*Bernoulli* **10**(6), 2004, 1039–1051

# Model selection for Gaussian regression with random design

LUCIEN BIRGÉ

*Laboratoire de Probabilités et Modèleles Aléatoires, UMR CNRS 7599, Boîte 188, Université Paris VI, 4 Place Jussieu, F-75252 Paris Cedex 05, France. E-mail: lb@ccr.jussieu.fr*

This paper is concerned with Gaussian regression with random design, where the observations are independent and indentically distributed. It is known from work by Le Cam that the rate of convergence of optimal estimators is closely connected to the metric structure of the parameter space with respect to the Hellinger distance. In particular, this metric structure essentially determines the risk when the loss function is a power of the Hellinger distance. For random design regression, one typically uses as loss function the squared $\mathbb{L}_2$-distance between the estimator and the parameter. If the parameter space is bounded with respect to the $\mathbb{L}_\infty$-norm, both distances are equivalent. Without this assumption, it may happen that there is a large distortion between the two distances, resulting in some unusual rates of convergence for the squared $\mathbb{L}_2$-risk, as noticed by Baraud. We explain this phenomenon and then show that the use of the Hellinger distance instead of the $\mathbb{L}_2$-distance allows us to recover the usual rates and to carry out model selection in great generality. An extension to the $\mathbb{L}_2$-risk is given under a boundedness assumption similar to that given by Wegkamp and by Yang.

*Keywords:* Besov spaces; Hellinger distance; minimax risk; model selection; random design regression

## 1. Introduction

One classical method for estimating an unknown (density or regression) function $s$ from $n$ observations is to construct a parametric model for $s$, i.e. a set $S$ of functions described by $D$ parameters, and proceed with the estimation as if $s$ actually belonged to the model, which results in an estimator $\hat{s} \in S$. The corresponding error is then the sum of a bias term, which comes from the fact that, typically, $s \notin S$ and a stochastic error corresponding to estimation within the model $S$ which is usually proportional to the number $D$ of parameters. Since $s$ is unknown, we generally do not have enough information to construct a good model leading to an estimator with a small risk. One way of solving this problem is to start with a large family $\{S_m, m \in \mathcal{M}\}$ of such models and the corresponding family of estimators $\{\hat{s}_m, m \in \mathcal{M}\}$, and then to use the data to select one model or equivalently one estimator in the family. This is what is called *model selection*. It can often be viewed as a *variable selection procedure* that selects a small number of functions (the variables) from a large set of such functions with cardinality possibly much bigger than the number of observations. The final estimator is then a combination of the selected functions. A simple example, when $s$ is defined on [0, 1], is as follows. One considers the points $x_j = jn^{-2}$ with $j \in \mathbb{N}$, $j \leqslant n^2$, and all functions of the form $\mathbb{1}_{[x_j, x_k)}$ with $j < k$. There are of course many

of them and one considers the models generated by linear combinations of a small number of these functions. We therefore look for a sparse representation of $s$ by combinations of some functions taken in a possibly very large set.

Many papers have been devoted to model selection for various statistical problems and we shall focus here on random design regression, i.e. a statistical framework in which we observe $n$ independent and identically distributed (i.i.d.) random pairs $(X_i, Y_i)$ with $Y_i \in \mathbb{R}$ and $X_i$ belonging to some measurable space $\mathcal{X}$ (typically a subset of $\mathbb{R}^k$). We assume that $X_i$ and $Y_i$ satisfy the relationship

$$Y_i = s(X_i) + \varepsilon_i, \qquad 1 \leqslant i \leqslant n, \tag{1.1}$$

where the random variables $\varepsilon_i$ are i.i.d., centred and independent of the $X_i$, and $s$ is an unknown function (parameter) to be estimated. We denote by $\mu$ the common distribution of the $X_i$ and assume hereafter that the unknown parameter $s$ belongs to some subset $\mathcal{S}$ of $\mathbb{L}_2(\mu)$ and that the distribution of $\varepsilon_i$ is normal with known variance $\sigma^2$. This is the precise framework that we shall call *Gaussian regression with random design*, denoting by $\|\cdot\|$ and $\|\cdot\|_\infty$ the norms in $\mathbb{L}_2(\mu)$ and $\mathbb{L}_\infty(\mu)$, respectively.

Although this problem is considered as a regression problem, it is technically closer to a density estimation problem since we have at hand $n$ i.i.d. observations with unknown distribution $P_s$ and density

$$\frac{\mathrm{d}P_s}{\mathrm{d}(\mu \otimes \lambda)}(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - s(x))^2}{2\sigma^2}\right) \tag{1.2}$$

with respect to $\mu \otimes \lambda$, where $\lambda$ denotes the Lebesgue measure on the real line. The main difference from classical density estimation is that the density depends in a complicated way on the infinite-dimensional parameter $s$ of interest. As we shall see, the difficulties connected with this model are mainly due to the distortion between the natural distance for the density problem, namely the Hellinger distance between distributions $P_s$ for $s \in \mathcal{S}$, and the $\mathbb{L}_2$-distance on the set $\mathcal{S}$ of parameters.

Despite the similarity of terminology, the previous framework appears to be technically different from the *Gaussian regression with fixed design* framework, in which one observes

$$Y_i = s(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\big(0, \sigma^2\big), \qquad 1 \leqslant i \leqslant n, \tag{1.3}$$

where the values $x_i \in \mathcal{X}$ are fixed and known and which merely corresponds to the estimation of the mean $\bar{s} = (s(x_1), \ldots, s(x_n))$ of the Gaussian vector $(Y_1, \ldots, Y_n)$ with distribution $Q_{\bar{s}}$. In this problem, the natural parameter set is not the function space $\mathcal{S}$ but its image $\mathcal{T} = \{\bar{s} = (s(x_1), \ldots, s(x_n)) | s \in \mathcal{S}\} \subset \mathbb{R}^n$. Since the Hellinger distance between the distributions $Q_{\bar{s}}$, $\bar{s} \in \mathcal{T}$, and the Euclidean distance on $\mathcal{T}$ associated with the norm $\|\bar{s}\|_n = n^{-1}\sum_{i=1}^n s(x_i)^2$ are well connected, the estimation of $\bar{s}$ does not lead to special difficulties, as shown in Birgé and Massart (2001) and Birgé (2003). If we wish to estimate the function $s$ itself, we have to reconstruct it from its values at the points $x_i$ which is clearly not always possible and, in any case, is a problem in approximation theory. Illustrations can be found in Baraud (2000).

Some specific problems connected with the random design framework (for the simplest case where the $X_i$ belong to $[0, 1]$) were brought to our attention by a recent paper of

Baraud (2002) and a conversation with the author. Classically, when estimating $s$, one uses the squared $\mathbb{L}_2$-distance as the loss function, which results in a risk function $\mathbb{E}_s\left[\|s-\hat{s}\|^2\right]$, where $\mathbb{E}_s$ denotes the expectation when $s$ obtains. Most results about model selection in this case require some boundedness assumption on both the parameter and the estimators, as in Juditsky and Nemirovski (2000), Yang (2000; 2001; 2004) and Wegkamp (2003). In principle, a precise bound on $\|s\|_\infty$ need not be known in order to construct the selection procedure, nevertheless, from a realistic point of view, the upper bound on the estimators can only be chosen in a reasonable way by the statistician if a bound on $\|s\|_\infty$ is known, at least approximately. Otherwise, one could choose too small a bound, which would result in high bias, or too large, which deteriorates the performance of the estimators. As we shall see below, if we wish to handle arbitrary parameter spaces, such upper bounds are more or less necessary when dealing with the squared $\mathbb{L}_2$-risk.

There are two noticeable exceptions to the use of upper bounds on the parameter space, which are Brown *et al*. (2002) and Baraud (2002). The former paper deals with equivalence of experiments between regression with random design and the white noise model, but equivalence only holds for compact balls in Hölder or Sobolev spaces of smoothness $\alpha > 1/2$, while we shall show below that problems occur when $\alpha < 1/2$. We recall here for further reference that the white noise model on $[0, 1]$ corresponds to observation of the process $Y_z$, $z \in [0, 1]$, where

$$Y_z = \int_0^z s(x)\mathrm{d}x + n^{-1/2} W(z), \qquad z \in [0, 1], \tag{1.4}$$

$W$ denotes the standard Brownian motion and $s \in \mathbb{L}_2([0, 1], \mathrm{d}x)$ is the unknown parameter to be estimated. The risk of an estimator $\hat{s}$ is again given by $\mathbb{E}_s\left[\|s-\hat{s}\|^2\right]$, with $\|\cdot\|$ the norm in $\mathbb{L}_2([0, 1], \mathrm{d}x)$.

Among more general results on model selection, Baraud (2002) recovers the usual $n^{-2\alpha/(2\alpha+1)}$ rate for estimation in Besov spaces $B^\alpha_{p,\infty}$ with index $\alpha$ larger than some limiting value $\alpha_l \in (1/p - 1/2, 1/p)$ when $1 \leqslant p < 2$, the precise value of $\alpha_l$ being given in Proposition 2 below. In the white noise model one derives this rate for $\alpha > 1/p - 1/2$ as in Donoho and Johnstone (1994). This limiting valve $\alpha_l$ of Baraud appears to be rather surprising and, in a private conversation, the author explained that, for $\alpha \leqslant \alpha_l$, he was still able to derive rates of convergence but which were 'suboptimal', i.e. slower than the usual ones, although he suspected they were unimprovable, apart from some logarithmic factors. According to Kerkyacharian (private conversation), it follows from Theorem 6.1 of Kerkyacharian and Picard (2000) that a similar limitation in the range of $\alpha$ holds for procedures based on thresholding of wavelet coefficients in density estimation.

We shall show below that this cut in the rates is actually unavoidable and due to the distortion between the $\mathbb{L}_2$ and Hellinger distances. As a consequence, we derive a lower bound for the rates of convergence over Besov balls with $\alpha \leqslant \alpha_l$, which coincides, up to logarithmic factors, with the upper bounds obtained by Baraud (personal communication). We then show that, if we use the squared Hellinger loss instead of the $\mathbb{L}_2$-loss, we are able to recover the usual rate in the range $\alpha > 1/p - 1/2$, provided that the unknown parameter $s$ belongs to $\mathbb{L}_\infty(\mu)$, although no information is required about $\|s\|_\infty$. In this case, the estimator that we construct belongs to $\mathbb{L}_\infty(\mu)$ by construction but may also have an arbitrary $\mathbb{L}_\infty$-norm.

The situation becomes quite different when we wish to carry out model selection using the squared $\mathbb{L}_2$-loss since we can only handle this problem when our estimator $\hat{s}$ is constructed in such a way that it satisfies $\|\hat{s}\|_\infty \leqslant A$ almost surely, which means that it should be constructed in order to belong to some given $\mathbb{L}_\infty$-ball of radius $A$ chosen by the statistician. In this case we have to know $\|s\|_\infty$ approximately in order to make an adequate choice of $A$.

# 2. The importance of boundedness assumptions

## 2.1. The relationships between Hellinger and $\mathbb{L}_2$-distances

We recall that if $P$ and $Q$ are two distributions, their Hellinger affinity $\rho(P, Q)$ and Hellinger distance $h(P, Q)$ are given by

$$\rho(P, Q) = \int \sqrt{\mathrm{d}P\,\mathrm{d}Q} = 1 - h^2(P, Q) \quad \text{with} \quad h^2(P, Q) = \frac{1}{2}\int \left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q}\right)^2.$$

If, in particular, $Q_{\bar{s}}$ denotes the joint distribution of the variables $Y_i$, $1 \leqslant i \leqslant n$, in the Gaussian regression with fixed design framework, as defined by (1.3), and $P_s$ the distribution of the process $Y_z$ in the white noise model (1.4), we obtain

$$\rho(Q_{\bar{t}}, Q_{\bar{u}}) = \exp\left(-\frac{\|t-u\|_n^2}{8\sigma^2}\right) \quad \text{and} \quad \rho(P_t, P_u) = \exp\left(-\frac{\|t-u\|^2}{8\sigma^2}\right),$$

and in both cases, $[-\log \rho(Q_{\bar{t}}, Q_{\bar{u}})]^{1/2}$ and $[-\log \rho(P_t, P_u)]^{1/2}$ are equivalent to the corresponding $\mathbb{L}_2$-type distances. In Gaussian regression with random design, the distribution $P_s$ of $(X_i, Y_i)$ has the density (1.2), which implies that

$$\rho(P_t, P_u) = \mathbb{E}\left[\exp\left(-\frac{(t-u)^2(X)}{8\sigma^2}\right)\right] = \int_0^1 \exp\left(-\frac{(t-u)^2(x)}{8\sigma^2}\right)\mathrm{d}\mu(x) \tag{2.1}$$

and

$$h^2(P_t, P_u) = 1 - \int_0^1 \exp\left(-\frac{(t-u)^2(x)}{8\sigma^2}\right)\mathrm{d}\mu(x). \tag{2.2}$$

We then derive from Jensen's inequality that

$$\rho(P_t, P_u) \geqslant \exp\left(-\frac{\|t-u\|^2}{8\sigma^2}\right) \quad \text{and} \quad h^2(P_t, P_u) \leqslant \frac{\|t-u\|^2}{8\sigma^2}. \tag{2.3}$$

Unfortunately, the reverse inequalities do not hold in general and $(-\log \rho)^{1/2}$ is not a distance as it is in either the Gaussian regression with fixed design framework (1.3) or the white noise model (1.4). It does not even satisfy an inequality of the form

$$\sqrt{-\log[\rho(P_t, P_u)]} \leqslant A\left(\sqrt{-\log[\rho(P_t, P_s)]} + \sqrt{-\log[\rho(P_s, P_u)]}\right), \tag{2.4}$$

for some (possibly large) positive constant $A$, uniformly with respect to $s$, $t$ and $u$. To see

this, let us assume that the $X_i$ are uniformly distributed on [0, 1] and consider the three functions $s$, $t = s + \gamma \mathbb{1}_{[0,1/2)}$ and $u = s - \gamma \mathbb{1}_{[1/2,1]}$ for some $\gamma > 0$. Then $t - u = \gamma \mathbb{1}_{[0,1]}$ and $\rho(P_t, P_u) = \exp\left[-\gamma^2/(8\sigma^2)\right]$, while

$$\rho(P_t, P_s) = \rho(P_s, P_u) = \frac{1}{2}\left[1 + \exp\left(-\frac{\gamma^2}{8\sigma^2}\right)\right] > \frac{1}{2}.$$

It follows that

$$\sqrt{-\log\left[\rho(P_t, P_s)\right]} + \sqrt{-\log\left[\rho(P_s, P_u)\right]} < 2\sqrt{\log 2},$$

while

$$\sqrt{-\log\left[\rho(P_t, P_u)\right]} = \frac{\gamma}{2\sigma\sqrt{2}}.$$

Therefore

$$\frac{\sqrt{-\log\left[\rho(P_t, P_u)\right]}}{\sqrt{-\log\left[\rho(P_t, P_s)\right]} + \sqrt{-\log\left[\rho(P_s, P_u)\right]}} > \frac{\gamma}{4\sigma\sqrt{2\log 2}}$$

tends to infinity with $\gamma$, proving that $(-\log \rho)^{1/2}$ cannot satisfy (2.4) whatever the value of $A$.

This phenomenon being due to the fact that $\|t - u\|_\infty$ can be arbitrary large, the situation changes if the parameters are restricted to belong to some $\mathbb{L}_\infty(\mu)$-ball.

**Proposition 1.** *Let $\mathcal{S}$ be a subset of some $\mathbb{L}_\infty(\mu)$-ball with centre $s_0$ and radius $r\sigma$, i.e. $\|s - s_0\|_\infty \leqslant r\sigma$ for all $s \in \mathcal{S}$. Then, for any $t$ and $u$ in $\mathcal{S}$,*

$$\rho(P_t, P_u) \leqslant 1 - \frac{\left[1 - \exp\left(-r^2/2\right)\right]\|t - u\|^2}{4r^2\sigma^2}. \tag{2.5}$$

*Consequently,*

$$\|t - u\| < 2.03(r \vee \mathrm{e})\sigma\, h(t, u). \tag{2.6}$$

***Proof.*** The second inequality being an easy consequence of the first since $h^2 = 1 - \rho$, it suffices to prove (2.5). We notice that $(t - u)^2(x)/(8\sigma^2) \leqslant r^2/2$ $\mu$ almost surely and use the fact that, if $Y$ is a random variable with values in $[0, M]$ and distribution $P_Y$, then

$$1 - \left(\int \exp(-y) 1 \mathrm{d}P_Y(y)\right) \geqslant M^{-1}(1 - \mathrm{e}^{-M})\mathbb{E}[Y].$$

This last inequality follows from the convexity of the function $x \mapsto \mathrm{e}^{-x}$ by integration of

$$\mathrm{e}^{-Y} = \mathrm{e}^{-(Y/M)M} \leqslant (Y/M)\mathrm{e}^{-M} + (1 - Y/M)\mathrm{e}^0. \qquad \square$$

As a consequence, if we restrict ourselves to a parameter space $\mathcal{S}$ contained in some $\mathbb{L}_\infty$-ball, (2.4) is satisfied for a suitable value of $A$ and $h$ is equivalent to the $\mathbb{L}_2$-distance on $\mathcal{S}$. This is the case considered by most authors and the easiest one. Indeed, since this is an estimation problem with $n$ i.i.d. observations, it can be handled by the techniques of Birgé (1983; 2003), resulting in control of the squared Hellinger risk for a suitable estimator $\hat{s}$. If this estimator

belongs to the same $\mathbb{L}_\infty$-ball, $\|s-\hat{s}\|$ can be bounded by $h(s,\hat{s})$ times a constant and the squared $\mathbb{L}_2$-risk is also under control. We shall explain more precisely in Section 3.2 how to make this work.

## 2.2. Some negative results

If $\|t-u\|_\infty$ is not bounded, the ratio $\|t-u\|/h(P_t, P_u)$ can be arbitrarily large and this accounts for the difficulties of evaluating the $\mathbb{L}_2$-risk in this situation and the results obtained by Baraud (2002) for the minimax risk over Besov balls of functions on [0, 1], when $\mu$ is the Lebesgue measure on [0, 1].

To understand what is going on, let us introduce the function $t=a\mathbb{1}_{[0,l]}$ with $a>0$ and $l\leqslant 1/2$ and the numbers $\alpha$, $p$ with $1\leqslant p<2$ and $\alpha=1/p-b$, $0<b<1/2$. We can then compute the $\mathbb{L}_p$-modulus of continuity $\omega(t,x)_p$ of $t$. According to DeVore and Lorentz (1993, p. 44),

$$\omega(t,x)_p=\sup_{0\leqslant h\leqslant x}\left[\int_0^{1-h}|t(y+h)-t(y)|^p\,\mathrm{d}y\right]^{1/p}=a(x\wedge l)^{1/p}.$$

It follows that the Besov seminorm of $t$ with respect to the Besov space $B^\alpha_{p,\infty}$ is

$$|t|^\alpha_p=\sup_{x>0}x^{-\alpha}\omega(t,x)_p=a\sup_{x>0}\{x^{1/p-\alpha}\wedge x^{-\alpha}l^{1/p}\}=al^{1/p-\alpha}=al^b. \tag{2.7}$$

Setting $u=-t$, we see that $\|t-u\|^2=4a^2l$ and, by (2.1),

$$\rho(P_t,P_u)=\int_0^l\exp\left(\frac{-4a^2}{8\sigma^2}\right)\mathrm{d}x+(1-l).$$

Then

$$h^2(P_t,P_u)=l\left[1-\exp\left(\frac{-a^2}{2\sigma^2}\right)\right]=l\left[1-\exp\left(\frac{-\|t-u\|^2}{8l\sigma^2}\right)\right].$$

For moderate values of $a/\sigma$ this is of the order of $\|t-u\|^2/\sigma^2$, but for large values of $a/\sigma$ it behaves like $l$, independently of $a$. In this case the ratio $\|t-u\|/h(P_t,P_u)$ is of order $a$ and can be arbitrarily large.

Let us now set $l=(2n)^{-1}$, $a=R(2n)^b$ with $R>0$. Then

$$h^2(P_t,P_u)<(2n)^{-1},\qquad \rho(P_t^n,P_u^n)>\left[1-(2n)^{-1}\right]^n\geqslant 1/2$$

and $\|t-u\|^2=4R^2(2n)^{2b-1}$. On the one hand, it follows from classical lower bounds on the risk dating back to Le Cam (1973) – see, for instance, Donoho and Liu (1991) – that any estimator $\hat{s}$ based on $n$ i.i.d. observations from (1.1) satisfies

$$\max\left\{\mathbb{E}_t\left[\|t-\hat{s}\|^2\right],\mathbb{E}_u\left[\|u-\hat{s}\|^2\right]\right\}\geqslant cR^2(2n)^{2b-1}, \tag{2.8}$$

for some positive universal constant $c$. On the other hand, by (2.7), the Besov seminorm of $t$ and $u$ is $R$. Therefore, by (2.8), for any estimator $\hat{s}$,

$$\sup_{\{s|\,|s|_p^\alpha \leqslant R\}} \mathbb{E}_s\left[\|s-\hat{s}\|^2\right] \geqslant cR^2(2n)^{2b-1} = c'R^2 n^{-1+2(1/p-\alpha)}. \tag{2.9}$$

We recall that the minimax risk over such Besov balls, in the white noise model (1.4), is known to be bounded by $CR^{2/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}$. It follows that the minimax risk in the regression model will be substantially larger, at least for large $n$, if $2(1/p-\alpha) > 1/(2\alpha+1)$ or equivalently $(1/p-\alpha)(2\alpha+1) > 1/2$. Elementary computations show that this is equivalent to

$$\frac{1}{2}\left[\frac{1}{p}-\frac{1}{2}-\sqrt{\left(\frac{1}{p}+\frac{1}{2}\right)^2-1}\right] < \alpha < \frac{1}{2}\left[\frac{1}{p}-\frac{1}{2}+\sqrt{\left(\frac{1}{p}+\frac{1}{2}\right)^2-1}\right].$$

The left-hand side is smaller than $1/p-1/2$ and one can easily check that the right-hand side is smaller than $1/p$. We have thus proved the following proposition:

**Proposition 2.** *For* $1 \leqslant p < 2$ *and*

$$\frac{1}{p}-\frac{1}{2} < \alpha < \alpha_l = \frac{1}{2}\left[\frac{1}{p}-\frac{1}{2}+\sqrt{\left(\frac{1}{p}+\frac{1}{2}\right)^2-1}\right],$$

*the rate of convergence, with respect to n, of the minimax risk over Besov balls of the form* $\{s|\,|s|_p^\alpha \leqslant R\}$ *cannot be better than* $n^{-1+2(1/p-\alpha)}$.

For $\alpha > \alpha_l$, Baraud (2002) recovers the usual $n^{-2\alpha/(2\alpha+1)}$ rate and his proof, when applied to the case $\alpha \leqslant \alpha_l$, gives an upper bound for the risk of order $n^{-1+2(1/p-\alpha)}$, up to some extra logarithmic factors (Baraud, personal communication), which means that, up to logarithmic factors, the rate $n^{-1+2(1/p-\alpha)}$ is minimax for $\alpha \leqslant \alpha_l$.

It is also worth noticing that a similar distortion of the risk, as compared to the white noise model, occurs when $R$ goes to infinity, for any $\alpha < 1$ and $n$, because of the $R^2$ factor in (2.9) instead of the usual $R^{2/(2\alpha+1)}$.

# 3. Upper bounds for the risk

As we mentioned in the Introduction, the problem of random design regression is a problem of estimation from an i.i.d. sample and the 'natural' loss function for such a problem is the Hellinger distance, as shown by Le Cam (1973; 1975; 1986). The distortion that may exist between $\mathbb{L}_2$ and Hellinger distances when dealing with unbounded functions leads to difficulties, as indicated in the previous section. Nevertheless, if we consider the squared Hellinger risk, instead of the squared $\mathbb{L}_2$-risk, we can essentially recover the usual rates of convergence if we assume that the true parameter belongs to $\mathbb{L}_\infty(\mu)$.

In order to prove this we need to recall some general results from the present author about model selection for i.i.d. variables. The framework is as follows: we observe $n$ i.i.d. random variables $Z_1, \ldots, Z_n$ on the measurable space $\mathcal{Z}$ with some unknown distribution

$P_s$, with $s$ belonging to some parameter set $M$, assuming that the mapping $s \mapsto P_s$ is one-to-one, which allows us to identify $M$ with a subset of the set of all distributions on $\mathcal{Z}$. Setting $h(t, u) = h(P_t, P_u)$ turns $M$ into a metric space with the Hellinger metric, and we shall denote by $\mathcal{B}_h(t, r)$ the open Hellinger ball with centre $t$ and radius $r$ in $M$. We also introduce a finite or countable family $\{S_m, m \in \mathcal{M}\}$ of discrete subsets of $M$. In what follows, $|S|$ denotes the cardinality of the set $S$. We proved in Birgé (2003) the following result.

**Theorem 1.** *Assume that, for each $m \in \mathcal{M}$, one can find numbers $\eta_m > 0$ and $D_m \geqslant 1/2$ such that*

$$|\{S_m \cap \mathcal{B}_h(t, x\eta_m)\}| \leqslant \exp\left(x^2 D_m\right) \qquad \text{for all } t \in M \quad \text{and} \quad x \geqslant 2. \tag{3.1}$$

*Assume, moreover, that the numbers $\eta_m$ and $D_m$ satisfy*

$$\eta_m^2 \geqslant 16.8 \frac{D_m}{n} \quad \text{for all } m \in \mathcal{M} \quad \text{and} \quad \sum_{m \in \mathcal{M}} \exp\left(-\frac{n\eta_m^2}{84}\right) = \Sigma < +\infty. \tag{3.2}$$

*Then one can construct an estimator $\hat{s} \in \bigcup_{m \in \mathcal{M}} S_m$ such that, for all $s \in M$,*

$$\mathbb{E}_s\left[h^2(s, \hat{s})\right] \leqslant 29[1 + 10^{-8}\Sigma] \inf_{m \in \mathcal{M}} \left\{h^2(s, S_m) \vee \eta_m^2\right\}, \tag{3.3}$$

*where $h(s, S_m) = \inf_{t \in S_m} h(s, t)$.*

## 3.1. Hellinger risk

Let us now return to our initial problem of estimating $s \in \mathbb{L}_\infty(\mu)$ with $n$ observations from (1.1). From now on, we shall denote by $d$ the $\mathbb{L}_2$-distance ($d(t, u) = \|t - u\|$) and by $\mathcal{B}_d$ the corresponding open balls. We shall also introduce the following definition

***Definition 1.*** *Let $S$ be a subset of some metric space $(M, d)$. We say that it has a Euclidean metric dimension bounded by $D$ (for the metric $d$) if, for any $\eta > 0$, one can find an $\eta$-net $S_\eta$ for $S$ (i.e. a subset of $M$ such that $d(s, S_\eta) \leqslant \eta$ for all $s \in S$) and, for any $t \in M$,*

$$|S_\eta \cap \mathcal{B}_d(t, x\eta)| \leqslant x^D \qquad \text{for all } x \geqslant 2.$$

In particular, one can check that if $S$ is a $k$-dimensional linear subspace of some Hilbert space $(M, d)$, its Euclidean dimension is bounded by $\log 3/\log 2$ if $k = 1$ and by $(\log 5/\log 2)k$ if $k > 1$. We can now prove:

**Theorem 2.** *Assume that we have at hand a finite or countable family $\{S'_{m'}\}_{m' \in \mathcal{M}'}$ of subsets of $\mathbb{L}_2(\mu)$ with respective Euclidean metric dimensions bounded by $\overline{D}_{m'} \geqslant 3/2$ and let $\{\Delta_{m'}\}_{m' \in \mathcal{M}'}$ be a family of non-negative weights satisfying*

$$\sum_{m' \in \mathcal{M}'} \exp\left(-\Delta_{m'}\right) = \Sigma' < +\infty. \tag{3.4}$$

*There exists an estimator $\hat{s}$ and a universal constant $C$ such that, for any $s \in \mathbb{L}_\infty(\mu)$,*

$$\mathbb{E}_s\bigl[h^2(s, \hat{s})\bigr] \leqslant C\bigl[1 + 10^{-7}\Sigma'\bigr] \inf_{m' \in \mathcal{M}'} \left\{ \frac{d^2(s, S'_{m'})}{\sigma^2} + \frac{\overline{D}_{m'} \vee \Delta_{m'}}{n} \left[ \log\left( \frac{\|s\|_\infty}{\sigma} \right) \vee 1 \right] \right\}.$$

***Proof.*** We wish to apply Theorem 3 to our situation, taking for $M$ the set of all distributions $P_s$ of $(X_i, Y_i)$, as given by (1.1), when $s \in \mathbb{L}_\infty(\mu)$. In order to do this, we introduce a *stratification* method, which involves replacing each initial model by several ones with specific properties. This method is not new: it appears in Yang and Barron (1998) and was used in Birgé (2003).

We first introduce a new index set $\mathcal{M} = \{(m', j),\ m' \in \mathcal{M}',\ j \in \mathbb{N}^\star\}$ ($\mathbb{N}^\star = \mathbb{N}\backslash\{0\}$) and define, for each $m = (m', j) \in \mathcal{M}$, $\eta_m = \sqrt{j}\eta_{m'}$ with $\eta_{m'} = [(16.8/n)(\overline{D}_{m'} \vee 5\Delta_{m'})]^{1/2}$. Then

$$\sum_{m \in \mathcal{M}} \exp\left( -\frac{n\eta_m^2}{84} \right) = \sum_{m' \in \mathcal{M}'} \exp\left( -\frac{n\eta_{m'}^2}{84} \right) \sum_{j \geqslant 1} \exp\left[ -(j-1)\frac{n\eta_{m'}^2}{84} \right]$$

$$\leqslant \sum_{m' \in \mathcal{M}'} \exp(-\Delta_{m'}) \sum_{j \geqslant 1} \exp\left[-0.3(j-1)\right],$$

where we have used the fact that $n\eta_{m'}^2/84 \geqslant \overline{D}_{m'}/5 \geqslant 0.3$. The second inequality in (3.2) is therefore satisfied with $\Sigma = \Sigma'/(1 - \mathrm{e}^{-0.3}) < 4\Sigma'$. We then define, for each $j \in \mathbb{N}^\star$, the operator $\theta_j$ from $\mathbb{L}_2(\mu)$ to $\mathbb{L}_\infty(\mu)$ by $\theta_j(t) = (t \wedge \sigma\mathrm{e}^j) \vee (-\sigma\mathrm{e}^j)$, which implies that $\|\theta_j(t)\|_\infty \leqslant \sigma\mathrm{e}^j$ for all $t$ and $j$. Given $m = (m', j) \in \mathcal{M}$, by assumption one can find a $\sigma\eta_m$-net $T_m$ for $S'_{m'}$ such that, for any $t \in \mathbb{L}_2(\mu)$,

$$|T_m \cap \mathcal{B}_d(t, x\sigma\eta_m)| \leqslant x^{\overline{D}_{m'}} \qquad \text{for all } x \geqslant 2. \tag{3.5}$$

We then set $T'_m = \{t \in T_m | d(t, \theta_j(t)) \leqslant 4\sigma\eta_m\}$ and $S_m = \{\theta_j(t),\ t \in T'_m\} \subset \mathbb{L}_\infty(\mu)$. If $S_m$ is empty, we remove it from the collection. It follows from (2.6) with $r = \mathrm{e}^j$ that, if $t$ and $u$ belong to $S_m$,

$$d(t, u) < 2.03\sigma\mathrm{e}^j h(t, u). \tag{3.6}$$

For $x \geqslant 2$ and $u \in \mathbb{L}_\infty(\mu)$ we consider $\mathcal{B} = \mathcal{B}_h(u, x\eta_m)$ and wish to bound $|\mathcal{B} \cap S_m|$ in order to check (3.1). Since there is nothing to prove if this is empty, we may assume that the intersection contains at least one point $u'$ and therefore, by (3.6),

$$\mathcal{B} \cap S_m \subset \mathcal{B}_h(u', 2x\eta_m) \cap S_m \subset \mathcal{B}_d\bigl(u', 4.06\sigma\mathrm{e}^j x\eta_m\bigr) \cap S_m. \tag{3.7}$$

Since, for any $t' \in S_m$, one can find some $t \in T_m$ with $t' = \theta_j(t)$ and $d(t, t') \leqslant 4\sigma\eta_m$,

$$\log|\mathcal{B}_d\bigl(u', 4.06\sigma\mathrm{e}^j x\eta_m\bigr) \cap S_m| \leqslant \log|\mathcal{B}_d\bigl(u', 4.06\sigma\mathrm{e}^j x\eta_m + 4\sigma\eta_m\bigr) \cap T_m|$$

and, by (3.5) and (3.7),

$$\log|\mathcal{B} \cap S_m| \leqslant \overline{D}_{m'}\log\bigl(4.06\mathrm{e}^j x + 4\bigr) < j\overline{D}_{m'}x^2 \quad \text{for } x \geqslant 2.$$

It follows that we can take $D_m = j\overline{D}_{m'}$ in (3.1) and that $\eta_m^2 \geqslant 16.8D_m/n$ as required for (3.2). Applying Theorem 1, we obtain for all $s \in \mathbb{L}_\infty(\mu)$,

$$\mathbb{E}_s\left[h^2(s, \hat{s})\right] \leqslant 29\left[1 + (4 \times 10^{-8}\Sigma')\right] \inf_{m \in \mathcal{M}} \left\{h^2(s, S_m) \vee \eta_m^2\right\}. \tag{3.8}$$

Now let $s$ and $m'$ be given and $j$ be the smallest positive integer satisfying $\|s\|_\infty \leqslant \sigma \mathrm{e}^j$ and $\mathrm{d}(s, S'_{m'}) \leqslant \sigma\sqrt{j}\eta_{m'}$. If $m = (m', j)$, then there exists $t \in T_m$ with $d(s, t) \leqslant 2\sigma\eta_m$. Since $\|s\|_\infty \leqslant \sigma \mathrm{e}^j$, obviously $d(s, \theta_j(t)) \leqslant d(s, t)$, hence $d(t, \theta_j(t)) \leqslant 2d(s, t) \leqslant 4\sigma\eta_m$ and $\theta_j(t) \in S_m$. Finally, $d(s, S_m) \leqslant 2\sigma\eta_m$ and $h^2(s, S_m) \leqslant d^2(s, S_m)/(8\sigma^2) \leqslant \eta_m^2/2$ by (2.3). The definition of $j$ (distinguishing between the cases $j = 1$ and $j > 1$) implies that

$$\eta_m^2 \leqslant 2\left\{\left(\eta_{m'}^2[\log(\|s\|_\infty/\sigma) \vee 1/2]\right) \vee \left(d(s, S'_{m'})/\sigma\right)^2\right\}.$$

Substitution in (3.8) leads, after some simplifications, to the desired bound for the Hellinger risk. □

The important point, in this result, is that the estimator $\hat{s}$ is universal in the sense that it only depends on the family $\{S'_{m'}\}_{m' \in \mathcal{M}'}$ and $\sigma$ and not on some prior upper bound on $\|s\|_\infty$ as is the case in most papers on the subject dealing with the $\mathbb{L}_2$-risk, and it is the use of the Hellinger distance that makes it possible. We are unable to obtain similar results for the $\mathbb{L}_2$-risk.

Note that, given some arbitrary measurable function $s_0$ on $\mathcal{X}$, we could alternatively, since all distances involved only depend on differences, base our construction on functions $t + s_0$ with $t \in S'_{m'}$. This would not change anything apart from the fact that the final result would involve $d(s - s_0, S'_{m'})$ and $\|s - s_0\|_\infty$. This can be useful if we suspect that the true $s$ is close to some known function $s_0$.

***A typical application.*** Of course, such a theorem has many applications and many model selection procedures which have been considered in previous papers of the author, such as Birgé and Massart (1997), Barron *et al.* (1999) and Birgé and Massart (2001), can be extended to the present regression framework since they are based on approximations by finite-dimensional linear spaces $S'_{m'}$ which therefore satisfy the assumptions of Theorem 2. In particular, all the strategies considered in Section 6 of Birgé and Massart (2001) can be transfered to the framework we study here. We shall content ourselves with considering the example of adaptation for arbitrary Besov balls, as discussed in Section 6 of Birgé and Massart (2001), to which we refer for the details of the construction.

Here $\mu$ is the Lebesgue measure on [0, 1], and the required family of approximating linear spaces $\{S'_{m'}\}_{m' \in \mathcal{M}'}$ has been defined in Birgé and Massart (2000): we start with a suitable basis of $\mathbb{L}_2([0, 1])$, generated by orthogonal wavelets, splines or piecewise polynomials, having some regularity $\alpha_0$ (which can be arbitrarily large) and, for each $D \geqslant D_0 > 1$ ($D_0$ depending on the chosen basis), construct a family $\mathcal{S}_D$ of $D$-dimensional linear spaces with $|\mathcal{S}_D| \leqslant \exp(c_1 D)$. We then prove that, if $s$ belongs to some Besov space $B^\alpha_{p,\infty}$ with $p > 0$ and $1/p - 1/2 < \alpha < \alpha_0$ with Besov seminorm $|s|^\alpha_p$, one can find some $s_D$ belonging to some linear space in $\mathcal{S}_D$ such that

$$\|s - s_D\| \leqslant c_2 |s|^\alpha_p D^{-\alpha}, \tag{3.9}$$

where $c_2$ depends on the basis, $\alpha$ and $p$. To apply Theorem 2, we set

$\{S'_{m'}\}_{m' \in \mathcal{M}'} = \cup_{D \geqslant D_0} \mathcal{S}_D$ and if $S'_{m'} \in \mathcal{S}_D$ we define $\overline{D}_{m'} = (\log 5/\log 2)D$ and $\Delta_{m'} = (c_1 + 1)D$. This allows us to construct an estimator $\hat{s}$ according to Theorem 2 and we finally obtain, after an optimization with respect to $D$, the following proposition:

**Proposition 3.** *One can find an estimator $\hat{s}$ based on a suitable wavelet, spline or piecewise polynomial basis, but independent of $\alpha$ and $p$, such that, if $p > 0$, $1/p - 1/2 < \alpha < \alpha_0$ and $s \in B^{\alpha}_{p,\infty} \cap \mathbb{L}_\infty([0, 1])$, then*

$$\mathbb{E}_s\left[h^2(s, \hat{s})\right] \leqslant C\left[\left(\frac{|s|^{\alpha}_p}{\sigma}\right)^{2/(1+2\alpha)} \left(\frac{B}{n}\right)^{2\alpha/(1+2\alpha)} \vee \frac{B}{n}\right],$$

with $B = \log\left(\sigma^{-1}\|s\|_\infty\right) \vee 1$ *and $C$ depending only on the basis, $\alpha$ and $p$.*

It is known that $n^{-2\alpha/(1+2\alpha)}$ is a lower bound for the rate of convergence of estimators over balls in $B^{\alpha}_{p,\infty}$ when one uses the squared $\mathbb{L}_2$-loss. The relevant lower-bounds arguments rely on the construction of suitable systems of small perturbations around zero. These functions belong to a ball of fixed radius in $\mathbb{L}_\infty([0, 1])$ and, on this ball, the Hellinger distance and the $\mathbb{L}_2$-distance are equivalent by (2.3) and (2.6). It follows that the same construction can be used to obtain analogous lower bounds for the squared Hellinger risk so that the rate $n^{-2\alpha/(1+2\alpha)}$ is actually also optimal in our case.

One can derive multidimensional analogues of Proposition 3.1 or consider more general classes of functions. These are easy exercises using the family of models provided by Birgé and Massart (2000; 2001).

Note that the previous results remain valid, using the same families $\{S'_{m'}\}_{m' \in \mathcal{M}'}$ of approximating spaces, if $\mu$ has a bounded density with respect to the Lebesgue measure $\lambda$. In such a case, (3.9) still holds with $c_2$ depending also on $\|\mathrm{d}\mu/\mathrm{d}\lambda\|_\infty$, and we can proceed as before.

## 3.2. $\mathbb{L}_2$-risk

It is likely that, if we content ourself with bounding the Hellinger risk, some readers will feel frustrated and ask what happens if we use the more familiar squared $\mathbb{L}_2$-loss. Unfortunately, we are unable to obtain an analogue of Theorem 2 for the $\mathbb{L}_2$-risk and, in order to bound it, we have to work with estimators $\hat{s}$ that are bounded in $\mathbb{L}_\infty(\mu)$ by a fixed constant, as Wegkamp (2003) and Yang (2000; 2001; 2004) do.

**Theorem 3.** *Assume that we have at hand a finite or countable family $\{S'_{m'}\}_{m' \in \mathcal{M}'}$ of subsets of, $\mathbb{L}_2(\mu)$ with respective Euclidean metric dimensions bounded by $\overline{D}_{m'} \geqslant 3/2$, and let $\{\Delta_{m'}\}_{m' \in \mathcal{M}'}$ be a family of non-negative weights satisfying*

$$\sum_{m' \in \mathcal{M}'} \exp\left[-\Delta_{m'}\right] \leqslant \Sigma'. \tag{3.10}$$

*Given a positive integer $J$, one can construct an estimator $\hat{s}$ satisfying* $\sup_{x\in\mathcal{X}}|\hat{s}(x)| \leqslant \sigma \mathrm{e}^J$ *and such that, for any* $s \in \mathbb{L}_\infty(\mu)$,

$$\mathbb{E}_s\big[\|s-\hat{s}\|^2\big] \leqslant CB^2[1+10^{-7}\Sigma']\Big[d^2(s, s_J) + \inf_{m'\in\mathcal{M}'}\big\{d^2(s_J, S'_{m'}) + \big(n^{-1}J\sigma^2\big)(\overline{D}_{m'} \vee \Delta_{m'})\big\}\Big],$$

*with* $B = \big(\sigma^{-1}\|s\|_\infty\big) \vee \mathrm{e}^J$ *and* $s_J = (s \wedge \sigma\mathrm{e}^J) \vee (-\sigma\mathrm{e}^J)$.

***Proof.*** Since we proceed more or less along the lines of the proof of Theorem 2, we shall omit some details. We just proceed as in this proof, except that we restrict the definition of the $m$'s to $j \geqslant J$, we replace $\theta_j$ by $\theta$ with $\theta(t) = (t \wedge \sigma\mathrm{e}^J) \vee (-\sigma\mathrm{e}^J)$ and set $T'_m = \{t \in T_m | d(t, \theta(t)) \leqslant 4\sigma\eta_m\}$ and $S_m = \{\theta(t),\ t \in T'_m\} \subset \mathbb{L}_\infty(\mu)$. This implies that $\mathrm{e}^j$ is replaced by $\mathrm{e}^J$ in the subsequent formulae but the values of $\eta_m$ and $D_m$ do not change and we still conclude that (3.8) holds. Since both $|s|$ and $|\hat{s}|$ are bounded by $B\sigma$, it follows from (2.6) that

$$\mathbb{E}_s\big[\|s-\hat{s}\|^2\big] \leqslant C_1[1+10^{-7}\Sigma']\sigma^2B^2 \inf_{m\in\mathcal{M}}\big\{h^2(s, S_m) + \eta_m^2\big\}. \tag{3.11}$$

Now, given $m' \in \mathcal{M}'$, choose $j \geqslant J$ minimal such that $\sqrt{j}\eta_{m'} \geqslant d(s_J, S'_{m'})/\sigma$ and set $m = (m', j)$. The definition of $j$ (distinguishing between the cases $j = J$ and $j > J$) implies that

$$\eta_m^2 \leqslant \big[2\sigma^{-2}d^2(s_J, S'_{m'})\big] \vee \big[J\eta_{m'}^2\big] \tag{3.12}$$

and the arguments used to conclude the proof of Theorem 2, with $s$ replaced by $s_J$, show that $h^2(s_J, S_m) \leqslant \eta_m^2/2$, which, together with (3.11) and (3.12), implies that

$$\mathbb{E}_s\big[\|s-\hat{s}\|^2\big] \leqslant C_2[1+10^{-7}\Sigma']\sigma^2B^2\left[h^2(s, s_J) + \inf_{m'\in\mathcal{M}'}\Big\{\big(\sigma^{-2}d^2(s_J, S'_{m'})\big) \vee \big(J\eta_{m'}^2\big)\Big\}\right].$$

The conclusion follows from our choice of $\eta_{m'}$ and (2.3), which implies that $h^2(s, s_J) \leqslant d^2(s, s_J)/(8\sigma^2)$. □

If $J \geqslant \log(\|s\|_\infty/\sigma)$, then $B = \mathrm{e}^J$, $s_J = s$ and we obtain

$$\mathbb{E}_s\big[\|s-\hat{s}\|^2\big] \leqslant CB^2[1+10^{-7}\Sigma'] \inf_{m'\in\mathcal{M}'}\big\{d^2(s, S'_{m'}) + \big(n^{-1}\sigma^2\log B\big)(\overline{D}_{m'} \vee \Delta_{m'})\big\}. \tag{3.13}$$

The optimal choice for $J$ is the smallest one, but such a choice actually requires a prior knowledge of $\|s\|_\infty$. If it is unknown, the choice of $J$ may be inadequate – either too large, which leads to an unnecessarily large value of $B$ in (3.13), or too small, which may imply a large value of $d(s, s_J)$.

# References

Baraud, Y. (2000) Model selection for regression on a fixed design. *Probab. Theory Related Fields*, **117**, 467–493.

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

Barron, A.R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related 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. (2003) Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Preprint no. 862, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI. http://www.proba.jussieu.fr/mathdoc/preprints/index.html.

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

Birgé, L. and Massart, P. (2000) An adaptive compression algorithm in Besov spaces. *Constr. Approx.*, **16**, 1–36.

Birgé, L. and Massart, P. (2001) Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203–268.

Brown, L.D., Cai, T.T., Low, M.G. and Zhang, C.-H. (2002) Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.*, **30**, 688–707.

DeVore, R.A. and Lorentz, G.G. (1993) *Constructive Approximation.* Berlin: Springer-Verlag.

Donoho, D.L. and Liu, R.C. (1991) Geometrizing rates of convergence II. *Ann. Statist.*, **19**, 633–667.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Juditsky, A. and Nemirovski, A.S. (2000) Functional aggregation for nonparametric regression. *Ann. Statist.*, **28**, 681–712.

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

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. In M. Puri (ed.), *Stochastic Processes and Related Topics, Vol. 1*, pp. 13–54. New York: Academic Press.

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

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

Yang, Y. (2000) Combining different procedures for adaptive regression. *J. Multivariate Anal.*, **74**, 135–161.

Yang, Y. (2001) Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, **96**, 574–588.

Yang, Y. (2004) Aggregating regression procedures for a better performance. *Bernoulli*, **10**, 25–47.

Yang, Y. and Barron, A.R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory*, **44**, 95–116.

Received December 2002 and revised April 2004