Least squares estimation with complexity penalties

By
Sara van de Geer
University of Leiden

Abstract. We examine the regression model $Y_i = g_0(z_i) + W_i$, $i = 1, \ldots, n$, and the penalized least squares estimator $\hat{g}_n = \arg\min_{g \in \mathcal{G}} \{\|Y - g\|^2 + \text{pen}^2(g)\}$, where $\text{pen}(g)$ is a penalty on the complexity of the function g. We show that a rate of convergence for $\hat{g}_n$ follows from the entropy of the sets $\mathcal{G}_*(\delta) = \{g \in \mathcal{G} : \|g - g_*\|^2 + \text{pen}^2(g) \leq \delta^2\}$, $\delta > 0$, where $g_* = \arg\min_{g \in \mathcal{G}} \{\|g - g_0\|^2 + \text{pen}^2(g)\}$ (say). As examples, we consider Sobolev and dimension penalties.

1. Introduction. We consider, for $n \geq 2$, the regression model

$$Y_i = g_0(z_i) + W_i, \; i = 1, \ldots, n,$$

where $Y_1, \ldots, Y_n$ are real-valued observations, $z_1, \ldots, z_n$ are explanatory variables, with values in some space $\mathcal{Z}$, the measurement errors $W_1, \ldots, W_n$ are independent and centered, and $g_0 : \mathcal{Z} \to \mathbf{R}$ is an unknown regression function.

If it is a priori known that g_0 is in some class of regression functions $\mathcal{G}$, we call $\mathcal{G}$ a model for g_0. Examples, in the case $\mathcal{Z} = \mathbf{R}$, are the class of all linear functions, or all polynomials of degree four, or all monotone functions, or all functions with integrable second derivative, etc. In van de Geer (1990), one can find a result which relates the speed of estimation to the entropy of the model $\mathcal{G}$ (see Theorem 1.1 below). Our goal in this paper is to generalize this theorem to the situation where one has virtually no a priori model for g_0. The idea is then to let the data speak for themselves, by allowing for a range of models and penalizing the complexity of a model. This is in the spirit of for instance Birgé and Massart (1997), Barron, Birgé and Massart (1999) or Lugosi and Nobel (1999). More generally, adaptation to unknown complexities can be done using various methods. Important references in this field are Efroimovich and Pinsker (1985), and Lepskii (1991, 1992).

There are numerous ways to describe complexity. It could be the dimension of the model, or the inverse of the number of derivatives one allows, or more generally the inverse of the smoothness parameter describing Besov spaces, and so on. We will show that various complexity penalties can be studied using one single approach. The result is a rate of convergence for the estimator, which can depend on the unknown complexity.

Our main aim is to provide a simple and general result, and insight into common features of some estimation methods. We will see that in fact, a slight re-formulation of existing theory for penalized least squares estimation will allow to include adaptive estimation methods. However, in special situations, our results can certainly be improved, yielding e.g. exact constants (see also the discussion in Section 4).

In this section, we introduce some notation, and we recall the result of van de Geer (1990) for least squares estimation without penalty in Theorem 1.1. Section 2 contains the main result. Theorem 2.1 generalizes Theorem 1.1, in such a way that ((almost) adaptive) rates of convergence for various kinds of complexity penalized estimators can be derived by entropy calculations. In Section 3, we will present some examples, such as the choice of a smoothness parameter or the selection of explanatory variables. We also consider soft thresholding penalties. In Section 4, we give a discussion of the results and some further references to related work.

Throughout the paper, g_0 denotes the true regression function, and $\mathcal{G}$ is some given class of regression functions. We will not always assume that $g_0 \in \mathcal{G}$. Moreover, in some cases $\mathcal{G}$ is very large, for example the class of *all* functions.

Let $Q_n = \sum_{i=1}^n \delta_{z_i}/n$ be the empirical measure of the explanatory variables $z_1, \ldots, z_n$. For $g : \mathcal{Z} \to \mathbf{R}$, we write the squared $L_2(Q_n)$ (semi-)norm as

$$\|g\|^2 = \frac{1}{n} \sum_{i=1}^n g(z_i)^2.$$

We let

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : \ \|g - g_0\| \le \delta\}, \tag{1.1}$$

denote a ball with radius δ around g_0, intersected with $\mathcal{G}$. We remark here, that we will extend the definition of $\mathcal{G}(\delta)$ later on, to include a penalty.

Write $H(u, \mathcal{G}(\delta))$ for the u-entropy of $\mathcal{G}(\delta)$. The entropy of a (pseudo-)metric space is defined as follows.

Definition. *Let T be a (subset of a) metric space. The u-covering number $N(u,T)$ is defined as the number of balls with radius u necessary to cover T. The u-entropy is then $H(u,T) = \log N(u,T) \vee 0$.*

Now, one may identify $\mathcal{G}(\delta)$ with a bounded subset of n-dimensional Euclidean space. Therefore, its entropy will always be finite. Without loss of generality, we may assume that the square root of its entropy is integrable. We denote the result by

$$J(\delta) = \int_0^\delta H^{1/2}(u, \mathcal{G}(\delta)) du. \tag{1.2}$$

We will assume throughout that the errors are uniformly sub-Gaussian:

$$\max_{1 \le i \le n} \mathbf{E} e^{|W_i|^2/K^2} \le 2, \tag{1.3}$$

where $K < \infty$ is some constant.

Now, consider the least squares estimator $\hat{g}_n$ (which we tacitly assume to exist):

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^n (Y_i - g(z_i))^2. \tag{1.4}$$

Later on we will introduce the penalized least squares estimator, for which we will use the same notation $\hat{g}_n$.

The following theorem relates the speed of estimation of the least squares estimator $\hat{g}_n$ to the entropy of $\mathcal{G}$. The theorem can be improved in some cases by taking in the definition of $J(\delta)$ the integral over $(\delta^2/c', \delta]$ (c' some suitable constant) instead of $(0, \delta]$. However, to avoid digressions, we will not present this improved version. (See also Subsection 3.3.)

Theorem 1.1. *Let $\mathcal{G}(\delta)$ be defined in (1.1). Consider the least squares estimator $\hat{g}_n$ defined in (1.4). Suppose that $g_0 \in \mathcal{G}$. Let $\Psi(\delta) \ge J(\delta) \vee \delta$ be an upper bound, such that $\Psi(\delta)/\delta^2$ decreases as δ increases. Then there exists a constant c depending only on K, such that for $\sqrt{n}\delta_n^2 \ge c\Psi(\delta_n)$ and for all $\delta \ge \delta_n$, we have*

$$\mathbf{P}(\|\hat{g}_n - g_0\|^2 \ge \delta^2) \le c \exp[-\frac{n\delta^2}{c^2}]. \tag{1.4}$$

This theorem can be found in van de Geer (1990, 2000). The theorem says that if $\mathcal{G}$ is large, the least squares estimator might converge rather slowly. Indeed, from van de Geer and Wegkamp (1996), we know that if $H(u, \mathcal{G}(\delta))/n$ does not tend to zero for all $0 < u \le \delta$, then the least squares estimator will not even be consistent for most error distributions. In other words, when little is known a priori about g_0, the least squares estimator will not behave well. A way out is to add a penalty to the least squares loss function.

2. Penalized least squares. We have now chosen for each $g \in \mathcal{G}$ a penalty on g, which we denote by $\text{pen}(g)$ (possibly $\text{pen}(g) = 0$ for all g, i.e. no penalty). The penalized least squares estimator is

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} \{\frac{1}{n} \sum_{i=1}^n (Y_i - g(z_i))^2 + \text{pen}^2(g)\}. \tag{2.1}$$

Write, for two functions $g \in \mathcal{G}$ and $\tilde{g}$,

$$\tau^2(g|\tilde{g}) = \|g - \tilde{g}\|^2 + \text{pen}^2(g). \tag{2.2}$$

Let g_* be a fixed function in $\mathcal{G}$. Although in Theorem 2.1, any $g_* \in \mathcal{G}$ will do, we actually have in mind choosing

$$g_* = \arg\min_{g \in \mathcal{G}} \tau^2(g|g_0), \tag{2.3}$$

because this choice gives the best bounds. If the minimum does not exist, one may take $g_* \in \mathcal{G}$ such that $\tau(g_*|g_0)$ is arbitrary close to the infimum of $\tau^2(g|g_0)$ over $\mathcal{G}$. Note that $\tau^2(g|g_0) = \|g - g_0\|^2 + \text{pen}^2(g)$. As we shall see in the examples, the above choice of g_* has an interpretation as trade-off between bias and variance, with $\|g - g_0\|^2$ representing the squared bias, and $\text{pen}^2(g)$ the variance.

Define now

$$\mathcal{G}_*(\delta) = \{g \in \mathcal{G} : \ \tau^2(g|g_*) \leq \delta^2\}, \ \delta > 0. \tag{2.4}$$

Let

$$J_*(\delta) = \int_0^\delta H^{1/2}(u, \mathcal{G}_*(\delta)) du. \tag{2.5}$$

Theorem 2.1. *Let $\Psi_*(\delta) \geq J_*(\delta) \vee \delta$ be some upper bound. Suppose that $\Psi_*(\delta)/\delta^2$ decreases as δ increases. There exists a constant c depending only on K, such that for*

$$\sqrt{n}\delta_n^2 \geq c\Psi_*(\delta_n), \tag{2.6}$$

we have for all $\delta \geq \delta_n$,

$$\begin{aligned} &\mathbf{P}\left(\tau^2(\hat{g}_n|g_0) \geq 2(\tau^2(g_*|g_0) + \delta^2)\right) \\ &\leq c \exp[-\frac{n\delta^2}{c^2}]. \end{aligned} \tag{2.7}$$

Proof. Since $g_* \in \mathcal{G}$, we have that

$$\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{g}_n(z_i))^2 + \text{pen}^2(\hat{g}_n) \leq \frac{1}{n}\sum_{i=1}^n (Y_i - g_*(z_i))^2 + \text{pen}^2(g_*).$$

Rewrite this to

$$\frac{2}{n}\sum_{i=1}^n W_i(\hat{g}_n(z_i) - g_*(z_i)) \geq \tau^2(\hat{g}_n|g_0) - \tau^2(g_*|g_0). \tag{2.8}$$

For $a, b \in \mathbf{R}$ one has $(a-b)^2 \geq a^2/2 - b^2$. So

$$\tau^2(\hat{g}_n|g_*) \geq \frac{1}{2}\tau^2(\hat{g}_n|g_0) - \tau^2(g_*|g_0),$$

and also

$$\tau^2(\hat{g}_n|g_0) \geq \frac{1}{2}\tau^2(\hat{g}_n|g_*) - \tau^2(g_*|g_0).$$

Now, if $\tau^2(\hat{g}_n|g_0) \geq 2(\tau^2(g_*|g_0) + \delta^2)$, we find that

$$\tau^2(\hat{g}_n|g_*) \geq \delta^2,$$

and moreover

$$\begin{aligned}
\tau^2(\hat{g}_n|g_0) &- \tau^2(g_*|g_0) \\
&= \frac{1}{3}\tau^2(\hat{g}_n|g_0) + \frac{2}{3}\tau^2(\hat{g}_n|g_0) - \tau^2(g_*|g_0) \\
&\geq \frac{1}{3}\tau^2(\hat{g}_n|g_0) + \frac{1}{3}\tau^2(g_*|g_0) \\
&\geq \frac{1}{6}\tau^2(\hat{g}_n|g_*).
\end{aligned} \tag{2.9}$$

We thus obtain that

$$\begin{aligned}
&\mathbf{P}\left(\tau^2(\hat{g}_n|g_0) \geq 2(\tau^2(g_*|g_0) + \delta^2)\right) \\
&\leq \mathbf{P}\left(\frac{2}{n}\sum_{i=1}^{n} W_i(\hat{g}_n(z_i) - g_*(z_i)) \geq \frac{1}{6}\tau^2(\hat{g}_n|g_*) \wedge \tau^2(\hat{g}_n|g_*) \geq \delta^2\right) \\
&\leq \sum_{l=1}^{\infty} \mathbf{P}\left(\sup_{g\in\mathcal{G}_*(2^l\delta)} |\frac{1}{n}\sum_{i=1}^{n} W_i(g(z_i) - g_*(z_i))| \geq \frac{1}{24}2^{2l}\delta^2\right) \\
&= \sum_{l=1}^{\infty} \mathbf{P}_l.
\end{aligned}$$

From Corollary 8.4 in van de Geer (2000), we know that for $\sqrt{n}a \geq C\Psi_*(R)$, where C depends on K only,

$$\mathbf{P}\left(\sup_{g\in\mathcal{G}_*(R)} |\frac{1}{n}\sum_{i=1}^{n} W_i(g(z_i) - g_*(z_i))| \geq a\right) \leq \exp[-\frac{na^2}{4C^2R^2}].$$

Now, we take $\delta \geq \delta_n$, and we have chosen $\sqrt{n}\delta_n^2 \geq c\Psi_*(\delta_n)$. If we choose the constant c large enough, this gives $\sqrt{n}\frac{1}{24}2^{2l}\delta^2 \geq C\Psi_*(2^l\delta)$ for all $l \in \{1, 2, \ldots\}$. So we may apply the above mentioned corollary to each $\mathbf{P}_l$ to obtain that

$$\begin{aligned}
\sum_{l=1}^{\infty} \mathbf{P}_l &\leq C\sum_{l=1}^{\infty} \exp[-\frac{n2^{2l}\delta^2}{2304C^2}] \\
&\leq c\exp[-\frac{n\delta^2}{c^2}],
\end{aligned}$$

for an appropriately chosen (large enough) constant c.

□

The following lemma is a simple consequence of Theorem 2.1.

Lemma 2.2. *Under the conditions of Theorem 2.1, we arrive at the inequality*

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2(\tau^2(g_*|g_0) + \delta_n^2) + \frac{c_0}{n}, \tag{2.10}$$

where c_0 is a constant only depending on K.

Proof. Write $\hat{\tau} = \tau(\hat{g}_n|g_0)$ and $\tau = \tau(g_*|g_0)$ for short. We have

$$\mathbf{E}\hat{\tau}^2 = \int_0^{\infty} \mathbf{P}(\hat{\tau}^2 \geq t)dt$$

$$\leq 2(\tau^2 + \delta_n^2) + \int_{t > 2(\tau^2+\delta_n^2)} \mathbf{P}(\hat{\tau}^2 \geq t) dt.$$

But clearly,

$$\int_{t>2(\tau^2+\delta_n^2)} \mathbf{P}(\hat{\tau}^2 \geq t)dt = \frac{1}{2}\int_{\delta > \delta_n} \mathbf{P}(\hat{\tau}^2 \geq 2(\tau^2+\delta^2))d\delta^2$$

$$\leq \frac{1}{2}\int_{\delta>\delta_n} c\exp[-\frac{n\delta^2}{c^2}]d\delta^2 = \frac{c^3}{2n}\exp[-\frac{n\delta_n^2}{c^2}] \leq \frac{c^3}{2ne},$$

where in the last inequality, we used that $n\delta_n^2 \geq c^2$. □

It can be easily verified that Theorem 2.1 reduces to Theorem 1.1, when both $\text{pen}(g) = 0$ for all $g \in \mathcal{G}$ and $g_0 \in \mathcal{G}$. It generalizes Theorem 1.1 to cover a broad class of penalized methods. In most cases however, Theorem 2.1 and Lemma 2.2 should be understood as a rough, but simple and general method to access rates of convergence. One cannot deduce a complete estimation recipe from it, mainly because (2.6) usually means that the choice of the penalty depends on c and hence on the distribution of the errors (via the constant K). Moreover, the entropy approach we used is rather slovenly on constants (and also in the entropy calculations itself it is not easy to obtain good constants). Therefore, we did not try to improve the factor 2 which appears in the theorem and the lemma following it. In special cases, ad hoc methods may lead to major improvements. See also the discussions in Section 3.3 and Section 4.

3. Examples.

3.1. Estimating a function in a Sobolev space. Let $\mathcal{Z} = [0,1]$, and let $\mathcal{G}$ be the class of functions with derivatives of all orders. Define the squared Sobolev semi-norm

$$I_s^2(g) = \int_0^1 |g^{(s)}(z)|^2 dz, \ g \in \mathcal{G}. \tag{3.1}$$

We assume that $g_0 \in \mathcal{G}$ with $I_s(g_0) < \infty$ for all $s \in \{1, 2, \ldots\}$. This is in fact no restriction because we are only concerned about estimating g_0 in the design points $z_1, \ldots, z_n$. We will apply Theorem 2.1. To avoid digressions, we take $g_* = g_0$ (which is possible by the assumption $g_0 \in \mathcal{G}$).

We will consider several penalties. In our entropy calculations, we need two approximation lemmas. The first lemma is a slight extension of results of Birman and Solomjak (1967), in that we express the dependence on s (albeit perhaps not in an optimal way). (See also Kolmogorov and Tihomirov (1959).) In the result we denote the uniform norm of a function $f : \mathcal{Z} \to \mathbf{R}$, by

$$|f|_\infty = \sup_{z \in \mathcal{Z}} |f(z)|. \tag{3.2}$$

Lemma 3.1. *Let $\mathcal{F}$ be the following class of s times differentiable functions on the unit interval:*

$$\mathcal{F} = \{f : I_s(f) \leq 1, |f|_\infty \leq 1\}. \tag{3.3}$$

Let $H_\infty(u, \mathcal{F})$ be the entropy of $\mathcal{F}$ for the metric induced by the uniform norm $|\cdot|_\infty$. Then for some constant A, not depending on s, we have

$$H_\infty(u, \mathcal{F}) \leq sA^2(\frac{1}{u})^{\frac{1}{s}}, \ u > 0. \tag{3.4}$$

We also need the entropy of a class of polynomials of finite degree. The entropy of a ball in finite-dimensional space is roughly speaking its dimension (see e.g. van de Geer (2000)):

Lemma 3.2. *Let $B_d(\delta)$ be a ball in d-dimensional Euclidean space with radius δ. Then*

$$H(u, B_d(\delta)) \leq d\log(\frac{5\delta}{u}), \ 0 < u \leq \delta. \tag{3.5}$$

Corollary 3.3. *We have for all δ, M and $s \in \{1, 2, \ldots\}$,*

$$H(u, \{g \in \mathcal{G} : \|g - g_0\| \le \delta,\ I_s(g) \le M\}) \le s \log(\frac{5(\delta + M)}{u}) + sA^2(\frac{M}{u})^{\frac{1}{s}},\ 0 < u \le \delta. \tag{3.6}$$

Before turning to various penalties, let us briefly present an application of Theorem 1.1. Suppose that in the regression model, $g_0 \in \mathcal{G}$, where $\mathcal{G} = \{g :\ I_s(g) \le M\}$, where s and M are given (with $M \ge 1$ say). Then from Theorem 1.1, the least squares estimator without penalty converges with rate $M^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}$. It is shown in van de Geer (1995) that this rate cannot improved. Comparing this result with e.g. inequality (3.11) below, we see that penalization allows one to adapt when M is not a priori known.

3.1.1. Penalty on $I_s(g)$ with s fixed. Consider the penalty

$$\text{pen}^2(g) = \lambda_s^2 I_s^2(g),\ g \in \mathcal{G}, \tag{3.7}$$

where $0 < \lambda_s \le 1$ is a smoothing parameter. To obtain a rate of convergence for the penalized estimator $\hat{g}_n$, we need to calculate the entropy of the class $\mathcal{G}_*(\delta)$. The result is given in the next lemma.

Lemma 3.4. *We have*

$$\mathbf{E}\tau^2(\hat{g}_n | g_0) \le 2\lambda_s^2 I_s^2(g_0) + \frac{C_s^2}{n\lambda_s^{1/s}} + \frac{c_0}{n}, \tag{3.8}$$

with C_s a constant depending on K and s and c_0 a constant depending on K.

Proof. Since

$$\mathcal{G}_*(\delta) \subset \{g \in \mathcal{G} :\ \|g - g_0\| \le \delta, I_s(g) \le \delta/\lambda_s\},$$

it follows that

$$H(u, \mathcal{G}_*(\delta)) \le s \log(\frac{10\delta}{u\lambda_s}) + sA^2(\frac{\delta}{u\lambda_s})^{\frac{1}{s}} \le A_s^2(\frac{\delta}{u\lambda_s})^{\frac{1}{s}},$$

for all $0 < u \le \delta$, and for some constant A_s depending on s. Therefore

$$J_*(\delta) \le A_s \frac{\delta}{\lambda_s^{1/2s}} = \Psi_*(\delta),\ \delta > 0.$$

With this choice of Ψ_*, we apply Lemma 2.2. □

Note that if $I_s(g_0)$ remains bounded in n, the choice $\lambda_s \asymp n^{-\frac{s}{2s+1}}$ leads to the usual rate

$$\mathbf{E}\|\hat{g}_n - g_0\|^2 = O(n^{-\frac{2s}{2s+1}}).$$

3.1.2. Penalty on $I_s(g)$ with s fixed, and on the smoothing parameter. Consider the penalty

$$\text{pen}^2(g) = \inf_{0<\lambda<\infty} \{\lambda^2 I_s^2(g) + \frac{\lambda_0^2}{n\lambda^{1/s}}\}, \tag{3.9}$$

where λ_0 is a (large enough) constant, to be specified later.

It is clear that the infimum in (3.9) is attained in

$$\lambda_g^2 = \left(\frac{\lambda_0^2}{2nsI_s^2(g)}\right)^{\frac{2s}{2s+1}},$$

so that

$$\text{pen}^2(g) = \left(I_s^2(g)(2s+1)\right)^{\frac{1}{2s+1}} \left(\frac{\lambda_0^2(2s+1)}{2ns}\right)^{\frac{2s}{2s+1}}. \tag{3.10}$$

Of course, one may first calculate the penalized estimator for all λ fixed. Let us denote the result by

$$\hat{g}_{n,\lambda} = \arg\min_{g \in \mathcal{G}} \{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(z_i))^2 + \lambda^2 I_s^2(g) \}.$$

The smoothing parameter λ is then chosen data dependent, by minimizing over λ the risk

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_{n,\lambda}(z_i))^2 + \lambda^2 I_s^2(\hat{g}_{n,\lambda}) + \frac{\lambda_0^2}{n\lambda^{1/s}}.$$

Lemma 3.5. *Take $\lambda_0^2 \geq c_1 s$, with c_1 a large enough constant depending on K. Then*

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2 \left(I_s^2(g_0)(2s+1) \right)^{\frac{1}{2s+1}} \left(\frac{\lambda_0^2(2s+1)}{2ns} \right)^{\frac{2s}{2s+1}} + \frac{c_0 s^2 \log n}{n}. \tag{3.11}$$

Proof. We will calculate the entropy of $\mathcal{G}_*(\delta)$. If $g \in \mathcal{G}_*(\delta)$, it holds that

$$I_s(g) \leq M_s,$$

where

$$M_s = \frac{(2ns)^s \delta^{2s+1}}{\lambda_0^{2s}(2s+1)^{\frac{2s+1}{2}}} \leq \frac{n^s \delta^{2s+1}}{\lambda_0^{2s}}.$$

Therefore, by Corollary 3.3, when $M_s \geq \delta$,

$$H(u, \mathcal{G}_*(\delta)) \leq s \log(\frac{10 M_s}{u}) + sA^2(\frac{M_s}{u})^{\frac{1}{s}}$$

$$\leq s \log(\frac{10 n^s \delta^{2s+1}}{u \lambda_0^{2s}}) + \frac{sA^2 n \delta^{2+\frac{1}{s}}}{u^{\frac{1}{s}} \lambda_0^2}.$$

If $M_s \leq \delta$ we obtain

$$H(u, \mathcal{G}_*(\delta)) \leq s \log(\frac{10\delta}{u}) + \frac{sA^2 n \delta^{2+\frac{1}{s}}}{u^{\frac{1}{s}} \lambda_0^2}.$$

It follows that for $M_s \geq \delta$,

$$\int_0^\delta H^{1/2}(u, \mathcal{G}_*(\delta)) du \leq 2A\sqrt{s} \frac{\sqrt{n}\delta^2}{\lambda_0} + A_0 \sqrt{s}\delta + s\delta \sqrt{\log(\frac{n\delta^2}{\lambda_0^2})},$$

with

$$A_0 = \int_0^1 \log^{1/2}(\frac{10}{u}) du.$$

For $M_s \leq \delta$, we find similarly

$$\int_0^\delta H^{1/2}(u, \mathcal{G}_*(\delta)) du \leq 2A\sqrt{s} \frac{\sqrt{n}\delta^2}{\lambda_0} + A_0 \sqrt{s}\delta.$$

With the choice

$$\Psi_*(\delta) = A_* \left(\sqrt{s} \frac{\sqrt{n}\delta^2}{\lambda_0} + \sqrt{s}\delta + s\delta \sqrt{\log(\frac{n\delta^2}{\lambda_0^2})} \right),$$

where $A_* = 2(A \vee A_0)$, we find

$$\sqrt{n}\delta_n^2 \geq c\Psi_*(\delta_n),$$

for

$$\lambda_0^2 \geq c_1 s, \ \delta_n \geq c_1 s \sqrt{\frac{\log n}{n}}.$$

Here c_1 is a constant depending on K. Application of Lemma 2.2 gives

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2\text{pen}^2(g_0) + \frac{c_0 s^2 \log n}{n}$$

$$= 2\left(I_s^2(g_0)(2s+1)\right)^{\frac{1}{2s+1}} \left(\frac{\lambda_0^2(2s+1)}{2ns}\right)^{\frac{2s}{2s+1}} + \frac{c_0 s^2 \log n}{n}.$$

□

Taking $\lambda_0^2 = c_1 s$ gives

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2\left(I_s^2(g_0)(2s+1)\right)^{\frac{1}{2s+1}} \left(\frac{c_1(2s+1)}{2n}\right)^{\frac{2s}{2s+1}} + \frac{c_0 s^2 \log n}{n}.$$

We may compare this with the exact asymptotic minimax constant obtained by Nussbaum (1985). He shows that

$$\liminf_{n\to\infty} \inf_{\tilde{g}_n} \sup_{g_0:\ I_s(g_0)\leq M} n^{\frac{2s}{2s+1}} \mathbf{E}\|\tilde{g}_n - g_0\|_\mu^2 = \left(M^2(2s+1)\right)^{\frac{1}{2s+1}} \left(\frac{s}{\pi(s+1)}\right)^{\frac{2s}{2s+1}}, \tag{3.12}$$

where μ is Lebesgue measure on $[0,1]$, $\|\cdot\|_\mu$ is the $L_2([0,1],\mu)$-norm, $M > 0$ is a now a given constant not depending on n, and where the infimum is over all estimators based on n observations of the regression model with equidistant design. The right hand side of (3.12) is Pinsker's constant. The constants in our bound (3.11) are not quite as good, especially when s is large. This may also be due to our crude bound (in terms of s) for the entropy. (See Section 4 for further discussion.)

3.1.3. Penalty on $I_s(g)$ and on s. We propose to penalize s in the following manner. Let $s_{\max} \in \{1,2,\ldots\}$ be given (we have in mind the situation where $s_{\max}$ grows with n). We take

$$\text{pen}^2(g) = \min_{1\leq s\leq s_{\max}} \lambda_s^2[I_s^2(g) + \lambda_0^2]. \tag{3.13}$$

Here λ_0 is a large enough constant, which we will specify later. We assume throughout that $\lambda_s = n^{-\frac{s}{2s+1}}$ for all s.

Lemma 3.6. *Take $\lambda_0^2 \geq c_1 s_{\max}^3$, with c_1 a large enough constant depending on K. Then*

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2\text{pen}^2(g_0) + \frac{c_0 s_{\max} \log n}{n}. \tag{3.14}$$

Proof. We will find an upper bound for $J_*(\delta)$ in this situation. Fix $\delta > 0$. If $\lambda_s > \delta/\lambda_0$ for all $s \leq s_{\max}$ we have that $\mathcal{G}(\delta) = \emptyset$. So let us assume that $\lambda_s \leq \delta/\lambda_0$ for some $s \leq s_{\max}$, and let $s(\delta)$ be the smallest value in $\{1,\ldots,s_{\max}\}$ for which this is true. Using Corollary 3.3, is now easy to see that

$$H(u,\mathcal{G}_*(\delta)) \leq \sum_{s=s(\delta)}^{s_{\max}} sA^2(\frac{\delta}{u\lambda_s})^{\frac{1}{s}} + s_{\max}\log(\frac{10\delta}{u\lambda_{s_{\max}}}).$$

Hence,

$$J_*(\delta) \leq \sum_{s=s(\delta)}^{s_{\max}} \int_0^\delta 2A\sqrt{s}(\frac{\delta}{u\lambda_s})^{\frac{1}{2s}} du + \sqrt{s_{\max}} \int_0^\delta \log^{1/2}(\frac{10\delta}{u\lambda_{s_{\max}}}) du$$

$$\leq 2A\sqrt{s_{\max}}\delta \sum_{s=s(\delta)}^{s_{\max}} (\frac{1}{\lambda_s})^{\frac{1}{2s}} + A_0\sqrt{s_{\max}}\delta\sqrt{\log n}$$

$$\leq 2As_{\max}^{3/2}\delta(\frac{1}{\lambda_{s(\delta)}})^{\frac{1}{2s(\delta)}} + A_0\sqrt{s_{\max}}\delta\sqrt{\log n}.$$

Here,

$$A_0 = \int_0^1 \log^{1/2}(\frac{10}{u})du.$$

Now, one easily sees that by our choice $\lambda_s = n^{-\frac{s}{2s+1}}$,

$$\lambda_s = \frac{1}{\sqrt{n}\lambda_s^{\frac{1}{2s}}}.$$

So

$$J_*(\delta) \leq 2As_{\max}^{3/2}\delta\sqrt{n}\lambda_{s(\delta)} + A_0\sqrt{s_{\max}}\delta\sqrt{\log n}$$

$$\leq 2As_{\max}^{3/2}\frac{\sqrt{n}\delta^2}{\lambda_0} + A_0\sqrt{s_{\max}}\delta\sqrt{\log n}.$$

We thus find that if we take

$$\lambda_0^2 \geq 4c^2A^2s_{\max}^3$$

we may choose $\delta_n = 2cA_0\sqrt{\frac{s_{\max}\log n}{n}}$, to find from Lemma 2.2,

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2\text{pen}^2(g_0) + \frac{c_0 s_{\max}\log n}{n}.$$

□

Recall that in this case,

$$\text{pen}^2(g_0) = \min_{s\leq s_{\max}} \{\lambda_s^2 I_s^2(g_0) + \lambda_0^2\}.$$

Let s_0 be the largest value of $s \leq s_{\max}$ where the above minimum is attained, and suppose that $I_{s_0}(g_0)$ remains bounded in n. Taking $\lambda_0^2 = c_1^2 s_{\max}^3$ one finds

$$\mathbf{E}\|\hat{g}_n - g_0\|^2 = O(s_{\max}^3 n^{-\frac{2s_0}{2s_0+1}}).$$

In this sense, the estimator with penalty on s is almost adaptive in s.

3.1.4. Penalty on $I_s(g)$ and on s, and on the smoothing parameter. Let again $1 \leq s_{\max} < \infty$ be an upper bound for the number of derivatives, and consider the penalty

$$\text{pen}^2(g) = \inf_{0<\lambda<\infty} \min_{1\leq s\leq s_{\max}} \{\lambda^2 I_s^2(g) + \frac{\lambda_0^2}{n\lambda^{1/s}}\}. \tag{3.15}$$

Lemma 3.7. *Take $\lambda_0^2 \geq c_1 s_{\max}^3$, where c_1 is a large enough constant depending on K. Then*

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2\text{pen}^2(g_0) + \frac{c_0 s_{\max}^2 \log n}{n}. \tag{3.16}$$

Proof. Clearly,

$$\mathcal{G}_*(\delta) = \cup_{s=1}^{s_{\max}} \mathcal{G}_*^{(s)}(\delta),$$

where

$$\mathcal{G}_*^{(s)}(\delta) = \{g \in \mathcal{G} : \ \|g - g_0\|^2 + \inf_{0<\lambda<\infty} \{\lambda^2 I_s^2(g) + \lambda_0^2/(n\lambda^{1/s})\} \leq \delta^2\}.$$

So, arguing as in the proof of Lemma 3.5,

$$H(u,\mathcal{G}_*(\delta)) \le \sum_{s=1}^{s_{\max}} sA^2(\frac{M_s}{u})^{\frac{1}{s}} + s_{\max}\log(\frac{5(\delta+M_{s_{\max}})}{u}),$$

where

$$M_s = \frac{(2ns)^s\delta^{2s+1}}{\lambda_0^{2s}(2s+1)^{\frac{2s+1}{2}}},\ s=1,\ldots,s_{\max}.$$

Thus

$$J_*(\delta) \le \sum_{s=1}^{s_{\max}} 2A\sqrt{s}\frac{\sqrt{n}\delta^2}{\lambda_0} + A_0\sqrt{s_{\max}}\delta + s_{\max}\delta\sqrt{\log(\frac{n\delta^2}{\lambda_0^2})}$$

$$\le A_*\left(s_{\max}^{3/2}\frac{\sqrt{n}\delta^2}{\lambda_0} + \sqrt{s_{\max}}\delta + s_{\max}\delta\sqrt{\log(\frac{n\delta^2}{\lambda_0^2})}\right) = \Psi_*(\delta).$$

The result now follows again from Lemma 2.2. □

3.2. Penalty on the dimension. Let $\mathcal{G}$ be the class of all functions on $z_1,\ldots,z_n$. To settle the notation, we assume the z_i are distinct. Let $\psi_1,\ldots,\psi_n$ be an orthonormal basis of $L_2(Q_n)$. We may write any function $g:\mathcal{Z}\to\mathbf{R}$ as

$$g(z_i) = \alpha_1\psi_1(z_i) + \ldots + \alpha_n\psi_n(z_i),\ i=1,\ldots,n,$$

where $\alpha_1,\ldots,\alpha_n$ are the coefficients of g.

Consider a fixed dimension d, and suppose we approximate g_0 by a function $g_{0,d}$ with d specified non-zero coefficients. The variance of the least squares estimator of $g_{0,d}$ is $\sigma^2 d/n$, σ^2 being the (average) variance of the measurement errors. The squared bias of this estimator is $\|g_{d,0}-g_0\|^2$. Therefore, the mean square error is

$$\|g_{d,0}-g_0\|^2 + \sigma^2\frac{d}{n}. \tag{3.17}$$

We will now show that a dimension penalty yields an estimator with mean square error which minimizes (3.17) over d, up to logarithmic factors in the non-nested case, and up to a constant in the nested case. In both situations, we take $g_* = \arg\min\tau^2(g|g_0)$.

3.2.1. Non-nested case. We define the dimension of g as the number of non-zero coefficients:

$$d_g = \#\{\alpha_k \ne 0\},$$

and take the penalty

$$\text{pen}^2(g) = \lambda_0^2\frac{d_g}{n}. \tag{3.18}$$

The penalized least squares estimator $\hat{g}_n$ is now the estimator with hard thresholding, i.e.,

$$\hat{g}_n = \sum_{k=1}^{n}\hat{\alpha}_k\psi_k,$$

where

$$\hat{\alpha}_k = \begin{cases}\tilde{\alpha}_k & \text{if } |\tilde{\alpha}_k| > \lambda_0/\sqrt{n}\\ 0 & \text{if } |\tilde{\alpha}_k| \le \lambda_0/\sqrt{n}\end{cases},$$

and where $\tilde{\alpha}_k$ is the empirical coefficient

$$\tilde{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_k(z_i).$$

We shall not use the explicit expression, so that generalizations to other estimation methods or other dimension penalties remain in sight.

Lemma 3.8. *Take $\lambda_0 \geq c_1\sqrt{\log n}$, where c_1 is a large enough constant depending on K. Then*

$$\mathbf{E}\tau^2(\hat{g}|g_0) \leq 2\tau^2(g_*|g_0) + \frac{c_0}{n}. \tag{3.19}$$

Proof. If $g \in \mathcal{G}_*(\delta)$, we must have that $d_g \leq \lfloor n\delta^2/\lambda_0^2 \rfloor = d(\delta)$, where $\lfloor a \rfloor$ denotes the integer (including zero) part of $a > 0$. There are

$$\binom{n}{d} \leq n^d$$

linear subspaces of dimension d. It is therefore easy to see from Lemma 3.2 that

$$H(u, \mathcal{G}_*(\delta)) \leq d(\delta)\log(\frac{5\delta}{u}) + d(\delta)\log n$$

$$\leq \frac{n\delta^2}{\lambda_0^2}\left(\log(\frac{5\delta}{u}) + \log n\right).$$

So

$$J_*(\delta) \leq A_0 \frac{\sqrt{n}\delta^2}{\lambda_0}\sqrt{\log n} \vee \delta = \Psi_*(\delta),$$

where

$$A_0 = \int_0^1 \log^{1/2}(\frac{5}{u})du + 1.$$

We find that $\sqrt{n}\delta_n \geq c\Psi_*(\delta)$ for $\sqrt{n}\delta_n \geq c$ and $\lambda_0 \geq cA_0\sqrt{\log n}$. □

3.2.2. Nested case. We now let the dimension of g be the last non-zero coefficient:

$$d_g = \max\{k : \alpha_k \neq 0\}.$$

The penalty is

$$\text{pen}^2(g) = \lambda_0^2 \frac{d_g}{n}. \tag{3.20}$$

Lemma 3.9. *For $\lambda_0 \geq c_1$, with c_1 a large enough constant depending on K, we have*

$$\mathbf{E}\tau^2(\hat{g}_n|g_0) \leq 2\tau^2(g_*|g_0) + \frac{c_0}{n}. \tag{3.21}$$

Proof. It is clear that $\mathcal{G}_*(\delta)$ is now the linear subspace $\{g = g_* + \sum_{k=1}^{d(\delta)} \alpha_k\psi_k\}$, with $d(\delta) = \lfloor \frac{n\delta^2}{\lambda_0^2} \rfloor$. So by Lemma 3.2,

$$H(u, \mathcal{G}_*(\delta)) \leq d(\delta)\log(\frac{5\delta}{u}) \leq \frac{n\delta^2}{\lambda_0^2}\log(\frac{5\delta}{u}).$$

Hence

$$J_*(\delta) \leq A_0\frac{\sqrt{n}\delta^2}{\lambda_0} \vee \delta = \Psi_*(\delta),$$

with

$$A_0 = \int_0^1 \log^{1/2}(\frac{5}{u})du.$$

□

3.3. Soft thresholding. As we already indicated in Section 2, Theorem 2.1 is simple and quite general, but cannot catch all the particularities of special cases. Here are some points of attention:
(i) If $\text{pen}(g_*|g_0)$ is of larger order than $\|g_* - g_0\|$, Theorem 2.1 might give rise to too pessimistic rates for $\|\hat{g}_n - g_0\|$.
(ii) In the proof of Theorem 2.1, we did not use sophisticated lower bounds for

$$\tau^2(\hat{g}_n|g_0) - \tau^2(g_*|g_0).$$

(iii) The entropy integral in (2.5) might be unnecessary large (but not larger than $n\delta A_0$, with the constant A_0 equal to $\int_0^1 \log^{1/2}(5/u)du$, because $\mathcal{G}_*(\delta)$ is a subset of a ball with radius δ in n-dimensional Euclidean space.) One may replace the definition (2.5) of $J_*(\delta)$ by

$$J_*(\delta) = \int_{\delta^2/c'}^{\delta} H^{1/2}(u, \mathcal{G}_*(\delta))du, \tag{3.22}$$

where c' is a constant depending on the distribution of the errors (see Birgé and Massart (1993), or van de Geer (1995, 2000)). Nevertheless, the result may be too rough (such as too many log-factors: see below (3.25) combined with Lemma 3.11).
(iv) The theorem relies on entropy methods to handle empirical processes. These methods do not lead to good constants. (Tracing back the behaviour of the constant c of Theorem 2.1 in e.g. the proof of Corollary 8.4 in van de Geer (2000) would lead to immense quantities.) Concentration inequalities (e.g., from Ledoux and Talagrand (1991)) can be used to improve on this.

We shall now consider an example where (i) occurs. Let, as in the previous example, $\mathcal{G}$ be the class of all functions on $z_1, \ldots, z_n$. Suppose for definiteness that the z_i are distinct, and let $\psi_1, \ldots, \psi_n$ be an orthonormal basis in $L_2(Q_n)$. We consider for some $\lambda_0 \geq 1$, the penalty

$$\text{pen}^2(g) = \frac{2\lambda_0}{\sqrt{n}} \sum_{k=1}^n |\alpha_k|, \; g = \sum_{k=1}^n \alpha_k \psi_k. \tag{3.23}$$

The penalized least squares estimator $\hat{g}_n$ is now the estimator with soft thresholding, i.e.,

$$\hat{g}_n = \sum_{k=1}^n \hat{\alpha}_k \psi_k,$$

where

$$\hat{\alpha}_k = \begin{cases} \tilde{\alpha}_k - \lambda_0/\sqrt{n} & \text{if } \tilde{\alpha}_k > \lambda_0/\sqrt{n} \\ \tilde{\alpha}_k + \lambda_0/\sqrt{n} & \text{if } \tilde{\alpha}_k < -\lambda_0/\sqrt{n} \\ 0 & \text{if } |\tilde{\alpha}_k| \leq \lambda_0/\sqrt{n} \end{cases},$$

and where $\tilde{\alpha}_k$ is the empirical coefficient

$$\tilde{\alpha}_k = \frac{1}{n} \sum_{i=1}^n Y_i \psi_k(z_i).$$

Thus, the soft thresholding estimator is very similar to the hard thresholding estimator. However, we shall again not use explicit expressions, so that the methodology can be extended to other estimation problems.

We shall now consider problem (i). For

$$g_* = \arg\min \tau^2(g|g_0), \tag{3.24}$$

one has

$$g_* = \sum_{k=1}^n \hat{\alpha}_{k,*}\psi_k,$$

where

$$\hat{\alpha}_{k,*} = \begin{cases} \alpha_{k,0} - \lambda_0/\sqrt{n} & \text{if } \alpha_{k,0} > \lambda_0/\sqrt{n} \\ \alpha_{k,0} + \lambda_0/\sqrt{n} & \text{if } \alpha_{k,0} < -\lambda_0/\sqrt{n} \\ 0 & \text{if } |\alpha_{k,0}| \le \lambda_0/\sqrt{n} \end{cases},$$

and where $\{\alpha_{k,0}\}$ are the coefficients of g_0. Let $\mathcal{K}_0 = \{\alpha_{k,0} : |\alpha_{k,0}| > \lambda_0/\sqrt{n}\}$. Then

$$\text{pen}^2(g_*) = \frac{2\lambda_0}{\sqrt{n}} \sum_{k \in \mathcal{K}_0} |\alpha_{k,*}|.$$

Therefore, as soon as g_0 has coefficients larger than one, $\text{pen}^2(g_*)$ is larger than $\lambda_0/\sqrt{n}$. So the best Theorem 2.1 can give here is the very slow $n^{-1/4}$ rate of convergence for $\|\hat{g}_n - g_0\|$. We are thus confronted with problem (i): Theorem 2.1 cannot yield good rates here.

Nevertheless, one may ask whether entropy calculations can be used to study the empirical process. To look at this more closely, we present a result of Loubes and van de Geer (2000), which says that the set

$$\mathcal{A} = \{\alpha \in \mathbf{R}^n : \sum_{k=1}^n |\alpha_k| \le 1\}$$

has entropy

$$H(u, \mathcal{A}) \le A^2 \frac{1}{u^2}\left(\log n + \log \frac{1}{u}\right),\ u > 0. \tag{3.25}$$

Putting this bound in the integral of the square root entropy (equation (3.22)) will give rise to unnecessary log-factors (see Lemma 3.11 below for the the behaviour of the empirical process).

Let us also address problem (ii). Define

$$\text{pen}_0^2(g) = \frac{2\lambda_0}{\sqrt{n}} \sum_{k \in \mathcal{K}_0} |\alpha_{k,0}|,\ g = \sum_{k=1}^n \alpha_k \psi_k. \tag{3.26}$$

Lemma 3.10. *We have for* $g = \sum_{k=1}^n \alpha_k \psi_k$,

$$\tau^2(g|g_0) - \tau^2(g_*|g_0) \ge \|g - g_0\|^2 - \|g_* - g_0\|^2 - \text{pen}_0^2(g - g_*) + \frac{2\lambda_0}{\sqrt{n}} \sum_{k \notin \mathcal{K}_0} |\alpha_k|. \tag{3.27}$$

Proof. This is true because

$$\text{pen}^2(g) - \text{pen}^2(g_*) = \frac{2\lambda_0}{\sqrt{n}} \sum_{k=1}^n (|\alpha_k| - |\alpha_{k,*}|)$$

$$\ge -\frac{2\lambda_0}{\sqrt{n}} \sum_{k \in \mathcal{K}_0} |\alpha_k - \alpha_{k,*}| + \frac{2\lambda_0}{\sqrt{n}} \sum_{k \notin \mathcal{K}_0} |\alpha_k|.$$

□

Next, we consider the empirical process.

Lemma 3.11. *We have*

$$\mathbf{P}\left(\sup_g \frac{\frac{1}{n}|\sum_{i=1}^n W_i(g(z_i) - g_*(z_i))|}{\text{pen}^2(g - g_*)} > \frac{3K\sqrt{\log n}}{\lambda_0}\right) \le 2n^{-\frac{5}{4}}. \tag{3.28}$$

Proof. Define

$$V_k = \frac{1}{n} \sum_{i=1}^n W_i \psi_k(z_i), \ k = 1, \ldots, n.$$

Then clearly, for $g = \sum_{k=1}^n \alpha_k \psi_k$,

$$\frac{1}{n} | \sum_{i=1}^n W_i (g(z_i) - g_*(z_i)) | = | \sum_{k=1}^n V_k (\alpha_k - \alpha_{k,*}) | \leq \max_{1 \leq k \leq n} |V_k| \sum_{k=1}^n |\alpha_k - \alpha_{k,*}|$$

$$= \max_{1 \leq k \leq n} |V_k| \text{pen}^2(g - g_*) \sqrt{n} / (2\lambda_0).$$

By Lemma 8.2 in van de Geer (2000), we know that the condition (1.3) on the errors implies that for each k,

$$\mathbf{P}(|V_k| > a) \leq 2 \exp[-\frac{na^2}{16K^2}], \ a > 0.$$

Take $a = 6K\sqrt{\log n/n}$ to find for each k,

$$\mathbf{P}(|V_k| > 6K\sqrt{\log n/n}) \leq 2n^{-\frac{9}{4}}.$$

Therefore,

$$\mathbf{P}(\max_{1 \leq k \leq n} |V_k| > 6K\sqrt{\log n/n}) \leq 2n^{-\frac{5}{4}}.$$

□

A similar result can be found using the entropy bound (3.22), but then one gets additional log-factors (see also Lemma 4.2 in Loubes and van de Geer (2000)).

Let us now define $d_0 = |\mathcal{K}_0|$, i.e. d_0 is the number of coefficients $\alpha_{k,0}$ of g_0 with $|\alpha_{k,0}| > \lambda_0/\sqrt{n}$. Note that in fact,

$$d_0 = d_{g_{**}},$$

where

$$g_{**} = \arg\min\{\|g - g_0\|^2 + \lambda_0^2 \frac{d_g}{n}\},$$

and where d_g is the number of non-zero coefficients of the function g.

Lemma 3.12. *We have, for $\lambda_0 \geq 6K\sqrt{\log n}$, almost surely for all n sufficiently large,*

$$\|\hat{g}_n - g_0\|^2 \leq 6 \left(\|g_* - g_0\|^2 + (4\lambda_0)^2 \frac{d_0}{n} \right). \tag{3.29}$$

Proof. We use the inequality

$$\frac{2}{n} \sum_{i=1}^n W_i (\hat{g}_n(z_i) - g_*(z_i)) \geq \tau^2(\hat{g}_n | g_0) - \tau^2(g_* | g_0). \tag{3.30}$$

By Lemma 3.11, almost surely for all n sufficiently large,

$$\frac{2}{n} | \sum_{i=1}^n W_i (\hat{g}_n(z_i) - g_*(z_i)) | \leq \frac{6K\sqrt{\log n}}{\lambda_0} \text{pen}^2(\hat{g}_n - g_0)$$

$$\leq \text{pen}^2(\hat{g}_n - g_*).$$

But

$$\text{pen}^2(\hat{g}_n - g_*) = \text{pen}_0^2(\hat{g}_n - g_*) + \frac{2\lambda_0}{\sqrt{n}} \sum_{k \notin \mathcal{K}_0} |\hat{\alpha}_k|.$$

Invoking this in (3.30), and using Lemma 3.10 gives

$$\|\hat{g}_n - g_0\|^2 \le \|g_* - g_0\|^2 + 2\text{pen}_0^2(\hat{g}_n - g_*).$$

By Cauchy-Schwarz,

$$\text{pen}_0^2(\hat{g}_n - g_*) \le 2\lambda_0 \sqrt{\frac{d_0}{n}} \|\hat{g}_n - g_*\|.$$

So we arrive at

$$\|\hat{g}_n - g_0\|^2 \le \|g_* - g_0\|^2 + 4\lambda_0 \sqrt{\frac{d_0}{n}} \|\hat{g}_n - g_*\|, \tag{3.31}$$

almost surely, for all n sufficiently large.

If $\|g_* - g_0\|^2 \le \frac{1}{5} 4\lambda_0 \sqrt{d_0/n}$, this gives

$$\|\hat{g}_n - g_0\|^2 \le \frac{6}{5} 4\lambda_0 \sqrt{\frac{d_0}{n}} \|\hat{g}_n - g_*\|,$$

So then

$$\|\hat{g}_n - g_*\| \le \|\hat{g}_n - g_0\| + \|g_* - g_0\|$$
$$\le \frac{(\sqrt{6}+1)^2}{5} 4\lambda_0 \sqrt{d_0/n},$$

and hence

$$\|\hat{g}_n - g_0\|^2 \le (\frac{(\sqrt{6}+1)^2}{5} + \frac{1}{5})^2 (4\lambda_0)^2 \frac{d_0}{n} \le 6(4\lambda_0)^2 \frac{d_0}{n}.$$

On the other hand, if $\|g_* - g_0\|^2 \ge \frac{1}{5} 4\lambda_0 \sqrt{d_0/n}$, we obtain from (3.31)

$$\|\hat{g}_n - g_0\|^2 \le 6\|g_* - g_0\|^2.$$

□

4. Conclusion. The examples show that the general approach of Theorem 2.1 can be useful in a variety of situations. There are clearly some drawbacks, such as the fact that penalized least squares is certainly not always the best method for obtaining adaptive estimators (alternatives are for example unbiased risk estimation, or using empirical complexities and sample splitting), the method of proof gives no explicit or very large constants, and penalties can be very large, i.e., larger than the estimation error (see in Section 3.3 for an illustration). On the other hand, this general method to access rates of convergence clarifies common features and can inspire the development of new methods. Moreover, the method of proof allows extensions to other estimation methods. For example, one may derive similar results for the least absolute deviations estimator, or quantile regression estimators. The issue of the constants then opens another window: to get good bounds, concentration inequalities for rather complicated empirical processes are required. Note furthermore that the least squares method often requires a good estimator of the variance of the errors, because constants depend on it. On the other hand, e.g. the least absolute deviations estimator can be based on universal constants. Simulation studies might help here to turn the theory into a practical method.

We have considered three examples. The first example is on Sobolev spaces. Since the results are based on entropy calculations only, it can be easily adjusted to cover, say, Besov spaces. Also, one may choose for a sequence space formulation. We remark that in general, the design (i.e. the configuration of the explanatory variables $z_1, \ldots, z_n$) will play a role in the entropy calculations or in the re-formulation in sequence spaces.

The penalties we have chosen in Section 3.1, are certainly not the best ones. For example, consider the sequence space formulation

$$\tilde{Y}_k = \alpha_{k,0} + V_k, \ k = 1, \ldots, n,$$

where the $\alpha_0 = (\alpha_{1,0}, \ldots, \alpha_{n,0})$ is unknown and where $V_1, \ldots, V_n$ are independent normally distributed variables with mean zero and variance σ^2/n. To estimate α_0, consider penalties of the form $\text{pen}^2(\alpha) = \sum_{k=1}^n w_k^2 \alpha_k^2$, with $w_1^2, \ldots, w_n^2$ given weights. In Nussbaum (1985), one can find the optimal choice of the weights in minimax sense for the situation where the model corresponds to estimating a function in a Sobolev space, with given number of derivatives s. This gives the minimax bound (3.12). The penalty (3.7) in Subsection 3.1.1. corresponds to the choice $w_k = \lambda_s k^s$, which is not the optimal choice as given in Nussbaum (1985).

Clearly, the penalty (3.10) in Subsection 3.1.2 is also not optimal in minimax sense. However, a small improvement in our result as compared to the minimax result is that when $I_s(g_0) = 0$, we arrive at almost the parametric rate.

The penalty of Subsection 3.1.4, for estimating s, is comparable to the one used by Kohler, Krzyżak and Schäfer (2000). They consider multivariate Sobolev classes, and employ a truncation device in their estimation method. Their results are similar the ones obtained in Subsection 3.1.4. The results can again be improved by other estimation methods. Golubev and Nussbaum (1996) show that the estimates used in Nussbaum (1985) can be made adaptive, in the sense that they no longer depend on the Sobolev radius M or on the number of derivatives s and yet attain the minimax bound (3.12). The method used for this purpose is unbiased risk minimization. Thus our result for estimating s is not optimal even in rate. We believe this is due to the estimation method, and not to our method of proof. On the other hand, the sequence space formulation and unbiased risk estimation relies heavily on the particular situation, and moreover, it might be not as common practice as adjusting a posteriori the degree of smoothness of a spline estimate.

Penalized methods can be seen as Bayesian methods, with the penalty playing the role of (say) minus the logarithm of the prior distribution. The penalized least squares estimator is then a posterior mode estimator. For the estimation of a function in Sobolev space, Belitser and Ghosal (2000) also consider adaptive Bayesian procedures. They use the sequence space formulation, and show that the probability mass the posterior distribution of the smoothness parameter s puts on values smaller than the "truth" (under-smoothing) tends to zero as $n \to \infty$. This result needs virtually no conditions on the prior on s. However, it is clear that this result as such gives no indication about the posterior mode estimator, which works only for special priors.

We have also considered some dimension penalties. (Almost) adaptivity of the hard and soft thresholding estimator is for example considered in Donoho and Johnstone (1996). The procedure can be refined for various wavelets, with the threshold depending on a particular resolution level. General (dimension) penalties are considered by Birgé and Massart (1997) and Barron, Birgé and Massart (1999). The latter treat a large variety of models and methods (e.g., maximum likelihood, least squares, least absolute deviations). In Subsection 3.2, we have defined dimension in such a way that it depends on the choice of the basis. Of course, the dimension of a linear space does not depend on the basis. Theorem 2.1 can be applied in various situations, allowing various linear spaces as a model.

The extension of the theory for soft thresholding type estimators has been considered in Loubes and van de Geer (2000). Here, entropy methods are used to get results for the empirical processes involved.

Keeping the approximation theory separate helps to gain insight in the statistical problem. We therefore did not present any detailed results on approximation theory, such as the best approximation in d-dimensional space of a Besov function with smoothness s. But, for example, it can be shown that Lemma 3.8 leads to a rate of the form $(\log n/n)^{-\frac{s}{2s+1}}$ (use e.g. Birgé and Massart (1996)). This means one has adaptation in rate, up to logarithmic factors.

References

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301-413

Belitser, E. and Ghosal, S. (2000). Adaptive Bayesian inference on the mean of an infinite dimensional normal distribution. Report MI 2000-06, University of Leiden

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory and Related Fields* **97**, 113-150

Birgé, L. and Massart, P. (1996). An adaptive compression algorithm in Besov spaces. Technical Report, Université de Paris-Sud

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In: *Festschrift for Lucien Le Cam* (Eds. D. Pollard, E. Torgersen and G.L. Yang). Springer, New York

BIRMAN, M. Š. and SOLOMJAK, M.Z. (1967). Piece-wise polynomial approximations of functions in the classes W_p^α. *Mathematics of the USSR Sbornik* **73**, 295-317

DONOHO, D.L. and JOHNSTONE, I.M. (1996), Neo-classcal minimax problems, thresholding and adaptive function estimation. *Bernoulli*, **2**, 39-62

EFROIMOVICH, S.Yu. and PINSKER, M.S. (1984). Learning algorithm for nonparametric filtering. *Automat. Remote Control* **45**, 1434-1440

GOLUBEV, G.K. and NUSSBAUM, M. (1996). Adaptive spline estimates for nonparametric regression models. *Theory Probab. Appl.* **37**, 521-529

KOHLER, M., KRZYŻAK, A. and SCHÄFER, D. (2000). Application of structural risk minimization to multivariate smoothing spline regression estimates. Preprint, Universität Stuttgart.

KOLMOGOROV, A.N. and TIHOMIROV, V.M. (1959). ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspehi Mat. Nauk.* **14** 3-86 [English transl.: *Amer. Math. Soc. Transl.* 2 (1961) **17**, 277-364]

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer Verlag, New York

LEPSKII, O. (1991). Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36**, 682-697

LEPSKII, O. (1992). Asymptotically minimax adaptive estimation II: Schemes without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37**, 433-448

LOUBES, J.-M. and VAN DE GEER, S.A. (2000). Adaptive estimation in regression, using soft thresholding type penalties. Report MI 2000-18, University of Leiden

LUGOSI, G. and NOBEL, A.B. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27**, 1830-1864

NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984-997

VAN DE GEER, S.A. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907-924

VAN DE GEER, S.A. (1995). The method of sieves and minimum contrast estimators. *Mathematical Methods of Statistics* **4**, 20-38

VAN DE GEER, S.A. and WEGKAMP, M. (1996). Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.* **24**, 2513-2523

VAN DE GEER, S.A. (2000). *Empirical processes in M-estimation.* Cambridge University Press

SARA A. VAN DE GEER
MATHEMATICAL INSTITUTE
UNIVERSITY OF LEIDEN
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS
GEER@MATH.LEIDENUNIV.NL