{m\in\mathcal{M}_{k,D}}\big\{\|\bar{s}_m-s\|_2^2+Dk/n\big\}.$$

As compared to the performance of the histogram estimator based on a single partition with D pieces belonging to $\mathcal{M}_{k,D}$, we loose at most a factor $\sqrt{k}$ or k.

The previous results are valid for all s but they can be improved if s has some regularity properties, due to the inclusion of $\mathcal{M}'_1$ and $\mathcal{M}'_2$ in $\mathcal{M}$. Let us now focus on the bounded regression setting, extensions to the i.i.d. setting with the distance v being more or less straightforward.

The partitions in $\mathcal{M}'_1$ are especially directed to the estimation of continuous functions, i.e. of functions with some modulus of continuity. Given such a modulus of continuity, i.e. a continuous non-decreasing function ω with $\omega(0)=0$ and $\omega(1)\leqslant1$ (see additional details in DeVore and Lorentz [27]), we denote by $\mathcal{S}_\omega$ the subset of M of those functions t such that $|t(x+y)-t(x)|\leqslant\omega(y)$ for all $x\in[0,1]$ and $0\leqslant y\leqslant1-x$. Then $x\omega^2(x)$ is increasing from 0 to $\omega^2(1)$ and, if μ is the Lebesgue measure, it follows from classical lower bounds arguments based on Assouad's Lemma (see [1,11] or [74]) that, for an n-sample and μ the uniform distribution, the minimax risk $R(\mathcal{S}_\omega,2)$ is bounded from below by

$$R(\mathcal{S}_\omega,2)\geqslant c(n\alpha_n)^{-1}\quad\text{with}\quad\begin{cases}\alpha_n\omega^2(\alpha_n)=n^{-1} & \text{if } \omega(1)\geqslant n^{-1/2};\\ \alpha_n=1 & \text{otherwise,}\end{cases} \tag{9.11}$$

c being some universal constant (for similar results in the context of density estimation with Hellinger loss, see [8], p. 210 or [5], Section 4.1.2). Since it is immediate to check that, for $m \in \mathcal{M}'_1$, $d_\infty(s, \bar{s}_m) \leqslant \omega(|m|^{-1})$ if $s \in \mathcal{S}_\omega$, we get from (9.10), independently of μ,

$$\mathbb{E}_s\big[\|\hat{s} - s\|_2^2\big] \leqslant C \inf_{m \in \mathcal{M}'_1} \big\{\omega^2\big(|m|^{-1}\big) + 2|m|/n\big\} \quad \text{for all } s \in \mathcal{S}_\omega.$$

Therefore, given $s \in \mathcal{S}_\omega$ and α_n defined by (9.11), which implies that $1 \geqslant \alpha_n \geqslant n^{-1}$, we can fix $m \in \mathcal{M}'_1$ with $|m| = \lceil \alpha_n^{-1} \rceil$ and get

$$\omega\big(|m|^{-1}\big) \leqslant \omega(\alpha_n) \leqslant (n\alpha_n)^{-1/2} \leqslant \sqrt{|m|/n} \quad \text{and} \quad |m| \leqslant 2\alpha_n^{-1}.$$

Then, whatever the distribution μ and $s \in \mathcal{S}_\omega$,

$$\mathbb{E}_s\big[\|\hat{s} - s\|_2^2\big] \leqslant 3Cn^{-1}|m| \leqslant 6C(n\alpha_n)^{-1},$$

which, up to a fixed constant, coincides with the lower bound (9.11) that we got in the case of a uniformly distributed design.

The partitions in $\mathcal{M}'_2$ are suitable for approximating functions belonging to some Besov spaces. We follow here the same path we used for the proof of Theorem 7. It follows from (8.22) with $q = +\infty$ that, if s is continuous and $s \in B^\alpha_{p,\infty}([0,1])$, $1 > \alpha > p^{-1}$, with $|s|_{B^\alpha_{p,\infty}} \leqslant R$, one can find some $\bar{s}_m$ which is a piecewise constant function build on a partition with less than $3D_j$ pieces and such that

$$\|s - \bar{s}_m\|_2 \leqslant \|s - \bar{s}_m\|_\infty \leqslant C(\alpha, p)|s|_{B^\alpha_{p,\infty}} 2^{-j\alpha}.$$

Using the bound (8.21) for D_j and optimizing with respect to j leads to the final bound

$$\mathbb{E}_s\big[\|\hat{s} - s\|_2^2\big] \leqslant C R^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \quad \text{for } R \geqslant n^{-1/2}. \tag{9.12}$$

If the unknown measure μ has a bounded density with respect to Lebesgue measure, so that $\mu([x,y]) \leqslant A(y-x)$ for $0 \leqslant x \leqslant y \leqslant 1$, then $\|s - \bar{s}_m\|_2^2 \leqslant A\int_0^1 [s(z) - \bar{s}_m(z)]^2\,\mathrm{d}z$ and the same argument based on (8.22) with $q = 2$ shows that (9.12) indeed holds for $1 > \alpha > p^{-1} - (1/2)$ and without the continuity assumption.

9.3. Convex aggregation

For simplicity, we restrict our study of convex aggregation to the i.i.d. setting with the variation distance for which the application of our method is almost straightforward. Let $\{t_1, \ldots, t_N\}$ be a finite subset of $\overline{M}$ (typically preliminary estimators). We would like to find the best convex combination of those points to estimate the distribution of the observations. Let us therefore choose for $\mathcal{M}$ the set of all non-void subsets m of $\{1, \ldots, N\}$ and, when $m = \{k_1, \ldots, k_{|m|}\}$, take for $\overline{S}_m$ the convex envelope of the t_k with $k \in m$, i.e.

$$\overline{S}_m = \left\{ \sum_{j=1}^{|m|} \lambda_j t_{k_j} \text{ with } \lambda_j \geqslant 0 \text{ for } 1 \leqslant j \leqslant |m| \text{ and } \sum_{j=1}^{|m|} \lambda_j = 1 \right\}. \tag{9.13}$$

Considering $\overline{M}$ as embedded into the normed linear space M' of finite signed measures, we can view $\overline{S}_m$ as a subset of an $|m|$-dimensional linear subspace of M' and it follows from Proposition 8 that it has a finite inner metric dimension bounded by $5|m|/3$. Hence, given $\eta_m > 0$, one can find a subset S_m of $\overline{S}_m$ which is an η_m-net for $\overline{S}_m$ and a D-model with parameters η_m, $5|m|/3$ and 1, and Assumption 3 is satisfied. Now observe that the number of elements of $\mathcal{M}$ with cardinality j is $\binom{N}{j} < (\mathrm{e}N/j)^j$. If we set

$$\eta_m^2 = 168|m|n^{-1}\big[(1/3) \vee \big\{1 + \log\big(N/|m|\big) + |m|^{-1}(\log N - 13.8)\big\}\big], \tag{9.14}$$

we can check that (8.12) holds with $\alpha = 2$ and $\Sigma < \mathrm{e}^{13.8} < 10^6$ and apply Corollary 6 with $d = v$ and $B' = 1$. Since $v(s, S_m) \leqslant v(s, \overline{S}_m) + \eta_m$, this proves

Theorem 10. *Let $t_1, \ldots, t_N$ be N given elements of the set $\overline{M}$ of all distributions on $\mathcal{X}$ and $X_1, \ldots, X_n$ be an n-sample from some unknown distribution P_s in $\overline{M}$. For m an arbitrary non-void subset of $\{1, \ldots, N\}$, let $\overline{S}_m$ denote*

the convex envelope of the t_k with $k \in m$ as defined by (9.13) *and η_m be given by* (9.14)*. One can build a T-estimator $\hat{s}(X_1, \ldots, X_n)$ such that, whatever $\overline{P}_s$,*

$$\mathbb{E}_s\left[v^q(s,\hat{s})\right] \leqslant 1.1(5^q) \inf_{m\in\mathcal{M}} \left\{v(s,\overline{S}_m) + \eta_m\right\}^q \quad \textit{for } 1 \leqslant q \leqslant 79.$$

In particular,

$$\mathbb{E}_s\left[v^q(s,\hat{s})\right] \leqslant C(q) \inf_{m\in\mathcal{M}} \left\{v(s,\overline{S}_m) + \sqrt{|m|n^{-1}\left[1+\log\left(N/|m|\right)\right]}\right\}^q.$$

It is worthwhile noticing that $\hat{s}$ simultaneously performs what is usually called "convex aggregation" (which corresponds to $|m| = N$) and estimator selection (which corresponds to $|m| = 1$), but also convex aggregation over proper subsets of $\{t_1, \ldots, t_N\}$.

Lower bounds for the risk of selection or convex aggregation in the case of random design regression have been obtained by Tsybakov in [56] and Yang in [70]. Although we deal with a different situation (density estimation with $\mathbb{L}_1$-loss), we may extrapolate them here for the sake of comparison. If the extrapolation is correct, we see that up to constants, our bound coincides with the lower bounds for selection and for convex aggregation with $N \leqslant \sqrt{n}$.

Acknowledgements

I would like to thank the participants of the Colloquium in Honour of Jean Bretagnolle, Didier Dacunha-Castelle and Ildar Ibragimov who, after the talk I gave in June 2001 on this topic, asked many questions which strongly pushed me to pursue in this direction. Special thanks also to Yannick Baraud, Albert Cohen, Fabienne Comte, Ron DeVore, Vladimir Koltchinskii, Pascal Massart, David Pollard, Vladimir Temlyakov and Sacha Tsybakov for encouragements, suggestions, comments and stimulating discussions. An anonymous referee's questions and suggestions greatly helped to improve the first version of this paper.

Appendix A

A.1. Proof of Proposition 2

Let $c = k^{1/4}$. Since $k \geqslant 128$,

$$c \geqslant 3.36 \quad \text{and} \quad \sqrt{3k} - \sqrt{k} - 4 - 1.21k^{1/4} > 0. \tag{A.1}$$

It follows that, whatever the true value of $s \in \mathcal{S}$, $\mathbb{P}_s[|X_0| \geqslant c + 1.21] < 0.114$ and that $\|\boldsymbol{X}'\|^2$ has a non-central $\chi^2(k)$ distribution. Since a non-central χ^2 variable is stochastically larger than a central one, it follows from Lemma 1 of Laurent and Massart [40] that

$$\mathbb{P}_s\left[\|\boldsymbol{X}'\|^2 \leqslant k - 2\sqrt{kx}\right] \leqslant \mathrm{e}^{-x} \quad \text{for } x > 0. \tag{A.2}$$

Setting $x = k/64 \geqslant 2$, we conclude, since $e^{-x} < 0.136$, that

$$\mathbb{P}_s[\Omega'] > 3/4 \quad \text{with } \Omega' = \left\{\|\boldsymbol{X}'\|^2 > 3k/4 \quad \text{and} \quad |X_0| < c + 1.21\right\}.$$

Now assume that the event Ω' holds. Since the m.l.e. $\hat{s}$ is the least squares estimator, $\hat{s}$ is the minimizer over $\mathcal{S}$ of $(X_0 - s_0)^2 + \|\boldsymbol{X}' - s'\|^2$. On Ω', $\|\boldsymbol{X}'\| > \sqrt{3k}/2 > \|s'\|$ and, given s_0, the minimum with respect to s' is obtained for $s' = 2\boldsymbol{X}'(1 - |s_0|/c)/\|\boldsymbol{X}'\|$ with value

$$f(s_0) = (X_0 - s_0)^2 + \left[\|\boldsymbol{X}'\| - 2\left(1 - |s_0|/c\right)\right]^2.$$

Since for $s_0 \neq 0$,

$$\begin{aligned}(c/2)s_0 f'(s_0) &= c(s_0 - X_0)s_0 + 2|s_0|\left[\|\boldsymbol{X}'\| - 2\left(1 - |s_0|/c\right)\right] > |s_0|\left[2\|\boldsymbol{X}'\| - 4 - c(c+1.21)\right] \\ &\geqslant |s_0|\left[\sqrt{3k} - 4 - \sqrt{k} - 1.21k^{1/4}\right]\end{aligned}$$

is non-negative by (A.1), $f(s_0)$ is minimal when $s_0 = 0$. Therefore, if Ω' holds, $\hat{s}_0 = 0$ and $\hat{s}' = 2\boldsymbol{X}'/\|\boldsymbol{X}'\|$. This implies that the quadratic risk at $s = (s_0, 0)$ of the m.l.e. is bounded from below by $(3/4)(s_0^2 + 4)$ with maximum

value $(3/4)\sqrt{k}+3$ when $|s_0|=c$. On the other hand, the estimator $\tilde{s}$ with $\tilde{s}_0=X_0$ and $\tilde{s}'=0$ has a quadratic risk which is uniformly bounded by 5.

A.1. Proof of Proposition 3

Observe that (5.3) implies, if $z \geqslant \bar{y}+w$, that

$$\mathbb{P}[Y+w>z]=\mathbb{P}[Y>z-w] \leqslant \alpha \exp\bigl[-\beta(z-w)^2\bigr]=\alpha \exp\bigl[-\beta(1-z^{-1}w)^2 z^2\bigr].$$

Therefore, using $1-w/z \geqslant 1-w/(\bar{y}+w)=\bar{y}/(\bar{y}+w)$, we get

$$\mathbb{P}[Y+w>z] \leqslant \alpha \exp\bigl[-\beta\bigl[\bar{y}/(\bar{y}+w)\bigr]^2 z^2\bigr],$$

which is the analogue of (5.3) with $Y+w$ replacing Y, $\bar{y}+w$ replacing $\bar{y}$ and $\beta[\bar{y}/(\bar{y}+w)]^2$ replacing β. Since $\beta[\bar{y}/(\bar{y}+w)]^2(\bar{y}+w)^2=\beta\bar{y}^2$, it suffices to prove (5.4) for $w=0$ and then make the relevant parameter changes. From (5.3), the change of variable $x=\beta y^{2/q}$ and Stirling's bound, $\Gamma(x+1)<\sqrt{\pi\,\mathrm{e}x}(x/\mathrm{e})^x$ valid for $x \geqslant 1/2$, we get

$$\begin{aligned}\mathbb{E}[Y^q]-\bar{y}^q &= \left(\int_0^\infty \mathbb{P}[Y^q \geqslant y]\,\mathrm{d}y - \bar{y}^q\right) \leqslant \int_{\bar{y}^q}^\infty \mathbb{P}[Y \geqslant y^{1/q}]\,\mathrm{d}y \\ &\leqslant \alpha \int_{\bar{y}^q}^\infty \exp(-\beta y^{2/q})\,\mathrm{d}y = \alpha\beta^{-q/2}\frac{q}{2}\int_{\beta\bar{y}^2}^\infty x^{q/2-1}\,\mathrm{e}^{-x}\,\mathrm{d}x \qquad (A.3)\\ &< \alpha\beta^{-q/2}\Gamma\left(\frac{q}{2}+1\right) \leqslant \alpha\beta^{-q/2}\sqrt{\frac{\pi\,\mathrm{e}q}{2}}\left(\frac{q}{2\,\mathrm{e}}\right)^{q/2}. \qquad (A.4)\end{aligned}$$

Alternatively, when $\beta\bar{y}^2 \geqslant q/2$, we can use the bound (Inequality 45 from Johnstone [36]) $\int_z^{+\infty} x^t\,\mathrm{e}^{-x}\,\mathrm{d}x < (z^{t+1}\,\mathrm{e}^{-z})(z-t)^{-1}$ for $z>t$, which gives $\int_{\beta\bar{y}^2}^\infty x^{q/2-1}\,\mathrm{e}^{-x}\,\mathrm{d}x < (\beta\bar{y}^2)^{q/2}\exp(-\beta\bar{y}^2)$ so that finally

$$\frac{\mathbb{E}[Y^q]-\bar{y}^q}{\alpha\bar{y}^q} \leqslant \begin{cases}\sqrt{\pi\,\mathrm{e}q/2}\bigl[q/(2\,\mathrm{e}\beta\bar{y}^2)\bigr]^{q/2} & \text{for all } \bar{y}>0,\\ (q/2)\exp(-\beta\bar{y}^2) & \text{if } \beta\bar{y}^2 \geqslant q/2,\end{cases}$$

which proves (5.4) for $w=0$. Both functions $x \mapsto (2\,\mathrm{e}x/q)^{-q/2}$ and $x \mapsto \mathrm{e}^{-x}$ are decreasing on $(0,+\infty)$ and they coincide for $x=q/2$ which implies that ζ_q is decreasing for $1 \leqslant q \leqslant 2\pi\,\mathrm{e}$. The choice of $c=0.612>1/2$ for $q>2\pi\,\mathrm{e}$ ensures that $\zeta_q(cq)<\zeta_q((cq)_-)$ so that ζ_q is still decreasing for all those values of q.

A.2. Bounding the errors of tests

All the results about tests that we use in this paper are based on the following easy but important lemma.

Lemma 7. *Let $X_1,\dots,X_n$ be n random variables on some measurable space $\mathcal{X}$, which, under both probabilities $\mathbb{P}$ or $\mathbb{Q}$, are independent, and let ϕ be a non-negative measurable function on $\mathcal{X}$ such that*

$$\mathbb{E}_{\mathbb{P}}\bigl[\sqrt{\phi(X_i)}\bigr] \leqslant \alpha \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}\bigl[1/\sqrt{\phi(X_i)}\bigr] \leqslant \beta \quad \text{for } 1 \leqslant i \leqslant n.$$

Then, for all $y \in \mathbb{R}$,

$$\mathbb{P}\left[\sum_{i=1}^n \log\phi(X_i) \geqslant ny\right] \leqslant \exp\bigl[n(\log\alpha - y)\bigr] \quad \text{and} \quad \mathbb{Q}\left[\sum_{i=1}^n \log\phi(X_i) \leqslant ny\right] \leqslant \exp\bigl[n(\log\beta + y)\bigr].$$

In particular, if the X_i are i.i.d. with distribution $\overline{P}$ under $\mathbb{P}$ and $\overline{Q}$ under $\mathbb{Q}$, then, for all $x \in \mathbb{R}$,

$$\mathbb{P}\left[\sum_{i=1}^n \log\left(\frac{\mathrm{d}\overline{Q}}{\mathrm{d}\overline{P}}\right)(X_i) \geqslant nx\right] \leqslant \exp\bigl[n\log\bigl[\rho(\overline{P},\overline{Q})\bigr]-(nx/2)\bigr] \qquad (A.5)$$

and

$$\mathbb{Q}\left[\sum_{i=1}^{n}\log\left(\frac{\mathrm{d}\overline{Q}}{\mathrm{d}\overline{P}}\right)(X_i)\leqslant nx\right]\leqslant\exp\bigl[n\log\bigl[\rho(\overline{P},\overline{Q})\bigr]+(nx/2)\bigr]. \tag{A.6}$$

Proof. It immediately follows from the elementary inequality

$$\mathbb{P}[\log Y\geqslant z]\leqslant \mathrm{e}^{-z}\mathbb{E}[Y],\quad \text{if } \mathbb{P}[Y\geqslant 0]=1 \tag{A.7}$$

and the independence of the X_i with an application to $\phi=\sqrt{\mathrm{d}\overline{Q}/\mathrm{d}\overline{P}}$. □

Proof of Proposition 4. We get from (A.7) and (5.17)

$$\begin{aligned}\mathbb{P}_s\bigl[\log(\mathrm{d}P_u/\mathrm{d}P_t)(\boldsymbol{X})\geqslant z\bigr]&\leqslant \mathrm{e}^{-z/2}\mathbb{E}_s\bigl[\exp\bigl[(1/2)\log(\mathrm{d}P_u/\mathrm{d}P_t)(\boldsymbol{X})\bigr]\bigr]\\&=\mathrm{e}^{-z/2}\mathbb{E}_0\Bigl[\sqrt{(\mathrm{d}P_u/\mathrm{d}P_t)(\boldsymbol{X})}(\mathrm{d}P_s/\mathrm{d}P_0)(\boldsymbol{X})\Bigr]\\&=\mathrm{e}^{-z/2}\mathbb{E}_0\left[\exp\left[-\frac{1}{2\sigma^2}\left(\frac{\|u\|^2-\|t\|^2}{2}+\|s\|^2-\langle\boldsymbol{X},u-t+2s\rangle\right)\right]\right].\end{aligned}$$

Since

$$\begin{aligned}&\frac{\|u\|^2-\|t\|^2}{2}+\|s\|^2-\langle\boldsymbol{X},u-t+2s\rangle\\&\quad=\left\|\frac{u-t}{2}+s\right\|^2-2\left\langle\boldsymbol{X},\frac{u-t}{2}+s\right\rangle+\frac{\|u\|^2-3\|t\|^2}{4}-\langle s,u-t\rangle+\frac{\langle u,t\rangle}{2},\end{aligned}$$

we get

$$\begin{aligned}\mathbb{P}_s\bigl[\log(\mathrm{d}P_u/\mathrm{d}P_t)(\boldsymbol{X})\geqslant z\bigr]&\leqslant \mathrm{e}^{-z/2}\mathbb{E}_0\bigl[p_{(u-t)/2+s}(\boldsymbol{X})\bigr]\exp\left[-\frac{1}{2\sigma^2}\left(\frac{\|u\|^2-3\|t\|^2}{4}-\langle s,u-t\rangle+\frac{\langle u,t\rangle}{2}\right)\right]\\&=\exp\left[-\frac{z}{2}-\frac{\|u\|^2-3\|t\|^2-4\langle s,u-t\rangle+2\langle u,t\rangle}{8\sigma^2}\right],\end{aligned}$$

the conclusion follows from the fact that

$$-\|u\|^2+3\|t\|^2+4\langle s,u-t\rangle-2\langle u,t\rangle=-\|t-u\|^2+4\langle s-t,u-t\rangle\leqslant-\|t-u\|\bigl(\|t-u\|-4\|s-t\|\bigr).$$

Proof of Proposition 5. It follows the ideas of the proof of Theorem 5 in [15]. Let us start with the random design case. Setting $Z_i=[Y_i-t(X_i)]^2-[Y_i-u(X_i)]^2$, we get the following decomposition:

$$\begin{aligned}Z_i&=\bigl[u(X_i)-t(X_i)\bigr]\bigl[2Y_i-t(X_i)-u(X_i)\bigr]\\&=\bigl[u(X_i)-t(X_i)\bigr]\bigl[2\bigl(s(X_i)-t(X_i)\bigr)+\bigl(t(X_i)-u(X_i)\bigr)+2\varepsilon_i\bigr]\\&=-\bigl[t(X_i)-u(X_i)\bigr]^2+2\bigl[u(X_i)-t(X_i)\bigr]\bigl[s(X_i)-t(X_i)\bigr]+2\varepsilon_i\bigl[u(X_i)-t(X_i)\bigr].\end{aligned} \tag{A.8}$$

Consequently, since $\mathbb{E}_s[\varepsilon_i|X_i]=0$ and $2|ab|\leqslant a^2/4+4b^2$,

$$\begin{aligned}\mathbb{E}_s[Z_i|X_i]&=-\bigl[t(X_i)-u(X_i)\bigr]^2+2\bigl[u(X_i)-t(X_i)\bigr]\bigl[s(X_i)-t(X_i)\bigr]\\&\leqslant-3\bigl[t(X_i)-u(X_i)\bigr]^2/4+4\bigl[s(X_i)-t(X_i)\bigr]^2,\end{aligned} \tag{A.9}$$

and finally, setting $y=-(1/4)\|t-u\|^2+4\|s-t\|^2$,

$$n^{-1}\mathbb{E}_s\left[\sum_{i=1}^{n}Z_i\right]\leqslant-\frac{3}{4}\|t-u\|^2+4\|s-t\|^2=y-\frac{1}{2}\|t-u\|^2. \tag{A.10}$$

It follows from its definition that $|Z_i| \leqslant 1$ and (A.8) implies that $|Z_i| \leqslant 2|t(X_i) - u(X_i)|$, hence, for any integer $k \geqslant 2$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s\big[|Z_i|^k\big] \leqslant \big(2\|t-u\|\big)^2 \leqslant (k!/2)\big(2\|t-u\|\big)^2 3^{2-k}.$$

It then follows from Bernstein's Inequality, as stated in [15], Lemma 8, that, for all $x \geqslant 0$,

$$\mathbb{P}_s\left[\sum_{i=1}^{n}\big(Z_i - \mathbb{E}[Z_i]\big) \geqslant nx\right] \leqslant \exp\left[\frac{-nx^2/2}{(2\|t-u\|)^2 + x/3}\right],$$

hence, by (A.10),

$$\mathbb{P}_s\left[\sum_{i=1}^{n} Z_i \geqslant n\big(x - \|t-u\|^2/2 + y\big)\right] \leqslant \exp\left[\frac{-3n}{2}\frac{x^2}{12\|t-u\|^2 + x}\right] \quad \text{for } x \geqslant 0.$$

Setting $x = z - y + \|t-u\|^2/2$, we derive, for $z \geqslant y - \|t-u\|^2/2$, that

$$\mathbb{P}_s\left[\sum_{i=1}^{n} Z_i \geqslant nz\right] \leqslant \exp\left[\frac{-3n}{2}\frac{(z + \|t-u\|^2/2 - y)^2}{12\|t-u\|^2 + (z + \|t-u\|^2/2 - y)}\right] = \exp\left[\frac{-3n}{4}\frac{[2(z-y) + \|t-u\|^2]^2}{2(z-y) + 25\|t-u\|^2}\right].$$

Since

$$\frac{[2(z-y) + \|t-u\|^2]^2}{2(z-y) + 25\|t-u\|^2} \geqslant \frac{1}{25}\left(\|t-u\|^2 + \frac{98(z-y)}{25}\right)$$

provided that the denominator is positive, which is the case here, we finally get

$$\mathbb{P}_s\left[\sum_{i=1}^{n} Z_i \geqslant nz\right] \leqslant \exp\left[\frac{-3n}{100}\left(\|t-u\|^2 + \frac{98(z-y)}{25}\right)\right] \quad \text{for } z \geqslant y - \frac{\|t-u\|^2}{2}.$$

Since this inequality also holds trivially when $z < y - \|t-u\|^2/2$, we have proved (5.21) from which we derive, since $y \leqslant 4\|s-t\|^2$, that

$$\mathbb{P}_s\big[\gamma'(t, \boldsymbol{X}) - \gamma'(u, \boldsymbol{X}) \geqslant nz\big] \leqslant \exp\left[\frac{3n}{100}\frac{98(4\|s-t\|^2 - z)}{25}\right] \quad \text{for all } z \in \mathbb{R},$$

and (5.23) follows. If $\|s-t\| \leqslant \|t-u\|/4$, then $y \leqslant 0$ and (5.21) implies that

$$\mathbb{P}_s\big[\gamma'(t, \boldsymbol{X}) - \gamma'(u, \boldsymbol{X}) \geqslant nz\big] \leqslant \exp\left[\frac{-3n}{100}\left(\|t-u\|^2 + \frac{98z}{25}\right)\right], \quad \text{for all } z \in \mathbb{R},$$

which gives (5.22) and concludes the proof for the random design. The proof for the fixed design case is identical: just replace X_i by x_i, $\mathbb{E}_s[\varepsilon_i|X_i]$ by $\mathbb{E}_s[\varepsilon_i]$ and $\mathbb{E}_s[Z_i|X_i]$ by $\mathbb{E}_s[Z_i]$ in (A.9).

Proof of Proposition 6. If $d = v$, let us consider in the metric space $(\overline{M}, v)$ of distributions on $\mathcal{X}$ the two closed balls $\mathcal{B}(t)$ and $\mathcal{B}(u)$ with respective centers t and u and radius $v(t,u)/4$. It follows from Section 7 of Huber [34] that there exists a least favorable pair (t_0, u_0), with $t_0 \in \mathcal{B}(t)$ and $u_0 \in \mathcal{B}(u)$, for testing between those balls, which means that, for all $z \in \mathbb{R}$,

$$\mathbb{P}_v\left[\log\left(\frac{\mathrm{d}\overline{P}_{u_0}}{\mathrm{d}\overline{P}_{t_0}}(X)\right) \leqslant z\right] \geqslant \mathbb{P}_{t_0}\left[\log\left(\frac{\mathrm{d}\overline{P}_{u_0}}{\mathrm{d}\overline{P}_{t_0}}(X)\right) \leqslant z\right] \quad \text{for all } v \in \mathcal{B}(t)$$

and

$$\mathbb{P}_v\left[\log\left(\frac{\mathrm{d}\overline{P}_{u_0}}{\mathrm{d}\overline{P}_{t_0}}(X)\right) \leqslant z\right] \leqslant \mathbb{P}_{u_0}\left[\log\left(\frac{\mathrm{d}\overline{P}_{u_0}}{\mathrm{d}\overline{P}_{t_0}}(X)\right) \leqslant z\right] \quad \text{for all } v \in \mathcal{B}(u).$$

Let then $\psi(t_0, u_0, \boldsymbol{X}) = 1 - \psi(u_0, t_0, \boldsymbol{X})$ be any likelihood ratio test between $\overline{P}_{t_0}$ and $\overline{P}_{u_0}$ of the form

$$\psi(t_0, u_0, \boldsymbol{X}) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} \log\big[(\mathrm{d}\overline{P}_{u_0}/\mathrm{d}\overline{P}_{t_0})(X_i)\big] < z, \\ 1 & \text{if } \sum_{i=1}^{n} \log\big[(\mathrm{d}\overline{P}_{u_0}/\mathrm{d}\overline{P}_{t_0})(X_i)\big] > z. \end{cases}$$

If $P_{t_0} = \overline{P}_{t_0}^{\otimes n}$ and $P_{u_0} = \overline{P}_{u_0}^{\otimes n}$, the classical properties of stochastic ordering imply that

$$\mathbb{P}_s\big[\psi(t_0, u_0, \boldsymbol{X}) = 1\big] \leqslant \mathbb{P}_{t_0}\big[\psi(t_0, u_0, \boldsymbol{X}) = 1\big] \quad \text{if } \bar{v}(s,t) \leqslant v(t,u)/4$$

and similarly,

$$\mathbb{P}_s\big[\psi(u_0, t_0, \boldsymbol{X}) = 1\big] \leqslant \mathbb{P}_{u_0}\big[\psi(u_0, t_0, \boldsymbol{X}) = 1\big] \quad \text{if } \bar{v}(s,u) \leqslant v(t,u)/4.$$

Since (A.5), (A.6) and (2.1) imply that

$$\mathbb{P}_{t_0}\big[\psi(t_0, u_0, \boldsymbol{X}) = 1\big] \leqslant \exp\big[-nh^2(\overline{P}_{u_0}, \overline{P}_{t_0}) - z/2\big]$$

and

$$\mathbb{P}_{u_0}\big[\psi(u_0, t_0, \boldsymbol{X}) = 1\big] \leqslant \exp\big[-nh^2(\overline{P}_{u_0}, \overline{P}_{t_0}) + z/2\big],$$

and, by (5.15),

$$h^2(\overline{P}_{u_0}, \overline{P}_{t_0}) \geqslant v^2(\overline{P}_{u_0}, \overline{P}_{t_0})/2 \geqslant \big[v(t,u)/2\big]^2/2,$$

the conclusion follows for the variation distance if we set $z = nx/4$.

If $d = h$, we consider, in the metric space $(\overline{M}, h)$ of distributions on $\mathcal{X}$, the two closed balls $\mathcal{B}(t)$ and $\mathcal{B}(u)$ with respective centers t and u and radius $h(t,u)/4$. It follows from Theorem 1, p. 485 of Le Cam [45] that one can find a non-negative measurable function ϕ on $\mathcal{X}$, such that

$$\int \phi \, \mathrm{d}\overline{P}_v \leqslant 1 - h^2(t,u)/4 \quad \text{if } v \in \mathcal{B}(t); \qquad \int (1/\phi) \, \mathrm{d}\overline{P}_v \leqslant 1 - h^2(t,u)/4 \quad \text{if } v \in \mathcal{B}(u).$$

The conclusion then follows from Lemma 7 with $y = x/4$ and

$$\psi(t, u, \boldsymbol{X}) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \log\big[\phi(X_i)\big] < nx/4, \\ 1 & \text{if } \sum_{i=1}^n \log\big[\phi(X_i)\big] > nx/4. \end{cases}$$

References

[1] P. Assouad, Deux remarques sur l'estimation, C. R. Acad. Sci. Paris, Sér. I Math. 296 (1983) 1021–1024.
[2] J.-Y. Audibert, Théorie statistique de l'apprentissage : une approche PAC-bayésienne, Thèse de doctorat, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, Paris, 2004.
[3] Y. Baraud, Model selection for regression on a random design, ESAIM Probab. Statist. 6 (2002) 127–146.
[4] A.R. Barron, Complexity regularization with applications to artificial neural networks, in: G. Roussas (Ed.), Nonparametric Functional Estimation, Kluwer, Dordrecht, 1991, pp. 561–576.
[5] A.R. Barron, L. Birgé, P. Massart, Risk bounds for model selection via penalization, Probab. Theory Related Fields 113 (1999) 301–415.
[6] A.R. Barron, T.M. Cover, Minimum complexity density estimation, IEEE Trans. Inform. Theory 37 (1991) 1034–1054.
[7] J. Beirlant, L. Györfi, On the asymptotic normality of the L_2-error in partitioning regression estimation, J. Statist. Plann. Inference 71 (1998) 93–107.
[8] L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 65 (1983) 181–237.
[9] L. Birgé, Sur un théorème de minimax et son application aux tests, Probab. Math. Statist. 3 (1984) 259–282.
[10] L. Birgé, Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées, Ann. Inst. H. Poincaré Sect. B 20 (1984) 201–223.
[11] L. Birgé, On estimating a density using Hellinger distance and some other strange facts, Probab. Theory Related Fields 71 (1986) 271–291.
[12] L. Birgé, Model selection for Gaussian regression with random design, Bernoulli 10 (2004) 1039–1051.
[13] L. Birgé, P. Massart, Rates of convergence for minimum contrast estimators, Probab. Theory Related Fields 97 (1993) 113–150.
[14] L. Birgé, P. Massart, From model selection to adaptive estimation, in: D. Pollard, E. Torgersen, G. Yang (Eds.), Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, Springer-Verlag, New York, 1997, pp. 55–87.
[15] L. Birgé, P. Massart, Minimum contrast estimators on sieves: exponential bounds and rates of convergence, Bernoulli 4 (1998) 329–375.
[16] L. Birgé, P. Massart, An adaptive compression algorithm in Besov spaces, Constr. Approx. 16 (2000) 1–36.
[17] L. Birgé, P. Massart, Gaussian model selection, J. Eur. Math. Soc. 3 (2001) 203–268.
[18] M.S. Birman, M.Z. Solomjak, Piecewise-polynomial approximation of functions of the classes W_p, Mat. Sb. 73 (1967) 295–317.
[19] L.D. Brown, M.G. Low, Asymptotic equivalence of nonparametric regression and white noise, Ann. Statist. 24 (1996) 2384–2398.
[20] F. Bunea, A.B. Tsybakov, M.H. Wegkamp, Aggregation for regression learning, Technical report 948, Laboratoire de Probabilités, Université Paris VI, 2004, http://www.proba.jussieu.fr/mathdoc/preprints/index.html# 2004.

[21] G. Castellan, Modified Akaike's criterion for histogram density estimation, Technical report 99.61, Université Paris-Sud, Orsay, 1999, http://www.math.u-psud.fr/~biblio/pub/1999/.
[22] G. Castellan, Sélection d'histogrammes à l'aide d'un critère de type Akaike, C. R. Acad. Sci. Paris 330 (2000) 729–732.
[23] O. Catoni, The mixture approach to universal model selection, Technical report LMENS-97-22, Ecole Normale Supérieure, Paris, 1997, http://www.dma.ens.fr/edition/publis/1997/titre97.html.
[24] O. Catoni, Statistical learning theory and stochastic optimization, in: J. Picard (Ed.), Lecture on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXXI – 2001, in: Lecture Note in Math., vol. 1851, Springer-Verlag, Berlin, 2004.
[25] H. Chernoff, A measure of asymptotic efficiency of tests of a hypothesis based on a sum of observations, Ann. Math. Statist. 23 (1952) 493–507.
[26] R.A. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov, Mathematical methods for supervised learning, Technical report 0422, IMI, University of South Carolina, Columbia, 2004, http://www.math.sc.edu/imip/preprints/04.html.
[27] R.A. DeVore, G.G. Lorentz, Constructive Approximation, Springer-Verlag, Berlin, 1993.
[28] L. Devroye, G. Lugosi, Combinatorial Methods in Density Estimation, Springer-Verlag, New York, 2001.
[29] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, D. Picard, Density estimation by wavelet thresholding, Ann. Statist. 24 (1996) 508–539.
[30] D.L. Donoho, R.C. Liu, B. MacGibbon, Minimax risk over hyperrectangles, and implications, Ann. Statist. 18 (1990) 1416–1437.
[31] P.P.B. Eggermont, V.N. LaRiccia, Maximum Penalized Likelihood Estimation, vol. I: Density Estimation, Springer, New York, 2001.
[32] P. Groeneboom, Some current developments in density estimation, in: J.W. de Bakker, M. Hazewinkel, J.K. Lenstra (Eds.), Mathematics and Computer Science, in: CWI Monograph, vol. 1, Elsevier, Amsterdam, 1986, pp. 163–192.
[33] L. Györfi, M. Kohler, A. Kryżak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer, New York, 2002.
[34] P.J. Huber, A robust version of the probability ratio test, Ann. Math. Statist. 36 (1965) 1753–1758.
[35] P.J. Huber, Robust Statistics, John Wiley, New York, 1981.
[36] I.M. Johnstone, Chi-square oracle inequalities, in: M.C.M. de Gunst, C.A.J. Klaassen, A.W. van der Vaart (Eds.), State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet, in: Lecture Notes Monograph Ser., vol. 36, Institute of Mathematical Statistics, 2001, pp. 399–418.
[37] A. Juditsky, A.S. Nemirovski, Functional aggregation for nonparametric estimation, Ann. Statist. 28 (2000) 681–712.
[38] G. Kerkyacharian, D. Picard, Thresholding algorithms, maxisets and well-concentrated bases, Test 9 (2000) 283–344.
[39] A.N. Kolmogorov, V.M. Tikhomirov, ε-entropy and ε-capacity of sets in function spaces, Amer. Math. Soc. Transl. (2) 17 (1961) 277–364.
[40] B. Laurent, P. Massart, Adaptive estimation of a quadratic functional by model selection, Ann. Statist. 28 (2000) 1302–1338.
[41] L.M. Le Cam, On the assumptions used to prove asymptotic normality of maximum likelihood estimates, Ann. Math. Statist. 41 (1970) 802–828.
[42] L.M. Le Cam, Limits of experiments, in: Proc. 6th Berkeley Symp. on Math. Stat. and Prob. I, 1972, pp. 245-261.
[43] L.M. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 1 (1973) 38–53.
[44] L.M. Le Cam, On local and global properties in the theory of asymptotic normality of experiments, in: M. Puri (Ed.), Stochastic Processes and Related Topics, vol. 1, Academic Press, New York, 1975, pp. 13–54.
[45] L.M. Le Cam, Asymptotic Methods in Statistical Decision Theory, Springer-Verlag, New York, 1986.
[46] L.M. Le Cam, Maximum likelihood: an introduction, Inter. Statist. Rev. 58 (1990) 153–171.
[47] L.M. Le Cam, Metric dimension and statistical estimation, CRM Proc. and Lecture Notes 11 (1997) 303–311.
[48] G.G. Lorentz, Approximation of Functions, Holt, Rinehart, Winston, New York, 1966.
[49] G.G. Lorentz, M. von Golitschek, Y. Makovoz, Constructive Approximation, Advanced Problems, Springer, Berlin, 1996.
[50] A.S. Nemirovski, Topics in non-parametric statistics, in: P. Bernard (Ed.), Lecture on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXVIII – 1998, in: Lecture Notes in Math., vol. 1738, Springer-Verlag, Berlin, 2000, pp. 85–297.
[51] M. Nussbaum, Asymptotic equivalence of density estimation and Gaussian white noise, Ann. Statist. 24 (1996) 2399–2430.
[52] A. Pinkus, n-widths in Approximation Theory, Springer-Verlag, Berlin, 1985.
[53] M.S. Pinsker, Optimal filtration of square-integrable signals in Gaussian noise, Problems Inform. Transmission 16 (1980) 120–133.
[54] X. Shen, W.H. Wong, Convergence rates of sieve estimates, Ann. Statist. 22 (1994) 580–615.
[55] B.W. Silverman, On the estimation of a probability density function by the maximum penalized likelihood method, Ann. Statist. 10 (1982) 795–810.
[56] A.B. Tsybakov, Optimal rates of aggregation, in: Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines, in: Lecture Notes in Artificial Intelligence, vol. 2777, Springer-Verlag, Berlin, 2003, pp. 303–313.
[57] S. van de Geer, Estimating a regression function, Ann. Statist. 18 (1990) 907–924.
[58] S. van de Geer, Hellinger-consistency of certain nonparametric maximum likelihood estimates, Ann. Statist. 21 (1993) 14–44.
[59] S. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, Cambridge, 2000.
[60] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[61] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.
[62] A. Wald, Note on the consistency of the maximum likelihood estimate, Ann. Math. Statist. 20 (1949) 595–601.
[63] M.H. Wegkamp, Model selection in nonparametric regression, Ann. Statist. 31 (2003) 252–273.
[64] W.H. Wong, X. Shen, Probability inequalities for likelihood ratios and convergence rates of sieve MLEs, Ann. Statist. 23 (1995) 339–362.
[65] Y. Yang, Minimax optimal density estimation, Ph.D. dissertation, Dept. of Statistics, Yale University, New Haven, 1996.
[66] Y. Yang, Mixing strategies for density estimation, Ann. Statist. 28 (2000) 75–87.
[67] Y. Yang, Combining different procedures for adaptive regression, J. Multivariate Anal. 74 (2000) 135–161.
[68] Y. Yang, Adaptive regression by mixing, J. Amer. Statist. Assoc. 96 (2001) 574–588.
[69] Y. Yang, How accurate can any regression procedure be?, Technical report, Iowa State University, Ames, 2001, http://www.public.iastate.edu/yyang/papers/index.html.

[70] Y. Yang, Aggregating regression procedures to improve performance, Bernoulli 10 (2004) 25–47.
[71] Y. Yang, A.R. Barron, An asymptotic property of model selection criteria, IEEE Trans. Inform. Theory 44 (1998) 95–116.
[72] Y. Yang, A.R. Barron, Information-theoretic determination of minimax rates of convergence, Ann. Statist. 27 (1999) 1564–1599.
[73] Y.G. Yatracos, Rates of convergence of minimum distance estimates and Kolmogorov's entropy, Ann. Statist. 13 (1985) 768–774.
[74] B. Yu, Assouad, Fano and Le Cam, in: D. Pollard, E. Torgersen, G. Yang (Eds.), Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, Springer-Verlag, New York, 1997, pp. 423–435.