A consistent algorithm to solve Lasso, elastic-net and Tikhonov regularization

Ernesto De Vito[a], Veronica Umanità[a], Silvia Villa[a]

[a]*Dipartimento di Matematica, Università di Genova, Via Dodecaneso 35, 16146, Genova, Italy*

Abstract

In the framework of supervised learning we prove that the iterative algorithm introduced in Umanità and Villa (2010) allows to estimate in a consistent way the relevant features of the regression function under the a priori assumption that it admits a sparse representation on a fixed dictionary.

Keywords: Learning theory, Regularization, Sparsity, Consistent estimator
2000 MSC: 65J20, 62G08, 68T05

1. Introduction

In the context of supervised learning theory, this paper studies the consistency of the algorithm proposed in Umanità and Villa (2010) in a deterministic framework. The algorithm is an iterative procedure for the minimization of the ℓ_2-regularized empirical error on the ℓ_1-ball with an early stopping rule where both the ℓ_1 and ℓ_2 norms are computed with respect to a (possibly infinite) dictionary of functions.

Supervised learning refers to a process that builds a function that best represents the relation between an input-output random pair (X, Y), with values in $\mathcal{X} \times \mathcal{Y}$, on the base of a sample of n i.i.d. copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y) (Vapnik (1998); Cucker and Smale (2002b); Poggio and Smale (2003)). In this paper we assume that $\mathcal{X}$ is a separable complete metric space and $\mathcal{Y}$ is a separable Hilbert space. The joint probability distribution is unknown, but we know that the regression function $f^*(x) = \mathbb{E}[Y_i \,|\, X_i = x]$ is of the form $f^* = \sum \beta_s^* \varphi_s$ with $\sum |\beta_s^*| < +\infty$ and $\{\varphi_s\}_{s\in\Gamma}$ a family of (bounded) functions from $\mathcal{X}$ to $\mathcal{Y}$ called *dictionary*.

Functions whose coefficient vector is in ℓ_1 are usually called sparse. The sparsity of the regression function is an appropriate assumption in several relevant applications (genomic data for example) and the problem of selecting a consistent estimator not only for prediction, but also for variable selection, is a relevant topic. This roughly amounts to ask the estimator being able to identify the features on which the regression function depends, and this topic has been studied by many points of view. Recently much effort has been devoted to the analysis of the case where the cardinality of the dictionary is significantly bigger than the number of examples, or even infinite, as in our model. In these situations, the classical Tikhonov regularization (Engl et al. (1996)) does not perform well and the most popular approach is based on the Lasso technique (Tibshirani (1996); Efron et al. (2004)): several related but different consistency properties of this regularization procedure have been proved under various hypotheses (Chen et al. (1998);

Candes and Tao (2007); Daubechies et al. (2004); Bunea (2008); Zhao and Yu (2006); Van de Geer (2008)). Anyway it is known that Lasso has some drawbacks, especially when there are correlated features. In fact, in this case different coefficients can give the same representation of f^*, and the Lasso tends to select only arbitrarily one non-zero coefficient from each group of correlated features instead of all the relevant ones. For this reason, other regularizing penalties have been proposed to select a particular representation of f^* depending on the required properties on the solution. For example the elastic-net penalty provides consistent and sparse estimators (De Mol et al. (2009)) and is thus preferable to the ℓ^1 - norm (Zou and Hastie (2005); Jin et al. (2009)). In fact, such a penalty is a weighted sum of the ℓ^1-norm, enforcing sparsity, and the square of the ℓ^2-norm of the vector coefficient, which promotes a grouping effect.

To estimate $\beta^* = (\beta^*_s)_s$ we use an iterative algorithm $(\beta^m_{\lambda,R,n})_n$ based on the one proposed in Umanità and Villa (2010) and consisting in the (approximate) minimization

$$\min_{\beta\in B_R} \sum_{i=1}^{n} \frac{1}{n} \left\| \left(\sum_s \beta_s \varphi_s(X_i) \right) - Y_i \right\|^2 + \lambda \|\beta\|_2^2 , \tag{1}$$

on the ℓ^1-ball B_R of radius R. Here we prove that, when the positive constant R is suitably large, there exists a choice of $m = m_n$ and of the regularization parameter $\lambda = \lambda_n$ such that $\beta^{m_n}_{\lambda_n,R,n}$ converges with probability one to $\beta^\dagger_R$ as the number of observations goes to ∞, where $\beta^\dagger_R$ is a regression vector of f^*, i.e. $f^* = \sum_s (\beta^\dagger_R)_s \varphi_s$. Note that the choice of m_n defines a stopping rule in the computation of the minimizer of (1), therefore the proposed algorithm belongs to the class of early stopping methods.

Since in general the dictionary is not assumed to be linearly independent, there are many different regression vectors, and the parameter R allows to move from the Lasso estimator, to the elastic-net and the Tikhonov estimator. Moreover the convergence of the algorithm on the coefficients ensures the consistency of the corresponding estimator of f^*.

Besides casting in a unified framework three different regularizing methods, this approach has the advantage of bypassing the problem of exactly computing the minimizer of the regularized empirical risk, as it is usually needed. In fact, we directly show the consistency of an approximation of this minimizer obtained through the application of a suitably early stopping rule, and this is particularly relevant from the applications point of view.

The paper is organized in the following way. In Sections 2 and Appendix Appendix A, referring to De Mol et al. (2009), we introduce the mathematical setting of the problem and the main tools we will use to solve it. In section 3 we propose the iterative projected algorithm analyzed in Umanità and Villa (2010); Combettes and Wajs (2005); Fornasier et al. (2008) to compute the solution of the constrained problem (1). Finally, through this algorithm, we produce an estimator of the regression function and show its consistency (Sections 3 and 4).

2. The model

In this section, following De Mol et al. (2009), we describe the general mathematical framework to deal with the problem of estimating the regression function in the context of supervised statistical learning.

Let $\mathcal{X}$ be a separable complete metric space and $\mathcal{Y}$ a real separable Hilbert space with norm and scalar product denoted by $\|\ \|$ and $\langle\cdot,\cdot\rangle$ respectively. Given a random input-output pair (X,Y) with probability distribution ρ, defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$ and taking values in the product $\mathcal{X}\times\mathcal{Y}$ we assume that (X,Y) fits the regression model

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

where $f^*:\mathcal{X}\to\mathcal{Y}$ is a measurable function and W is a random noise in $\mathcal{Y}$ satisfying

$$\mathbf{E}[W\,|\,X] = 0 \tag{2}$$

$$\mathbf{E}\left[\exp\left(\frac{\|W\|}{L}\right) - \frac{\|W\|}{L} - 1\,|\,X\right] \le \frac{\sigma^2}{2L^2} \tag{3}$$

for some positive constants σ and L.
It follows from (2) that $f^*:\mathcal{X}\to\mathcal{Y}$ is the regression function, i.e. $f^*(x)=\mathbf{E}[Y\,|\,X=x]$ for almost all $x\in\mathcal{X}$, while equation (3) implies (see Van der Vaart and Wellner (1996))

$$\mathbf{E}[\|W\|^m\,|X] \le \frac{1}{2}m!\sigma^2L^{m-2} \qquad \forall\, m\ge 2. \tag{4}$$

As mentioned in the introduction f^* is assumed to belong to a specific hypothesis space $\mathcal{H}$ that we now describe.

Let $(\varphi_s)_{s\in\Gamma}$ be a countable *dictionary* of measurable features $\varphi_s:\mathcal{X}\to\mathcal{Y}$ such that

$$\forall\ x\in\mathcal{X} \quad \kappa(x) := \sum_{s\in\Gamma}|\varphi_s(x)|^2 \le \kappa \tag{5}$$

for some positive constant κ. Assumption (5) ensures that for any $\beta\in\ell_2(\Gamma)$ the series $\sum_{s\in\Gamma}\beta_s\varphi_s$ defines a bounded function $f_\beta:\mathcal{X}\to\mathcal{Y}$ (the series is summable in $\mathcal{Y}$ uniformly on $\mathcal{X}$). Our main assumption is that the regression function f^* admits a sparse representation with respect to the dictionary $\{\varphi_s\}$, namely

$$f^* = \sum_{s\in\Gamma}\beta^*_s\varphi_s \quad \text{for at least one } \beta^*\in\ell^1(\Gamma). \tag{6}$$

This implies that f^* belongs to the Hilbert space (De Mol et al. (2009))

$$\mathcal{H} := \{f_\beta := \sum_{s\in\Gamma}\beta_s\varphi_s\ :\ \beta\in\ell^2(\Gamma)\},$$

whose elements are bounded functions on $\mathcal{X}$ thanks to the inequality $\sup_{x\in\mathcal{X}}|f_\beta(x)|\le\kappa^{1/2}\|\beta\|_2$. Since we consider $\mathcal{H}$ as *hypothesis space* in which we search for an estimator of f^*, the map $\beta\mapsto f_\beta$ allows us to cast the problem in $\ell^2(\Gamma)$. Since the features $\{\varphi_s\}_{s\in\Gamma}$ can be linearly dependent, the set of the *regression vectors* of f^*

$$\mathcal{C} = \{\beta\in\ell^2(\Gamma)\ :\ f^*(X) = \sum_{s\in\Gamma}\beta_s\varphi_s(X)\} \tag{7}$$

(which β^* belongs to) in general is not a singleton. As a consequence, different algorithms can select different elements in $\mathcal{C}$. The scheme we propose is the following.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. copies of (X, Y). For a fixed a positive parameter R, for $\lambda > 0$ and positive constants $(\gamma_n^{(m)})_m$, we introduce a family of estimators of f^* by setting

$$\begin{cases} \beta^0_{\lambda,R,n} = 0 \\ \beta^{m+1}_{\lambda,R,n} = P_R\left[(1 - \lambda\gamma_n^{(m)})\beta^m_{\lambda,R,n} + \gamma_n^{(m)}\hat{\beta}^m_{\lambda,R,n}\right], \end{cases} \tag{8}$$

where

$$(\hat{\beta}^m_{\lambda,R,n})_s := \frac{1}{n}\sum_{i=1}^n \langle Y_i - \sum_{g\in\Gamma}(\beta^m_{\lambda,R,n})_g \varphi_g(X_i), \varphi_s(X_i)\rangle$$

and P_R is the projection onto $B_R := \{\beta \in \ell^2(\Gamma) : \|\beta\|_1 \leq R\}$, the ℓ^1-ball of radius R in $\ell^2(\Gamma)$.

Defining

$$R_\infty := \min_{\beta\in\mathcal{C}} \|\beta\|_1, \tag{9}$$

we show in the next section that, for $R \geq R_\infty$ and for a suitable choice of $m = m_n$ and $\lambda = \lambda_n$, the sequence $(\beta^m_{\lambda,R,n})_n$ converges to the regression vector

$$\beta^\dagger_R := \operatorname*{argmin}_{\mathcal{C}\cap B_R} \|\beta\|_2^2. \tag{10}$$

Note that the condition $R \geq R_\infty$ is necessary and sufficient to guarantee $\mathcal{C} \cap B_R \neq \emptyset$. Moreover, $\beta^\dagger_R$ is well defined since $\mathcal{C}$ and B_R are closed and convex subspaces and $\|\cdot\|_2^2$ is coercive and strictly convex, so that the set of minimizers is nonempty and is reduced to a singleton.

As R grows, $\beta^\dagger_R$ ranges from the Lasso solution to the Tikhonov one passing through the elastic-net. More precisely, $\beta^\dagger_R$ is respectively (see Theorem 8 of Umanità and Villa (2010)):

1. the element of minimal ℓ^2-norm among the solutions of the ℓ^1-regularization, that is
$$\beta^\dagger_R = \operatorname*{argmin}_{\beta\in\mathcal{M}_1} \|\beta\|_2^2, \qquad \text{if } R = R_\infty,$$
where $\mathcal{M}_1 := \operatorname{argmin}_{\beta\in\mathcal{C}} \|\beta\|_1$;
2. the Tikhonov representation of f^*, namely
$$\beta^\dagger_R = \beta^\dagger := \operatorname*{argmin}_{\mathcal{C}} \|\beta\|_2^2, \qquad \text{if } R \geq R_0 := \left\|\beta^\dagger\right\|_1;$$
3. the elastic-net representation of f^*, i.e.
$$\beta^\dagger_R = \operatorname*{argmin}_{\beta\in\mathcal{C}} p_\tau(\beta), \qquad \text{if } R_\infty < R < R_0,$$
where $p_\tau(\beta) := 2\tau\|\beta\|_1 + \|\beta\|_2^2$ (see De Mol et al. (2009); Zou and Hastie (2005)). Of course there is a relationship between R and τ, but such relationship is not explicit. For a discussion of this fact see Fornasier et al. (2008); Umanità and Villa (2010).

Remark 1. *Hypothesis* (6) *guarantees* $\mathcal{C} \neq \emptyset$. *On the other hand every* $\beta^* \in \mathcal{C}$ *satisfies* (6), *and* (10) *selects a unique element in* $\mathcal{C}$ *to which the algorithm converges. The choice of* R *allows for choosing an appropriate regression vector of* f^*. *In fact, varying the parameter* R, *we can identify different features of* f^* *according to some available a priori information on the solution. However* R_0 *and* R_∞ *are a priori information, and to develop adaptive methods for the a posteriori estimate of these quantities would be an interesting topic (see De Vito et al. (2010)).*

3. Consistency for selection and prediction

This section is devoted to the rigorous statement of the main convergence results. We start introducing some notations.

We denote by $L^2(\Omega, \mathbb{P}; \mathcal{Y})$ the Hilbert space of square-integrable random variables taking values in $\mathcal{Y}$ with the usual L^2-norm. Given $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ a measurable function, then

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

We denote by $HS(\ell^2)$ the Hilbert space of Hilbert-Schmidt operators on $\ell^2(\Gamma)$ endowed with the norm $\|\cdot\|_{HS}$.

For all $\omega \in \Omega$ we define the following Hilbert-Schmidt operators

$$\Phi_{\mathbb{P}} : \ell^2(\Gamma) \to L^2(\Omega, \mathbb{P}; \mathcal{Y}), \qquad \Phi_{\mathbb{P}}\beta = f_\beta(X)$$

and

$$\Phi_n(\omega) : \ell^2(\Gamma) \to \mathcal{Y}^n, \qquad (\Phi_n(\omega))\,\beta = (f_\beta(X_1(\omega)), \dots, f_\beta(X_n(\omega)))\,. \tag{11}$$

The operators $\Phi_n(\omega)$ and $\Phi_{\mathbb{P}}$ are well-defined thanks to the results in Subsection Appendix A of the Appendix. Moreover $\Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}} \in HS(\ell^2)$.

We can rewrite the family $\beta_{\lambda,R,n}^m$ of estimators as

$$\begin{cases} \beta_{\lambda,R,n}^0 = 0 \\ \beta_{\lambda,R,n}^{m+1}(\omega) = P_R\left[(1 - \lambda\gamma_n^{(m)})\beta_{\lambda,R,n}^m(\omega) + \gamma_n^{(m)}\Phi_n^*(\omega)(\mathbf{Y}(\omega) - \Phi_n(\omega)\beta_{\lambda,R,n}^m(\omega))\right], \end{cases} \tag{12}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)$. In this way we obtain a family of random variables on $\ell^2(\Gamma)$. This can be proved by induction using that the projection is continuous and the map $\omega \mapsto \Phi_n^*(\omega)(\mathbf{Y}(\omega) - \Phi_n(\omega)\beta_{\lambda,R,n}^m(\omega))$ is measurable (see Lemma 1 in the Appendix).

In the following we will consider $R \geq R_\infty$.

Theorem 1. *Fix* $R \geq R_\infty$ *and consider* $\lambda = \lambda_n \in (0,1]$ *satisfying* $\lim_n \lambda_n = 0$ *and* $\lim_n \sqrt{n}\lambda_n(\log n)^{-1} = +\infty$. *Choose* $(\gamma_n^{(m)})_{n,m}$ *such that*

$$0 < \underline{\gamma} := \inf_{n,m} \gamma_n^{(m)} \leq \sup_{n,m} \gamma_n^{(m)} =: \overline{\gamma} < 2/(1+\kappa). \tag{13}$$

Then there exists a sequence m_n *with* $\lim_{n\to+\infty} m_n = +\infty$ *such that*

$$\lim_{n\to+\infty} \left\| \beta_{\lambda_n,R,n}^{m_n} - \beta_R^\dagger \right\|_2 = 0 \qquad \textit{with probability } 1, \tag{14}$$

where $\beta_{\lambda_n,R,n}^{m_n}$ *is defined in* (12) *and* $\beta_R^\dagger$ *in* (10).

Remark 2. *In order to have finite sample bounds it is necessary to have an estimate of the approximation error* $\left\|\beta_{\lambda,R} - \beta_R^\dagger\right\|_2$ *(see* (18)*). A rate of convergence cannot be obtained in the general case as proved in the "no free lunch theorem", Györfi et al. (2002). To obtain an estimate of the convergence rate it is necessary to have some a priori information on* $\beta_R^\dagger$*, given by the so called* source conditions*. In the non-quadratic case the problem of determining source counditions giving a rate of convergence of polynomial type is an open problem. The theory is completely clear only in the quadratic case, see Cucker and Smale (2002a) and the discussion in De Mol et al. (2009).*

Remark 3. The hypothesis $\lambda_n \in (0, 1]$ is not essential, and can be removed. In this case the condition $\sup_{n,m} \gamma_n^{(m)} = \overline{\gamma} < 2/(1+\kappa)$ must be replaced with $\sup_n \gamma_n^{(m)} = \overline{\gamma} < 2/(\sup \lambda_n + \kappa)$. Note that, since $\lambda_n \to 0$, it follows $2/(\sup \lambda_n + \kappa) > 0$. In a similar way the initialization $\beta^0_{\lambda,R,n} = 0$ simplifies the proofs, but is not mandatory, and can be replaced with an arbitrary one, on condition that $(\beta^0_{\lambda,R,n})_n$ is bounded.

As a consequence of Theorem 1, we obtain consistency for prediction of the estimator corresponding to the coefficients $\beta^{m_n}_{\lambda_n,R,n}$. We adapt the definition of a consistent estimator to our context (see Def. 7.9 in Schervish (1995)).

Definition 1. *We say that an estimator* $\omega \mapsto f_n(\omega) \in \mathcal{H}$ *of the regression function* f^* *is* consistent *if*

$$\mathbb{P}\left\{\omega \,:\, \lim_n \mathbf{E}\left[\|(f_n(\omega))(X) - f^*(X)\|^2\right] = 0\right\} = 1.$$

Corollary 1. *The map* $f_n : \Omega \to \mathcal{H}$ *given by*

$$\omega \mapsto f_{\beta^{m_n}_{\lambda_n,R,n}(\omega)} \in \mathcal{H},$$

where $(m_n)_n$ *and* $(\lambda_n)_n$ *are sequences obtained applying Theorem 1, is a consistent estimator of* f^**.*

Proof. Since $(f_n(\omega))(X) = \Phi_{\mathbb{P}}(\beta^{m_n}_{\lambda_n,R,n}(\omega))$ and $f^*(X) = \Phi_{\mathbb{P}}(\beta_R^\dagger)$ by (7) and the fact that $\beta_R^\dagger \in \mathcal{C}$, we get

$$\mathbf{E}\left[\|(f_n(\omega))(X) - f^*(X)\|^2\right] = \left\|\Phi_{\mathbb{P}}(\beta^{m_n}_{\lambda_n,R,n}(\omega) - \beta_R^\dagger)\right\|^2_{\mathbb{P}} \leq \|\Phi_{\mathbb{P}}\|^2 \left\|\beta^{m_n}_{\lambda_n,R,n}(\omega) - \beta_R^\dagger\right\|^2_2,$$

which converges to 0 with probability 1 by Theorem 1. Consistency is thus proved. □

Some comments are in order. Theorem 1 provides the consistency for variable selection of $\beta^{m_n}_{\lambda_n,R,n}$, which we measure in terms of the ℓ^2-norm. Putting together this result with the consistency property stated in Corollary 1, we are able to exhibit a unique estimator converging to the regression function whose coefficients are the asymptotic solutions of the Tikhonov regularization, the elastic-net (see De Mol et al. (2009); Zou and Hastie (2005)) and the Lasso respectively. Note that, usually the problem of consistency of the Lasso regularization is rather complex to deal with because of the non uniqueness of the minimal ℓ^1- norm solution; here, we bypass the problem obtaining an algorithm which selects a particular element in the set of these solutions.

3.1. Derivation of the algorithm

We conclude this section by proving that the choice of the constants $\gamma_n^{(m)}$ in Theorem 1 allows us to view the proposed family of estimators $\{\beta^m_{\lambda,R,n}\}_m$(for fixed n) as an approximation in probability of the minimizer $\omega \mapsto \beta_{\lambda,R,n}(\omega)$ on B_R of the regularized empirical risk $\mathcal{E}^n_\lambda(\omega,\cdot)$ defined by

$$\begin{array}{rccl} \mathcal{E}^n_\lambda: & \Omega \times \ell^2(\Gamma) & \to & [0,+\infty) \\ & (\omega,\beta) & \mapsto & \|\Phi_n(\omega)(\beta) - \mathbf{Y}(\omega)\|_n^2 + \lambda \|\beta\|_2^2 \end{array} . \tag{15}$$

Note that $\beta_{\lambda,R,n}(\omega)$ is well defined since $\mathcal{E}^n_\lambda(\omega,\cdot)$ is a lower semicontinuous, coercive and strictly convex functional for all $\omega \in \Omega$ and B_R is a closed and convex subset of $\ell^2(\Gamma)$. In addition $\beta_{\lambda,R,n}$ is a random variable thanks to Rockafellar (1976), Theorem 2K.
It is a well known fact that $\beta_{\lambda,R,n}(\omega)$ can be computed through the iterative projected algorithm (which is a particular case of the forward-backward scheme)

$$\beta^{m+1}_{\lambda,R,n}(\omega) = P_R\left[(1-\lambda\gamma_n^{(m)}(\omega))\beta^m_{\lambda,R,n}(\omega) + \gamma_n^{(m)}(\omega)\Phi_n^*(\omega)(\mathbf{Y}(\omega) - \Phi_n(\omega)\beta^m_{\lambda,R,n}(\omega)\right] \tag{16}$$

with

$$0 < \inf_m \gamma_n^{(m)}(\omega) \le \sup_m \gamma_n^{(m)}(\omega) < 2/(\|\Phi_n^*(\omega)\Phi_n(\omega)\| + \lambda), \tag{17}$$

since for fixed n, the sequence of random variables $(\beta^m_{\lambda,R,n})_m$ is pointwise convergent to the estimator $\beta_{\lambda,R,n}$ (see Combettes and Wajs (2005); Fornasier et al. (2008) and Theorem 6 in Umanità and Villa (2010)). Moreover, the possibility of choosing the step-size $\gamma_n^{(m)}$ adaptively improves the convergence rate (see Fornasier et al. (2008)).

Our algorithm (12) can be obtained by (16) taking the constants $\gamma_n^{(m)}(\omega)$ regardless of ω and satisfying equation (13). Note that, in this way, $\gamma_n^{(m)}(\omega) := \gamma_n^{(m)}$ fulfills condition (17) for all ω thanks to equation (A.9), and so $(\beta^m_{\lambda,R,n})_m$ is pointwise convergent to $\beta_{\lambda,R,n}$.

4. Proof of Theorem 1

We decompose the quantity in (14) in the sum of two terms, one being deterministic, and the other depending on the sampling. Since

$$\mathcal{C} = \operatorname*{argmin}_{\beta\in\ell^2(\Gamma)} \|\Phi_{\mathbb{P}}\beta - Y\|^2_{\mathbb{P}},$$

and $\beta_R^\dagger$ belongs to $\mathcal{C}$, $\beta_R^\dagger$ minimizes the discrepancy $\|\Phi_{\mathbb{P}}\beta - Y\|^2_{\mathbb{P}}$ on $\ell^2(\Gamma)$. Hence, as usually happens in the inverse problems theory, $\beta_R^\dagger$ can be approximated by the unique minimizer $\beta_{\lambda,R}$ on B_R of the regularized expected risk $\mathcal{E}_\lambda$, $\lambda > 0$, where

$$\mathcal{E}_\lambda(\beta) = \|\Phi_{\mathbb{P}}(\beta) - Y\|^2_{\mathbb{P}} + \lambda\|\beta\|_2^2, \quad \beta \in \ell^2(\Gamma).$$

Therefore it is natural to consider the following decomposition:

$$\|\beta^{m+1}_{\lambda,R,n}(\omega) - \beta_R^\dagger\|_2 \le \|\beta^{m+1}_{\lambda,R,n}(\omega) - \beta_{\lambda,R}\|_2 + \|\beta_{\lambda,R} - \beta_R^\dagger\|_2, \tag{18}$$

for fixed $\omega \in \Omega$. The first term is an approximation of the so called *sample error*, while the second is named *approximation error*. Regarding the latter, Theorem 44 of Dontchev and Zolezzi (1993) gives

$$\lim_{\lambda\to 0} \|\beta_{\lambda,R} - \beta_R^\dagger\|_2 = 0, \tag{19}$$

when $R \geq R_\infty$, and Combettes and Wajs (2005); Umanità and Villa (2010) gives

$$\beta_{\lambda,R} = P_R\left[(1-\lambda\gamma)\beta_{\lambda,R} + \gamma\Phi_{\mathbb{P}}^*(Y - \Phi_{\mathbb{P}}\beta_{\lambda,R})\right] \tag{20}$$

for all $\gamma \in \mathbb{R}$.

Next we bound the first term in (18).

Proposition 1. *Consider $\lambda_n \in (0,1]$, $\bar{\gamma}, \underline{\gamma}$ as in (13) and let*

$$q_n := \max\left\{1 - \underline{\gamma}\lambda_n, \bar{\gamma}\left(\lambda_n + \kappa\right) - 1\right\}. \tag{21}$$

Then $q_n < 1$ and for all $\omega \in \Omega$ the quantity $\|\beta_{\lambda,R,n}^{m+1}(\omega) - \beta_{\lambda,R}\|_2$ is bounded from above by

$$\|\beta_{\lambda,R}\|_2\, q_n^{m+1} + \left(\|\Phi_n^*(\omega)\mathbf{W}(\omega)\|_2 + \|\beta_{\lambda,R} - \beta_R^\dagger\|_2 \cdot \|\Phi_n^*(\omega)\Phi_n(\omega) - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS}\right)\frac{\bar{\gamma}}{1-q_n},$$

where $\mathbf{W}(\omega) := (W_1(\omega), \ldots, W_n(\omega))$, $W_i := Y_i - f^*(X_i)$.

Proof. Let $\omega \in \Omega$: since ω remains fixed in the whole proof, we omit the explicit dependence on it in the following definitions. Let $n \in \mathbb{N}$ and define a bounded operator $A_n^{(m)}$ on $\ell^2(\Gamma)$ by setting

$$A_n^{(m)} := \left(1 - \lambda\gamma_n^{(m)}\right) I - \gamma_n^{(m)}\Phi_n^*\Phi_n.$$

If we consider

$$\begin{aligned} T_n^{(m)}(\beta) &:= P_R\left[(1-\lambda\gamma_n^{(m)})\beta + \gamma_n^{(m)}\Phi_n^*(\mathbf{Y} - \Phi_n\beta)\right] \\ &= P_R\left[A_n^{(m)}\beta + \gamma_n^{(m)}\Phi_n^*\mathbf{Y}\right], \end{aligned}$$

$$T_\gamma(\beta) := P_R\left[(1-\lambda\gamma)\beta + \gamma\Phi_{\mathbb{P}}^*(Y - \Phi_{\mathbb{P}}\beta)\right], \qquad (\gamma > 0),$$

for all $\beta \in \ell^2(\Gamma)$, then $\beta_{\lambda,R,n}^{m+1} = T_n^{(m)}\left(\beta_{\lambda,R,n}^m\right)$ and $\beta_{\lambda,R} = T_\gamma(\beta_{\lambda,R})$ for all $\gamma > 0$ (by equation (20)). Therefore,

$$\begin{aligned} \|\beta_{\lambda,R,n}^{m+1} - \beta_{\lambda,R}\|_2 &\leq \|T_n^{(m)}\left(\beta_{\lambda,R,n}^m\right) - T_n^{(m)}(\beta_{\lambda,R})\|_2 + \|T_n^{(m)}(\beta_{\lambda,R}) - \beta_{\lambda,R}\|_2 \\ &\leq \|A_n^{(m)}\| \cdot \|\beta_{\lambda,R,n}^m - \beta_{\lambda,R}\|_2 + \|T_n^{(m)}(\beta_{\lambda,R}) - \beta_{\lambda,R}\|_2 \end{aligned} \tag{22}$$

thanks to the non-expansiveness of P_R. Choosing $\gamma = \gamma_n^{(m)}$ we have

$$\begin{aligned} \|T_n^{(m)}(\beta_{\lambda,R}) - \beta_{\lambda,R}\|_2 &= \|T_n^{(m)}(\beta_{\lambda,R}) - T_{\gamma_n^{(m)}}(\beta_{\lambda,R})\|_2 \\ &\leq \gamma_n^{(m)}\|\Phi_n^*\mathbf{Y} - \Phi_{\mathbb{P}}^*Y - \left(\Phi_n^*\Phi_n - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\right)\beta_{\lambda,R}\|_2; \end{aligned}$$

now, since $Y = f^*(X) + W = \Phi_{\mathbb{P}}\beta_R^\dagger + W$ by (10) and $\Phi_{\mathbb{P}}^*W = 0$ by assumption (2), we get

$$\Phi_{\mathbb{P}}^*Y = \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\beta_R^\dagger$$

and by (A.9)

$$\Phi_n^*\mathbf{Y} = \Phi_n^*\Phi_n\beta_R^\dagger + \Phi_n^*\mathbf{W},$$

so that

$$\begin{aligned}\|T_n^{(m)}(\beta_{\lambda,R}) - \beta_{\lambda,R}\|_2 &\leq \gamma_n^{(m)} \| \left(\Phi_n^*\Phi_n - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\right)\left(\beta_R^\dagger - \beta_{\lambda,R}\right) + \Phi_n^*\mathbf{W}\| \\ &\leq \gamma_n^{(m)} \|\Phi_n^*\Phi_n - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS} \cdot \|\beta_{\lambda,R} - \beta_R^\dagger\|_2 \\ &+ \gamma_n^{(m)} \|\Phi_n^*\mathbf{W}\|_2.\end{aligned}$$

Substituting in equation (22) we obtain

$$\begin{aligned}\|\beta_{\lambda,R,n}^{m+1} - \beta_{\lambda,R}\|_2 &\leq \|A_n^{(m)}\| \cdot \|\beta_{\lambda,R,n}^m - \beta_{\lambda,R}\|_2 + \gamma_n^{(m)} \|\Phi_n^*\Phi_n - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS} \cdot \|\beta_{\lambda,R} - \beta_R^\dagger\|_2 \\ &+ \gamma_n^{(m)} \|\Phi_n^*\mathbf{W}\|_2.\end{aligned}$$

Iterating and recalling that $\beta_{\lambda,R,n}^0 = 0$ we get

$$\begin{aligned}\|\beta_{\lambda,R,n}^{m+1} - \beta_{\lambda,R}\|_2 &\leq \prod_{j=0}^{m} \|A_n^{(j)}\| \cdot \|\beta_{\lambda,R}\|_2 \\ &+ \left(\|\Phi_n^*\mathbf{W}\|_2 + \|\Phi_n^*\Phi_n - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS}\|\beta_{\lambda,R} - \beta_R^\dagger\|_2\right) \cdot \\ &\cdot \left(\gamma_n^{(m)} + \sum_{j=0}^{m-1} \gamma_n^{(j)} \prod_{k=j+1}^{m} \|A_n^{(k)}\|\right). \qquad (23)\end{aligned}$$

On the other hand it holds

$$\begin{aligned}\|A_n^{(k)}\| &= \left\| \left(1 - \lambda_n\gamma_n^{(k)}\right) I - \gamma_n^{(k)}(\Phi_n^*\Phi_n) \right\| \\ &\leq \max\left\{ \left|1 - \underline{\gamma}\lambda_n\right|, \left|1 - \bar{\gamma}\left(\lambda_n + \kappa\right)\right| \right\},\end{aligned}$$

where the second bound holds for all $k \in \mathbb{N}$ and $\omega \in \Omega$ thanks to Proposition 3 in Appendix. Evaluating explicitly the maximum in the previous equation we get

$$\|A_n^{(k)}\| \leq \max\left\{1 - \underline{\gamma}\lambda_n, \bar{\gamma}\left(\lambda_n + \kappa\right) - 1\right\} = q_n$$

for all k. By (13) it immediately follows that $q_n < 1$. Finally, we have

$$\sum_{j=0}^{m-1} \gamma_n^{(j)} \prod_{k=j+1}^{m} \|A_n^{(k)}\|_2 + \gamma_n^{(m_n)} \leq \sum_{j=0}^{m} \bar{\gamma} q_n^{m-j} = \bar{\gamma}\frac{1 - q_n^{m+1}}{1 - q_n} \leq \bar{\gamma}\frac{1}{1 - q_n}. \qquad (24)$$

Substituting in inequality (23) we get the thesis. □

We can now prove Theorem 1.

Proof. Let $(q_n)_n \subseteq (0,1)$ be the sequence defined in Proposition 1. Since

$$\lim_{m \to +\infty} q_n^m = 0 \qquad \text{for all } n \in \mathbf{N},$$

via a diagonal procedure, it is possible to select a subsequence $(m_n)_n$ such that

$$\lim_{n \to +\infty} q_n^{m_n} = 0.$$

By the triangular inequality we have (see equations (19) and (20))

$$\|\beta^{m_n}_{\lambda_n,R,n} - \beta^\dagger_R\|_2 \le \|\beta^{m_n}_{\lambda_n,R,n} - \beta_{\lambda_n,R}\|_2 + \|\beta_{\lambda_n,R} - \beta^\dagger_R\|_2,$$

where $\lim_n \|\beta_{\lambda_n,R} - \beta^\dagger_R\|_2 = 0$ by equation (19). Therefore, it is enough to prove that $\lim_n \|\beta^{m_n}_{\lambda_n,R,n} - \beta_{\lambda_n,R}\|_2 = 0$ with probability 1.
Proposition 1 allows us to bound from above the quantity $\|\beta^{m_n}_{\lambda_n,R,n}(\omega) - \beta_{\lambda_n,R}\|_2$ by

$$q_n^{(m_n+1)} \|\beta_{\lambda,R}\|_2 + \left(\|\Phi_n^*(\omega)\mathbf{W}(\omega)\|_2 + \|\Phi_n^*(\omega)\Phi_n(\omega) - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS}\|\beta_{\lambda_n,R} - \beta^\dagger_R\|_2\right)\frac{\bar\gamma}{1-q_n},$$

where the first term goes to zero for $n \to \infty$. Denoting by $M := \sup_{n\ge1} \|\beta_{\lambda_n,R} - \beta^\dagger_R\|_2 < \infty$ (see equation (19)), the second term in the above equation is smaller than

$$\left(\frac{\sqrt{n}}{\log n}\|\Phi_n^*(\omega)\mathbf{W}(\omega)\|_2 + \frac{\sqrt{n}}{\log n}\|\Phi_n^*(\omega)\Phi_n(\omega) - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS}M\right)\frac{\log n}{\sqrt{n}}\frac{\bar\gamma}{1-q_n},$$

and

$$\lim_n \frac{\sqrt{n}}{\log n}\|\Phi_n^*(\omega)\mathbf{W}(\omega)\|_2 = \lim_n \frac{\sqrt{n}}{\log n}\|\Phi_n^*(\omega)\Phi_n(\omega) - \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\|_{HS} = 0$$

with probability 1 thanks to Corollary 2. Finally, since the definition of q_n in (21) gives

$$\frac{\sqrt{n}}{\log n}(1-q_n) = \frac{\sqrt{n}}{\log n}\min\left\{\underline{\gamma}\lambda_n, 2 - \bar\gamma(\lambda_n+\kappa)\right\},$$

the assumptions on λ_n and (13) imply $\sqrt{n}(1-q_n)/\log n \to \infty$, so that we can conclude

$$\lim_n \|\beta^{m_n}_{\lambda_n,R,n} - \beta^\dagger_R\|_2 = 0 \quad \text{with probability 1.}$$

□

Appendix A. Auxiliary results

In this section we reported some statements and proofs of known facts for the sake of completeness.

We introduce the map $\Phi_x : \ell^2(\Gamma) \to \mathcal{Y}$ by setting

$$\Phi_x\beta := f_\beta(x), \qquad x \in \mathcal{X},\ \beta \in \ell^2(\Gamma). \tag{A.1}$$

In particular from Proposition 3 and Lemma 1 in De Mol et al. (2009), for every $x \in \mathcal{X}$ the map Φ_x is a Hilbert-Schmidt operator such that

$$\mathrm{tr}(\Phi_x^*\Phi_x) = \mathrm{tr}(\Phi_x\Phi_x^*) = \kappa(x) \le \kappa \tag{A.2}$$

and its adjoint $\Phi_x^* : \mathcal{Y} \to \ell^2(\Gamma)$ is given by:

$$(\Phi_x^* y)_\gamma := \langle y, \varphi_\gamma(x)\rangle, \qquad y \in \mathcal{Y},\ s \in \Gamma. \tag{A.3}$$

Moreover we will consider the maps

$$\begin{aligned} &\Phi_X\beta : \Omega \to \mathcal{Y}, && \Phi_X\beta := f_\beta \circ X \\ &\Phi_X^* Z : \Omega \to \ell^2(\Gamma), && (\Phi_X^* Z)(\omega) = \Phi^*_{X(\omega)}(Z(\omega)) \end{aligned}$$

for all $\omega \in \Omega,\ \beta \in \ell^2(\Gamma)$ and random variables $Z : \Omega \to \mathcal{Y}$.

Remark 4. *The functions defined above are well defined random variables.*

Proof. Concerning $\Phi_X\beta$, it is enough to prove its measurability. By definition we have $\Phi_X\beta(\omega) = \sum_{\gamma\in\Gamma}\beta_\gamma\varphi_\gamma(X(\omega))$. Since the functions φ_γ and X are measurable, the same holds for $\Phi_X\beta$.

The measurability of Φ_X^*Z follows from the fact that the map $\omega \mapsto \langle y, \Phi_X^*(Z)(\omega)\rangle$ is measurable for each $y \in \ell^2(\Gamma)$ and $\ell^2(\Gamma)$ is a separable space. □

Below we recall some useful results shown in De Mol et al. (2009) (see Lemma 1 and Proposition 1).

Proposition 2. *The following facts hold.*

1. *For all $\beta \in \ell^2(\Gamma)$, $\Phi_X\beta$ belongs to $L^2(\Omega, \mathbb{P}; \mathcal{Y})$ and*

$$\Phi_{\mathbb{P}} : \ell^2(\Gamma) \to L^2(\Omega, \mathbb{P}; \mathcal{Y}), \qquad \Phi_{\mathbb{P}}\beta = \Phi_X\beta$$

   is a Hilbert-Schmidt operator such that

$$\mathrm{tr}(\Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}) = \mathrm{tr}(\Phi_{\mathbb{P}}\Phi_{\mathbb{P}}^*) = \mathbf{E}[\kappa(X)] \le \kappa. \tag{A.4}$$

2. *$\Phi_X^*\Phi_X : \Omega \to HS(\ell^2)$, defined by setting $\Phi_X^*\Phi_X(\omega) := \Phi_{X(\omega)}^*\Phi_{X(\omega)}$ is a random variable with*

$$\mathbf{E}[\Phi_X^*\Phi_X] = \Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}. \tag{A.5}$$

3. *Y belongs to $L^2(\Omega, \mathbb{P}; \mathcal{Y})$, Φ_X^*Y has finite expectation and*

$$\Phi_{\mathbb{P}}^*Y = \mathbf{E}[\Phi_X^*Y] \tag{A.6}$$
$$(\Phi_{\mathbb{P}}^*Y)_\gamma = \mathbf{E}[\langle Y, \varphi_\gamma(X)\rangle] \tag{A.7}$$
$$(\Phi_{\mathbb{P}}^*\Phi_{\mathbb{P}}\beta)_\gamma = \mathbf{E}[\langle \Phi_{\mathbb{P}}\beta, \varphi_\gamma(X)\rangle]. \tag{A.8}$$

Proof. We only prove in details the measurability of the map $\Phi_X^*\Phi_X$; all the other proofs can be found in De Mol et al. (2009) (see Lemma 1 and Proposition 1). Since $HS(\ell^2)$ is a separable Hilbert space, it is enough to prove that $\omega \mapsto [\Phi_{X(\omega)}^*(\Phi_{X(\omega)})\beta]_\gamma = \langle \sum_{i\in\Gamma}\beta_i\varphi_i(X(\omega)), \varphi_\gamma(X(\omega))\rangle$ is measurable for each $\beta \in \ell^2(\Gamma)$, $\gamma \in \Gamma$. This follows from the measurability of φ_γ and of the scalar product. □

Let now $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n-observed i.i.d. copies of (X, Y), and consider the Hilbert space $\mathcal{Y}^n$ with the scalar product

$$\langle (z_1, \ldots, z_n), (w_1, \ldots, w_n)\rangle_n := \frac{1}{n}\sum_{i=1}^n \langle z_i, w_i\rangle .$$

Proposition 3. *For all $\omega \in \Omega$ the map $\Phi_n(\omega)$ defined in* (11) *is a Hilbert-Schmidt operator with adjoint $\Phi_n(\omega)^* : \mathcal{Y}^n \to \ell^2(\Gamma)$ given by*

$$\Phi_n(\omega)^*(z_1, \ldots, z_n) = \frac{1}{n}\sum_{i=1}^n \Phi_{X_i(\omega)}^* z_i.$$

Moreover, the random variable

$$\Phi_n^*\Phi_n : \Omega \to HS(\ell^2)$$

satisfies

$$\Phi_n^*\Phi_n = \frac{1}{n}\sum_{i=1}^{n} \Phi_{X_i}^*\Phi_{X_i}, \qquad \operatorname{tr}((\Phi_n^*\Phi_n)(\omega)) = \frac{1}{n}\sum_{i=1}^{n} \kappa(X_i) \leq \kappa. \tag{A.9}$$

Proof. Since

$$\begin{aligned}\sum_{s\in\Gamma} \| \left(\Phi_n(\omega)\right) e_\gamma \|_n^2 &= \sum_{s\in\Gamma} \frac{1}{n}\sum_{i=1}^{n} \|(\Phi_{X_i} e_\gamma)(\omega)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \operatorname{tr}(\Phi_{X_i(\omega)}^*\Phi_{X_i(\omega)}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \kappa(X_i(\omega)) \leq \kappa\end{aligned}$$

by equation (A.2), any $\Phi_n(\omega)$ is a Hilbert-Schmidt operator and the second equation in (A.9) is fulfilled.

Given $(z_1, \ldots, z_n) \in \mathcal{Y}^n$ and $\beta \in \ell^2(\Gamma)$ we have

$$\begin{aligned}\langle \Phi_n(\omega)^*(z_1, \ldots, z_n), \beta \rangle_2 &= \langle (z_1, \ldots, z_n), \Phi_n(\omega)\beta \rangle_n = \frac{1}{n}\sum_{i=1}^{n} \left\langle z_i, \Phi_{X_i(\omega)}\beta \right\rangle \\ &= \sum_{i=1}^{n} \left\langle \Phi_{X_i(\omega)}^* z_i, \beta \right\rangle_2,\end{aligned}$$

so that $\Phi_n(\omega)^*(z_1, \ldots, z_n) = \frac{1}{n}\sum_{i=1}^{n} \Phi_{X_i(\omega)}^* z_i$.

Finally,

$$\left((\Phi_n^*\Phi_n)(\omega)\right)\beta = \left(\Phi_n(\omega)^*\Phi_n(\omega)\right)\beta = \frac{1}{n}\sum_{i=1}^{n} \Phi_{X_i(\omega)}^*\Phi_{X_i(\omega)}\beta = \frac{1}{n}\sum_{i=1}^{n} \left((\Phi_{X_i}^*\Phi_{X_i})(\omega)\right)\beta$$

holds for every $\omega \in \Omega$ and $\beta \in \ell^2(\Gamma)$, and so first equation (A.9) follows. □

In order to prove that the proposed estimators $\beta_{\lambda,R,n}^m$ are random variables in $\ell^2(\Gamma)$, we need the following result.

Lemma 1. *Given the random variables $\alpha : \Omega \to \ell^2(\Gamma)$ and $\mathbf{Z} = (Z_1, \ldots, Z_n) : \Omega \to \mathcal{Y}^n$, the following maps are random variables too:*

$$\begin{aligned}&\Phi_n\alpha : \Omega \to \mathcal{Y}^n, && (\Phi_n\alpha)(\omega) := \left(\Phi_n(\omega)\right)\alpha(\omega) \\ &\Phi_n^*\mathbf{Z} : \Omega \to \ell^2(\Gamma), && \left(\Phi_n^*\mathbf{Z}\right)(\omega) := \left(\Phi_n^*(\omega)\right)\mathbf{Z}(\omega).\end{aligned} \tag{A.10}$$

The proof is similar to Remark 4.

In the proof of Theorem 1 we use the following result based on the concentration inequalities in Hilbert spaces and showed in Pinelis (1999).

Lemma 2. *Let $(\xi_i)_{i=1}^n$ be a sequence of i.i.d. zero mean random variables with values in a real separable Hilbert space such that*

$$\mathbf{E}[\|\xi_i\|^m] \leq \frac{1}{2} m! M^2 H^{m-2} \qquad \forall\, m \geq 2, \tag{A.11}$$

with M and H positive constants. Then,

$$\lim_n \frac{1}{\sqrt{n}\log n} \left\| \sum_{i=1}^n \xi_i \right\| = 0$$

with probability 1.

Proof. Theorem 8.6 in Pinelis (1994) (see also Pinelis (1999)) assures that, for all $n \geq 1$ and $\epsilon > 0$, the inequality

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\| \geq \epsilon \right) \leq 2 e^{-\frac{n\epsilon^2}{M^2 + H\epsilon + M\sqrt{M^2 + 2H\epsilon}}}$$

holds. Therefore, given $\epsilon > 0$, we also obtain

$$\mathbb{P}\left(\frac{1}{\sqrt{n}\log n} \left\| \sum_{i=1}^n \xi_i \right\| \geq \epsilon \right) = \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\| \geq \epsilon \frac{\log n}{\sqrt{n}} \right) \leq 2 e^{-A(n,\epsilon)} = 2 \left(\frac{1}{n} \right)^{\frac{A(n,\epsilon)}{\log n}},$$

with

$$A(n,\epsilon) := \frac{\epsilon^2 (\log n)^2}{M^2 + H\epsilon \frac{\log n}{\sqrt{n}} + M\sqrt{M^2 + 2H\epsilon \frac{\log n}{\sqrt{n}}}}.$$

It follows that

$$\sum_{n \geq 1} \mathbb{P}\left(\frac{1}{\sqrt{n}\log n} \left\| \sum_{i=1}^n \xi_i \right\| \geq \epsilon \right) \leq 2 \sum_{n \geq 1} \left(\frac{1}{n} \right)^{\frac{A(n,\epsilon)}{\log n}}.$$

Since

$$\frac{A(n,\epsilon)}{\log n} = \frac{\epsilon^2 \log n}{M^2 + H\epsilon \frac{\log n}{\sqrt{n}} + M\sqrt{M^2 + 2H\epsilon \frac{\log n}{\sqrt{n}}}}$$

tends to $+\infty$, the series $\sum_{n \geq 1} \left(\frac{1}{n}\right)^{\frac{A(n,\epsilon)}{\log n}}$ is convergent, and then the Borel-Cantelli lemma gives the thesis. □

Corollary 2. *Given $\mathbf{W} = (W_1, \ldots W_n)$ as in Proposition 1, we have*

$$\mathbb{P}\left(\lim_n \frac{\sqrt{n}}{\log n} \|\Phi_n^*(\omega)\mathbf{W}(\omega)\|_2 = 0 \right) = \mathbb{P}\left(\lim_n \frac{\sqrt{n}}{\log n} \|\Phi_n^*(\omega)\Phi_n(\omega) - \Phi_{\mathbb{P}}^* \Phi_{\mathbb{P}}\|_{HS} = 0 \right) = 1.$$

Proof. Proposition 3 implies

$$\Phi_n^* \mathbf{W} = \frac{1}{n} \sum_{i=1}^n \Phi_{X_i}^* W_i, \qquad \Phi_n^* \Phi_n - \Phi_{\mathbb{P}}^* \Phi_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^n \left(\Phi_{X_i}^* \Phi_{X_i} - \Phi_{\mathbb{P}}^* \Phi_{\mathbb{P}} \right).$$

Now, since $X_1, \dots, X_n$ are i.i.d. and

$$\begin{aligned} \mathbf{E}[\Phi^*_{X_i} W_i] &= \mathbf{E}[\mathbf{E}[\Phi^*_{X_i} W_i \,|X_i\,]] = 0 \\ \mathbf{E}[\Phi^*_{X_i} \Phi_{X_i}] &= \Phi^*_{\mathbb{P}} \Phi_{\mathbb{P}} \end{aligned}$$

by Proposition 2 and assumption (2), the random variables $\left\{\Phi^*_{X_i} W_i\right\}_i$ and $\left\{\Phi^*_{X_i}\Phi_{X_i} - \Phi^*_{\mathbb{P}}\Phi_{\mathbb{P}}\right\}_i$ are i.i.d. and have zero mean. Moreover, for all $m \geq 2$ they satisfy

$$\mathbf{E}[\left\|\Phi^*_{X_i} W_i\right\|_2^m] = \mathbf{E}[(\sum_{s\in\Gamma} |\,\langle \varphi_\gamma(X_i), W_i\rangle\,|^2)^{m/2}] \leq \kappa^{m/2}\mathbf{E}[\|W_i\|^m] \leq \kappa^{m/2}\frac{m!}{2}\sigma^2 L^{m-2}$$

thanks to (5) and (4), and

$$\mathbf{E}[\left\|\Phi^*_{X_i}\Phi_{X_i} - \Phi^*_{\mathbb{P}}\Phi_{\mathbb{P}}\right\|_{HS}^m] \leq (2\kappa)^m \leq \frac{m!}{2}(2\kappa)^2\kappa^{m-2}$$

since

$$\left\|\Phi^*_{X_i}\Phi_{X_i}\right\|_{HS} \leq \mathrm{tr}(\Phi^*_{X_i}\Phi_{X_i}) \leq \kappa, \qquad \left\|\Phi^*_{\mathbb{P}}\Phi_{\mathbb{P}}\right\|_{HS} \leq \mathrm{tr}(\Phi^*_{\mathbb{P}}\Phi_{\mathbb{P}}) \leq \kappa$$

(see (A.2), (A.4)) and the inequality $2^{m-1} \leq m!$ holds.

Applying Lemma 2 to variables $\Phi^*_{X_i} W_i$ and $\Phi^*_{X_i}\Phi_{X_i} - \Phi^*_{\mathbb{P}}\Phi_{\mathbb{P}}$ we get the thesis. □

References

Bunea, F., 2008. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. Electron. J. Stat. 2, 1153–1197.

Candes, E., Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. 35 (6), 2313–2351.

Chen, S., Donoho, D., Saunders, M., 1998. Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20 (1), 33–61.

Combettes, P. L., Wajs, V. R., 2005. Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. 4 (4), 1168–1200 (electronic).

Cucker, F., Smale, S., 2002a. Best choices for regularization parameters in learning theory: on the bias-variance problem. Found. Comput. Math. 2 (4), 413–428.

Cucker, F., Smale, S., 2002b. On the mathematical foundations of learning. Bull. Amer. Math. Soc. (N.S.) 39 (1), 1–49 (electronic).

Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics 57, 1413–1457.

De Mol, C., De Vito, E., Rosasco, L., 2009. Elastic-net regularization in learning theory. J. Complexity 25 (2), 201–230.

De Vito, E., Pereverzev, S., Rosasco, L., 2010. Adaptive kernel methods via the balancing principle, to appear.

Dontchev, A. L., Zolezzi, T., 1993. Well-posed optimization problems. Vol. 1543 of Lecture Notes in Mathematics. Springer-Verlag.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

Engl, H. W., Hanke, M., Neubauer, A., 1996. Regularization of inverse problems. Mathematics and its Applications 375.

Fornasier, M., Daubechies, I., Loris, I., 2008. Accelerated projected gradient methods for linear inverse problems with sparsity constraints. J. Fourier Anal. Appl.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002. A distribution-free theory of nonparametric regression. Springer Series in Statistics. Springer-Verlag, New York.

Jin, B., Lorenz, D. A., Schiffler, S., 2009. Elastic-net regularization: error estimates and active set methods.
URL `http://arXiv:0905.0796`

Pinelis, I., 1994. Optimum bounds for the distributions of martingales in Banach spaces. Ann. Probab. 22 (4), 1679–1706.

Pinelis, I., 1999. Correction: optimum bounds for the distributions of martingales in banach spaces. Ann. Probab. 27 (4), 2119.

Poggio, T., Smale, S., 2003. The mathematics of learning: Dealing with data. Amer. Math. Soc. Notice 50 (5), 537–544.

Rockafellar, R. T., 1976. Integral functionals, normal integrands and measurable selections. In: Nonlinear operators and the calculus of variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975). Springer, Berlin, pp. 157–207. Lecture Notes in Math., Vol. 543.

Schervish, M. J., 1995. Theory of statistics. Springer Series in Statistics. Springer-Verlag, New York.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 56, 267–288.

Umanità, V., Villa, S., 2010. Elastic-net regularization: iterative algorithms and asymptotic behavior of solutions. To appear.

Van de Geer, S. A., 2008. High-dimensional generalized linear models and the lasso. Ann. Statistics 36 (2), 614–645.

Van der Vaart, A. W., Wellner, A. W., 1996. Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag.

Vapnik, V. N., 1998. Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, a Wiley-Interscience Publication.

Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.

Zou, Z., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320.