An l1-Oracle Inequality for the Lasso

Pascal Massart, Caroline Meynet

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An ℓ_1-Oracle Inequality for the Lasso

Pascal Massart — Caroline Meynet

N° 7356

Juillet 2010

Optimization, Learning and Statistical Methods

ISSN 0249-6399 ISRN INRIA/RR--7356--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

An ℓ_1-Oracle Inequality for the Lasso

Pascal Massart * † , Caroline Meynet * †

Theme : Optimization, Learning and Statistical Methods
Équipes-Projets SELECT

Rapport de recherche n° 7356 — Juillet 2010 — 49 pages

Abstract: These last years, while many efforts have been made to prove that the Lasso behaves like a variable selection procedure at the price of strong assumptions on the geometric structure of these variables, much less attention has been paid to the analysis of the performance of the Lasso as a regularization algorithm. Our first purpose here is to provide a result in this direction by proving that the Lasso works almost as well as the deterministic Lasso provided that the regularization parameter is properly chosen. This result does not require any assumption at all, neither on the structure of the variables nor on the regression function.

Our second purpose is to introduce a new estimator particularly adapted to deal with infinite countable dictionaries. This estimator is constructed as an ℓ_0-penalized estimator among a sequence of Lasso estimators associated to a dyadic sequence of growing truncated dictionaries. The selection procedure automatically chooses the best level of truncation of the dictionary so as to make the best tradeoff between approximation, ℓ_1-regularization and sparsity. From a theoretical point of view, we shall provide an oracle inequality satisfied by this selected Lasso estimator.

All the oracle inequalities presented in this paper are obtained via the application of a single general theorem of model selection among a collection of nonlinear models. The key idea that enables us to apply this general theorem is to see ℓ_1-regularization as a model selection procedure among ℓ_1-balls.

Finally, rates of convergence achieved by the Lasso and the selected Lasso estimators on a wide class of functions are derived from these oracle inequalities, showing that these estimators perform at least as well as greedy algorithms.

Key-words: Lasso, ℓ_1-oracle inequalities, Model selection by penalization, ℓ_1-balls, Generalized linear Gaussian model.

* Université Paris-Sud, Laboratoire de Mathématiques, UMR 8628, 91405 Orsay, France
† INRIA Saclay Ile-de-France, Projet SELECT

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Une inégalité oracle ℓ_1 pour le Lasso

Résumé : Ces dernières années, alors que de nombreux efforts ont été faits pour prouver que le Lasso agit comme une procédure de sélection de variables au prix d'hypothèses contraignantes sur la structure géométrique de ces variables, peu de travaux analysant la performance du Lasso en tant qu'algorithme de régularisation ℓ_1 ont été réalisés. Notre premier objectif est de fournir un résultat dans cette voie en prouvant que le Lasso se comporte presque aussi bien que le Lasso déterministe à condition que le paramètre de régularisation soit bien choisi. Ce résultat ne nécessite aucune hypothèse, ni sur la structure des variables, ni sur la fonction de régression.

Notre second objectif est de contruire un nouvel estimateur particulièrement adapté à l'utilisation de dictionnaires infinis. Cet estimateur est construit par pénalisation ℓ_0 d'une suite d'estimateurs Lasso associés à une suite dyadique croissante de dictionnaires tronqués. L'algorithme correspondant choisit automatiquement le niveau de troncature garantissant le meilleur compromis entre approximation, régularisation ℓ_1 et parcimonie. D'un point de vue théorique, nous établissons une inégalité oracle satisfaite par cet estimateur.

Toutes les inégalités oracles présentées dans cet article sont obtenues en appliquant un théorème de sélection de modèles parmi un ensemble de modèles non linéaires, grâce à l'idée clé qui consiste à envisager la régularisation ℓ_1 comme une procédure de sélection de modèles parmi des boules ℓ_1.

Enfin, nous déduisons de ces inégalités oracles des vitesses de convergence sur de larges classes de fonctions montrant en particulier que les estimateurs Lasso sont aussi performants que les algorithmes greedy.

Mots-clés : Lasso, inégalités oracles ℓ_1, sélection de modèles par pénalisation, boules ℓ_1, modèles linéaires gaussiens généralisés.

Contents

1 Introduction

We consider the problem of estimating a regression function f belonging to a Hilbert space $\mathbb{H}$ in a fairly general Gaussian framework which includes the fixed design regression or the white noise frameworks. Given a dictionary $\mathcal{D} = \{\phi_j\}_j$ of functions in $\mathbb{H}$, we aim at constructing an estimator $\hat{f} = \hat{\theta}.\phi := \sum_j \hat{\theta}_j \, \phi_j$ of f which enjoys both good statistical properties and computational performance even for large or infinite dictionaries.

For high-dimensional dictionaries, direct minimization of the empirical risk can lead to overfitting and we need to add a complexity penalty to avoid it. One could use an ℓ_0-penalty, i.e. penalize the number of non-zero coefficients $\hat{\theta}_j$ of $\hat{f}$ (see [4] for instance) so as to produce interpretable sparse models but there is no efficient algorithm to solve this non-convex minimization problem when the size of the dictionary becomes too large. On the contrary, ℓ_1-penalization leads to convex optimization and is thus computationally feasible even for high-dimensional data. Moreover, due to its geometric properties, ℓ_1-penalty tends to produce some coefficients that are exactly zero and hence often behaves like

an ℓ_0-penalty. These are the main motivations for introducing ℓ_1-penalization rather than other penalizations.

In the linear regression framework, the idea of ℓ_1-penalization was first introduced by Tibshirani [18] who considered the so-called Lasso estimator (Least Absolute Shrinkage and Selection Operator). Then, lots of studies on this estimator have been carried out, not only in the linear regression framework but also in the nonparametric regression setup with quadratic or more general loss functions (see [3], [14], [19] among others). In the particular case of the fixed design Gaussian regression models, if we observe n i.i.d. random couples $(x_1, Y_1), \ldots, (x_n, Y_n)$ such that

$$Y_i = f(x_i) + \sigma\xi_i, \quad i = 1, \ldots, n, \tag{1.1}$$

and if we consider a dictionary $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ of size p, the Lasso estimator is defined as the following ℓ_1-penalized least squares estimator

$$\hat{f}_p := \hat{f}_p(\lambda_p) = \underset{h \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \|Y - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)}, \tag{1.2}$$

where $\|Y - h\|^2 := \sum_{i=1}^n (Y_i - h(x_i))^2 / n$ is the empirical risk of h, $\mathcal{L}_1(\mathcal{D}_p)$ is the linear span of $\mathcal{D}_p$ equipped with the ℓ_1-norm $\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} := \inf\{\|\theta\|_1 = \sum_{j=1}^p |\theta_j| : \ h = \theta.\phi = \sum_{j=1}^p \theta_j \phi_j\}$ and $\lambda_p > 0$ is a regularization parameter.

Since ℓ_1-penalization can be seen as a "convex relaxation" of ℓ_0-penalization, many efforts have been made to prove that the Lasso behaves like a variable selection procedure by establishing sparsity oracle inequalities showing that the ℓ_1-solution mimicks the "ℓ_0-oracle" (see for instance [3] for the prediction loss in the case of the quadratic nonparametric Gaussian regression model). Nonetheless, all these results require strong restrictive assumptions on the geometric structure of the variables. We refer to [6] for a detailed overview of all these restrictive assumptions.

In this paper, we shall explore another approach by analyzing the performance of the Lasso as a regularization algorithm rather than a variable selection procedure. This shall be done by providing an ℓ_1-oracle type inequality satisfied by this estimator (see Theorem 3.2). In the particular case of the fixed design Gaussian regression model, this result says that if $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ with $\max_{j=1,\ldots,p} \|\phi_j\| \leq 1$, then there exists an absolute constant $C > 0$ such that for all $\lambda_p \geq 4\sigma\, n^{-1/2}(\sqrt{\ln p} + 1)$, the Lasso estimator defined by (1.2) satisfies

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\right] \leq C \left[\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) + \frac{\sigma\lambda_p}{\sqrt{n}}\right]. \tag{1.3}$$

This simply means that, provided that the regularization parameter λ_p is properly chosen, the Lasso estimator works almost as well as the deterministic Lasso. Notice that, unlike the sparsity oracle inequalities, the above result does not require any assumption neither on the target function f nor on the structure of the variables ϕ_j of the dictionary $\mathcal{D}_p$, except simple normalization that we can always assume by considering $\phi_j / \|\phi_j\|$ instead of ϕ_j. This ℓ_1-oracle type inequality is not entirely new. Indeed, on the one hand, Barron and al. [9] have

provided a similar risk bound but in the case of a truncated Lasso estimator under the assumption that the target function is bounded by a constant. On the other hand, Rigollet and Tsybakov [16] are proposing a result with the same flavour but with the subtle difference that it is expressed as a probability bound which does not imply (1.3) (see a more detailed explanation in Section 3.2).

We shall derive (1.3) from a fairly general model selection theorem for non linear models, interpreting ℓ_1-regularization as an ℓ_1-balls model selection criterion (see Section 7). This approach will allow us to go one step further than the analysis of the Lasso estimator for finite dictionaries. Indeed, we can deal with infinite dictionaries in various situations.

In the second part of this paper, we shall thus focus on infinite countable dictionaries. The idea is to order the variables of the infinite dictionary $\mathcal{D}$ thanks to the a priori knowledge we can have of these variables, then write the dictionary $\mathcal{D} = \{\phi_j\}_{j\in\mathbb{N}^*} = \{\phi_1, \phi_2, \dots\}$ according to this order, and consider the dyadic sequence of truncated dictionaries $\mathcal{D}_1 \subset \cdots \subset \mathcal{D}_p \subset \cdots \subset \mathcal{D}$ where $\mathcal{D}_p = \{\phi_1, \dots, \phi_p\}$ for $p \in \{2^J, J \in \mathbb{N}\}$. Given this sequence $(\mathcal{D}_p)_p$, we introduce an associated sequence of Lasso estimators $(\hat{f}_p)_p$ with regularization parameters λ_p depending on p, and choose $\hat{f}_{\hat{p}}$ as an ℓ_0-penalized estimator among this sequence by penalizing the size of the truncated dictionaries $\mathcal{D}_p$. This selected Lasso estimator $\hat{f}_{\hat{p}}$ is thus based on an algorithm choosing automatically the best level of truncation of the dictionary and is constructed to make the best tradeoff between approximation, ℓ_1-regularization and sparsity. From a theoretical point of view, we shall establish an oracle inequality satisfied by this selected Lasso estimator. Of course, although introduced for infinite dictionaries, this estimator remains well defined for finite dictionaries and it may be profitable to exploit its good properties and to use it rather than the classical Lasso for such dictionaries.

In a third part of this paper, we shall focus on the rates of convergence of the sequence of the Lassos and the selected Lasso estimator introduced above. We shall provide rates of convergence of these estimators for a wide range of function classes described by mean of interpolation spaces $\mathcal{B}_{q,r}$ that are adapted to the truncation of the dictionary and constitute an extension of the intersection between weak-$\mathcal{L}_q$ spaces and Besov spaces $\mathcal{B}^r_{2,\infty}$ for non orthonormal dictionaries. Our results will prove that the Lasso estimators $\hat{f}_p$ for p large enough and the selected Lasso estimator $\hat{f}_{\hat{p}}$ perform as well as the greedy algorithms described by Barron and al. in [1]. Besides, our convergence results shall highlight the advantage of using the selected Lasso estimator rather than Lassos. Indeed, we shall prove that the Lasso estimators $\hat{f}_p$, like the greedy algorithms in [1], are efficient only for p large enough compared to the unknown parameters of smoothness of f whereas $\hat{f}_{\hat{p}}$ always achieves good rates of convergence whenever the target function f belongs to some interpolation space $\mathcal{B}_{q,r}$. In particular, we shall check that these rates of convergence are optimal by establishing a lower bound of the minimax risk over the intersection between $\mathcal{L}_q$ spaces and Besov spaces $\mathcal{B}^r_{2,\infty}$ in the orthonormal case.

We shall end this paper by providing some theoretical results on the performance of the Lasso for particular infinite uncountable dictionaries such as those used for neural networks. Although Lasso solutions can not be computed in practice for such dictionaries, our purpose is just to point out the fact that the

Lasso theoretically performs as well as the greedy algorithms in [1], by establishing rates of convergence based on an ℓ_1-oracle type inequality similar to (1.3) satisfied by the Lasso for such dictionaries.

The article is organized as follows. The notations and the generalized linear Gaussian framework in which we shall work throughout the paper are introduced in Section 2. In Section 3, we consider the case of finite dictionaries and analyze the performance of the Lasso as a regularization algorithm by providing an ℓ_1-oracle type inequality which highlights the fact that the Lasso estimator works almost as well as the deterministic Lasso provided that the regularization parameter is large enough. In section 4, we study the case of infinite countable dictionaries and establish a similar oracle inequality for the selected Lasso estimator $\hat{f}_{\hat{p}}$. In section 5, we derive from these oracle inequalities rates of convergence of the Lassos and the selected Lasso estimator for a variety of function classes. Some theoretical results on the performance of the Lasso for the infinite uncountable dictionaries used to study neural networks in the artificial intelligence field are mentioned in Section 6. Finally, Section 7 is devoted to the explanation of the key idea that enables us to derive all our oracle inequalities from a single general model selection theorem and to the statement of this general theorem. The proofs are postponed until Section 8.

2 Models and notations

2.1 General framework and statistical problem

Let us first describe the generalized linear Gaussian model we shall work with. We consider a separable Hilbert space $\mathbb{H}$ equipped with a scalar product $\langle .,. \rangle$ and its associated norm $\|.\|$.

Definition 2.1. [Isonormal Gaussian process] *A centered Gaussian process $(W(h))_{h\in\mathbb{H}}$ is isonormal if its covariance is given by $\mathbb{E}[W(g)W(h)] = \langle g, h\rangle$ for all $g, h \in \mathbb{H}$.*

The statistical problem we consider is to approximate an unknown target function f in $\mathbb{H}$ when observing a process $(Y(h))_{h\in\mathbb{H}}$ defined by

$$Y(h) = \langle f, h\rangle + \varepsilon W(h), \quad h \in \mathbb{H}, \tag{2.1}$$

where $\varepsilon > 0$ is a fixed parameter and W is an isonormal process. This framework is convenient to cover both finite-dimensional models and the infinite-dimensional white noise model as described in the following examples.

Example 2.2. [Fixed design Gaussian regression model] Let $\mathcal{X}$ be a measurable space. One observes n i.i.d. random couples $(x_1, Y_1), \ldots, (x_n, Y_n)$ of $\mathcal{X} \times \mathbb{R}$ such that

$$Y_i = f(x_i) + \sigma\xi_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where the covariates $x_1, \ldots, x_n$ are deterministic elements of $\mathcal{X}$, the errors ξ_i are i.i.d. $\mathcal{N}(0, 1)$, $\sigma > 0$ and $f : \mathcal{X} \mapsto \mathbb{R}$ is the unknown regression function to be estimated. If one considers $\mathbb{H} = \mathbb{R}^n$ equipped with the scalar product $\langle u, v\rangle = \sum_{i=1}^n u_i\, v_i/n$, defines $y = (Y_1, \ldots, Y_n)^T$, $\xi = (\xi_1, \ldots, \xi_n)^T$ and improperly denotes $h = (h(x_1), \ldots, h(x_n))$ for every $h : \mathcal{X} \mapsto \mathbb{R}$, then $W(h) := \sqrt{n}\,\langle \xi, h\rangle$

defines an isonormal Gaussian process on $\mathbb{H}$ and $Y(h) := \langle y, h\rangle$ satisfies (2.1) with $\varepsilon := \sigma/\sqrt{n}$.
Let us notice that

$$\|h\| := \sqrt{\frac{1}{n}\sum_{i=1}^{n} h^2(x_i)} \tag{2.3}$$

corresponds to the $\mathbb{L}_2$-norm with respect to the measure $\nu_x := \sum_{i=1}^n \delta_{x_i}/n$ with δ_u the Dirac measure at u. It depends on the sample size n and on the training sample via $x_1, \ldots, x_n$ but we omit this dependence in notation (2.3).

Example 2.3. [The white noise framework] In this case, one observes $\zeta(x)$ for $x \in [0,1]$ given by the stochastic differential equation

$$d\zeta(x) = f(x)\,dx + \varepsilon\,dB(x) \text{ with } \zeta(0) = 0,$$

where B is a standard Brownian motion, f is a square-integrable function and $\varepsilon > 0$. If we define $W(h) = \int_0^1 h(x)\,dB(x)$ for every $h \in \mathbb{L}_2([0,1])$, then W is an isonormal process on $\mathbb{H} = \mathbb{L}_2([0,1])$, and $Y(h) = \int_0^1 h(x)\,d\zeta(x)$ obeys to (2.1) provided that $\mathbb{H}$ is equipped with its usual scalar product $\langle f, h\rangle = \int_0^1 f(x)h(x)\,dx$. Typically, f is a signal and $d\zeta(x)$ represents the noisy signal received at time x. This framework easily extends to a d-dimensional setting if one considers some multivariate Brownian sheet B on $[0,1]^d$ and takes $\mathbb{H} = \mathbb{L}_2\left([0,1]^d\right)$.

2.2 Penalized least squares estimators

To solve the general statistical problem (2.1), one can consider a dictionary $\mathcal{D}$, i.e. a given finite or infinite set of functions $\phi_j \in \mathbb{H}$ that arise as candidate basis functions for estimating the target function f, and construct an estimator $\hat{f} = \hat{\theta}.\phi := \sum_{j,\,\phi_j\in\mathcal{D}} \hat{\theta}_j\,\phi_j$ in the linear span of $\mathcal{D}$. All the matter is to choose a "good" linear combination in the following meaning. It makes sense to aim at constructing an estimator as the best approximating point of f by minimizing $\|f - h\|$ or, equivalently, $-2\langle f, h\rangle + \|h\|^2$. However f is unknown, so one may instead minimize the empirical least squares criterion

$$\gamma(h) := -2Y(h) + \|h\|^2. \tag{2.4}$$

But since we are mainly interested in very large dictionaries, direct minimization of the empirical least squares criterion can lead to overfitting. To avoid it, one can rather consider a penalized risk minimization problem and consider

$$\hat{f} \in \arg\min_h \gamma(h) + \text{pen}(h), \tag{2.5}$$

where $\text{pen}(h)$ is a positive penalty to be chosen. Finally, since the resulting estimator $\hat{f}$ depends on the observations, its quality can be measured by its quadratic risk $\mathbb{E}[\|f - \hat{f}\|^2]$.

The penalty $\text{pen}\,(h)$ can be chosen according to the statistical target. In the recent years, the situation where the number of variables ϕ_j can be very large (as compared to ε^{-2}) has received the attention of many authors due to the increasing number of applications for which this can occur. Micro-array data analysis

or signal reconstruction from a dictionary of redundant wavelet functions are typical examples for which the number of variables either provided by Nature or considered by the statistician is large. Then, an interesting target is to select the set of the "most significant" variables ϕ_j among the initial collection. In this case, a convenient choice for the penalty is the ℓ_0-penalty that penalizes the number of non-zero coefficients $\hat{\theta}_j$ of $\hat{f}$, thus providing sparse estimators and interpretable models. Nonetheless, except when the functions ϕ_j are orthonormal, there is no efficient algorithm to solve this minimization problem in practice when the dictionary becomes too large. On the contrary, ℓ_1-penalization, that is to say $\mathrm{pen}(h) \propto \|h\|_{\mathcal{L}_1(\mathcal{D})} := \inf\left\{\|\theta\|_1 = \sum_{j,\,\phi_j\in\mathcal{D}} |\theta_j| \quad \text{such that } h = \theta.\phi\right\}$, leads to convex optimization and is thus computationally feasible even for high-dimensional data. Moreover, due to its geometric properties, ℓ_1-penalty tends to produce some coefficients that are exactly zero and thus often behaves like an ℓ_0-penalty, hence the popularity of ℓ_1-penalization and its associated estimator the Lasso defined by

$$\hat{f}(\lambda) = \underset{h\in\mathcal{L}_1(\mathcal{D})}{\arg\min}\ \gamma(h) + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}, \quad \lambda > 0,$$

where $\mathcal{L}_1(\mathcal{D})$ denotes the set of functions h in the linear span of $\mathcal{D}$ with finite ℓ_1-norm $\|h\|_{\mathcal{L}_1(\mathcal{D})}$.

3 The Lasso for finite dictionaries

While many efforts have been made to prove that the Lasso behaves like a variable selection procedure at the price of strong (though unavoidable) assumptions on the geometric structure of the dictionary (see [3] or [6] for instance), much less attention has been paid to the analysis of the performance of the Lasso as a regularization algorithm. The analysis we propose below goes in this very direction. In this section, we shall consider a finite dictionary $\mathcal{D}_p$ of size p and provide an ℓ_1-oracle type inequality bounding the quadratic risk of the Lasso estimator by the infimum over $\mathcal{L}_1(\mathcal{D}_p)$ of the tradeoff between the approximation term $\|f-h\|^2$ and the ℓ_1-norm $\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}$.

3.1 Definition of the Lasso estimator

We consider the generalized linear Gaussian model and the statistical problem (2.1) introduced in the last section. Throughout this section, we assume that $\mathcal{D}_p = \{\phi_1, \ldots, \phi_p\}$ is a finite dictionary of size p. In this case, any h in the linear span of $\mathcal{D}_p$ has finite ℓ_1-norm

$$\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} := \inf\left\{\|\theta\|_1 = \sum_{j=1}^{p} |\theta_j| \ ,\ \theta\in\mathbb{R}^p \text{ such that } h = \theta.\phi\right\} \tag{3.1}$$

and thus belongs to $\mathcal{L}_1(\mathcal{D}_p)$. We propose to estimate f by a penalized least squares estimator as introduced at (2.5) with a penalty $\mathrm{pen}(h)$ proportional to $\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}$. This estimator is the so-called Lasso estimator $\hat{f}_p$ defined by

$$\hat{f}_p := \hat{f}_p(\lambda_p) = \underset{h\in\mathcal{L}_1(\mathcal{D}_p)}{\arg\min}\ \gamma(h) + \lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}, \tag{3.2}$$

where $\lambda_p > 0$ is a regularization parameter and $\gamma(h)$ is defined by (2.4).

Remark 3.1. Let us notice that the general definition (3.2) coincides with the usual definition of the Lasso in the particular case of the classical fixed design Gaussian regression model presented in Example 2.2,

$$Y_i = f(x_i) + \sigma\xi_i, \quad i = 1, \dots, n.$$

Indeed, if we define $y = (Y_1, \dots, Y_n)^T$, we have

$$\gamma(h) = -2Y(h) + \|h\|^2 = -2\langle y, h\rangle + \|h\|^2 = \|y - h\|^2 - \|y\|^2,$$

so we deduce from (3.2) that the Lasso satisfies

$$\hat{f}_p = \underset{h \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left(\|y - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right). \tag{3.3}$$

Let us now consider for all $h \in \mathcal{L}_1(\mathcal{D}_p)$, $\Theta_h := \{\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p,\ h = \theta.\phi = \sum_{j=1}^p \theta_j \phi_j\}$. Then, we get from (3.1) that

$$\begin{aligned}
\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|y - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) &= \inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|y - h\|^2 + \lambda_p \inf_{\theta \in \Theta_h} \|\theta\|_1 \right) \\
&= \inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \inf_{\theta \in \Theta_h} \left(\|y - h\|^2 + \lambda_p \|\theta\|_1 \right) \\
&= \inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \inf_{\theta \in \Theta_h} \left(\|y - \theta.\phi\|^2 + \lambda_p \|\theta\|_1 \right) \\
&= \inf_{\theta \in \mathbb{R}^p} \left(\|y - \theta.\phi\|^2 + \lambda_p \|\theta\|_1 \right).
\end{aligned}$$

Therefore, we get from (3.3) that $\hat{f}_p = \hat{\theta}_p.\phi$ where $\hat{\theta}_p = \arg\min_{\theta \in \mathbb{R}^p} \|y - \theta.\phi\|^2 + \lambda_p \|\theta\|_1$, which corresponds to the usual definition of the Lasso estimator for the fixed design Gaussian regression models with finite dictionaries of size p (see [3] for instance).

3.2 The ℓ_1-oracle inequality

Let us now state the main result of this section.

Theorem 3.2. *Assume that* $\max_{j=1,\dots,p} \|\phi_j\| \leq 1$ *and that*

$$\lambda_p \geq 4\varepsilon \left(\sqrt{\ln p} + 1 \right). \tag{3.4}$$

Consider the corresponding Lasso estimator $\hat{f}_p$ *defined by* (3.2).
Then, there exists an absolute positive constant C *such that, for all* $z > 0$, *with probability larger than* $1 - 3.4\,\mathrm{e}^{-z}$,

$$\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C \left[\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \lambda_p \varepsilon (1 + z) \right]. \tag{3.5}$$

Integrating (3.5) *with respect to* z *leads to the following* ℓ_1*-oracle type inequality in expectation,*

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \right] \leq C \left[\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \lambda_p \varepsilon \right]. \tag{3.6}$$

This ℓ_1-oracle type inequality highlights the fact that the Lasso (i.e. the "noisy" Lasso) behaves almost as well as the deterministic Lasso provided that the regularization parameter λ_p is properly chosen. The proof of Theorem 3.2 is detailed in Section 8 and we refer the reader to Section 7 for the description of the key observation that has enabled us to establish it. In a nutshell, the basic idea is to view the Lasso as the solution of a penalized least squares model selection procedure over a countable collection of models consisting of ℓ_1-balls. Inequalities (3.5) and (3.6) are thus deduced from a general model selection theorem borrowed from [5] and presented in Section 7 as Theorem 7.1.

Remark 3.3.

1. Notice that unlike the sparsity oracle inequalities with ℓ_0-penalty established by many authors ([3], [19], [14] among others), the above result does not require any assumption neither on the target function f nor on the structure of the variables ϕ_j of the dictionary $\mathcal{D}_p$, except simple normalization that we can always assume by considerating $\phi_j/\|\phi_j\|$ instead of ϕ_j.

2. Although such ℓ_1-oracle type inequalities have already been studied by a few authors, no such general risk bound has yet been put forward. Indeed, Barron and al. [9] have provided a risk bound like (3.6) but they restrict to the case of a truncated Lasso estimator under the assumption that the target function is bounded by a constant. For their part, Rigollet and Tsybakov [16] are proposing an oracle inequality for the Lasso similar to (3.5) which is valid under the same assumption as the one of Theorem 3.2, i.e. simple normalization of the variables of the dictionary, but their bound in probability can not be integrated to get an bound in expectation as the one we propose at (3.6). Indeed, first notice that the constant measuring the level of confidence of their risk bound appears inside the infimum term as a multiplicative factor of the ℓ_1-norm whereas the constant z measuring the level of confidence of our risk bound (3.5) appears as an additive constant outside the infimum term so that the bound in probability (3.5) can easily be integrated with respect to z, which leads to the bound in expectation (3.6). Besides, the main drawback of the result given by Tsybakov and Rigollet is that the lower bound of the regularization parameter λ_p they propose (i.e. $\lambda_p \geq \sqrt{8(1+z/\ln p)}\,\varepsilon\sqrt{\ln p}$) depends on the level of condidence z, with the consequence that their choice of the Lasso estimator $\hat{f}_p = \hat{f}_p(\lambda_p)$ also depends on this level of confidence. On the contrary, our lower bound $\lambda_p \geq 4\varepsilon(\sqrt{\ln p}+1)$ does not depend on z so that we are able to get the result (3.5) satisfied with high probability by an estimator $\hat{f}_p = \hat{f}_p(\lambda_p)$ independent of the level of confidence of this probability.

3. Theorem 3.2 is interesting from the point of view of approximation theory. Indeed, as we shall see in Proposition 5.6, it shows that the Lasso performs as well as the greedy algorithms studied in [1] and [9].

4. We can check that the upper bound (3.6) is sharp. Indeed, assume that $p \geq 2$, that $f \in \mathcal{L}_1(\mathcal{D}_p)$ with $\|f\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq R$ and that $R \geq \varepsilon$. Consider the Lasso estimator $\hat{f}_p$ for $\lambda_p = 4\varepsilon(\sqrt{\ln p}+1)$. Then, by bounding the

infimum term in the right-hand side of (3.6) by the value at $h = f$, we get that

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2\right] \leq C\lambda_p \left(\|f\|_{\mathcal{L}_1(\mathcal{D}_p)} + \varepsilon\right) \leq 8CR\varepsilon \left(\sqrt{\ln p} + 1\right), \tag{3.7}$$

where $C > 0$. Now, it is established in Proposition 5 in [2] that there exists $\kappa > 0$ such that the minimax risk over the ℓ_1-balls $S_{R,p} = \{h \in \mathcal{L}_1(\mathcal{D}_p), \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq R\}$ satisfies

$$\inf_{\tilde{h}} \sup_{h \in S_{R,p}} \mathbb{E}\left[\|h - \tilde{h}\|^2\right] \geq \kappa \inf\left(R\varepsilon\sqrt{1 + \ln\left(p\varepsilon R^{-1}\right)},\, p\varepsilon^2,\, R^2\right), \tag{3.8}$$

where the infimum is taken over all possible estimators $\tilde{h}$. Comparing the upper bound (3.7) to the lower bound (3.8), we see that the ratio between them is bounded independently of ε for all $S_{R,p}$ such that the signal to noise ratio $R\varepsilon^{-1}$ is between $\sqrt{\ln p}$ and p. This proves that the Lasso estimator $\hat{f}_p$ is approximately minimax over such sets $S_{R,p}$.

4 A selected Lasso estimator for infinite countable dictionaries

In many applications such as micro-array data analysis or signal reconstruction, we are now faced with situations in which the number of variables of the dictionary is always increasing and can even be infinite. Consequently, it is desirable to find competitive estimators for such infinite dimensional problems. Unfortunately, the Lasso is not well adapted to infinite dictionaries. Indeed, from a practical point of view, there is no algorithm to approximate the Lasso solution over an infinite dictionary because it is not possible to evaluate the infimum of $\gamma(h) + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}$ over the whole set $\mathcal{L}_1(\mathcal{D})$ for an infinite dictionary $\mathcal{D}$, but only over a finite subset of it. Moreover, from a theoretical point of view, it is difficult to prove good results on the Lasso for infinite dictionaries, except in rare situations when the variables have a specific structure (see Section 6 on neural networks).

In order to deal with an infinite countable dictionary $\mathcal{D}$, one may order the variables of the dictionary, write the dictionary $\mathcal{D} = \{\phi_j\}_{j \in \mathbb{N}^*} = \{\phi_1, \phi_2, \dots\}$ according to this order, then truncate $\mathcal{D}$ at a given level p to get a finite subdictionary $\{\phi_1, \dots, \phi_p\}$ and finally estimate the target function by the Lasso estimator $\hat{f}_p$ over this subdictionary. This procedure implies two difficulties. First, one has to put an order on the variables of the dictionary, and then all the matter is to decide at which level one should truncate the dictionary to make the best tradeoff between approximation and complexity. Here, our purpose is to resolve this last dilemma by proposing a selected Lasso estimator based on an algorithm choosing automatically the best level of truncation of the dictionary once the variables have been ordered. Of course, the algorithm and thus the estimation of the target function will depend on which order the variables have been classified beforehand. Notice that the classification of the variables can reveal to be more or less difficult according to the problem under consideration. Nonetheless, there are a few applications where there may be an obvious order for the variables, for instance in the case of dictionaries of wavelets.

In this section, we shall first introduce the selected Lasso estimator that we propose to approximate the target function in the case of infinite countable dictionaries. Then, we shall provide an oracle inequality satisfied by this estimator. This inequality is to be compared to Theorem 3.2 established for the Lasso in the case of finite dictionaries. Its proof is again an application of the general model selection Theorem 7.1. Finally, we make a few comments on the possible advantage of using this selected Lasso estimator for finite dictionaries in place of the classical Lasso estimator.

4.1 Definition of the selected Lasso estimator

We still consider the generalized linear Gaussian model and the statistical problem (2.1) introduced in Section 2. We recall that, to solve this problem, we use a dictionary $\mathcal{D} = \{\phi_j\}_j$ and seek for an estimator $\hat{f} = \hat{\theta}.\phi = \sum_{j,\,\phi_j \in \mathcal{D}} \hat{\theta}_j\, \phi_j$ solution of the penalized risk minimization problem,

$$\hat{f} \in \underset{h \in \mathcal{L}_1(\mathcal{D})}{\arg\min}\, \gamma(h) + \mathrm{pen}(h), \tag{4.1}$$

where $\mathrm{pen}(h)$ is a suitable positive penalty. Here, we assume that the dictionary is infinite countable and that it is ordered,

$$\mathcal{D} = \{\phi_j\}_{j \in \mathbb{N}^*} = \{\phi_1, \phi_2, \dots\}.$$

Given this order, we can consider the sequence of truncated dictionaries $(\mathcal{D}_p)_{p \in \mathbb{N}^*}$ where

$$\mathcal{D}_p := \{\phi_1, \dots, \phi_p\} \tag{4.2}$$

corresponds to the subdictionary of $\mathcal{D}$ truncated at level p, and the associated sequence of Lasso estimators $(\hat{f}_p)_{p \in \mathbb{N}^*}$ defined in Section 3.1,

$$\hat{f}_p := \hat{f}_p(\lambda_p) = \underset{h \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min}\ \gamma(h) + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)}, \tag{4.3}$$

where $(\lambda_p)_{p \in \mathbb{N}^*}$ is a sequence of regularization parameters whose values will be specified below. Now, we shall choose a final estimator as an ℓ_0-penalized estimator among a subsequence of the Lasso estimators $(\hat{f}_p)_{p \in \mathbb{N}^*}$. Let us denote by Λ the set of dyadic integers,

$$\Lambda = \{2^J, J \in \mathbb{N}\}, \tag{4.4}$$

and define

$$\hat{f}_{\hat{p}} = \underset{p \in \Lambda}{\arg\min} \left[\gamma(\hat{f}_p) + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} + \mathrm{pen}(p)\right] \tag{4.5}$$

$$= \underset{p \in \Lambda}{\arg\min} \left[\underset{h \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left(\gamma(h) + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) + \mathrm{pen}(p)\right], \tag{4.6}$$

where $\mathrm{pen}(p)$ penalizes the size p of the truncated dictionary $\mathcal{D}_p$ for all $p \in \Lambda$. From (4.6) and the fact that $\mathcal{L}_1(\mathcal{D}) = \cup_{p \in \Lambda} \mathcal{L}_1(\mathcal{D}_p)$, we see that this selected Lasso estimator $\hat{f}_{\hat{p}}$ is a penalized least squares estimator solution of (4.1) where, for any $p \in \Lambda$ and $h \in \mathcal{L}_1(\mathcal{D}_p)$, $\mathrm{pen}(h) = \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} + \mathrm{pen}(p)$ is a combination of both ℓ_1-regularization and ℓ_0-penalization. We see from (4.5) that the algorithm automatically chooses the rank $\hat{p}$ so that $\hat{f}_{\hat{p}}$ makes the best tradeoff between approximation, ℓ_1-regularization and sparsity.

Remark 4.1. Notice that from a theoretical point of view, one could have defined $\hat{f}_{\hat{p}}$ as an ℓ_0-penalized estimator among the whole sequence of Lasso estimators $(\hat{f}_p)_{p\in\mathbb{N}^*}$ (or more generally among any subsequence of $(\hat{f}_p)_{p\in\mathbb{N}^*}$) instead of $(\hat{f}_p)_{p\in\Lambda}$. Nonetheless, to compute $\hat{f}_{\hat{p}}$ efficiently, it is interesting to limit the number of computations of the sequence of Lasso estimators $\hat{f}_p$ especially if we choose an ℓ_0-penalty $\mathrm{pen}(p)$ that does not grow too fast with p, typically $\mathrm{pen}(p) \propto \ln p$, which will be the case in the next theorem. That is why we have chosen to consider a dyadic truncation of the dictionary $\mathcal{D}$.

4.2 An oracle inequality for the selected Lasso estimator

By applying the same general model selection theorem (Theorem 7.1) as for the establishment of Theorem 3.2, we can provide a risk bound satisfied by the estimator $\hat{f}_{\hat{p}}$ with properly chosen penalties λ_p and $\mathrm{pen}(p)$ for all $p \in \Lambda$. The sequence of ℓ_1-regularization parameters $(\lambda_p)_{p\in\Lambda}$ is simply chosen from the lower bound given by (3.4) while a convenient choice for the ℓ_0-penalty will be $\mathrm{pen}(p) \propto \ln p$.

Theorem 4.2. *Assume that* $\sup_{j\in\mathbb{N}^*} \|\phi_j\| \leq 1$. *Set for all* $p \in \Lambda$,

$$\lambda_p = 4\varepsilon\left(\sqrt{\ln p}+1\right), \qquad \mathrm{pen}(p) = 5\varepsilon^2 \ln p, \tag{4.7}$$

and consider the corresponding selected Lasso estimator $\hat{f}_{\hat{p}}$ *defined by* (4.6). *Then, there exists an absolute constant* $C > 0$ *such that*

$$\begin{aligned}&\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p})\right]\\ &\leq C\left[\inf_{p\in\Lambda}\left(\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right)+\mathrm{pen}(p)\right)+\varepsilon^2\right].\end{aligned} \tag{4.8}$$

Remark 4.3. Our primary motivation for introducing the selected Lasso estimator described above was to construct an estimator adapted from the Lasso and fitted to solve problems of estimation dealing with infinite dictionaries. Nonetheless, we can notice that such a selected Lasso estimator remains well-defined and can also be interesting for estimation in the case of finite dictionaries. Indeed, let $\mathcal{D}_{p_0}$ be a given finite dictionary of size p_0. Assume for simplicity that $\mathcal{D}_{p_0}$ is of cardinal an integer power of two: $p_0 = 2^{J_0}$. Instead of working with the Lasso estimator defined by

$$\hat{f}_{p_0} = \operatorname*{arg\,min}_{h\in\mathcal{L}_1(\mathcal{D}_{p_0})} \gamma(h) + \lambda_{p_0}\|h\|_{\mathcal{L}_1(\mathcal{D}_{p_0})},$$

with $\lambda_{p_0} = 4\varepsilon\left(\sqrt{\ln p_0}+1\right)$ being chosen from the lower bound of Theorem 3.2, one can introduce a sequence of dyadic truncated dictionaries $\mathcal{D}_1 \subset \cdots \subset \mathcal{D}_p \subset \cdots \subset \mathcal{D}_{p_0}$, and consider the associated selected Lasso estimator defined by

$$\hat{f}_{\hat{p}} = \operatorname*{arg\,min}_{p\in\Lambda_0}\left[\operatorname*{arg\,min}_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\gamma(h)+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right)+\mathrm{pen}(p)\right],$$

where $\Lambda_0 = \{2^J, J = 0, \dots, J_0\}$ and where the sequences $\lambda_p = 4\varepsilon\left(\sqrt{\ln p}+1\right)$ and $\mathrm{pen}(p) = 5\varepsilon^2 \ln p$ are chosen from Theorem 4.2. The estimator $\hat{f}_{\hat{p}}$ can be seen as an ℓ_0-penalized estimator among the sequence of Lasso estimators $(\hat{f}_p)_{p\in\Lambda_0}$ associated to the truncated dictionaries $(\mathcal{D}_p)_{p\in\Lambda_0}$,

$$\hat{f}_p = \underset{h\in\mathcal{L}_1(\mathcal{D}_p)}{\arg\min}\ \gamma(h) + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)}.$$

In particular, notice that the selected Lasso estimator $\hat{f}_{\hat{p}}$ and the Lasso estimator $\hat{f}_{p_0}$ coincide when $\hat{p} = p_0$ and that in any case the definition of $\hat{f}_{\hat{p}}$ guarantees that $\hat{f}_{\hat{p}}$ makes a better tradeoff between approximation, ℓ_1-regularization and sparsity than $\hat{f}_{p_0}$. Furthermore, the risk bound (4.8) remains satisfied by $\hat{f}_{\hat{p}}$ for a finite dictionary $\mathcal{D}_{p_0}$ if we replace $\mathcal{D}$ by $\mathcal{D}_{p_0}$ and Λ by Λ_0.

5 Rates of convergence of the Lasso and selected Lasso estimators

In this section, our purpose is to provide rates of convergence of the Lasso and the selected Lasso estimators introduced in Section 3 and Section 4. Since in learning theory one has no or not much a priori knowledge of the smoothness of the unknown target function f in the Hilbert space $\mathbb{H}$, it is essential to aim at establishing performance bounds for a wide range of function classes. Here, we shall analyze rates of convergence whenever f belongs to some real interpolation space between a subset of $\mathcal{L}_1(\mathcal{D})$ and the Hilbert space $\mathbb{H}$. This will provide a full range of rates of convergence related to the unknown smoothness of f. In particular, we shall prove that both the Lasso and the selected Lasso estimators perform as well as the greedy algorithms presented by Barron and al. in [1]. Furthermore, we shall check that the selected Lasso estimator is simultaneously approximately minimax when the dictionary is an orthonormal basis of $\mathbb{H}$ for a suitable signal to noise ratio.

Throughout the section, we keep the same framework as in Section 4.1. In particular, $\mathcal{D} = \{\phi_j\}_{j\in\mathbb{N}^*}$ shall be a given infinite countable ordered dictionary. We consider the sequence of truncated dictionaries $(\mathcal{D}_p)_{p\in\mathbb{N}^*}$ defined by (4.2), the associated sequence of Lasso estimators $(\hat{f}_p)_{p\in\mathbb{N}^*}$ defined by (4.3) and the selected Lasso estimator $\hat{f}_{\hat{p}}$ defined by (4.6) with $\lambda_p = 4\varepsilon(\sqrt{\ln p}+1)$ and $\mathrm{pen}(p) = 5\varepsilon^2 \ln p$ and where Λ still denotes the set of dyadic integers defined by (4.4).

The rates of convergence for the sequence of the Lasso and the selected Lasso estimators will be derived from the oracle inequalities established in Theorem 3.2 and Theorem 4.2 respectively. We know from Theorem 3.2 that, for all $p \in \mathbb{N}^*$, the quadratic risk of the Lasso estimator $\hat{f}_p$ is bounded by

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2\right] \le C\left[\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2 + \lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) + \lambda_p\varepsilon\right], \qquad (5.1)$$

where C is an absolute positive constant, while we know from Theorem 4.2 that the quadratic risk of the selected Lasso estimator $\hat{f}_{\hat{p}}$ is bounded by

$$\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2\right] \leq C\left[\inf_{p\in\Lambda}\left(\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right)+\mathrm{pen}(p)\right)+\varepsilon^2\right], \tag{5.2}$$

where C is an absolute positive constant. Thus, to bound the quadratic risks of the estimators $\hat{f}_{\hat{p}}$ and $\hat{f}_p$ for all $p\in\mathbb{N}^*$, we can first focus on bounding for all $p\in\mathbb{N}^*$,

$$\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right)=\|f-f_p\|^2+\lambda_p\|f_p\|_{\mathcal{L}_1(\mathcal{D}_p)}, \tag{5.3}$$

where we denote by f_p the deterministic Lasso for the truncated dictionary $\mathcal{D}_p$ defined by

$$f_p=\operatorname*{arg\,min}_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right). \tag{5.4}$$

This first step will be handled in Section 5.1 by considering suitable interpolation spaces. Then in Section 5.2, we shall pass on the rates of convergence of the deterministic Lassos to the Lasso and the selected Lasso estimators thanks to the upper bounds (5.1) and (5.2). By looking at these upper bounds, we can expect the selected Lasso estimator to achieve much better rates of convergence than the Lasso estimators. Indeed, for a fixed value of $p\in\mathbb{N}^*$, we can see that the risk of the Lasso estimator $\hat{f}_p$ is roughly of the same order as the rate of convergence of the corresponding deterministic Lasso f_p, whereas the risk of $\hat{f}_{\hat{p}}$ is bounded by the infimum over *all* $p\in\Lambda$ of penalized rates of convergence of the deterministic Lassos f_p.

5.1 Interpolation spaces

Remember that we are first looking for an upper bound of $\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right)$ for all $p\in\mathbb{N}^*$. In fact, this quantity is linked to another one in the approximation theory, which is the so-called $K_{\mathcal{D}_p}$-functional defined below. This link is specified in the following essential lemma.

Lemma 5.1. *Let D be some finite or infinite dictionary. For any $\lambda\geq 0$ and $\delta>0$, consider*

$$L_D(f,\lambda):=\inf_{h\in\mathcal{L}_1(D)}\left(\|f-h\|^2+\lambda\|h\|_{\mathcal{L}_1(D)}\right)$$

and the K_D-functional defined by

$$K_D(f,\delta):=\inf_{h\in\mathcal{L}_1(D)}\left(\|f-h\|+\delta\|h\|_{\mathcal{L}_1(D)}\right). \tag{5.5}$$

Then,

$$\frac{1}{2}\inf_{\delta>0}\left(K_D^2(f,\delta)+\frac{\lambda^2}{2\delta^2}\right)\leq L_D(f,\lambda)\leq\inf_{\delta>0}\left(K_D^2(f,\delta)+\frac{\lambda^2}{4\delta^2}\right). \tag{5.6}$$

Let us now introduce a whole range of interpolation spaces $\mathcal{B}_{q,r}$ that are intermediate spaces between subsets of $\mathcal{L}_1(\mathcal{D})$ and the Hilbert space $\mathbb{H}$ on which the $K_{\mathcal{D}_p}$-functionals (and thus the rates of convergence of the deterministic Lassos f_p) are controlled for all $p\in\mathbb{N}^*$.

Definition 5.2. **[Spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{q,r}$]** *Let $R > 0$, $r > 0$, $1 < q < 2$ and $\alpha = 1/q - 1/2$.*
We say that a function g belongs to the space $\mathcal{L}_{1,r}$ if there exists $C > 0$ such that for all $p \in \mathbb{N}^*$, there exists $g_p \in \mathcal{L}_1(\mathcal{D}_p)$ such that

$$\|g_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C$$

and

$$\|g - g_p\| \leq C\,|\mathcal{D}_p|^{-r} = Cp^{-r}. \tag{5.7}$$

The smallest C such that this holds defines a norm $\|g\|_{\mathcal{L}_{1,r}}$ on the space $\mathcal{L}_{1,r}$.
We say that g belongs to $\mathcal{B}_{q,r}(R)$ if, for all $\delta > 0$,

$$\inf_{h \in \mathcal{L}_{1,r}} \left(\|g - h\| + \delta \|h\|_{\mathcal{L}_{1,r}} \right) \leq R\,\delta^{2\alpha}. \tag{5.8}$$

We say that $g \in \mathcal{B}_{q,r}$ if there exists $R > 0$ such that $g \in \mathcal{B}_{q,r}(R)$. In this case, the smallest R such that $g \in \mathcal{B}_{q,r}(R)$ defines a norm on the space $\mathcal{B}_{q,r}$ and is denoted by $\|g\|_{\mathcal{B}_{q,r}}$.

Remark 5.3. Note that the spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{q,r}$ depend on the choice of the whole dictionary $\mathcal{D}$ as well as on the way it is ordered, but we shall omit this dependence so as to lighten the notations. The set of spaces $\mathcal{L}_{1,r}$ can be seen as substitutes for the whole space $\mathcal{L}_1(\mathcal{D})$ that are adapted to the truncation of the dictionary. In particular, the spaces $\mathcal{L}_{1,r}$ are smaller than the space $\mathcal{L}_1(\mathcal{D})$ and the smaller the value of $r > 0$, the smaller the distinction between them. In fact, looking at (5.7), we can see that working with the spaces $\mathcal{L}_{1,r}$ rather than $\mathcal{L}_1(\mathcal{D})$ will enable us to have a certain amount of control (measured by the parameter r) as regards what happens beyond the levels of truncation.

Thanks to the property of the interpolation spaces $\mathcal{B}_{q,r}$ and to the equivalence established in Lemma 5.1 between the rates of convergence of the deterministic Lassos and the $K_{\mathcal{D}_p}$-functional, we are now able to provide the following upper bound of the rates of convergence of the deterministic Lassos when the target function belongs to some interpolation space $\mathcal{B}_{q,r}$.

Lemma 5.4. *Let $1 < q < 2$, $r > 0$ and $R > 0$. Assume that $f \in \mathcal{B}_{q,r}(R)$.*
Then, there exists $C_q > 0$ depending only on q such that, for all $p \in \mathbb{N}^*$,

$$\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) \leq C_q \max\left(R^q \lambda_p^{2-q}\,,\, \left(Rp^{-r}\right)^{\frac{2q}{2-q}} \lambda_p^{\frac{4(1-q)}{2-q}} \right). \tag{5.9}$$

Remark 5.5. **[Orthonormal case]** Let us point out that the abstract interpolation spaces $\mathcal{B}_{q,r}$ are in fact natural extensions to non-orthonormal dictionaries of function spaces that are commonly studied in statistics to analyze the approximation performance of estimators in the orthonormal case, that is to say Besov spaces, strong-$\mathcal{L}_q$ spaces and weak-$\mathcal{L}_q$ spaces. More precisely, recall that if $\mathbb{H}$ denotes a Hilbert space and $\mathcal{D} = \{\phi_j\}_{j \in \mathbb{N}^*}$ is an orthonormal basis of $\mathbb{H}$, then, for all $r > 0$, $q > 0$ and $R > 0$, we say that $g = \sum_{j=1}^{\infty} \theta_j\,\phi_j$ belongs to the Besov space $\mathcal{B}^r_{2,\infty}(R)$ if

$$\sup_{J \in \mathbb{N}^*} \left(J^{2r} \sum_{j=J}^{\infty} \theta_j^2 \right) \leq R^2, \tag{5.10}$$

while g is said to belong to $\mathcal{L}_q(R)$ if

$$\sum_{j=1}^{\infty} |\theta_j|^q \leq R^q, \tag{5.11}$$

and a slightly weaker condition is that g belongs to $w\mathcal{L}_q(R)$, that is to say

$$\sup_{\eta>0} \left(\eta^q \sum_{j=1}^{\infty} \mathbb{1}_{\{|\theta_j|>\eta\}} \right) \leq R^q. \tag{5.12}$$

Then, we prove in Section 8 that for all $1 < q < 2$ and $r > 0$, there exists $C_{q,r} > 0$ depending only on q and r such that the following inclusions of spaces hold for all $R > 0$ when $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$:

$$\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R) \subset w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R) \subset \mathcal{B}_{q,r}(C_{q,r}\, R). \tag{5.13}$$

In particular, these inclusions shall turn out to be useful to check the optimality of the rates of convergence of the selected Lasso estimator in Section 5.3.

5.2 Upper bounds of the quadratic risk of the estimators

The rates of convergence of the deterministic Lassos f_p given in Lemma 5.4 can now be passed on to the Lasso estimators $\hat{f}_p$, $p \in \mathbb{N}^*$, and to the selected Lasso estimator $\hat{f}_{\hat{p}}$ thanks to the oracle inequalities (5.1) and (5.2) respectively.

Proposition 5.6. *Let $1 < q < 2$, $r > 0$ and $R > 0$. Assume that $f \in \mathcal{B}_{q,r}(R)$. Then, there exists $C_q > 0$ depending only on q such that, for all $p \in \mathbb{N}^*$,*

- *if $\left(\sqrt{\ln p} + 1\right)^{\frac{q-1}{q}} \leq R\varepsilon^{-1} \leq p^{\frac{2r}{q}} \left(\sqrt{\ln p} + 1\right)$, then*

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2\right] \leq C_q\, R^q \left(\varepsilon \left(\sqrt{\ln p} + 1\right)\right)^{2-q}, \tag{5.14}$$

- *if $R\varepsilon^{-1} > p^{\frac{2r}{q}} \left(\sqrt{\ln p} + 1\right)$, then*

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2\right] \leq C_q \left(R\, p^{-r}\right)^{\frac{2q}{2-q}} \left(\varepsilon \left(\sqrt{\ln p} + 1\right)\right)^{\frac{4(1-q)}{2-q}}, \tag{5.15}$$

- *if $R\varepsilon^{-1} < \left(\sqrt{\ln p} + 1\right)^{\frac{q-1}{q}}$, then*

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2\right] \leq C_q\, \varepsilon^2 \left(\sqrt{\ln p} + 1\right). \tag{5.16}$$

Proposition 5.7. *Let $1 < q < 2$ and $r > 0$. Assume that $f \in \mathcal{B}_{q,r}(R)$ with $R > 0$ such that $R\varepsilon^{-1} \geq \max\left(\mathrm{e}, (4r)^{-1} q\right)$.*
Then, there exists $C_{q,r} > 0$ depending only on q and r such that the quadratic risk of $\hat{f}_{\hat{p}}$ satisfies

$$\mathbb{E}\left[\|f - \hat{f}_{\hat{p}}\|^2\right] \leq C_{q,r}\, R^q \left(\varepsilon \sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}. \tag{5.17}$$

Remark 5.8.

1. Notice that the assumption $R\varepsilon^{-1} \geq \max\left(\mathrm{e}, (4r)^{-1}q\right)$ of Proposition 5.7 is not restrictive since it only means that we consider non-degenerate situations when the signal to noise ratio is large enough, which is the only interesting case to use the selected Lasso estimator. Indeed, if $R\varepsilon^{-1}$ is too small, then the estimator equal to zero will always be better than any other non-zero estimators, in particular Lasso estimators.

2. Proposition 5.7 highlights the fact that the selected Lasso estimator can simultaneously achieve rates of convergence of order $\left(\varepsilon\sqrt{\ln\left(\|f\|_{\mathcal{B}_{q,r}}\varepsilon^{-1}\right)}\right)^{2-q}$ for all classes $\mathcal{B}_{q,r}$ without knowing which class contains f. Besides, comparing the upper bound (5.17) to the lower bound (5.20) established in the next section for the minimax risk when the dictionary $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$ and $r < 1/q - 1/2$, we see that they can match up to a constant if the signal to noise ratio is large enough. This proves that the rate of convergence (5.17) achieved by $\hat{f}_{\hat{p}}$ is optimal.

3. Analyzing the different results of Proposition 5.6, we can notice that, unlike the selected Lasso estimator, the Lasso estimators are not adaptative. In particular, comparing (5.14) to (5.17), we see that the Lassos $\hat{f}_p$ are likely to achieve the optimal rate of convergence (5.17) only for p large enough, more precisely p such that $R\varepsilon^{-1} \leq p^{2r/q}(\sqrt{\ln p} + 1)$. For smaller values of p, truncating the dictionary at level p affects the rate of convergence as it is shown at (5.15). The problem is that q and r are unknown since they are the parameters characterizing the smoothness of the unknown target function. Therefore, when one chooses a level p of truncation of the dictionary, one does not know if $R\varepsilon^{-1} \leq p^{2r/q}(\sqrt{\ln p} + 1)$ and thus if the corresponding Lasso estimator $\hat{f}_p$ has a good rate of convergence. When working with the Lassos, the statistician is faced with a dilemma since one has to choose p large enough to get an optimal rate of convergence, but the larger p the less sparse and interpretable the model. The advantage of using the selected Lasso estimator rather than the Lassos is that, by construction of $\hat{f}_{\hat{p}}$, we are sure to get an estimator making the best tradeoff between approximation, ℓ_1-regularization and sparsity and achieving desirable rates of convergence for any target function belonging to some interpolation space $\mathcal{B}_{q,r}$.

4. Looking at the different results from (5.14) to (5.17), we can notice that the parameter q has much more influence on the rates of convergence than the parameter r since the rates are of order depending only on the parameter q while the dependence on r appears only in the multiplicative factor. Nonetheless, note that the smoother the target function with respect to the parameter r, the smaller the number of variables necessary to keep to get a good rate of convergence for the Lasso estimators. Indeed, on the one hand, it is easy to check that $\mathcal{B}_{q,r}(R) \subset \mathcal{B}_{q,r'}(R)$ for $r > r' > 0$ which means that the smoothness of f increases with r, while on the other hand, $p^{2r/q}(\sqrt{\ln p} + 1)$ increases with respect to r so that the larger r the smaller p satisfying the constraint necessary for the Lasso $\hat{f}_p$ to achieve the optimal rate of convergence (5.14).

5. Proposition 5.6 shows that the Lassos $\hat{f}_p$ perform as well as the greedy algorithms studied by Barron and al. in [1]. Indeed, in the case of the fixed design Gaussian regression model introduced in Example 2.2 with a sample of size n, we have $\varepsilon = \sigma/\sqrt{n}$ and (5.14) yields that the Lasso estimator $\hat{f}_p$ achieves a rate of convergence of order $R^q \left(n^{-1} \ln p\right)^{1-q/2}$ provided that $R\varepsilon^{-1}$ is well-chosen, which corresponds to the rate of convergence established by Barron and al. for the greedy algorithms. Similarly to our result, Barron and al. need to assume that the dictionary is large enough so as to ensure such rates of convergence. In fact, they consider truncated dictionaries of size p greater than $n^{1/(2r)}$ with $n^{1/q-1/2} \geq \|f\|_{\mathcal{B}_{q,r}}$. Under these assumptions, we recover the upper bound we impose on $R\varepsilon^{-1}$ to get the rate (5.14).

Remark 5.9. [Orthonormal case]

1. Notice that the rates of convergence provided for the Lasso estimators in Proposition 5.6 are a generalization to non-orthonormal dictionaries of the well-known performance bounds of soft-thresholding estimators in the orthonormal case. Indeed, when the dictionary $\mathcal{D} = \{\phi_j\}_j$ is an orthonormal basis of $\mathbb{H}$, if we set $\Theta_p := \{\theta = (\theta_j)_{j\in\mathbb{N}^*},\ \theta = (\theta_1, \dots, \theta_p, 0, \dots, 0, \dots)\}$ and calculate the subdifferential of the function $\theta \in \Theta_p \mapsto \gamma(\theta.\phi) + \lambda_p\|\theta\|_1$, where the function γ is defined by (2.4), we easily get that $\hat{f}_p = \hat{\theta}_p.\phi$ with $\hat{\theta}_p = (\widehat{\theta}_{p,1}, \dots, \widehat{\theta}_{p,p}, 0, \dots, 0, \dots)$ where for all $j = 1, \dots, p$,

$$\widehat{\theta}_{p,j} \;=\; \begin{cases} Y(\phi_j) - \lambda_p/2 & \text{if } Y(\phi_j) > \lambda_p/2 = 2\varepsilon\left(\sqrt{\ln p} + 1\right), \\ Y(\phi_j) + \lambda_p/2 & \text{if } Y(\phi_j) < -\lambda_p/2 = 2\varepsilon\left(\sqrt{\ln p} + 1\right), \\ 0 & \text{else,} \end{cases}$$

   where Y is defined by (2.1). Thus, the Lasso estimators $\hat{f}_p$ correspond to soft-thresholding estimators with thresholds of order $\varepsilon\sqrt{\ln p}$, and Proposition 5.6 together with the inclusions of spaces (5.13) enable to recover the well-known rates of convergence of order $(\varepsilon\sqrt{\ln p})^{2-q}$ for such thresholding estimators when the target function belongs to $w\mathcal{L}_q \cap \mathcal{B}^r_{2,\infty}$ (see for instance [7] for the establishment of such rates of convergence for estimators based on wavelet thresholding in the white noise framework).

2. Let us stress that, in the orthonormal case, since the Lasso estimators $\hat{f}_p$ correspond to soft-thresholding estimators with thresholds of order $\varepsilon\sqrt{\ln p}$, then the selected Lasso estimator $\hat{f}_{\hat{p}}$ can be viewed as a soft-thresholding estimator with adapted threshold $\varepsilon\sqrt{\ln \hat{p}}$.

5.3 Lower bounds in the orthonormal case

To complete our study on the rates of convergence, we propose to establish a lower bound of the minimax risk in the orthonormal case so as to prove that the selected Lasso estimator is simultaneously approximately minimax over spaces $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ in the orthonormal case for suitable signal to noise ratio $R\varepsilon^{-1}$.

Proposition 5.10. *Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$. Let $1 < q < 2$, $0 < r < 1/q - 1/2$ and $R > 0$ such that $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, u^2)$ where*

$$u := \frac{1}{r} - q\left(1 + \frac{1}{2r}\right) > 0. \tag{5.18}$$

Then, there exists an absolute constant $\kappa > 0$ such that the minimax risk over $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ satisfies

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)} \mathbb{E}\left[\|f - \tilde{f}\|^2\right] \geq \kappa\, u^{1-\frac{q}{2}}\, R^q \left(\varepsilon \sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}, \tag{5.19}$$

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

Remark 5.11.

1. Notice that the lower bound (5.19) depends much more on the parameter q than on the parameter r that only appears as a multiplicative factor through the term u. In fact, the assumption $f \in \mathcal{B}^r_{2,\infty}(R)$ is just added to the assumption $f \in \mathcal{L}_q(R)$ in order to control the size of the high-level components of f in the orthonormal basis $\mathcal{D}$ (see the proof of Lemma 8.5 to convince yourself), but this additional parameter of smoothness $r > 0$ can be taken arbitrarily small and has little effect on the minimax risk.

2. It turns out that the constraint $r < 1/q - 1/2$ of Proposition 5.10 is quite natural. Indeed, assume that $r > 1/q - 1/2$. Then, on the one hand it is easy to check that, for all $R > 0$, $\mathcal{B}^r_{2,\infty}(R') \subset \mathcal{L}_q(R)$ with $R' = (1 - 2^{ru})^{1/q} R$ where u is defined by (5.18), and thus $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R') = \mathcal{B}^r_{2,\infty}(R')$. On the other hand, noticing that $R' < R$, we have $\mathcal{B}^r_{2,\infty}(R') \subset \mathcal{B}^r_{2,\infty}(R)$ and thus $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R') \subset \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. Consequently, $\mathcal{B}^r_{2,\infty}(R') \subset \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R) \subset \mathcal{B}^r_{2,\infty}(R)$, and the intersection space $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ is no longer a real intersection between a strong-$\mathcal{L}_q$ space and a Besov space $\mathcal{B}^r_{2,\infty}$ but rather a Besov space $\mathcal{B}^r_{2,\infty}$ itself. In this case, the lower bound of the minimax risk is known to be of order $\varepsilon^{4r/(2r+1)}$ (see [11] for instance), which is no longer of the form (5.19).

Now, we can straightforwardly deduce from (5.13) and Proposition 5.10 the following result which proves that the rate of convergence (5.17) achieved by the selected Lasso estimator is optimal.

Proposition 5.12. *Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of $\mathbb{H}$. Let $1 < q < 2$, $0 < r < 1/q - 1/2$ and $R > 0$ such that $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, u^2)$ where*

$$u := \frac{1}{r} - q\left(1 + \frac{1}{2r}\right) > 0.$$

Then, there exists $C_{q,r} > 0$ depending only on q and r such that the minimax risk over $\mathcal{B}_{q,r}(R)$ satisfies

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{B}_{q,r}(R)} \mathbb{E}\left[\|f - \tilde{f}\|^2\right] \geq C_{q,r}\, R^q \left(\varepsilon \sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}, \tag{5.20}$$

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

Remark 5.13. Looking at (5.13), one could have obtained a result similar to (5.20) by bounding from below the minimax risk over $w\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)$ instead of $\mathcal{L}_q(R)\cap\mathcal{B}^r_{2,\infty}(R)$ as it is done in Proposition 5.10. We refer the interested reader to Theorem 1 in [17] for the establishment of such a result.

6 The Lasso for uncountable dictionaries : neural networks

In this section, we propose to provide some theoretical results on the performance of the Lasso when considering some particular infinite uncountable dictionaries such as those used for neural networks in the fixed design Gaussian regression models. Of course, there is no algorithm to approximate the Lasso solution for infinite dictionaries, so the following results are just to be seen as theoretical performance of the Lasso. We shall provide an ℓ_1-oracle type inequality satisfied by the Lasso and deduce rates of convergence of this estimator whenever the target function belongs to some interpolation space between $\mathcal{L}_1(\mathcal{D})$ and the Hilbert space $\mathbb{H}=\mathbb{R}^n$. These results will again prove that the Lasso theoretically performs as well as the greedy algorithms introduced in [1].

In the artificial intelligence field, the introduction of artificial neural networks have been motivated by the desire to model the human brain by a computer. They have been applied successfully to pattern recognition (radar systems, face identification...), sequence recognition (gesture, speech...), image analysis, adaptative control, and their study can enable the reconstruction of software agents (in computer, video games...) or autonomous robots for instance. Artificial neural networks receive a number of input signals and produce an output signal. They consist of multiple layers of weighted-sum units, called neurons, which are of the type

$$\phi_{a,b}:\ \mathbb{R}^d\mapsto\mathbb{R},\quad x\mapsto\chi\left(\langle a,x\rangle+b\right), \tag{6.1}$$

where $a\in\mathbb{R}^d$, $b\in\mathbb{R}$ and χ is the Heaviside function $\chi(x)=\mathbb{1}_{\{x>0\}}$ or more generally a sigmoid function. Here, we shall restrict to the case of χ being the Heaviside function. In other words, if we consider the infinite uncountable dictionary $\mathcal{D}=\{\phi_{a,b};\ a\in\mathbb{R}^d,b\in\mathbb{R}\}$, then a neural network is a real-valued function defined on $\mathbb{R}^d$ belonging to the linear span of $\mathcal{D}$.

Let us now consider the fixed design Gaussian regression model introduced in Example 2.2 with neural network regression function estimators. Given a training sequence $\{(x_1,Y_1),\dots,(x_n,Y_n)\}$, we assume that $Y_i=f(x_i)+\sigma\xi_i$ for all $i=1,\dots,n$ and we study the Lasso estimator over the set of neural network regression function estimators in $\mathcal{L}_1(\mathcal{D})$,

$$\hat{f}:=\hat{f}(\lambda)=\underset{h\in\mathcal{L}_1(\mathcal{D})}{\arg\min}\ \|Y-h\|^2+\lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}, \tag{6.2}$$

where $\lambda>0$ is a regularization parameter, $\mathcal{L}_1(\mathcal{D})$ is the linear span of $\mathcal{D}$ equipped with the ℓ_1-norm

$$\|h\|_{\mathcal{L}_1(\mathcal{D})}:=\inf\left\{\|\theta\|_1=\sum_{a\in\mathbb{R}^d,b\in\mathbb{R}}|\theta_{a,b}|,\quad h=\theta.\phi=\sum_{a\in\mathbb{R}^d,b\in\mathbb{R}}\theta_{a,b}\,\phi_{a,b}\right\} \tag{6.3}$$

and $\|Y-h\|^2:=\sum_{i=1}^n\left(Y_i-h(x_i)\right)^2/n$ is the empirical risk of h.

6.1 An ℓ_1-oracle type inequality

Despite the fact that the dictionary $\mathcal{D}$ for neural networks is infinite uncountable, we are able to establish an ℓ_1-oracle type inequality satisfied by the Lasso which is similar to the one provided in Theorem 3.2 in the case of a finite dictionary. This is due to the very particular structure of the dictionary $\mathcal{D}$ which is only composed of functions derived from the Heaviside function. This property enables us to achieve theoretical results without truncating the whole dictionary into finite subdictionaries contrary to the study developed in Section 4 where we considered arbitrary infinite countable dictionaries (see Remark 8.3 for more details). The following ℓ_1-oracle type inequality is once again a direct application of the general model selection Theorem 7.1 already used to prove both Theorem 3.2 and Theorem 4.2.

Theorem 6.1. *Assume that*

$$\lambda \geq \frac{28\,\sigma}{\sqrt{n}}\left(\sqrt{\ln\left((n+1)^{d+1}\right)}+4\right).$$

Consider the corresponding Lasso estimator $\hat{f}$ defined by (6.2).
Then, there exists an absolute constant $C>0$ such that

$$\mathbb{E}\left[\|f-\hat{f}\|^2+\lambda\|\hat{f}\|_{\mathcal{L}_1(\mathcal{D})}\right] \leq C\left[\inf_{h\in\mathcal{L}_1(\mathcal{D})}\left(\|f-h\|^2+\lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right)+\lambda\frac{\sigma}{\sqrt{n}}\right].$$

6.2 Rates of convergence in real interpolation spaces

We can now deduce theoretical rates of convergence for the Lasso from Theorem 6.1. Since we do not truncate the dictionary $\mathcal{D}$, we shall not consider the spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{q,r}$ that we introduced in the last section because they were adapted to the truncation of the dictionary. Here, we can work with the whole space $\mathcal{L}_1(\mathcal{D})$ instead of $\mathcal{L}_{1,r}$ and the spaces $\mathcal{B}_{q,r}$ will be replaced by bigger spaces $\mathcal{B}_q$ that are the real interpolation spaces between $\mathcal{L}_1(\mathcal{D})$ and $\mathbb{H}=\mathbb{R}^n$.

Definition 6.2. **[Space $\mathcal{B}_q$]** *Let $1<q<2$, $\alpha=1/q-1/2$ and $R>0$. We say that a function g belongs to $\mathcal{B}_q(R)$ if, for all $\delta>0$,*

$$\inf_{h\in\mathcal{L}_1(\mathcal{D})}\left(\|g-h\|+\delta\|h\|_{\mathcal{L}_1(\mathcal{D})}\right) \leq R\,\delta^{2\alpha}. \tag{6.4}$$

We say that $g\in\mathcal{B}_q$ if there exists $R>0$ such that $g\in\mathcal{B}_q(R)$. In this case, the smallest R such that $g\in\mathcal{B}_q(R)$ defines a norm on the space $\mathcal{B}_q$ and is denoted by $\|g\|_{\mathcal{B}_q}$.

The following proposition shows that the Lasso simultaneously achieves desirable levels of performance on all classes $\mathcal{B}_q$ without knowing which class contains f.

Proposition 6.3. *Let $1<q<2$.*
Assume that $f\in\mathcal{B}_q(R)$ with $R\geq\sigma\left[\ln\left((n+1)^{d+1}\right)\right]^{\frac{q-1}{2q}}/\sqrt{n}$.
Consider the Lasso estimator $\hat{f}$ defined by (6.2) with

$$\lambda=\frac{28\,\sigma}{\sqrt{n}}\left(\sqrt{\ln\left((n+1)^{d+1}\right)}+4\right).$$

Then, there exists $C_q > 0$ depending only on q such that the quadratic risk of $\hat{f}$ satisfies

$$\mathbb{E}\left[\|f - \hat{f}\|^2\right] \leq C_q \, R^q \left[\frac{\ln\left((n+1)^{d+1}\right)}{n}\right]^{1-\frac{q}{2}}.$$

Remark 6.4. Notice that the above rates of convergence are of the same order as those provided for the Lasso in Proposition 5.6 for a suitable signal to noise ratio in the case of an infinite countable dictionary with $\varepsilon = \sigma/\sqrt{n}$. Besides, we recover the same rates of convergence as those obtained by Barron and al. in [1] for the greedy algorithms when considering neural networks. Notice that our results can be seen as the analog in the Gaussian framework of their results which are valid under the assumption that the output variable Y is bounded but not necessarily Gaussian.

7 A model selection theorem

Let us end this paper by describing the main idea that has enabled us to establish all the oracle inequalities of Theorem 3.2, Theorem 4.2 and Theorem 6.1 as an application of a single general model selection theorem, and by presenting this general theorem.

We keep the notations introduced in Section 2. In particular, recall that one observes a process $(Y(h))_{h\in\mathbb{H}}$ defined by $Y(h) = \langle f, h\rangle + \varepsilon W(h)$ for all $h \in \mathbb{H}$, where $\varepsilon > 0$ is a fixed parameter and W is an isonormal process, and that we define $\gamma(h) := -2Y(h) + \|h\|^2$.

The basic idea is to view the Lasso estimator as the solution of a penalized least squares model selection procedure over a properly defined countable collection of models with ℓ_1-penalty. The key observation that enables one to make this connection is the simple fact that $\mathcal{L}_1(\mathcal{D}) = \bigcup_{R>0}\{h \in \mathcal{L}_1(\mathcal{D}),\ \|h\|_{\mathcal{L}_1(\mathcal{D})} \leq R\}$, so that for any finite or infinite given dictionary $\mathcal{D}$, the Lasso $\hat{f}$ defined by

$$\hat{f} = \operatorname*{arg\,min}_{h\in\mathcal{L}_1(\mathcal{D})}\left(\gamma(h) + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right)$$

satisfies

$$\gamma(\hat{f}) + \lambda\|\hat{f}\|_{\mathcal{L}_1(\mathcal{D})} = \inf_{h\in\mathcal{L}_1(\mathcal{D})}\gamma(h) + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})} = \inf_{R>0}\left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D})}\leq R}\gamma(h) + \lambda R\right).$$

Then, to obtain a countable collection of models, we just discretize the family of ℓ_1-balls $\{h \in \mathcal{L}_1(\mathcal{D}),\ \|h\|_{\mathcal{L}_1(\mathcal{D})} \leq R\}$ by setting for any integer $m \geq 1$,

$$S_m = \left\{h \in \mathcal{L}_1(\mathcal{D}),\ \|h\|_{\mathcal{L}_1(\mathcal{D})} \leq m\varepsilon\right\},$$

and define $\hat{m}$ as the smallest integer such that $\hat{f}$ belongs to $S_{\hat{m}}$, i.e.

$$\hat{m} = \left\lceil \frac{\|\hat{f}\|_{\mathcal{L}_1(\mathcal{D})}}{\varepsilon} \right\rceil. \tag{7.1}$$

It is now easy to derive from the definitions of $\hat{m}$ and $\hat{f}$ and from the fact that $\mathcal{L}_1(\mathcal{D}) = \bigcup_{m\geq 1} S_m$ that

$$\begin{aligned}\gamma(\hat{f}) + \lambda\hat{m}\varepsilon &\leq \gamma(\hat{f}) + \lambda\left(\|\hat{f}\|_{\mathcal{L}_1(\mathcal{D})} + \varepsilon\right)\\ &= \inf_{h\in\mathcal{L}_1(\mathcal{D})}\left(\gamma(h) + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right) + \lambda\varepsilon\\ &= \inf_{m\geq 1}\left(\inf_{h\in S_m}\left(\gamma(h) + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right)\right) + \lambda\varepsilon\\ &\leq \inf_{m\geq 1}\left(\inf_{h\in S_m}\gamma(h) + \lambda m\varepsilon\right) + \lambda\varepsilon,\end{aligned}$$

that is to say

$$\gamma(\hat{f}) + \mathrm{pen}(\hat{m}) \leq \inf_{m\geq 1}\left(\inf_{h\in S_m}\gamma(h) + \mathrm{pen}(m)\right) + \rho \tag{7.2}$$

with $\mathrm{pen}(m) = \lambda m\varepsilon$ and $\rho = \lambda\varepsilon$. This means that $\hat{f}$ is equivalent to a ρ-approximate penalized least squares estimator over the sequence of models given by the collection of ℓ_1-balls $\{S_m,\ m \geq 1\}$. This property will enable us to derive ℓ_1-oracle type inequalities by applying a general model selection theorem that guarantees such inequalities provided that the penalty $\mathrm{pen}(m)$ is large enough. This general theorem, stated below as Theorem 7.1, is borrowed from [5] and is a restricted version of an even more general model selection theorem that the interested reader can find in [15], Theorem 4.18. For the sake of completeness, the proof of Theorem 7.1 is recalled in Section 8.

Theorem 7.1. *Let $\{S_m\}_{m\in\mathcal{M}}$ be a countable collection of convex and compact subsets of a Hilbert space $\mathbb{H}$. Define, for any $m \in \mathcal{M}$,*

$$\Delta_m := \mathbb{E}\left[\sup_{h\in S_m} W(h)\right], \tag{7.3}$$

and consider weights $\{x_m\}_{m\in\mathcal{M}}$ such that

$$\Sigma := \sum_{m\in\mathcal{M}} e^{-x_m} < \infty.$$

Let $K > 1$ and assume that, for any $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq 2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m\varepsilon x_m}\right). \tag{7.4}$$

Given non negative ρ_m, $m \in \mathcal{M}$, define a ρ_m-approximate penalized least squares estimator as any $\hat{f} \in S_{\hat{m}}$, $\hat{m} \in \mathcal{M}$, such that

$$\gamma(\hat{f}) + \mathrm{pen}(\hat{m}) \leq \inf_{m\in\mathcal{M}}\left(\inf_{h\in S_m}\gamma(h) + \mathrm{pen}(m) + \rho_m\right).$$

Then, there is a positive constant $C(K)$ such that for all $f \in \mathbb{H}$ and $z > 0$, with probability larger than $1 - \Sigma\,\mathrm{e}^{-z}$,

$$\begin{aligned}&\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m})\\ &\leq C(K)\left[\inf_{m\in\mathcal{M}}\left(\inf_{h\in S_m}\|f - h\|^2 + \mathrm{pen}(m) + \rho_m\right) + (1+z)\varepsilon^2\right].\end{aligned} \tag{7.5}$$

Integrating this inequality with respect to z leads to the following risk bound

$$\mathbb{E}\left[\|f-\hat{f}\|^2+\mathrm{pen}(\hat{m})\right] \leq C(K)\left[\inf_{m\in\mathcal{M}}\left(\inf_{h\in S_m}\|f-h\|^2+\mathrm{pen}(m)+\rho_m\right)+(1+\Sigma)\varepsilon^2\right]. \quad (7.6)$$

8 Proofs

8.1 Oracle inequalities

We first prove the general model selection Theorem 7.1. Its proof is based on the concentration inequality for the suprema of Gaussian processes established in [5]. Then, deriving Theorem 3.2, Theorem 4.2 and Theorem 6.1 from Theorem 7.1 is an exercise. Indeed, using the key observation that the Lasso and the selected Lasso estimators are approximate penalized least squares estimators over a collection of ℓ_1-balls with a convenient penalty, it only remains to determine a lower bound on this penalty to guarantee condition (7.4) and then to apply the conclusion of Theorem 7.1.

8.1.1 Proof of Theorem 7.1

Let $m\in\mathcal{M}$. Since S_m is assumed to be a convex and compact subset, we can consider f_m the projection of f onto S_m, that is the unique element of S_m such that $\|f-f_m\|=\inf_{h\in S_m}\|f-h\|$. By definition of $\hat{f}$, we have

$$\gamma(\hat{f})+\mathrm{pen}(\hat{m})\leq\gamma(f_m)+\mathrm{pen}(m)+\rho_m.$$

Since $\|f\|^2+\gamma(h)=\|f-h\|^2-2\varepsilon W(h)$, this implies that

$$\|f-\hat{f}\|^2+\mathrm{pen}(\hat{m})\leq\|f-f_m\|^2+2\varepsilon\left(W(\hat{f})-W(f_m)\right)+\mathrm{pen}(m)+\rho_m. \quad (8.1)$$

For all $m'\in\mathcal{M}$, let $y_{m'}$ be a positive number whose value will be specified below and define for every $h\in S_{m'}$

$$2w_{m'}(h)=\left(\|f-f_m\|+\|f-h\|\right)^2+y_{m'}^2. \quad (8.2)$$

Finally, set

$$V_{m'}=\sup_{h\in S_{m'}}\left(\frac{W(h)-W(f_m)}{w_{m'}(h)}\right).$$

Taking these definitions into account, we get from (8.1) that

$$\|f-\hat{f}\|^2+\mathrm{pen}(\hat{m})\leq\|f-f_m\|^2+2\varepsilon w_{\hat{m}}(\hat{f})V_{\hat{m}}+\mathrm{pen}(m)+\rho_m. \quad (8.3)$$

The essence of the proof is the control of the random variables $V_{m'}$ for all possible values of m'. To this end, we may use the concentration inequality for the suprema of Gaussian processes (see [5]) which ensures that, given $z>0$, for all $m'\in\mathcal{M}$,

$$\mathbb{P}\left[V_{m'}\geq\mathbb{E}\left[V_{m'}\right]+\sqrt{2v_{m'}(x_{m'}+z)}\right]\leq\mathrm{e}^{-(x_{m'}+z)}, \quad (8.4)$$

where

$$v_{m'} = \sup_{h \in S_{m'}} \mathrm{Var}\left[\frac{W(h) - W(f_m)}{w_{m'}(h)}\right] = \sup_{h \in S_{m'}} \frac{\|h - f_m\|^2}{w_{m'}^2(h)} \ .$$

From (8.2), $w_{m'}(h) \geq (\|f - f_m\| + \|f - h\|)\, y_{m'} \geq \|h - f_m\| y_{m'}$, so $v_{m'} \leq y_{m'}^{-2}$ and summing the inequalities (8.4) over $m' \in \mathcal{M}$, we get that for every $z > 0$ there is an event Ω_z with $\mathbb{P}(\Omega_z) > 1 - \Sigma \mathrm{e}^{-z}$ such that on Ω_z, for all $m' \in \mathcal{M}$,

$$V_{m'} \leq \mathbb{E}\left[V_{m'}\right] + y_{m'}^{-1}\sqrt{2(x_{m'} + z)}. \tag{8.5}$$

Let us now bound $\mathbb{E}\left[V_{m'}\right]$. We may write

$$\mathbb{E}\left[V_{m'}\right] \leq \mathbb{E}\left[\frac{\sup_{h \in S_{m'}}\left(W(h) - W(f_{m'})\right)}{\inf_{h \in S_{m'}} w_{m'}\left(h\right)}\right] + \mathbb{E}\left[\frac{\left(W(f_{m'}) - W(f_m)\right)_+}{\inf_{h \in S_{m'}} w_{m'}(h)}\right] \ . \tag{8.6}$$

But from the definition of $f_{m'}$, we have for all $h \in S_{m'}$

$$\begin{aligned} 2w_{m'}(h) &\geq \left(\|f - f_m\| + \|f - f_{m'}\|\right)^2 + y_{m'}^2 \\ &\geq \|f_{m'} - f_m\|^2 + y_{m'}^2 \\ &\geq \left(y_{m'}^2 \vee 2y_{m'}\|f_{m'} - f_m\|\right). \end{aligned}$$

Hence, on the one hand via (7.3) and recalling that W is centered, we get

$$\begin{aligned} \mathbb{E}\left[\frac{\sup_{h \in S_{m'}}\left(W(h) - W(f_{m'})\right)}{\inf_{h \in S_{m'}} w_{m'}(h)}\right] &\leq 2\, y_{m'}^{-2}\, \mathbb{E}\left[\sup_{h \in S_{m'}}\left(W(h) - W(f_{m'})\right)\right] \\ &= 2\, y_{m'}^{-2}\, \Delta_{m'}, \end{aligned}$$

and on the other hand, using the fact that $\left(W(f_{m'}) - W(f_m)\right)/\|f_m - f_{m'}\|$ is a standard normal variable, we get

$$\mathbb{E}\left[\frac{\left(W(f_{m'}) - W\ (f_m)\right)_+}{\inf_{h \in S_{m'}} w_{m'}(h)}\right] \leq y_{m'}^{-1}\, \mathbb{E}\left[\frac{W(f_{m'}) - W(f_m)}{\|f_m - f_{m'}\|}\right]_+ \leq y_{m'}^{-1}\,(2\pi)^{-1/2}\,.$$

Collecting these inequalities, we get from (8.6) that for all $m' \in \mathcal{M}$,

$$\mathbb{E}\left[V_{m'}\right] \leq 2\Delta_{m'} y_{m'}^{-2} + (2\pi)^{-1/2}\, y_{m'}^{-1} \ .$$

Hence, setting $\delta = \left((4\pi)^{-1/2} + \sqrt{z}\right)^2$, (8.5) implies that on the event Ω_z, for all $m' \in \mathcal{M}$,

$$\begin{aligned} V_{m'} &\leq y_{m'}^{-1}\left[2\Delta_{m'} y_{m'}^{-1} + \sqrt{2x_{m'}} + (2\pi)^{-1/2} + \sqrt{2z}\right] \\ &= y_{m'}^{-1}\left[2\Delta_{m'} y_{m'}^{-1} + \sqrt{2x_{m'}} + \sqrt{2\delta}\right]. \end{aligned} \tag{8.7}$$

Given $K' \in \left(1, \sqrt{K}\right]$ to be chosen later, we now define

$$y_{m'}^2 = 2K'^2\varepsilon^2\left[\left(\sqrt{x_{m'}} + \sqrt{\delta}\right)^2 + K'^{-1}\varepsilon^{-1}\Delta_{m'} + \sqrt{K'^{-1}\varepsilon^{-1}\Delta_{m'}}\left(\sqrt{x_{m'}} + \sqrt{\delta}\right)\right].$$

With this choice of $y_{m'}$, it is not hard to check that (8.7) warrants that on the event Ω_z, $\varepsilon V_{m'} \leq {K'}^{-1}$ for all $m' \in \mathcal{M}$, which in particular implies that $\varepsilon V_{\hat{m}} \leq {K'}^{-1}$, and we get from (8.3) and (8.2) that

$$\begin{aligned}
&\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) \\
&\leq \|f - f_m\|^2 + 2{K'}^{-1} w_{\hat{m}}(\hat{f}) + \mathrm{pen}(m) + \rho_m \\
&= \|f - f_m\|^2 + {K'}^{-1} \left[\left(\|f - f_m\| + \|f - \hat{f}\| \right)^2 + y_{\hat{m}}^2 \right] + \mathrm{pen}(m) + \rho_m. \quad (8.8)
\end{aligned}$$

Moreover, using repeatedly the elementary inequalities $(a+b)^2 \leq (1+\theta)a^2 + (1+\theta^{-1})b^2$ or equivalently $2ab \leq \theta a^2 + \theta^{-1} b^2$ for various values of $\theta > 0$, we derive that on the one hand

$$\left(\|f - f_m\| + \|f - \hat{f}\| \right)^2 \leq \sqrt{K'} \left(\|f - \hat{f}\|^2 + \frac{\|f - f_m\|^2}{\sqrt{K'} - 1} \right),$$

and on the other hand

$$K'^{-1} y_{\hat{m}}^2 \leq 2K'^2 \varepsilon^2 \left[\varepsilon^{-1} \Delta_{\hat{m}} + x_{\hat{m}} + \sqrt{\varepsilon^{-1} \Delta_{\hat{m}} x_{\hat{m}}} + B(K') \left(\frac{1}{2\pi} + 2z \right) \right],$$

where $B(K') = (K'-1)^{-1} + \left(4K'(K'^2 - 1) \right)^{-1}$.

Hence, setting $A(K') = 1 + K'^{-1/2} \left(\sqrt{K'} - 1 \right)^{-1}$, we deduce from (8.8) that on the event Ω_z,

$$\begin{aligned}
&\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) \\
&\leq A(K') \|f - f_m\|^2 + K'^{-1/2} \|f - \hat{f}\|^2 + 2K'^2 \varepsilon \left[\Delta_{\hat{m}} + \varepsilon x_{\hat{m}} + \sqrt{\varepsilon \Delta_{\hat{m}} x_{\hat{m}}} \right] \\
&\quad + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 K'^2 B(K') \left(\frac{1}{2\pi} + 2z \right),
\end{aligned}$$

or equivalently

$$\begin{aligned}
&\left(1 - K'^{-1/2} \right) \|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) - 2K'^2 \varepsilon \left[\Delta_{\hat{m}} + \varepsilon x_{\hat{m}} + \sqrt{\varepsilon \Delta_{\hat{m}} x_{\hat{m}}} \right] \\
&\leq A(K') \|f - f_m\|^2 + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 B(K') \left(\frac{1}{2\pi} + 2z \right).
\end{aligned}$$

Because of condition (7.4) on the penalty function, this implies that

$$\begin{aligned}
&\left(1 - K'^{-1/2} \right) \|f - \hat{f}\|^2 + \left(1 - K'^2 K^{-1} \right) \mathrm{pen}(\hat{m}) \\
&\leq A(K') \|f - f_m\|^2 + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 B(K') \left(\frac{1}{2\pi} + 2z \right).
\end{aligned}$$

Now choosing $K' = K^{2/5}$, we get that

$$\begin{aligned}
&\left(1 - K^{-1/5} \right) \left(\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) \right) \\
&\leq A(K^{2/5}) \|f - f_m\|^2 + \mathrm{pen}(m) + \rho_m + 2\varepsilon^2 B(K^{2/5}) \left(\frac{1}{2\pi} + 2z \right).
\end{aligned}$$

So, there exists a positive constant $C := C(K)$ depending only on K such that for all $z > 0$, on the event Ω_z,

$$\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) \leq C \left(\inf_{m \in \mathcal{M}} \left(\|f - f_m\|^2 + \mathrm{pen}(m) + \rho_m \right) + \varepsilon^2 (1 + z) \right),$$

which proves (7.5). Integrating this inequality with respect to z straightforwardly leads to the risk bound (7.6). □

8.1.2 Proof of Theorem 3.2

Fix $p \in \mathbb{N}^*$. Let $\mathcal{M} = \mathbb{N}^*$ and consider the collection of ℓ_1-balls for $m \in \mathcal{M}$,

$$S_m = \left\{ h \in \mathcal{L}_1(\mathcal{D}_p),\ \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon \right\}.$$

We have noticed at (7.2) that the Lasso estimator $\hat{f}_p$ is a ρ-approximate penalized least squares estimator over the sequence $\{S_m,\ m \geq 1\}$ for $\mathrm{pen}(m) = \lambda_p m\varepsilon$ and $\rho = \lambda_p \varepsilon$. So, it only remains to determine a lower bound on λ_p that guarantees that $\mathrm{pen}(m)$ satisfies condition (7.4).

Let $h \in S_m$ and consider $\theta = (\theta_1, \ldots, \theta_p)$ such that $h = \theta.\phi = \sum_{j=1}^p \theta_j \phi_j$ and $\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} = \|\theta\|_1$. The linearity of W implies that

$$W(h) = \sum_{j=1}^p \theta_j W(\phi_j) \leq \sum_{j=1}^p |\theta_j| \, |W(\phi_j)| \leq m\varepsilon \max_{j=1,\ldots,p} |W(\phi_j)|. \tag{8.9}$$

From Definition 2.1, $\mathrm{Var}[W(\phi_j)] = \mathbb{E}\left[W^2(\phi_j)\right] = \|\phi_j\|^2 \leq 1$ for all $j = 1, \ldots, p$. So, the variables $W(\phi_j)$ and $(-W(\phi_j))$, $j = 1, \ldots, p$, are $2p$ centered normal variables with variance less than 1 and thus (see Lemma 2.3 in [15] for instance),

$$\mathbb{E}\left[\max_{j=1,\ldots,p} |W(\phi_j)| \right] = \mathbb{E}\left[\left(\max_{j=1,\ldots,p} W(\phi_j) \right) \vee \left(\max_{j=1,\ldots,p} (-W(\phi_j)) \right) \right] \leq \sqrt{2 \ln(2p)}.$$

Therefore, we deduce from (8.9) that

$$\Delta_m := \mathbb{E}\left[\sup_{h \in S_m} W(h) \right] \leq m\varepsilon \sqrt{2 \ln(2p)} \leq \sqrt{2} m\varepsilon \left(\sqrt{\ln p} + \sqrt{\ln 2} \right). \tag{8.10}$$

Now, choose the weights of the form $x_m = \gamma m$ where $\gamma > 0$ is specified below. Then, $\sum_{m \geq 1} e^{-x_m} = 1/(e^\gamma - 1) := \Sigma_\gamma < +\infty$.

Defining $K = 4\sqrt{2}/5 > 1$ and $\gamma = (1 - \sqrt{\ln 2})/K$, and using the inequality $2\sqrt{ab} \leq \eta a + \eta^{-1} b$ with $\eta = 1/2$, we get that

$$\begin{aligned}
2K\varepsilon \left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m} \right) &\leq K\varepsilon \left(\frac{5}{2} \Delta_m + 4 x_m \varepsilon \right) \\
&\leq 4m\varepsilon^2 \left(\sqrt{\ln p} + \sqrt{\ln 2} + K\gamma \right) \\
&\leq 4m\varepsilon^2 \left(\sqrt{\ln p} + 1 \right) \\
&\leq \lambda_p m\varepsilon
\end{aligned}$$

as soon as

$$\lambda_p \geq 4\varepsilon \left(\sqrt{\ln p} + 1\right). \tag{8.11}$$

For such values of λ_p, condition (7.4) on the penalty function is satisfied and we may apply Theorem 7.1. Taking into account the definition of $\hat{m}$ at (7.1) and noticing that $\varepsilon^2 \leq \lambda_p \varepsilon/4$ for λ_p satisfying (8.11), we get from (7.5) that there exists some $C > 0$ such that for all $z > 0$, with probability larger than $1 - \Sigma_\gamma \, \mathrm{e}^{-z} \geq 1 - 3.4\, \mathrm{e}^{-z}$,

$$\begin{aligned}
&\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \\
&\leq C \left[\inf_{m \geq 1} \left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon} \|f - h\|^2 + \lambda_p m \varepsilon \right) + \lambda_p \varepsilon + (1+z)\varepsilon^2 \right] \\
&\leq C \left[\inf_{m \geq 1} \left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon} \|f - h\|^2 + \lambda_p m \varepsilon \right) + \lambda_p \varepsilon (1+z) \right],
\end{aligned} \tag{8.12}$$

while the risk bound (7.6) leads to

$$\begin{aligned}
&\mathbb{E}\left[\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\right] \\
&\leq C \left[\inf_{m \geq 1} \left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon} \|f - h\|^2 + \lambda_p m \varepsilon \right) + \lambda_p \varepsilon + (1+\Sigma_\gamma)\varepsilon^2 \right] \\
&\leq C \left[\inf_{m \geq 1} \left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon} \|f - h\|^2 + \lambda_p m \varepsilon \right) + \lambda_p \varepsilon \right].
\end{aligned} \tag{8.13}$$

Finally, to get the desired bounds (3.5) and (3.6), just notice that for all $R > 0$, by considering $m_R = \lceil R/\varepsilon \rceil \in \mathbb{N}^*$, we have for all $g \in \mathcal{L}_1(\mathcal{D}_p)$ such that $\|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq R$,

$$\begin{aligned}
\inf_{m \geq 1} \left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon} \|f - h\|^2 + \lambda_p m \varepsilon \right) &\leq \|f - g\|^2 + \lambda_p \, m_R \, \varepsilon \\
&\leq \|f - g\|^2 + \lambda_p R + \lambda_p \varepsilon,
\end{aligned}$$

so that

$$\begin{aligned}
&\inf_{m \geq 1} \left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon} \|f - h\|^2 + \lambda_p m \varepsilon \right) \\
&\leq \inf_{R > 0} \left(\inf_{\|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq R} \|f - g\|^2 + \lambda_p R \right) + \lambda_p \varepsilon \\
&= \inf_{g \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - g\|^2 + \lambda_p \|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \lambda_p \varepsilon,
\end{aligned} \tag{8.14}$$

and combining (8.14) with (8.12) and (8.13) leads to

$$\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C \left[\inf_{g \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - g\|^2 + \lambda_p \|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \lambda_p \varepsilon (1+z) \right]$$

and

$$\mathbb{E}\left[\|f - \hat{f}_p\|^2 + \lambda_p \|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\right] \leq C \left[\inf_{g \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f - g\|^2 + \lambda_p \|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \lambda_p \varepsilon \right],$$

where $C > 0$ is some absolute constant. □

8.1.3 Proof of Theorem 4.2

Let $\mathcal{M} = \mathbb{N}^* \times \Lambda$ and consider the set of ℓ_1-balls for all $(m,p) \in \mathcal{M}$,

$$S_{m,p} = \left\{ h \in \mathcal{L}_1(\mathcal{D}_p),\ \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq m\varepsilon \right\}.$$

Define $\hat{m}$ as the smallest integer such that $\hat{f}_{\hat{p}}$ belongs to $S_{\hat{m},\hat{p}}$, i.e.

$$\hat{m} = \left\lceil \frac{\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})}}{\varepsilon} \right\rceil. \tag{8.15}$$

Let $\alpha > 0$ be a constant to be chosen later. From the definitions of $\hat{m}$, $\lambda_{\hat{p}}$ and $\mathrm{pen}(\hat{p})$, and using the fact that for all $p \in \Lambda$, $\sqrt{\ln p} \leq (\ln p)/\sqrt{\ln 2}$, we have

$$\begin{aligned}
\gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) &\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \lambda_{\hat{p}}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + 4\varepsilon^2\sqrt{\ln \hat{p}} + 4\varepsilon^2 + 5\alpha\varepsilon^2 \ln \hat{p} \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \left(\frac{4}{5\sqrt{\ln 2}} + \alpha \right) 5\varepsilon^2 \ln \hat{p} + 4\varepsilon^2 \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \left(\frac{4}{5\sqrt{\ln 2}} + \alpha \right) \mathrm{pen}(\hat{p}) + 4\varepsilon^2.
\end{aligned}$$

Now, if we choose $\alpha = 1 - 4/(5\sqrt{\ln 2}) \in]0,1[$, we get from the definition of $\hat{f}_{\hat{p}}$ and the fact that $\mathcal{L}_1(\mathcal{D}_p) = \bigcup_{m \in \mathbb{N}^*} S_{m,p}$, that

$$\begin{aligned}
\gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) &\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p}) + 4\varepsilon^2 \\
&\leq \inf_{p \in \Lambda} \left[\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\gamma(h) + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \mathrm{pen}(p) \right] + 4\varepsilon^2 \\
&\leq \inf_{p \in \Lambda} \left[\inf_{m \in \mathbb{N}^*} \left(\inf_{h \in S_{m,p}} \gamma(h) + \lambda_p m\varepsilon \right) + \mathrm{pen}(p) \right] + 4\varepsilon^2 \\
&\leq \inf_{(m,p) \in \mathcal{M}} \left[\inf_{h \in S_{m,p}} \gamma(h) + \lambda_p m\varepsilon + \mathrm{pen}(p) \right] + 4\varepsilon^2,
\end{aligned}$$

that is to say

$$\gamma(\hat{f}_{\hat{p}}) + \mathrm{pen}(\hat{m},\hat{p}) \leq \inf_{(m,p) \in \mathcal{M}} \left[\inf_{h \in S_{m,p}} \gamma(h) + \mathrm{pen}(m,p) + \rho_p \right],$$

with $\mathrm{pen}(m,p) := \lambda_p m\varepsilon + \alpha\,\mathrm{pen}(p)$ and $\rho_p := (1-\alpha)\,\mathrm{pen}(p) + 4\varepsilon^2$. This means that $\hat{f}_{\hat{p}}$ is equivalent to a ρ_p-approximate penalized least squares estimator over the sequence of models $\{S_{m,p},\ (m,p) \in \mathcal{M}\}$. By applying Theorem 7.1, this property will enable us to derive a performance bound satisfied by $\hat{f}_{\hat{p}}$ provided that $\mathrm{pen}(m,p)$ is large enough. So, it remains to choose weights $x_{m,p}$ so that condition (7.4) on the penalty function is satisfied with $\mathrm{pen}(m,p) = \lambda_p m\varepsilon + \alpha\,\mathrm{pen}(p)$.

Let us choose the weights of the form $x_{m,p} = \gamma m + \beta \ln p$ where $\gamma > 0$ and $\beta > 0$ are numerical constants specified later. Then,

$$\begin{aligned}\Sigma_{\gamma,\beta} := \sum_{(m,p)\in\mathcal{M}} e^{-x_{m,p}} &= \left(\sum_{m\in\mathbb{N}^*} e^{-\gamma m}\right)\left(\sum_{p\in\Lambda} e^{-\beta \ln p}\right) \\ &= \left(\sum_{m\in\mathbb{N}^*} e^{-\gamma m}\right)\left(\sum_{J\in\mathbb{N}} e^{-\beta \ln 2^J}\right) \\ &= \frac{1}{(e^\gamma - 1)\,(1 - 2^{-\beta})} < +\infty.\end{aligned}$$

Moreover, for all $(m,p) \in \mathcal{M}$, we can prove similarly as (8.10) that

$$\Delta_{m,p} := \mathbb{E}\left[\sup_{h\in S_{m,p}} W(h)\right] \leq \sqrt{2} m\varepsilon \left(\sqrt{\ln p} + \sqrt{\ln 2}\right).$$

Now, defining $K = 4\sqrt{2}/5 > 1$, $\gamma = (1 - \sqrt{\ln 2})/K > 0$ and $\beta = (5\alpha)/(4K) > 0$, and using the inequality $2\sqrt{ab} \leq \eta a + \eta^{-1} b$ with $\eta = 1/2$, we have

$$\begin{aligned}&2K\varepsilon\left(\Delta_{m,p} + \varepsilon x_{m,p} + \sqrt{\Delta_{m,p}\varepsilon x_{m,p}}\right) \\ &\leq K\varepsilon\left(\frac{5}{2}\,\Delta_{m,p} + 4x_{m,p}\varepsilon\right) \\ &\leq 4\varepsilon^2\left(m\sqrt{\ln p} + m\sqrt{\ln 2} + K\gamma m + K\beta \ln p\right) \\ &\leq 4\varepsilon^2\left(m\left(\sqrt{\ln p} + 1\right) + K\beta \ln p\right) \\ &\leq 4\varepsilon^2\left(m\left(\sqrt{\ln p} + 1\right) + \frac{5\alpha}{4}\ln p\right) \\ &\leq \lambda_p m\varepsilon + \alpha\,\mathrm{pen}(p).\end{aligned}$$

Thus, condition (7.4) is satisfied and we can apply Theorem 7.1 with $\mathrm{pen}(m,p) = \lambda_p m\varepsilon + \alpha\,\mathrm{pen}(p)$ and $\rho_p = (1-\alpha)\,\mathrm{pen}(p) + 4\varepsilon^2$, which leads to the following risk bound:

$$\begin{aligned}&\mathbb{E}\left[\|f - \hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p})\right] \\ &\leq C\left[\inf_{(m,p)\in\mathcal{M}}\left(\inf_{h\in S_{m,p}}\|f-h\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p)\right) + (5 + \Sigma_{\gamma,\beta})\,\varepsilon^2\right] \\ &\leq C\left[\inf_{(m,p)\in\mathcal{M}}\left(\inf_{h\in S_{m,p}}\|f-h\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p)\right) + \varepsilon^2\right], \qquad (8.16)\end{aligned}$$

where $C > 0$ denotes some numerical constant. The infimum of this risk bound can easily be extended to $\inf_{p\in\Lambda}\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}$. Indeed, let $p_0 \in \Lambda$ and $R > 0$, and consider $m_R = \lceil R/\varepsilon \rceil \in \mathbb{N}^*$. Then for all $g \in \mathcal{L}_1(\mathcal{D}_{p_0})$ such that $\|g\|_{\mathcal{L}_1(\mathcal{D}_{p_0})} \leq R$,

we have $g \in S_{m_R,p_0}$, and thus

$$\begin{aligned}
&\inf_{(m,p)\in\mathcal{M}} \left(\inf_{h\in S_{m,p}} \|f-h\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p) \right) \\
&\leq \|f-g\|^2 + \lambda_{p_0} m_R \varepsilon + \mathrm{pen}(p_0) \\
&\leq \|f-g\|^2 + \lambda_{p_0} (R+\varepsilon) + \mathrm{pen}(p_0) \\
&\leq \|f-g\|^2 + \lambda_{p_0} R + \left(\frac{4}{5\sqrt{\log 2}} + 1 \right) \mathrm{pen}(p_0) + 4\varepsilon^2. \qquad (8.17)
\end{aligned}$$

So, we deduce from (8.16) and (8.17) that there exists $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \right] \\
&\leq C \left[\inf_{p\in\Lambda} \left(\inf_{R>0} \left(\inf_{\substack{g\in\mathcal{L}_1(\mathcal{D}_p), \\ \|g\|_{\mathcal{L}_1(\mathcal{D}_p)}\leq R}} \|f-g\|^2 + \lambda_p R \right) + \mathrm{pen}(p) \right) + \varepsilon^2 \right] \\
&\leq C \left[\inf_{p\in\Lambda} \left(\inf_{g\in\mathcal{L}_1(\mathcal{D}_p)} \left(\|f-g\|^2 + \lambda_p \|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \mathrm{pen}(p) \right) + \varepsilon^2 \right]. \qquad (8.18)
\end{aligned}$$

Finally, let us notice that from the fact that $\alpha \in]0,1[$ and from (8.15), we have

$$\begin{aligned}
&\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p}) \right] \\
&\leq \frac{1}{\alpha}\, \mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \alpha\,\mathrm{pen}(\hat{p}) \right] \\
&\leq \frac{1}{\alpha}\, \mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \right]. \qquad (8.19)
\end{aligned}$$

Combining (8.18) with (8.19) leads to the result. □

8.1.4 Proof of Theorem 6.1

The proof of Theorem 6.1 is again an application of Theorem 7.1 and it is thus very similar to the proof of Theorem 3.2. In particular, it is still based on the key idea that the Lasso estimator $\hat{f}$ is an approximate penalized least squares estimator over the collection of ℓ_1-balls for $m \in \mathbb{N}^*$, $S_m = \{h \in \mathcal{L}_1(\mathcal{D}),\ \|h\|_{\mathcal{L}_1(\mathcal{D})} \leq m\sigma/\sqrt{n}\}$. The main difference is that the dictionary $\mathcal{D}$ considered for Theorem 3.2 was finite while the dictionary $\mathcal{D} = \{\phi_{a,b},\ a \in \mathbb{R}^d, b \in \mathbb{R}\}$ is infinite. Consequently, we can not use the same tools to check the assumptions of Theorem 7.1, more precisely to provide an upper bound of $\mathbb{E}\left[\sup_{h\in S_m} W(h)\right]$. Here, we shall bound this quantity by using Dudley's criterion (see Theorem 3.18 in [15] for instance) and we shall thus first establish an upper bound of the t-packing number of $\mathcal{D}$ with respect to $\|.\|$.

Definition 8.1. [**t-packing numbers**] *Let $t > 0$ and let $\mathcal{G}$ be a set of functions $\mathbb{R}^d \mapsto \mathbb{R}$. We call t-packing number of $\mathcal{G}$ with respect to $\|.\|$, and denote by $N(t, \mathcal{G}, \|.\|)$, the maximal $m \in \mathbb{N}^*$ such that there exist functions $g_1, \ldots, g_m \in \mathcal{G}$ with $\|g_i - g_j\| \geq t$ for all $1 \leq i < j \leq m$.*

Lemma 8.2. *Let $t > 0$. Then, the t-packing number of $\mathcal{D}$ with respect to $\|.\|$ is upper bounded by*

$$N(t, \mathcal{D}, \|.\|) \leq (n+1)^{d+1} \left(\frac{4+t}{t} \right).$$

Proof. The inequality can easily be deduced from the intermediate result (9.10) in the proof of Lemma 9.3 in [8]. We recall here this result. Let $\mathcal{G}$ be a set of functions $\mathbb{R}^d \mapsto \mathbb{R}$. If $\mathcal{G}$ is a linear vector space of dimension D, then, for every $R > 0$ and $t > 0$, the t-packing number of $\{g \in \mathcal{G}, \|g\| \leq R\}$ with respect to $\|.\|$ is upper bounded by

$$N\left(t, \{g \in \mathcal{G}, \|g\| \leq R\}, \|.\|\right) \leq \left(\frac{4R+t}{t}\right)^D.$$

We can apply this result to the linear span of $\phi_{a,b}$, that we denote by $\mathcal{F}_{a,b}$, for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. From (6.1), we have $\sup_x |\phi_{a,b}(x)| \leq 1$, so $\|\phi_{a,b}\|^2 = \sum_{i=1}^n \phi_{a,b}^2(x_i)/n \leq 1$ and we get that

$$\mathcal{D} = \bigcup_{a\in\mathbb{R}^d, b\in\mathbb{R}} \{\phi_{a,b}\} \subset \bigcup_{a\in\mathbb{R}^d, b\in\mathbb{R}} \{g \in \mathcal{F}_{a,b}, \|g\| \leq 1\},$$

with

$$N\left(t, \{g \in \mathcal{F}_{a,b}, \|g\| \leq 1\}, \|.\|\right) \leq \frac{4+t}{t},$$

for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. To end the proof, just notice that there are at most $(n+1)^{d+1}$ hyperplanes in $\mathbb{R}^d$ separating the points $(x_1, \ldots, x_n)$ in different ways (see Chapter 9 in [8] for instance), with the result that there are at most $(n+1)^{d+1}$ ways of selecting $\phi_{a,b}$ in $\mathcal{D}$ that will be different on the sample $(x_1, \ldots, x_n)$. Therefore, we get that for all $t > 0$,

$$N\left(t, \mathcal{D}, \|.\|\right) \leq (n+1)^{d+1} \frac{4+t}{t}.$$

□

Remark 8.3. Let us point out the fact that we are able to get such an upper bound of $N\left(t, \mathcal{D}, \|.\|\right)$ thanks to the particular structure of the dictionary $\mathcal{D}$. Indeed, for all $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $x \in \mathbb{R}^d$, $\phi_{a,b}(x) \in \{0, 1\}$, but there are at most $(n+1)^{d+1}$ hyperplanes in $\mathbb{R}^d$ separating the observed points $(x_1, \ldots, x_n)$ in different ways, so there are at most $(n+1)^{d+1}$ ways of selecting $\phi_{a,b} \in \mathcal{D}$ which will give different functions on the sample $(x_1, \ldots, x_n)$. In particular, this property enables us to bound the packing numbers of $\mathcal{D}$ without truncation of the dictionary. This would not be possible for an arbitrary infinite (countable or uncountable) dictionary and truncation of the dictionary into finite subdictionaries was necessary to achieve our theoretical results in Section 4 when considering an arbitrary infinite countable dictionary.

The following technical lemma will also be used in the proof of Theorem 6.1.

Lemma 8.4.

$$\int_0^1 \sqrt{\ln\left(\frac{1}{t}\right)}\, dt \leq \sqrt{\pi}.$$

Proof. By integration by parts and by defining $u = \sqrt{2\ln(1/t)}$, we have

$$\int_0^1 \sqrt{\ln(1/t)}\, dt = \left[t\sqrt{\ln(1/t)}\right]_0^1 + \int_0^1 \frac{1}{2\sqrt{\ln(1/t)}}\, dt = \frac{1}{\sqrt{2}} \int_0^{+\infty} \mathrm{e}^{-u^2/2}\, du.$$

But, if Z is a standard Gaussian variable, we have

$$\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-u^2/2} \, du \, = \, \mathbb{P}(Z \geq 0) \leq 1,$$

hence the result. □

Proof of Theorem 6.1 Let us define $\varepsilon = \sigma/\sqrt{n}$. Consider the collection of ℓ_1-balls for $m \in \mathbb{N}^*$,

$$S_m = \left\{ h \in \mathcal{L}_1(\mathcal{D}),\ \|h\|_{\mathcal{L}_1(\mathcal{D})} \leq m\varepsilon \right\}.$$

We have noticed in Section 7 that the Lasso estimator $\hat{f}$ is a ρ-approximate penalized least squares estimator over the sequence $\{S_m,\ m \geq 1\}$ for $\mathrm{pen}(m) = \lambda m\varepsilon$ and $\rho = \lambda\varepsilon$. So, it only remains to determine a lower bound on λ that guarantees that $\mathrm{pen}(m)$ satisfies condition (7.4) and to apply the conclusion of Theorem 7.1.

Let $h \in S_m$. From (6.3), for all $\delta > 0$, there exist coefficients $\theta_{a,b}$ such that $h = \sum_{a,b} \theta_{a,b}\, \phi_{a,b}$ and $\sum_{a,b} |\theta_{a,b}| \leq m\varepsilon + \delta$. By using the linearity of W, we get that

$$W(h) = \sum_{a,b} \theta_{a,b}\, W(\phi_{a,b}) \leq \sup_{a,b} |W(\phi_{a,b})| \sum_{a,b} |\theta_{a,b}| \leq (m\varepsilon + \delta) \sup_{a,b} |W(\phi_{a,b})|.$$

Then, by Dudley's criterion (see Theorem 3.18 in [15] for instance), we have

$$\begin{aligned} \Delta_m := \mathbb{E}\left[\sup_{h \in S_m} W(h) \right] &\leq (m\varepsilon + \delta)\, \mathbb{E}\left[\sup_{a,b} |W(\phi_{a,b})| \right] \\ &\leq 12(m\varepsilon + \delta) \int_0^{\alpha} \sqrt{\ln\left(N\left(t, \mathcal{D}, \|.\|\right)\right)}\ dt, \end{aligned}$$

where $\alpha^2 = \sup_{a,b} \mathbb{E}\left[W^2(\phi_{a,b})\right] = \sup_{a,b} \|\phi_{a,b}\|^2 = \sup_{a,b} \left(\sum_{i=1}^n \phi_{a,b}^2(x_i)/n \right) \leq 1$ from (6.1). So,

$$\Delta_m \leq 12(m\varepsilon + \delta) \int_0^{1} \sqrt{\ln\left(N\left(t, \mathcal{D}, \|.\|\right)\right)}\ dt.$$

Moreover, by using Lemma 8.2 and Lemma 8.4, we get that

$$\begin{aligned} &\int_0^1 \sqrt{\ln\left(N\left(t, \mathcal{D}, \|.\|\right)\right)}\ dt \\ &\leq \int_0^1 \sqrt{\ln\left[(n+1)^{d+1}\, \frac{4+t}{t} \right]}\ dt \\ &= \int_0^1 \sqrt{\ln\left((n+1)^{d+1}\right) + \ln(4+t) + \ln\left(\frac{1}{t}\right)}\ dt \\ &\leq \sqrt{\ln\left((n+1)^{d+1}\right)} + \int_4^5 \sqrt{\ln(t)}\ dt + \int_0^1 \sqrt{\ln\left(\frac{1}{t}\right)}\ dt \\ &\leq \sqrt{\ln\left((n+1)^{d+1}\right)} + \sqrt{\ln 5} + \sqrt{\pi}. \end{aligned}$$

Thus,

$$\Delta_m \leq 12(m\varepsilon+\delta)\left[\sqrt{\ln\left((n+1)^{d+1}\right)}+C\right],$$

where $C=\sqrt{\ln 5}+\sqrt{\pi}\in]0,4[$.

Now, choose the weights of the form $x_m=\gamma m$ where γ is a positive numerical constant specified below. Then $\sum_{m\geq 1} e^{-x_m}=1/\left(e^{\gamma}-1\right):=\Sigma_\gamma<+\infty$.

Defining $K=14/13>1$, $\gamma=13\,(4-C)/4>0$, and using the inequality $2\sqrt{ab}\leq \eta a+\eta^{-1}b$ with $\eta=1/6$, we get that

$$\begin{aligned}
&2K\varepsilon\left(\Delta_m+\varepsilon x_m+\sqrt{\Delta_m\varepsilon x_m}\right)\\
&\leq K\varepsilon\left(\frac{13}{6}\,\Delta_m+8x_m\varepsilon\right)\\
&\leq K(m\varepsilon+\delta)\varepsilon\left(26\left[\sqrt{\ln\left((n+1)^{d+1}\right)}+C\right]+8\gamma\right)\\
&\leq 28(m\varepsilon+\delta)\varepsilon\left(\sqrt{\ln\left((n+1)^{d+1}\right)}+C+4-C\right)\\
&< 28(m\varepsilon+\delta)\varepsilon\left(\sqrt{\ln\left((n+1)^{d+1}\right)}+4\right).
\end{aligned}$$

Since this inequality is true for all $\delta>0$, we get when δ tends to 0 that

$$2K\varepsilon\left(\Delta_m+\varepsilon x_m+\sqrt{\Delta_m\varepsilon x_m}\right)\leq 28m\varepsilon^2\left(\sqrt{\ln\left((n+1)^{d+1}\right)}+4\right)\leq \lambda m\varepsilon$$

as soon as

$$\lambda\geq 28\varepsilon\left(\sqrt{\ln\left((n+1)^{d+1}\right)}+4\right). \tag{8.20}$$

For such values of λ, condition (7.4) on the penalty function is satisfied and me may apply Theorem 7.1 with $\mathrm{pen}(m)=\lambda m\varepsilon$ and $\rho=\lambda\varepsilon$ for all $m\geq 1$. Taking into account the definition of $\hat{m}$ at (7.1) and noticing that $\varepsilon^2\leq\lambda\varepsilon/112$ for λ satisfying (8.20), the risk bound (7.6) leads to

$$\begin{aligned}
&\mathbb{E}\left[\|f-\hat{f}\|^2+\lambda\|\hat{f}\|_{\mathcal{L}_1(\mathcal{D})}\right]\\
&\leq C\left[\inf_{m\geq 1}\left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D})}\leq m\varepsilon}\|f-h\|^2+\lambda m\varepsilon\right)+\lambda\varepsilon+(1+\Sigma_\gamma)\varepsilon^2\right]\\
&\leq C\left[\inf_{m\geq 1}\left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D})}\leq m\varepsilon}\|f-h\|^2+\lambda m\varepsilon\right)+\lambda\varepsilon\right],
\end{aligned}$$

where $C>0$ is some absolute constant. We end the proof as the one of Theorem 3.2. □

8.2 Rates of convergence

8.2.1 Proofs of the upper bounds

We first prove a crucial equivalence between the rates of convergence of the deterministic Lassos and $K_{\mathcal{D}_p}$-functionals, which shall able us to provide an

upper bound of the rates of convergence of the deterministic Lassos when the target function belongs to some interpolation space $\mathcal{B}_{q,r}$. Then, we shall pass on these rates of convergence to the Lasso and the selected Lasso estimators. Looking at the proofs of Proposition 5.6 and Proposition 5.7, we can see that the rate of convergence of a Lasso estimator is nothing else than the rate of convergence of the corresponding deterministic Lasso, whereas we can choose the best penalized rate of convergence of all the deterministic Lassos to get the rate of convergence of the selected Lasso estimator, which explains why this estimator can achieve a much better rate of convergence than any Lasso estimator.

Proof of Lemma 5.1. Let us first prove the right-hand side inequality of (5.6). For all $h \in \mathcal{L}_1(D)$, $\lambda \geq 0$ and $\delta > 0$, we have

$$\|f-h\|^2 + \delta^2 \|h\|^2_{\mathcal{L}_1(D)} + \frac{\lambda^2}{4\delta^2} \;\leq\; \left(\|f-h\| + \delta\|h\|_{\mathcal{L}_1(D)}\right)^2 + \frac{\lambda^2}{4\delta^2}.$$

Taking the infimum on all $\delta > 0$ on both sides, and noticing that the infimum on the left-hand side is achieved for $\delta^2 = \lambda / \left(2\|h\|_{\mathcal{L}_1(D)}\right)$, we get that

$$\|f-h\|^2 + \lambda\|h\|_{\mathcal{L}_1(D)} \;\leq\; \inf_{\delta>0}\left[\left(\|f-h\| + \delta\|h\|_{\mathcal{L}_1(D)}\right)^2 + \frac{\lambda^2}{4\delta^2}\right].$$

Then, taking the infimum on all $h \in \mathcal{L}_1(D)$, we get that

$$\begin{aligned}
&\inf_{h\in\mathcal{L}_1(D)} \left(\|f-h\|^2 + \lambda\|h\|_{\mathcal{L}_1(D)}\right) \\
&\leq \inf_{\delta>0}\left[\inf_{h\in\mathcal{L}_1(D)} \left(\|f-h\| + \delta\|h\|_{\mathcal{L}_1(D)}\right)^2 + \frac{\lambda^2}{4\delta^2}\right] \\
&= \inf_{\delta>0}\left[\left(\inf_{h\in\mathcal{L}_1(D)} \left(\|f-h\| + \delta\|h\|_{\mathcal{L}_1(D)}\right)\right)^2 + \frac{\lambda^2}{4\delta^2}\right] \\
&= \inf_{\delta>0}\left(K_D^2(f,\delta) + \frac{\lambda^2}{4\delta^2}\right),
\end{aligned}$$

which proves the right-hand side inequality of (5.6). Let us now prove similarly the left-hand side inequality of (5.6). By definition of $L_D(f,\lambda)$, for all $\eta > 0$, there exists h_η such that $L_D(f,\lambda) \leq \|f-h_\eta\|^2 + \lambda\|h_\eta\|_{\mathcal{L}_1(D)} \leq L_D(f,\lambda) + \eta$. For all $\delta > 0$, we have

$$\begin{aligned}
K_D^2(f,\delta) + \frac{\lambda^2}{2\delta^2} \;&=\; \inf_{h\in\mathcal{L}_1(D)} \left(\|f-h\| + \delta\|h\|_{\mathcal{L}_1(D)}\right)^2 + \frac{\lambda^2}{2\delta^2} \\
&\leq \left(\|f-h_\eta\| + \delta\|h_\eta\|_{\mathcal{L}_1(D)}\right)^2 + \frac{\lambda^2}{2\delta^2} \\
&\leq 2\left(\|f-h_\eta\|^2 + \delta^2\|h_\eta\|^2_{\mathcal{L}_1(D)}\right) + \frac{\lambda^2}{2\delta^2}\;.
\end{aligned}$$

Taking the infimum on all $\delta > 0$ on both sides, and noticing that the infimum on the right-hand side is achieved for $\delta^2 = \lambda / \left(2\|h_\eta\|_{\mathcal{L}_1(D)}\right)$, we get that

$$\inf_{\delta>0}\left(K_D^2(f,\delta) + \frac{\lambda^2}{2\delta^2}\right) \leq 2\left(\|f-h_\eta\|^2 + \lambda\|h_\eta\|_{\mathcal{L}_1(D)}\right) \leq 2\left(L_D(f,\lambda) + \eta\right).$$

We get the expected inequality when η tends to zero. □

Proof of Lemma 5.4. Let $p \in \mathbb{N}^*$ and $\delta > 0$. Applying (5.6) with $D = \mathcal{D}_p$ and $\lambda = \lambda_p$, we have

$$\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f-h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) \leq \inf_{\delta > 0} \left(K^2_{\mathcal{D}_p}(f,\delta) + \frac{\lambda_p^2}{4\delta^2} \right). \tag{8.21}$$

So, it just remains to bound $K^2_{\mathcal{D}_p}(f,\delta)$ when $f \in \mathcal{B}_{q,r}(R)$. Let $\alpha := 1/q - 1/2$. By definition of $f \in \mathcal{B}_{q,r}(R)$, for all $\delta > 0$, there exists $g \in \mathcal{L}_{1,r}$ such that $\|f-g\| + \delta \|g\|_{\mathcal{L}_{1,r}} \leq R\,\delta^{2\alpha}$. So, we have

$$\|f-g\| \leq R\,\delta^{2\alpha} \tag{8.22}$$

and

$$\|g\|_{\mathcal{L}_{1,r}} \leq R\,\delta^{2\alpha-1}. \tag{8.23}$$

Then, by definition of $g \in \mathcal{L}_{1,r}$, there exists $g_p \in \mathcal{L}_1(\mathcal{D}_p)$ such that

$$\|g_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq \|g\|_{\mathcal{L}_{1,r}} \tag{8.24}$$

and

$$\|g-g_p\| \leq \|g\|_{\mathcal{L}_{1,r}}\, p^{-r}. \tag{8.25}$$

Then, we get from (8.22), (8.25) and (8.23) that

$$\|f-g_p\| \leq \|f-g\| + \|g-g_p\| \leq R\,\left(\delta^{2\alpha} + \delta^{2\alpha-1}\, p^{-r}\right), \tag{8.26}$$

and we deduce from (5.5), (8.26), (8.24) and (8.23) that

$$K_{\mathcal{D}_p}(f,\delta) \leq \|f-g_p\| + \delta\,\|g_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq R\left(2\delta^{2\alpha} + \delta^{2\alpha-1}\, p^{-r}\right).$$

So, we get from (8.21) that

$$\begin{aligned}
\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f-h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) &\leq \inf_{\delta > 0} \left(R^2 \left(2\delta^{2\alpha} + \delta^{2\alpha-1}\, p^{-r}\right)^2 + \frac{\lambda_p^2}{4\delta^2} \right) \\
&\leq \inf_{\delta > 0} \left(2R^2 \left(4\delta^{4\alpha} + \delta^{4\alpha-2}\, p^{-2r}\right) + \frac{\lambda_p^2}{4\delta^2} \right).
\end{aligned} \tag{8.27}$$

Let us now consider δ_0 such that $8R^2{\delta_0}^{4\alpha} = \lambda_p^2 \left(4{\delta_0}^2\right)^{-1}$, and δ_1 such that $2R^2\delta_1^{4\alpha-2}\, p^{-2r} = \lambda_p^2 \left(4{\delta_1}^2\right)^{-1}$, that is to say $\delta_0 = \left(\lambda_p \left(4\sqrt{2}R\right)^{-1}\right)^{1/(2\alpha+1)}$ and $\delta_1 = \left(\lambda_p\, p^r \left(2\sqrt{2}R\right)^{-1}\right)^{1/(2\alpha)}$. We can notice that there exists $C_q > 0$ depending only on q such that $\delta_0^{4\alpha-2}\, p^{-2r} \leq C_q \delta_0^{4\alpha}$ for all p checking $\lambda_p\, p^{r(2\alpha+1)} \geq R$, while $\delta_1^{4\alpha} \leq C_q \delta_1^{4\alpha-2}\, p^{-2r}$ for all p checking $\lambda_p\, p^{r(2\alpha+1)} < R$. Therefore, we deduce from (8.27) that there exists $C_q > 0$ depending only on q such that for all p checking $\lambda_p\, p^{r(2\alpha+1)} \geq R$,

$$\begin{aligned}
\inf_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f-h\|^2 + \lambda_p \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) &\leq 2R^2 \left(4\delta_0^{4\alpha} + \delta_0^{4\alpha-2}\, p^{-2r}\right) + \frac{\lambda_p^2}{4\delta_0^2} \\
&\leq C_q\, R^2 \delta_0^{4\alpha} \\
&\leq C_q\, R^{\frac{2}{2\alpha+1}} \lambda_p^{\frac{4\alpha}{2\alpha+1}} \\
&= C_q\, R^q \lambda_p^{2-q},
\end{aligned} \tag{8.28}$$

while for all p checking $\lambda_p\, p^{r(2\alpha+1)} < R$,

$$\begin{aligned}\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) &\le 2R^2\left(4\delta_1^{4\alpha}+\delta_1^{4\alpha-2}\,p^{-2r}\right)+\frac{\lambda_p^2}{4\delta_1^2}\\ &\le C_q\,R^2\delta_1^{4\alpha-2}\,p^{-2r}\\ &\le C_q\,R^{\frac{1}{\alpha}}\,p^{-\frac{r}{\alpha}}\,\lambda_p^{2-\frac{1}{\alpha}}\\ &= C_q\left(Rp^{-r}\right)^{\frac{2q}{2-q}}\lambda_p^{\frac{4(1-q)}{2-q}}. \qquad (8.29)\end{aligned}$$

Inequalities (8.28) and (8.29) can be summarized into the following result:

$$\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) \le C_q\max\left(R^q\lambda_p^{2-q}\,,\,\left(Rp^{-r}\right)^{\frac{2q}{2-q}}\lambda_p^{\frac{4(1-q)}{2-q}}\right).$$

□

Proof of Proposition 5.6. From (5.1) and (5.9), we know that there exists some constant $C_q > 0$ depending only on q such that, for all $p \in \mathbb{N}^*$, the quadratic risk of $\hat{f}_p$ is bounded by

$$\begin{aligned}\mathbb{E}\left[\|f-\hat{f}_p\|^2\right] &\le C_q\left(\max\left(R^q\lambda_p^{2-q}\,,\,\left(Rp^{-r}\right)^{\frac{2q}{2-q}}\lambda_p^{\frac{4(1-q)}{2-q}}\right)+\lambda_p\varepsilon\right)\\ &\le C_q\max\left(R^q\lambda_p^{2-q}\,,\,\left(Rp^{-r}\right)^{\frac{2q}{2-q}}\lambda_p^{\frac{4(1-q)}{2-q}}\,,\,\lambda_p\varepsilon\right).\end{aligned}$$

By remembering that $\lambda_p = 4\varepsilon(\sqrt{\ln p}+1)$ and by comparing the three terms inside the maximum according to the value of p, we get (5.14), (5.15) and (5.16). □

Proof of Proposition 5.7. From (5.2) and (5.9), we know that there exists some constant $C_q > 0$ depending only on q such that the quadratic risk of $\hat{f}_{\hat{p}}$ is bounded by

$$\begin{aligned}&\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2\right]\\ &\le C_q\left[\inf_{p\in\Lambda}\left(\max\left(R^q\lambda_p^{2-q}\,,\,\left(Rp^{-r}\right)^{\frac{2q}{2-q}}\lambda_p^{\frac{4(1-q)}{2-q}}\right)+\varepsilon^2\ln p\right)+\varepsilon^2\right]\\ &\le C_q\inf_{p\in\Lambda\setminus\{1\}}\left(\max\left(R^q\left(\varepsilon\sqrt{\ln p}\right)^{2-q}\,,\,\left(Rp^{-r}\right)^{\frac{2q}{2-q}}\left(\varepsilon\sqrt{\ln p}\right)^{\frac{4(1-q)}{2-q}}\right)+\varepsilon^2\ln p\right), \qquad (8.30)\end{aligned}$$

where we use the fact that for all $p \ge 2$, we have $\lambda_p = 4\varepsilon(\sqrt{\ln p}+1) \le 4(1+1/\sqrt{\ln 2})\varepsilon\sqrt{\ln p}$ and $\varepsilon^2 \le \varepsilon^2(\ln p)/\ln 2$. We now choose p such that the two terms inside the maximum are approximately of the same order. More precisely, let us define

$$J_{q,r} = \left\lceil \frac{q}{2r}\log_2\left(R\varepsilon^{-1}\right)\right\rceil,$$

where $\lceil x\rceil$ denotes the smallest integer greater than x, and $p_{q,r} := 2^{J_{q,r}}$. Since we have assumed $R\varepsilon^{-1} \ge \mathrm{e}$, we have $p_{q,r} \in \Lambda\setminus\{1\}$ and we deduce from (8.30)

that

$$\begin{aligned}&\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2\right]\\&\leq C_q\left(\max\left(R^q\left(\varepsilon\sqrt{\ln p_{q,r}}\right)^{2-q},\left(Rp_{q,r}^{-r}\right)^{\frac{2q}{2-q}}\left(\varepsilon\sqrt{\ln p_{q,r}}\right)^{\frac{4(1-q)}{2-q}}\right)+\varepsilon^2\ln p_{q,r}\right).\end{aligned}\tag{8.31}$$

Now, let us give an upper bound of each term of the right-hand side of (8.31) by assuming that $R\varepsilon^{-1}\geq\max\left(\mathrm{e},(4r)^{-1}q\right)$. First, we have by definition of $p_{q,r}$ that

$$\ln p_{q,r}\leq\ln 2+\frac{q}{2r}\ln\left(R\varepsilon^{-1}\right).$$

Moreover, for all $x>0$, $\ln 2\leq 2\,x\ln x$ and by assumption $R\varepsilon^{-1}\geq q/(4r)$, so

$$\ln 2\leq\frac{q}{2r}\ln\left(\frac{q}{4r}\right)\leq\frac{q}{2r}\ln\left(R\varepsilon^{-1}\right),$$

and thus we get that

$$\ln p_{q,r}\leq\frac{q}{r}\ln\left(R\varepsilon^{-1}\right).\tag{8.32}$$

Then, we deduce from (8.32) that the first term of (8.31) is upper bounded by

$$R^q\left(\varepsilon\sqrt{\ln p_{q,r}}\right)^{2-q}\leq R^q\left(\frac{q}{r}\right)^{1-\frac{q}{2}}\left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}.\tag{8.33}$$

For the second term of (8.31), using (8.32), the fact that $p_{q,r}\geq(R\varepsilon^{-1})^{\frac{q}{2r}}$, that $\frac{4(1-q)}{2-q}\leq 2-q$ and that $\ln(R\varepsilon^{-1})\geq 1$, we get that

$$\begin{aligned}\left(Rp_{q,r}^{-r}\right)^{\frac{2q}{2-q}}\left(\varepsilon\sqrt{\ln p_{q,r}}\right)^{\frac{4(1-q)}{2-q}}&\leq R^q\varepsilon^{2-q}\left(\sqrt{\frac{q}{r}\ln\left(R\varepsilon^{-1}\right)}\right)^{\frac{4(1-q)}{2-q}}\\&\leq R^q\left(\frac{q}{r}\right)^{\frac{2(1-q)}{2-q}}\left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}.\end{aligned}\tag{8.34}$$

For the third term of (8.31), we have

$$\varepsilon^2\ln p_{q,r}\leq\frac{q}{r}\,\varepsilon^2\ln\left(R\varepsilon^{-1}\right)=\frac{q}{r}\left(\frac{\ln\left[(R\varepsilon^{-1})^2\right]}{2(R\varepsilon^{-1})^2}\right)^{q/2}R^q\left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}.$$

Now, let us introduce

$$g:\,]0,+\infty[\mapsto\mathbb{R},\;\;x\mapsto\frac{\ln x}{x}.$$

It is easy to check that $g(x^2)\leq 1/x$ for all $x>0$. Using this property and the fact that $R\varepsilon^{-1}\geq\mathrm{e}$, we get that

$$\frac{\ln\left[(R\varepsilon^{-1})^2\right]}{(R\varepsilon^{-1})^2}=g\left((R\varepsilon^{-1})^2\right)\leq\frac{1}{R\varepsilon^{-1}}\leq\frac{1}{\mathrm{e}},$$

and thus

$$\varepsilon^2\ln p_{q,r}\leq\frac{q}{r}\left(\frac{1}{2\mathrm{e}}\right)^{q/2}R^q\left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}.\tag{8.35}$$

Then, we deduce from (8.31), (8.33), (8.34) and (8.35) that there exists $C_{q,r} > 0$ depending only on q and r such that

$$\mathbb{E}\left[\|f - \hat{f}_{\hat{p}}\|^2\right] \leq C_{q,r}\, R^q \left(\varepsilon\sqrt{\ln\left(R\varepsilon^{-1}\right)}\right)^{2-q}.$$

□

Proof of Proposition 6.3. Set $\varepsilon = \sigma/\sqrt{n}$. From Theorem 6.1, we have

$$\mathbb{E}\left[\|f - \hat{f}\|^2\right] \leq C\left[\inf_{h\in\mathcal{L}_1(\mathcal{D})}\left(\|f-h\|^2 + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right) + \lambda\varepsilon\right], \tag{8.36}$$

where C is an absolute positive constant. Then, if $f \in \mathcal{B}_q(R)$, we get from (6.4) and (5.6) with $D = \mathcal{D}$ that

$$\inf_{h\in\mathcal{L}_1(\mathcal{D})}\left(\|f-h\|^2 + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right) \leq \inf_{\delta>0}\left(R^2\delta^{4\alpha} + \frac{\lambda^2}{4\delta^2}\right).$$

The infimum on the right-hand side is achieved for $\delta = (\lambda/(2R))^{1/(2\alpha+1)}$ and the last inequality leads to

$$\begin{aligned}\inf_{h\in\mathcal{L}_1(\mathcal{D})}\left(\|f-h\|^2 + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}\right) &\leq 2R^{\frac{2}{2\alpha+1}}\left(\frac{\lambda}{2}\right)^{\frac{4\alpha}{2\alpha+1}} \\ &= 2^{\frac{1-2\alpha}{1+2\alpha}}\, R^{\frac{2}{2\alpha+1}}\, \lambda^{\frac{4\alpha}{2\alpha+1}} \\ &= 2^{q-1} R^q \lambda^{2-q}.\end{aligned}$$

Thus, we deduce from (8.36) that there exists some $C_q > 0$ depending only on q such that

$$\begin{aligned}&\mathbb{E}\left[\|f - \hat{f}\|^2\right] \\ &\leq C_q\left[R^q\lambda^{2-q} + \lambda\varepsilon\right] \\ &\leq C_q\left[R^q\left(\varepsilon\left(\sqrt{\ln\left((n+1)^{d+1}\right)} + 4\right)\right)^{2-q} + \varepsilon^2\left(\sqrt{\ln\left((n+1)^{d+1}\right)} + 4\right)\right] \\ &\leq C_q\left[R^q\left(\varepsilon\sqrt{\ln\left((n+1)^{d+1}\right)}\right)^{2-q} + \varepsilon^2\sqrt{\ln\left((n+1)^{d+1}\right)}\right] \qquad (8.37) \\ &\leq C_q \max\left(R^q\left(\varepsilon\sqrt{\ln\left((n+1)^{d+1}\right)}\right)^{2-q}, \varepsilon^2\sqrt{\ln\left((n+1)^{d+1}\right)}\right) \\ &\leq C_q\, R^q\left(\varepsilon\sqrt{\ln\left((n+1)^{d+1}\right)}\right)^{2-q}, \qquad (8.38)\end{aligned}$$

where we get (8.37) by using the fact $4 \leq 5\sqrt{\ln 2} \leq 5\sqrt{\ln\left((n+1)^{d+1}\right)}$ for $n \geq 1$ and $d \geq 1$ and (8.38) thanks to the assumption $R\varepsilon^{-1} \geq \left[\ln\left((n+1)^{d+1}\right)\right]^{\frac{q-1}{2q}}$. □

8.2.2 Proofs of the lower bounds in the orthonormal case

To prove that the rates of convergence (5.17) achieved by the selected Lasso estimator on the classes $\mathcal{B}_{q,r}$ are optimal, we propose to establish a lower bound

of the minimax risk over $\mathcal{B}_{q,r}$ when the dictionary is an orthonormal basis of $\mathbb{H}$ and to check that it is of the same order as the rates (5.17). The first central point is to prove Remark 5.5, that is to say the inclusion in the orthonormal case of the space $w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ in the space $\mathcal{B}_{q,r}(C_{q,r}R)$ for all $R > 0$ and some $C_{q,r} > 0$ depending only on q and r. Taking this inclusion into account, we shall then focus on establishing a lower bound of the minimax risk over the smaller space $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$, which shall reveal to be an easy task, and which entails the same lower bound over the bigger space $\mathcal{B}_{q,r}$.

Proof of Remark 5.5. Let $R > 0$. The first inclusion comes from the simple inclusion $\mathcal{L}_q(R) \subset w\mathcal{L}_q(R)$. Let us prove the second inclusion here. Assume that $f \in w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. For all $p \in \mathbb{N}^*$ and $\beta > 0$, define

$$f_{p,\beta} := \underset{h \in \mathcal{L}_1(\mathcal{D}_p)}{\arg\min} \left(\|f - h\|^2 + \beta \|h\|_{\mathcal{L}_1(\mathcal{D}_p)} \right). \tag{8.39}$$

The proof will be divided in two main parts. First, we shall choose β such that $f_{p,\beta} \in \mathcal{L}_{1,r}$. Secondly, we shall choose p such that $\|f - f_{p,\beta}\| + \delta \|f_{p,\beta}\|_{\mathcal{L}_{1,r}} \leq C_{q,r} R \, \delta^{2\alpha}$ for all $\delta > 0$, some $C_{q,r} > 0$ and $\alpha = 1/q - 1/2$, which shall prove that $f \in \mathcal{B}_{q,r}(C_{q,r}R)$. To establish our results, we shall need an upper bound of $\|f - f_{p,\beta}\|$ and $\|f_{p,\beta}\|_{\mathcal{L}_1(\mathcal{D}_p)}$. These bounds are provided by Lemma 8.5 stated below.

Let us first choose β such that $f_{p,\beta} \in \mathcal{L}_{1,r}$. From Lemma 8.5, we have

$$\|f - f_{p,\beta}\| \leq R(p+1)^{-r} + \sqrt{C_q} \, R^{q/2} \beta^{1-q/2}.$$

Now choose β such that $R(p+1)^{-r} = \sqrt{C_q} \, R^{q/2} \beta^{1-q/2}$, that is to say

$$\beta_p = R \left(\sqrt{C_q} \, (p+1)^r \right)^{-\frac{2}{2-q}}. \tag{8.40}$$

Then, we have

$$\|f - f_{p,\beta_p}\| \leq 2R(p+1)^{-r} \leq 2R(2p)^{-r} = 2^{1-r} R p^{-r}. \tag{8.41}$$

Let us now check that $f_{p,\beta_p} \in \mathcal{L}_{1,r}$. Define

$$C_p := \max \left\{ 2^{2-r} R, \max_{p' \in \mathbb{N}^*, p' \leq p} \|f_{p',\beta_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \right\}. \tag{8.42}$$

Let $p' \in \mathbb{N}^*$. By definition of $f_{p',\beta_{p'}}$, we have $f_{p',\beta_{p'}} \in \mathcal{L}_1(\mathcal{D}_{p'})$. If $p' \leq p$, then we deduce from (8.41) that

$$\|f_{p,\beta_p} - f_{p',\beta_{p'}}\| \leq \|f_{p,\beta_p} - f\| + \|f - f_{p',\beta_{p'}}\| \leq 2^{1-r} R \left(p^{-r} + p'^{-r} \right) \leq C_p p'^{-r},$$

and we have $\|f_{p',\beta_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \leq C_p$ by definition of C_p. If $p' > p$, then $\mathcal{L}_1(\mathcal{D}_p) \subset \mathcal{L}_1(\mathcal{D}_{p'})$ and $f_{p,\beta_p} \in \mathcal{L}_1(\mathcal{D}_{p'})$ with $\|f_{p,\beta_p}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \leq \|f_{p,\beta_p}\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C_p$ and $\|f_{p,\beta_p} - f_{p,\beta_p}\| = 0 \leq C_p p'^{-r}$. So, $f_{p,\beta_p} \in \mathcal{L}_{1,r}$.

Now, it only remains to choose a convenient $p \in \mathbb{N}^*$ so as to prove that $f \in \mathcal{B}_{q,r}(R)$.
Let us first give an upper bound of $\|f_{p,\beta_p}\|_{\mathcal{L}_{1,r}}$ for all $p \in \mathbb{N}^*$. By definition of $\|f_{p,\beta_p}\|_{\mathcal{L}_{1,r}}$ and the above upper bounds, we have $\|f_{p,\beta_p}\|_{\mathcal{L}_{1,r}} \leq C_p$. So, we just

have to bound C_p. Let $p' \in \mathbb{N}^*, p' \leq p$. From Lemma 8.5, we know that there exists $C_q > 0$ depending only on q such that $\|f_{p',\beta_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} \leq C_q R^q \beta_{p'}^{1-q}$. So, we get from (8.40) that

$$\begin{aligned}\|f_{p',\beta_{p'}}\|_{\mathcal{L}_1(\mathcal{D}_{p'})} &\leq C_q R \left(\sqrt{C_q}\,(p'+1)^r\right)^{\frac{2(q-1)}{2-q}} \\ &\leq C_q R \left(\sqrt{C_q}(p+1)^r\right)^{\frac{2(q-1)}{2-q}} \\ &= C_q^{\frac{1}{2-q}} R\,(2p)^{\frac{2(q-1)r}{2-q}}\,,\end{aligned}$$

and we deduce from (8.42) that

$$C_p \leq \max\left(2^{2-r}R, C_q^{\frac{1}{2-q}} R\,(2p)^{\frac{2(q-1)r}{2-q}}\right) \leq C_{q,r} R p^{\frac{2(q-1)r}{2-q}}$$

where $C_{q,r} > 0$ depends only on q and r. Thus, we have

$$\|f_{p,\beta_p}\|_{\mathcal{L}_{1,r}} \leq C_{q,r} R p^{\frac{2(q-1)r}{2-q}}. \tag{8.43}$$

Then, we deduce from (8.41) and (8.43) that for all $p \in \mathbb{N}^*$ and $\delta > 0$,

$$\begin{aligned}\inf_{h \in \mathcal{L}_{1,r}} \|f-h\| + \delta\|h\|_{\mathcal{L}_{1,r}} &\leq \|f - f_{p,\beta_p}\| + \delta\|f_{p,\beta_p}\|_{\mathcal{L}_{1,r}} \\ &\leq 2^{1-r} R p^{-r} + \delta C_{q,r} R p^{\frac{2(q-1)r}{2-q}}. \end{aligned} \tag{8.44}$$

We now choose $p \geq 2$ such that $p^{-r} \approx \delta\, p^{\frac{2(q-1)r}{2-q}}$. More precisely, set $p = 2^J$ where $J = \left\lceil (2-q)(qr)^{-1} \log_2(\delta^{-1}) \right\rceil$. With this value of p, we get that there exists $C'_{q,r} > 0$ depending only on q and r such that (8.44) is upper bounded by $C'_{q,r} R\, \delta^{(2-q)/q} = C'_{q,r} R\, \delta^{2\alpha}$. This means that $f \in \mathcal{B}_{q,r}(C'_{q,r}\, R)$, hence (5.13). □

Lemma 8.5. *Assume that the dictionary $\mathcal{D}$ is an orthonormal basis of the Hilbert space $\mathbb{H}$ and that there exist $1 < q < 2$, $r > 0$ and $R > 0$ such that $f \in w\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$. For all $p \in \mathbb{N}^*$ and $\beta > 0$, define*

$$f_{p,\beta} := \operatorname*{arg\,min}_{h \in \mathcal{L}_1(\mathcal{D}_p)} \left(\|f-h\|^2 + \beta\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right).$$

Then, there exists $C_q > 0$ depending only on q such that for all $p \in \mathbb{N}^*$ and $\beta > 0$,

$$\|f_{p,\beta}\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C_q R^q \beta^{1-q}$$

and

$$\|f - f_{p,\beta}\| \leq R(p+1)^{-r} + \sqrt{C_q}\, R^{q/2} \beta^{1-q/2}.$$

The proof of Lemma 8.5 uses the two following technical lemmas.

Lemma 8.6. *For all $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$ and $\gamma > 0$,*

$$\sum_{j=1}^p a_j^2\, \mathbb{1}_{\{|a_j| \leq \gamma\}} \leq 2 \sum_{j=1}^p \int_0^\gamma t\, \mathbb{1}_{\{|a_j| > t\}}\, dt.$$

Proof.

$$
\begin{aligned}
&2\sum_{j=1}^{p}\int_0^{\gamma} t\,\mathbb{1}_{\{|a_j|>t\}}\,dt \\
&= 2\sum_{j=1}^{p}\left[\left(\int_0^{\gamma} t\,\mathbb{1}_{\{|a_j|>t\}}\,dt\right)\mathbb{1}_{\{|a_j|>\gamma\}} + \left(\int_0^{\gamma} t\,\mathbb{1}_{\{|a_j|>t\}}\,dt\right)\mathbb{1}_{\{|a_j|\le\gamma\}}\right] \\
&= 2\sum_{j=1}^{p}\left[\left(\int_0^{\gamma} t\,dt\right)\mathbb{1}_{\{|a_j|>\gamma\}} + \left(\int_0^{|a_j|} t\,dt\right)\mathbb{1}_{\{|a_j|\le\gamma\}}\right] \\
&= \sum_{j=1}^{p}\left(\gamma^2\,\mathbb{1}_{\{|a_j|>\gamma\}} + a_j^2\,\mathbb{1}_{\{|a_j|\le\gamma\}}\right) \\
&\ge \sum_{j=1}^{p} a_j^2\,\mathbb{1}_{\{|a_j|\le\gamma\}}.
\end{aligned}
$$

□

Lemma 8.7. *For all $a=(a_1,\dots,a_p)\in\mathbb{R}^p$ and $\gamma>0$,*

$$
\sum_{j=1}^{p}|a_j|\mathbb{1}_{\{|a_j|>\gamma\}} = \gamma\sum_{j=1}^{p}\mathbb{1}_{\{|a_j|>\gamma\}} + \sum_{j=1}^{p}\int_{\gamma}^{+\infty}\mathbb{1}_{\{|a_j|>t\}}\,dt.
$$

Proof.

$$
\sum_{j=1}^{p}\int_{\gamma}^{+\infty}\mathbb{1}_{\{|a_j|>t\}}\,dt = \sum_{j=1}^{p}\left(\int_{\gamma}^{|a_j|}dt\right)\mathbb{1}_{\{|a_j|>\gamma\}} = \sum_{j=1}^{p}(|a_j|-\gamma)\,\mathbb{1}_{\{|a_j|>\gamma\}}.
$$

□

Proof of Lemma 8.5. Let denote by $\{\theta_j^*\}_{j\in\mathbb{N}^*}$ the coefficients of the target function f in the basis $\mathcal{D}=\{\phi_j\}_{j\in\mathbb{N}^*}$, $f=\theta^*.\phi=\sum_{j\in\mathbb{N}^*}\theta_j^*\,\phi_j$. We introduce for all $p\in\mathbb{N}^*$,

$$
\Theta_p := \left\{\theta=(\theta_j)_{j\in\mathbb{N}^*},\ \theta=(\theta_1,\dots,\theta_p,0,\dots,0,\dots)\right\}.
$$

Let $\beta>0$. Since $f_{p,\beta}\in\mathcal{L}_1(\mathcal{D}_p)$, there exists a unique $\theta^{p,\beta}\in\Theta_p$ such that $f_{p,\beta}=\theta^{p,\beta}.\phi$. Moreover, from (3.1) and using the orthonormality of the basis functions ϕ_j, we have

$$
\theta^{p,\beta} = \arg\min_{\theta\in\Theta_p}\left(\|\theta^*.\phi-\theta.\phi\|^2+\beta\|\theta\|_1\right) = \arg\min_{\theta\in\Theta_p}\left(\|\theta^*-\theta\|^2+\beta\|\theta\|_1\right). \tag{8.45}
$$

By calculating the subdifferential of the function $\theta\in\mathbb{R}^p\mapsto\|\theta^*-\theta\|^2+\beta\|\theta\|_1$, we get that the solution of the convex minimization problem (8.45) is $\theta^{p,\beta}=(\theta_1^{p,\beta},\dots,\theta_p^{p,\beta},0,\dots,0,\dots)$ where for all $j=1,\dots,p$,

$$
\theta_j^{p,\beta} = \begin{cases} \theta_j^*-\beta/2 & \text{if } \theta_j^*>\beta/2, \\ \theta_j^*+\beta/2 & \text{if } \theta_j^*<-\beta/2, \\ 0 & \text{else.} \end{cases}
$$

Then, we have

$$\begin{aligned}\|f - f_{p,\beta}\|^2 &= \|\theta^* - \theta^{p,\beta}\|^2 \\ &= \sum_{j=1}^{\infty} \left(\theta_j^* - \theta_j^{p,\beta}\right)^2 \\ &= \sum_{j=p+1}^{\infty} \theta_j^{*2} + \sum_{j=1}^{p} \theta_j^{*2} \mathbb{1}_{\{|\theta_j^*| \le \beta/2\}} + \sum_{j=1}^{p} \frac{\beta^2}{4} \mathbb{1}_{\{|\theta_j^*| > \beta/2\}} \\ &\le \underbrace{\sum_{j=p+1}^{\infty} \theta_j^{*2}}_{(i)} + \underbrace{\sum_{j=1}^{p} \theta_j^{*2} \mathbb{1}_{\{|\theta_j^*| \le \beta/2\}}}_{(ii)} + \underbrace{\frac{\beta}{2} \sum_{j=1}^{p} |\theta_j^*| \mathbb{1}_{\{|\theta_j^*| > \beta/2\}}}_{(iii)} \quad . \end{aligned} \tag{8.46}$$

while

$$\begin{aligned}\|f_{p,\beta}\|_{\mathcal{L}_1(\mathcal{D}_p)} &= \sum_{j=1}^{\infty} |\theta_j^{p,\beta}| \\ &= \sum_{j=1}^{p} \left(|\theta_j^*| - \frac{\beta}{2}\right) \mathbb{1}_{\{|\theta_j^*| > \beta/2\}} \\ &\le \sum_{j=1}^{p} |\theta_j^*| \mathbb{1}_{\{|\theta_j^*| > \beta/2\}} = (iii) \; . \end{aligned} \tag{8.47}$$

Now, since f is assumed to belong to $\mathcal{B}^r_{2,\infty}(R)$, we get from (5.10) that (i) is bounded by

$$\sum_{j=p+1}^{\infty} \theta_j^{*2} \le R^2 (p+1)^{-2r}. \tag{8.48}$$

Let us now bound (ii) and (iii) thanks to the assumption $f \in w\mathcal{L}_q(R)$. By applying Lemma 8.6 and Lemma 8.7 with $a_j = \theta_j^*$ for all $j = 1, \ldots, p$ and $\gamma = \beta/2$, and by using the fact that $\sum_{j=1}^{p} \mathbb{1}_{\{|\theta_j^*| > t\}} \le \sum_{j=1}^{\infty} \mathbb{1}_{\{|\theta_j^*| > t\}} \le R^q t^{-q}$ for all $t > 0$ if $f \in w\mathcal{L}_q(R)$, we get that (ii) is bounded by

$$\begin{aligned}\sum_{j=1}^{p} \theta_j^{*2} \mathbb{1}_{\{|\theta_j^*| \le \beta/2\}} &\le 2 \sum_{j=1}^{p} \int_0^{\beta/2} t \mathbb{1}_{\{|\theta_j^*| > t\}} \, dt \\ &\le 2R^q \int_0^{\beta/2} t^{1-q} \, dt \\ &= \frac{2^{q-1}}{2-q} \, R^q \beta^{2-q}, \end{aligned} \tag{8.49}$$

while (iii) is bounded by

$$\begin{aligned}\sum_{j=1}^{p} |\theta_j^*| \mathbb{1}_{\{|\theta_j^*| > \beta/2\}} &= \frac{\beta}{2} \sum_{j=1}^{p} \mathbb{1}_{\{|\theta_j^*| > \beta/2\}} + \sum_{j=1}^{p} \int_{\beta/2}^{+\infty} \mathbb{1}_{\{|\theta_j^*| > t\}} \, dt \\ &\le R^q \left(\frac{\beta}{2}\right)^{1-q} + R^q \int_{\beta/2}^{+\infty} t^{-q} \, dt \\ &= \frac{q \, 2^{q-1}}{q-1} \, R^q \beta^{1-q}. \end{aligned} \tag{8.50}$$

Gathering together (8.47) and (8.50) on the one hand and (8.46), (8.48), (8.49) and (8.50) on the other hand, we get that there exists $C_q > 0$ depending only on q such that

$$\|f_{p,\beta}\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C_q R^q \beta^{1-q}$$

and

$$\|f - f_{p,\beta}\|^2 \leq R^2(p+1)^{-2r} + C_q R^q \beta^{2-q}.$$

Finally,

$$\|f - f_{p,\beta}\| \leq \sqrt{R^2(p+1)^{-2r} + C_q R^q \beta^{2-q}} \leq R(p+1)^{-r} + \sqrt{C_q}\, R^{q/2} \beta^{1-q/2}.$$

□

Proof of Proposition 5.10. Let us define

$$M = \varepsilon\sqrt{u \ln(R\,\varepsilon^{-1})}, \quad p = 2^J, \quad d = 2^K,$$

with

$$J = \left\lfloor \frac{2-q}{2r} \log_2\left(RM^{-1}\right) \right\rfloor$$

and

$$K = \left\lfloor q \log_2\left(RM^{-1}\right) \right\rfloor.$$

Let us first check that M is well-defined and that $d \leq p$ under the assumptions of Proposition 5.10. Under the assumption $r < 1/q - 1/2$, we have $u > 0$, and since $R\varepsilon^{-1} \geq \mathrm{e}^2 \geq \mathrm{e}$, M is well-defined. Moreover, since $r < 1/q - 1/2$, we have $(2-q)/(2r) > q$, so it only remains to check that $RM^{-1} \geq \mathrm{e}$ so as to prove that $d \leq p$. We shall in fact prove the following stronger result:

Result ($\diamondsuit$): If $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, u^2)$, then $R\varepsilon^{-1} / \left(\ln(R\varepsilon^{-1})\right) \geq u$.

This result indeed implies that, under the assumption $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, u^2)$,

$$RM^{-1} = \frac{R\varepsilon^{-1}}{\sqrt{u \ln(R\varepsilon^{-1})}} = \sqrt{R\varepsilon^{-1}} \sqrt{\frac{R\varepsilon^{-1}}{u \ln(R\varepsilon^{-1})}} \geq \mathrm{e} \times 1 \geq \mathrm{e}.$$

Let us prove Result ($\diamondsuit$). Introduce the function

$$g :]0, +\infty[\mapsto \mathbb{R}, \ x \mapsto \frac{x}{\ln x}.$$

It is easy to check that g is non-decreasing on $[\mathrm{e}, +\infty[$ and that $g(x^2) \geq x$ for all $x > 0$. Now, assume that $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, u^2)$. Using the properties of g, we deduce that if $u \geq \mathrm{e}$ then $R\varepsilon^{-1} \geq u^2 \geq \mathrm{e}^2 \geq \mathrm{e}$ and

$$\frac{R\varepsilon^{-1}}{\ln(R\varepsilon^{-1})} = g(R\varepsilon^{-1}) \geq g(u^2) \geq u,$$

while if $u < \mathrm{e}$ then $R\varepsilon^{-1} \geq \mathrm{e}^2 \geq \mathrm{e}$ and

$$\frac{R\varepsilon^{-1}}{\ln(R\varepsilon^{-1})} = g(R\varepsilon^{-1}) \geq g(\mathrm{e}^2) \geq \mathrm{e} > u,$$

hence Result ($\diamondsuit$).

Now, consider the following hypercube $\Theta(p,d,M)$ defined by

$$\left\{ \sum_{j=1}^{\infty} \theta_j \, \phi_j, \ (\theta_1, \ldots, \theta_p) \in [0,M]^p, \ \theta_j = 0 \text{ for } j \geq p+1, \ \sum_{j=1}^{p} \mathbb{1}_{\{\theta_j \neq 0\}} = d \right\}$$

$$= \left\{ M \sum_{j=1}^{\infty} \beta_j \, \phi_j, \ (\beta_1, \ldots, \beta_p) \in [0,1]^p, \ \beta_j = 0 \text{ for } j \geq p+1, \ \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} = d \right\}.$$

The essence of the proof is just to ckeck that $\Theta(p,d,M) \subset \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$, which shall enable us to bound from below the minimax risk over $\mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ by the lower bound of the minimax risk over $\Theta(p,d,M)$ provided in [2].

Let $h \in \Theta(p,d,M)$. We write $h = \sum_{j=1}^{\infty} \theta_j \phi_j = M \sum_{j=1}^{\infty} \beta_j \phi_j$.

$$\sum_{j=1}^{\infty} |\theta_j|^q = M^q \sum_{j=1}^{p} \beta_j^q \, \mathbb{1}_{\{\beta_j \neq 0\}} \leq M^q \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} \leq M^q d \leq M^q \left(RM^{-1}\right)^q \leq R^q.$$

Thus, $h \in \mathcal{L}_q(R)$.

Let $J_0 \in \mathbb{N}^*$. If $J_0 > p$, then,

$$J_0^{2r} \sum_{j=J_0}^{\infty} \theta_j^2 \leq J_0^{2r} \sum_{j=p+1}^{\infty} \theta_j^2 = 0 \leq R^2.$$

Now consider $J_0 \leq p$. Then,

$$\begin{aligned} J_0^{2r} \sum_{j=J_0}^{\infty} \theta_j^2 &= J_0^{2r} M^2 \sum_{j=J_0}^{p} \beta_j^2 \mathbb{1}_{\{\beta_j \neq 0\}} \\ &\leq J_0^{2r} M^2 \sum_{j=J_0}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} \\ &\leq p^{2r} M^2 d \\ &\leq \left(RM^{-1}\right)^{2-q} M^2 \left(RM^{-1}\right)^q \\ &\leq R^2. \end{aligned}$$

Thus, $h \in \mathcal{B}^r_{2,\infty}(R)$. Therefore, $\Theta(p,d,M) \subset \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)$ and

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)} \mathbb{E}\left[\|f - \tilde{f}\|^2\right] \geq \inf_{\tilde{f}} \sup_{f \in \Theta(p,d,M)} \mathbb{E}\left[\|f - \tilde{f}\|^2\right]. \tag{8.51}$$

Now, from Theorem 5 in [2], we know that the minimax risk over $\Theta\,(p,d,M)$ satisfies

$$\begin{aligned} \inf_{\tilde{f}} \sup_{f \in \Theta(p,d,M)} \mathbb{E}\left[\|f - \tilde{f}\|^2\right] &\geq \kappa d \min\left(M^2, \varepsilon^2\left(1 + \ln\left(\frac{p}{d}\right)\right)\right) \\ &\geq \kappa \frac{\left(RM^{-1}\right)^q}{2} \min\left(M^2, \varepsilon^2\left(1 + \ln\left(\frac{p}{d}\right)\right)\right) \\ &\geq \kappa' R^q M^{-q} \min\left(M^2, \varepsilon^2\left(1 + \ln\left(\frac{p}{d}\right)\right)\right), \end{aligned} \tag{8.52}$$

where $\kappa > 0$ and $\kappa' > 0$ are absolute constants. Moreover, we have

$$\begin{aligned}
\varepsilon^2 \left(1 + \ln\left(\frac{p}{d}\right)\right) &\geq \varepsilon^2 \left(1 + \ln\left[\frac{\left(RM^{-1}\right)^{\frac{2-q}{2r}}}{2\left(RM^{-1}\right)^q}\right]\right) \\
&= \varepsilon^2 \left(1 + \ln\left[\left(RM^{-1}\right)^u\right] - \ln 2\right) \\
&\geq \varepsilon^2 \ln\left[\left(RM^{-1}\right)^u\right] \\
&= \varepsilon^2 \ln\left[\left(R\varepsilon^{-1}\right)^u \left(\varepsilon M^{-1}\right)^u\right] \\
&= M^2 + \varepsilon^2 \ln\left[\left(\varepsilon M^{-1}\right)^u\right] \\
&= M^2 - \frac{u}{2}\varepsilon^2 \ln\left[u \ln\left(R\varepsilon^{-1}\right)\right]. \qquad (8.53)
\end{aligned}$$

But the assumption $R\varepsilon^{-1} \geq \max(\mathrm{e}^2, u^2)$ implies that (8.53) is greater than $M^2/2$. Indeed, first notice that

$$M^2 - \frac{u}{2}\varepsilon^2 \ln\left[u \ln\left(R\varepsilon^{-1}\right)\right] \geq M^2/2 \;\Leftrightarrow\; \frac{R\varepsilon^{-1}}{\ln(R\varepsilon^{-1})} \geq u, \qquad (8.54)$$

and then apply Result ($\diamondsuit$) above. Thus, we deduce from (8.51), (8.52), (8.53) and (8.54) that there exists $\kappa'' > 0$ such that

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{L}_q(R) \cap \mathcal{B}^r_{2,\infty}(R)} \mathbb{E}\left[\|f - \tilde{f}\|^2\right] \geq \kappa'' R^q M^{2-q} = \kappa'' u^{1-\frac{q}{2}} R^q \left(\varepsilon\sqrt{\ln(R\varepsilon^{-1})}\right)^{2-q}.$$

□

References

[1] Barron, A.R., Cohen, A., Dahmen, W., DeVore, R.A. Approximation and learning by greedy algorithms. *Annals of Statistics*, Vol. 36, No. 1, 64–94 (2008).

[2] Birgé, L. and Massart, P. Gaussian model selection. *Journal of the European Mathematical Society*, No. 3, 203–268 (2001).

[3] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, Vol. 37, No. 4, 1705–1732 (2009).

[4] Birgé, L. and Massart, P. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138, 33–73 (2007).

[5] Boucheron, S., Lugosi, G. and Massart, P. *Concentration inequalities with applications*. To appear.

[6] Bühlmann, P. and van de Geer, S. On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, 3, 1360–1392 (2009).

[7] Cohen, A., DeVore, R., Kerkyacharin, G. and Picard, D. Maximal spaces with given rate of convergence for thresholding algorithms. *Applied and Computational Harmonic Analysis*, 11, 167–191 (2001).

[8] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. *A distribution free theory of nonparametric regression*. Springer-Verlag, New-York (2002).

[9] Huang, C., Cheang, G.H.L. and Barron, A.R. Risk of penalized least squares, greedy selection and ℓ_1 penalization for flexible function libraries. Preprint (2008).

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

[11] Donoho, D.L. and Johnstone, I.M. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, Vol. 36, No. 3, 879–921 (1998).

[12] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. Least Angle Regression. *Annals of Statistics*, Vol. 32, No. 2, 407–499 (2004).

[13] Härdle, W., Kerkyacharin, G., Picard, D. and Tsybakov, A. *Wavelets, Approximation, and Statistical applications*. Springer-Verlag, Paris-Berlin (1998).

[14] Koltchinskii, V. Sparsity in penalized empirical risk minimization. *Annals of Statistics*, Vol. 45, No. 1, 7–57 (2009).

[15] Massart, P. *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics 1896, Springer Berlin-Heidelberg (2007).

[16] Rigollet, P. and Tsybakov, A. Exponential Screening and optimal rates of sparse estimation. Preprint (2010).

[17] Rivoirard, V. Nonlinear estimation over weak Besov spaces and minimax Bayes method. *Bernoulli*, Vol. 12, No. 4, 609–632 (2006).

[18] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, 58, 267–288 (1996).

[19] van de Geer, S.A. High dimensional generalized linear models and the Lasso. *Annals of Statistics*, Vol. 36, No. 2, 614–645 (2008).

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399