Elastic-net regularization in learning theory

Citation: De Mol, Christine, Ernesto De Vito, and Lorenzo Rosasco. "Elastic-Net Regularization in Learning Theory." Journal of Complexity 25, no. 2 (April 2009): 201–230. © 2009 Elsevier Inc.

As Published: http://dx.doi.org/10.1016/j.jco.2009.01.002

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/96186

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Journal of Complexity 25 (2009) 201–230

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Elastic-net regularization in learning theory

Christine De Mol [a], Ernesto De Vito [b,c], Lorenzo Rosasco [d,e,*]

[a] *Department of Mathematics and ECARES, Université Libre de Bruxelles, Campus Plaine CP 217, Bd du Triomphe, 1050 Brussels, Belgium*

[b] *Dipartimento di Scienze per l'Architettura, Università di Genova, Stradone Sant'Agostino, 37, 16123, Genova, Italy*

[c] *INFN, Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy*

[d] *Center for Biological and Computational Learning, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, United States*

[e] *Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, Via Dodecaneso 35,16146 Genova, Italy*

ARTICLE INFO

Article history:
Received 4 August 2008
Accepted 9 January 2009
Available online 30 January 2009

Keywords:
Learning
Regularization
Sparsity
Elastic net

ABSTRACT

Within the framework of statistical learning theory we analyze in detail the so-called elastic-net regularization scheme proposed by Zou and Hastie [H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B, 67(2) (2005) 301–320] for the selection of groups of correlated variables. To investigate the statistical properties of this scheme and in particular its consistency properties, we set up a suitable mathematical framework. Our setting is random-design regression where we allow the response variable to be vector-valued and we consider prediction functions which are linear combinations of elements (*features*) in an infinite-dimensional dictionary. Under the assumption that the regression function admits a sparse representation on the dictionary, we prove that there exists a particular "*elastic-net representation*" of the regression function such that, if the number of data increases, the elastic-net estimator is consistent not only for prediction but also for variable/feature selection. Our results include finite-sample bounds and an adaptive scheme to select the regularization parameter. Moreover, using convex analysis tools, we derive an iterative thresholding algorithm for computing the elastic-net solution which is different from the optimization procedure originally proposed in the above-cited work.

* Corresponding author at: Center for Biological and Computational Learning, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, United States.

E-mail addresses: demol@ulb.ac.be (C. De Mol), devito@dima.unige.it (E. De Vito), lrosasco@mit.edu (L. Rosasco).

0885-064X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2009.01.002

1. Introduction

We consider the standard framework of supervised learning, that is non-parametric regression with random design. In this setting, there is an input–output pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with unknown probability distribution P, and the goal is to find a prediction function $f_n : \mathcal{X} \to \mathcal{Y}$, based on a training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent random pairs distributed as (X, Y). A good solution f_n is such that, given a new input $x \in \mathcal{X}$, the value $f_n(x)$ is a good prediction of the true output $y \in \mathcal{Y}$. When choosing the square loss to measure the quality of the prediction, as we do throughout this paper, this means that the expected risk $\mathbb{E}\left[|Y - f_n(X)|^2\right]$ is *small*, or, in other words, that f_n is a *good* approximation of the regression function $f^*(x) = \mathbb{E}[Y \mid X = x]$ minimizing this risk.

In many learning problems, a major goal besides prediction is that of *selecting the variables* that are *relevant to achieving good predictions*. In the problem of variable selection we are given a set $(\psi_\gamma)_{\gamma\in\Gamma}$ of functions from the input space $\mathcal{X}$ into the output space $\mathcal{Y}$ and we aim at selecting those functions which are needed to represent the regression function, where the *representation* is typically given by a linear combination. The set $(\psi_\gamma)_{\gamma\in\Gamma}$ is usually called a *dictionary* and its elements *features*. We can think of the features as measurements used to represent the input data, as providing some relevant parametrization of the input space, or as a (possibly overcomplete) dictionary of functions used to represent the prediction function. In modern applications, the number p of features in the dictionary is usually very large, possibly much larger than the number n of examples in the training set. This situation is often referred to as the "large p, small n paradigm" [1], and a key to obtaining a meaningful solution in such a case is the requirement that the prediction function f_n is a linear combination of only a *few* elements in the dictionary, i.e. f_n admits a *sparse* representation.

The above setting can be illustrated by two examples of applications we are currently working on and which provide an underlying motivation for the theoretical framework developed in the present paper. The first application is a classification problem in computer vision, namely face detection [2–4]. The training set contains images of faces and non-faces and each image is represented by a very large redundant set of features capturing the local geometry of faces, for example wavelet-like dictionaries or other local descriptors. The aim is to find a good predictor able to detect faces in new images.

The second application is the analysis of microarray data, where the features are the expression level measurements of the genes in a given sample or patient, and the output is either a classification label discriminating between two or more pathologies or a continuous index indicating, for example, the gravity of an illness. In this problem, besides prediction of the output for examples-to-come, another important goal is the identification of the features that are the most relevant to building the estimator and would constitute a gene signature for a certain disease [5,6]. In both applications, the number of features we have to deal with is much larger than the number of examples and assuming sparsity of the solution is a very natural requirement.

The problem of variable/feature selection has a long history in statistics and it is known that the brute-force approach (trying all possible subsets of features), though theoretically appealing, is computationally unfeasible. A first strategy to overcome this problem is provided by greedy algorithms. A second route, which we follow in this paper, makes use of sparsity-based regularization schemes (convex relaxation methods). The most well-known example of such schemes is probably the so-called *Lasso regression* [7] – also referred to in the signal processing literature as Basis Pursuit Denoising [8] – where a coefficient vector β_n is estimated as the minimizer of the empirical risk penalized with the ℓ_1-norm, namely

$$\beta_n = \operatorname*{argmin}_{\beta=(\beta_\gamma)_{\gamma\in\Gamma}} \left(\frac{1}{n} \sum_{i=1}^{n} |Y_i - f_\beta(X_i)|^2 + \lambda \sum_{\gamma\in\Gamma} |\beta_\gamma| \right),$$

where $f_\beta = \sum_{\gamma\in\Gamma} \beta_\gamma \psi_\gamma$, λ is a suitable positive regularization parameter and $(\psi_\gamma)_{\gamma\in\Gamma}$ a given set of features. An extension of this approach, called *bridge regression*, amounts to replacing the ℓ_1-penalty by an ℓ_p-penalty [9]. It has been shown that this kind of penalty can still achieve sparsity when p is bigger, but very close to 1 (see [10]). For this class of techniques, both consistency and computational aspects have been studied. Non-asymptotic bounds within the framework of statistical learning have been studied in several papers [11–17,10]. A common feature of these results is that they assume

that the dictionary is finite (with cardinality possibly depending on the number of examples) and satisfies some assumptions about the linear independence of the relevant features – see [10] for a discussion on this point – whereas $\mathcal{Y}$ is usually assumed to be $\mathbb{R}$. Several numerical algorithms have also been proposed to solve the optimization problem underlying Lasso regression and are based e.g. on quadratic programming [8], on the so-called LARS algorithm [18] or on iterative soft-thresholding (see [19] and references therein).

Despite its success in many applications, the Lasso strategy has some drawback in variable selection problems where there are highly correlated features and we need to identify all the relevant ones. This situation is of uttermost importance for e.g. microarray data analysis since, as well known, there is a lot of functional dependency between genes which are organized in small interacting networks. The identification of such groups of correlated genes involved in a specific pathology is desirable to make progress in the understanding of the underlying biological mechanisms.

Motivated by microarray data analysis, Zou and Hastie [20] proposed the use of a penalty which is a weighted sum of the ℓ_1-norm and the square of the ℓ_2-norm of the coefficient vector β. The first term enforces the sparsity of the solution, whereas the second term ensures democracy among groups of correlated variables. In [20] the corresponding method is called *(naive) elastic net*. The method allows selecting groups of correlated features when the groups are not known in advance (algorithms to enforce group sparsity with *preassigned* groups of variables have been proposed in e.g. [21–23] using other types of penalties).

In the present paper we study several properties of the elastic-net regularization scheme for vector-valued regression in a random design. In particular, we prove consistency under some adaptive and non-adaptive choices for the regularization parameter. As concerns variable selection, we assess the accuracy of our estimator for the vector β with respect to the ℓ_2-norm, whereas the prediction ability of the corresponding function $f_n = f_{\beta_n}$ is measured by the expected risk $\mathbb{E}\left[|Y - f_n(X)|^2\right]$. To derive such error bounds, we characterize the solution of the variational problem underlying elastic-net regularization as the fixed point of a contractive map and, as a byproduct, we derive an explicit iterative thresholding procedure to compute the estimator. As explained below, in the presence of highly collinear features, the presence of the ℓ_2-penalty, besides enforcing grouped selection, is crucial to ensure stability with respect to random sampling.

In the remainder of this section, we define the main ingredients for elastic-net regularization within our general framework, discuss the underlying motivations for the method and then outline the main results established in the paper.

As an extension of the setting originally proposed in [20], we allow the dictionary to have an infinite number of features. In such a case, to cope with infinite sums, we need some assumptions on the coefficients. We assume that the prediction function we have to determine is a linear combination of the features $(\psi_\gamma)_{\gamma\in\Gamma}$ in the dictionary and that the series

$$f_\beta(x) = \sum_{\gamma\in\Gamma} \beta_\gamma \psi_\gamma(x),$$

converges absolutely for all $x \in \mathcal{X}$ and for all sequences $\beta = (\beta_\gamma)_{\gamma\in\Gamma}$ satisfying $\sum_{\gamma\in\Gamma} u_\gamma \beta_\gamma^2 < \infty$, where u_γ are given positive weights. The latter constraint can be viewed as a constraint on the *regularity* of the functions f_β we use to approximate the regression function. For infinite-dimensional sets, as for example wavelet bases or splines, suitable choices of the weights correspond to the assumption that f_β is in a Sobolev space (see Section 2 for more details about this point). Such a requirement of regularity is common when dealing with infinite-dimensional spaces of functions, as happens in approximation theory, signal analysis and inverse problems.

To ensure the convergence of the series defining f_β, we assume that

$$\sum_{\gamma\in\Gamma} \frac{|\psi_\gamma(x)|^2}{u_\gamma} \text{ is finite for all } x \in X. \tag{1}$$

Notice that for finite dictionaries, the series becomes a finite sum and the previous condition as well as the introduction of weights becomes superfluous.

To simplify the notation and the formulation of our results, and without any loss in generality, we will in the following rescale the features by defining $\varphi_\gamma = \psi_\gamma/\sqrt{u_\gamma}$, so that on this *rescaled dictionary*, $f_\beta = \sum_{\gamma\in\Gamma} \tilde{\beta}_\gamma \varphi_\gamma$ will be represented by means of a vector $\tilde{\beta}_\gamma = \sqrt{u_\gamma}\beta_\gamma$ belonging to ℓ_2; condition (1) then becomes $\sum_{\gamma\in\Gamma} |\varphi_\gamma(x)|^2 < +\infty$, for all $x \in X$. From now on, we will only use this rescaled representation and we drop the tilde on the vector β.

Let us now define our estimator as the minimizer of the empirical risk penalized with a (weighted) elastic-net penalty, that is, a combination of the squared ℓ_2-norm and a weighted ℓ_1-norm of the vector β. More precisely, we define the elastic-net penalty as follows.

Definition 1. Given a family $(w_\gamma)_{\gamma\in\Gamma}$ of weights $w_\gamma \geq 0$ and a parameter $\varepsilon \geq 0$, let $p_\varepsilon : \ell_2 \to [0, \infty]$ be defined as

$$p_\varepsilon(\beta) = \sum_{\gamma\in\Gamma} (w_\gamma|\beta_\gamma| + \varepsilon\beta_\gamma^2) \tag{2}$$

which can also be rewritten as $p_\varepsilon(\beta) = \|\beta\|_{1,w} + \varepsilon\,\|\beta\|_2^2$, where $\|\beta\|_{1,w} = \sum_{\gamma\in\Gamma} w_\gamma|\beta_\gamma|$.

The weights w_γ allow us to enforce more or less sparsity on different groups of features. We assume that they are prescribed in a given problem, so that we do not need to explicitly indicate the dependence of $p_\varepsilon(\beta)$ on these weights. The elastic-net estimator is defined by the following minimization problem.

Definition 2. Given $\lambda > 0$, let $\mathcal{E}_n^\lambda : \ell_2 \to [0, +\infty]$ be the empirical risk penalized by the penalty $p_\varepsilon(\beta)$

$$\mathcal{E}_n^\lambda(\beta) = \frac{1}{n}\sum_{i=1}^n |Y_i - f_\beta(X_i)|^2 + \lambda p_\varepsilon(\beta), \tag{3}$$

and let $\beta_n^\lambda \in \ell_2$ be *the* or *a* minimizer of (3) on ℓ_2

$$\beta_n^\lambda = \operatorname*{argmin}_{\beta\in\ell_2} \mathcal{E}_n^\lambda(\beta). \tag{4}$$

The positive parameter λ is a regularization parameter controlling the trade-off between the empirical error and the penalty. Clearly, β_n^λ also depends on the parameter ε, but we do not write explicitly this dependence since ε will always be fixed.

Setting $\varepsilon = 0$ in (3), we obtain as a special case an infinite-dimensional extension of the Lasso regression scheme. On the other hand, setting $w_\gamma = 0,\ \forall\gamma$, the method reduces to ℓ_2-regularized least-squares regression – also referred to as *ridge regression* – with a generalized linear model. The ℓ_1-penalty has selection capabilities since it enforces sparsity of the solution, whereas the ℓ_2-penalty induces a linear shrinkage on the coefficients leading to stable solutions. The positive parameter ε controls the trade-off between the ℓ_1-penalty and the ℓ_2-penalty.

We will show that, if $\varepsilon > 0$, the minimizer β_n^λ always exists and is unique. In the paper we will focus on the case $\varepsilon > 0$. Some of our results, however, still hold for $\varepsilon = 0$, possibly under some supplementary conditions, as will be indicated in due time.

As previously mentioned one of the main advantages of the elastic-net penalty is that it allows achieving stability with respect to random sampling. To illustrate this property more clearly, we consider a toy example where the (rescaled) dictionary has only two elements φ_1 and φ_2 with weights $w_1 = w_2 = 1$. The effect of random sampling is particularly dramatic in the presence of highly correlated features. To illustrate this situation, we assume that φ_1 and φ_2 exhibit a special kind of linear dependency, namely that they are linearly dependent on the input data $X_1, \ldots, X_n$: $\varphi_2(X_i) = \tan\theta_n\, \varphi_1(X_i)$ for all $i = 1, \ldots, n$, where we have parametrized the coefficient of proportionality by means of the angle $\theta_n \in [0, \pi/2]$. Notice that this angle is a random variable since it depends on the input data.

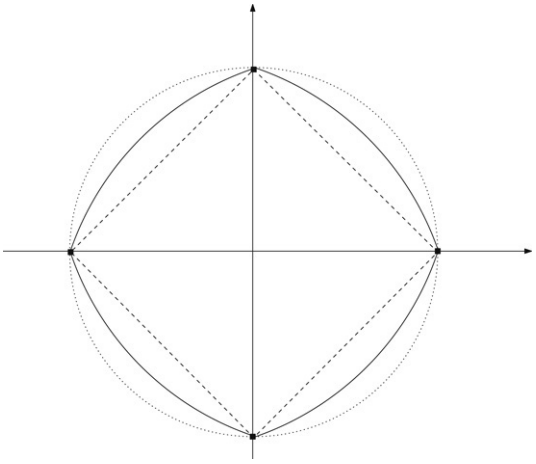

Fig. 1. The ε-ball with $\varepsilon > 0$ (solid line), the square (ℓ_1-ball), which is the ε-ball with $\varepsilon = 0$ (dashed line), and the disc (ℓ_2-ball), which is the ε-ball with $\varepsilon \to \infty$ (dotted line).

Observe that the minimizers of (3) must lie at a tangency point between a level set of the empirical error and a level set of the elastic-net penalty. The level sets of the empirical error are all parallel straight lines with slope $-\cot\theta_n$, as depicted by a dashed line in the two panels of Fig. 2, whereas the level sets of the elastic-net penalty are *elastic-net balls* (ε-*balls*) with center at the origin and corners at the intersections with the axes, as depicted in Fig. 1. When $\varepsilon = 0$, i.e. with a pure ℓ_1-penalty (Lasso), the ε-ball is simply *a square* (dashed line in Fig. 1) and we see that the unique tangency point will be the *top corner* if $\theta_n > \pi/4$ (the point T in the two panels of Fig. 2), or the *right corner* if $\theta_n < \pi/4$. For $\theta_n = \pi/4$ (that is, when φ_1 and φ_2 coincide on the data), the minimizer of (3) is no longer unique since the level sets will touch along an edge of the square. Now, if θ_n randomly tilts around $\pi/4$ (because of the random sampling of the input data), we see that the Lasso estimator is not stable since it randomly jumps between the top and the right corner. If $\varepsilon \to \infty$, i.e. with a pure ℓ_2-penalty (ridge regression), the ε-ball becomes a disc (dotted line in Fig. 1) and the minimizer is the point of the straight line having minimal distance from the origin (the point Q in the two panels of Fig. 2). The solution always exists, is stable under random perturbations, but it is never sparse (if $0 < \theta_n < \pi/2$).

The situation changes if we consider the elastic-net estimator with $\varepsilon > 0$ (the corresponding minimizer is the point P in the two panels of Fig. 2). The presence of the ℓ_2-term ensures a smooth and stable behavior when the Lasso estimator becomes unstable. More precisely, let $-\cot\theta_+$ be the slope of the right tangent at the top corner of the elastic-net ball ($\theta_+ > \pi/4$), and $-\cot\theta_-$ the slope of the upper tangent at the right corner ($\theta_- < \pi/4$). As depicted in top panel of Fig. 2, the minimizer will be the top corner if $\theta_n > \theta_+$. It will be the right corner if $\theta_n < \theta_-$. In both cases the elastic-net solution is sparse. On the other hand, if $\theta_- \le \theta_n \le \theta_+$ the minimizer has both components β_1 and β_2 different from zero — see the bottom panel of Fig. 2; in particular, $\beta_1 = \beta_2$ if $\theta_n = \pi/4$. Now we observe that if θ_n randomly tilts around $\pi/4$, the solution smoothly moves between the top corner and the right corner. However, the price we paid to get such stability is a decrease in sparsity, since the solution is sparse only when $\theta_n \notin [\theta_-, \theta_+]$.

The previous elementary example could be refined in various ways to show the essential role played by the ℓ_2-penalty to overcome the instability effects inherent to the use of the ℓ_1-penalty for variable selection in a random-design setting.

Remark 1. Stability in the case of collinear features can also be achieved by using an ℓ_p-penalty with $p > 1$ instead of $p = 1$. However, since such penalty term is differentiable, the corresponding ℓ_p-balls in our two-dimensional example are delimited by a smooth curve without any *corner* and, as a consequence, sparse solutions are not obtained in the presence of collinear features. Nevertheless,

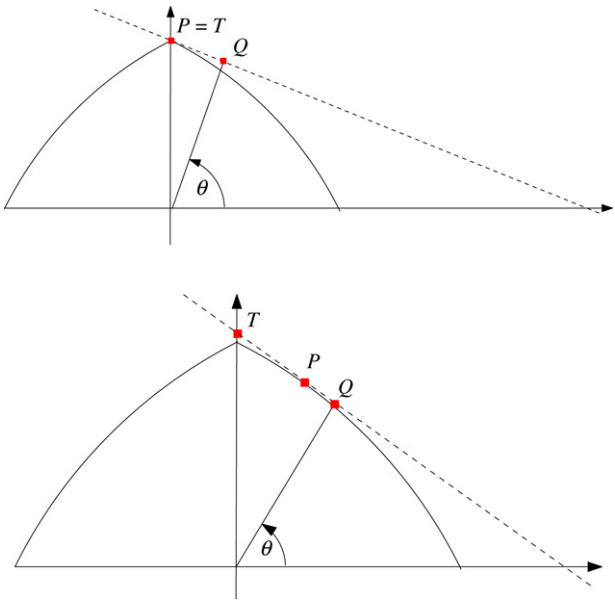

Fig. 2. Estimators in the two-dimensional example: T = Lasso, P = elastic net and Q = ridge regression. Top panel: $\theta_+ < \theta < \pi/2$. Bottom panel: $\pi/4 < \theta < \theta_+$.

when assuming that the relevant features are linearly independent, sparsity could still be enforced by means of ℓ_p-penalties as shown in [10].

We now conclude this introductory section by a summary of the main results which will be derived in the core of the paper. A key result will be to show that for $\varepsilon > 0$, β_n^λ is the fixed point of the following contractive map

$$\beta = \frac{1}{\tau + \varepsilon\lambda}\mathbf{S}_\lambda\left((\tau I - \Phi_n^*\Phi_n)\beta + \Phi_n^* Y\right)$$

where τ is a suitable relaxation constant, $\Phi_n^*\Phi_n$ and $\Phi_n^* Y$ are respectively the matrix and the vector with entries

$$(\Phi_n^*\Phi_n)_{\gamma,\gamma'} = \frac{1}{n}\sum_{i=1}^n \left\langle \varphi_\gamma(X_i), \varphi_{\gamma'}(X_i)\right\rangle \quad \text{and} \quad (\Phi_n^* Y)_\gamma = \frac{1}{n}\sum_{i=1}^n \left\langle \varphi_\gamma(X_i), Y_i\right\rangle,$$

($\langle\cdot,\cdot\rangle$ denotes the scalar product in the output space $\mathcal{Y}$). Moreover, $\mathbf{S}_\lambda(\beta)$ is the soft-thresholding operator acting componentwise as follows

$$[\mathbf{S}_\lambda(\beta)]_\gamma = \begin{cases} \beta_\gamma - \dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma > \dfrac{\lambda w_\gamma}{2} \\ 0 & \text{if } |\beta_\gamma| \le \dfrac{\lambda w_\gamma}{2} \\ \beta_\gamma + \dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma < -\dfrac{\lambda w_\gamma}{2}. \end{cases}$$

As a consequence of the Banach fixed point theorem, β_n^λ can be computed by means of an iterative algorithm. This procedure is completely different from the modification of the LARS algorithm used in [20] and is akin instead to the algorithm developed in [19].

Another interesting property which we will derive from the above equation is that the non-zero components of β_n^λ are such that $w_\gamma \le \frac{C}{\lambda}$, where C is a constant depending on the data. Hence the only active features are those for which the corresponding weight lies below the threshold C/λ. If

the features are organized into finite subsets of increasing complexity (as happens for example for wavelets) and the weights tend to infinity with increasing feature complexity, then the number of active features is finite and can be determined for any given data set. Let us recall that in the case of ridge regression, the so-called *representer theorem*, see [24], ensures that we only have to solve in practice a finite-dimensional optimization problem, even when the dictionary is infinite-dimensional (as in kernel methods). This is no longer true, however, with an ℓ_1-type regularization and, for practical purposes, one would need to truncate infinite dictionaries. A standard way to do this is to consider only a finite subset of m features, with m possibly depending on n — see for example [12,15]. Notice that such a procedure implicitly assumes some order in the features and makes sense only if the retained features are the most relevant ones. For example, in [25], it is assumed that there is a natural exhaustion of the hypothesis space with nested subspaces spanned by finite-dimensional subsets of features of increasing size. In our approach we adopt a different strategy, namely the encoding of such information in the elastic-net penalty by means of suitable weights in the ℓ_1-norm.

The main result of our paper concerns the consistency for variable selection of β_n^λ. We prove that, if the regularization parameter $\lambda = \lambda_n$ satisfies the conditions $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty}(n\lambda_n^2 - 2\log n) = +\infty$, then

$$\lim_{n\to\infty} \left\| \beta_n^{\lambda_n} - \beta^\varepsilon \right\|_2 = 0 \text{ with probability one,}$$

where the vector β^ε, which we call the *elastic-net representation of* f_β, is the minimizer of

$$\min_{\beta\in\ell_2} \left(\sum_{\gamma\in\Gamma} w_\gamma |\beta_\gamma| + \varepsilon \sum_{\gamma\in\Gamma} |\beta_\gamma|^2 \right) \text{ subject to } f_\beta = f^*.$$

The vector β^ε exists and is unique provided that $\varepsilon > 0$ and the regression function f^* admits a *sparse representation on the dictionary*, i.e. $f^* = \sum_{\gamma\in\Gamma} \beta_\gamma^* \varphi_\gamma$ for at least a vector $\beta^* \in \ell_2$ such that $\sum_{\gamma\in\Gamma} w_\gamma |\beta_\gamma^*|$ is finite. Notice that, when the features are linearly dependent, there is a problem of identifiability since there are many vectors β such that $f^* = f_\beta$. The elastic-net regularization scheme forces $\beta_n^{\lambda_n}$ to converge to β^ε. This is precisely what happens for linear inverse problems where the regularized solution converges to the minimum-norm solution of the least-squares problem. As a consequence of the above convergence result, one easily deduces the consistency of the corresponding prediction function $f_n := f_{\beta_n^{\lambda_n}}$, that is, $\lim_{n\to\infty} \mathbb{E}\left[|f_n - f^*|^2\right] = 0$ with probability one. When the regression function does not admit a sparse representation, we can still prove the previous consistency result for f_n provided that the linear span of the features is sufficiently rich. Finally, we use a data-driven choice for the regularization parameter, based on the so-called balancing principle [26], to obtain non-asymptotic bounds which are adaptive to the unknown regularity of the regression function.

The rest of the paper is organized as follows. In Section 2, we set up the mathematical framework of the problem. In Section 3, we analyze the optimization problem underlying elastic-net regularization and the iterative thresholding procedure we propose to compute the estimator. Finally, Section 4 contains the statistical analysis with our main results concerning the estimation of the errors on our estimators as well as their consistency properties under appropriate a priori and adaptive strategies for choosing the regularization parameter.

2. Mathematical setting of the problem

2.1. Notations and assumptions

In this section we describe the general setting of the regression problem we want to solve and specify all the required assumptions.

We assume that $\mathcal{X}$ is a separable metric space and that $\mathcal{Y}$ is a (real) separable Hilbert space, with norm and scalar product denoted respectively by $|\cdot|$ and $\langle\cdot,\cdot\rangle$. Typically, $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $\mathcal{Y}$ is $\mathbb{R}$. Recently, however, there has been an increasing interest in vector-valued regression problems [27, 28] and multiple supervised learning tasks [29,30]: in both settings $\mathcal{Y}$ is taken to be $\mathbb{R}^m$. Also infinite-dimensional output spaces are of interest as e.g. in the problem of estimating glycemic response during

a time interval depending on the amount and type of food; in such a case, $\mathcal{Y}$ is the space L^2 or some Sobolev space. Other examples of applications in an infinite-dimensional setting are given in [31].

Our first assumption concerns the set of features.

Assumption 1. The family of features $(\varphi_\gamma)_{\gamma\in\Gamma}$ is a countable set of measurable functions $\varphi_\gamma : \mathcal{X} \to \mathcal{Y}$ such that

$$\forall x \in \mathcal{X} \quad k(x) = \sum_{\gamma\in\Gamma} |\varphi_\gamma(x)|^2 \le \kappa, \tag{5}$$

for some finite number κ.

The index set Γ is countable, but we do not assume any order. As for the convergence of series, we use the notion of summability: given a family $(v_\gamma)_{\gamma\in\Gamma}$ of vectors in a normed vector space V, $v = \sum_{\gamma\in\Gamma} v_\gamma$ means that $(v_\gamma)_{\gamma\in\Gamma}$ is summable[1] with sum $v \in V$.

Assumption 1 can be seen as a condition on the class of functions that can be recovered by the elastic-net scheme. As already noted in the Introduction, we have at our disposal an arbitrary (countable) dictionary $(\psi_\gamma)_{\gamma\in\Gamma}$ of measurable functions, and we try to approximate f^* with linear combinations $f_\beta(x) = \sum_{\gamma\in\Gamma} \beta_\gamma \psi_\gamma(x)$ where the set of coefficients $(\beta_\gamma)_{\gamma\in\Gamma}$ satisfies some *decay condition* equivalent to a *regularity condition* on the functions f_β. We make this condition precise by assuming that there exists a sequence of positive weights $(u_\gamma)_{\gamma\in\Gamma}$ such that $\sum_{\gamma\in\Gamma} u_\gamma \beta_\gamma^2 < \infty$ and, for any of such vectors $\beta = (\beta_\gamma)_{\gamma\in\Gamma}$, that the series defining f_β converges absolutely for all $x \in \mathcal{X}$. These two facts follow from the requirement that the set of rescaled features $\varphi_\gamma = \frac{\psi_\gamma}{\sqrt{u_\gamma}}$ satisfies $\sum_{\gamma\in\Gamma} |\varphi_\gamma(x)|^2 < \infty$. Condition (5) is a little bit stronger since it requires that $\sup_{x\in\mathcal{X}} \sum_{\gamma\in\Gamma} |\varphi_\gamma(x)|^2 < \infty$, so that we also have that the functions f_β are bounded. To simplify the notation, in the rest of the paper, we only use the (rescaled) features φ_γ and, with this choice, the regularity condition on the coefficients $(\beta_\gamma)_{\gamma\in\Gamma}$ becomes $\sum_{\gamma\in\Gamma} \beta_\gamma^2 < \infty$.

An example of features satisfying condition (5) is given by a family of *rescaled* wavelets on $\mathcal{X} = [0, 1]$. Let $\{\psi_{jk} \mid j = 0, 1 \ldots;\ k \in \Delta_j\}$ be a orthonormal wavelet basis in $L^2([0, 1])$ with regularity C^r, $r > \frac{1}{2}$, where for $j \ge 1$ $\{\psi_{jk} \mid k \in \Delta_j\}$ is the orthonormal wavelet basis (with suitable boundary conditions) spanning the detail space at level j. To simplify notation, it is assumed that the set $\{\psi_{0k} \mid k \in \Delta_0\}$ contains both the wavelets and the scaling functions at level $j = 0$. Fix s such that $\frac{1}{2} < s < r$ and let $\varphi_{jk} = 2^{-js}\psi_{jk}$. Then

$$\sum_{j=0}^{\infty} \sum_{k\in\Delta_j} |\varphi_{jk}(x)|^2 = \sum_{j=0}^{\infty} \sum_{k\in\Delta_j} 2^{-2js} |\psi_{jk}(x)|^2 \le C \sum_{j=0}^{\infty} 2^{-2js} 2^j = C \frac{1}{1 - 2^{1-2s}} = \kappa,$$

where C is a suitable constant depending on the number of wavelets that are non-zero at a point $x \in [0, 1]$ for a given level j, and on the maximum values of the scaling function and of the mother wavelet; see [32] for a similar setting.

Condition (5) allows defining the hypothesis space in which we search for the estimator. Let ℓ_2 be the Hilbert space of the families $(\beta_\gamma)_{\gamma\in\Gamma}$ of real numbers such that $\sum_{\gamma\in\Gamma} \beta_\gamma^2 < \infty$, with the usual scalar product $\langle\cdot,\cdot\rangle_2$ and the corresponding norm $\|\cdot\|_2$. We will denote by $(e_\gamma)_{\gamma\in\Gamma}$ the canonical basis of ℓ_2 and by $\mathrm{supp}(\beta) = \{\gamma \in \Gamma \mid \beta_\gamma \ne 0\}$ the support of β. The Cauchy–Schwarz inequality and condition (5) ensure that, for any $\beta = (\beta_\gamma)_{\gamma\in\Gamma} \in \ell_2$, the series

$$\sum_{\gamma\in\Gamma} \beta_\gamma \varphi_\gamma(x) = f_\beta(x)$$

[1] That is, for all $\eta > 0$, there is a finite subset $\Gamma_0 \subset \Gamma$ such that $\left\| v - \sum_{\gamma\in\Gamma'} v_\gamma \right\|_V \le \eta$ for all finite subsets $\Gamma' \supset \Gamma_0$. If $\Gamma = \mathbb{N}$, the notion of summability is equivalent to requiring the series to converge unconditionally (i.e. its terms can be permuted without affecting convergence). If the vector space is finite-dimensional, summability is equivalent to absolute convergence, but in the infinite-dimensional setting, there are summable series which are not absolutely convergent.

is summable in $\mathcal{Y}$ uniformly on $\mathcal{X}$ with

$$\sup_{x\in\mathcal{X}} |f_\beta(x)| \le \|\beta\|_2 \, \kappa^{\frac{1}{2}}. \tag{6}$$

Later on, in Proposition 3, we will show that the hypothesis space $\mathcal{H} = \left\{f_\beta \mid \beta \in \ell_2\right\}$ is then a vector-valued reproducing kernel Hilbert space on $\mathcal{X}$ with a bounded kernel [33], and that $(\varphi_\gamma)_{\gamma\in\Gamma}$ is a normalized tight frame for $\mathcal{H}$. In the example of the wavelet features one can easily check that $\mathcal{H}$ is the Sobolev space H^s on $[0, 1]$ and $\|\beta\|_2$ is equivalent to $\|f_\beta\|_{H^s}$.

The second assumption concerns the regression model.

Assumption 2. The random couple (X, Y) in $\mathcal{X} \times \mathcal{Y}$ obeys the regression model

$$Y = f^*(X) + W$$

where

$$f^* = f_{\beta^*} \quad \text{for some } \beta^* \in \ell_2 \text{ with } \sum_{\gamma\in\Gamma} w_\gamma |\beta^*_\gamma| < +\infty \tag{7}$$

and

$$\mathbb{E}\,[W \mid X] = 0 \tag{8}$$

$$\mathbb{E}\left[\exp\left(\frac{|W|}{L}\right) - \frac{|W|}{L} - 1 \,\middle|\, X\right] \le \frac{\sigma^2}{2L^2} \tag{9}$$

with $\sigma, L > 0$. The family $(w_\gamma)_{\gamma\in\Gamma}$ forms the positive weights defining the elastic-net penalty $p_\varepsilon(\beta)$ in (2).

Observe that $f^* = f_{\beta^*}$ is always a bounded function by (6). Moreover condition (7) is a further regularity condition on the regression function and will not be needed for some of the results derived in the paper. Assumption (9) is satisfied by bounded, Gaussian or sub-Gaussian noise. In particular, it implies

$$\mathbb{E}\left[|W|^m|X\right] \le \frac{1}{2} m!\, \sigma^2 L^{m-2}, \quad \forall m \ge 2, \tag{10}$$

see [34], so that W has a finite second moment. It follows that Y has a finite first moment and (8) implies that f^* is the regression function $\mathbb{E}\,[Y \mid X = x]$.

Condition (7) controls both the sparsity and the regularity of the regression function. If $\inf_{\gamma\in\Gamma} w_\gamma = w_0 > 0$, it is sufficient to require that $\|\beta^*\|_{1,w}$ is finite. Indeed, the Hölder inequality gives that

$$\|\beta\|_2 \le \frac{1}{w_0} \|\beta\|_{1,w}. \tag{11}$$

If $w_0 = 0$, we also need $\|\beta^*\|_2$ to be finite. In the example of the (rescaled) wavelet features a natural choice for the weights is $w_{jk} = 2^{ja}$ for some $a \in \mathbb{R}$, so that $\|\beta\|_{1,w}$ is equivalent to the norm $\|f_\beta\|_{B^{\tilde{s}}_{1,1}}$, with $\tilde{s} = a + s + \frac{1}{2}$, in the Besov space $B^{\tilde{s}}_{1,1}$ on $[0, 1]$ (for more details, see e.g. the appendix in [19]). In such a case, (7) is equivalent to requiring that $f^* \in H^s \cap B^{\tilde{s}}_{1,1}$.

Finally, our third assumption concerns the training sample.

Assumption 3. The sequence of random pairs $(X_n, Y_n)_{n\ge1}$ are independent and identically distributed (*i.i.d.*) according to the distribution of (X, Y).

In the following, we let P be the probability distribution of (X, Y), and $L^2_{\mathcal{Y}}(P)$ be the Hilbert space of (measurable) functions $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ with the norm

$$\|f\|_P^2 = \int_{\mathcal{X}\times\mathcal{Y}} |f(x, y)|^2 \, \mathrm{d}P(x, y).$$

With a slight abuse of notation, we regard the random pair (X, Y) as a function on $\mathcal{X} \times \mathcal{Y}$, that is, $X(x, y) = x$ and $Y(x, y) = y$. Moreover, we denote by $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$ the empirical distribution and by $L^2_{\mathcal{Y}}(\mathbb{P}_n)$ the corresponding (finite-dimensional) Hilbert space with norm

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} |f(X_i, Y_i)|^2.$$

2.2. Operators defined by the set of features

The choice of a quadratic loss function and the Hilbert structure of the hypothesis space suggest using some tools from the theory of linear operators. In particular, the function f_β depends linearly on β and can be regarded as an element of both $L^2_{\mathcal{Y}}(P)$ and of $L^2_{\mathcal{Y}}(\mathbb{P}_n)$. Hence it defines two operators, whose properties are summarized by the next two propositions, based on the following lemma.

Lemma 1. *For any fixed $x \in \mathcal{X}$, the map $\Phi_x : \ell_2 \to \mathcal{Y}$ defined by*

$$\Phi_x \beta = \sum_{\gamma \in \Gamma} \varphi_\gamma(x)\beta_\gamma = f_\beta(x)$$

is a Hilbert–Schmidt operator, its adjoint $\Phi_x^* : \mathcal{Y} \to \ell_2$ acts as

$$(\Phi_x^* y)_\gamma = \langle y, \varphi_\gamma(x)\rangle \quad \gamma \in \Gamma,\ y \in \mathcal{Y}. \tag{12}$$

In particular $\Phi_x^*\Phi_x$ is a trace-class operator with

$$\mathrm{Tr}\,(\Phi_x^*\Phi_x) \le \kappa. \tag{13}$$

Moreover, $\Phi_X^* Y$ is an ℓ_2-valued random variable with

$$\left\|\Phi_X^* Y\right\|_2 \le \kappa^{\frac{1}{2}} |Y|, \tag{14}$$

and $\Phi_X^*\Phi_X$ is an $\mathcal{L}_{\mathrm{HS}}$-valued random variable with

$$\left\|\Phi_X^*\Phi_X\right\|_{\mathrm{HS}} \le \kappa, \tag{15}$$

where $\mathcal{L}_{\mathrm{HS}}$ denotes the separable Hilbert space of the Hilbert–Schmidt operators on ℓ_2, and $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert–Schmidt norm.

Proof. Clearly Φ_x is a linear map from ℓ_2 to $\mathcal{Y}$. Since $\Phi_x e_\gamma = \varphi_\gamma(x)$, we have

$$\sum_{\gamma\in\Gamma} |\Phi_x e_\gamma|^2 = \sum_{\gamma\in\Gamma} |\varphi_\gamma(x)|^2 \le \kappa,$$

so that Φ_x is a Hilbert–Schmidt operator and $\mathrm{Tr}\,(\Phi_x^*\Phi_x) \le \kappa$ by (5). Moreover, given $y \in \mathcal{Y}$ and $\gamma \in \Gamma$

$$(\Phi_x^* y)_\gamma = \langle \Phi_x^* y, e_\gamma\rangle_2 = \langle y, \varphi_\gamma(x)\rangle$$

which is (12). Finally, since $\mathcal{X}$ and $\mathcal{Y}$ are separable, the map $(x, y) \to \langle y, \varphi_\gamma(x)\rangle$ is measurable, then $(\Phi_X^* Y)_\gamma$ is a real random variable and, since ℓ_2 is separable, $\Phi_X^* Y$ is an ℓ_2-valued random variable with

$$\left\|\Phi_X^* Y\right\|_2^2 = \sum_{\gamma\in\Gamma} \langle Y, \varphi_\gamma(X)\rangle^2 \le \kappa |Y|^2.$$

A similar proof holds for $\Phi_X^*\Phi_X$, recalling that any trace-class operator is in $\mathcal{L}_{\mathrm{HS}}$ and $\left\|\Phi_X^*\Phi_X\right\|_{\mathrm{HS}} \le \mathrm{Tr}\,(\Phi_X^*\Phi_X)$. □

The following proposition defines the distribution-dependent operator Φ_P as a map from ℓ_2 into $L^2_{\mathcal{Y}}(P)$.

Proposition 1. *The map $\Phi_P : \ell_2 \to L^2_{\mathcal{Y}}(P)$, defined by $\Phi_P \beta = f_\beta$, is a Hilbert–Schmidt operator and*

$$\Phi_P^* Y = \mathbb{E}\left[\Phi_X^* Y\right] \tag{16}$$

$$\Phi_P^* \Phi_P = \mathbb{E}\left[\Phi_X^* \Phi_X\right] \tag{17}$$

$$\mathrm{Tr}\,(\Phi_P^* \Phi_P) = \mathbb{E}\,[k(X)] \le \kappa. \tag{18}$$

Proof. Since f_β is a bounded (measurable) function, $f_\beta \in L^2_{\mathcal{Y}}(P)$ and

$$\sum_{\gamma\in\Gamma} \left\| \Phi_P e_\gamma \right\|_P^2 = \sum_{\gamma\in\Gamma} \mathbb{E}\left[|\varphi_\gamma(X)|^2\right] = \mathbb{E}\,[k(X)] \le \kappa.$$

Hence Φ_P is a Hilbert–Schmidt operator with $\mathrm{Tr}\,(\Phi_P^*\Phi_P) = \sum_{\gamma\in\Gamma} \left\| \Phi_P e_\gamma \right\|_P^2$ so that (18) holds. By (9) W has a finite second moment and by (6) $f^* = f_{\beta^*}$ is a bounded function, hence $Y = f^*(X) + W$ is in $L^2_{\mathcal{Y}}(P)$. Now for any $\beta \in \ell_2$ we have

$$\left\langle \Phi_P^* Y, \beta \right\rangle_2 = \langle Y, \Phi_P \beta\rangle_P = \mathbb{E}\,[\langle Y, \Phi_X \beta\rangle] = \mathbb{E}\left[\left\langle \Phi_X^* Y, \beta\right\rangle_2\right].$$

On the other hand, by (14), $\Phi_X^* Y$ has finite expectation, so that (16) follows. Finally, given $\beta, \beta' \in \ell_2$

$$\left\langle \Phi_P^* \Phi_P \beta', \beta\right\rangle_2 = \left\langle \Phi_P \beta', \Phi_P \beta\right\rangle_P = \mathbb{E}\left[\left\langle \Phi_X \beta', \Phi_X \beta\right\rangle\right] = \mathbb{E}\left[\left\langle \Phi_X^* \Phi_X \beta', \beta\right\rangle_2\right]$$

so that (17) is clear, since $\Phi_X^* \Phi_X$ has finite expectation as a consequence of the fact that it is a bounded $\mathcal{L}_{\mathrm{HS}}$-valued random variable. □

Replacing P by the empirical measure we get the sample version of the operator.

Proposition 2. *The map $\Phi_n : \ell_2 \to L^2_{\mathcal{Y}}(\mathbb{P}_n)$ defined by $\Phi_n \beta = f_\beta$ is Hilbert–Schmidt operator and*

$$\Phi_n^* Y = \frac{1}{n} \sum_{i=1}^n \Phi_{X_i}^* Y_i \tag{19}$$

$$\Phi_n^* \Phi_n = \frac{1}{n} \sum_{i=1}^n \Phi_{X_i}^* \Phi_{X_i} \tag{20}$$

$$\mathrm{Tr}\,(\Phi_n^* \Phi_n) = \frac{1}{n} \sum_{i=1}^n k(X_i) \le \kappa. \tag{21}$$

The proof of Proposition 2 is analogous to the proof of Proposition 1, except that P is to be replaced by $\mathbb{P}_n$.

By (12) with $y = \varphi_{\gamma'}(x)$, we have that the matrix elements of the operator $\Phi_x^* \Phi_x$ are $(\Phi_x^* \Phi_x)_{\gamma\gamma'} = \left\langle \varphi_{\gamma'}(x), \varphi_\gamma(x)\right\rangle$ so that $\Phi_n^* \Phi_n$ is the empirical mean of the Gram matrix of the set $(\varphi_\gamma)_{\gamma\in\Gamma}$, whereas $\Phi_P^* \Phi_P$ is the corresponding mean with respect to the distribution P. Notice that if the features are linearly dependent in $L^2_{\mathcal{Y}}(\mathbb{P}_n)$, the matrix $\Phi_n^* \Phi_n$ has a non-trivial kernel and hence is not invertible. More important, if Γ is countably infinite, $\Phi_n^* \Phi_n$ is a compact operator, so that its inverse (if it exists) is not bounded. On the contrary, if Γ is finite and $(\varphi_\gamma)_{\gamma\in\Gamma}$ are linearly independent in $L^2_{\mathcal{Y}}(\mathbb{P}_n)$, then $\Phi_n^* \Phi_n$ is invertible. A similar reasoning holds for the matrix $\Phi_P^* \Phi_P$. To control whether these matrices have a bounded inverse or not, we introduce a lower spectral bound $\kappa_0 \ge 0$, such that

$$\kappa_0 \le \inf_{\beta\in\ell_2 | \|\beta\|_2 = 1} \left\langle \Phi_P^* \Phi_P \beta, \beta\right\rangle_2$$

and, with probability 1,

$$\kappa_0 \le \inf_{\beta\in\ell_2 | \|\beta\|_2 = 1} \left\langle \Phi_n^* \Phi_n \beta, \beta\right\rangle_2.$$

Clearly we can have $\kappa_0 > 0$ only if Γ is finite and the features $(\varphi_\gamma)_{\gamma\in\Gamma}$ are linearly independent both in $L^2_{\mathcal{Y}}(\mathbb{P}_n)$ and $L^2_{\mathcal{Y}}(P)$.

On the other hand, (18) and (21) give the crude upper spectral bounds

$$\sup_{\beta\in\ell_2|\|\beta\|_2=1} \left\langle \Phi_P^*\Phi_P\beta, \beta\right\rangle_2 \le \kappa,$$
$$\sup_{\beta\in\ell_2|\|\beta\|_2=1} \left\langle \Phi_n^*\Phi_n\beta, \beta\right\rangle_2 \le \kappa.$$

One could improve these estimates by means of a tight bound on the largest eigenvalue of $\Phi_P^*\Phi_P$.

We end this section by showing that, under the assumptions we made, a structure of reproducing kernel Hilbert space emerges naturally. Let us denote by $\mathcal{Y}^{\mathcal{X}}$ the space of functions from $\mathcal{X}$ to $\mathcal{Y}$.

Proposition 3. *The linear operator $\Phi : \ell_2 \to \mathcal{Y}^{\mathcal{X}}$, $\Phi\beta = f_\beta$, is a partial isometry from ℓ_2 onto the vector-valued reproducing kernel Hilbert space $\mathcal{H}$ on $\mathcal{X}$, with reproducing kernel $K : \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathcal{Y})$*

$$K(x,t)y = (\Phi_x\Phi_t^*)y = \sum_{\gamma\in\Gamma}\varphi_\gamma(x)\left\langle y, \varphi_\gamma(t)\right\rangle \quad x,t\in\mathcal{X},\ y\in\mathcal{Y}, \tag{22}$$

the null space of Φ is

$$\ker\Phi = \left\{\beta\in\ell_2 \mid \sum_{\gamma\in\Gamma}\varphi_\gamma(x)\beta_\gamma = 0 \quad \forall x\in X\right\}, \tag{23}$$

and the family $(\varphi_\gamma)_{\gamma\in\Gamma}$ is a normalized tight frame in $\mathcal{H}$, namely

$$\sum_{\gamma\in\Gamma}\left|\left\langle f,\varphi_\gamma\right\rangle_{\mathcal{H}}\right|^2 = \|f\|^2_{\mathcal{H}} \quad \forall f\in\mathcal{H}.$$

Conversely, let $\mathcal{H}$ be a vector-valued reproducing kernel Hilbert space with reproducing kernel K such that $K(x,x) : \mathcal{Y}\to\mathcal{Y}$ is a trace-class operator for all $x\in\mathcal{X}$, with trace bounded by κ. If $(\varphi_\gamma)_{\gamma\in\Gamma}$ is a normalized tight frame in $\mathcal{H}$, then (5) holds.

Proof. Proposition 2.4 of [33] (with $\mathcal{K} = \mathcal{Y}$, $\widehat{\mathcal{H}} = \ell_2$, $\gamma(x) = \Phi_x^*$ and $A = \Phi$) gives that Φ is a partial isometry from ℓ_2 onto the reproducing kernel Hilbert space $\mathcal{H}$, with reproducing kernel $K(x,t)$. On the other hand (23) is clear. Since Φ is a partial isometry with range $\mathcal{H}$ and $\Phi e_\gamma = \varphi_\gamma$ where $(e_\gamma)_{\gamma\in\Gamma}$ is a basis in ℓ_2, then $(\varphi_\gamma)_{\gamma\in\Gamma}$ is a normalized tight frame in $\mathcal{H}$.

To show the converse result, given $x\in X$ and $y\in\mathcal{Y}$, we apply the definition of a normalized tight frame to the function K_xy defined by $(K_xy)(t) = K(t,x)y$. K_xy belongs to $\mathcal{H}$ by the definition of a reproducing kernel Hilbert space and is such that the following reproducing property $\langle f, K_xy\rangle_{\mathcal{H}} = \langle f(x), y\rangle$ holds for any $f\in\mathcal{H}$. Then

$$\langle K(x,x)y, y\rangle = \|K_xy\|^2_{\mathcal{H}} = \sum_{\gamma\in\Gamma}\left|\left\langle K_xy,\varphi_\gamma\right\rangle_{\mathcal{H}}\right|^2 = \sum_{\gamma\in\Gamma}\left|\left\langle y,\varphi_\gamma(x)\right\rangle\right|^2,$$

where we used twice the reproducing property. Now, if $(y_i)_{i\in I}$ is a basis in $\mathcal{Y}$ and $x\in\mathcal{X}$

$$\sum_{\gamma\in\Gamma}|\varphi_\gamma(x)|^2 = \sum_{\gamma\in\Gamma}\sum_{i\in I}\left|\left\langle y_i,\varphi_\gamma(x)\right\rangle\right|^2 = \sum_{i\in I}\langle K(x,x)y_i, y_i\rangle = \mathrm{Tr}\,(K(x,x)) \le \kappa. \quad \square$$

3. Minimization of the elastic-net functional

In this section, we study the properties of the elastic-net estimator β_n^λ defined by (4). First of all, we characterize the minimizer of the elastic-net functional (3) as the unique fixed point of a contractive map. Moreover, we characterize some sparsity properties of the estimator and propose a natural iterative soft-thresholding algorithm to compute it. Our algorithmic approach is totally different from

the method proposed in [20], where β_n^λ is computed by first reducing the problem to the case of a pure ℓ_1-penalty and then applying the LARS algorithm [18].

In the following we make use of the following vector notation. Given a sample of n i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, and using the operators defined in the previous section, we can rewrite the elastic-net functional (3) as

$$\mathcal{E}_n^\lambda(\beta) = \|\Phi_n\beta - Y\|_n^2 + \lambda p_\varepsilon(\beta), \tag{24}$$

where the $p_\varepsilon(\cdot)$ is the elastic-net penalty defined by (2).

3.1. Fixed point equation

The main difficulty in minimizing (24) is that the functional is not differentiable because of the presence of the ℓ_1-term in the penalty. Nonetheless the convexity of such a term enables us to use tools from subdifferential calculus. Recall that, if $F : \ell_2 \to \mathbb{R}$ is a convex functional, the subgradient at a point $\beta \in \ell_2$ is the set of elements $\eta \in \ell_2$ such that

$$F(\beta + \beta') \geq F(\beta) + \langle \eta, \beta' \rangle_2 \quad \forall \beta' \in \ell_2.$$

The subgradient at β is denoted by $\partial F(\beta)$, see [35]. We compute the subgradient of the convex functional $p_\varepsilon(\beta)$, using the following definition of $\mathrm{sgn}(t)$

$$\begin{cases} \mathrm{sgn}(t) = 1 & \text{if } t > 0 \\ \mathrm{sgn}(t) \in [-1, 1] & \text{if } t = 0 \\ \mathrm{sgn}(t) = -1 & \text{if } t < 0. \end{cases} \tag{25}$$

We first state the following lemma.

Lemma 2. *The functional $p_\varepsilon(\cdot)$ is a convex, lower semi-continuous (l.s.c.) functional from ℓ_2 into $[0, \infty]$. Given $\beta \in \ell_2$, a vector $\eta \in \partial p_\varepsilon(\beta)$ if and only if*

$$\eta_\gamma = w_\gamma \mathrm{sgn}(\beta_\gamma) + 2\varepsilon\beta_\gamma \quad \forall \gamma \in \Gamma \quad \text{and} \quad \sum_{\gamma\in\Gamma} \eta_\gamma^2 < +\infty.$$

Proof. Define the map $F : \Gamma \times \mathbb{R} \to [0, \infty]$

$$F(\gamma, t) = w_\gamma |t| + \varepsilon t^2.$$

Given $\gamma \in \Gamma$, $F(\gamma, \cdot)$ is a convex, continuous function and its subgradient is

$$\partial F(\gamma, t) = \left\{ \tau \in \mathbb{R} \mid \tau = w_\gamma \mathrm{sgn}(t) + 2\varepsilon t \right\},$$

where we used the fact that the subgradient of $|t|$ is given by $\mathrm{sgn}(t)$. Since

$$p_\varepsilon(\beta) = \sum_{\gamma\in\Gamma} F(\gamma, \beta_\gamma) = \sup_{\Gamma' \text{ finite}} \sum_{\gamma\in\Gamma'} F(\gamma, \beta_\gamma)$$

and $\beta \mapsto \beta_\gamma$ is continuous, a standard result of convex analysis [35] ensures that $p_\varepsilon(\cdot)$ is convex and lower semi-continuous.

The computation of the subgradient is standard. Given $\beta \in \ell_2$ and $\eta \in \partial p_\varepsilon(\beta) \subset \ell_2$, by the definition of a subgradient,

$$\sum_{\gamma\in\Gamma} F(\gamma, \beta_\gamma + \beta'_\gamma) \geq \sum_{\gamma\in\Gamma} F(\gamma, \beta_\gamma) + \sum_{\gamma\in\Gamma} \eta_\gamma \beta'_\gamma \quad \forall \beta' \in \ell_2.$$

Given $\gamma \in \Gamma$, choosing $\beta' = te_\gamma$ with $t \in \mathbb{R}$, it follows that η_γ belongs to the subgradient of $F(\gamma, \beta_\gamma)$, that is,

$$\eta_\gamma = w_\gamma \mathrm{sgn}(\beta_\gamma) + 2\varepsilon\beta_\gamma. \tag{26}$$

Conversely, if (26) holds for all $\gamma \in \Gamma$, by the definition of a subgradient

$$F(\gamma, \beta_\gamma + \beta'_\gamma) \geq F(\gamma, \beta_\gamma) + \eta_\gamma \beta'_\gamma.$$

By summing over $\gamma \in \Gamma$ and taking into account the fact that $(\eta_\gamma \beta_\gamma)_{\gamma\in\Gamma} \in \ell_1$, then

$$p_\varepsilon(\beta + \beta') \geq p_\varepsilon(\beta) + \langle \eta, \beta' \rangle_2 . \quad \square$$

To state our main result about the characterization of the minimizer of (24), we need to introduce the soft-thresholding function $s_\lambda : \mathbb{R} \to \mathbb{R}, \lambda > 0$ which is defined by

$$s_\lambda(t) = \begin{cases} t - \dfrac{\lambda}{2} & \text{if } t > \dfrac{\lambda}{2} \\ 0 & \text{if } |t| \leq \dfrac{\lambda}{2} \\ t + \dfrac{\lambda}{2} & \text{if } t < -\dfrac{\lambda}{2}, \end{cases} \tag{27}$$

and the corresponding nonlinear thresholding operator $\mathbf{S}_\lambda : \ell_2 \to \ell_2$ acting componentwise as

$$[\mathbf{S}_\lambda(\beta)]_\gamma = s_{\lambda w_\gamma}(\beta_\gamma) . \tag{28}$$

We note that the soft-thresholding operator satisfies

$$\mathbf{S}_{a\lambda}(a\beta) = a\mathbf{S}_\lambda(\beta) \qquad a > 0, \beta \in \ell_2, \tag{29}$$

$$\|\mathbf{S}_\lambda(\beta) - \mathbf{S}_\lambda(\beta')\|_2 \leq \|\beta - \beta'\|_2 \qquad \beta, \beta' \in \ell_2. \tag{30}$$

These properties are immediate consequences of the fact that

$$s_{a\lambda}(at) = a s_\lambda(t) \quad a > 0, t \in \mathbb{R}$$
$$|s_\lambda(t) - s_\lambda(t')| \leq |t - t'| \quad t, t' \in \mathbb{R}.$$

Notice that (30) with $\beta' = 0$ ensures that $\mathbf{S}_\lambda(\beta) \in \ell_2$ for all $\beta \in \ell_2$.

We are ready to prove the following theorem.

Theorem 1. *Given $\varepsilon \geq 0$ and $\lambda > 0$, a vector $\beta \in \ell_2$ is a minimizer of the elastic-net functional (3) if and only if it solves the nonlinear equation*

$$\frac{1}{n}\sum_{i=1}^{n} \langle Y_i - (\Phi_n\beta)(X_i), \varphi_\gamma(X_i) \rangle - \varepsilon\lambda\beta_\gamma = \frac{\lambda}{2} w_\gamma \operatorname{sgn}(\beta_\gamma) \quad \forall \gamma \in \Gamma, \tag{31}$$

or, equivalently,

$$\beta = \mathbf{S}_\lambda\left((1 - \varepsilon\lambda)\beta + \Phi_n^*(Y - \Phi_n\beta)\right). \tag{32}$$

If $\varepsilon > 0$ the solution always exists and is unique. If $\varepsilon = 0$, $\kappa_0 > 0$ and $w_0 = \inf_{\gamma\in\Gamma} w_\gamma > 0$, the solution still exists and is unique.

Proof. If $\varepsilon > 0$ the functional $\mathcal{E}_n^\lambda$ is strictly convex, finite at 0, and it is coercive by

$$\mathcal{E}_n^\lambda(\beta) \geq p_\varepsilon(\beta) \geq \lambda\varepsilon \|\beta\|_2^2 .$$

Observing that $\|\Phi_n\beta - Y\|_n^2$ is continuous and, by Lemma 2, the elastic-net penalty is l.s.c., then $\mathcal{E}_n^\lambda$ is l.s.c. and, since ℓ_2 is reflexive, there is a unique minimizer β_n^λ in ℓ_2. If $\varepsilon = 0$, $\mathcal{E}_n^\lambda$ is convex, but the fact that $\kappa_0 > 0$ ensures that the minimizer is unique. Its existence follows from the observation that

$$\mathcal{E}_n^\lambda(\beta) \geq p_\varepsilon(\beta) \geq \lambda \|\beta\|_{1,w} \geq \lambda w_0 \|\beta\|_2 ,$$

where we used (11). In both cases the convexity of $\mathcal{E}_n^\lambda$ implies that β is a minimizer if and only if $0 \in \partial\mathcal{E}_n^\lambda(\beta)$. Since $\|\Phi_n\beta - Y\|_n^2$ is continuous, Corollary III.2.1 of [35] ensures that the subgradient is linear. Observing that $\|\Phi_n\beta - Y\|_n^2$ is differentiable with derivative $2\Phi_n^*\Phi_n\beta - 2\Phi_n^*Y$, we get

$$\partial\mathcal{E}_n^\lambda(\beta) = 2\Phi_n^*\Phi_n\beta - 2\Phi_n^*Y + \lambda\partial p_\varepsilon(\beta),$$

so that (31) follows taking into account the explicit form of $\partial p_\varepsilon(\beta)$, $\Phi_n^*\Phi_n\beta$ and Φ_n^*Y, given by Lemma 2 and Proposition 2, respectively.

We now prove (32), which is equivalent to the set of equations

$$\beta_\gamma = \mathcal{S}_{\lambda w_\gamma}\left((1-\varepsilon\lambda)\beta_\gamma + \frac{1}{n}\sum_{i=1}^n \big\langle Y_i - (\Phi_n\beta)(X_i), \varphi_\gamma(X_i)\big\rangle\right) \quad \forall\gamma \in \Gamma. \tag{33}$$

Setting $\beta'_\gamma = \big\langle Y - \Phi_n\beta, \varphi_\gamma(X)\big\rangle_n - \varepsilon\lambda\beta_\gamma$, we have $\beta_\gamma = \mathcal{S}_{\lambda w_\gamma}\left(\beta_\gamma + \beta'_\gamma\right)$ if and only if

$$\beta_\gamma = \begin{cases} \beta_\gamma + \beta'_\gamma - \dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma + \beta'_\gamma > \dfrac{\lambda w_\gamma}{2} \\ 0 & \text{if } |\beta_\gamma + \beta'_\gamma| \le \dfrac{\lambda w_\gamma}{2} \\ \beta_\gamma + \beta'_\gamma + \dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma + \beta'_\gamma < -\dfrac{\lambda w_\gamma}{2}, \end{cases}$$

that is,

$$\begin{cases} \beta'_\gamma = \dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma > 0 \\ |\beta'_\gamma| \le \dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma = 0 \quad \text{or else} \quad \beta'_\gamma = \dfrac{\lambda w_\gamma}{2}\operatorname{sgn}(\beta_\gamma) \\ \beta'_\gamma = -\dfrac{\lambda w_\gamma}{2} & \text{if } \beta_\gamma < 0 \end{cases}$$

which is equivalent to (31). □

The following corollary gives some more information about the characterization of the solution as the fixed point of a contractive map. In particular, it provides an explicit expression for the Lipschitz constant of this map and it shows how it depends on the spectral properties of the empirical mean of the Gram matrix and on the regularization parameter λ.

Corollary 1. *Let $\varepsilon \ge 0$ and $\lambda > 0$. Pick any arbitrary $\tau > 0$. Then β is a minimizer of $\mathcal{E}_n^\lambda$ in ℓ_2 if and only if it is a fixed point of the following Lipschitz map $\mathcal{T}_n : \ell_2 \to \ell_2$, namely*

$$\beta = \mathcal{T}_n\beta \quad \text{where } \mathcal{T}_n\beta = \frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left((\tau I - \Phi_n^*\Phi_n)\beta + \Phi_n^*Y\right). \tag{34}$$

With the choice $\tau = \frac{\kappa_0+\kappa}{2}$, the Lipschitz constant is bounded by

$$q = \frac{\kappa - \kappa_0}{\kappa + \kappa_0 + 2\varepsilon\lambda} \le 1.$$

In particular, with this choice of τ and if $\varepsilon > 0$ or $\kappa_0 > 0$, $\mathcal{T}_n$ is a contraction.

Proof. Clearly β is a minimizer of $\mathcal{E}_n^\lambda$ if and only if it is a minimizer of $\frac{1}{\tau+\varepsilon\lambda}\mathcal{E}_n^\lambda$, which means that, in (32), we can replace λ with $\frac{\lambda}{\tau+\varepsilon\lambda}$, Φ_n by $\frac{1}{\sqrt{\tau+\varepsilon\lambda}}\Phi_n$ and Y by $\frac{1}{\sqrt{\tau+\varepsilon\lambda}}Y$. Hence β is a minimizer of $\mathcal{E}_n^\lambda$ if and only if it is a solution of

$$\beta = \mathbf{S}_{\frac{\lambda}{\tau+\varepsilon\lambda}}\left(\left(1 - \frac{\varepsilon\lambda}{\tau+\varepsilon\lambda}\right)\beta + \frac{1}{\tau+\varepsilon\lambda}\Phi_n^*(Y - \Phi_n\beta)\right).$$

Therefore, by (29) with $a = \frac{1}{\tau+\varepsilon\lambda}$, β is a minimizer of $\mathcal{E}_n^\lambda$ if and only if $\beta = \mathcal{T}_n\beta$.

We show that $\mathcal{T}_n$ is Lipschitz and calculate explicitly a bound on the Lipschitz constant. By assumption we have $\kappa_0 I \le \Phi_n^*\Phi_n \le \kappa I$; then, by the Spectral Theorem,

$$\left\|\tau I - \Phi_n^*\Phi_n\right\|_{\ell_2,\ell_2} \le \max\left\{|\tau-\kappa_0|, |\tau-\kappa|\right\},$$

where $\|\cdot\|_{\ell_2,\ell_2}$ denotes the operator norm of a bounded operator on ℓ_2. Hence, using (30), we get

$$\begin{aligned}\left\|\mathcal{T}_n\beta - \mathcal{T}_n\beta'\right\|_2 &\le \frac{1}{\tau+\varepsilon\lambda}\left\|(\tau I - \Phi_n^*\Phi_n)(\beta-\beta')\right\|_2\\ &\le \max\left\{\left|\frac{\tau-\kappa_0}{\tau+\varepsilon\lambda}\right|, \left|\frac{\tau-\kappa}{\tau+\varepsilon\lambda}\right|\right\}\left\|\beta-\beta'\right\|_2\\ &=: q\left\|\beta-\beta'\right\|_2.\end{aligned}$$

The minimum of q with respect to τ is obtained for

$$\frac{\tau-\kappa_0}{\tau+\varepsilon\lambda} = \frac{\kappa-\tau}{\tau+\varepsilon\lambda},$$

that is, $\tau = \frac{\kappa+\kappa_0}{2}$, and, with this choice, we get

$$q = \frac{\kappa-\kappa_0}{\kappa+\kappa_0+2\varepsilon\lambda}. \quad \square$$

By inspecting the proof, we notice that the choice $\tau = \frac{\kappa_0+\kappa}{2}$ provides the best possible Lipschitz constant under the assumption that $\kappa_0 I \le \Phi_n^*\Phi_n \le \kappa I$. If $\varepsilon > 0$ or $\kappa_0 > 0$, $\mathcal{T}_n$ is a contraction and β_n^λ can be computed by means of the Banach fixed point theorem. If $\varepsilon = 0$ and $\kappa_0 = 0$, $\mathcal{T}_n$ is only non-expansive, so that proving the convergence of the successive approximation scheme is not straightforward.[2]

Let us now write down explicitly the iterative procedure suggested by Corollary 1 to compute β_n^λ. Define the iterative scheme by

$$\begin{aligned}&\beta^0 = 0,\\ &\beta^\ell = \frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left((\tau I - \Phi_n^*\Phi_n)\beta^{\ell-1} + \Phi_n^*Y\right)\end{aligned}$$

with $\tau = \frac{\kappa_0+\kappa}{2}$. The following corollary shows that the β^ℓ converges to β_n^λ when ℓ goes to infinity.

Corollary 2. *Assume that $\varepsilon > 0$ or $\kappa_0 > 0$. For any $\ell \in \mathbb{N}$ the following inequality holds*

$$\left\|\beta^\ell - \beta_n^\lambda\right\|_2 \le \frac{(\kappa-\kappa_0)^\ell}{(\kappa+\kappa_0+2\varepsilon\lambda)^\ell(\kappa_0+\varepsilon\lambda)}\left\|\Phi_n^*Y\right\|_2. \tag{35}$$

In particular, $\lim_{\ell\to\infty}\left\|\beta^\ell - \beta_n^\lambda\right\|_2 = 0$.

Proof. Since $\mathcal{T}_n$ is a contraction with Lipschitz constant $q = \frac{\kappa-\kappa_0}{\kappa+\kappa_0+2\varepsilon\lambda} < 1$, the Banach fixed point theorem applies and the sequence $\left(\beta^\ell\right)_{\ell\in\mathbb{N}}$ converges to the unique fixed point of $\mathcal{T}_n$, which is β_n^λ by Corollary 1. Moreover we can use the Lipschitz property of $\mathcal{T}_n$ to write

$$\begin{aligned}\left\|\beta^\ell - \beta_n^\lambda\right\|_2 &\le \left\|\beta^\ell - \beta^{\ell+1}\right\|_2 + \left\|\beta^{\ell+1} - \beta_n^\lambda\right\|_2\\ &\le q\left\|\beta^{\ell-1} - \beta^\ell\right\|_2 + q\left\|\beta^\ell - \beta_n^\lambda\right\|_2\\ &\le q^\ell\left\|\beta^0 - \beta^1\right\|_2 + q\left\|\beta^\ell - \beta_n^\lambda\right\|_2,\end{aligned}$$

[2] Interestingly, it was proved in [19] using different arguments that the same iterative scheme can still be used for the case $\varepsilon = 0$ and $\kappa_0 = 0$.

so that we immediately get

$$\left\|\beta^{\ell}-\beta_n^{\lambda}\right\|_2 \le \frac{q^{\ell}}{1-q}\left\|\beta^1-\beta^0\right\|_2 \le \frac{(\kappa-\kappa_0)^{\ell}}{(\kappa_0+\kappa+2\varepsilon\lambda)^{\ell}(\kappa_0+\varepsilon\lambda)}\left\|\Phi_n^* Y\right\|_2$$

since $\beta^0=0$, $\beta^1=\frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left(\Phi_n^* Y\right)$ and $1-q=\frac{2(\kappa_0+\varepsilon\lambda)}{\kappa_0+\kappa+2\varepsilon\lambda}$. □

Let us remark that bound (35) provides a natural stopping rule for the number of iterations, namely to select ℓ such that $\left\|\beta^{\ell}-\beta_n^{\lambda}\right\|_2 \le \eta$, where η is a bound on the distance between the estimator β_n^{λ} and the true solution. For example, if $\left\|\Phi_n^* Y\right\|_2$ is bounded by M and if $\kappa_0=0$, the stopping rule is

$$\ell_{\text{stop}} \ge \frac{\log\frac{M}{\varepsilon\lambda\eta}}{\log(1+\frac{2\varepsilon\lambda}{\kappa})} \quad \text{so that } \left\|\beta^{\ell_{\text{stop}}}-\beta_n^{\lambda}\right\|_2 \le \eta.$$

Note that in the case of an infinite-dimensional dictionary, the above iteration involves infinite-dimensional matrices. In Section 3.2 we will show that under mild assumptions on the weights it is always possible to reduce the problem to a finite-dimensional one.

Finally we notice that all previous results also hold when considering the distribution-dependent version of the method. The following proposition summarizes the results in this latter case.

Proposition 4. *Let $\varepsilon \ge 0$ and $\lambda > 0$. Pick any arbitrary $\tau > 0$. Then a vector $\beta \in \ell_2$ is a minimizer of*

$$\mathcal{E}^{\lambda}(\beta) = \mathbb{E}\left[|\Phi_P\beta - Y|^2\right] + \lambda p_{\varepsilon}(\beta).$$

if and only if it is a fixed point of the following Lipschitz map, namely

$$\beta = \mathcal{T}\beta \quad \text{where } \mathcal{T}\beta = \frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left((\tau I - \Phi_P^*\Phi_P)\beta + \Phi_P^* Y\right). \tag{36}$$

If $\varepsilon > 0$ or $\kappa_0 > 0$, the minimizer is unique.

If it is unique, we denote it by β^{λ}:

$$\beta^{\lambda} = \operatorname*{argmin}_{\beta\in\ell_2}\left(\mathbb{E}\left[|\Phi_P\beta - Y|^2\right] + \lambda p_{\varepsilon}(\beta)\right). \tag{37}$$

We add a comment. Under Assumption 2 and the definition of β^{ε}, the statistical model is $Y = \Phi_P\beta^{\varepsilon} + W$ where W has zero mean, so that β^{λ} is also the minimizer of

$$\|\Phi_P\beta - \Phi_P\beta^{\varepsilon}\|_P^2 + \lambda p_{\varepsilon}(\beta). \tag{38}$$

3.2. Sparsity properties

The results of the previous section immediately yield a crude estimate of the number and localization of the non-zero coefficients of our estimator. Indeed, although the set of features could be infinite, β_n^{λ} has only a finite number of coefficients different from zero provided that the sequence of weights is bounded away from zero.

Corollary 3. *Assume that the family of weights satisfies $\inf_{\gamma\in\Gamma} w_\gamma > 0$, then for any $\beta \in \ell_2$, the support of $\mathbf{S}_\lambda(\beta)$ is finite. In particular, β_n^{λ}, β^{ℓ} and β^{λ} are all finitely supported.*

Proof. Let $w_0 = \inf_{\gamma\in\Gamma} w_\gamma > 0$. Since $\sum_{\gamma\in\Gamma}|\beta_\gamma|^2 < +\infty$, there is a finite subset $\Gamma_0 \subset \Gamma$ such that $|\beta_\gamma| \le \frac{\lambda}{2}w_0 \le \frac{\lambda}{2}w_\gamma$ for all $\gamma \notin \Gamma_0$. This implies that

$$\mathcal{S}_{\lambda w_\gamma}\left(\beta_\gamma\right) = 0 \quad \text{for } \gamma \notin \Gamma_0,$$

by the definition of soft-thresholding, so that the support of $\mathbf{S}_\lambda(\beta)$ is contained in Γ_0. Eqs. (32) and (36) and the definition of β^{ℓ} imply that β_n^{λ}, β^{λ} and β^{ℓ} have finite support. □

However, the supports of β^ℓ and β_n^λ are not known a priori and to compute β^ℓ one would need to store the infinite matrix $\Phi_n^*\Phi_n$. The following corollary suggests a strategy to overcome this problem.

Corollary 4. *Given $\varepsilon \geq 0$ and $\lambda > 0$, let*

$$\Gamma_\lambda = \left\{ \gamma \in \Gamma \mid \left\|\varphi_\gamma\right\|_n \neq 0 \quad \text{and} \quad w_\gamma \leq \frac{2\,\|Y\|_n\,(\left\|\varphi_\gamma\right\|_n + \sqrt{\varepsilon\lambda})}{\lambda} \right\}$$

then

$$\operatorname{supp}(\beta_n^\lambda) \subset \Gamma_\lambda. \tag{39}$$

Proof. If $\left\|\varphi_\gamma\right\|_n = 0$, clearly $\beta_\gamma = 0$ is a solution of (31). Let $M = \|Y\|_n$; the definition of β_n^λ as the minimizer of (24) yields the bound $\mathcal{E}_n^\lambda(\beta_n^\lambda) \leq \mathcal{E}_n^\lambda(0) = M^2$, so that

$$\left\|\Phi_n\beta_n^\lambda - Y\right\|_n \leq M \qquad p_\varepsilon(\beta_n^\lambda) \leq \frac{M^2}{\lambda}.$$

Hence, for all $\gamma \in \Gamma$, the second inequality gives that $\varepsilon\lambda(\beta_n^\lambda)_\gamma^2 \leq M^2$, and we have

$$\left|\left\langle Y - \Phi_n\beta_n^\lambda, \varphi_\gamma(X)\right\rangle_n - \varepsilon\lambda(\beta_n^\lambda)_\gamma\right| \leq M\left(\left\|\varphi_\gamma\right\|_n + \sqrt{\varepsilon\lambda}\right)$$

and, therefore, by (31),

$$|\operatorname{sgn}((\beta_n^\lambda)_\gamma)| \leq \frac{2M(\left\|\varphi_\gamma\right\|_n + \sqrt{\varepsilon\lambda})}{\lambda w_\gamma}.$$

Since $|\operatorname{sgn}((\beta_n^\lambda)_\gamma)| = 1$ when $(\beta_n^\lambda)_\gamma \neq 0$, this implies that $(\beta_n^\lambda)_\gamma = 0$ if $\frac{2M(\left\|\varphi_\gamma\right\|_n+\sqrt{\varepsilon\lambda})}{\lambda w_\gamma} < 1$. □

Now, let Γ' be the set of indices γ such that the corresponding feature $\varphi_\gamma(X_i) \neq 0$ for some $i = 1, \ldots, n$. If the family of corresponding weights $(w_\gamma)_{\gamma\in\Gamma'}$ goes to infinity,[3] then Γ_λ is always finite.

Remark 2. This last property has immediate computational implications. Since $\operatorname{supp}(\beta_n^\lambda) \subset \Gamma_\lambda$, one can replace Γ with Γ_λ in the definition of Φ_n so that $\Phi_n^*\Phi_n$ is a finite matrix and Φ_n^*Y is a finite vector. In particular the iterative procedure given by Corollary 1 can then be implemented by means of finite matrices.

Finally, by inspecting the proof above one sees that a similar result holds true for the distribution-dependent minimizer β^λ. Its support is always finite, as already noticed, and moreover is included in the following set

$$\left\{ \gamma \in \Gamma \mid \left\|\varphi_\gamma\right\|_P \neq 0 \quad \text{and} \quad w_\gamma \leq \frac{2\,\|Y\|_P\,(\left\|\varphi_\gamma\right\|_P + \sqrt{\varepsilon\lambda})}{\lambda} \right\}.$$

4. Probabilistic error estimates

In this section we provide an error analysis for the elastic-net regularization scheme. Our primary goal is the *variable selection problem*, so that we need to control the error $\left\|\beta_n^{\lambda_n} - \beta\right\|_2$, where λ_n is a suitable choice of the regularization parameter as a function of the data, and β is an explanatory vector encoding the features that are relevant to reconstructing the regression function f^*, that is, such that $f^* = \Phi_P\beta$. Although Assumption (7) implies that the above equation has at least a solution β^* with

[3] The sequence $(w_\gamma)_{\gamma\in\Gamma'}$ goes to infinity, if for all $M > 0$ there exists a finite set Γ_M such that $|w_\gamma| > M, \forall\gamma \notin \Gamma_M$.

$p_\varepsilon(\beta^*) < \infty$, nonetheless, the operator Φ_P is injective only if $(\varphi_\gamma(X))_{\gamma\in\Gamma}$ is ℓ_2-linearly independent in $L^2_y(P)$. As usually done for inverse problems, to restore uniqueness we choose, among all the vectors β such that $f^* = \Phi_P\beta$, the vector β^ε which is the minimizer of the elastic-net penalty. The vector β^ε can be regarded as the *best* representation of the regression function f^* according to the elastic-net penalty and we call it the *elastic-net representation*. Clearly this representation will depend on ε.

Next we focus on the following error decomposition (for any fixed positive λ),

$$\left\|\beta_n^\lambda - \beta^\varepsilon\right\|_2 \le \left\|\beta_n^\lambda - \beta^\lambda\right\|_2 + \left\|\beta^\lambda - \beta^\varepsilon\right\|_2, \tag{40}$$

where β^λ is given by (37). The first error term in the right-hand side of the above inequality is due to finite sampling and will be referred to as the *sample error*, whereas the second error term is deterministic and is called the *approximation error*. In Section 4.2 we analyze the sample error via concentration inequalities and we consider the behavior of the approximation error as a function of the regularization parameter λ. The analysis of these error terms leads us to discuss the choice of λ and to derive statistical consistency results for elastic-net regularization. In Section 4.3 we discuss a priori and a posteriori (adaptive) parameter choices.

4.1. Identifiability condition and elastic-net representation

The following proposition provides a way to define a unique solution of the equation $f^* = \Phi_P\beta$. Let

$$\mathcal{B} = \left\{\beta \in \ell_2 \mid \Phi_P\beta = f^*(X)\right\} = \beta^* + \ker\Phi_P$$

where $\beta^* \in \ell_2$ is given by (7) in Assumption 2 and

$$\ker\Phi_P = \{\beta \in \ell_2 \mid \Phi_P\beta = 0\} = \left\{\beta \in \ell_2 \mid f_\beta(X) = 0 \text{ with probability } 1\right\}.$$

Proposition 5. *If $\varepsilon > 0$ or $\kappa_0 > 0$, there is a unique $\beta^\varepsilon \in \ell_2$ such that*

$$p_\varepsilon(\beta^\varepsilon) = \inf_{\beta\in\mathcal{B}} p_\varepsilon(\beta). \tag{41}$$

Proof. If $\kappa_0 > 0$, $\mathcal{B}$ reduces to a single point, so that there is nothing to prove. If $\varepsilon > 0$, $\mathcal{B}$ is a closed subset of a reflexive space. Moreover, by Lemma 2, the penalty $p_\varepsilon(\cdot)$ is strictly convex, l.s.c. and, by (7) of Assumption 2, there exists at least one $\beta^* \in \mathcal{B}$ such that $p_\varepsilon(\beta^*)$ is finite. Since $p_\varepsilon(\beta) \ge \varepsilon\|\beta\|_2^2$, $p_\varepsilon(\cdot)$ is coercive. A standard result of convex analysis implies that the minimizer exists and is unique. □

4.2. Consistency: Sample and approximation errors

The main result of this section is a probabilistic error estimate for $\left\|\beta_n^\lambda - \beta^\lambda\right\|_2$, which will provide a choice $\lambda = \lambda_n$ for the regularization parameter as well as a convergence result for $\left\|\beta_n^{\lambda_n} - \beta^\varepsilon\right\|_2$.

We first need to establish two lemmas. The first one shows that the sample error can be studied in terms of the following quantities

$$\left\|\Phi_n^*\Phi_n - \Phi_P^*\Phi_P\right\|_{\mathrm{HS}} \quad \text{and} \quad \left\|\Phi_n^* W\right\|_2 \tag{42}$$

measuring the perturbation due to random sampling and noise (we recall that $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm of a Hilbert–Schmidt operator on ℓ_2). The second lemma provides suitable probabilistic estimates for these quantities.

Lemma 3. *Let $\varepsilon \ge 0$ and $\lambda > 0$. If $\varepsilon > 0$ or $\kappa_0 > 0$, then*

$$\left\|\beta_n^\lambda - \beta^\lambda\right\|_2 \le \frac{1}{\kappa_0 + \varepsilon\lambda}\left(\left\|(\Phi_n^*\Phi_n - \Phi_P^*\Phi_P)(\beta^\lambda - \beta^\varepsilon)\right\|_2 + \left\|\Phi_n^* W\right\|_2\right). \tag{43}$$

Proof. Let $\tau = \frac{\kappa_0+\kappa}{2}$ and recall that β_n^λ and β^λ satisfy (34) and (36), respectively. Taking into account (30) we get

$$\left\|\beta_n^\lambda - \beta^\lambda\right\|_2 \leq \frac{1}{\tau+\varepsilon\lambda}\left\|(\tau\beta_n^\lambda - \Phi_n^*\Phi_n\beta_n^\lambda + \Phi_n^*Y) - (\tau\beta^\lambda - \Phi_P^*\Phi_P\beta^\lambda + \Phi_P^*Y)\right\|_2. \tag{44}$$

By Assumption 2 and the definition of β^ε, $Y = f^*(X) + W$, and $\Phi_P\beta^\varepsilon$ and $\Phi_n\beta^\varepsilon$ both coincide with the function f^*, regarded as an element of $L^2_{\mathcal{Y}}(P)$ or $L^2_{\mathcal{Y}}(\mathbb{P}_n)$ respectively. Moreover by (8) $\Phi_P^*W = 0$, so that

$$\Phi_n^*Y - \Phi_P^*Y = (\Phi_n^*\Phi_n - \Phi_P^*\Phi_P)\beta^\varepsilon + \Phi_n^*W.$$

Moreover

$$(\tau I - \Phi_n^*\Phi_n)\beta_n^\lambda - (\tau I - \Phi_P^*\Phi_P)\beta^\lambda = (\tau I - \Phi_n^*\Phi_n)(\beta_n^\lambda - \beta^\lambda) - (\Phi_n^*\Phi_n - \Phi_P^*\Phi_P)\beta^\lambda.$$

From the assumption on $\Phi_n^*\Phi_n$ and the choice $\tau = \frac{\kappa+\kappa_0}{2}$, we have $\left\|\tau I - \Phi_n^*\Phi_n\right\|_{\ell_2,\ell_2} \leq \frac{\kappa-\kappa_0}{2}$, so that (44) gives

$$(\tau+\varepsilon\lambda)\left\|\beta_n^\lambda - \beta^\lambda\right\|_2 \leq \left\|(\Phi_n^*\Phi_n - \Phi_P^*\Phi_P)(\beta^\lambda - \beta^\varepsilon)\right\|_2 + \left\|\Phi_n^*W\right\|_2 + \frac{\kappa-\kappa_0}{2}\left\|\beta_n^\lambda - \beta^\lambda\right\|_2.$$

Bound (43) is established by observing that $\tau + \varepsilon\lambda - (\kappa-\kappa_0)/2 = \kappa_0 + \varepsilon\lambda$. □

The probabilistic estimates for (42) are straightforward consequences of the law of large numbers for vector-valued random variables. More precisely, we recall the following probabilistic inequalities based on a result of [36,37]; see also Th. 3.3.4 of [38] as well as [39] for concentration inequalities for Hilbert-space-valued random variables.

Proposition 6. *Let $(\xi_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. zero-mean random variables taking values in a real separable Hilbert space $\mathcal{H}$ and satisfying*

$$\mathbb{E}[\|\xi_i\|_{\mathcal{H}}^m] \leq \frac{1}{2}m!M^2H^{m-2} \quad \forall m \geq 2, \tag{45}$$

where M and H are two positive constants. Then, for all $n \in \mathbb{N}$ and $\eta > 0$

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_{\mathcal{H}} \geq \eta\right] \leq 2e^{-\frac{n\eta^2}{M^2+H\eta+M\sqrt{M^2+2H\eta}}} = 2e^{-n\frac{M^2}{H^2}g\left(\frac{H\eta}{M^2}\right)} \tag{46}$$

where $g(t) = \frac{t^2}{1+t+\sqrt{1+2t}}$, or, for all $\delta > 0$,

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_{\mathcal{H}} \leq \left(\frac{H\delta}{n} + \frac{M\sqrt{2\delta}}{\sqrt{n}}\right)\right] \geq 1 - 2e^{-\delta}. \tag{47}$$

Proof. Bound (46) is given in [36] with a wrong factor, see [37]. To show (47), observe that the inverse of the function $\frac{t^2}{1+t+\sqrt{1+2t}}$ is the function $t + \sqrt{2t}$ so that the equation

$$2e^{-n\frac{M^2}{H^2}g\left(\frac{H\eta}{M^2}\right)} = 2e^{-\delta}$$

has the solution

$$\eta = \frac{M^2}{H}\left(\frac{H^2\delta}{nM^2} + \sqrt{2\frac{H^2\delta}{nM^2}}\right). \quad \square$$

Lemma 4. *With probability greater than* $1 - 4e^{-\delta}$, *the following two inequalities hold, for any* $\lambda > 0$ *and* $\varepsilon > 0$,

$$\left\|\Phi_n^* W\right\|_2 \leq \left(\frac{L\sqrt{\kappa}\delta}{n} + \frac{\sigma\sqrt{\kappa}\sqrt{2\delta}}{\sqrt{n}}\right) \leq \underbrace{\frac{\sqrt{2\kappa\delta}(\sigma + L)}{\sqrt{n}}}_{\text{if } \delta \leq n} \tag{48}$$

and

$$\left\|\Phi_n^* \Phi_n - \Phi_P^* \Phi_P\right\|_{HS} \leq \left(\frac{\kappa\delta}{n} + \frac{\kappa\sqrt{2\delta}}{\sqrt{n}}\right) \leq \underbrace{\frac{3\kappa\sqrt{\delta}}{\sqrt{n}}}_{\text{if } \delta \leq n}. \tag{49}$$

Proof. Consider the ℓ_2-valued random variable $\Phi_X^* W$. From (8), $\mathbb{E}\left[\Phi_X^* W\right] = \mathbb{E}\left[\mathbb{E}\left[\Phi_X^* W | X\right]\right] = 0$ and, for any $m \geq 2$,

$$\mathbb{E}\left[\left\|\Phi_X^* W\right\|_2^m\right] = \mathbb{E}\left[\left(\sum_{\gamma \in \Gamma} \left|\langle \varphi_\gamma(X), W\rangle\right|^2\right)^{\frac{m}{2}}\right] \leq \kappa^{\frac{m}{2}} \mathbb{E}\left[|W|^m\right] \leq \kappa^{\frac{m}{2}} \frac{m!}{2} \sigma^2 L^{m-2},$$

due to (5) and (10). Applying (47) with $H = \sqrt{\kappa} L$ and $M = \sqrt{\kappa}\sigma$, and recalling definition (19), we get that

$$\|\Phi_n^* W\|_2 \leq \frac{\sqrt{\kappa} L \delta}{n} + \frac{\sqrt{\kappa}\sigma\sqrt{2\delta}}{\sqrt{n}}$$

with probability greater than $1 - 2e^{-\delta}$.

Consider the random variable $\Phi_X \Phi_X^*$ taking values in the Hilbert space of Hilbert–Schmidt operators (where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm). One has that $\mathbb{E}\left[\Phi_X \Phi_X^*\right] = \Phi_P \Phi_P^*$ and, by (13)

$$\left\|\Phi_X \Phi_X^*\right\|_{\mathrm{HS}} \leq \mathrm{Tr}\left(\Phi_X \Phi_X^*\right) \leq \kappa.$$

Hence

$$\begin{aligned}\mathbb{E}\left[\left\|\Phi_X \Phi_X^* - \Phi_P \Phi_P^*\right\|_{\mathrm{HS}}^m\right] &\leq \mathbb{E}\left[\left\|\Phi_X \Phi_X^* - \Phi_P \Phi_P^*\right\|_{\mathrm{HS}}^2\right] (2\kappa)^{m-2} \\ &\leq \frac{m!}{2} \kappa^2 \kappa^{m-2},\end{aligned}$$

by $m! \geq 2^{m-1}$. Applying (47) with $H = M = \kappa$

$$\left\|\Phi_n \Phi_n^* - \Phi_P \Phi_P^*\right\|_{\mathrm{HS}} \leq \frac{\kappa\delta}{n} + \frac{\kappa\sqrt{2\delta}}{\sqrt{n}},$$

with probability greater than $1 - 2e^{-\delta}$. The simplified bounds are clear provided that $\delta \leq n$. □

Remark 3. In both (48) and (49), the condition $\delta \leq n$ allows simplifying the bounds enlightening the dependence on n and the confidence level $1 - 4e^{-\delta}$. In the following results we always assume that $\delta \leq n$, but we stress the fact that this condition is only needed to simplify the form of the bounds. Moreover, observe that, for a fixed confidence level, this requirement on n is very weak – for example, to achieve a 99% confidence level, we only need to require that $n \geq 6$.

The next proposition gives a bound on the sample error. This bound is uniform in the regularization parameter λ in the sense that there exists an event independent of λ such that its probability is greater than $1 - 4e^{-\delta}$ and (50) holds true.

Proposition 7. *Assume that $\varepsilon > 0$ or $\kappa_0 > 0$. Let $\delta > 0$ and $n \in \mathbb{N}$ such that $\delta \leq n$, for any $\lambda > 0$ the bound*

$$\left\|\beta_n^\lambda - \beta^\lambda\right\|_2 \leq \frac{c\sqrt{\delta}}{\sqrt{n}(\kappa_0 + \varepsilon\lambda)} \left(1 + \left\|\beta^\lambda - \beta^\varepsilon\right\|_2\right) \tag{50}$$

holds with probability greater than $1 - 4e^{-\delta}$, where $c = \max\left\{\sqrt{2\kappa}(\sigma + L), 3\kappa\right\}$.

Proof. Plug bounds (49) and (48) in (43), taking into account that

$$\left\|(\Phi_n^*\Phi_n - \Phi_P^*\Phi_P)(\beta^\lambda - \beta^\varepsilon)\right\|_2 \leq \left\|\Phi_n^*\Phi_n - \Phi_P^*\Phi_P\right\|_{\mathrm{HS}} \left\|\beta^\lambda - \beta^\varepsilon\right\|_2. \quad \square$$

By inspecting the proof, one sees that the constant κ_0 in (43) can be replaced by any constant κ_λ such that

$$\kappa_0 \leq \kappa_\lambda \leq \inf_{\beta \in \ell_2 | \|\beta\|_2 = 1} \left\| \sum_{\gamma \in \Gamma_\lambda} \beta_\gamma \varphi_\gamma \right\|_n^2 \text{ with probability 1,}$$

where Γ_λ is the set of *active features* given by Corollary 4. If $\kappa_0 = 0$ and $\kappa_\lambda > 0$, i.e. when Γ_λ is finite and the active features are linearly independent, one can improve bound (52) below. Since we mainly focus on the case of linearly-dependent dictionaries we will not discuss this point any further.

The following proposition shows that the approximation error $\left\|\beta^\lambda - \beta^\varepsilon\right\|_2$ tends to zero when λ tends to zero.

Proposition 8. *If $\varepsilon > 0$ then*

$$\lim_{\lambda \to 0} \left\|\beta^\lambda - \beta^\varepsilon\right\|_2 = 0.$$

Proof. It is enough to prove the result for an arbitrary sequence $(\lambda_j)_{j\in\mathbb{N}}$ converging to 0. Putting $\beta^j = \beta^{\lambda_j}$, since $\|\Phi_P\beta - Y\|_P^2 = \|\Phi_P\beta - f^*(X)\|_P^2 + \|f^*(X) - Y\|_P^2$, by the definition of β^j as the minimizer of (37) and the fact that β^ε solves $\Phi_P\beta = f^*$, we get

$$\left\|\Phi_P\beta^j - f^*(X)\right\|_P^2 + \lambda_j p_\varepsilon(\beta^j) \leq \left\|\Phi_P\beta^\varepsilon - f^*(X)\right\|_P^2 + \lambda_j p_\varepsilon(\beta^\varepsilon) = \lambda_j p_\varepsilon(\beta^\varepsilon).$$

Condition (7) of Assumption 1 ensures that $p_\varepsilon(\beta^\varepsilon)$ is finite, so that

$$\left\|\Phi_P\beta^j - f^*(X)\right\|_P^2 \leq \lambda_j p_\varepsilon(\beta^\varepsilon) \quad \text{and} \quad p_\varepsilon(\beta^j) \leq p_\varepsilon(\beta^\varepsilon).$$

Since $\varepsilon > 0$, the last inequality implies that $(\beta^j)_{j\in\mathbb{N}}$ is a bounded sequence in ℓ_2. Hence, possibly passing to a subsequence, $(\beta^j)_{j\in\mathbb{N}}$ converges weakly to some β_*. We claim that $\beta_* = \beta^\varepsilon$. Since $\beta \mapsto \|\Phi_P\beta - f^*(X)\|_P^2$ is l.s.c.

$$\left\|\Phi_P\beta_* - f^*(X)\right\|_P^2 \leq \liminf_{j\to\infty} \left\|\Phi_P\beta^j - f^*(X)\right\|_P^2 \leq \liminf_{j\to\infty} \lambda_j p_\varepsilon(\beta^\varepsilon) = 0,$$

that is $\beta_* \in \mathcal{B}$. Since $p_\varepsilon(\cdot)$ is l.s.c.,

$$p_\varepsilon(\beta_*) \leq \liminf_{j\to\infty} p_\varepsilon(\beta^j) \leq p_\varepsilon(\beta^\varepsilon).$$

By the definition of β^ε, it follows that $\beta_* = \beta^\varepsilon$ and, hence,

$$\lim_{j\to\infty} p_\varepsilon(\beta^j) = p_\varepsilon(\beta^\varepsilon). \tag{51}$$

To prove that β^j converges to β^ε in ℓ_2, it is enough to show that $\lim_{j\to\infty} \left\|\beta^j\right\|_2 = \|\beta^\varepsilon\|_2$. Since $\|\cdot\|_2$ is l.s.c., $\liminf_{j\to\infty} \left\|\beta^j\right\|_2 \geq \|\beta^\varepsilon\|_2$. Hence we are left with proving that $\limsup_{j\to\infty} \left\|\beta^j\right\|_2 \leq \|\beta^\varepsilon\|_2$. Assume the contrary. This implies that, possibly passing to a subsequence,

$$\lim_{j\to\infty} \left\|\beta^j\right\|_2 > \|\beta^\varepsilon\|_2$$

and, using (51),

$$\lim_{j\to\infty}\sum_{\gamma\in\Gamma} w_\gamma|\beta^j_\gamma| < \sum_{\gamma\in\Gamma} w_\gamma|\beta^\varepsilon|.$$

However, since $\beta \mapsto \sum_{\gamma\in\Gamma} w_\gamma|\beta_\gamma|$ is l.s.c.

$$\liminf_{j\to\infty}\sum_{\gamma\in\Gamma} w_\gamma|\beta^j_\gamma| \geq \sum_{\gamma\in\Gamma} w_\gamma|\beta^\varepsilon|. \quad \square$$

From (50) and the triangular inequality, we easily deduce that

$$\left\|\beta^\lambda_n - \beta^\varepsilon\right\|_2 \leq \frac{c\sqrt{\delta}}{\sqrt{n}(\kappa_0+\varepsilon\lambda)}\left(1+\left\|\beta^\lambda - \beta^\varepsilon\right\|_2\right) + \left\|\beta^\lambda - \beta^\varepsilon\right\|_2 \tag{52}$$

with probability greater than $1-4\mathrm{e}^{-\delta}$. Since the tails are exponential, the above bound and the Borel–Cantelli lemma imply the following theorem, which states that the estimator β^λ_n converges to the solution β^ε, for a suitable choice of the regularization parameter λ.

Theorem 2. *Assume that $\varepsilon > 0$ and $\kappa_0 = 0$. Let λ_n be a choice of λ as a function of n such that $\lim_{n\to\infty}\lambda_n = 0$ and $\lim_{n\to\infty} n\lambda_n^2 - 2\log n = +\infty$. Then*

$$\lim_{n\to\infty}\left\|\beta^{\lambda_n}_n - \beta^\varepsilon\right\|_2 = 0 \text{ with probability 1.}$$

If $\kappa_0 > 0$, the above convergence result holds for any choice of λ_n such that $\lim_{n\to\infty}\lambda_n = 0$.

Proof. The only non-trivial statement concerns the convergence with probability 1. We give the proof only for $\kappa_0 = 0$, the other one being similar. Let $(\lambda_n)_{n\geq 1}$ be a sequence such that $\lim_{n\to\infty}\lambda_n = 0$ and $\lim_{n\to\infty} n\lambda_n^2 - 2\log n = +\infty$. Since $\lim_{n\to\infty}\lambda_n = 0$, Proposition 8 ensures that $\lim_{n\to\infty}\left\|\beta^{\lambda_n} - \beta^\varepsilon\right\|_2 = 0$. Hence, it is enough to show that $\lim_{n\to\infty}\left\|\beta^{\lambda_n}_n - \beta^{\lambda_n}\right\|_2 = 0$ with probability 1. Let $D = \sup_{n\geq 1}\varepsilon^{-1}c(1+\left\|\beta^{\lambda_n} - \beta^\varepsilon\right\|_2)$, which is finite since the approximation error goes to zero if λ tends to zero. Given $\eta > 0$, let $\delta = n\lambda_n^2\frac{\eta^2}{D^2} \leq n$ for n large enough, so that bound (50) holds providing that

$$\mathbb{P}\left[\left\|\beta^{\lambda_n}_n - \beta^{\lambda_n}\right\|_2 \geq \eta\right] \leq 4\mathrm{e}^{-n\lambda_n^2\frac{\eta^2}{D^2}}.$$

The condition that $\lim_{n\to\infty} n\lambda_n^2 - 2\log n = +\infty$ implies that the series $\sum_{n=1}^\infty \mathrm{e}^{-n\lambda_n^2\frac{\eta^2}{D^2}}$ converges and the Borel–Cantelli lemma gives the thesis. $\square$

Remark 4. The two conditions on λ_n in the above theorem are clearly satisfied with the choice $\lambda_n = (1/n)^r$ with $0 < r < \frac{1}{2}$. Moreover, by inspecting the proof, one can easily check that to have the convergence of $\beta^{\lambda_n}_n$ to β^ε in probability, it is enough to require that $\lim_{n\to\infty}\lambda_n = 0$ and $\lim_{n\to\infty} n\lambda_n^2 = +\infty$.

Let $f_n = f_{\beta^{\lambda_n}_n}$. Since $f^* = f_{\beta^\varepsilon}$ and $\mathbb{E}\left[|f_n(X) - f^*(X)|^2\right] = \left\|\Phi_P(\beta^{\lambda_n}_n - \beta^\varepsilon)\right\|^2_P$, the above theorem implies that

$$\lim_{n\to\infty}\mathbb{E}\left[|f_n(X) - f^*(X)|^2\right] = 0$$

with probability 1, that is, the consistency of the elastic-net regularization scheme with respect to the square loss.

Let us remark that we are also able to prove such consistency without assuming (7) in Assumption 2. To this end, we need the following lemma, which is of interest by itself.

Lemma 5. *Instead of Assumption 2, assume that the regression model is given by*

$$Y = f^*(X) + W,$$

where $f^* : \mathcal{X} \to \mathcal{Y}$ is a bounded function and W satisfies (8) and (9). For fixed λ and $\varepsilon > 0$, with probability greater than $1 - 2e^{-\delta}$ we have

$$\left\| \Phi_n^*(f^\lambda - f^*) - \Phi_P^*(f^\lambda - f^*) \right\|_2 \leq \left(\frac{\sqrt{\kappa} D_\lambda \delta}{n} + \frac{\sqrt{2\kappa\delta} \left\| f^\lambda - f^* \right\|_P}{\sqrt{n}} \right), \tag{53}$$

where $f^\lambda = f_{\beta^\lambda}$ and $D_\lambda = \sup_{x \in \mathcal{X}} |f^\lambda(x) - f^*(x)|$.

We notice that in (53), the function $f^\lambda - f^*$ is regarded both as an element of $L^2_{\mathcal{Y}}(\mathbb{P}_n)$ and as an element of $L^2_{\mathcal{Y}}(P)$.

Proof. Consider the ℓ_2-valued random variable

$$Z = \Phi_X^*(f^\lambda(X) - f^*(X)) \qquad Z_\gamma = \left\langle f^\lambda(X) - f^*(X), \varphi_\gamma(X) \right\rangle.$$

A simple computation shows that $\mathbb{E}[Z] = \Phi_P^*(f^\lambda - f^*)$ and

$$\|Z\|_2 \leq \sqrt{\kappa} |f^\lambda(X) - f^*(X)|.$$

Hence, for any $m \geq 2$,

$$\begin{aligned}
\mathbb{E}\left[\|Z - \mathbb{E}[Z]\|_2^m \right] &\leq \mathbb{E}\left[\|Z - \mathbb{E}[Z]\|_2^2 \right] \left(2\sqrt{\kappa} \sup_{x \in \mathcal{X}} |f^\lambda(x) - f^*(x)| \right)^{m-2} \\
&\leq \kappa \mathbb{E}\left[|f^\lambda(X) - f^*(X)|^2 \right] \left(2\sqrt{\kappa} \sup_{x \in \mathcal{X}} |f^\lambda(x) - f^*(x)| \right)^{m-2} \\
&\leq \tfrac{m!}{2} (\sqrt{\kappa} \left\| f^\lambda - f^* \right\|_P)^2 (\sqrt{\kappa} D_\lambda)^{m-2}.
\end{aligned}$$

Applying (47) with $H = \sqrt{\kappa} D_\lambda$ and $M = \sqrt{\kappa} \left\| f^\lambda - f^* \right\|_P$, we obtain bound (53). □

Observe that under Assumption (7) and by the definition of β^ε one has that $D_\lambda \leq \sqrt{\kappa} \left\| \beta^\lambda - \beta^\varepsilon \right\|_2$, so that (53) becomes

$$\left\| (\Phi_n^* \Phi_n - \Phi_P^* \Phi_P)(\beta^\lambda - \beta^\varepsilon) \right\|_2 \leq \left(\frac{\kappa\delta \left\| \beta^\lambda - \beta^\varepsilon \right\|_2}{n} + \frac{\sqrt{2\kappa\delta} \left\| \Phi_P(\beta^\lambda - \beta^\varepsilon) \right\|_P}{\sqrt{n}} \right).$$

Since Φ_P is a compact operator this bound is tighter than the one deduced from (49). However, the price we pay is that the bound does not hold uniformly in λ. We are now able to state the universal strong consistency of the elastic-net regularization scheme.

Theorem 3. *Assume that (X, Y) satisfy (8) and (9) and that the regression function f^* is bounded. If the linear span of features $(\varphi_\gamma)_{\gamma \in \Gamma}$ is dense in $L^2_{\mathcal{Y}}(P)$ and $\varepsilon > 0$, then*

$$\lim_{n \to \infty} \mathbb{E}\left[|f_n(X) - f^*(X)|^2 \right] = 0 \text{ with probability } 1,$$

provided that $\lim_{n \to \infty} \lambda_n = 0$ and $\lim_{n \to \infty} n\lambda_n^2 - 2 \log n = +\infty$.

Proof. As above we bound separately the approximation error and the sample error. As for the first term, let $f^\lambda = f_{\beta^\lambda}$. We claim that $\mathbb{E}\left[|f^\lambda(X) - f^*(X)|^2 \right]$ goes to zero when λ goes to zero. Given $\eta > 0$, the fact that the linear span of the features $(\varphi_\gamma)_{\gamma \in \Gamma}$ is dense in $L^2_{\mathcal{Y}}(P)$ implies that there is $\beta^\eta \in \ell_2$ such that $p_\varepsilon(\beta^\eta) < \infty$ and

$$\mathbb{E}\left[|f_{\beta^\eta}(X) - Y|^2 \right] \leq \mathbb{E}\left[|f^*(X) - Y|^2 \right] + \eta.$$

Let $\lambda_\eta = \frac{\eta}{1+p_\varepsilon(\beta^\eta)}$, then, for any $\lambda \le \lambda_\eta$,

$$\begin{aligned}\mathbb{E}\left[|f^\lambda(X) - f^*(X)|^2\right] &\le \left(\mathbb{E}\left[|f^\lambda(X) - Y|^2\right] - \mathbb{E}\left[|f^*(X) - Y|^2\right]\right) + \lambda p_\varepsilon(\beta^\lambda)\\ &\le \left(\mathbb{E}\left[|f_{\beta^\eta}(X) - Y|^2\right] - \mathbb{E}\left[|f^*(X) - Y|^2\right]\right) + \lambda p_\varepsilon(\beta^\eta)\\ &\le \eta + \eta.\end{aligned}$$

As for the sample error, we let $f_n^\lambda = f_{\beta_n^\lambda}$ (so that $f_n = f_n^{\lambda_n}$) and observe that

$$\mathbb{E}\left[|f^\lambda(X) - f_n^\lambda(X)|^2\right] = \left\|\Phi_P(\beta_n^\lambda - \beta^\lambda)\right\|_P^2 \le \kappa \left\|\beta_n^\lambda - \beta^\lambda\right\|_2^2.$$

We bound $\left\|\beta_n^\lambda - \beta^\lambda\right\|_2$ by (53) observing that

$$\begin{aligned}D_\lambda &= \sup_{x\in\mathcal{X}} |f^\lambda(x) - f^*(x)| \le \sup_{x\in\mathcal{X}} |f_{\beta^\lambda}(x)| + \sup_{x\in\mathcal{X}} |f^*(x)|\\ &\le \sqrt{\kappa}\left\|\beta^\lambda\right\|_2 + \sup_{x\in\mathcal{X}} |f^*(x)| \le D\frac{1}{\sqrt{\lambda}}\end{aligned}$$

where D is a suitable constant and where we used the crude estimate

$$\lambda\varepsilon\left\|\beta^\lambda\right\|_2^2 \le \mathcal{E}^\lambda(\beta^\lambda) \le \mathcal{E}^\lambda(0) = \mathbb{E}\left[|Y|^2\right].$$

Hence (53) yields

$$\left\|\Phi_n^*(f^\lambda - f^*) - \Phi_P^*(f^\lambda - f^*)\right\|_2 \le \left(\frac{\sqrt{\kappa}\delta D}{\sqrt{\lambda}n} + \frac{\sqrt{2\kappa\delta}\left\|f^\lambda(X) - f^*(X)\right\|_P}{\sqrt{n}}\right). \tag{54}$$

Observe that the proof of (43) does not depend on the existence of β^ε provided that we replace both $\Phi_P\beta_n^\lambda \in L^2_y(P)$ and $\Phi_n\beta_n^\lambda \in L^2_y(\mathbb{P}_n)$ with f^*, and we take into account that both $\Phi_P\beta^\lambda \in L^2_y(P)$ and $\Phi_n\beta^\lambda \in L^2_y(\mathbb{P}_n)$ are equal to f^λ. Hence, plugging (54) and (48) in (43) we have that with probability greater than $1 - 4e^{-\delta}$

$$\left\|\beta_n^\lambda - \beta^\lambda\right\|_2 \le \frac{D\sqrt{\delta}}{\kappa_0 + \varepsilon\lambda}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{\lambda}n} + \frac{\left\|f^\lambda(X) - f^*(X)\right\|_P}{\sqrt{n}}\right)$$

where D is a suitable constant and $\delta \le n$. The thesis now follows by combining the bounds on the sample and approximation errors and repeating the proof of Theorem 2. □

To have an explicit convergence rate, one needs an explicit bound on the approximation error $\left\|\beta^\lambda - \beta^\varepsilon\right\|_2$, for example of the form $\left\|\beta^\lambda - \beta^\varepsilon\right\|_2 = O(\lambda^r)$. This is out of the scope of the paper. We report only the following simple result.

Proposition 9. *Assume that the features φ_γ are in finite number and linearly independent. Let $N^* = |\mathrm{supp}(\beta^\varepsilon)|$ and $w^* = \sup_{\gamma\in\mathrm{supp}(\beta^\varepsilon)}\{w_\gamma\}$, then*

$$\left\|\beta^\lambda - \beta^\varepsilon\right\|_2 \le DN^*\lambda.$$

With the choice $\lambda_n = \frac{1}{\sqrt{n}}$, for any $\delta > 0$ and $n \in \mathbb{N}$ with $\delta \le n$

$$\left\|\beta_n^{\lambda_n} - \beta^\varepsilon\right\|_2 \le \frac{c\sqrt{\delta}}{\sqrt{n}\kappa_0}\left(1 + \frac{DN^*}{\sqrt{n}}\right) + \frac{DN^*}{\sqrt{n}}, \tag{55}$$

with probability greater than $1 - 4e^{-\delta}$, where $D = \frac{w^* + 2\varepsilon\|\beta^\varepsilon\|_\infty}{2\kappa_0}$ and $c = \max\left\{\sqrt{2\kappa}(\sigma + L), 3\kappa\right\}$.

Proof. Observe that the assumption on the set of features is equivalent to assuming that $\kappa_0 > 0$. First, we bound the approximation error $\left\|\beta^\lambda - \beta^\varepsilon\right\|_2$. As usual, with the choice $\tau = \frac{\kappa_0+\kappa}{2}$, (36) gives

$$\beta^\lambda - \beta^\varepsilon = \frac{1}{\tau+\varepsilon\lambda}\left[\mathbf{S}_\lambda\left((\tau I - \Phi_P^*\Phi_P)\beta^\lambda + \Phi_P^*\Phi_P\beta^\varepsilon\right) - \mathbf{S}_\lambda\left(\tau\beta^\varepsilon\right) + \mathbf{S}_\lambda\left(\tau\beta^\varepsilon\right) - \tau\beta^\varepsilon\right] - \frac{\varepsilon\lambda}{\tau+\varepsilon\lambda}\beta^\varepsilon.$$

Property (30) implies that

$$\left\|\beta^\lambda - \beta^\varepsilon\right\|_2 \le \frac{1}{\tau+\varepsilon\lambda}\left(\left\|(\tau I - \Phi_P^*\Phi_P)(\beta^\lambda - \beta^\varepsilon)\right\|_2 + \left\|\mathbf{S}_\lambda\left(\tau\beta^\varepsilon\right) - \tau\beta^\varepsilon\right\|_2\right) + \frac{\varepsilon\lambda}{\tau+\varepsilon\lambda}\left\|\beta^\varepsilon\right\|_2.$$

Since $\|\tau I - \Phi_P^*\Phi_P\| \le \frac{\kappa-\kappa_0}{2}$, $\left\|\beta^\varepsilon\right\|_2 \le N^*\left\|\beta^\varepsilon\right\|_\infty$ and

$$\left\|\mathbf{S}_\lambda\left(\tau\beta^\varepsilon\right) - \tau\beta^\varepsilon\right\|_2 \le w^* N^* \frac{\lambda}{2},$$

one has

$$\begin{aligned}\left\|\beta^\lambda - \beta^\varepsilon\right\|_2 &\le \frac{\kappa+\kappa_0+2\varepsilon\lambda}{2(\kappa_0+\varepsilon\lambda)}\left(\frac{2}{\kappa+\kappa_0+2\varepsilon\lambda}w^*N^*\frac{\lambda}{2} + \frac{2\varepsilon\lambda}{\kappa_0+\kappa+2\varepsilon\lambda}\left\|\beta^\varepsilon\right\|_2\right)\\ &\le \left(\frac{w^* + 2\varepsilon\left\|\beta^\varepsilon\right\|_\infty}{2\kappa_0}\right)N^*\lambda = DN^*\lambda.\end{aligned}$$

Bound (55) is then a straightforward consequence of (52). □

4.3. Adaptive choice

In this section, we suggest an adaptive choice of the regularization parameter λ. The main advantage of this selection rule is that it does not require any knowledge of the behavior of the approximation error. To this end, it is useful to replace the approximation error with the following upper bound

$$\mathcal{A}(\lambda) = \sup_{0<\lambda'\le\lambda}\left\|\beta^{\lambda'} - \beta^\varepsilon\right\|_2. \tag{56}$$

The following simple result holds.

Lemma 6. *Given $\varepsilon > 0$, $\mathcal{A}$ is an increasing continuous function and*

$$\left\|\beta^\lambda - \beta^\varepsilon\right\|_2 \le \mathcal{A}(\lambda) \le A < \infty$$
$$\lim_{\lambda\to 0+}\mathcal{A}(\lambda) = 0.$$

Proof. First of all, we show that $\lambda \mapsto \beta^\lambda$ is a continuous function. Fix $\lambda > 0$; for any h such that $\lambda + h > 0$, (36) with $\tau = \frac{\kappa_0+\kappa}{2}$ and Corollary 1 give

$$\begin{aligned}\left\|\beta^{\lambda+h} - \beta^\lambda\right\|_2 &\le \left\|\mathcal{T}_{\lambda+h}(\beta^{\lambda+h}) - \mathcal{T}_{\lambda+h}(\beta^\lambda)\right\|_2 + \left\|\mathcal{T}_{\lambda+h}(\beta^\lambda) - \mathcal{T}_\lambda(\beta^\lambda)\right\|_2\\ &\le \frac{\kappa-\kappa_0}{\kappa+\kappa_0+2\varepsilon(\lambda+h)}\left\|\beta^{\lambda+h} - \beta^\lambda\right\|_2\\ &\quad + \left\|\frac{1}{\tau+\varepsilon(\lambda+h)}\mathbf{S}_{\lambda+h}\left(\beta'\right) - \frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left(\beta'\right)\right\|_2\end{aligned}$$

where $\beta' = (\tau I - \Phi_P^*\Phi_P)\beta^\lambda + \Phi_P^*Y$ does not depend on h and we wrote $\mathcal{T}_\lambda$ to make explicit the dependence of the map $\mathcal{T}$ on the regularization parameter. Hence

$$\begin{aligned}&\left\|\beta^{\lambda+h} - \beta^\lambda\right\|_2\\ &\le \frac{\tau+\varepsilon(\lambda+h)}{\kappa_0+\varepsilon(\lambda+h)}\left(\left|\frac{1}{\tau+\varepsilon(\lambda+h)} - \frac{1}{\tau+\varepsilon\lambda}\right|\left\|\beta'\right\|_2 + \frac{1}{\tau+\varepsilon\lambda}\left\|\mathbf{S}_{\lambda+h}\left(\beta'\right) - \mathbf{S}_\lambda\left(\beta'\right)\right\|_2\right).\end{aligned}$$

The claim follows by observing that (assuming for simplicity that $h > 0$)

$$\left\|\mathbf{S}_{\lambda+h}\left(\beta'\right)-\mathbf{S}_{\lambda}\left(\beta'\right)\right\|_2^2 = \sum_{w_\gamma\lambda\le|\beta'_\gamma|<w_\gamma(\lambda+h)}|\beta'_\gamma-\operatorname{sgn}(\beta'_\gamma)w_\gamma\lambda|^2+\sum_{|\beta'_\gamma|\ge w_\gamma(\lambda+h)}w_\gamma^2h^2$$
$$\le h^2\sum_{|\beta'_\gamma|\ge w_\gamma\lambda}w_\gamma^2\le h^2\sum_{|\beta'_\gamma|\ge w_\gamma\lambda}(\beta'_\gamma/\lambda)^2\le h^2\left\|\beta'\right\|_2^2/\lambda^2,$$

which goes to zero if h tends to zero.

Now, by the definition of β^λ and β^ε

$$\varepsilon\lambda\left\|\beta^\lambda\right\|_2^2\le\mathbb{E}\left[|\Phi_P\beta^\lambda-f^*(X)|^2\right]+\lambda p_\varepsilon(\beta^\lambda)\le\mathbb{E}\left[|\Phi_P\beta^\varepsilon-f^*(X)|^2\right]+\lambda p_\varepsilon(\beta^\varepsilon)=\lambda p_\varepsilon(\beta^\varepsilon),$$

so that

$$\left\|\beta^\lambda-\beta^\varepsilon\right\|_2\le\|\beta^\varepsilon\|_2+\frac{1}{\sqrt{\varepsilon}}p_\varepsilon(\beta^\varepsilon)=:A.$$

Hence $\mathcal{A}(\lambda)\le A$ for all λ. Clearly $\mathcal{A}(\lambda)$ is an increasing function of λ; the fact that $\left\|\beta^\lambda-\beta^\varepsilon\right\|_2$ is continuous and goes to zero with λ ensures that the same holds true for $\mathcal{A}(\lambda)$. □

Notice that we replaced the approximation error with $\mathcal{A}(\lambda)$ just for a technical reason, namely to deal with an increasing function of λ. If we have a monotonic decay rate at our disposal, such as $\left\|\beta^\lambda-\beta^\varepsilon\right\|_2\asymp\lambda^a$ for some $a>0$ and for $\lambda\to0$, then clearly $\mathcal{A}(\lambda)\asymp\lambda^a$.

Now, we fix $\varepsilon>0$ and $\delta\ge2$ and we assume that $\kappa_0=0$. Then we simplify bound (52) observing that

$$\left\|\beta_n^\lambda-\beta^\varepsilon\right\|_2\le C\left(\frac{1}{\sqrt{n}\varepsilon\lambda}+\mathcal{A}(\lambda)\right)\tag{57}$$

where $C=c\sqrt{\delta}(1+A)$; the bound holds with probability greater than $1-4e^{-\delta}$ uniformly for all $\lambda>0$.

When λ increases, the first term in (57) decreases whereas the second increases; hence to have a tight bound a *natural* choice of the parameter consists of balancing the two terms in the above bound, taking

$$\lambda_n^{\text{opt}}=\sup\left\{\lambda\in]0,\infty[\,|\,\mathcal{A}(\lambda)=\frac{1}{\sqrt{n}\varepsilon\lambda}\right\}.$$

Since $\mathcal{A}(\lambda)$ is continuous, $\frac{1}{\sqrt{n}\varepsilon\lambda_n^{\text{opt}}}=\mathcal{A}(\lambda_n^{\text{opt}})$ and the resulting bound is

$$\left\|\beta_n^\lambda-\beta^\varepsilon\right\|_2\le\frac{2C}{\sqrt{n}\varepsilon\lambda_n^{\text{opt}}}.\tag{58}$$

This method for choosing the regularization parameter clearly requires the knowledge of the approximation error. To overcome this drawback, we discuss a data-driven choice for λ that allows achieving the rate (58) *without* requiring any prior information on $\mathcal{A}(\lambda)$. For this reason, such a choice is said to be *adaptive*. The procedure we present is also referred to as an *a posteriori* choice since it depends on the given sample and not only on its cardinality n. In other words, the method is purely data-driven.

Let us consider a discrete set of values for λ defined by the geometric sequence

$$\lambda_i=\lambda_0 2^i\quad i\in\mathbb{N}\ \lambda_0>0.$$

Notice that we may replace the sequence $\lambda_0 2^i$ by any other geometric sequence $\lambda_i=\lambda_0 q^i$ with $q>1$; this would only lead to a more complicated constant in (60). Define the parameter λ_n^+ as follows

$$\lambda_n^+=\max\left\{\lambda_i\left|\left\|\beta_n^{\lambda_j}-\beta_n^{\lambda_{j-1}}\right\|_2\le\frac{4C}{\sqrt{n}\varepsilon\lambda_{j-1}}\quad\text{for all } j=0,\dots,i\right.\right\}\tag{59}$$

(with the convention that $\lambda_{-1} = \lambda_0$). This strategy for choosing λ is inspired by a procedure originally proposed in [40] for Gaussian white noise regression and which has been widely discussed in the context of deterministic as well as stochastic inverse problems (see [26,41]). In the context of non-parametric regression from random design, this strategy has been considered in [42] and the following proposition is a simple corollary of a result contained in [42].

Proposition 10. *Provided that* $\lambda_0 < \lambda_n^{\text{opt}}$, *the following bound holds with probability greater than* $1 - 4e^{-\delta}$

$$\left\| \beta_n^{\lambda_n^+} - \beta^\varepsilon \right\|_2 \leq \frac{20C}{\sqrt{n}\varepsilon\lambda_n^{\text{opt}}}. \tag{60}$$

Proof. The proposition results from Theorem 2 in [42]. For completeness, we report here a proof adapted to our setting. Let Ω be the event such that (57) holds for any $\lambda > 0$; we have that $\mathbb{P}[\Omega] \geq 1 - 4e^{-\delta}$ and we fix a sample point in Ω.

The definition of λ_n^{opt} and the assumption $\lambda_0 < \lambda_n^{\text{opt}}$ ensure that $\mathcal{A}(\lambda_0) \leq \frac{1}{\sqrt{n}\varepsilon\lambda_0}$. Hence the set $\left\{\lambda_i \mid \mathcal{A}(\lambda_i) \leq \frac{1}{\sqrt{n}\varepsilon\lambda_i}\right\}$ is not empty and we can define

$$\lambda_n^* = \max\left\{\lambda_i \mid \mathcal{A}(\lambda_i) \leq \frac{1}{\sqrt{n}\varepsilon\lambda_i}\right\}.$$

The fact that $(\lambda_i)_{i\in\mathbb{N}}$ is a geometric sequence implies that

$$\lambda_n^* \leq \lambda_n^{\text{opt}} < 2\lambda_n^*, \tag{61}$$

while (57) with the definition of λ_n^* ensures that

$$\left\| \beta_n^{\lambda_n^*} - \beta^\varepsilon \right\|_2 \leq C\left(\frac{1}{\sqrt{n}\varepsilon\lambda_n^*} + \mathcal{A}(\lambda_n^*)\right) \leq \frac{2C}{\sqrt{n}\varepsilon\lambda_n^*}. \tag{62}$$

We show that $\lambda_n^* \leq \lambda_n^+$. Indeed, for any $\lambda_j < \lambda_n^*$, using (57) twice, we get

$$\begin{aligned}\left\| \beta_n^{\lambda_n^*} - \beta_n^{\lambda_j} \right\|_2 &\leq \left\| \beta_n^{\lambda_j} - \beta^\varepsilon \right\|_2 + \left\| \beta_n^{\lambda_n^*} - \beta^\varepsilon \right\|_2 \\ &\leq C\left(\frac{1}{\sqrt{n}\varepsilon\lambda_j} + \mathcal{A}(\lambda_j) + \frac{1}{\sqrt{n}\varepsilon\lambda_n^*} + \mathcal{A}(\lambda_n^*)\right) \leq \frac{4C}{\sqrt{n}\varepsilon\lambda_j},\end{aligned}$$

where the last inequality holds since $\lambda_j < \lambda_n^* \leq \lambda_n^{\text{opt}}$ and $\mathcal{A}(\lambda) \leq \frac{1}{\sqrt{n}\varepsilon\lambda}$ for all $\lambda < \lambda_n^{\text{opt}}$. Now $2^m\lambda_0 \leq \lambda_n^* \leq \lambda_n^+ = 2^{m+k}$ for some $m, k \in \mathbb{N}$, so that

$$\begin{aligned}\left\| \beta_n^{\lambda_n^+} - \beta_n^{\lambda_n^*} \right\|_2 &\leq \sum_{\ell=0}^{k-1} \left\| \beta_n^{m+1+\ell} - \beta_n^{m+\ell} \right\|_2 \leq \sum_{\ell=0}^{k-1} \frac{4C}{\sqrt{n}\varepsilon\lambda_{m+\ell}} \\ &\leq \frac{4C}{\sqrt{n}\varepsilon\lambda_n^*} \sum_{\ell=0}^{\infty} \frac{1}{2^\ell} = \frac{4C}{\sqrt{n}\varepsilon\lambda_n^*}\, 2.\end{aligned}$$

Finally, recalling (61) and (62), we get bound (60):

$$\left\| \beta_n^{\lambda_n^+} - \beta^\varepsilon \right\|_2 \leq \left\| \beta_n^{\lambda_n^+} - \beta_n^{\lambda_n^*} \right\|_2 + \left\| \beta_n^{\lambda_n^*} - \beta^\varepsilon \right\|_2 \leq \frac{8C}{\sqrt{n}\varepsilon\lambda_n^*} + \frac{2C}{\sqrt{n}\varepsilon\lambda_n^*} \leq 20C\frac{1}{\sqrt{n}\varepsilon\lambda_n^{\text{opt}}}. \quad \square$$

Notice that the a priori condition $\lambda_0 < \lambda_n^{\text{opt}}$ is satisfied, for example, if $\lambda_0 < \frac{1}{A\varepsilon\sqrt{n}}$.

To illustrate the implications of the last proposition, let us suppose that

$$\left\| \beta^\lambda - \beta^\varepsilon \right\|_2 \asymp \lambda^a \tag{63}$$

for some unknown $a \in]0, 1]$. One has then that $\lambda_n^{\text{opt}} \asymp n^{-\frac{1}{2(a+1)}}$ and $\left\| \beta_n^{\lambda_n^+} - \beta^\varepsilon \right\|_2 \asymp n^{-\frac{a}{2(a+1)}}$.

We end noting that, if we specialize our analysis to least squares regularized with a pure ℓ_2-penalty (i.e. setting $w_\gamma = 0, \forall \gamma \in \Gamma$), then our results lead to the error estimate in the norm of the reproducing kernel space $\mathcal{H}$ obtained in [43,44]. Indeed, in such a case, β^ε is the generalized solution $\beta^\dagger$ of the equation $\Phi_P \beta = f^*$ and the approximation error satisfies (63) under the a priori assumption that the regression vector $\beta^\dagger$ is in the range of $(\Phi_P^* \Phi_P)^a$ for some $0 < a \leq 1$ (the fractional power makes sense since $\Phi_P^* \Phi_P$ is a positive operator). Under this assumption, it follows that $\left\| \beta_n^{\lambda_n^+} - \beta^\varepsilon \right\|_2 \asymp n^{-\frac{a}{2(a+1)}}$. To compare this bound with the results in the literature, recall that both $f_n = f_{\beta_n^{\lambda_n^+}}$ and $f^* = f_{\beta^\dagger}$ belong to the reproducing kernel Hilbert space $\mathcal{H}$ defined in Proposition 3. In particular, one can check that $\beta^\dagger \in \text{ran}(\Phi_P^* \Phi_P)^a$ if and only if $f^* \in \text{ran}\, L_K^{\frac{2a+1}{2}}$, where $L_K : L^2_{\mathcal{Y}}(P) \to L^2_{\mathcal{Y}}(P)$ is the integral operator whose kernel is the reproducing kernel K [45]. Under this condition, the following bound holds

$$\left\| f_n - f^* \right\|_{\mathcal{H}} \leq \left\| \beta_n^{\lambda_n^+} - \beta^\varepsilon \right\|_2 \asymp n^{-\frac{a}{2(a+1)}},$$

which gives the same rate as in Theorem 2 of [43] and Corollary 17 of [44].

Acknowledgments

We thank Alessandro Verri for helpful suggestions and discussions. Christine De Mol acknowledges support by the "Action de Recherche Concertée" Nb 02/07-281, the VUB-GOA 62 grant and the National Bank of Belgium BNB; she is also grateful to the DISI, Università di Genova for hospitality during a semester in which the present work was initiated. Ernesto De Vito and Lorenzo Rosasco have been partially supported by the FIRB project RBIN04PARL and by the EU Integrated Project Health-e-Child IST-2004-027749.

References

[1] E. Candès, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Ann. Statist. 35 (6) (2007) 2313–2351.
[2] A. Destrero, C. De Mol, F. Odone, A. Verri, A regularized approach to feature selection for face detection, in: Proceedings ACCV07, 2007, pages II: 881–890.
[3] A. Destrero, S. Mosci, C. De Mol, A. Verri, F. Odone, Feature selection for high-dimensional data, Comput. Manag. Sci. 6 (1) (2009) 25–40.
[4] A. Destrero, C. De Mol, F. Odone, A. Verri, A sparsity-enforcing method for learning face features, IEEE Trans. Image Process. 18 (1) (2009) 188–201.
[5] C. De Mol, S. Mosci, M. Traskine, A. Verri, A regularized method for selecting nested groups of relevant genes from microarray data, J. Comp. Biol. (in press). Preprint available at: http://arxiv.org/abs/0809.1777.
[6] A. Barla, S. Mosci, L. Rosasco, A. Verri, A method for robust variable selection with significance assessment, in: ESANN 2008, 2008. Preprint available at: http://www.disi.unige.it/person/MosciS/PAPERS/esann.pdf.
[7] R. Tibshirani, Regression selection and shrinkage via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[8] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1998) 33–61.
[9] W. Fu, Penalized regressions: The bridge versus the lasso, J. Comput. Graph. Statist. 7 (3) (1998) 397–416.
[10] V. Koltchinskii, Sparsity in penalized empirical risk minimization, Annales de l'Institut Henri Poincaré B, Probab. Statist. 45 (1) (2009) 7–57.
[11] K. Knight, W. Fu, Asymptotics for lasso-type estimators, Ann. Statist. 28 (5) (2000) 1356–1378.
[12] F. Bunea, A. Tsybakov, M. Wegkamp, Aggregation and sparsity via l1 penalized least squares, in: Proc. 19th Annu. Conference on Comput. Learning Theory, Springer, 2006, pp. 379–391.
[13] J.-M. Loubes, S. van de Geer, Adaptive estimation with soft thresholding penalties, Statist. Neerlandica 56 (4) (2002) 454–479.
[14] B. Tarigan, S.A. van de Geer, Classifiers of support vector machine type with l_1 complexity regularization, Bernoulli 12 (6) (2006) 1045–1076.
[15] E. Greenshtein, Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint, Ann. Statist. 34 (5) (2006) 2367–2386.
[16] S.A. van de Geer, High-dimensional generalized linear models and the lasso, Ann. Statist. 36 (2) (2008) 614–645.
[17] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.
[18] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2004) 407–499.
[19] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math. 57 (11) (2004) 1413–1457.

[20] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2) (2005) 301–320.
[21] A. Owen, A robust hybrid of lasso and ridge regression, in: Contemporary Mathematics, vol. 443, American Mathematical Society, Providence, Rhode Island, 2007, pp. 59–72.
[22] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B Series B 68 (2006) 49–67.
[23] M. Fornasier, H. Rauhut, Recovery algorithms for vector-valued data with joint sparsity constraints, SIAM J. Numer. Anal. 46 (2) (2008) 577–613.
[24] G. Wahba, Spline models for observational data, in: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, PA, 1990.
[25] A. Barron, A. Cohen, W. Dahmen, R. DeVore, Adaptive approximation and learning by greedy algorithms, Ann. Statist. 36 (1) (2008) 64–94.
[26] F. Bauer, S. Pereverzev, Regularization without preliminary knowledge of smoothness and error behavior, European J. Appl. Math. 16 (2005) 303–317.
[27] C.A. Micchelli, M. Pontil, On learning vector-valued functions, Neural Comput. 17 (1) (2005) 177–204.
[28] L. Baldassarre, A. Barla, B. Gianesin, M. Marinelli, Vector valued regression for iron overload estimation, in: Proceedings of ICPR 2008, Tampa, FL, USA.
[29] C.A. Micchelli, M. Pontil, T. Evgeniou, Learning multiple tasks with kernel methods, J. Mach. Learn. Res. 6 (2005) 615–637.
[30] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, vol. 19, MIT Press, Cambridge, MA, 2007, pp. 41–48.
[31] A. Caponnetto, C.A. Micchelli, M. Pontil, Y. Ying, Universal multi-task kernels, J. Mach. Learn. Res. 9 (2008) 1615–1646.
[32] U. Amato, A. Antoniadis, M. Pensky, Wavelet kernel penalized estimation for non-equispaced design regression, Statist. Comput. 16 (1) (2006) 37–55.
[33] C. Carmeli, E. De Vito, A. Toigo, Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem, Anal. Appl. (Singap.) 4 (4) (2006) 377–408.
[34] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, in: Springer Series in Statistics, Springer-Verlag, New York, 1996.
[35] I. Ekeland, T. Turnbull, Infinite-dimensional Optimization and Convexity, in: Chicago Lectures in Mathematics, The University of Chicago Press, Chicago, 1983.
[36] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. 22 (4) (1994) 1679–1706.
[37] I. Pinelis, Correction: ''Optimum bounds for the distributions of martingales in Banach spaces'', Ann. Probab. 22 (4) (1994) 1679–1706. MR1331198 (96b:60010); Ann. Probab. 27 (4) (1999) 2119.
[38] V. Yurinsky, Sums and Gaussian Vectors, in: Lecture Notes in Mathematics, vol. 1617, Springer-Verlag, Berlin, 1995.
[39] I.F. Pinelis, A.I. Sakhanenko, Remarks on inequalities for probabilities of large deviations, Theory Probab. Appl. 30 (1) (1985) 143–148.
[40] O. Lepskii, On a problem of adaptive estimation in Gaussian white noise, Theory Probab. Appl. 35 (1990) 454–466.
[41] E. Schock, S.V. Pereverzev, On the adaptive selection of the parameter in regularization of ill-posed problems, SIAM J. Numer. Anal. 43 (2005) 2060–2076.
[42] E. De Vito, S. Pereverzev, L. Rosasco, Adaptive learning via the balancing principle, BCL paper-275/CSAIL and Technical Report-TR-2008-062, Massachusetts Institute of Technology, 2008. Preprint available at: http://dspace.mit.edu/bitstream/handle/1721.1/42896/MIT-CSAIL-TR-2008-062.pdf.
[43] S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2) (2007) 153–172.
[44] F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, J. Complexity 23 (1) (2007) 52–72.
[45] A. Caponnetto, E. De Vito, Optimal rates for regularized least-squares algorithm, Found. Comput. Math. 7 (3) (2007) 331–368.