Universités de Paris 6 & Paris 7 - CNRS (UMR 7599)

PRÉPUBLICATIONS DU LABORATOIRE

DE PROBABILITÉS & MODÈLES ALÉATOIRES

4, place Jussieu - Case 188 - 75 252 Paris cedex 05

http://www.proba.jussieu.fr

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

L. BIRGÉ

NOVEMBRE 2003

Prépublication n^o 862

Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris VI & Université Paris VII,
4, place Jussieu, Case 188, F-75252 Paris Cedex 05.

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

Lucien Birgé

Université Paris VI
Laboratoire de Probabilités et Modèles Aléatoires – C.N.R.S. (UMR 7599)

27/10/2003

Abstract

This paper is devoted to the description and study of a family of estimators, that we shall call T-estimators (T for tests), for minimax estimation and model selection. Their construction is based on former ideas about deriving estimators from some families of tests due to Le Cam (1973 and 1975) and Birgé (1983, 1984a and b) and about complexity based model selection from Barron and Cover (1991).

It is well-known that maximum likelihood estimators or, 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 for the estimators to work correctly. Our method, which derives an estimator from many simultaneous tests between some probability balls in a suitable metric space, tends to solve many of these difficulties. Its robustness properties allow to deal with minimax estimation and model selection in a unified way, since bounding the minimax risk amounts to perform the method with a single, well-chosen, model. This results 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 (we shall concentrate here on the i.i.d. and the Gaussian sequence settings only) and can handle essentially all types of models, linear or not, parametric and non-parametric, simultaneously. From these viewpoints, it is much more flexible than traditional methods. In particular, we shall be able to derive some adaptation results for density estimation over Besov balls that do not seem to be accessible to classical methods. The counterpart for these nice properties is the very high computational complexity of our construction which makes our estimators look more like theoretical than practical tools.

[0] *AMS 1991 subject classifications.* Primary 62F35; secondary 62G05.

Key words and phrases. 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 restrictive, assumptions but may behave in a terrible way otherwise. This fact has been recognized for a long time; examples about the m.l.e. and further references can be found in Le Cam (1990).

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 in Section 2.1.3 below.

After some years of study of minimum contrast estimators, the present author became convinced of the need for a more flexible and less demanding alternative method for estimation and model selection. He 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. He 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) van de Geer (1993 and 2000), Birgé and Massart (1997), Barron, Birgé and Massart (1999), Castellan (1999 and 2000) or Wegkamp (2003). 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 (2002) or be uniformly bounded as in Yang (2000).

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 and homoscedastic Gaussian sequences that we shall consider in this paper and to Gaussian regression with random design which has been studied in Birgé (2002).

The resulting estimators, which we shall 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 $\hat{s}$ (de-

pending 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 belongs 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 (not just one), it automatically provides a model selection procedure, tending to choose the best model (in a suitable sense) among the family. Therefore good choices of the families of models result in adaptive estimators. From this point of view, one superiority of T-estimators over more convential selection methods is the fact that they can cope with fairly arbitrary countable families of models, possibly nonlinear or infinite dimensional. In particular, one can mix conventional parametric models with those used for nonparametric 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.

Let us now briefly sketch the main ideas underlying our construction. The starting point is a careful analysis of the way the m.l.e. does (or doesn't) work. The essential is to realize that computing the m.l.e. (and, more generally, any minimum contrast estimator) over some parameter set $\mathcal{S}$ amounts to perform a large number of simultaneous tests. Indeed, if we denote by $\Lambda(t, \boldsymbol{X})$ the likelihood of the observation $\boldsymbol{X}$ when the parameter t obtains, we can consider all likelihood ratio tests between distinct

points $t \neq u$ in $\mathcal{S}$ which accept t or u according to the fact that $\Lambda(t, \boldsymbol{X}) > \Lambda(u, \boldsymbol{X})$ or $\Lambda(t, \boldsymbol{X}) < \Lambda(u, \boldsymbol{X})$ (assuming no ties for simplicity). The m.l.e. $\hat{s}$ over $\mathcal{S}$ is the point which is accepted against any other. As a consequence, the performances of the m.l.e. depend on the properties of those tests. But they are too many to be controlled simultaneously without some continuity properties of the likelihood process, hence the restrictive assumptions needed to make the empirical process arguments work.

A natural idea to avoid testing between so many points is (assuming that $\mathcal{S}$ is compact) to restrict the maximization of the likelihood to some finite approximation S of $\mathcal{S}$. Unfortunately, this does not work in general because likelihood ratio tests are not robust and one cannot say much about the performances of the test between t and u when s obtains, even if s is close to t or u, unless one again puts some restrictive assumptions on the likelihood ratios. In order to make the argument work we have to replace the likelihood ratio tests by robust ones (tests between balls) and then find a way to build an estimator from a family of such tests.

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 Le Cam (1973 and 1975) he 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 Birgé (1983, 1984a, 1984b and 1986), using an alternative construction, still based on testing, we extended Le Cam's results with an emphasis on the minimax risk for nonparametric 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 Le Cam (1997), these ideas remained widely unknown since then (with the exception of Groeneboom, 1986) and, to our knowledge, nobody (including the present author) tried to apply the method to other stochastic frameworks like the Gaussian one, 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 (1985), Yang and Barron (1999) and Devroye and Lugosi (2001).

Some fundamental ideas for complexity-based model selection, which are also somewhat related to testing, appeared later in Barron and Cover (1991) and Barron (1991). 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 Barron and Cover (1991) and subsequent works will allow us to substantially improve and generalize the constructions of Birgé (1983 and 1984b) in particular towards model selection and adaptive estimation.

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 Birgé (1983) and Yang and Barron (1999). 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

The next section first illustrates, 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 a 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 various difficulties encountered. Section 3 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 4, 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 two frameworks we consider here: independent variables and Gaussian sequences. 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 and which we call its *metric dimension*. Section 6 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.

The last section is devoted to various applications. In particular, we show how to mix models for parametric and nonparametric estimation. In the Gaussian sequence framework, we show that T-estimators not only allow to recover all the results of Birgé and Massart (2001) 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 Birgé and Massart (2001) about the White noise framework could be transfered to density estimation with Hellinger loss in the same way. To conclude the section, we show how to use our method for aggregation of preliminary estimators, for instance to select a partition for histogram estimation.

1.5 Two illustrations with independent variables

Let us conclude this section with two specific applications, as an appetizer for the reader. 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 $X_1, \ldots, X_n$, although independent, do not necessarily follow the assumed model.

Proposition 1 *Let $X_1, \ldots, X_n$, $n \geq 200$, be independent random variables with arbitrary unknown distributions $\overline{P}_i$, $1 \leq i \leq n$ 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 $\hat{\theta}(X_1, \ldots, X_n)$ such that, whatever the distributions $\overline{P}_i$,*

$$\mathbb{E}\left[\sup_{1\leq i\leq n} h^2\left(\overline{P}_i, \mathcal{U}_{\hat{\theta}}\right)\right] \leq C \inf_{\theta>0}\left\{\sup_{1\leq i\leq n} h^2\left(\overline{P}_i, \mathcal{U}_\theta\right) + \frac{[\log(|\log\theta|) - n/200 - 14.8] \vee 1}{n}\right\},$$

where C denotes a universal constant.

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. $X_1, \ldots, X_n$ 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 $[\log(|\log\theta_0|) - n/200 - 14.8] \vee 1$ (which is 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 only slightly wrong in the sense that $\sup_{1\leq i\leq n} h^2\left(\overline{P}_i, \mathcal{U}_{\theta_0}\right) \leq 2/n$ for some $\theta_0 > 0$, the risk of our estimator remains of order $n^{-1}([\log(|\log\theta_0|) - n/200 - 14.8] \vee 1)$ while the risk of the m.l.e. may become larger than 0.45 as shown in Section 2.1.3 below.

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

$$\mathbb{E}_s\left[\|s - \hat{s}\|_1^q\right] \leq C(r,\alpha,p,q,d) R^{dq/(2\alpha+d)} n^{-q\alpha/(2\alpha+d)} \quad \textit{for all } q \geq 1.$$

As far as we know, all results about such estimation problems (without additional boundedness assumptions), even those dealing with the minimax risk for known α and p, are limited to the range $\alpha > d/p$, while our method allows to handle the larger scale of Besov spaces given by $\alpha > (d/p - d)_+$. Such risk bounds can be found in Donoho, Johnstone, Kerkyacharian and Picard (1996). The more recent improved results of Kerkyacharian and Picard (2000) do not apply to the $\mathbb{L}_1$-loss.

2 The difficulties connected with maximum likelihood estimation and some possible remedies

2.1 A few illustrations of the deficiencies of the m.l.e.

The m.l.e. is known to behave in an optimal way for parametric estimation under suitable regularity assumptions — see, for instance, Le Cam (1970) or the book by van der Vaart (1998) — and to have the right rate of convergence in nonparametric situations under specific entropy assumptions — van de Geer (1990, 1993 and 2000),

Birgé and Massart (1993), Shen and Wong (1994) and Wong and Shen (1995) —. 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 (1990) in a paper 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 the 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 Birgé and Massart (1993, Section 4) and Devroye and Lugosi (2001, 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 bracketting conditions as in Van de Geer (1993 and 2000) — 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.

2.1.1 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 $X_1, \ldots, X_n$ 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 $p = \int_{-\infty}^{0} f(t)\, dt$, 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.

2.1.2 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 Birgé and Massart (1998) relying on the construction of rather complicated non-parametric parameter spaces. A much simpler and nonasymptotic illustration of the suboptimality of the m.l.e. when the parameter space is too "massive", although it involves a convex compact parameter space, 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, $c = k^{1/4}$ and*

$$\mathcal{S} = \left\{ s \in \mathbb{R}^{k+1} \;\middle|\; |s_0| \leq c \quad \textit{and} \quad \|s'\| \leq 2(1 - |s_0|/c) \right\}.$$

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

$$\sup_{s \in \mathcal{S}} \mathbb{E}_s \left[\|s - \hat{s}\|^2 \right] \geq (3/4)\sqrt{k} + 3 \qquad \textit{and} \qquad \inf_{\tilde{s}} \sup_{s \in \mathcal{S}} \mathbb{E}_s \left[\|s - \tilde{s}\|^2 \right] \leq 5.$$

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 the Appendix.

2.1.3 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 $n \geq 3$ 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 $0 < \theta \leq 1$. 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

$$h^2(P, Q) = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ} \right)^2, \quad \rho(P, Q) = \int \sqrt{dP dQ} = 1 - h^2(P, Q), \quad (2.1)$$

where dP and dQ 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 $\mathbb{E}_\theta \left[h^2 \left(\mathcal{U}_\theta, \mathcal{U}_{X_{(n)}} \right) \right] = 1/(2n+1)$. Let us now suppose that $\overline{P}$ does not belong to $\mathcal{S}$ but has the density

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

with respect to Lebesgue measure. Since $h^2 \left(\overline{P}, \mathcal{U}_{1/10} \right) < 2n^{-1}$, one would expect the increase of risk due to this small deviation from the parametric assumption to be $O(1/n)$ if the m.l.e. were robust. This is not the case: with probability $1-(1-2/n)^n > 1 - e^{-2}$, $X_{(n)} \geq 9/10$ and therefore $\rho \left(\overline{P}, \mathcal{U}_{X_{(n)}} \right) \leq \sqrt{2}/3$. It follows that

$$\mathbb{E}_{\overline{P}} \left[h^2 \left(\overline{P}, \mathcal{U}_{X_{(n)}} \right) \right] > \left(1 - \sqrt{2}/3 \right) \left(1 - e^{-2} \right) > 0.45.$$

2.2 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. In this section, we assume that we observe 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 $\mathcal{X}$, where s belongs to some parameter set $\mathcal{S}$. The corresponding probabilities are denoted 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 $h(s,t) = h\left(\overline{P}_s, \overline{P}_t\right)$ 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 $\left\{\overline{P}_s, s \in \mathcal{S}\right\}$ is dominated with respective densities $d\overline{P}_s$ and, to be consistent with the minimum contrast estimators approach, we shall denote by $\Lambda_n(t, \boldsymbol{X}) = -\sum_{i=1}^n \log\left(d\overline{P}_t(X_i)\right)$ 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 $t \in S$.

2.2.1 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 $\mathcal{S}$ by a finite subset S such that for $s \in \mathcal{S}$ one can find $t \in S$ with $h(s,t) \leq \eta$ 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

$$\mathbb{P}_s[\Lambda_n(u, \boldsymbol{X}) = \Lambda_n(t, \boldsymbol{X})] = 0 \quad \text{for all } s \in \mathcal{S}, \;\; t, u \in S, \quad t \neq u. \tag{2.2}$$

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 analysis which goes back to Wald (1949). The first step is to observe that, whatever t and u, the errors of likelihood ratio tests between $\overline{P}_t$ and $\overline{P}_u$ are bounded by

$$\mathbb{P}_t[\Lambda_n(u, \boldsymbol{X}) \leq \Lambda_n(t, \boldsymbol{X})] \leq \exp\left[n \log\left[\rho\left(\overline{P}_u, \overline{P}_t\right)\right]\right] \leq \exp\left[-nh^2(u,t)\right], \tag{2.3}$$

which follows from (8.3) in the Appendix and (2.1). More precise results in this direction can be found in Chernoff (1952).

Now, given $\eta > 0$, $K \geq 1$ and $s \in S$, we want to bound $\mathbb{P}_s\left[h(s, \hat{s}) \geq K\eta\right]$. For this, we set $S_k = \left\{u \in S \;\middle|\; 2^{k/2} K\eta \leq h(s,u) < 2^{(k+1)/2} K\eta\right\}$ and denote by $|S_k|$ the cardinality of S_k. We derive from (2.3) that

$$\begin{aligned}
&\mathbb{P}_s\left[h(s, \hat{s}) \geq K\eta\right] \\
&\quad \leq \mathbb{P}_s\left[\exists u \in S \text{ with } h(s,u) \geq K\eta \text{ and } \Lambda_n(u, \boldsymbol{X}) \leq \Lambda_n(s, \boldsymbol{X})\right] \qquad (2.4) \\
&\quad = \sum_{k=0}^{+\infty} \mathbb{P}_s[\exists u \in S_k \text{ with } \Lambda_n(u, \boldsymbol{X}) \leq \Lambda_n(s, \boldsymbol{X})] \\
&\quad \leq \sum_{k=0}^{+\infty} |S_k| \sup_{u \in S_k} \mathbb{P}_s[\Lambda_n(u, \boldsymbol{X}) \leq \Lambda_n(s, \boldsymbol{X})] \;\leq\; \sum_{k=0}^{+\infty} |S_k| \exp\left[-2^k n K^2 \eta^2\right]. \qquad (2.5)
\end{aligned}$$

In order to get a small bound for the right-hand side of (2.5), one should first require that the first term of the series, $|S_0| \exp\left[-nK^2\eta^2\right]$, be small, which will determine

the choice of η. In particular, one should require that $nK^2\eta^2 \geq 1$. 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 $|S_k| \leq |S_0| \exp(2^{k-1})$ would do.

2.2.2 An alternative point of view on the previous analysis

If $s \in \mathcal{S} \setminus S$, one can find $t \in S$ with $h(s,t) \leq \eta$, hence $\mathbb{P}_s\left[h(s,\hat{s}) \geq (K+1)\,\eta\right] \leq \mathbb{P}_s\left[h(t,\hat{s}) \geq K\eta\right]$ and the previous arguments could be extended straightforwardly, at least for K large enough, if we could bound $\mathbb{P}_s[\Lambda_n(u,\boldsymbol{X}) \leq \Lambda_n(t,\boldsymbol{X})]$ by an analogue of (2.3) when $h(s,t) \leq \eta$ 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 rather restrictive assumptions on the likelihood ratios, such a property does not hold, as shown by the following counterexample. Let μ denote the Lebesgue measure on $[0,1]$, $\overline{P}_u = u \cdot \mu$ for any density u with respect to μ, $s = \mathbb{1}_{[0,1]}$, $\lambda = 1-(2n)^{-1}$ and $t = \lambda^{-1}\mathbb{1}_{[0,\lambda]}$. One can check that $h^2(s,t) < (3n)^{-1}$, which implies that there exists no perfect test between s and t, whatever n, therefore that s is quite close to t. Let u be any density such that $\|\log u\|_\infty < +\infty$. Then, even if $h(t,u)$ is close to one, which means that u is far away from t,

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

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 (2.2), 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 $t \in S$, $\mathcal{R}_t = \{u \in S \,|\, \Lambda_n(u,\boldsymbol{X}) < \Lambda_n(t,\boldsymbol{X})\}$ 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 $t,u \in S$, $\mathcal{D}_{\boldsymbol{X}}(t) \vee \mathcal{D}_{\boldsymbol{X}}(u) \geq h(t,u)$ and therefore that $h(t,\hat{s}) \leq \mathcal{D}_{\boldsymbol{X}}(t) \vee \mathcal{D}_{\boldsymbol{X}}(\hat{s}) \leq \mathcal{D}_{\boldsymbol{X}}(t)$. Finally, if $h(s,t) \leq \eta$,

$$\begin{aligned}\mathbb{P}_s[h(s,\hat{s}) \geq (K+1)\eta] &\leq \mathbb{P}_s[h(t,\hat{s}) \geq K\eta] \\ &\leq \mathbb{P}_s[\mathcal{D}_{\boldsymbol{X}}(t) \geq K\eta] \;=\; \mathbb{P}_s\left[\exists u \in \mathcal{R}_t \text{ with } h(t,u) \geq K\eta\right],\end{aligned}$$

since S is finite. This is equivalent to (2.4) but the proof of (2.5) cannot procede 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 absolutely no reason anymore that one could express $\hat{s}$ as $\operatorname{argmin}_{t\in S} \gamma_n(t,\boldsymbol{X})$ for some contrast function γ_n. Nevertheless, the bound

$$\mathbb{P}_s[h(s,\hat{s}) \geq (K+1)\eta] \leq \mathbb{P}_s\left[\exists u \in \mathcal{R}_t \text{ with } h(t,u) \geq K\eta\right] \tag{2.6}$$

still holds and if we can bound the errors of the new robust tests by some suitable analogue of (2.3), one could procede as before from (2.6) to some analogue of (2.5). Typically, we shall require that the robust tests we use satisfy the following error bound for some constants c and κ:

$$\mathbb{P}_s\left[\text{the test between } t \text{ and } u \text{ decides } u\right] \leq \exp\left[-cnh^2(t,u)\right] \quad \text{if } h(t,u) \geq \kappa h(s,t). \tag{2.7}$$

Deriving an estimator of $s \in \mathcal{S}$ from a family of tests satisfying (2.7) 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 $t \in S$, all the tests between t and the points $u \in S$ far enough from t will accept t with large probability and $\mathcal{D}_{\boldsymbol{X}}(t)$ should therefore not be large. On the other hand, if $u \in S$ is far enough from s, the test between t and u will accept t 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 (2.3) like (2.7), but it has been known for a long time that such tests do exist, as shown by Le Cam (1975) and Birgé (1984a). 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 (2.5) (counting the number of points of S contained in balls and bounding the errors of tests), at the chore of our method.

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

3.1 A general statistical framework

We observe some random element $\boldsymbol{X} \in \mathcal{E}$ and consider a set $\{P_t, t \in M\}$ of distributions on $\mathcal{E}$ parametrized by a semi-metric space (M, d). By semi-metric, we mean that the function d is a semi-distance, i.e. satisfies

$$d(t,t) = 0 \qquad \text{and} \qquad d(t,u) = d(u,t) \geq 0 \quad \text{for all } t,u \in M, \tag{3.1}$$

but not necessarily a version of the triangular inequality

$$d(t,u) \leq A[d(t,r) + d(r,u)] \quad \text{for all } r,t,u \in M \text{ and some } A \geq 1. \tag{3.2}$$

Of course, for most applications, in particular those developed in this paper, d will be a genuine distance satisfying (3.2) with $A = 1$. Nevertheless, part of the results we shall prove here only require that (3.1) hold 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 (3.2) 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 $A = 1$.

We assume that the parametrization $t \mapsto P_t$ is one-to-one which will allow us to systematically identify t and P_t, M with a set of distributions on $\mathcal{E}$ and write indifferently $d(P_t, P_u)$ or $d(t,u)$. We shall use the classical notations for open and closed balls in M with center t and radius $r \geq 0$:

$$\mathcal{B}_d(t,r) = \{u \in M \,|\, d(u,t) < r\} \quad \text{and} \quad \overline{\mathcal{B}}_d(t,r) = \{u \in M \,|\, d(u,t) \leq r\}, \tag{3.3}$$

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 by $d(t,S) = \inf_{u\in S} d(t,u)$.

We denote by $\mathbb{P}_s$ the probability that gives $\boldsymbol{X}$ its true distribution P_s and by $\mathbb{E}_s$ the corresponding expectation operator. To any measurable map $\hat{s}$ from $\mathcal{E}$ to M corresponds the estimator $\hat{s}(\boldsymbol{X})$. By *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.

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 will often be used to denote constants that may vary from line to line.

3.2 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 represents our *model(s)* for the true s;

ii) a family of tests between the points of S;

iii) a small nonnegative parameter ε.

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 $\boldsymbol{X} \in \mathcal{E}$ and two distributions $P_t, P_u \in M$, a (non-randomized) test between P_t and P_u (or equivalently 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. For instance, the likelihood ratio tests that we used in Section 2.2.1 should be defined by

$$\psi(t,u,\boldsymbol{X}) = \begin{cases} 0 & \text{if} \quad \Lambda_n(t,\boldsymbol{X}) < \Lambda_n(u,\boldsymbol{X}) \\ 1 & \text{if} \quad \Lambda_n(t,\boldsymbol{X}) > \Lambda_n(u,\boldsymbol{X}), \end{cases}$$

the value of ψ being irrelevant when the likelihood ratio is equal to one in view of (2.2). We can now define a T-estimator in the following way.

Definition 2 *Let $\varepsilon \geq 0$ be given, S be a countable subset of M and $\{\psi(t,u,\boldsymbol{X})\}$ 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 $\mathcal{R}_t = \{u \in S, u \neq t \,|\, \psi(t,u,\boldsymbol{X}) = 1\}$, we define the random function $\mathcal{D}_{\boldsymbol{X}}$ on S by*

$$\mathcal{D}_{\boldsymbol{X}}(t) = \begin{cases} \sup\limits_{u\in\mathcal{R}_t} \big\{d(t,u)\big\} & \text{if } \mathcal{R}_t \neq \emptyset; \\ \\ 0 & \text{if } \mathcal{R}_t = \emptyset. \end{cases} \tag{3.4}$$

Assuming that $\inf_{u\in S} \mathcal{D}_{\boldsymbol{X}}(u) < +\infty$, we call T-estimator *(alternatively T_ε-estimator when we want to emphasize the dependence on ε) derived from S and the family of tests $\{\psi(t,u,\boldsymbol{X})\}$ any measurable application $\hat{s}(\boldsymbol{X})$ with values in S satisfying $\mathcal{D}_{\boldsymbol{X}}(\hat{s}(\boldsymbol{X})) \leq \inf_{t\in S} \mathcal{D}_{\boldsymbol{X}}(t) + \varepsilon$.*

Note that a T_ε-estimator exists if $\inf_{u\in S} \mathcal{D}_{\boldsymbol{X}}(u) < +\infty$ and $\varepsilon > 0$, the first condition being always satisfied under the assumptions we consider here. When $\varepsilon = 0$ and S is infinite T_0-estimators do not automatically exist. In any case, unicity is not warranted.

The definition implies the following trivial, but basic properties:

Lemma 1 *The function $\mathcal{D}_{\boldsymbol{X}}$ and any T_ε-estimator $\hat{s}(\boldsymbol{X})$ satisfy for all $(t,u) \in S^2$,*

$$\mathcal{D}_{\boldsymbol{X}}(t) \vee \mathcal{D}_{\boldsymbol{X}}(u) \geq d(t,u) \qquad \textit{and} \qquad d(\hat{s}(\boldsymbol{X}),t) \leq \mathcal{D}_{\boldsymbol{X}}(t) + \varepsilon.$$

Moreover, if d is a distance,

$$d(s,\hat{s}(\boldsymbol{X})) \leq \inf_{t\in S} \left\{ d(s,t) + \mathcal{D}_{\boldsymbol{X}}(t) \right\} + \varepsilon \quad \textit{for all } s \in M. \tag{3.5}$$

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 of the point t: if it is large, one should believe that the true s is far from t. From this point of view, (3.5) says that a T-estimator makes the best compromise among the points in S between the distance from t to the true s and its plausibility.

4 T-estimators based on a single model

4.1 Working within the general framework

In this section, the set S will be *the* model for the unknown parameter s. Our aim is to prove some large deviation results for the minimizer(s) of $\mathcal{D}_{\boldsymbol{X}}$ that are quite similar to those which one can prove for minimum contrast estimators, with the noticeable superiority of our construction that the assumptions we require are much weaker than the usual ones.

4.1.1 The assumptions

In order to ensure the existence of T-estimators and evaluate their performances, we need two assumptions, one which controls the errors of the tests ψ and provides an analogue of (2.7) and one relative to the "massiveness" of S which bounds the number of points of S that are contained in balls. They are as follows.

Assumption 1 *We say that a subset S of M satisfies Assumption 1 if there exists a function δ from $M \times S$ to $[0,+\infty]$ such that $\delta(s,t) \geq \kappa d(s,t)$ for some $\kappa \geq 4$ and two constants $a, B > 0$ such that one can find, for each pair $(t,u) \in S^2$ with $t \neq u$, a test function $\psi(t,u,\boldsymbol{X})$ (with $\psi(u,t,\boldsymbol{X}) = 1 - \psi(t,u,\boldsymbol{X})$) satisfying*

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

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 $B\exp\left[-ad^2(t,u)\right]$ whatever $u \neq t$ and this remains true as long as s remains "close enough" to t in the sense that $\delta(s,t) \leq d(t,u)$. 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, $\delta = \kappa d$, 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 that not all (semi-)distances do suit our needs: the construction of tests that satisfy the previous assumption is only possible for some special (semi-)distances. An illustration of this fact will be given in Section 5.5 below.

Assumption 2 *The set S satisfies Assumption 2(η, D, B') for $\eta, B' > 0$ and $D \geq 1/2$ if*

$$|S \cap \mathcal{B}_d(t,r)| \leq B' \exp\left[D[(r/\eta) \vee 2]^2\right] \quad \textit{for all } r > 0 \textit{ and } t \in S. \tag{4.1}$$

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 η, D and B' and lattices in Euclidean spaces as well. Further examples will be given in Section 5.

Some easy consequences of this assumption, which we shall repetedly use in the sequel, are as follows.

Lemma 2 *If S satisfies Assumption 2(η, D, B'), then S is at most countable and it also satisfies Assumption 2(η', D', B') for all $\eta' > 0$ and $D' = D\left[(\eta'/\eta)^2 \vee 1\right]$.*

If d is a distance, then

$$|S \cap \mathcal{B}_d(t,r)| \leq B' \exp\left[4D[(r/\eta) \vee 1]^2\right] \quad \textit{for all } r > 0 \textit{ and } t \in M, \tag{4.2}$$

and for any function δ from $M \times S$ to $[0, +\infty]$ such that $\delta(s,t) \geq \kappa d(s,t)$ for some positive κ, there exists a well-defined "projection" 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 "projection" operator π from M to S with $d(s, \pi(s)) = d(s,S) = \inf_{t \in S} d(s,t)$.

Proof: The first part of the lemma is clear. If d is a distance and $S \cap \mathcal{B}(t,r)$ is not empty, it contains at least one point u and is therefore included in $S \cap \mathcal{B}(u, 2r)$ with $u \in S$. Hence (4.2) follows from Assumption 2. The existence of π' is immediate although it may not be unique. In order to solve the problem of ties, it suffices to index S by the integers and choose for $\pi'(s)$ the element with the smallest index, the same being true for π. □

4.1.2 General risk bounds for T-estimators

Under the previous assumptions, we can derive the existence of T_0-estimators and bound their risk. To this end, we shall introduce, for each $q \geq 1$ a special function ζ_q, the importance of which is justified by the following

Proposition 3 *Let Y be a nonnegative random variable such that*

$$\mathbb{P}[Y > y] \leq \alpha \exp\left(-\beta y^2\right) \quad \textit{for } y \geq \overline{y} > 0, \tag{4.3}$$

where α, β denote some positive constants. Then, whatever $\lambda \geq 0$ and $q \geq 1$,

$$\begin{aligned} \mathbb{E}[(Y + \lambda\overline{y})^q] &\leq (1+\lambda)^q \overline{y}^q \left[1 + \alpha\zeta_q\left(\beta\overline{y}^2\right)\right] &(4.4)\\ &\leq (1+\lambda)^q \left[\overline{y}^q + \alpha\beta^{-q/2}\sqrt{\pi e q/2}\left[q/(2e)\right]^{q/2}\right], &(4.5)\end{aligned}$$

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

$$\zeta_q(x) = \sqrt{\frac{\pi e q}{2}}\left[\frac{q}{2ex}\right]^{q/2}\mathbb{1}_{(0,cq)}(x) + \frac{q}{2}e^{-x}\mathbb{1}_{[cq,+\infty)}(x); \quad c = \begin{cases} 1/2 & \text{if } q \le 2\pi e; \\ 0.612 & \text{if } q > 2\pi e. \end{cases} \tag{4.6}$$

The proof is given in the Appendix.

Theorem 2 *Let d be a distance and S satisfy Assumptions 1 and 2(η, D, B') with $\delta = \kappa d$ and $2a\eta^2 \ge 3D$. Then, for all $s \in M$, $\mathbb{P}_s$-a.s. there exists T_0-estimators $\hat{s}(\boldsymbol{X})$ derived from S and any of them satisfies, for all $q \ge 1$,*

$$\mathbb{E}_s\left[d^q(s,\hat{s})\right] \le \left(1+\kappa^q\right)\left[d(s,S)\vee\eta\right]^q + \overline{C}(\pi e q/2)^{1/2}(5q/e)^{q/2}a^{-q/2}, \tag{4.7}$$

with $\overline{C} = 2.2BB'$, and

$$\mathbb{E}_s\left[d^q(s,\hat{s})\right] \le \left(1+\kappa^q\right)\left[d(s,S)\vee\eta\right]^q\left[1+\overline{C}\zeta_q\left(\kappa^2 a\eta^2/6\right)\right]. \tag{4.8}$$

These bounds actually require some comments. Since they are easier to analyze for moderate values of q, we shall, for simplicity, restrict ourselves to the quadratic risk ($q=2$) and assume that $B=1$ (which is the typical situation), getting from (4.7)

$$\mathbb{E}_s\left[d^2(s,\hat{s})\right] \le \left(1+\kappa^2\right)\left[d(s,S)\vee\eta\right]^2 + 22B'\sqrt{\pi/e}\,a^{-1} \tag{4.9}$$

and from (4.8), since $\kappa^2 a\eta^2/6 \ge \kappa^2 D/4 \ge 4D \ge 2$,

$$\mathbb{E}_s\left[d^2(s,\hat{s})\right] \le \left(1+\kappa^2\right)\left[d(s,S)\vee\eta\right]^2\left[1+2.2B'\exp\left(-\kappa^2 a\eta^2/6\right)\right]. \tag{4.10}$$

i) Both bounds reveal the relevance of the parameter B' which controls what we shall call the "remainder" terms. Indeed, it is easy to see that Assumptions 2 is over parametrized since we could always normalize B' to one. This is clear if $B' < 1$ and if $B' > 1$, we could merely change D to $D + (\log B')/4$. If B' is small, this would obviously be a net loss for the risk bounds. In the opposite case, the operation improves the remainder term but deteriorates the main one because of the requirement $2a\eta^2 \ge 3D$ which typically forces us to enlarge η if we enlarge D. In situations where $a\eta^2$ is large, B' can easily be absorbed by the exponential factor in (4.10) while enlarging η is not a good strategy. This accounts a posteriori for the introduction of B', even if we shall often normalize it to one in the examples.

ii) The "remainder" term plays a different role in the two bounds. In (4.9) it is an additive term of order a^{-1} independent of η and $d(s,S)$ (for instance, as we shall see, of order n^{-1} for i.i.d. observations). In (4.10), it has a multiplicative effect and is negligeable as compared to one if, for instance, D is larger than B'.

iii) If we omit the remainder term and forget about $d(s,S)$ which is clearly unknown, it appears that one would like to minimize $\kappa\eta$. A first step is generally to take $2a\eta^2 \simeq 3D$ when this is feasible. Then it remains to make κ^2/a as small as possible, which involves the formulation of Assumption 1. In most cases and the example of Proposition 4 below will clearly illustrate this fact, there is no unique choice for the pair (κ, a) and, to some extent, the larger κ, the larger a. Our choices for those pairs in the applications below are motivated by the minimization of κ^2/a.

iv) Under the additional assumption that $d(s,S) \le \kappa'\eta$, one could handle in the same way more general loss functions and get bounds for $\mathbb{E}_s\left[\ell\left(\eta^{-1}d(s,\hat{s})\right)\right]$ provided that the function ℓ does not grow too fast at infinity. We leave this extension to the reader.

4.1.3 First applications

As an immediate consequence of (4.8) and the fact that $\zeta_q\left(\kappa^2 a\eta^2/6\right) \leq \zeta_q(2)$, one 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

$$R(\hat{s}, \mathcal{S}, q) = \sup_{s\in\mathcal{S}} \mathbb{E}_s\left[d^q(s, \hat{s})\right] \qquad \text{and} \qquad R(\mathcal{S}, q) = \inf_{\tilde{s}} R(\tilde{s}, \mathcal{S}, q), \tag{4.11}$$

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

Corollary 1 *Let (M, d) be a metric space satisfying Assumption 1 with $\delta = \kappa d$. Let $S \subset M$ satisfy Assumption 2(η, D, B') with $2a\eta^2 \geq 3D$ and $\hat{s}$ be a T_0-estimator derived from S (which exists a.s.), then*

$$R(\mathcal{S}, q) \leq R(\hat{s}, \mathcal{S}, q) \leq C(q, \kappa, BB') \left[\sup_{s\in\mathcal{S}} d(s, S) \vee \eta\right]^q \quad \text{for } q \geq 1. \tag{4.12}$$

We can similarly get risk bounds for maximum likelihood estimators over S when the likelihood ratio tests satisfy some robustness properties and, more generally, for minimum contrast estimators, the case of maximum likelihood estimators corresponding to $\gamma(t, \boldsymbol{X}) = -\log[(dP_t/d\mu)(\boldsymbol{X})]$ for some dominating measure μ.

Theorem 3 *Let γ be a function from $S \times \mathcal{E}$ to $[-\infty, +\infty)$ such that, for all $s \in M$,*

$$\mathbb{P}_s[\gamma(t, \boldsymbol{X}) = \gamma(u, \boldsymbol{X})] = 0 \quad \text{for all } (t, u) \in S^2, \ \ t \neq u, \tag{4.13}$$

and let the tests ψ given by

$$\psi(t, u, \boldsymbol{X}) = \begin{cases} 0 & \text{if } \gamma(t, \boldsymbol{X}) < \gamma(u, \boldsymbol{X}); \\ 1 & \text{if } \gamma(t, \boldsymbol{X}) > \gamma(u, \boldsymbol{X}) \end{cases} \tag{4.14}$$

satisfy Assumption 1 for some distance d and $\delta = \kappa d$. If Assumption 2(η, D, B') holds with $2a\eta^2 \geq 3D$, $\mathbb{P}_s$-a.s. there exists a unique T_0-estimator $\hat{s}(\boldsymbol{X})$ which is the unique minimizer with respect to $t \in S$ of the function $\gamma(t, \boldsymbol{X})$ and for all $s \in M$ and all $q \geq 1$, the risk of $\hat{s}$ is bounded by (4.7) and (4.8).

Before we prove the previous theorems (in Section 4.3), let us see how they can be applied to some specific stochastic frameworks.

4.2 Application to independent observations and Gaussian sequences

The previous results will apply easily if M itself satisfies Assumption 1, in which case one just has to find a suitable subset S of M satisfying Assumption 2 (η, D, B') with $2a\eta^2 \geq 3D$. This does happen for various statistical frameworks. To keep this paper to an acceptable size, we shall only consider three simple illustrations here, namely the independent, the i.i.d. and the Gaussian settings. The case of Gaussian regression with random design has been considered in Birgé (2002). Further examples will be given in subsequent papers.

4.2.1 The independent and i.i.d. settings

The independent setting In this setting, we observe a set $\boldsymbol{X} = (X_1, \ldots, X_n)$ of n independent random variables X_i with values in $\mathcal{X}$. We denote by $\overline{M}$ the set of all distributions on $\mathcal{X}$, set $\mathcal{E} = \mathcal{X}^n$ and take for M the set of all product distributions on $\mathcal{X}^n$: $M = \left\{P_t = \bigotimes_{i=1}^n \overline{P}_i,\ \overline{P}_i \in \overline{M} \text{ for } 1 \le i \le n\right\}$. As usual, we identify t and P_t and M with the set of ts.

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_is 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 $(X_1, \ldots, X_n)$ with marginal distributions $\overline{P}_t$ on $\mathcal{X}$. In this paper, we shall systematically restrict to considering models $S \subset M$ that are actually subsets of $\overline{M}$. This 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 $\overline{h}$ of the coordinates or the sup-variation distance $\overline{v}$. 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

$$v(P, Q) = \frac{1}{2} \int |dP - dQ| = \sup_A |P(A) - Q(A)|, \tag{4.15}$$

where the supremum is over all measurable sets. It is well-known from Le Cam (1973) that the two distances satisfy the inequalities

$$h^2(P, Q) \le v(P, Q) \le h(P, Q)\sqrt{2 - h^2(P, Q)}. \tag{4.16}$$

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

$$\overline{h}(t, u) = \sup_{1 \le i \le n} h\left(\overline{P}_i, \overline{Q}_i\right) \qquad \text{and} \qquad \overline{v}(t, u) = \sup_{1 \le i \le n} v\left(\overline{P}_i, \overline{Q}_i\right).$$

In particular, if $u \in \overline{M}$, i.e. $P_u = \overline{P}_u^{\otimes n}$, then

$$\overline{h}(t, u) = \sup_{1 \le i \le n} h\left(\overline{P}_i, \overline{P}_u\right) \qquad \text{and} \qquad \overline{v}(t, u) = \sup_{1 \le i \le n} v\left(\overline{P}_i, \overline{P}_u\right),$$

and if both t and $u \in \overline{M}$, then

$$\overline{h}(t, u) = h\left(\overline{P}_t, \overline{P}_u\right) \qquad \text{and} \qquad \overline{v}(t, u) = v\left(\overline{P}_t, \overline{P}_u\right), \tag{4.17}$$

which allows us to identify $\overline{h}$ with h and $\overline{v}$ with v on $\overline{M}$ and turn it to 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 P_t for i.i.d. samples $(X_1, \ldots, X_n)$ or equivalently restrict ourseleves 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 $d\overline{P}_t/d\mu = t$, hence $P_t = (t \cdot \mu)^{\otimes n}$.

4.2.2 The Gaussian setting

The Gaussian setting corresponds to the so-called "Gaussian sequence" framework which consists of the observation of a sequence $\boldsymbol{X} = (X_i)_{i\geq 1}$ of independent Gaussian variables with known variance σ^2 and respective means s_i, i.e. $X_i \sim \mathcal{N}(s_i, \sigma^2)$ with $s = (s_i)_{i\geq 1} \in M = \boldsymbol{l}_2(\mathbb{N}^\star)$ ($\mathbb{N}^\star = \mathbb{N}\setminus\{0\}$). 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 again by P_s the 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 $\overline{P}_0 = \mathcal{N}\left(0,\sigma^2\right)$ for reference measure, getting

$$\frac{dP_t}{dP_0}(\boldsymbol{X}) = \exp\left[\frac{1}{\sigma^2}\left(\langle t, \boldsymbol{X}\rangle - \frac{\|t\|^2}{2}\right)\right]. \tag{4.18}$$

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 $\mathbb{R}^n$. Setting $s_i = 0$ for $i > n$ immediately leads to the Gaussian setting. 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

$$Y(z) = \int_0^z s(x)\,dx + \sigma W(z), \quad 0 \leq z \leq 1, \tag{4.19}$$

where s is an unknown function in $\mathbb{L}_2([0,1],dx)$ and W is a Wiener process with $W(0) = 0$. Choosing some orthonormal basis $\{\varphi_i, i \geq 1\}$ of $\mathbb{L}_2([0,1],dx)$ and defining $s_i = \int_0^1 s(x)\varphi_i(x)\,dx$, $X_i = \int_0^1 \varphi_i(x)\,dY(x)$ leads to the Gaussian setting. The function s in (4.19) can be identified with the sequence $(s_i)_{i\geq 1}$ of its Fourier coefficients with respect to the basis $\{\varphi_i, i \geq 1\}$ via Plancherel's formula. Since this correspondance is an isometry, it allows us to view the White noise framework (4.19) as an alternative representation of the Gaussian setting with parameter space $M = \mathbb{L}_2([0,1],dx)$ and distance d corresponding to the $\mathbb{L}_2$-norm. Much more on this is to be found in Sections 1 and 6 of Birgé and Massart (2001).

4.2.3 The existence of robust tests

It is not difficult to check that Assumption 1 holds in the Gaussian setting, since then likelihood ratio tests are naturally robust as shown by the following proposition. The case $x = 0$ only is relevant for deriving Assumption 1, but the more general bounds given here, which use x to balance between the two errors of each test, will be required to deal with model selection in Section 6.

Proposition 4 *Let $\boldsymbol{X} = (X_i)_{i\geq 1} \in \mathbb{R}^{\mathbb{N}^\star}$ 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 $s,t,u \in \boldsymbol{l}_2(\mathbb{N}^\star)$,*

$$\mathbb{P}_s\left[\log\left(\frac{dP_u}{dP_t}\right)(\boldsymbol{X}) \geq 2x\right] \leq \exp\left[-x - \frac{\|t-u\|(\|t-u\| - 4\|s-t\|)}{8\sigma^2}\right].$$

In particular,

$$\sup_{\left\{s\in\boldsymbol{l}_2(\mathbb{N}^\star)\,\middle|\,\|s-t\|\le\|t-u\|/6\right\}} \mathbb{P}_s\left[\log\left(\frac{dP_u}{dP_t}\right)(\boldsymbol{X})\ge 2x\right]\le\exp\left[-x-\frac{\|t-u\|^2}{24\sigma^2}\right]$$

and, whatever $s\in\boldsymbol{l}_2(\mathbb{N}^\star)$,

$$\mathbb{P}_s\left[\log\left(\frac{dP_u}{dP_t}\right)(\boldsymbol{X})\ge 2x\right]\le\exp\left[-x+\frac{\|t-s\|^2}{2\sigma^2}\right].$$

In the 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 (1965) — see also Huber (1981, Section 10.3) — for the variation distance and by Le Cam (1975) and Birgé (1984a) for the Hellinger distance.

Proposition 5 *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* $\boldsymbol{X}=(X_1,\ldots,X_n)$ *is a set of independent random variables with distribution* $P_s=\bigotimes_{i=1}^n\overline{P}_i$, *then*

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

and

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

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

4.2.4 Application to the independent and Gaussian settings

The families of tests provided by the previous propositions (with $x=0$) do satisfy Assumption 1 with $B=1$ for suitable values of a and δ. In the Gaussian setting, we get $a=\left(24\sigma^2\right)^{-1}$ and $\delta=6d_2$, while for the independent setting with $S\subset\overline{M}$, we get

$$a=n/4,\ \ \delta=4\overline{h}\quad\text{(Hellinger case)}\quad\text{or}\quad a=n/8,\ \ \delta=4\overline{v}\quad\text{(variation case)}.$$

Applying Theorems 2 and 3 to the independent and Gaussian settings respectively leads to the following corollaries.

Corollary 2 *In the independent setting, let* S *be a subset of the set* $\overline{M}$ *of all possible distributions for* X_1 *(endowed with either the Hellinger distance* h *or the variation distance* v*) which satisfies Assumption 2*(η,D,B') *with* $\eta^2\ge 6D/n$ *(Hellinger case) or* $\eta^2\ge 12D/n$ *(variation case). We can build* T*-estimators* $\hat{s}$ *that satisfy either*

$$\mathbb{E}_s\left[\overline{h}^q(s,\hat{s})\right]\le C(q,B')\left[\inf_{t\in S}\overline{h}^q(s,t)\vee\eta^q\right]\quad\text{for all } q\ge 1\ \text{ and }\ s\in M,$$

or

$$\mathbb{E}_s\left[\overline{v}^q(s,\hat{s})\right]\le C(q,B')\left[\inf_{t\in S}\overline{v}^q(s,t)\vee\eta^q\right]\quad\text{for all } q\ge 1\ \text{ and }\ s\in M.$$

Corollary 3 *In the Gaussian setting, let S be a subset of $\boldsymbol{l}_2(\mathbb{N}^\star)$ (endowed with the distance d_2) which satisfies Assumption 2(η, D, B') with $\eta^2 \geq 36D\sigma^2$. Then the maximum likelihood estimator $\hat{s}$ over S a.s. exists, is unique and satisfies*

$$\mathbb{E}_s\left[\|s-\hat{s}\|^q\right] \leq C(q,B')\left[\inf_{t\in S}\|s-t\|^q \vee \eta^q\right] \quad \textit{for all } q \geq 1 \textit{ and } s \in M.$$

4.2.5 T-estimators based on 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 2.1.3 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 $\left[0, e^t\right]$, since then

$$h^2(t,u) = h^2\left(\overline{P}_t, \overline{P}_u\right) = 1 - \exp(-|t-u|/2) \leq |t-u|/2. \tag{4.20}$$

Given some $\gamma \in \mathbb{R}$, we shall set

$$\eta^2 = 16.8D/n \quad \text{with } D \geq 1/2; \quad J = \sup\left\{j \in \mathbb{N} \,\middle|\, j \leq 4.5\exp[(4D)\vee(n/84)]\right\}; \tag{4.21}$$

$$I = \left[\gamma, \gamma + 4J\eta^2\right] \qquad \text{and} \qquad S = \left\{\gamma + 2\eta^2(1+2j),\ j \in \mathbb{N},\ j \leq J-1\right\}. \tag{4.22}$$

It follows from (4.20) that $\inf_{t\in S} h\left(\overline{P}_t, \overline{P}_u\right) \leq \eta$ for all $u \in I$ and consequently that, whatever the distribution $P_s = \bigotimes_{i=1}^n \overline{P}_i$ of $\boldsymbol{X}$,

$$\inf_{t\in S}\overline{h}^2(P_s,P_t) = \inf_{t\in S}\sup_{1\leq i\leq n} h^2\left(\overline{P}_i,\overline{P}_t\right) \leq 2\left[\eta^2 + \inf_{t\in I}\sup_{1\leq i\leq n} h^2\left(\overline{P}_i,\overline{P}_t\right)\right]. \tag{4.23}$$

In order to check that S satisfies Assumption 2, 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 (4.21) and (4.22). Then, whatever the probability $P_s = \bigotimes_{i=1}^n \overline{P}_i$,*

$$|S \cap \mathcal{B}_{\overline{h}}(P_s, r)| \leq 4.5\exp\left[D[(r/\eta)\vee 2]^2\right] \quad \textit{for all } r > 0. \tag{4.24}$$

Proof: We shall distinguish between two situations. When $r^2 \geq 1/5$, then $(r/\eta)^2 \geq n/(84D)$ and (4.24) follows from (4.21) since $|S| = J$. If $r^2 < 1/5$, there is obviously nothing to prove if $r \leq \inf_{t\in S}\overline{h}(P_s,P_t)$ and we can therefore assume that there exists some $t \in S$ such that $\mathcal{B}_{\overline{h}}(P_s,r) \subset \mathcal{B}_{\overline{h}}(P_t,2r)$. It then follows from the definition of S and (4.20) that

$$|S \cap \mathcal{B}_{\overline{h}}(P_s,r)| \leq |S \cap \mathcal{B}_{\overline{h}}(P_t,2r)| \leq 2j+1,$$

where the integer j is defined by

$$1 - \exp\left(-2j\eta^2\right) < 4r^2 \leq 1 - \exp\left(-2[j+1]\eta^2\right).$$

Hence

$$|S \cap \mathcal{B}_{\overline{h}}(P_s,r)| \leq -\eta^{-2}\log\left(1-4r^2\right) + 1 < (5\log 5)(r/\eta)^2 + 1$$

since $-\log\left(1-4r^2\right) < (5\log 5)r^2$ for $r^2 < 1/5$. Finally (4.24) follows from the lower bound $D \geq 1/2$. □

The lemma implies that S satisfies Assumption $2\,(\eta, D, 4.5)$. 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 (4.23), the risk bound

$$\mathbb{E}_s\left[\overline{h}^2(s,\hat{s})\right] \leq C\left[\inf_{\theta\in\Theta}\sup_{1\leq i\leq n} h^2\left(\overline{P}_i, \mathcal{U}_\theta\right) + n^{-1}\right],$$

where Θ denotes the interval $\left[\exp(\gamma), \exp\left(\gamma + 4J\eta^2\right)\right]$. 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.

4.3 Proofs of Theorems 2 and 3

The control of $d(s,\hat{s})$ mainly follows from a large deviation inequality for the function $\mathcal{D}_{\boldsymbol{X}}$ which holds even if d is only a semi-distance.

Proposition 6 *Let Assumption 1 and Assumption $2\,(\eta, D, B')$ hold with $D \leq 2a\eta^2/3$. Let $s \in M$, $s' \in S$ be such that $\delta(s,s') < +\infty$ and $y_0 = (4\eta) \vee \delta(s,s')$. Then*

$$\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') > y\right] < 2.2BB'\exp\left(-ay^2/6\right) \quad \textit{for } y \geq y_0, \tag{4.25}$$

and, $\mathbb{P}_s$-a.e., there exists at least one T_0-estimator.

Proof: Let us set $\theta = 5/4$ and $S_k = \left\{t \in S \;\middle|\; \theta^{k/2}y \leq d(s',t) < \theta^{(k+1)/2}y\right\}$ with $y \geq y_0$. Then

$$\begin{aligned}
\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') > y\right] &\leq \mathbb{P}_s\left[\,\exists t \in S \text{ with } d(t,s') \geq y \text{ and } \psi(s',t,\boldsymbol{X}) = 1\right] \\
&= \sum_{k=0}^{+\infty}\mathbb{P}_s\left[\exists t \in S_k \text{ with } \psi(s',t,\boldsymbol{X}) = 1\right] \\
&\leq \sum_{k=0}^{+\infty}|S_k|\sup_{t\in S_k}\mathbb{P}_s\left[\psi(s',t,\boldsymbol{X}) = 1\right]. \qquad (4.26)
\end{aligned}$$

Since $S_k \subset \mathcal{B}_d\left(s, \theta^{(k+1)/2}y\right)$ and $\theta^{(k+1)/2}y \geq \sqrt{\theta}y > 2\eta$, it follows from Assumption $2\,(\eta, D, B')$ that $|S_k| \leq B'\exp\left[\theta^{k+1}(y/\eta)^2 D\right]$. If $t \in S_k$, then $d(s',t) \geq \theta^{k/2}y \geq y_0 \geq \delta(s,s')$ and we can apply Assumption 1 to derive from (4.26) and the bound on D that

$$\begin{aligned}
\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') > y\right] &\leq BB'\sum_{k=0}^{+\infty}\exp\left[\theta^{k+1}\frac{y^2}{\eta^2}D - a\theta^k y^2\right] \\
&\leq BB'\sum_{k=0}^{+\infty}\exp\left[-\frac{a\theta^k y^2}{3}(3-2\theta)\right].
\end{aligned}$$

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

$$\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') > y\right] \leq BB'\exp\left[-\frac{ay^2}{6}\right]\sum_{k=0}^{+\infty}\exp\left[\frac{ay^2}{6}\left(1-\theta^k\right)\right]$$

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

which proves (4.25). Therefore $\mathcal{D}_{\boldsymbol{X}}(s')$ is a.s. finite and by Lemma 1, the set of points $t \in S$ such that $\mathcal{D}_{\boldsymbol{X}}(t) \leq \mathcal{D}_{\boldsymbol{X}}(s')$ is a subset of $S \cap \overline{\mathcal{B}}_d(s', \mathcal{D}_{\boldsymbol{X}}(s'))$ hence finite by (4.1). It follows that the set of elements $u \in S$ such that $\mathcal{D}_{\boldsymbol{X}}(u) = \inf_{t \in S} \mathcal{D}_{\boldsymbol{X}}(t)$ is nonempty and finite. Since S is countable, it is possible to select an element $\hat{s}$ of this set in a measurable way. □

Proof of Theorem 2 Its assumptions implying those of Proposition 6, (4.25) holds for all $s' \in S$. Setting $s' = \pi(s)$ where π is the operator provided by Lemma 2, we get $d(s, S) = d(s, s')$ and derive from (3.5) with $\varepsilon = 0$ that $d(s, \hat{s}) \leq \mathcal{D}_{\boldsymbol{X}}(s') + d(s, s')$. In view of (4.25) we may then bound the risk of $\hat{s}$ via Proposition 3 with $Y = \mathcal{D}_{\boldsymbol{X}}(s')$, $\alpha = 2.2BB'$, $\beta = a/6$, $\overline{y} = \kappa[d(s, s') \vee \eta] \geq y_0$ and $\lambda = \kappa^{-1}$. We derive (4.8) from (4.4) since $a\overline{y}^2/6 \geq \kappa^2 a\eta^2/6$ and ζ_q is decreasing, and (4.7) from (4.5) since $3^{q/2}\left(1 + \kappa^{-1}\right)^q < 5^{q/2}$. □

Proof of Theorem 3 We can again apply Proposition 6, getting some T_0-estimator $\hat{s}(\boldsymbol{X})$. It follows from the definition of $\mathcal{D}_{\boldsymbol{X}}$ and the form of our tests that $\gamma(\hat{s}, \boldsymbol{X})$ is (a.s.) smaller than $\gamma(t, \boldsymbol{X})$ at any point $t \in S$ with $d(\hat{s}, t) > \mathcal{D}_{\boldsymbol{X}}(\hat{s})$. Since the number of remaining points is finite by (4.1), a.s. there exists a minimizer $\tilde{s}$ of γ. Since S is countable, all values of $\gamma(t, \boldsymbol{X})$ are different a.s. by (4.13). Hence $\tilde{s}$ is unique, therefore satisfies $\mathcal{D}_{\boldsymbol{X}}(\tilde{s}) = 0$ so that $\hat{s} = \tilde{s}$. We apply the results of Theorem 2 to conclude. □

5 Metric dimension and minimax risk

If we consider a stochastic framework (like the Gaussian and i.i.d. settings) for which M itself satisfies Assumption 1, we can bound the minimax risk over subsets $\mathcal{S}$ of M via Corollary 1. Optimizing the upper bound in (4.12) for given values of q, κ, B and B' amounts to minimize $\sup_{s \in \mathcal{S}} d(s, S) \vee \eta$ with respect to those S and η such that S satisfies Assumption 2 (η, D, B') with $\eta^2 \geq 3D/(2a)$. Since, by Lemma 2, for any $\eta' > \eta$, one can always replace the pair η, D by η', D' satisfying the same relationship, one can assume that $\eta \geq \sup_{s \in \mathcal{S}} d(s, S)$. Therefore, we have to look for the minimal value of η such that there exists some S satisfying Assumption 2 (η, D, B') with $\eta^2 \geq 3D/(2a)$ and $d(s, S) \leq \eta$ 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 sets S, 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: $\sup_{s \in \mathcal{S}} d(s, S') \leq \eta$. Since such sets are often not countable, they cannot be used directly for our construction and 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 4.2.5. It is therefore important to understand how one can derive T-estimators from "natural" approximation spaces or simple parametric families.

5.1 Introducing metric dimensions

The previous reasoning assumed a fixed value of B' and, as we noticed in Section 4.1.2, there are some possible balances in Assumption 2 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 5. If one wants to use the influence of B' efficiently, it is better to go back to the general point of view we took for Section 4.

Let us now recall the following definitions from Approximation Theory.

Definition 3 *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, whatever $s \in \mathcal{S}$, one can find $t \in S$ such that $d(s, t) \leq \eta$. 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 $\mathcal{S}$ need 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 for $\mathcal{S}$ satisfying Assumption 2 $(\eta, D, 1)$ with $D \leq 2a\eta^2/3$ only depend on some metric properties of $\mathcal{S}$ that describe its "massiveness". In view of Assumption 2, it may seem natural to caracterize this massiveness by the function D' defined on $(0, +\infty)$ by

$$D'(\eta) = \eta^2 \inf_{S_\eta} \sup_{t \in S_\eta;\, r \geq 2\eta} r^{-2} \log\left(|S_\eta \cap \mathcal{B}(t, r)|\right),$$

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 $D'(\eta) \leq \eta^2 \log 3$ for $\eta < 1/2$ and $D'(1/2) \geq (\log 2)/4$. 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 smooth upper bound $\widetilde{D}$ for D' that we call *a bound for the metric dimension* of $\mathcal{S}$. The (minor) restrictions imposed to $\widetilde{D}$ in the next definition have been chosen in order to make our further analysis simpler. They are actually harmless in view of the use we make of $\widetilde{D}$.

Definition 4 *Let $\mathcal{S}$ be a subset of some metric space (M, d) and $\widetilde{D}$ be a right-continuous function from $(0, +\infty)$ to $[1/2, +\infty]$, not identically equal to $+\infty$. We shall 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 $\mathcal{S}$ such that*

$$|S_\eta \cap \mathcal{B}(t, r)| \leq \exp\left[\widetilde{D}(\eta)[(r/\eta) \vee 2]^2\right] \quad \textit{for all } r > 0 \;\; \textit{and} \;\; t \in M, \tag{5.1}$$

and that it has a finite metric dimension bounded by $\overline{D} \in \mathbb{R}$ *if (5.1) holds with $1/2 \leq \widetilde{D}(\eta) = \overline{D} < +\infty$ for all $\eta > 0$.*

Let us notice here that if $\mathcal{S}$ has a metric dimension bounded by $\widetilde{D}$ then this remains true for any subset of $\mathcal{S}$. Clearly, the set S_η provided by the previous definition satisfies Assumption $2\,(\eta, \widetilde{D}(\eta), 1)$. Some sort of a reciprocal is provided by (4.2) of Lemma 2, which implies that if an η-net S_η for $\mathcal{S}$ satisfies Assumption $2\,(\eta, D_\eta, 1)$, then one can bound the metric dimension of $\mathcal{S}$ by $\widetilde{D}(\eta) = 4D_\eta$. The difference is due to the fact that in Assumption 2 we only consider $t \in S_\eta$.

5.2 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 (1973, 1975) who introduced the following notion of *metric dimension* to measure the massiveness of a set. He defines the $D(\eta)$ metric dimension of $\mathcal{S}$ as the smallest number z such that any set in $\mathcal{S}$ with diameter $2x \geq 2\eta$ can be covered by no more than 2^z sets of diameter not larger than x. Birgé (1983, Assumption H1 p. 186) introduces 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 $j \geq j_0 \geq 1$ by $2^{jD(\eta)}$. This is actually very similar to (4.1), apart from the fact that in our assumption we replaced x^D by the less restrictive $\exp\left(x^2 D\right)$. The initial definitions by Le Cam and Birgé 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

$$|S_\eta \cap \mathcal{B}(t, r)| \leq (r/\eta)^{c_2 k} \quad \text{for all } r \geq 2\eta \ \text{ and } \ t \in \mathcal{S}.$$

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.

5.3 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 $\mathcal{S}$ satisfying Assumption $2\,(\eta, \widetilde{D}(\eta), 1)$. As a consequence, one can bound the minimax risk on $\mathcal{S}$ as soon as one can get a bound for its metric dimension. The following result is an immediate consequence of Theorem 2 (with $B = B' = 1$).

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*

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

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

$$\mathbb{E}_s\left[d^q(s, \hat{s})\right] \leq C(q)\kappa^q \left[d(s, \mathcal{S}) + \tilde{\eta}\right]^q \quad \text{for all } s \in M \text{ and } q \geq 1.$$

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

Such a theorem actually emphasizes an important property of T-estimators: 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 bounded risk with an extra $d^q(s,\mathcal{S})$ added term corresponding to a misspecification in the statistical model. This can be viewed as an important robustness property of T-estimators that is definitely not shared by the classical m.l.e. as was demonstrated in Section 2.1.3.

It is often possible to show, under some additional assumptions, that the upper bounds for the minimax risk provided by Theorem 4 have the correct order of magnitude when $\widetilde{D}$ has been chosen in a close to minimal way. Such lower bounds arguments have been developed in Birgé (1983, 1984b and 1986) and, more recently, in Yang and Barron (1999). We shall not insist on this here referring the interested reader to those papers.

One can of course, using the relevant values of a and κ, immediately translate the results of Theorem 4 to the i.i.d. and Gaussian settings as for Corollaries 2 and 3. The independent setting could be handled in the same way but we omit it for simplicity. For the i.i.d. setting, we essentially recover (in a slightly more general form), the results of Birgé (1983) (see his Proposition 3.1 and Corollary 2.6).

Corollary 4 *Assume we are in the i.i.d. setting and the subset $\mathcal{S}$ of all distributions on $\mathcal{X}$ has a metric dimension bounded by $\widetilde{D}$ with respect to either the Hellinger or the variation distance. Let $\eta_n = \inf\left\{\eta > 0 \,\middle|\, n\eta^2 \geq 6\widetilde{D}(\eta)\right\}$ (Hellinger case) or $\eta_n = \inf\left\{\eta > 0 \,\middle|\, n\eta^2 \geq 12\widetilde{D}(\eta)\right\}$ (variation case). Then one can build in each case a T-estimator $\hat{s}$ satisfying, for all $s \in M$ and $q \geq 1$,*

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

In particular $R(\mathcal{S},q) \leq C(q)\eta_n^q$.

In the Gaussian setting, we get

Corollary 5 *Assume we are in the Gaussian setting and $\mathcal{S} \subset \boldsymbol{l}_2(\mathbb{N}^\star)$ has a metric dimension bounded by $\widetilde{D}$ with respect to the distance d_2 corresponding to the norm in $\boldsymbol{l}_2(\mathbb{N}^\star)$. Let $\eta_\sigma = \inf\left\{\eta > 0 \,\middle|\, (\eta/\sigma)^2 \geq 36\widetilde{D}(\eta)\right\}$. Then one can build a T-estimator $\hat{s}$ satisfying for all $s \in M$ and $q \geq 1$,*

$$\mathbb{E}_s\left[\|s-\hat{s}\|^q\right] \leq C(q)\left[\eta_\sigma + d_2(s,\mathcal{S})\right]^q.$$

In particular $R(\mathcal{S},q) \leq C(q)\eta_\sigma^q$.

This last 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 Birgé (1983) via our Proposition 4.

5.4 A few typical illustrations

5.4.1 Finite dimensional normed linear spaces

One purpose of Approximation Theory (see, for instance Pinkus, 1985) is, given some function space $\mathcal{S}$, to provide k-dimensional linear spaces S_k such that $\sup_{s\in\mathcal{S}} d(s,S_k) \leq$

ε 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. One easy result in this direction is as follows.

Proposition 7 *Let S_k be a k-dimensional linear subspace of some normed linear space M. If M is a Hilbert space, then S_k has a finite metric dimension bounded by $k(\log 5)/4 < 0.403k$ (by 1/2 if $k = 1$). Otherwise, it is bounded by $k \log 5 < 5k/3$.*

Proof: It is based on the classical Lemma 4 below — see, for instance, Lemma 1 of Lorentz (1966) —. Indeed, there exists a maximal η-separated subset S of S_k, which is therefore an η-net for S_k and it satisfies, by (5.3),

$$|\{u \in S \,|\, \|t-u\| < r\}| < \exp\left[k(r/\eta)^2(\log 5)/4\right] \quad \text{for all } t \in S_k \text{ and } r \geq 2\eta. \quad (5.2)$$

If M is a Hilbert space, Pythagoras' Theorem implies that (5.2) still holds for $t \in M$. Otherwise, we apply Lemma 2 to get the bound. □

Lemma 4 *If S_k is a k-dimensional normed linear space and S an η-separated subset of S_k, then*

$$|\{u \in S \,|\, \|t-u\| \leq x\eta\}| \leq (2x+1)^k \quad \textit{for all } x > 0 \textit{ and all } t \in S_k. \quad (5.3)$$

As a straightforward consequence of Proposition 7, Theorem 4 implies that if $\mathcal{S}$ is any subset of a k-dimensional normed linear space (or more generally has a finite metric dimension bounded by k) the minimax quadratic risk $R(\mathcal{S}, 2)$ is bounded by $Ca^{-1}k$ (Ck/n in the i.i.d. setting and $C\sigma^2 k$ in the Gaussian setting, as expected).

It is easy to improve Bound 5.2 for Euclidean spaces of dimensions one and two, using an explicit construction for the nets. If $k = 1$, using the lattice $S = 2\eta\mathbb{Z}$, we can replace $(\log 5)/4$ by $(\log 3)/4$ in (5.2). If $k = 2$, we can cover $\mathbb{R}^2$ with hexagons and take for S their centers, which results in the dimension bound $(\log 7)/4$ instead of $(\log 5)/2$. The example of $k = 1$ shows that (5.2) is not too pessimistic. A lower-bound argument again based on volumes shows that, if S is an arbitrary η-net for $\mathbb{R}^k$, then $|S \cap \mathcal{B}(t, 3\eta)| \geq 2^k$. This is a rather poor but straightforward lower bound which implies that the metric dimension of $\mathbb{R}^k$ is, in any case, of the order of k. When M is a Hilbert space, one can also build η-nets for $S_k, k \geq 3$, in a constructive way but this results in the suboptimal bound $0.458k$ for the metric dimension.

Proposition 8 *Let $S \subset S_k \subset M$ be a lattice of the form $S = \left[\left(2\eta/\sqrt{k}\right)\mathbb{Z}\right]^k$, where S_k denotes a k-dimensional linear subspace (identified to $\mathbb{R}^k$) of some Hilbert space M and η is a positive number. Then S is an η-net for S_k and whatever $t \in M$,*

$$|S \cap \mathcal{B}(t, r)| < (\pi k)^{-1/2} \exp\left[0.458k(r/\eta)^2\right] \quad \textit{for } r \geq 2\eta.$$

The result follows immediately from the next lemma, which can be proved as Lemma 2 from Birgé and Massart (1998).

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

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

5.4.2 Parametric models

The initial construction of Le Cam (1973) 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, the following property, which appears in Lemma 7 of Le Cam (1973), holds:

Property 1 *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 $\mathbb{R}^k$, two positive constants $b \leq b'$ and an increasing function ξ satisfying $\xi(x\lambda) \leq x^\beta \xi(\lambda)$, $\beta > 0$, for all $x \geq 2$, $\lambda > 0$, such that*

$$b\|\theta - \theta'\| \leq \xi\left(d(t(\theta), t(\theta'))\right) \leq b'\|\theta - \theta'\| \quad \text{for all } \theta, \theta' \in \Theta. \tag{5.4}$$

One can then mimic the proof of Proposition 7 to get

Proposition 9 *Under Property 1, the subset $\mathcal{S}$ of (M, d) has a metric dimension bounded by $k[\log(b'/b) + \beta \log 5] \vee (1/2)$.*

Note that, for the i.i.d. setting, Property 1 can only hold for bounded sets Θ since d being either h or v is bounded.

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

Property 2 *A subset $\mathcal{V}$ of $\mathcal{S}$ enjoys Property 2 if, whatever $\eta, r > 0$, $t \in M$, and S_η an η-separated subset of $\mathcal{S}$,*

$$|\mathcal{V} \cap S_\eta \cap \mathcal{B}(t, r)| \leq \exp\left[D[(r/\eta) \vee 1]^2\right],$$

for some $D > 0$.

If $\mathcal{S}$ is a k-dimensional manifold, for any $s \in \mathcal{S}$, one can find a viciniy 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 7 that there exists a vicinity $\mathcal{V}_s$ of s which enjoys Property 2 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 $(s_n)_{n\geq 1}$ in $\mathcal{S}$ converging to s_0 and a sequence $(r_n)_{n\geq 1}$ converging to 0 such that $\mathcal{S} \cap \mathcal{B}(s_n, r_n)$ does not enjoy Property 2. This would contradict the fact that $\mathcal{V}_{s_0}$ enjoys Property 2 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 $\mathcal{S}$ via the following

Proposition 10 *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 2 with the same value of D, the metric dimension of $\mathcal{S}$ is bounded by*

$$\widetilde{D}(\eta) = \frac{D(\bar{r}/\eta)^2 + \log K}{4 \vee [\bar{r}/(2\eta)]^2} \vee 1/2, \tag{5.5}$$

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

Proof: If $\eta \geq \bar{r}$, 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 (5.5) holds. If $\eta < \bar{r}$, we take for S a maximal η-separated subset of $\mathcal{S}$ and it follows from Property 2 that $|S \cap \mathcal{B}(s, \bar{r})| \leq \exp\left[D(\bar{r}/\eta)^2\right]$

for all $s \in \mathcal{S}$, which implies that $|S| \le K \exp\left[D(\bar{r}/\eta)^2\right]$. Let t be an arbitrary point in M. We now distinguish between two cases. If $2\eta \le r < \bar{r}/2$, 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 $s \in \mathcal{S}$, hence $|S \cap \mathcal{B}(s,2r)| \le \exp\left[4D(r/\eta)^2\right]$ by Property 2 and (5.5) holds since $\bar{r} > 4\eta$. If $r \ge (\bar{r}/2) \vee 2\eta$, then

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

and (5.5) holds again. □

5.4.3 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 5 *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 nonincreasing 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 Kolmogorov and Tikhomirov (1961), Lorentz (1966), Birman and Solomjak (1967) and Lorentz, von Golitschek and Makovoz (1996) 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 $M = \mathbb{R}^k$ 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.

5.4.4 Ellipsoids

Let us now consider a situation where the function $\widetilde{D}(\eta)$ converges to infinity when η goes to 0. Given a nonincreasing sequence $\boldsymbol{a} = (a_i)_{i\ge1}$ 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

$$\mathcal{E}(\boldsymbol{a}) = \left\{ s = (s_i)_{i\ge1} \;\middle|\; \sum_{i=1}^{+\infty} \left(\frac{s_i}{a_i}\right)^2 \le 1 \right\}, \tag{5.6}$$

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 $\sum_{i=k+1}^{+\infty} s_i^2 \le a_{k+1}^2$ for $k \ge 0$ 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 $\ge 1/2$ for $\eta \ge a_1$. For $\eta \ge \sqrt{2}a_{k+1}$ with $k \ge 1$, we set $\lambda = \eta\sqrt{2/k} \ge 2a_{k+1}k^{-1/2}$ 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

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

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

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

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

$$\widetilde{D}(\eta) = 0.52 \inf \left\{ k \in \mathbb{N}^\star \;\middle|\; \sqrt{2} a_{k+1} \leq \eta \right\}. \tag{5.7}$$

We can then derive from Corollary 5 and Lemma 6 below the following upper bounds for the minimax risk over $\mathcal{E}(\boldsymbol{a})$:

$$R(\mathcal{E}(\boldsymbol{a}), q) \leq C(q) \left(\inf_{k \geq 1} \left\{ a_{k+1} \vee \sigma \sqrt{k} \right\} \right)^q \quad \text{for } q \geq 1. \tag{5.8}$$

This bound is similar to the one we got for the i.i.d. setting in Birgé (1983, 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, Liu and MacGibbon (1990) — see also Birgé and Massart (2001, Section 6.2) —. An exact asymptotic evaluation of the minimax risk over ellipsoids has been given by Pinsker (1980).

Lemma 6 *Let $(b_k)_{k \geq 1}$ be some nonincreasing 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 $G(x) = \inf\{k \geq 1 \,|\, b_k \leq x\}$. Then, for all $t > 0$,*

$$\inf \left\{ x \;\middle|\; x^2 \geq t G(x) \right\} = \inf_{k \geq 1} \left(b_k \vee \sqrt{tk} \right).$$

The elementary proof will be ommitted.

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 4.2.2. It is then of common practice to put some Sobolev-type restriction on the unknown s, of the form $\|s^{(\alpha)}\| \leq R$. This amounts to assume that s belongs to the ellipsoid $\mathcal{E}(\boldsymbol{a}) = \mathcal{E}'(\alpha, R)$ defined by (5.6) with coefficients

$$a_1 = +\infty \qquad \text{and} \qquad a_{2j} = a_{2j+1} = R(2\pi j)^{-\alpha} \quad \text{for } j \geq 1. \tag{5.9}$$

We refer to Birgé and Massart (2001, Section 1.1.4) for additional details. This and (5.7) lead to the following upper bound $\widetilde{D}$ for the metric dimension of $\mathcal{E}'(\alpha, R)$:

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

or equivalently,

$$\frac{\widetilde{D}_{\alpha,R}(\eta)}{0.52} = 2 \left\lceil \frac{1}{2\pi} \left(\frac{\sqrt{2} R}{\eta} \right)^{1/\alpha} \right\rceil - 1 \quad \text{with } \lceil x \rceil = \inf\{n \in \mathbb{N} \,|\, n \geq x\}. \tag{5.10}$$

The resulting upper bound for the minimax risk then derives from (5.8):

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

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

5.5 The importance of choosing the "right" distance

One should be very careful with the application of the previous results and definitely not take for granted that *if the parameter set has a bounded metric dimension, then the minimax risk is under control*. Such results have been proved here for the distance d_2 in the Gaussian setting and for the Hellinger or the variation distances in the i.i.d. setting. They can be extended to other situations but are definitely not valid in general.

It is, for instance, quite common to use the squared $\mathbb{L}_2$-loss when estimating densities with respect to Lebesgue measure. This point of view has been criticized in particular by Devroye — see Devroye (1987) — who advertises for the $\mathbb{L}_1$-loss. The main point is that the $\mathbb{L}_2$-distance is not a distance between probability measures and depends on the choice of the dominating measure. The following example shows that there is no hope to build a theory of the risk with $\mathbb{L}_2$-loss under purely metric assumptions because Assumption 1 does not hold generally in the i.i.d. setting when we take for d the $\mathbb{L}_2$-distance.

Proposition 11 *Let us consider the parameter set $\mathcal{S} = (0, 1/3]$ and, for $s \in \mathcal{S}$, define the distribution $\overline{P}_s$ by its density with respect to Lebesgue measure on $[0, 1]$:*

$$d\overline{P}_s/dx = f_s = s^{-2}\mathbb{1}_{[0,s^3]} + \left(s^2 + s + 1\right)^{-1}\mathbb{1}_{(s^3,1]}.$$

Suppose we have at disposal n i.i.d. observations with density f_s, $s \in \mathcal{S}$. Then one can build for each $n \geq 1$ a T-estimator $\hat{s}_n$ such that

$$\sup_{n\geq 1}\sup_{s\in\mathcal{S}} \mathbb{E}_s\left[nh^2(s, \hat{s}_n)\right] < +\infty.$$

On the other hand, although the metric dimension of $\mathcal{S}$ with respect to the $\mathbb{L}_2$-distance is finite, $\sup_{s\in\mathcal{S}} \mathbb{E}_s\left[\|s - \tilde{s}_n\|^2\right] = +\infty$, whatever n and the estimator $\tilde{s}_n$.

6 Working with several models

6.1 Why do we need several models

The limitation of our previous approach which amounts to base our estimation procedure on a single well-chosen discrete model S appears clearly in the applications. When we tried to estimate a uniform distribution on $[0, e^t]$ in Section 4.2.5, we had to restrict to a compact set of values for t and when we designed a minimax (up to constants) estimator for the ellipsoid $\mathcal{E}'(\alpha, R)$ in Section 5.4.4, we had to know the values of α and R in order to build the estimator. It would clearly be more satisfactory to be able to use several models simultaneously in the construction of our estimator. 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.

In order to motivate our new construction, let us consider, in the Gaussian setting, a somewhat extreme, but nevertheless instructive example. Let us suppose that we hesitate between two assumptions for the unknown parameter s. The simplest one,

which is likely to be true but far from certain is that s belongs to some one-dimensional linear space $\mathcal{S}_1$. The second assumption, which is certain, is $s \in \mathcal{S}_N$ where $\mathcal{S}_N \supset \mathcal{S}_1$ is a linear space of large dimension N. If the first assumption is true, it follows from the considerations of Section 5.4.1 that we could base our construction on a discrete subset of $\mathcal{S}_1$ satisfying Assumption 2 with $D = 1/2, B' = 1$ and get, by Corollary 5, a risk bound of the form

$$\mathbb{E}_s \left[\|s - \hat{s}\|^2 \right] \leq C \left[d_2^2(s, \mathcal{S}_1) + \sigma^2 \right].$$

This is perfect if s actually belongs to $\mathcal{S}_1$ or is close enough to it, but may be otherwise terrible. On the other hand, working with a suitable discrete subset of $\mathcal{S}_N$ would lead to the risk bound $CN\sigma^2$ whatever $s \in \mathcal{S}_N$, which is not satisfactory if s does belong to $\mathcal{S}_1$. One possible way of solving this dilemna would be to test between the two assumptions but this becomes much harder if we have to choose between many arbitrary and possibly non-linear sets with bounded metric dimensions, instead of two linear spaces. Moreover, since the very definition of T-estimators includes testing, it seems natural to incorporate the tests between the different parameter spaces in the construction of the estimator. In order to understand how to do this, let us go back to our example, assuming that $s \in \mathcal{S}_1 \subset \mathcal{S}_N$. Let us denote by S_1 and S_N the discrete subsets of $\mathcal{S}_1$ and $\mathcal{S}_N$ respectively that we use to build the T-estimators under each assumption and assume, for simplicity, that s belongs to both sets S_1 and S_N. If we want to build an estimator, as we did before, using both sets together, although $s \in S_1$, there are so many points in S_N that are at a distance of order $\sigma\sqrt{N}$ of s that there is a large probability that one of the tests between s and one of those points does reject s and that the resulting estimator be at a distance of order $\sigma\sqrt{N}$ of s. What happens here is exactly what happens with the m.l.e. in Section 2.2. The fact that, because of its large dimension, there are many points of S_N around any point of S_1 gives an advantage to S_N. We actually need to compensate for this advantage. Instead of always using tests between t and u with a threshold 0, considering the sign of $\log[t(\boldsymbol{X})/u(\boldsymbol{X})]$, we may want to give an advantage to t when it belongs to a set of small (metric) dimension and u belongs to a set of larger dimension. It suffices for this to change the threshold and accept t when $\log[t(\boldsymbol{X})/u(\boldsymbol{X})] > x$ with $x < 0$. This will change the equilibrium between the two errors of the tests and compensate for the fact that there are many more points in the bigger set. This is actually how penalized maximum likelihood works: in order to use simultaneously many sets S_m corresponding to different dimensions, it attaches to each one a penalty $\mathrm{pen}(m)$. Performing penalized maximum likelihood then amounts to choosing an estimator $\tilde{s}$ satisfying

$$\tilde{s} = \hat{s}_{\hat{m}} \quad \text{and} \quad \log \tilde{s}(\boldsymbol{X}) - \mathrm{pen}(\hat{m}) = \sup_{m \in \mathcal{M}} \sup_{t \in S_m} \{\log t(\boldsymbol{X}) - \mathrm{pen}(m)\}.$$

By doing so, when we compare the penalized likelihoods at $t \in S_m$ and $u \in S_{m'}$, we compare $\log t(\boldsymbol{X}) - \mathrm{pen}(m)$ with $\log u(\boldsymbol{X}) - \mathrm{pen}(m')$ or equivalently $\log[t(\boldsymbol{X})/u(\boldsymbol{X})]$ with $\mathrm{pen}(m) - \mathrm{pen}(m')$ and if $\mathrm{pen}(m)$ is an increasing function of the dimension, we do exactly what was expected: compensate for the dimension.

Of course, the penalized m.l.e. 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 Barron, Birgé and Massart (1999), Castellan (1999 and 2000), van

de Geer (2000) or Eggermont and LaRiccia (2001). Even in the Gaussian setting, it is difficult to get general results for the penalized m.l.e. with arbitrary models. 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.

6.2 The new assumptions

The set up we use here is the one we defined in Section 4.1 with the only difference that S is now written as a finite or countable union of countable subsets S_m of M: $S = \bigcup_{m\in\mathcal{M}} S_m$, the set $\{S_m, m \in \mathcal{M}\}$ representing a family of possible models for the unknown s.

Our first assumption bounds the errors of the tests $\psi(t, u, \boldsymbol{X})$ but not in the same way as in Section 4.1. In Assumption 1, t and u were treated in a symmetric way in the sense that both errors of the tests were bounded by the same quantity $B\exp\left[-ad^2(t, u)\right]$. We now want to introduce, via some function η from S to $\mathbb{R}^+$, a dissymmetry in our tests and be able to favour one of the two points. To do this, we attach to each point t in S a penalty $\eta^2(t)$ and, when testing between t and u favour the point with the smaller penalty, which means that the error at this point will be smaller than the other error.

Assumption 3 *There exists a function δ from $M \times S$ to $[0, +\infty]$ such that $\delta(s, t) \geq \kappa d(s, t)$ for some $\kappa \geq 4$ and two constants $a, B > 0$ such that, given a non-negative function η on S, one can find for each pair $(t, u) \in S^2$ with $t \neq u$, a test function $\psi(t, u, \boldsymbol{X})$ (with $\psi(u, t, \boldsymbol{X}) = 1 - \psi(t, u, \boldsymbol{X})$) which satisfies*

$$\sup_{\{s\in M\,|\,\delta(s,t)\leq d(t,u)\}} \mathbb{P}_s[\psi(t, u, \boldsymbol{X}) = 1] \leq B\exp\left[-a\left(d^2(t, u) - \eta^2(t) + \eta^2(u)\right)\right]. \quad (6.1)$$

Note here that Assumption 1 merely corresponds to the special situation $\eta(t) = 0$ for all $t \in S$. We can immediately derive from Propositions 4 and 5, that this assumption is satisfied in the Gaussian and independent settings.

The metric properties of S are described by the following generalization of Assumption 2.

Assumption 4 *There exists $B' > 0$ such that, for each $m \in \mathcal{M}$, one can find $\eta_m > 0$ and $D_m \geq 1/2$ satisfying*

$$|\{S_m \cap \mathcal{B}_d(t, r)\}| \leq B'\exp\left[D_m[(r/\eta_m) \vee 2]^2\right] \quad \textit{for all } r > 0 \textit{ and } t \in S. \quad (6.2)$$

The assumption implies that S is countable and that each S_m satisfies Assumption 2 (η_m, D_m, B'). In particular, Lemma 2 applies to each S_m. Conversely, if each S_m satisfies Assumption 2 (η_m, D_m, B') and d is a distance, Assumption 4 holds with D_m replaced by $4D_m$ in (6.2) by Lemma 2.

6.3 The estimator and its performances

As before, we still want to define T-estimators via the minimization of the function $\mathcal{D}_{\boldsymbol{X}}$ given by (3.4) as explained in Definition 2 but we first have to specify which penalty function η we shall use to build our family of tests. We want to connect it with the η_ms of Assumption 4 so that $\eta(t) = \eta_m$ when t belongs to S_m. Unfortunately,

the previous sentence is definitely ambiguous because t may belong to several S_ms simultaneously, which leads to the following precise definition:

$$\eta(t) = \inf\{\eta_m \,|\, m \in \mathcal{M} \text{ and } t \in S_m\}. \tag{6.3}$$

From now on, we shall assume that η is given by (6.3).

We can now prove a general result about the existence and performances of T-estimators which is the extension to the case of several models of Theorem 2. The relationship between D and $a\eta^2$ should now hold for each model S_m and uniformly in m, but we now require an additional assumption, namely (6.6) below, which bounds the "complexity" of our family of models in the sense that it controls the growth of the sequence of numbers $N_j = \left|\left\{m \in \mathcal{M} \,\middle|\, j-1 \le \eta_m^2 < j\right\}\right|$. This condition should be viewed as an analogue of (3.1) in Barron and Cover (1991), (19) in Birgé and Massart (1997), (2.2) in Barron, Birgé and Massart (1999) or (3.3) in Birgé and Massart (2001). In particular it implies that

$$|\{m \in \mathcal{M} \,|\, \eta_m \le z\}| < +\infty \quad \text{for all } z > 0. \tag{6.4}$$

Theorem 5 *Let (M,d) be a metric space and Assumptions 3 and 4 hold with $\delta = \kappa d$,*

$$a\eta_m^2 \ge 21 D_m/5 \quad \textit{for all } m \in \mathcal{M} \tag{6.5}$$

and

$$\sum_{m\in\mathcal{M}} \exp\left[-a\eta_m^2/21\right] = \Sigma < +\infty. \tag{6.6}$$

Let s be an arbitrary element of M. The function $\mathcal{D}_{\boldsymbol{X}}$ is $\mathbb{P}_s$-a.s. finite on S and there exits T_ε-estimators $\hat{s}(\boldsymbol{X})$ for $\varepsilon > 0$ and also for $\varepsilon = 0$ if $\mathcal{M}$ is finite. If $0 \le \varepsilon \le 0.48 a^{-1/2}$ the risk of any of them satisfies, for $q \ge 1$, both bounds

$$\mathbb{E}_s[d^q(s,\hat{s})] \le \left[\kappa + \frac{4}{3}\right]^q \inf_{m\in\mathcal{M}} \left\{d(s,S_m) \vee \eta_m\right\}^q + \frac{BB'\Sigma}{7}\sqrt{\frac{\pi e q}{2}}\left[\frac{4q}{3e}\right]^{q/2} a^{-q/2}; \tag{6.7}$$

$$\mathbb{E}_s[d^q(s,\hat{s})] \le \left[\kappa + \frac{4}{3}\right]^q \inf_{m\in\mathcal{M}} \left\{\left[d(s,S_m) \vee \eta_m\right]^q \left[1 + \frac{BB'\Sigma}{7}\zeta_q\left(\frac{2\kappa^2 a\eta_m^2}{3}\right)\right]\right\}, \tag{6.8}$$

with ζ_q given by (4.6). In particular,

$$\mathbb{E}_s[d^q(s,\hat{s})] \le \left[1 + \frac{BB'\Sigma}{10^8}\right]\left[\kappa + \frac{4}{3}\right]^q \inf_{m\in\mathcal{M}} \{d(s,S_m) \vee \eta_m\}^q \quad \textit{for } 1 \le q \le 70. \tag{6.9}$$

The risk bounds (6.7) and (6.8) are the analogues of those we got in Theorem 2 with the terms involving $BB'\Sigma$ playing the role of either additive of multiplicative remainder terms. On the one hand, the additive term in (6.7) does not involve η_m or D_m and behaves as $a^{-q/2}$. On the other hand, even for values of q as large as 70, the term involving $BB'\Sigma$ in (6.8) should be considered as negligeable unless $BB'\Sigma$ is really huge, as seen from (6.9). Indeed, for the typical value $BB' = 1$, one can allow very large values of Σ without any noticeable effect on the risk bound (6.8). The choice of 70 has nothing special apart from the facts that it is already larger than the powers involved in risk functions one would currently use, that it leads to a simple bound and that the remainder term increases rather quickly for larger values

of q. Therefore, we shall restrict ourselves to $1 \leq q \leq 70$ in the sequel for simplicity. Of course, all the results could be extended to larger values of q, starting from either (6.7) or (6.8).

As a consequence of Theorem 5, we can build T-estimators from a suitable discretization of some collection of approximating models such those provided by Approximation Theory. The following result is easily comparable to more classical ones about the performances of penalized estimators as in Barron, Birgé and Massart (1999) or Birgé and Massart (2001).

Corollary 6 *Let Assumption 3 hold with $\delta = \kappa d$ and $\left\{\overline{S}_m\right\}_{m\in\mathcal{M}}$ 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 nonnegative weights such that*

$$\sum_{m\in\mathcal{M}} \exp[-\Delta_m/5] = \Sigma. \tag{6.10}$$

There exists a T-estimator $\hat{s}$ satisfying, for all $s \in M$ and $1 \leq q \leq 70$,

$$\mathbb{E}_s\left[d^q(s,\hat{s})\right] \leq \left[1+\frac{B\Sigma}{10^8}\right]\left[\kappa+\frac{4}{3}\right]^q \inf_{m\in\mathcal{M}}\left\{d\left(s,\overline{S}_m\right)+\sqrt{\frac{21}{5a}\left(\overline{D}_m \vee \Delta_m\right)}\right\}^q. \tag{6.11}$$

Proof: For each m, set $\eta_m^2 = (21/5)a^{-1}\left(\overline{D}_m \vee \Delta_m\right)$. Then (6.6) holds and (6.5) is satisfied with $D_m = \overline{D}_m$ since $\overline{D}_m \geq 1/2$ by the definition of the metric dimension. Moreover, this definition also implies that, for each $m \in \mathcal{M}$, there exists $S_m \subset M$ which is an η_m-net for $\overline{S}_m$ and satisfies Assumption 4 with $D_m = \overline{D}_m$ and $B' = 1$. We may therefore apply Theorem 5 to get (6.9). We conclude by noticing that $d(s,S_m) \vee \eta_m \leq d\left(s,\overline{S}_m\right) + \eta_m$. □

Remark: One could, alternatively, adopt a Bayesian point of view, introducing some prior distribution ν on $\mathcal{M}$ with $\nu_m = -\log(\nu(\{m\})) > 0$ for each m. Setting $\Delta_m/5 = \nu_m - 16$ leads to $\Sigma = e^{16} < 10^8/11$ and (6.11) becomes

$$\mathbb{E}_s\left[d^q(s,\hat{s})\right] \leq \left[1+\frac{B}{11}\right]\left[\kappa+\frac{4}{3}\right]^q \inf_{m\in\mathcal{M}}\left\{d\left(s,\overline{S}_m\right)+\sqrt{\frac{21}{a}\left[\frac{\overline{D}_m}{5} \vee (\nu_m - 16)\right]}\right\}^q.$$

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 Birgé and Massart (2001) for penalized least squares.

Finally, we shall consider the special situation where the tests $\psi(t,u,\boldsymbol{X})$ of Assumption 3 derive from some contrast function γ, as in (4.14). This occurs in particular for the Gaussian setting as we shall see below.

Theorem 6 *Let d be a distance, Assumption 4 hold and, for all $(t,u) \in S^2$, $t \neq u$,*

$$\sup_{\{s\in M \,|\, d(s,t)\leq\kappa^{-1}d(t,u)\}} \mathbb{P}_s[\gamma(u,\boldsymbol{X}) \leq \gamma(t,\boldsymbol{X})] \leq B\exp\left[-a\left(d^2(t,u)-\eta^2(t)+\eta^2(u)\right)\right], \tag{6.12}$$

for some fixed $a, B > 0$ and $\kappa \geq 4$. Assume, moreover, that, whatever $s \in M$,

$$\mathbb{P}_s[\gamma(u,\boldsymbol{X}) \leq \gamma(t,\boldsymbol{X})] \leq B\exp\left[a\left(\kappa' d^2(s,t)+\eta^2(t)-\eta^2(u)\right)\right], \tag{6.13}$$

for some $\kappa' > 0$. Let the constants η_m, D_m and a satisfy (6.5) and

$$\sum_{m \in \mathcal{M}} \exp\left[-a\kappa'' \eta_m^2\right] < +\infty \quad \textit{with } 0 \leq \kappa'' < 1/21. \tag{6.14}$$

Then, $\mathbb{P}_s$-a.s., there exists at least one point $\hat{s} \in S$ such that $\gamma(\hat{s}, \boldsymbol{X}) = \inf_{t \in S} \gamma(t, \boldsymbol{X})$ and any such point satisfies (6.7), (6.8) and (6.9).

6.4 Proofs of Theorems 5 and 6

6.4.1 Some preliminary results

As in Section 4.3, we shall first prove some intermediate results of large deviations for $\mathcal{D}_{\boldsymbol{X}}$ which only require that d be a semi-metric.

Proposition 12 *Let Assumptions 3 and 4 hold with the penalty function η given by (6.3) and constants a, η_m and D_m satisfying (6.5) and (6.6). If $s \in M$ and $s' \in S$ are such that $\delta(s, s') < +\infty$, then the function $\mathcal{D}_{\boldsymbol{X}}$ defined on S by (3.4) satisfies*

$$\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') > y\right] \leq \frac{BB'\Sigma}{7} \exp\left[-\frac{2ay^2}{3}\right] \quad \textit{for } y \geq y_0 = [4\eta(s')] \vee \delta(s, s'). \tag{6.15}$$

If, moreover, whatever $s \in M$ and $t \neq u \in S$,

$$\mathbb{P}_s[\psi(t, u, \boldsymbol{X}) = 1] \leq B \exp\left[a\left(\kappa' d^2(s,t) + \eta^2(t) - \eta^2(u)\right)\right], \tag{6.16}$$

for some $\kappa' > 0$ and (6.14) holds, then

$$\alpha(y) = \mathbb{P}_s\left[\ \exists\, t \in S \textit{ with } \psi(s', t, \boldsymbol{X}) = 1 \textit{ and } \eta(t) \geq y\right] \underset{y \to +\infty}{\longrightarrow} 0. \tag{6.17}$$

Proof: Using (6.4) we can index the set $\mathcal{M}$ as $\{m_j, j \in \mathbb{N}\}$ in nonincreasing order of the η_m's, so that $i < j$ implies $\eta_{m_i} \leq \eta_{m_j}$. Then we derive from the S_ms a partition $\{S'_m, m \in \mathcal{M}\}$ of S by setting, using this indexing of $\mathcal{M}$,

$$S'_{m_k} = S_{m_k} \bigcap \left(\bigcup_{j<k} S_{m_j}\right)^c \quad \text{for all } k \in \mathbb{N}. \tag{6.18}$$

Since $S'_m \subset S_m$, the S'_ms still satisfy Assumption 4 with the same constants η_m, D_m and B'. Moreover, our ordering of $\mathcal{M}$ has been chosen in such a way that whatever $t \in S$, $\eta(t) = \eta_p$ if $t \in S'_p$, therefore we derive from (3.4) and (6.1) that

$$\begin{aligned}
\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') > y\right] &\leq \mathbb{P}_s\left[\ \exists\, t \in S \text{ with } d(t, s') \geq y \text{ and } \psi(s', t, \boldsymbol{X}) = 1\right] \\
&\leq \sum_{p \in \mathcal{M}} \sum_{\substack{t \in S'_p \\ d(t,s') \geq y}} \mathbb{P}_s[\psi(s', t, \boldsymbol{X}) = 1] \\
&\leq B \sum_{p \in \mathcal{M}} \sum_{\substack{t \in S'_p \\ d(t,s') \geq y}} \exp\left[-a\left(d^2(t, s') - \eta(s')^2 + \eta_p^2\right)\right], \quad (6.19)
\end{aligned}$$

since $d(t,s') \geq y \geq \delta(s,s')$. To prove (6.15), it is then enough, by (6.6), to show that

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geq y}} \exp\left[-a\left(d^2(t,s')-\eta^2(s')+\eta_p^2\right)\right] \leq \frac{B'}{7}\exp\left[-\frac{2ay^2}{3}-\frac{a\eta_p^2}{21}\right]. \tag{6.20}$$

To prove (6.20), we shall systematically use the inequalities

$$ay^2 \geq ay_0^2 \geq 16a\eta^2(s') \geq 16\times 21 D_m/5 \geq 168/5, \tag{6.21}$$

which follow from (6.5) when $s' \in S'_m$, and distinguish between two cases.

Case 1: $y \geq 2\eta_p$
Let us observe that, if $z \geq 2\eta_p$, Assumption 4 and (6.5) imply that

$$\left|\left\{t\in S'_p \;\middle|\; d(t,s')<z\right\}\right| \leq B'\exp\left[(z/\eta_p)^2 D_p\right] \leq B'\exp\left[5az^2/21\right].$$

Setting $\theta = 91/80$, we derive from this bound and (6.21) that

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

Hence, by (6.21),

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geq y}} \exp\left[-a\left(d^2(t,s')-\eta^2(s')+\eta_p^2\right)\right] < 1.036B'\exp\left[-a\left(2y^2/3+\eta_p^2\right)\right]. \tag{6.22}$$

Case 2: $y < 2\eta_p$
In this case we split the sum in (6.19) into two parts. For $d(t,s') \geq 2\eta_p$, we can apply the results of Case 1, i.e. (6.22) with y replaced by $2\eta_p$ and then the assumption $y < 2\eta_p$, to get

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geq 2\eta_p}} \exp\left[-a\left(d^2(t,s')-\eta^2(s')+\eta_p^2\right)\right] < 1.036B'\exp\left[-a\left(2y^2/3+\eta_p^2\right)\right].$$

For $y \leq d(t,s') < 2\eta_p$, we use (6.5) and (6.21) again to derive

$$\begin{aligned}
\sum_{\substack{t\in S'_p \\ y\leq d(t,s')<2\eta_p}} \exp\left[-a\left(d^2(t,s')-\eta^2(s')+\eta_p^2\right)\right] &\leq B'\exp\left[4D_p - ay^2 + a\eta^2(s') - a\eta_p^2\right] \\
&\leq B'\exp\left[-a\left(15y^2/16+\eta_p^2/21\right)\right].
\end{aligned}$$

Adding the sums for $d(t,s') \geq 2\eta_p$ 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, whatever $y \geq y_0$,

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geq y}} \exp\left[-a\left(d^2(t,s') - \eta^2(s') + \eta_p^2\right)\right]$$
$$\leq \quad B' \exp\left[-\frac{2ay^2}{3} - \frac{a\eta_p^2}{21}\right]\left(1.036\exp\left[-\frac{20a\eta_p^2}{21}\right] + \exp\left[-\frac{13ay^2}{48}\right]\right)$$

and (6.20) follows from (6.21).

To prove (6.17), we introduce the set $\mathcal{M}_y = \{p \in \mathcal{M} \,|\, \eta_p \geq y\}$. Since $\eta(t) = \eta_p$ if $t \in S'_p$,

$$\alpha(y) \leq \sum_{p\in\mathcal{M}_y} \sum_{\substack{t\in S'_p \\ d(t,s')\geq 2\eta_p}} \mathbb{P}_s[\psi(s',t,\boldsymbol{X}) = 1] + \sum_{p\in\mathcal{M}_y} \sum_{\substack{t\in S'_p \\ d(t,s')< 2\eta_p}} \mathbb{P}_s[\psi(s',t,\boldsymbol{X}) = 1].$$

Since $\eta_p \geq y$, we can bound the first term using (6.20):

$$\sum_{\substack{t\in S'_p \\ d(t,s')\geq 2\eta_p}} \mathbb{P}_s[\psi(s',t,\boldsymbol{X}) = 1] \leq \frac{BB'}{7}\exp\left[-\left(8ay^2/3\right) - \left(a\eta_p^2/21\right)\right]. \tag{6.23}$$

For the second term, we use (6.16), (6.5), $d(s,s') \leq \kappa^{-1}\delta(s,s') \leq y_0/4$ and (6.21) to get

$$\sum_{\substack{t\in S'_p \\ d(t,s')< 2\eta_p}} \mathbb{P}_s[\psi(s',t,\boldsymbol{X}) = 1]$$
$$\begin{aligned} &\leq BB'\exp\left[4D_p + a\kappa' d^2(s,s') + a\eta(s')^2 - a\eta_p^2\right] \\ &\leq BB'\exp\left[-\left(a\eta_p^2/21\right) + \left(ay_0^2/16\right)(\kappa'+1)\right] \\ &< BB'\exp\left[-a\kappa''\eta_p^2 - ay^2(1/21-\kappa'') + \left(ay_0^2/16\right)(\kappa'+1)\right]. \end{aligned} \tag{6.24}$$

Summing (6.23) and (6.24) with respect to $p \in \mathcal{M}_y$ and letting y go to infinity allows to deduce (6.17) from (6.14). □

Remark: Since by (6.21), $2ay^2/3 \geq 112/5$, the right-hand side of (6.15) is bounded by $2.7 \times 10^{-11} BB'\Sigma$. Therefore, unless $BB'\Sigma$ is very large, (6.15) essentially says that $\mathbb{P}_s\left[\mathcal{D}_{\boldsymbol{X}}(s') \leq [4\eta(s')] \vee \delta(s,s')\right] \simeq 1$.

6.4.2 Proof of Theorem 5

The assumptions imply, by Proposition 12, that (6.15) holds for $y \geq y_0$ and, therefore, $\mathcal{D}_{\boldsymbol{X}}(s') < +\infty$, $\mathbb{P}_s$-a.s. for any $s' \in S$, which proves the existence of T_ε-estimators $\hat{s}(\boldsymbol{X})$, at least for $\varepsilon > 0$. Moreover, by Lemma 1,

$$\left\{t \in S \,|\, \mathcal{D}_{\boldsymbol{X}}(t) \leq \mathcal{D}_{\boldsymbol{X}}(s')\right\} \subset \bigcup_{m\in\mathcal{M}} \left[S_m \cap \overline{\mathcal{B}}(s', \mathcal{D}_{\boldsymbol{X}}(s'))\right].$$

When $\mathcal{M}$ is finite, this is a.s. a finite set since $S_m \cap \overline{\mathcal{B}}(s', \mathcal{D}_{\boldsymbol{X}}(s'))$ is finite for each m and there exists at least one T_0-estimator.

Let us now fix some $m \in \mathcal{M}$ and set $\overline{y} = \kappa[d(s, S_m) \vee \eta_m]$ and $s' = \pi_m(s)$ where π_m is a "projection" operator from M to S_m provided by Lemma 2 via Assumption 4. Then $d(s, s') = d(s, S_m) \leq \overline{y}/\kappa$, $\eta(s') \leq \eta_m$, $\overline{y} \geq y_0$ and $\overline{y}/\kappa \geq \eta_m \geq \sqrt{2.1/a}$ by (6.21), hence $\varepsilon < \overline{y}/(3\kappa)$. Then (3.5) implies that

$$d(s, \hat{s}(\boldsymbol{X})) \leq d(s, s') + \mathcal{D}_{\boldsymbol{X}}(s') + \varepsilon \leq \mathcal{D}_{\boldsymbol{X}}(s') + 4\overline{y}/(3\kappa).$$

We now use (6.15) and apply Proposition 3, with $Y = \mathcal{D}_{\boldsymbol{X}}(s')$, $\alpha = BB'\Sigma/7$, $\beta = 2a/3$ and $\lambda = 4/(3\kappa)$ to get the analogues of (4.4) and (4.5). We then argue as in the proof of Theorem 2, using $(3/4)[1 + 4/(3\kappa)]^2 < 4/3$ in (4.5) and $\overline{y} \geq \kappa\eta_m$ in (4.4). An optimization with respect to m, which is arbitrary in $\mathcal{M}$, then leads to (6.7) and (6.8). Since $2\kappa^2 a\eta_m^2/3 \geq 224D_m/5 \geq 22.4$ whatever m, (6.9) follows from a study of the function $q \mapsto \zeta_q(22.4)$.

6.4.3 Proof of Theorem 6

We first want to prove that there exists a.s. some $\hat{s}(\boldsymbol{X}) \in S$ such that $\gamma(\hat{s}, \boldsymbol{X}) = \inf_{t \in S} \gamma(t, \boldsymbol{X})$. Let us define the tests $\psi(t, u, \boldsymbol{X})$, for all $(t, u) \in S^2$, $t \neq u$, by

$$\psi(t, u, \boldsymbol{X}) = \begin{cases} 0 & \text{if } \gamma(t, \boldsymbol{X}) < \gamma(u, \boldsymbol{X}); \\ 1 & \text{if } \gamma(t, \boldsymbol{X}) > \gamma(u, \boldsymbol{X}); \end{cases} \tag{6.25}$$

(with an arbitrary value in case of equality) and fix some $s' \in S$. Replacing $\psi(t, u, \boldsymbol{X}) = 1$ by $\gamma(u, \boldsymbol{X}) \leq \gamma(t, \boldsymbol{X})$ in the proof of Proposition 12, we can conclude in the same way that $\mathcal{D}_{\boldsymbol{X}}(s') < +\infty$ a.s. and (6.17) holds. If s' is a minimizer of γ, there is nothing to prove. Otherwise such a minimizer should belong to the nonempty set $\{t \in S \,|\, \gamma(t, \boldsymbol{X}) < \gamma(s', \boldsymbol{X})\}$. If t belongs to this set, $\mathcal{D}_{\boldsymbol{X}}(s') \geq d(s', t)$. Hence, setting $\mathcal{M}'_y = \{m \in \mathcal{M} \,|\, \eta_m < y\}$, we get

$$\{t \in S \,|\, \gamma(t, \boldsymbol{X}) < \gamma(s', \boldsymbol{X}) \text{ and } \eta(t) < y\} \subset \bigcup_{m \in \mathcal{M}'_y} \left[S_m \cap \overline{\mathcal{B}}(s', \mathcal{D}_{\boldsymbol{X}}(s'))\right]. \tag{6.26}$$

The set $S_m \cap \overline{\mathcal{B}}(s', \mathcal{D}_{\boldsymbol{X}}(s'))$ is finite for each m and $\mathcal{M}'_y$ as well for any $y > 0$ by (6.4). This implies that, with a probability at least $1 - \alpha(y)$, the set of ts such that $\gamma(t, \boldsymbol{X}) < \gamma(s', \boldsymbol{X})$ is finite and there exists some minimizer $\hat{s}(\boldsymbol{X})$ of $\gamma(\cdot, \boldsymbol{X})$. Letting y go to infinity, we conclude from (6.17) that $\hat{s}(\boldsymbol{X})$ exists with probability one.

Now, for any such minimizer $\hat{s}$, we have, following again the proof of Proposition 12,

$$\begin{aligned} \mathbb{P}_s[d(\hat{s}, s') \geq y] &\leq \mathbb{P}_s\left[\, \exists\, t \in S \text{ with } d(t, s') \geq y \text{ and } \gamma(t, \boldsymbol{X}) \leq \gamma(s', \boldsymbol{X})\right] \\ &\leq (BB'\Sigma/7) \exp\left[-2ay^2/3\right], \end{aligned}$$

if $y \geq y_0$ and we conclude as in the proof of Theorem 5 with $\varepsilon = 0$.

7 Some applications

7.1 Application to the Gaussian setting

In the Gaussian setting, given a discrete subset S of $M = \boldsymbol{l}_2(\mathbb{N}^\star)$ satisfying Assumption 4, we can define on S the following penalized contrast function

$$\gamma(t, \boldsymbol{X}) = \frac{\|t\|^2}{2} - \langle t, \boldsymbol{X} \rangle + \frac{\eta^2(t)}{12}. \tag{7.1}$$

In view of the form of the Gaussian likelihood ratios as given by (4.18), the associated tests given by (6.25) are merely the likelihood ratio tests defined by

$$\psi(t,u,\boldsymbol{X}) = \begin{cases} 0 & \text{if } \log\left(\dfrac{dP_u}{dP_t}\right)(\boldsymbol{X}) + \dfrac{\eta^2(t)-\eta^2(u)}{12\sigma^2} < 0; \\ 1 & \text{if } \log\left(\dfrac{dP_u}{dP_t}\right)(\boldsymbol{X}) + \dfrac{\eta^2(t)-\eta^2(u)}{12\sigma^2} > 0. \end{cases} \tag{7.2}$$

By Proposition 4, they satisfy the error bounds (6.12) and (6.13) with

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

whatever the value of ψ when $\log[(dP_u/dP_t)(\boldsymbol{X})] = \left(\eta^2(u) - \eta^2(t)\right) / \left(12\sigma^2\right)$ and we may therefore apply Theorem 6 to the function γ given by (7.1). There exists a minimizer $\hat{s}$ of γ over S and it is a.s. unique since $\mathbb{P}_s[\gamma(t,\boldsymbol{X}) = \gamma(u,\boldsymbol{X})] = 0$ for $t \neq u$ and S is countable. This T_0-estimator $\hat{s}$ is merely a penalized least squares estimator on S with penalty proportional to $\eta(\cdot)$.

Starting with a family of general models with controlled metric dimensions instead of discrete models, we can derive from Theorem 5 the following result, the proof of which is analogue to the one of Corollary 6.

Corollary 7 *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*

$$\eta_m^2 \geq 101\sigma^2\widetilde{D}_m(\eta_m) \text{ for all } m \in \mathcal{M} \quad \text{and} \quad \sum_{m\in\mathcal{M}} \exp\left[-\frac{\eta_m^2}{504\sigma^2}\right] = \Sigma < +\infty. \tag{7.3}$$

Then one can build a T-estimator $\hat{s}$ which is a penalized least squares estimator on some suitable countable set S and which satisfies, for all $s \in M$ and $1 \leq q \leq 70$,

$$\mathbb{E}_s\left[\|s-\hat{s}\|^q\right] \leq \left[1 + \frac{\Sigma}{10^8}\right]\left(\frac{22}{3}\right)^q \inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\|s-t\| + \eta_m\right\}^q.$$

If, in particular, the sets $\overline{S}_m$ have respective finite metric dimensions bounded by $\overline{D}_m$ and (6.10) holds, one can choose $\eta_m^2 = 101\sigma^2(\overline{D}_m \vee \Delta_m)$ and get

$$\mathbb{E}_s\left[\|s-\hat{s}\|^q\right] \leq \left[1 + \frac{\Sigma}{10^8}\right]\left(\frac{22}{3}\right)^q \inf_{m\in\mathcal{M}}\left\{d_2(s,\overline{S}_m) + 10.05\sigma\sqrt{\overline{D}_m \vee \Delta_m}\right\}^q. \tag{7.4}$$

Let us observe that the results given in the second part of this corollary completely parallel, and actually generalize, since we are not restricted to linear models $\overline{S}_m$, the results of Theorem 2 of Birgé and Massart (2001) 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 7 and set $\overline{D}_m = D_m/2$ and $\Delta_m = 5[L_mD_m + \log(\Sigma'/\Sigma)]$, then (6.10) is satisfied, $\overline{S}_m$ has a finite metric dimension bounded by $\overline{D}_m$ and (7.4) holds. The resulting risk bound is (apart from the constants) the exact analogue of (3.5) in Birgé and Massart (2001). 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 Birgé and Massart (2001) remain valid for T-estimators. The novelty is that Corollary 7 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 two possibilities.

7.1.1 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 $i \geq 1$. 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 Birgé and Massart (2001) 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 Δ_ms satisfying (6.10), 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 Birgé and Massart (2001).

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

$$d_2^2(\theta, \theta') = \sum_{i \geq 1} \left[\frac{1}{e^{i\theta}} - \frac{1}{e^{i\theta'}}\right]^2 = \frac{1}{e^{2\theta} - 1} + \frac{1}{e^{2\theta'} - 1} - \frac{2}{e^{\theta+\theta'} - 1} < \frac{1}{e^{2\theta} - 1}\,. \qquad (7.5)$$

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

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

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

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

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

$$(8/5)x^{-3} \leq g''(x) \leq 2x^{-3} \quad \text{for } x \leq 3 \quad \text{and} \quad e^{-x} \leq g''(x) \leq (5/4)e^{-x} \quad \text{for } x \geq 3.$$

It finally follows that

$$(\theta' - \theta)^2 f(\theta') \leq d_2^2(\theta, \theta') \leq (5/4)(\theta' - \theta)^2 f(\theta) \quad \text{for all } \theta < \theta', \qquad (7.6)$$

for a decreasing function f given by

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

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

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

Then we set

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

It follows from these definitions that, whatever $\theta \in I_k$ one can find some $\theta' \in \{\theta_{k,j},\, 0 \leq j \leq J_k\} \cup \{\theta_{k+1}\}$ such that $|\theta - \theta'| \leq 4\eta e^{k/2}$, hence by (7.6), $d_2(\theta, \theta') < \eta$ and

$$\sup_{\theta \in I_k} \left[\left(\inf_{0 \leq j \leq J_k} d_2(\theta, \theta_{k,j}) \right) \wedge d_2(\theta, \theta_{k+1}) \right] < \eta. \tag{7.7}$$

Moreover, the definition of θ_k implies that

$$\theta_{k+1} - \theta_k = \begin{cases} 1/2 & \text{for } k \geq 0; \\ 3/2 - 5^{-1/3}e^{2/3} \approx 0.361 > (1/2)e^{-1/3} & \text{for } k = -1; \\ 5^{-1/3}\left(e^{4/3} - e\right)e^{k/3} \approx 0.629 e^{k/3} & \text{for } k \leq -2. \end{cases} \tag{7.8}$$

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

$$J_k < G(k)/\eta \quad \text{with} \quad G(k) = \begin{cases} (1/16)\exp(-k/2) & \text{for } k \geq 0; \\ 0.079\exp(-k/6) & \text{for } k < 0. \end{cases} \tag{7.9}$$

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

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

One can check in the same way that this inequality remains true if $K < 0$ and $K+12 \geq 0$ or if $K \geq 0$. Then, by (7.5), $d(\theta_{K+12}, \theta) < \eta$ for $\theta > \theta_{K+12}$ and it follows from the previous arguments that the set $S = \{\theta_{k,j},\, k < K, 0 \leq j \leq J_k\} \cup \{\theta_K, \ldots, \theta_{K+12}\}$ is an η-net for $\overline{S}$. Since, for $k < K$, $|S \cap I_k| = J_k + 1 \leq 2G(k)/\eta$ and for $k \geq K$, $|S \cap I_k| = 1$, we get, for $k \geq 0$,

$$|S \cap [\theta_k, +\infty)| \leq \frac{1}{8\eta}\sum_{j \geq k} \exp\left(\frac{-j}{2}\right) + 13 < \frac{0.318}{\eta}\exp\left(\frac{-k}{2}\right) + 13 \tag{7.10}$$

and, for $k < 0$,

$$\begin{aligned} |S \cap [\theta_k, +\infty)| &\leq \frac{0.158}{\eta}\sum_{j=k}^{-1}\exp\left(\frac{-j}{6}\right) + \frac{1}{8\eta}\sum_{j \geq 0}\exp\left(\frac{-j}{2}\right) + 13 \\ &\leq \frac{0.158}{\eta}\sum_{j \geq k}\exp\left(\frac{-j}{6}\right) + 13 \;<\; \frac{1.03}{\eta}\exp\left(\frac{-k}{6}\right) + 13. \end{aligned} \tag{7.11}$$

We are now in a position to bound the cardinality of $S \cap \mathcal{B}(\theta', r)$ for $\theta' \in \overline{S}$ and $r \geq 2\eta$. Note that the part of the ball that matters is merely the interval $(\underline{\theta}, \overline{\theta})$ with

$\underline{\theta} < \theta' < \overline{\theta}$ and $d_2(\theta', \underline{\theta}) = d_2(\theta', \overline{\theta}) = r$. If the whole interval is contained in some I_k, then by (7.6),

$$\overline{\theta} - \underline{\theta} \leq 2r \left[f(\theta_{k+1})^{-1/2} \right] \leq 2re^{2+k/2},$$

hence

$$|(\underline{\theta}, \overline{\theta}) \cap S| = |(\underline{\theta}, \overline{\theta}) \cap I_k| \leq e^2 r/(4\eta) + 1.$$

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

$$|(\underline{\theta}, \overline{\theta}) \cap S| \leq |(\underline{\theta}, \overline{\theta}) \cap I_{k-1}| + |S \cap [\theta_k, +\infty)| \leq e^2 r/(4\eta) + 1 + |S \cap [\theta_k, +\infty)|$$

and by (7.6),

$$2r \geq d_2(\theta_k, \theta_{k+1}) \geq e^{-2-k/2}(\theta_{k+1} - \theta_k).$$

If $k < 0$, it follows from (7.8) that $\theta_{k+1} - \theta_k > (1/2)e^{k/3}$, hence $e^{-k/6} < 4e^2 r$ and by (7.11),

$$|(\underline{\theta}, \overline{\theta}) \cap S| \leq e^2 r/(4\eta) + 14 + 30.5r/\eta < \exp\left[1.125(r/\eta)^2\right] \quad \text{for } r \geq 2\eta.$$

One can check that the same bound holds for $k \geq 0$ as well as in the cases we started with. It finally follows from Lemma 2 that the metric dimension of $\overline{S}$ is bounded by 4.5.

7.1.2 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 (1982), Wahba (1990), Eggermont and LaRiccia (2001) and Györfi et al. (2002).

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 5.4.4: 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 (5.9) 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 7 with $\eta_m^2 = 53\,\sigma^2\,2^{2m/(2\alpha+1)}$. It follows from (5.10) with $R = 2^m\sigma$ that

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

for all $m \in \mathcal{M}$. Then (7.3) holds with

$$\Sigma = \Sigma(\alpha) < \sum_{m \geq 0} \exp\left[-2^{2m/(2\alpha+1)}/9.51\right]$$

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

$$\mathbb{E}_s\left[\|s - \hat{s}\|^q\right] \leq (22/3)^q \left[1 + 10^{-8}\Sigma(\alpha)\right] \inf_{m \in \mathcal{M}} \left\{d_2(s, \overline{S}_m) + \eta_m\right\}^q \quad \text{for } 1 \leq q \leq 70.$$

If we choose $m = \inf\left\{ j \in \mathbb{N} \;\middle|\; \|s^{(\alpha)}\| \le 2^j \sigma \right\}$, then $s \in \overline{S}_m$, $2^m \le 2\left[\left(\sigma^{-1}\|s^{(\alpha)}\|\right) \vee 1\right]$ and we conclude that

$$\mathbb{E}_s\left[\|s - \hat{s}\|^q\right] \le C(q)\left[1 + 10^{-8}\Sigma(\alpha)\right]\sigma^q\left[\left(\sigma^{-1}\|s^{(\alpha)}\|\right) \vee 1\right]^{q/(2\alpha+1)},$$

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 corresponds to η_m), 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 Györfi et al. (2002) 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 Birgé and Massart (2001). 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.

7.2 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 satisfying Assumption 3 are not, in general, likelihood ratio tests 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 8 *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 $\overline{v}$ or the sup-Hellinger distance $\overline{h}$, and that we have at disposal a finite or countable family of discrete subsets $\{S_m\}_{m\in\mathcal{M}}$ of the set of all distributions for i.i.d. variables. Let those sets S_m satisfy Assumption 4 and*

$$\eta_m^2 \ge 16.8\alpha D_m/n \quad \textit{for all } m \in \mathcal{M}; \qquad \sum_{m\in\mathcal{M}} \exp\left[-\frac{n\eta_m^2}{84\alpha}\right] = \Sigma < +\infty \tag{7.12}$$

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

$$\mathbb{E}_s\left[d^q(s,\hat{s})\right] \le \left[1 + 10^{-8}B'\Sigma\right](16/3)^q \inf_{m\in\mathcal{M}}\left\{d(s, S_m) \vee \eta_m\right\}^q \quad \textit{for } 1 \le q \le 70,$$

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

Proof: It follows from Proposition 5 that in the independent setting, Assumption 3 holds with $B = 1$, $\delta = 4d$ and $a = n/8$ when $d = \overline{v}$, $a = n/4$ when $d = \overline{h}$. In view of these values, our assumptions on η_m and D_m imply (6.5) and (6.6) and the conclusion follows from (6.9) of Theorem 5. □

Models based on uniform distributions As we noticed in Section 4.2.5, it is not possible, whatever $D > 0$, to find a single net satisfying Assumption 2 (η, D, B') for the whole set $\overline{S} = \left\{\overline{P}_t^{\otimes n}, t \in \mathbb{R}\right\}$, where $\overline{P}_t = \mathcal{U}_{e^t}$ denotes the uniform distribution on $[0, \exp(t)]$ 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 metric dimension into a countable number of pieces, each one with finite metric dimension (but not necessarily the same). Here, we replace $\overline{S}$ (identified to $\mathbb{R}$) by the union of submodels $\left\{\overline{S}_m, m \in \mathbb{Z}\right\}$ with $\overline{S}_m = [m\Gamma, (m+1)\Gamma)$ and $\Gamma = 2.94 \exp[n/200]$.

Proof of Proposition 1: In order to apply Corollary 8 with $d = \overline{h}$, we first apply the construction of Section 4.2.5 to each interval $\overline{S}_m$ with $D_m = 10 \log([|m|/K] \vee 1) + 1/2$, $K = 10^6$, and $\eta_m^2 = 16.8 D_m/n$. The resulting value of J, as defined by (4.21) with $D = D_m$, satisfies, since $n \geq 200$, $J > 4.5 \exp[n/84] - 1 > 4.4 \exp[n/84]$ and the corresponding interval I, with $\gamma = m\Gamma$, has a length

$$4J\eta_m^2 > 17.6 \exp[n/84] \times (8.4/n) > 2.94 \exp[n/200] = \Gamma \quad \text{for } n \geq 200.$$

It follows that $\overline{S}_m \subset I$ and the set S_m provided by (4.22) (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 4 with $B' = 4.5$. Moreover, our choice for D_m and η_m imply that (7.12) is satisfied with

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

Finally, since $K = 10^6$, we get from Corollary 8 that, whatever the true probability P_s,

$$\mathbb{E}_s \left[\overline{h}^2(s, \hat{s}) \right] \leq 1.2(16/3)^2 \inf_{m \in \mathcal{M}} \left\{ \overline{h}(s, S_m) \vee \eta_m \right\}^2 \leq C \inf_{m \in \mathcal{M}} \left\{ \overline{h}^2(s, \overline{S}_m) + D_m/n \right\}.$$

If the original parameter $\theta = e^t$ satisfies $\log \theta \in \overline{S}_m$, then $(\log \theta)/\Gamma \in [m, m+1)$ and

$$D_m \leq \left[C' \left(\log(|\log \theta|) - (n/200) - 14.8 \right) \right] \vee 1,$$

which concludes the proof with $\hat{\theta} = \exp(\hat{s})$. □

7.3 Density estimation

7.3.1 From Approximation Theory to discrete models

Density estimation is a major problem in the i.i.d. setting which has been the subject of thousands 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 want to explain how, starting from a family of models with good approximation properties, but which are sets of functions, we can turn it into a family of models which are sets of densities (and

can therefore be used for our estimation purposes) and retain the same properties. The following proposition actually applies to 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 13 *Let M_0 and $\overline{S}$ be two subsets of some metric space (M', d), $\overline{S}$ with finite metric dimension bounded by $\overline{D}$. For any $\eta > 0$, one can find a discrete subset S of M_0 such that*

$$|S \cap \mathcal{B}(t, r)| \leq \exp\left[9.01\overline{D}(r/\eta)^2\right] \quad \textit{for all } t \in M' \textit{ and } r \geq 2\eta \tag{7.13}$$

and

$$d(s, S) \leq 3.03\left[d\left(s, \overline{S}\right) + \eta\right] \quad \textit{for all } s \in M'. \tag{7.14}$$

Proof: Once again, our main tool will be a stratification procedure, replacing $\overline{S}$ by a family of sets S_j and setting $S = \bigcup_j S_j$. By definition, there exists some discrete subset T' of M' which is an η-net for $\overline{S}$ satisfying

$$|T' \cap \mathcal{B}(t, r)| \leq \exp\left[(r/\eta)^2\overline{D}\right] \quad \text{for all } t \in M' \text{ and } r \geq 2\eta. \tag{7.15}$$

Given $\varepsilon = 10^{-4}$, we can build an operator π from T' to M_0 such that $d(t', \pi(t')) \leq (1+\varepsilon)d(t', M_0)$. Let us first set $\eta_0 = 0$,

$$\eta_j = e^{(j-1)/38}\eta, \quad T'_j = \{t' \in T' \,|\, \eta_{j-1} \leq d(t', M_0) < \eta_j\} \quad \text{for } j \geq 1$$

and $S_1 = \pi(T'_1)$. Then, for $j > 1$, we define inductively the sets S_j, choosing for S_j a maximal η_j-separated subset, and therefore an η_j-net, of

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

We finally set $S = \bigcup_{j \geq 1} S_j$. It follows from this construction that, if $t' \in T'$, then $t' \in T'_j$ for some $j \geq 1$ and

$$d(\pi(t'), S) \leq \eta_j = e^{1/38}\eta_{j-1} \leq e^{1/38} d(t', M_0).$$

Moreover, for any $s \in M_0$, it follows from Lemma 2 that one can find $t' \in T'$ with $d(s, T') = d(s, t') \geq d(t', M_0)$. Therefore

$$\begin{aligned} d(s, S) &\leq d(s, t') + d(t', \pi(t')) + d(\pi(t'), S) \\ &\leq d(s, T') + \left(1 + \varepsilon + e^{1/38}\right) d(t', M_0) \;\leq\; 3.03 d(s, T') \end{aligned}$$

and (7.14) follows. Let us now turn to (7.13). Let $t \in M'$, $r \geq 2\eta$ be given and $J > 1$ be defined by $\eta_J < 2r \leq \eta_{J+1}$. Then, by the definition of S_j and the fact that the diameter of the open ball $\mathcal{B}(t, r)$ is smaller than $2r$, either

$$S'_J \cap \mathcal{B}(t, r) = \emptyset \quad \text{with } S'_J = \bigcup_{1 \leq j \leq J} S_j \qquad \text{or} \qquad \left(\bigcup_{j \geq J+1} S_j\right) \cap \mathcal{B}(t, r) = \emptyset.$$

In the first case, $|S \cap \mathcal{B}(t, r)| \leq 1$. In the second case, $|S \cap \mathcal{B}(t, r)| = |S'_J \cap \mathcal{B}(t, r)|$. If $u \in S'_J$, then $u = \pi(u')$ for some $u' \in T'_j$ with $j \leq J$, hence $d(u, u') < (1+\varepsilon)\eta_J < 2(1+\varepsilon)r$. Finally, by (7.15),

$$|S'_J \cap \mathcal{B}(t, r)| \leq |T' \cap \mathcal{B}(t, (3+2\varepsilon)r)| \leq \exp\left[9.01(r/\eta)^2\overline{D}\right],$$

which proves (7.13). □

7.3.2 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 a density estimator from an i.i.d. sample from a distribution with density s with respect to μ and bound its Hellinger risk.

Theorem 7 *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 $\left\{\overline{S}_m\right\}_{m\in\mathcal{M}}$ 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 $\{\Delta_m\}_{m\in\mathcal{M}}$ be a family of nonnegative weights satisfying (6.10). Let $X_1,\ldots,X_n$ be an i.i.d. sample from some distribution $\overline{P}_s$ with density s with respect to μ. One can build a T-estimator $\hat{s}(X_1,\ldots,X_n)$ satisfying, for all $s\in M$ and $1\leq q\leq 70$,*

$$\mathbb{E}_s\left[h^q(s,\hat{s})\right]\leq C(q)\left[1+10^{-8}\Sigma\right]\inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\left\|\sqrt{s}-t\right\|_2+\sqrt{\left[\overline{D}_m\vee\Delta_m\right]/n}\right\}^q. \tag{7.16}$$

Proof: For each m, set $D_m=9.01\overline{D}_m$ and $\eta_m^2=(84/5)n^{-1}[D_m\vee\Delta_m]$. It immediately follows that (7.12) holds. Let d_μ denote the distance in $\mathbb{L}_2(\mu)$ and g the mapping from M to $\mathbb{L}_2(\mu)$ given by $g(t)=\sqrt{t}$. Then g is one-to-one from M to $M_\mu=g(M)$. Since M_μ is a subset of the metric space $\mathbb{L}_2(\mu)$, we may apply Proposition 13 (with $M'=\mathbb{L}_2(\mu)$, $M_0=M_\mu$ and $\eta=\eta_m\sqrt{2}$) to each set $\overline{S}_m$ getting a discrete subset S'_m of M_μ with the following properties

$$\left|S'_m\cap\mathcal{B}_{d_\mu}\left(t',r\sqrt{2}\right)\right|\leq\exp\left[9.01(r/\eta_m)^2\overline{D}_m\right]\quad\text{for all } t'\in\mathbb{L}_2(\mu)\text{ and } r\geq 2\eta_m$$

and

$$d_\mu\left(t',S'_m\right)\leq 3.03\left[d_\mu\left(t',\overline{S}_m\right)+\eta_m\sqrt{2}\right]\quad\text{for all } t'\in M_\mu.$$

Defining $S_m=g^{-1}(S'_m)$ and using the fact that $\sqrt{2}h\left(g^{-1}(t'),g^{-1}(u')\right)=d_\mu(t',u')$ one gets, setting $t=g^{-1}(t')$ for $t'\in M_\mu$,

$$|S_m\cap\mathcal{B}_h(t,r)|\leq\exp\left[9.01(r/\eta_m)^2\overline{D}_m\right]\quad\text{for all } t\in M\text{ and } r\geq 2\eta_m \tag{7.17}$$

and

$$h(t,S_m)\leq 3.03\left[d_\mu\left(\sqrt{t},\overline{S}_m\right)/\sqrt{2}+\eta_m\right]\quad\text{for all } t\in M.$$

We then derive from (7.17) that Assumption 4 holds with $B'=1$ and the conclusion follows from Corollary 8. □

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 bracketting, $\mathbb{L}_\infty$-bounds on s or the introduction of Kullback-Leibler divergences — compare with Yang and Barron (1998, Theorem 1), Birgé, Barron and Massart (1999, Theorem 2), Castellan (1999 and 2000) or van de Geer (2000) —.

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 7 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

$$\begin{aligned} \|s - \hat{s}_{\overline{R}}\|_2^2 &= \int \left(\sqrt{s} + \sqrt{\hat{s}_{\overline{R}}}\right)^2 \left(\sqrt{s} - \sqrt{\hat{s}_{\overline{R}}}\right)^2 d\mu \;\leq\; 4\overline{R} \int \left(\sqrt{s} - \sqrt{\hat{s}_{\overline{R}}}\right)^2 \\ &\leq 4\overline{R} \int \left(\sqrt{s} - \sqrt{\hat{s}}\right)^2 = 8\overline{R} h^2(s, \hat{s}) \end{aligned}$$

and a bound for $\mathbb{E}_s\left[\|s - \hat{s}_{\overline{R}}\|_2^2\right]$ could be derived from (7.16). 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.

To illustrate the power of Theorem 7, we give one application relying on the following proposition which partly summarizes the results of Birgé and Massart (2000). We refer to this paper and the book by DeVore and Lorentz (1993) for details on Besov spaces.

Proposition 14 *Given positive integers d and r, one can find for each $j \geq 0$ a family $\left\{\overline{S}_m\right\}_{m \in \mathcal{M}_j}$ of D_j-dimensional linear spaces of functions on $\mathbb{R}^d$ with the following properties:*

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

$$D_j \leq c_1 + c_2 2^{jd} \qquad \text{and} \qquad |\mathcal{M}_j| \leq \exp\left(c_3 2^{jd}\right),$$

where the constants $c_i \geq 1$ only depend on r and d;

ii) for any $p > 0$, $q \geq 1$ and α with $r > \alpha > (d/p - d/q)_+$ and any function t belonging to the Besov space $B^\alpha_{p,\infty}([0,1]^d)$ 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

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

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

Restricting ourselves to the case $d = 1$ for simplicity, we can apply Theorem 7 to the family of models $\left\{\overline{S}_m\right\}_{m \in \mathcal{M}}$ with $\mathcal{M} = \cup_{j \geq 0} \mathcal{M}_j$ provided by the previous proposition. If $m \in \mathcal{M}_j$, we choose $\Delta_m = 5\left(c_3 2^j + 10^{-6} j\right)$ and $\overline{D}_m = \left(c_1 + c_2 2^j\right)/2$ according to Proposition 7. Applying Proposition 14 with $t = \sqrt{s}$ and $q = 2$, we derive from Theorem 7 that, if $|\sqrt{s}|_{B^\alpha_{p,\infty}} \leq R$ with $r > \alpha > (1/p - 1/2)_+$,

$$\mathbb{E}_s\left[h^2(s, \hat{s})\right] \leq C_1 \inf_{j \geq 0} \left\{C(r, \alpha, p) R^2 2^{-2j\alpha} + c_4 n^{-1} 2^j\right\}.$$

An optimization with respect to j leads to the following

Theorem 8 *Let $X_1, \ldots, X_n$ be an n-sample from some distribution $\overline{P}_s$ with density s with respect to Lebesgue measure on $[0,1]$ and let the integer $r > 0$ be given. One can build a T-estimator $\hat{s}(X_1, \ldots, X_n)$ such that, if the Besov semi-norm of $\sqrt{s}$ satisfies $|\sqrt{s}|_{B^\alpha_{p,\infty}} \leq R$ for some $p > 0$, $r > \alpha > (1/p - 1/2)_+$ and $R \geq 1/\sqrt{n}$, then*

$$\mathbb{E}_s\left[h^2(s, \hat{s})\right] \leq C(r, \alpha, p) R^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)}. \tag{7.18}$$

Note that the use of Hellinger distance allows to get adaptation for the whole domain $1/p \geq \alpha > (1/p - 1/2)_+$ which is, to our knowledge, new for density estimation. One can prove in the same way a multidimensional analogue.

One can also apply to the i.i.d. setting the results of Birgé and Massart (2001, 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 8. 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 (7.18) 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 nonparametric models, getting the parametric rate if the true distribution is in the parametric model and the nonparametric one otherwise.

7.3.3 A parallel with the White noise framework

One should stress the fact that the results of Theorem 7 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 7 also applies to the White noise framework via the identification mentioned in Section 4.2.2 leading to the following

Corollary 9 *Assume we are in the White noise framework, observing the process Y given by (4.19) with unknown parameter s and that we have at disposal a finite or countable family of subsets $\left\{\overline{S}_m\right\}_{m\in\mathcal{M}}$ of $\mathbb{L}_2([0,1], dx)$ with respective finite metric dimensions bounded by $\overline{D}_m$. If the family of nonnegative weights $\{\Delta_m\}_{m\in\mathcal{M}}$ satisfies (6.10), one can build a T-estimator $\hat{s}(Y)$ satisfying, for all $s \in \mathbb{L}_2([0,1], dx)$ and $1 \leq q \leq 70$,*

$$\mathbb{E}_s\left[\|s - \hat{s}\|_2^q\right] \leq C(q)\left[1 + 10^{-8}\Sigma\right] \inf_{m\in\mathcal{M}} \left\{ \inf_{t\in\overline{S}_m} \|s - t\|_2 + \sqrt{\left[\overline{D}_m \vee \Delta_m\right]/n} \right\}^q. \tag{7.19}$$

Clearly, (7.16) is the exact analogue of (7.19). 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 Birgé and Massart (2001), 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 placed on $\sqrt{s}$ and our Theorem 8 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 (1972 and 1986) and illustrated, for instance, by Brown and Low (1996) or Nussbaum (1996), among other examples. Our paralellism has some limitations: it holds up to constants, we restrict to loss functions of the form $\|s - \hat{s}\|_2^q$ and $h^q(s, \hat{s})$ (although this could easily be generalized) 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).

7.3.4 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 variation distance since for two probabilities P and Q with respective densities f and g, $\|f-g\|_1 = 2v(P,Q)$. Combining Corollary 8 and Proposition 13 we get the following analogue of Theorem 7 for the independent setting. The proof being quite similar, it will be omitted.

Theorem 9 *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 nonnegative weights satisfying (6.10). 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 $s\in M$ and $1\le q\le 70$,*

$$\mathbb{E}_s\left[\overline{v}^q(s,\hat{s})\right] \le C(q)\left[1+10^{-8}\Sigma\right]\inf_{m\in\mathcal{M}}\left\{\inf_{t\in\overline{S}_m}\overline{v}(s,t)+\sqrt{\left[\overline{D}_m\vee\Delta_m\right]/n}\right\}^q,$$

with

$$\overline{v}(s,t) = \sup_{1\le i\le n} v(\overline{P}_i, t\cdot\mu) \quad \textit{for } s\in M \quad \textit{and} \quad t\in\overline{M}_\mu.$$

Theorem 1 immediately follows from this last result and Proposition 14 by a proof which is completely similar to the proof of Theorem 8 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 $|s'|_{B^\alpha_{p,\infty}} \le R$ we get, when $d=1$,

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

7.4 Aggregation of estimators

The fact that our method can be applied to select an estimator from some family has been suggested to us by Yannick Baraud and Sacha Tsybakov (private conversations). For simplicity, we shall just deal with the i.i.d. case, as defined in Section 4.2.1, using variation distance v, the case of Hellinger being quite similar. The typical situation involves a sample of size $2n$. We use the first half of the sample to build the estimators to be aggregated and the second half to select one among them or mix them together in some way. The construction of suitable T-estimators being based on the second half-sample only, one can merely work conditionnally to the first half. From this point of view, the initial estimators are just points in $\overline{M}$, as in Tsybakov (2003), which we assume here. One should compare the following results with those of Catoni (1997 and 1999), Nemirovski (2000) and Yang (2000).

7.4.1 Selecting an estimator from a countable family

Suppose we are given a countable number $\{t_m, m\in\mathcal{M}\}$ of points in $\overline{M}$ (possibly preliminary estimators) and we want to select one among them. We can use here a

Bayesian approach, starting with a prior distribution ν on $\mathcal{M}$ and then setting $\nu_m = -\log(\nu(\{m\})) > 0$, $S_m = \{t_m\}$, $D_m = 1/2$, $B' = e^{-2}$ and $\eta_m^2 = (168/n)[(1/10) \vee (\nu_m - 18)]$, so that the assumptions of Corollary 8 are satisfied with $\Sigma = e^{18}$. We then get

Theorem 10 *Let $X_1, \ldots, X_n$ be an i.i.d. sample from some unknown distribution $\overline{P}_s$ on $\mathcal{X}$, $\{t_m, m \in \mathcal{M}\}$ be a countable subset of the set $\overline{M}$ of all distributions on $\mathcal{X}$, and ν be a probability on $\mathcal{M}$ with $\nu_m = -\log(\nu(\{m\})) > 0$ for each m. One can build a selection procedure $\hat{m}(X_1, \ldots, X_n)$ with values in $\mathcal{M}$ such that, whatever $\overline{P}_s \in \overline{M}$ and $1 \leq q \leq 70$,*

$$\mathbb{E}_s\left[v^q\left(s, t_{\hat{m}}\right)\right] \leq 1.1(16/3)^q \inf_{m \in \mathcal{M}} \left\{ v\left(s, t_m\right) \vee 13\sqrt{n^{-1}[(1/10) \vee (\nu_m - 18)]} \right\}^q .$$

Such a result could be used, for instance, for bandwith selection.

7.4.2 Convex aggregation

Let now $\{t_1, \ldots, t_N\}$ 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 nonvoid 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_ks with $k \in m$, i.e.

$$\overline{S}_m = \left\{ \sum_{j=1}^{|m|} \lambda_j t_{k_j} \quad \text{with } \lambda_j \geq 0 \text{ for } 1 \leq j \leq |m| \quad \text{and} \quad \sum_{j=1}^{|m|} \lambda_j = 1 \right\}. \tag{7.20}$$

Since $\overline{S}_m$ is isometric to a subset of an $|m|$-dimensional normed linear space, it follows from Proposition 7 that it has a finite metric dimension bounded by $5|m|/3$. Hence, given $\eta_m > 0$, one can find an η_m-net S_m of $\overline{S}_m$ which satisfies Assumption 4 with $D_m = 5|m|/3$ and $B' = 1$. Now observe that the number of elements of $\mathcal{M}$ with cardinality j is $\binom{N}{j} < (eN/j)^j$. If we set

$$\eta_m^2 = (168|m|/n)\left[(1/3) \vee [1 + \log(N/|m|) + (\log N - 16)/|m|\,]\right], \tag{7.21}$$

we can then check that (7.12) holds with $\alpha = 2$ and $\Sigma < e^{16}$ and apply Corollary 8 with $d = v$. Since $v(s, S_m) \leq v(s, \overline{S}_m) + \eta_m$, this proves

Theorem 11 *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 $\overline{P}_s$ in $\overline{M}$. For m an arbitrary nonvoid subset of $\{1, \ldots, N\}$, let $\overline{S}_m$ denote the convex envelope of the t_ks with $k \in m$ as defined by (7.20) and η_m be given by (7.21). One can build a T-estimator $\hat{s}(X_1, \ldots, X_n)$ such that, whatever $\overline{P}_s$,*

$$\mathbb{E}_s\left[v^q(s, \hat{s})\right] \leq 1.1(16/3)^q \inf_{m \in \mathcal{M}} \left\{ v(s, \overline{S}_m) + \eta_m \right\}^q \quad \text{for } 1 \leq q \leq 70.$$

In particular,

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

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 $\{t_1, \ldots, t_N\}$.

7.4.3 Partition selection for histograms

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 $m = \{I_1, \ldots, I_D\}$ of $\mathcal{X}$, we can consider the histogram estimator $\hat{s}_m \cdot \mu$ of $\overline{P}_s$ with density $\hat{s}_m$ with respect to μ given by

$$\hat{s}_m(X_1, \ldots, X_n) = \frac{1}{n} \sum_{j=1}^{D} \frac{N_j}{\mu(I_j)} \mathbb{1}_{I_j}, \quad \text{with } N_j = \sum_{i=1}^{n} \mathbb{1}_{I_j}(X_i). \tag{7.22}$$

One easily shows that

$$\mathbb{E}_s \left[v\left(\overline{P}_s, \hat{s}_m \cdot \mu\right) \right] \leq v\left(\overline{P}_s, \bar{s}_m \cdot \mu\right) + \frac{1}{2}\sqrt{\frac{D-1}{n}} \quad \text{with } \bar{s}_m = \sum_{j=1}^{D} \frac{\overline{P}_s(I_j)}{\mu(I_j)} \mathbb{1}_{I_j}. \tag{7.23}$$

Suppose now given $2n$ i.i.d. observations with distribution $\overline{P}_s$, a countable family $\mathcal{M}$ of finite partitions of $\mathcal{X}$ and a probability measure ν on $\mathcal{M}$. One can first build all the histograms $\hat{s}_m(X_1, \ldots, X_n)$ based on the first n observations as in (7.22) and then select one among them, $\hat{m}(X_{n+1}, \ldots, X_{2n})$ using the last n observations which results in a density estimator $\tilde{s} = \hat{s}_{\hat{m}}$. Applying Theorem 10 with $q = 1$ conditionally on $X_1, \ldots, X_n$ gives

$$\begin{aligned} &\mathbb{E}_s \left[v\left(\overline{P}_s, \tilde{s} \cdot \mu\right) \middle| X_1, \ldots, X_n \right] \\ &\qquad \leq \; C \inf_{m \in \mathcal{M}} \left\{ v\left(\overline{P}_s, \hat{s}_m \cdot \mu\right) \vee \left(n^{-1/2} \sqrt{1 \vee (\nu_m - 18)} \right) \right\}. \end{aligned}$$

Integrating with respect to $X_1, \ldots, X_n$ and applying (7.23) leads to

$$\mathbb{E}_s \left[v\left(\overline{P}_s, \tilde{s} \cdot \mu\right) \right] \leq C \inf_{m \in \mathcal{M}} \left\{ v\left(\overline{P}_s, \bar{s}_m \cdot \mu\right) + n^{-1/2} \sqrt{|m| \vee (\nu_m - 18)} \right\}. \tag{7.24}$$

If $|m| \geq 2 \vee [c(\nu_m - 18)]$ for all $m \in \mathcal{M}$ and some $c > 0$, by (7.23), this corresponds, up to some multiplicative constant depending on c, to the risk of the best histogram among the family. This result, which allows arbitrary families of partitions and requires no assumption at all on $\overline{P}_s$ should be compared, for instance, with Castellan (1999 and 2000) and Devroye and Lugosi (2001).

Let us now focus on the case $\mathcal{X} = [0, 1)$ and μ is Lebesgue measure. We only consider partitions $m_{\mathcal{J}} = \{I_1, \ldots, I_D\}$ of $[0, 1)$ generated by increasing sequences $\mathcal{J} = \{0 = x_0 < x_1 < \ldots < x_D = 1\}$ with $I_j = [x_{j-1}, x_j)$. For each positive integer k, we set $\mathcal{J}_k = \left\{ j2^{-k}, j = 0, \ldots, 2^k \right\}$ and, if $2 \leq D \leq 2^k$, 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 $\mathcal{M} = \bigcup_{k \geq 1} \left(\bigcup_{D=2}^{2^k} \mathcal{M}_{k,D} \right)$. Since $|\mathcal{M}_{k,D}| \leq \begin{pmatrix} 2^k - 1 \\ D - 1 \end{pmatrix} < 2^{k(D-1)}$,

$$\sum_{k \geq 1} \sum_{D=2}^{2^k} |\mathcal{M}_{k,D}| e^{-Dk} < \sum_{k \geq 1} 2^{-k} \sum_{D \geq 2} \exp[-Dk(1 - \log 2)] < 5/4$$

and we can define a probability measure ν on $\mathcal{M}$ setting $\nu(\{m\}) = ce^{-Dk}$ with $c > 4/5$, hence $\nu_m < Dk + \log(5/4)$. Using (7.24), we conclude that there exists a partition selection procedure $\hat{m}$ such that, whatever $\overline{P}_s$,

$$\mathbb{E}_s \left[v\left(\overline{P}_s, \hat{s}_{\hat{m}} \cdot \mu\right) \right] \leq C \inf_{k \geq 1} \inf_{2 \leq D \leq 2^k} \inf_{m \in \mathcal{M}_{k,D}} \left\{ v\left(\overline{P}_s, \bar{s}_m \cdot \mu\right) + \sqrt{Dk/n} \right\}.$$

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

8 Appendix

8.1 Proof of Proposition 2

Since $k \geq 128$,

$$c \geq 3.36 \qquad \text{and} \qquad \sqrt{3k} - \sqrt{k} - 4 - 1.21k^{1/4} > 0. \tag{8.1}$$

It follows that, whatever the true value of $s \in \mathcal{S}$, $\mathbb{P}_s[|X_0| \geq c + 1.21] < 0.114$ and that $\|\boldsymbol{X}'\|^2$ has a non-central $\chi^2(k)$ distribution. Since a noncentral χ^2 variable is stochastically larger than a central one, it follows from Laurent and Massart (2000, Lemma 1) that

$$\mathbb{P}_s\left[\|\boldsymbol{X}'\|^2 \leq k - 2\sqrt{kx}\right] \leq e^{-x} \quad \text{for } x > 0. \tag{8.2}$$

Setting $x = k/64 \geq 2$, 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

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

Since for $s_0 \neq 0$,

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

is nonnegative by (8.1), $f(s_0)$ 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 $(3/4)\left(s_0^2 + 4\right)$ 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.

8.2 Proof of Proposition 3

First (4.3) implies that

$$\mathbb{P}[Y + \lambda\overline{y} > z] = \mathbb{P}[Y > z - \lambda\overline{y}] \leq \alpha \exp\left[-\beta(z - \lambda\overline{y})^2\right] \quad \text{for } z \geq (1 + \lambda)\overline{y}.$$

Moreover, if $z \geq (1 + \lambda)\overline{y}$, then $z - \lambda\overline{y} = z(1 - \lambda\overline{y}/z) \geq z/(1 + \lambda)$, so that finally

$$\mathbb{P}[Y + \lambda\overline{y} > z] \leq \alpha \exp\left[-\frac{\beta z^2}{(1 + \lambda)^2}\right] \quad \text{for } z \geq (1 + \lambda)\overline{y},$$

which is exactly (4.3) with different parameters. It therefore suffices to consider the case $\lambda = 0$ and then replace $\overline{y}$ by $(1+\lambda)\overline{y}$ and β by $\beta(1+\lambda)^{-2}$.

Using (4.3) and the change of variable $x = \beta y^{2/q}$, we get

$$\begin{aligned}\frac{1}{\alpha}\left(\mathbb{E}[Y^q] - \overline{y}^q\right) &= \frac{1}{\alpha}\left(\int_0^\infty \mathbb{P}[Y^q \geq y]\,dy - \overline{y}^q\right) \leq \frac{1}{\alpha}\int_{\overline{y}^q}^\infty \mathbb{P}\left[Y \geq y^{1/q}\right]\,dy \\ &\leq \int_{\overline{y}^q}^\infty \exp\left(-\beta y^{2/q}\right)\,dy = \beta^{-q/2}\frac{q}{2}\int_{\beta\overline{y}^2}^\infty x^{q/2-1}e^{-x}\,dx.\end{aligned}$$

There are then various ways to bound the last integral I. Either we use $qI/2 \leq \Gamma(q/2+1)$ and Stirling's expansion: $\Gamma(x+1) \leq \sqrt{\pi e x}(x/e)^x$ for $x \geq 1/2$, or, when $\beta\overline{y}^2 \geq q/2$, we alternatively use the bound — Johnstone (2001), Inequality (45) — $\int_z^{+\infty} x^t e^{-x}\,dx < \left(z^{t+1}e^{-z}\right)(z-t)^{-1}$ for $z > t$, which gives $I < \left(\beta\overline{y}^2\right)^{q/2}\exp\left(-\beta\overline{y}^2\right)$ so that

$$\frac{\mathbb{E}[Y^q] - \overline{y}^q}{\alpha\overline{y}^q} \leq \begin{cases} \sqrt{\pi e q/2}\left(2e\beta\overline{y}^2/q\right)^{-q/2} & \text{for all } \overline{y} > 0; \\ (q/2)\exp\left(-\beta\overline{y}^2\right) & \text{if } \beta\overline{y}^2 \geq q/2. \end{cases}$$

This implies that (4.4) holds with $\lambda = 0$. Both functions $x \mapsto (2ex/q)^{-q/2}$ and $x \mapsto e^{-x}$ are decreasing on $(0, +\infty)$ and they coincide for $x = q/2$ which implies that ζ_q is decreasing for $1 \leq q \leq 2\pi e$. The choice of $c = 0.612 > 1/2$ for $q > 2\pi e$ ensures that $\zeta_q(cq) < \zeta_q((cq)_-)$ so that ζ_q is still decreasing for all those values of q. To prove (4.5), we observe that the function $f(x) = x^{q/2}\zeta_q(x)$ is constant on $(0, cq)$ and has a downward jump at cq like ζ_q. Since $f(x) = (q/2)x^{q/2}e^{-x}$ for $x \geq cq \geq q/2$, which is a decreasing function for $x \geq q/2$, the maximum of f is its value on $(0, cq)$, i.e. $\sqrt{\pi eq/2}(2e/q)^{-q/2}$. This implies that

$$\overline{y}^q\zeta_q\left(\beta\overline{y}^2\right) = \beta^{-q/2}\left(\beta\overline{y}^2\right)^{q/2}\zeta_q\left(\beta\overline{y}^2\right) \leq \beta^{-q/2}\sqrt{\pi eq/2}(2e/q)^{-q/2},$$

which completes the proof of (4.5).

8.3 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 $X_1, \ldots, X_n$ 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 nonnegative measurable function on $\mathcal{X}$ such that*

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

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

$$\mathbb{P}\left[\sum_{i=1}^n \log\phi(X_i) \geq ny\right] \leq \exp[n(\log\alpha - y)]$$

and

$$\mathbb{Q}\left[\sum_{i=1}^n \log\phi(X_i) \leq ny\right] \leq \exp[n(\log\beta + y)].$$

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

$$\mathbb{P}\left[\sum_{i=1}^{n} \log\left(\frac{d\overline{Q}}{d\overline{P}}\right)(X_i) \geq nx\right] \leq \exp\left[n \log[\rho(\overline{P}, \overline{Q})] - (nx/2)\right] \tag{8.3}$$

and

$$\mathbb{Q}\left[\sum_{i=1}^{n} \log\left(\frac{d\overline{Q}}{d\overline{P}}\right)(X_i) \leq nx\right] \leq \exp\left[n \log[\rho(\overline{P}, \overline{Q})] + (nx/2)\right]. \tag{8.4}$$

Proof: It immediately follows from the elementary inequality

$$\mathbb{P}[\log Y \geq z] \leq e^{-z}\mathbb{E}[Y], \quad \text{if } \mathbb{P}[Y \geq 0] = 1 \tag{8.5}$$

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

Proof of Proposition 4 We get from (8.5) and (4.18)

$$\begin{aligned}
&\mathbb{P}_s[\log(dP_u/dP_t)(\boldsymbol{X}) \geq 2x] \\
&\qquad \leq \; e^{-x}\mathbb{E}_s\left[\exp[(1/2)\log(dP_u/dP_t)(\boldsymbol{X})]\right] \\
&\qquad = \; e^{-x}\mathbb{E}_0\left[\sqrt{(dP_u/dP_t)(\boldsymbol{X})}(dP_s/dP_0)(\boldsymbol{X})\right] \\
&\qquad = \; e^{-x}\mathbb{E}_0\left[\exp\left[-\frac{1}{2\sigma^2}\left(\frac{\|u\|^2 - \|t\|^2}{2} + \|s\|^2 - \langle \boldsymbol{X}, u - t + 2s\rangle\right)\right]\right].
\end{aligned}$$

Since

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

we get

$$\begin{aligned}
&\mathbb{P}_s[\log(dP_u/dP_t)(\boldsymbol{X}) \geq 2x] \\
&\qquad \leq \; e^{-x}\mathbb{E}_0\left[p_{\frac{u-t}{2}+s}(\boldsymbol{X})\right]\exp\left[-\frac{1}{2\sigma^2}\left(\frac{\|u\|^2 - 3\|t\|^2}{4} - \langle s, u - t\rangle + \frac{\langle u, t\rangle}{2}\right)\right] \\
&\qquad = \; \exp\left[-x - \frac{1}{8\sigma^2}\left(\|u\|^2 - 3\|t\|^2 - 4\langle s, u - t\rangle + 2\langle u, t\rangle\right)\right].
\end{aligned}$$

The conclusion follows from the fact that

$$\begin{aligned}
-\|u\|^2 + 3\|t\|^2 + 4\langle s, u - t\rangle - 2\langle u, t\rangle \;&=\; -\|t - u\|^2 + 4\langle s - t, u - t\rangle \\
&\leq\; -\|t - u\|(\|t - u\| - 4\|s - t\|).
\end{aligned}$$

Proof of Proposition 5 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 Huber (1965, Section 7) 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, whatever $x \in \mathbb{R}$,

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

and

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

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

$$\psi(t_0, u_0, \boldsymbol{X}) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \log\left[(d\overline{P}_{u_0}/d\overline{P}_{t_0})(X_i)\right] < 2x, \\ 1 & \text{if } \sum_{i=1}^n \log\left[(d\overline{P}_{u_0}/d\overline{P}_{t_0})(X_i)\right] > 2x. \end{cases}$$

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

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

and similarly,

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

Since (8.3), (8.4) and (2.1) imply that

$$\mathbb{P}_{t_0}[\psi(t_0, u_0, \boldsymbol{X}) = 1] \le \exp\left[-nh^2(\overline{P}_{u_0}, \overline{P}_{t_0}) - x\right]$$

and

$$\mathbb{P}_{u_0}[\psi(u_0, t_0, \boldsymbol{X}) = 1] \le \exp\left[-nh^2(\overline{P}_{u_0}, \overline{P}_{t_0}) + x\right],$$

and, by (4.16),

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

the conclusion follows for the variation distance.

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 (1986) that one can find a nonnegative measurable function ϕ on $\mathcal{X}$, such that

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

The conclusion then follows from Lemma 7 with

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

and $ny = x$.

8.4 Proof of Proposition 11

Let us first evaluate, for $0 < s < t \leq 1/3$,

$$h^2(s,t) = h^2(P_s, P_t) = \frac{1}{2}\int_0^1 \left(\sqrt{f_s(x)} - \sqrt{f_t(x)}\right)^2 dx.$$

Setting $\beta_s = \left(s^2+s+1\right)^{-1} \in [9/13, 1)$, we get

$$\begin{aligned} 2h^2(s,t) &= s^3\left(s^{-1}-t^{-1}\right)^2 + \left(t^3-s^3\right)\left(t^{-1}-\sqrt{\beta_s}\right)^2 + \left(1-t^3\right)\left(\sqrt{\beta_s}-\sqrt{\beta_t}\right)^2 \\ &= (t-s)\frac{s}{t}\left(1-\frac{s}{t}\right) + (t-s)\left[1+\frac{s}{t}+\left(\frac{s}{t}\right)^2\right]\left(1-t\sqrt{\beta_s}\right)^2 \\ &\quad + \left(1-t^3\right)\left(\sqrt{\beta_s}-\sqrt{\beta_t}\right)^2. \end{aligned}$$

Note that the monotonicity of $s \mapsto \beta_s$ implies that

$$4/9 < \left(1-t\sqrt{\beta_s}\right)^2 < 1, \qquad \sqrt{\beta_s}+\sqrt{\beta_t} > 2\sqrt{\beta_{1/3}} = 6/\sqrt{13}$$

and

$$0 < \beta_s - \beta_t = \frac{(t-s)(s+t+1)}{(s^2+s+1)(t^2+t+1)} < t-s. \tag{8.6}$$

It follows that

$$0 < \left(\sqrt{\beta_s}-\sqrt{\beta_t}\right)^2 = \frac{(\beta_s-\beta_t)^2}{\left(\sqrt{\beta_s}+\sqrt{\beta_t}\right)^2} < \frac{13}{36}(t-s)^2 = \frac{13t}{36}(t-s)\left(1-\frac{s}{t}\right)$$

and

$$0 < \left(1-t^3\right)\left(\sqrt{\beta_s}-\sqrt{\beta_t}\right)^2 < \frac{13}{108}(t-s)\left(1-\frac{s}{t}\right).$$

Hence, if we set $z = s/t \in (0,1)$, $2h^2(s,t) = (t-s)g(s,t)$, with

$$g(s,t) = z(1-z) + c_1\left(1+z+z^2\right) + c_2(1-z), \quad \frac{4}{9} < c_1 < 1 \quad \text{and} \quad 0 < c_2 < \frac{13}{108}.$$

It follows that

$$4/9 < c_1 < g(s,t) < 1 + 2z + c_2(1-z) < 3$$

and we conclude that whatever s and t in $(0, 1/3]$,

$$h^2(s,t) = C(s,t)|s-t| \quad \text{with } 2/9 < C(s,t) < 3/2.$$

It immediately follows that the set $S = \{t_j,\ j \geq 0\}$ with $t_j = (2j+1)2\eta^2/3$ is an η-net for the family $\{f_s,\ s \in \mathcal{S}\}$. On the other hand, given $t_j \in S$ and $r \geq 2\eta$, in order that $t \in \mathcal{B}(t_j, r)$, it is required that $h^2(t_j,t) = C(t_j,t)|t_j - t| < r^2$ which implies that $|t_j - t| < (9/2)r^2$ and therefore

$$|S \cap \mathcal{B}(t_j, r)| \leq 1 + (27/4)(r/\eta)^2 \leq \exp\left[0.84(r/\eta)^2\right].$$

It follows that S satisfies Assumption 2 $(\eta, 0.84, 1)$ and Theorem 2 implies the upper bound for the squared Hellinger risk.

Let us now procede with the $\mathbb{L}_2$-distance d defined by $d^2(s,t) = \int [f_s(x) - f_t(x)]^2 dx$. We get

$$\begin{aligned} d^2(s,t) &= s^3\left(s^{-2}-t^{-2}\right)^2 + \left(t^3-s^3\right)\left(t^{-2}-\beta_s\right)^2 + \left(1-t^3\right)(\beta_s-\beta_t)^2 \\ &= \left(\frac{1}{s}-\frac{1}{t}\right)\left(1-\frac{s}{t}\right)\left(1+\frac{s}{t}\right)^2 \\ &\quad + \left(\frac{1}{s}-\frac{1}{t}\right)\left[\frac{s}{t}+\left(\frac{s}{t}\right)^2+\left(\frac{s}{t}\right)^3\right]\left(1-t^2\beta_s\right)^2 \\ &\quad + \left(\frac{1}{s}-\frac{1}{t}\right)\left(1-\frac{s}{t}\right)st^2\left(1-t^3\right)\left(\frac{\beta_s-\beta_t}{t-s}\right)^2. \end{aligned}$$

Since $8/9 < 1 - t^2\beta_s < 1$ and, by (8.6),

$$0 < st^2\left(1-t^3\right)\left(\frac{\beta_s-\beta_t}{t-s}\right)^2 < \frac{1}{27},$$

we conclude that $d^2(s,t) = \left(s^{-1}-t^{-1}\right) g(s,t)$ with $0 < z = s/t < 1$,

$$g(z) = (1-z)(1+z)^2 + c_1\left(z+z^2+z^3\right) + c_2(1-z), \quad \frac{8}{9} < c_1 < 1 \quad \text{and} \quad 0 < c_2 < \frac{1}{27}.$$

It follows that

$$g(z) > 1 + z - z^2 - z^3 + (8/9)\left(z+z^2+z^3\right) > 1$$

and

$$g(z) < 1 + 2z + (1/27)(1-z) < 3,$$

which finally implies that whatever s and t in $(0, 1/3]$,

$$d^2(s,t) = C(s,t)\left|s^{-1}-t^{-1}\right| \quad \text{with } 1 < C(s,t) < 3.$$

Now setting $S = \{t_j,\ j \geq 0\}$ with $t_j^{-1} = 3 + 2j\eta^2/3$ we deduce as before that S satisfies Assumption $2\,(\eta, D, 1)$ for some D independent of η. It nevertheless follows from the fact that $h^2(s,t) \to 0$ while $\|s-t\|^2 \to +\infty$ when s and t tend to zero and classical arguments of Le Cam (1973) or Donoho and Liu (1987) that the minimax risk over $\mathcal{S}$ is infinite when we use the $\mathbb{L}_2$-loss.

Acknowledgements I would like to thank the participants of the Colloquium in Honour of Jean Bretagnolle, Didier Dacunha-Castelle and Ildar Ibragimov who asked questions after the talk I gave in June 2001 on this topic, in particular David Pollard, since they pushed me to pursue in this direction. Special thanks also to my collegues Yannick Baraud, Fabienne Comte, Pascal Massart and Sacha Tsybakov for their encouragements, suggestions and comments on some earlier version of the paper.

References

BARAUD, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127-146.

BARRON, A.R. (1991). Complexity regularization with applications to artificial neural networks. In *Nonparametric Functional Estimation* (G. Roussas, ed.). Kluwer, Dordrecht, 561-576.

BARRON, A.R. and COVER, T.M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37**, 1034-1054.

BARRON, A.R., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-415.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie Verw. Geb.* **65**, 181-237.

BIRGÉ, L. (1984a). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3**, 259-282.

BIRGÉ, L. (1984b). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Sect. B* **20**, 201-223.

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields* **71**, 271-291.

BIRGÉ, L. (2002). Model selection for Gaussian regression with random design. Technical Report No 783. Laboratoire de Probabilités, Université Paris VI.
http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2002

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Th. Rel. Fields* **97**, 113-150.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.), 55-87. Springer-Verlag, New York.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329-375.

BIRGÉ, L. and MASSART, P. (2000). An adaptive compression algorithm in Besov spaces. *Constructive Approximation* **16** 1-36.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268.

BIRMAN, M.S. and SOLOMJAK, M.Z. (1967). Piecewise-polynomial approximation of functions of the classes W_p. *Mat. Sbornik* **73**, 295-317.

BROWN, L.D. and LOW, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384-2398.

CASTELLAN, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report 99.61. Université Paris-Sud, Orsay.
http://www.math.u-psud.fr/ biblio/pub/1999/

CASTELLAN, G. (2000). Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C.R.A.S.* **330**, 729-732.

CATONI, O. (1997). The mixture approach to universal model selection. Technical Report LMENS-97-22. Ecole Normale Supérieure, Paris.
http://www.dma.ens.fr/edition/publis/1997/titre97.html

CATONI, O. (1999). Universal aggregation rules with sharp oracle inequalities. Technical Report No 510. Laboratoire de Probabilités, Université Paris VI.
http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999

CHERNOFF, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based

on a sum of observations. *Ann. Math. Stat.* **23**, 493-507.

DeVORE, R.A. and LORENTZ, G.G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin.

DEVROYE, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York.

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding *Ann. Statist.* **24**, 508-539.

DONOHO, D.L. and LIU, R.C. (1987). Geometrizing rates of convergence I. Technical report 137. Department of Statistics, University of California, Berkeley.

DONOHO, D.L., LIU, R.C. and MacGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18**, 1416-1437.

EGGERMONT, P.P.B. and LaRICCIA, V.N. (2001). *Maximum Penalized Likelihood Estimation, Volume I: Density Estimation.* Springer, New York.

GROENEBOOM, P. (1986). Some current developments in density estimation, in *Mathematics and Computer Science, CWI Monograph, volume 1* (J.W. de Bakker, M. Hazewinkel, J.K. Lenstra, eds.), Elsevier, Amsterdam, 163-192.

GYÖRFI, L., KOHLER, M., KRYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York.

HUBER, P.J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36**, 1753-1758.

HUBER, P.J. (1981). *Robust Statistics.* John Wiley, New York.

JOHNSTONE, I. (2001). Chi-square oracle inequalities, in *State of the Art in Probability and Statistics, Festschrift for Willem R. van Zwet* (Mathisca C.M. de Gunst, Chris A.J. Klaassen, Aad W. van der Vaart, eds.), Institute of Mathematical Statistics, Lecture Notes–Monograph Series **36**, 399-418.

KERKYACHARIAN, G. and PICARD, D. (2000). Thresholding algorithms, maxisets and well-concentrated bases. *Test* **9**, 283-344.

KOLMOGOROV, A.N. and TIKHOMIROV, V.M. (1961). ε -entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Transl. (2)* **17**, 277-364.

LAURENT, B. and MASSART, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 1302-1338.

Le CAM, L.M. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Stat.* **41**, 802-828.

Le CAM, L.M. (1972). Limits of experiments. *Proc. 6th Berkeley Symp. on Math. Stat. and Prob., I*, 245-261.

Le CAM, L.M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** , 38-53.

Le CAM, L.M. (1975). On local and global properties in the theory of asymptotic normality of experiments. *Stochastic Processes and Related Topics, Vol. 1* (M. Puri, ed.), 13-54. Academic Press, New York.

Le CAM, L.M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

Le CAM, L.M. (1990). Maximum likelihood: an introduction. *Inter. Statist. Review* **58** , 153-171.

Le CAM, L.M. (1997). Metric dimension and statistical estimation. *CRM Proc. and*

Lecture Notes **11** , 303-311.

LORENTZ, G.G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Society* **72**, 902-937.

LORENTZ, G.G., von GOLITSCHEK, M. and MAKOVOZ, Y. (1996). *Constructive Approximation, Advanced Problems*. Springer, Berlin.

NEMIROVSKI, A.S. (2000). Topics in Non-Parametric Statistics. In *Lecture on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998* (P. Bernard, ed.), 85-297. Lecture Note in Mathematics 1738, Springer-Verlag, Berlin.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24**, 2399-2430.

PINKUS, A. (1985). *n-widths in Approximation Theory*. Springer-Verlag, Berlin.

PINSKER, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems of Information Transmission* **16**, 120-133.

SHEN, X. and WONG, W.H. (1994). Convergence rates of sieve estimates. *Ann. Statist.* **22**, 580-615.

SILVERMAN, B.W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795-810.

TSYBAKOV, A. (2003). Optimal Rates of Aggregation. In *Proceedings of COLT-2003*. Lecture Note in Computer Science, Springer-Verlag, Berlin (to appear).

van de GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907-924.

van de GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimates. *Ann. Statist.* **21**, 14-44.

van de GEER, S. (2000). *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.

van der VAART, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

WAHBA, G. (1990). *Spline Models for Observational Data*. S.I.A.M., Philadelphia.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** , 595-601.

WEGKAMP, M. (2003). Model selection in nonparametric regression. *Ann. Statist.* **31**, 252-273.

WONG, W.H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339-362.

YANG, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75-87.

YANG, Y. and BARRON, A.R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* **44**, 95-116.

YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564-1599.

YATRACOS, Y.G. (1985). Rates of convergence of minimum distance estimates and Kolmogorov's entropy. *Ann. Statist.* **13**, 768-774.

Lucien BIRGÉ
UMR 7599 "Probabilités et modèles aléatoires"
Laboratoire de Probabilités, boîte 188
Université Paris VI, 4 Place Jussieu
F-75252 Paris Cedex 05
France
e-mail: LB@CCR.JUSSIEU.FR