Electronic Journal of Statistics
Vol. 5 (2011) 669–687
ISSN: 1935-7524
DOI: 10.1214/11-EJS623

The Lasso as an ℓ_1-ball model selection procedure

Pascal Massart and Caroline Meynet

Département de Mathématiques
Université Paris-Sud
91405 Orsay, France

e-mail: pascal.massart@math.u-psud.fr; caroline.meynet@math.u-psud.fr

Abstract: 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 these variables, much less attention has been paid to the oracle inequalities for the Lasso involving the ℓ_1-norm of the target vector. Such inequalities proved in the literature show that, provided that the regularization parameter is properly chosen, the Lasso approximately mimics the deterministic Lasso. Some of them do not require any assumption at all, neither on the structure of the variables nor on the regression function. Our first purpose here is to provide a conceptually very simple result in this direction in the framework of Gaussian models with non-random regressors.

Our second purpose is to propose 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 is choosing automatically 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.

The oracle inequalities presented in this paper are obtained via the application of a general theorem of model selection among a collection of nonlinear models which is a direct consequence of the Gaussian concentration inequality. 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.

Keywords and phrases: Lasso, ℓ_1-oracle inequalities, model selection by penalization, ℓ_1-balls, generalized linear Gaussian model.

Received June 2010.

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 penalty to avoid it. One appropriate choice would be to use an ℓ_0-penalty by penalizing the number of non-zero coefficients $\hat{\theta}_j$ of $\hat{f}$ (see [5] for instance) so as to produce sparse estimators and interpretable models, but this minimization problem is non-convex and thus computationally unfeasible when the size of the dictionary becomes too large. On the contrary, ℓ_1-penalization leads to convex optimization, so there are efficient algorithms to approximate the solution of this problem even for high-dimensional data (see [12] for instance). Besides, by running these algorithms, one can notice that ℓ_1-penalty tends to produce sparse solutions and thus to behave like an ℓ_0-penalty. This phenomenon is due to the geometric properties of the ℓ_1-norm. For these reasons, ℓ_1-penalization and its associated solution, the so-called Lasso, have been widely used in the recent years as surrogate for ℓ_0-penalization.

In this paper, we look at the Lasso as an ℓ_1-regularization algorithm rather than a variable selection procedure. We analyze its performance by providing an ℓ_1-oracle inequality (see Theorem 3.1). It will prove that, provided that the regularization parameter is properly chosen, the Lasso performs almost as well as the deterministic Lasso, and what is noticeable is that this ℓ_1-result requires no assumption neither on the unknown target function nor on the variables ϕ_j of the dictionary (except simple normalization that we can always assume by considering $\phi_j/\|\phi_j\|$ instead of ϕ_j), contrary to the usual ℓ_0-oracle inequalities in the literature that are valid only under restrictive conditions.

The establishment of ℓ_1-oracle inequalities for the Lasso is not entirely new. In fact, a few authors such as Barron and al. [15], Bartlett and al. [2] or Rigollet and Tsybakov [20] have recently studied such ℓ_1-bounds, but there are some differences between their results and ours (see page 673 for more details). By stating Theorem 3.1, our aim is to add to the existing literature results on the ℓ_1-performance of the Lasso in simple, yet important, cases such as the fixed design Gaussian regression or the white noise models. We shall establish both a bound in probability and a bound in expectation, and our results shall be valid with no assumption neither on the target function nor on the variables of the dictionary (except simple normalization). Besides, we propose a method of analysis which is quite different from the methods used in the papers mentioned above. We shall derive our results from a fairly general model selection theorem for non linear models, interpreting ℓ_1-regularization as an ℓ_1-balls model selection criterion (see Appendix A). This approach will allow us to go one step further than the analysis of the Lasso for finite dictionaries and to deal with infinite dictionaries in various situations.

In the second part of this paper, we shall thus focus on infinite countable dictionaries. While infinite dictionaries are more and more used in many applications such as micro-array data analysis or signal reconstruction, it proves difficult to calibrate the regularization parameter of the Lasso and thus to establish good theoretical results on the performance of this estimator for such dictionaries. To solve this problem, we propose a procedure that provides an optimal level of truncation of the whole infinite dictionary as well as an efficient

estimator in the linear span of this optimal finite subdictionary. For orthonormal dictionaries, this estimator is nothing else than a soft-thresholding estimator with an adaptive threshold.

The article is organized as follows. The framework and statistical problem are introduced in Section 2. In Section 3, we study the case of finite dictionaries and analyze the performance of the Lasso as an ℓ_1-regularization algorithm by providing an ℓ_1-oracle inequality showing that the Lasso estimator works almost as well as the deterministic Lasso provided that the regularization parameter is chosen large enough. In section 4, we look at the case of infinite countable dictionaries and introduce a procedure based on Lasso type penalization combined with an additional complexity penalty that produces an efficient selected Lasso estimator. The explanation of the key idea that enables us to derive all our oracle inequalities from a single general model selection theorem is postponed until Appendix A. Finally, the oracle inequalities are proved in Appendix B.

2. General framework and statistical problem

Let $\mathbb{H}$ be a separable Hilbert space equipped with a scalar product $\langle .,. \rangle$ and its associated norm $\|.\|$. The statistical problem we consider is to estimate 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 $(W(h))_{h\in\mathbb{H}}$ is an isonormal process, that is to say a centered Gaussian process with covariance given by $\mathbb{E}[W(g)W(h)] = \langle g, h \rangle$ for all $g, h \in \mathbb{H}$. This framework is convenient to cover both finite-dimensional models, such as the classical fixed design Gaussian regression model, and infinite-dimensional models, such as the Gaussian white noise model (see [17] for details on these models).

To solve the statistical problem (2.1), we shall introduce 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 consider estimators $\hat{f} = \hat{\theta}.\phi := \sum_{j,\, \phi_j \in \mathcal{D}} \hat{\theta}_j \, \phi_j$ 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

$$\gamma(h) := -2Y(h) + \|h\|^2. \tag{2.2}$$

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, we shall rather consider estimators solution of a penalized risk minimization problem,

$$\hat{f} \in \arg\min_h \gamma(h) + \mathrm{pen}(h), \tag{2.3}$$

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

In Sections 3 and 4.3, we shall consider ℓ_1-penalization, that is to say pen$(h) \propto \inf\{\|\theta\|_1 = \sum_{j,\,\phi_j\in\mathcal{D}} |\theta_j| \text{ such that } h = \theta.\phi\}$, while we shall suggest in Section 4 a penalty pen(h) combination of an ℓ_1-penalty and a complexity penalty.

3. The Lasso for finite dictionaries

In this section, we provide an ℓ_1-oracle inequality satisfied by the Lasso in the case of finite dictionaries.

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 (2.3) with a penalty pen(h) proportional to $\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}$. This estimator is the so-called Lasso estimator $\hat{f}_p$ defined by

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

where $\lambda_p > 0$ is some regularization parameter and $\gamma(h)$ is defined by (2.2).

3.2. An ℓ_1-oracle inequality

Let us now state the main result of this section. This ℓ_1-oracle inequality highlights the fact that, provided that the regularization parameter λ_p is properly chosen, the Lasso, which is the solution of the ℓ_1-penalized empirical risk minimization problem, behaves as well as the deterministic Lasso, that is to say the solution of the ℓ_1-penalized true risk minimization problem, up to an error term of order $O(\varepsilon^2)$ where $O(.)$ depends on the complexity of the dictionary.

Theorem 3.1. *Assume that* $\max_{j=1,\ldots,p} \|\phi_j\| \le 1$ *and that*

$$\lambda_p \ge 4\varepsilon\left(\sqrt{\ln p} + 1\right). \tag{3.3}$$

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 $1 - 3.4\,\mathrm{e}^{-z}$,

$$\|f-\hat{f}_p\|^2+\lambda_p\|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} \leq C\left[\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2+\lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right)+\lambda_p\,\varepsilon(1+z)\right]. \tag{3.4}$$

Integrating (3.4) with respect to z leads to the following ℓ_1-oracle inequality in expectation,

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

Remark 3.2. These last years, the Lasso has essentially been developed as an approach to sparse recovery based on convex optimization and thus the main focus on this estimator has been on the establishment of ℓ_0-oracle inequalities so as to study its performance as a variable selection procedure. Here, Theorem 3.1 does not take into account sparsity and rather provides information about the performance of the Lasso as an ℓ_1-regularization algorithm by providing ℓ_1-oracle inequalities satisfied by this estimator. Notice that the ℓ_1-oracle inequalities of Theorem 3.1 are valid for regularization parameters of the same order (3.3) as the usual regularization parameters considered for the establishment of ℓ_0-oracle inequalities (see [4] among others). Let us also stress that, contrary to the ℓ_0-results that require some restrictive assumptions on the dictionary and that are interesting only if the target function can be well approximated by a sparse function in the linear span of the dictionary, the ℓ_1-oracle inequalities (3.4) and (3.5) are established with no assumption neither on the target function 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. This shows that, whereas one can not be sure whether the conditions for the Lasso to be a good variable selection procedure are fulfilled or not, one is always guaranteed that the Lasso achieves high-performance as regards ℓ_1-regularization.

In fact, ℓ_1-oracle inequalities of the same type as (3.4) or (3.5) have already been studied by a few authors such as Barron and al. [15], Bartlett and al. [2] or Rigollet and Tsybakov [20], but all these results present dissimilarities with Theorem 3.1. Let us have a look at these differences.

In [20], Rigollet and Tsybakov are proposing an oracle inequality for the Lasso similar to (3.4) which is valid under the same assumption as the one of Theorem 3.1, 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 (3.5). Indeed, 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 (3.4) appears as an additive constant outside the infimum term so that the bound in probability (3.4) can easily be integrated with respect to z, which leads

to the bound in expectation (3.5). Besides, the lower bound of the regularization parameter λ_p proposed by Tsybakov and Rigollet ($\lambda_p \geq \sqrt{8(1+z/\ln p)}\,\varepsilon\sqrt{\ln p}$) 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 (3.4) satisfied with high probability by an estimator $\hat{f}_p = \hat{f}_p(\lambda_p)$ independent of the level of confidence of this probability.

As regards Bartlett and al. [2], they have obtained an oracle inequality for the Lasso of the same type as (3.4) in the context of linear regression. Nonetheless, they have considered the case of random design $(X,Y) \in \mathbb{R}^d \times \mathbb{R}$ rather than our setting with fixed design and Gaussian noise. Therefore, they have to overcome rather substantial (and interesting) difficulties in the analysis of empirical processes involved in the problem. In particular, their method of analysis requires a uniform concentration phenomenon that forces them to make strong assumptions, namely that both X and Y are bounded almost surely by a constant independent of n. Moreover, they get a lower bound on the regularization parameter with an extra ln-factor compared to (3.3).

However, the oracle inequality of Theorem 3.1 is proved with undetermined constant C whereas the ℓ_1-oracle inequalities in both [20] and [2] are sharp, i.e. with $C = 1$.

Barron and al. [15] have also studied risk bounds for ℓ_1-penalized estimators in the case of random design. Rather than assuming that Y is bounded as it is done by Bartlett and al., they make the assumption that the errors satisfy some Bernstein's moment condition, but on the other hand, they assume that the target function is bounded by a constant and the risk bound they provide is not satisfied by the Lasso itself but only by a truncated Lasso estimator.

The proof of Theorem 3.1 is detailed in Appendix B and we refer the reader to Appendix A 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 (3.4) and (3.5) are then deduced from a general model selection theorem stated as Theorem A.1 in Appendix A. Let us point out that this approach will allow us to go one step further than the analysis of the Lasso for finite dictionaries and to deal with infinite dictionaries as we shall see in Section 4.

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. So it is desirable to find competitive estimators for such infinite dimensional problems, but (except in rare situations when the variables have a specific structure: see Remark 4.3 on neural networks) it proves very difficult to establish good theoretical results on the performance of the Lasso solution over an infinite dictionary. Indeed, when

considering a finite dictionary of size p, Theorem 3.1 guarantees that for a regularization parameter greater than a certain quantity depending on the size p, the corresponding Lasso estimator achieves good performance results, but for an infinite dictionary there is no size p and thus no lower bound on the regularization parameter to guarantee good performance of the corresponding Lasso estimator. Our goal here is to propose a procedure to calibrate the regularization parameter by providing an optimal size $\hat{p}$ in a sense described below.

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 important case of dictionaries of wavelets.

For the particular case of an orthonormal dictionary where the truncated Lasso estimators are nothing else than soft-thresholding estimators with a fixed threshold, the selected Lasso estimator is a soft-thresholding estimator with an adaptive threshold which is automatically chosen by the algorithm constructing this estimator. So, our procedure provides a new contribution to the crucial problem of choosing the threshold when working with soft-thresholding estimators.

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. To solve this problem, we recall that 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 empirical 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

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

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,

$$\hat{f}_p := \hat{f}_p(\lambda_p) = \underset{h\in\mathcal{L}_1(\mathcal{D}_p)}{\arg\min}\ \gamma(h) + \lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}, \tag{4.2}$$

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 $(\hat{f}_p)_{p\in\mathbb{N}^*}$. More precisely, let us denote by Λ the set of dyadic integers,

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

and define

$$\begin{aligned}\hat{p} &= \underset{p\in\Lambda}{\arg\min}\left[\gamma(\hat{f}_p) + \lambda_p\|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)} + \mathrm{pen}(p)\right]\\ &= \underset{p\in\Lambda}{\arg\min}\left[\underset{h\in\mathcal{L}_1(\mathcal{D}_p)}{\arg\min}\left(\gamma(h) + \lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) + \mathrm{pen}(p)\right],\end{aligned} \tag{4.3}$$

where $\mathrm{pen}(p)$ penalizes the size p of the truncated dictionary $\mathcal{D}_p$ for all $p \in \Lambda$. Then, the final estimator we consider is the selected Lasso estimator $\hat{f}_{\hat{p}}$. From (4.3) 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 all $p \in \Lambda$ 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 complexity penalization. We also see from (4.3) 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. 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 a complexity penalty $\mathrm{pen}(p)$ that does not grow too fast with p. In the sequel, we shall consider a penalty $\mathrm{pen}(p) \propto \ln p$. So, taking a dyadic truncation $\mathcal{D}_p := \{\phi_1, \ldots, \phi_p\} = \{\phi_1, \ldots, \phi_{2^J}\}$ of the dictionary $\mathcal{D}$ enables to get a complexity penalty $\mathrm{pen}(p) \propto \ln p = J\ln 2$ which grows linearly at each step J of the algorithm, thus leading to a more efficient algorithm. That is why we have chosen to work with a dyadic subsequence of dictionaries rather than with other subsequences.

Although 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, notice that this estimator remains well-defined and can also be interesting for estimation in the case of large finite dictionaries. Indeed, if we consider a finite dictionary of

size p_0, then it can be advantageous to work with the selected Lasso estimator $\hat{f}_{\hat{p}}$ rather than with the Lasso estimator $\hat{f}_{p_0}$ since the definition of $\hat{f}_{\hat{p}}$ guarantees that $\hat{f}_{\hat{p}}$ always makes a better tradeoff between approximation, ℓ_1-regularization and sparsity than $\hat{f}_{p_0}$. Besides, $\hat{f}_{\hat{p}}$ is always sparser than $\hat{f}_{p_0}$ since $\hat{p} \leq p_0$. In particular, notice that $\hat{f}_{\hat{p}}$ and $\hat{f}_{p_0}$ coincide when $\hat{p} = p_0$.

4.2. An oracle inequality

By applying the same general model selection theorem (Theorem A.1) as for the establishment of Theorem 3.1, 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 (3.3) while a convenient choice for the complexity 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* $p \in \Lambda$,

$$\lambda_p = c_1\varepsilon\left(\sqrt{\ln p} + 1\right), \qquad \mathrm{pen}(p) = c_2\varepsilon^2 \ln p, \tag{4.4}$$

where $c_1 \geq 4$ *and* $c_2 > c_1/\sqrt{\ln 2}$.
Consider the corresponding selected Lasso estimator $\hat{f}_{\hat{p}}$ *defined by* (4.3).
Then, there exists an absolute constant $C > 0$ *such that*

$$\begin{aligned}&\mathbb{E}\left[\|f - \hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p})\right]\\ &\leq C\left[\inf_{p\in\Lambda}\left(\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\|f-h\|^2 + \lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) + \mathrm{pen}(p)\right) + \varepsilon^2\right].\end{aligned} \tag{4.5}$$

4.3. The Lasso for particular infinite uncountable dictionaries

As explained at the beginning of this section, it is generally very difficult to establish good theoretical results on the performance of the Lasso for infinite dictionaries. Yet, let us just point out here that it can be easier to prove such results for some particular infinite dictionaries whose structure is nice enough. For example, it is the case for neural networks in the fixed design Gaussian regression models. Recall that a neural network is a real-valued function defined on $\mathbb{R}^d$ belonging to the linear span of the dictionary $\mathcal{D} = \{\phi_{a,b}\,;\ a \in \mathbb{R}^d, b \in \mathbb{R}\}$ where

$$\phi_{a,b} :\ \mathbb{R}^d \mapsto \mathbb{R}, \quad x \mapsto \mathbb{1}_{\{\langle a,x\rangle + b > 0\}}. \tag{4.6}$$

Now, given a training sequence $\{(x_1, Y_1), \ldots, (x_n, Y_n)\}$, if we assume that $Y_i = f(x_i) + \sigma\xi_i$ for all $i = 1, \ldots, n$, then the Lasso estimator over the set of neural network regression function estimators in $\mathcal{L}_1(\mathcal{D})$ is defined by

$$\hat{f} := \hat{f}(\lambda) = \operatorname*{arg\,min}_{h\in\mathcal{L}_1(\mathcal{D})} \|Y - h\|^2 + \lambda\|h\|_{\mathcal{L}_1(\mathcal{D})}, \tag{4.7}$$

where $\lambda > 0$ is a regularization parameter, $\|Y-h\|^2 := \sum_{i=1}^n (Y_i - h(x_i))^2/n$ is the empirical risk of h and $\mathcal{L}_1(\mathcal{D})$ is the linear span of $\mathcal{D}$ equipped with the ℓ_1-norm $\|h\|_{\mathcal{L}_1(\mathcal{D})} := \inf\{\|\theta\|_1 = \sum_{a\in\mathbb{R}^d, b\in\mathbb{R}} |\theta_{a,b}|,\ h = \theta.\phi = \sum_{a\in\mathbb{R}^d, b\in\mathbb{R}} \theta_{a,b}\,\phi_{a,b}\}$.

Despite the fact that the dictionary $\mathcal{D}$ is infinite uncountable, we are able to establish an ℓ_1-oracle inequality satisfied by the Lasso which is similar to the one provided in Theorem 3.1 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 above where we considered arbitrary infinite countable dictionaries. The following ℓ_1-oracle inequality is once again a direct application of the general model selection Theorem A.1 already used to prove both Theorem 3.1 and Theorem 4.2.

Theorem 4.3. *Assume that*

$$\lambda \geq \kappa\sigma\sqrt{\frac{d}{n}} \tag{4.8}$$

for some absolute constant $\kappa > 0$ large enough and consider the corresponding Lasso estimator $\hat{f}$ defined by (4.7).
Then, there exists an absolute constant $C > 0$ such that

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

Appendix A: 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 3.1, Theorem 4.2 and Theorem 4.3 as an application of a single general model selection theorem, and by stating and proving this general theorem. We keep the notations introduced in Section 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 $\hat{f}$ 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.

$$\hat{m} = \left\lceil \frac{\|\hat{f}\|_{\mathcal{L}_1(\mathcal{D})}}{\varepsilon} \right\rceil. \tag{A.1}$$

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

$$\gamma(\hat{f}) + \mathrm{pen}(\hat{m}) \leq \inf_{m\geq 1}\left(\inf_{h\in S_m}\gamma(h) + \mathrm{pen}(m)\right) + \rho \tag{A.2}$$

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 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 A.1, is a restricted version of an even more general model selection theorem that the interested reader can find in [17], Theorem 4.18.

Theorem A.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 $m \in \mathcal{M}$,*

$$\Delta_m := \mathbb{E}\left[\sup_{h\in S_m} W(h)\right], \tag{A.3}$$

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 $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geq 2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m\varepsilon x_m}\right). \tag{A.4}$$

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 $1 - \Sigma\, \mathrm{e}^{-z}$,

$$\begin{aligned} &\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) \\ &\leq C(K) \left[\inf_{m \in \mathcal{M}} \left(\inf_{h \in S_m} \|f - h\|^2 + \mathrm{pen}(m) + \rho_m \right) + (1+z)\varepsilon^2 \right]. \end{aligned} \tag{A.5}$$

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

$$\begin{aligned} &\mathbb{E}\left[\|f - \hat{f}\|^2 + \mathrm{pen}(\hat{m}) \right] \\ &\leq C(K) \left[\inf_{m \in \mathcal{M}} \left(\inf_{h \in S_m} \|f - h\|^2 + \mathrm{pen}(m) + \rho_m \right) + (1+\Sigma)\varepsilon^2 \right]. \end{aligned} \tag{A.6}$$

Appendix B: Proof of the ℓ_1-oracle inequalities

Deriving Theorem 3.1, Theorem 4.2 and Theorem 4.3 from Theorem A.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 (A.4) and then to apply the conclusion of Theorem A.1.

B.1. Proof of Theorem 3.1

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 (A.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 (A.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

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

Recalling that W is isonormal (see (2.1)), we have $\mathrm{Var}\left[W(\phi_j)\right] = \|\phi_j\|^2 \leq 1$ 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 [17] 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 (B.1) that

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

Now, choose the weights of the form $x_m = \gamma m$ where $\gamma > 0$ is specified below. Then, $\sum_{m\ge1} e^{-x_m} = 1/\left(e^{\gamma}-1\right) := \Sigma_\gamma < +\infty$.

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

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

as soon as

$$\lambda_p \ge 4\varepsilon\left(\sqrt{\ln p}+1\right). \tag{B.3}$$

For such values of λ_p, condition (A.4) on the penalty function is satisfied and we may apply Theorem A.1 with $\mathrm{pen}(m) = \lambda_p m\varepsilon$ and $\rho = \lambda_p\varepsilon$. Taking into account the definition of $\hat{m}$ at (A.1) and noticing that $\varepsilon^2 \le \lambda_p\varepsilon/4$ for λ_p satisfying (B.3), we get from (A.5) that there exists some $C > 0$ such that for all $z > 0$, with probability larger than $1 - \Sigma_\gamma\,\mathrm{e}^{-z} \ge 1 - 3.4\,\mathrm{e}^{-z}$,

$$\begin{aligned}
&\|f-\hat{f}_p\|^2 + \lambda_p\|\hat{f}_p\|_{\mathcal{L}_1(\mathcal{D}_p)}\\
&\le C\left[\inf_{m\ge1}\left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\le m\varepsilon}\|f-h\|^2 + \lambda_p m\varepsilon\right) + \lambda_p\varepsilon + (1+z)\varepsilon^2\right]\\
&\le C\left[\inf_{m\ge1}\left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\le m\varepsilon}\|f-h\|^2 + \lambda_p m\varepsilon\right) + \lambda_p\varepsilon(1+z)\right].
\end{aligned} \tag{B.4}$$

Finally, to get the desired bound (3.4), just notice that for all $g \in \mathcal{L}_1(\mathcal{D}_p)$, by considering $m_g = \lceil \|g\|_{\mathcal{L}_1(\mathcal{D}_p)}/\varepsilon\rceil \in \mathbb{N}^*$ so that $\|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \le m_g\varepsilon$, we have

$$\begin{aligned}
\inf_{m\ge1}\left(\inf_{\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\le m\varepsilon}\|f-h\|^2 + \lambda_p m\varepsilon\right) &\le \|f-g\|^2 + \lambda_p\, m_g\,\varepsilon\\
&\le \|f-g\|^2 + \lambda_p\|g\|_{\mathcal{L}_1(\mathcal{D}_p)} + \lambda_p\varepsilon,
\end{aligned} \tag{B.5}$$

and combining (B.4) with (B.5) leads to

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

Similarly, we get the risk bound (3.5) from (A.6). □

B.2. 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.

$$\hat{m} = \left\lceil \frac{\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})}}{\varepsilon} \right\rceil. \tag{B.6}$$

Let $\alpha = 1 - c_1/(c_2\sqrt{\ln 2})$. From (B.6) and (4.4), using the fact that for all $p \in \Lambda$, $\sqrt{\ln p} \leq (\ln p)/\sqrt{\ln 2}$, the definitions of α and $\hat{f}_{\hat{p}}$ and the fact that $\mathcal{L}_1(\mathcal{D}_p) = \bigcup_{m\in\mathbb{N}^*} S_{m,p}$, we get that

$$\begin{aligned}
&\gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \lambda_{\hat{p}}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + c_1\varepsilon^2\left(\sqrt{\ln\hat{p}} + 1\right) + \alpha c_2\varepsilon^2 \ln\hat{p} \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \left(\frac{c_1}{c_2\sqrt{\ln 2}} + \alpha\right) c_2\varepsilon^2 \ln\hat{p} + c_1\varepsilon^2 \\
&\leq \gamma(\hat{f}_{\hat{p}}) + \lambda_{\hat{p}}\|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p}) + c_1\varepsilon^2 \\
&\leq \inf_{p\in\Lambda}\left[\inf_{h\in\mathcal{L}_1(\mathcal{D}_p)}\left(\gamma(h) + \lambda_p\|h\|_{\mathcal{L}_1(\mathcal{D}_p)}\right) + \mathrm{pen}(p)\right] + c_1\varepsilon^2 \\
&\leq \inf_{p\in\Lambda}\left[\inf_{m\in\mathbb{N}^*}\left(\inf_{h\in S_{m,p}}\gamma(h) + \lambda_p m\varepsilon\right) + \mathrm{pen}(p)\right] + c_1\varepsilon^2 \\
&\leq \inf_{(m,p)\in\mathcal{M}}\left[\inf_{h\in S_{m,p}}\gamma(h) + \lambda_p m\varepsilon + \mathrm{pen}(p)\right] + c_1\varepsilon^2,
\end{aligned}$$

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 $\rho_p := (1-\alpha)\,\mathrm{pen}(p) + c_1\varepsilon^2$ (notice that thanks to the assumption $c_2 > c_1/\sqrt{\ln 2}$, we have $\alpha \in]0,1[$, so $\mathrm{pen}(m,p) > 0$ and $\rho_p > 0$). 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 A.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.

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

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

Moreover, for all $(m,p) \in \mathcal{M}$, we can prove similarly as (B.2) 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, define $K = c_1[\sqrt{2}(2 + (c_1 - 2)^{-1})]^{-1}$ (notice that $K > 1$ thanks to the assumption $c_1 \geq 4$), $\gamma = (1 - \sqrt{\ln 2})/K > 0$ and $\beta = (c_2\alpha)/(c_1 K) > 0$. Taking into account these definitions and using the inequality $2\sqrt{ab} \leq \eta a + \eta^{-1} b$ with $\eta = (c_1 - 2)^{-1}$, $a = \Delta_{m,p}$ and $b = \varepsilon x_{m,p}$, we get that

$$\begin{aligned}&2K\varepsilon\left(\Delta_{m,p} + \varepsilon x_{m,p} + \sqrt{\Delta_{m,p}\varepsilon x_{m,p}}\right) \\ &\leq K\varepsilon\left(\left(2 + (c_1 - 2)^{-1}\right)\Delta_{m,p} + c_1 x_{m,p}\varepsilon\right) \\ &\leq c_1\varepsilon^2\left(m\sqrt{\ln p} + m\sqrt{\ln 2} + K\gamma m + K\beta \ln p\right) \\ &\leq c_1\varepsilon^2\left(m\left(\sqrt{\ln p} + 1\right) + \frac{c_2\alpha}{c_1}\ln p\right) \\ &\leq \lambda_p m\varepsilon + \alpha\,\mathrm{pen}(p) \\ &= \mathrm{pen}(m,p).\end{aligned}$$

Thus, condition (A.4) is satisfied and we can apply Theorem A.1 with $\mathrm{pen}(m,p) = \lambda_p m\varepsilon + \alpha\,\mathrm{pen}(p)$ and $\rho_p = (1-\alpha)\,\mathrm{pen}(p) + c_1\varepsilon^2$, which leads to:

$$\begin{aligned}&\mathbb{E}\left[\|f - \hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p})\right] \\ &\leq C\left[\inf_{(m,p)\in\mathcal{M}}\left(\inf_{h\in S_{m,p}}\|f - h\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p)\right) + \left(c_1 + 1 + \Sigma_{\gamma,\beta}\right)\varepsilon^2\right] \\ &\leq C\left[\inf_{(m,p)\in\mathcal{M}}\left(\inf_{h\in S_{m,p}}\|f - h\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p)\right) + \varepsilon^2\right], \qquad \text{(B.7)}\end{aligned}$$

where $C > 0$ denotes some absolute 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 $g \in \mathcal{L}_1(\mathcal{D}_{p_0})$,

and consider $m_g = \lceil \|g\|_{\mathcal{L}_1(\mathcal{D}_{p_0})}/\varepsilon \rceil \in \mathbb{N}^*$ so that $g \in S_{m_g,p_0}$. Then,

$$\begin{aligned}
&\inf_{(m,p)\in\mathcal{M}} \left(\inf_{h\in S_{m,p}} \|f-h\|^2 + \lambda_p m\varepsilon + \mathrm{pen}(p) \right) \\
&\leq \|f-g\|^2 + \lambda_{p_0} m_g \varepsilon + \mathrm{pen}(p_0) \\
&\leq \|f-g\|^2 + \lambda_{p_0} \left(\|g\|_{\mathcal{L}_1(\mathcal{D}_{p_0})} + \varepsilon \right) + \mathrm{pen}(p_0) \\
&\leq \|f-g\|^2 + \lambda_{p_0} \|g\|_{\mathcal{L}_1(\mathcal{D}_{p_0})} + \left(\frac{c_1}{c_2\sqrt{\ln 2}} + 1 \right) \mathrm{pen}(p_0) + c_1\varepsilon^2. \qquad \text{(B.8)}
\end{aligned}$$

So, we deduce from (B.7) and (B.8) that there exists $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \right] \\
&\leq C \left[\inf_{p\in\Lambda} \left(\inf_{g\in\mathcal{L}_1(\mathcal{D}_p)} \left(\|f-g\|^2 + \lambda_p \|g\|_{\mathcal{L}_1(\mathcal{D}_p)} \right) + \mathrm{pen}(p) \right) + \varepsilon^2 \right]. \qquad \text{(B.9)}
\end{aligned}$$

Finally, let us notice that from the fact that $\alpha \in]0,1[$ and from (B.6), we have

$$\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}} \|\hat{f}_{\hat{p}}\|_{\mathcal{L}_1(\mathcal{D}_{\hat{p}})} + \mathrm{pen}(\hat{p}) \right] \leq \frac{1}{\alpha}\,\mathbb{E}\left[\|f-\hat{f}_{\hat{p}}\|^2 + \lambda_{\hat{p}}\hat{m}\varepsilon + \alpha\,\mathrm{pen}(\hat{p}) \right]. \qquad \text{(B.10)}$$

Combining (B.9) with (B.10) leads to the result. □

B.3. Proof of Theorem 4.3

Let us recall that, for $t > 0$, the t-packing number $\mathcal{N}(t,\mathcal{G},N)$ of a set $\mathcal{G}$ with respect to a norm $N(.)$ is the maximal $m \in \mathbb{N}^*$ such that there exist $g_1,\ldots,g_m \in \mathcal{G}$ with $N(g_i - g_j) \geq t$ for all $1 \leq i < j \leq m$, while the t-entropy number is defined by $H(t,\mathcal{G},N) := \ln(\mathcal{N}(t,\mathcal{G},N))$.

Lemma B.1. *Let $t > 0$ and let $\mathcal{D} = \{\phi_{a,b} : a \in \mathbb{R}^d, b \in \mathbb{R}\}$ be a dictionary of neurons where $\phi_{a,b}$ is defined by (4.6). Then,*

$$\int_0^1 \sqrt{H(t,\mathcal{D},\|.\|)}\; dt \leq C\sqrt{d+1},$$

where $C > 0$ is an absolute constant ($C \geq 22$ is convenient).

Proof. The result just comes from the fact that $\mathcal{D}$ is a subset of the boolean n-cube with Vapnik-Chervonenkis dimension $d+1$. Indeed, for all $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, let us denote by $A_{a,b}$ the affine half-space of $\mathbb{R}^d$ defined by $A_{a,b} = \{x \in \mathbb{R}^d : \langle a,x\rangle + b > 0\}$ and consider the associated VC class $\mathcal{A} = \{A_{a,b}, a \in \mathbb{R}^d, b \in \mathbb{R}\}$ which is of dimension $d+1$. Also introduce

$$\mathcal{A}(x_1^n) := \left\{ \left(\mathbb{1}_{\{x_1\in A\}},\ldots,\mathbb{1}_{\{x_n\in A\}} \right), A \in \mathcal{A} \right\} \subset \{0,1\}^n$$

equipped with the ℓ_1-norm $\|.\|_{1,n}$ defined by

$$\|u\|_{1,n} = \frac{1}{n}\sum_{i=1}^{n}|u_i|$$

for all $u = (u_1, \ldots, u_n) \in \mathcal{A}(x_1^n)$. Then, for all $\phi_{a,b} \in \mathcal{D}$, (4.6) implies that $\phi_{a,b} = \mathbb{1}_{A_{a,b}}$ and $\|\phi_{a,b}\| = \sqrt{\|u_{a,b}\|_{1,n}}$ where $u_{a,b} = (\mathbb{1}_{\{x_1 \in A_{a,b}\}}, \ldots, \mathbb{1}_{\{x_n \in A_{a,b}\}}) \in \mathcal{A}(x_1^n)$. Thus, we get that

$$H\left(t, \mathcal{D}, \|.\|\right) \leq H\left(\sqrt{t}, \mathcal{A}(x_1^n), \|.\|_{1,n}\right).$$

Moreover, we can easily get from the upper bound of the entropy for a VC class of dimension $d+1$ provided by Haussler in [14] that

$$\int_0^1 \sqrt{H\left(\sqrt{t}, \mathcal{A}(x_1^n), \|.\|_{1,n}\right)}\, dt \leq C\sqrt{d+1},$$

where $C > 0$ is an absolute constant ($C \geq 22$ is convenient), hence the result. □

Proof of Theorem 4.3 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 Appendix A 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 (A.4) and to apply the conclusion of Theorem A.1.

Let $h \in S_m$. 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 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 [17] for instance), we have

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

where $\alpha^2 = \sup_{a,b}\mathbb{E}[W^2(\phi_{a,b})] = \sup_{a,b}\|\phi_{a,b}\|^2 = \sup_{a,b}(\sum_{i=1}^n \phi_{a,b}^2(x_i)/n) \leq 1$ from (4.6). So, $\alpha \leq 1$ and we get from Lemma B.1 that there exists $c > 0$ ($c \geq 12 \times 22 = 264$ is convenient) such that

$$\Delta_m \leq 12(m\varepsilon+\delta)\int_0^1\sqrt{H\left(t,\mathcal{D},\|.\|\right)}\ dt \leq c(m\varepsilon+\delta)\sqrt{d+1}. \tag{B.11}$$

Now if we choose weights $x_m = cm$, then $\sum_{m\geq 1} e^{-x_m} := \Sigma_c < +\infty$, and using the inequality $2\sqrt{ab} \leq a+b$ we get from (B.11) that for all $K > 1$,

$$\begin{aligned} 2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m}\right) &\leq 3K\varepsilon\left(\Delta_m + \varepsilon x_m\right) \\ &\leq 3K\varepsilon\left(c(m\varepsilon+\delta)\sqrt{d+1} + cm\varepsilon\right) \\ &\leq 3cK\varepsilon(m\varepsilon+\delta)\left(\sqrt{d+1}+1\right) \\ &\leq 3c(\sqrt{2}+1)K\varepsilon(m\varepsilon+\delta)\sqrt{d}. \end{aligned}$$

Since this inequality is true for all $\delta > 0$, we get when δ tends to 0 that there exists $\kappa > 0$ ($\kappa = 3c(\sqrt{2}+1)K$) such that

$$2K\varepsilon\left(\Delta_m + \varepsilon x_m + \sqrt{\Delta_m \varepsilon x_m}\right) \leq \kappa m\varepsilon^2\sqrt{d} \leq \lambda m\varepsilon$$

as soon as

$$\lambda \geq \kappa\varepsilon\sqrt{d}. \tag{B.12}$$

For such values of λ, condition (A.4) on the penalty function is satisfied and me may apply Theorem A.1 with $\text{pen}(m) = \lambda m\varepsilon$ and $\rho = \lambda\varepsilon$ for all $m \geq 1$. We end the proof similarly as the one of Theorem 3.1. □

Acknowledgements

We thank the associate editor and the referees for their helpful comments on the paper, especially about Section 4.3 on neural networks.

References

[1] Barron, A.R., Cohen, A., Dahmen, W. and DeVore, R.A. Approximation and learning by greedy algorithms. *Annals of Statistics*, Vol. 36, No. 1, 64–94 (2008). MR2387964

[2] Bartlett, P.L., Mendelson, S. and Neeman, J. ℓ_1-regularized linear regression: persistence and oracle inequalities. Preprint (2009).

[3] Birgé, L. and Massart, P. Gaussian model selection. *Journal of the European Mathematical Society*, No. 3, 203–268 (2001). MR1848946

[4] 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). MR2533469

[5] Birgé, L. and Massart, P. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138, 33–73 (2007). MR2288064

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

[7] Bühlmann, P. and van de Geer, S. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, Vol. 3, 1360–1392 (2009). MR2576316

[8] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, Vol. 1, 169–194 (2007). MR2312149

[9] 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). MR1848302

[10] DeVore, R.A. and Lorentz, G.G. *Constructive Approximation*. Springer-Verlag, Berlin (1993). MR1261635

[11] Donoho, D.L. and Johnstone, I.M. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, Vol. 36, No. 3, 879–921 (1998). MR1635414

[12] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. Least Angle Regression. *Annals of Statistics*, Vol. 32, No. 2, 407–499 (2004). MR2060166

[13] Härdle, W., Kerkyacharin, G., Picard, D. and Tsybakov, A. *Wavelets, approximation, and statistical applications*. Springer-Verlag, Paris-Berlin (1998). MR1618204

[14] Haussler, D. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory Series A*, 69, 217–232 (1995). MR1313896

[15] 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). MR2711791

[16] Koltchinskii, V. Sparsity in penalized empirical risk minimization. *Annals of Statistics*, Vol. 45, No. 1, 7–57 (2009). MR2500227

[17] 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). MR2319879

[18] Massart, P. and Meynet, C. An ℓ_1-oracle inequality for the Lasso. *arXiv*, 1007.4791 (2010).

[19] Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, Vol. 37, No. 1, 246–270 (2009). MR2488351

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

[21] Rivoirard, V. Nonlinear estimation over weak Besov spaces and minimax Bayes method. *Bernoulli*, Vol. 12, No. 4, 609–632 (2006). MR2248230

[22] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, 58, 267–288 (1996). MR1379242

[23] van de Geer, S.A. High dimensional generalized linear models and the Lasso. *Annals of Statistics*, Vol. 36, No. 2, 614–645 (2008). MR2396809

[24] Zhang, C.H. and Huang, J. Model-selection consistency of the Lasso in high-dimensional linear regression. *Annals of Statistics*, Vol. 36, 1567–1594 (2008). MR2435448

[25] Zhao, P. and Yu, B. On model selection consistency of Lasso. *J. Machine Learning Res.*, 7, 2541–2567 (2007). MR2274449