Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2011-041 September 26, 2011
CBCL-303

Nonparametric Sparsity and Regularization

Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Nonparametric Sparsity and Regularization

Sofia Mosci[‡], Lorenzo Rosasco[♯,⋆], Matteo Santoro[⋆] Alessandro Verri[‡], Silvia Villa[†]

‡ - *DISI, University of Genova, Italy*

♯ - *CBCL, McGovern Institute, Massachusetts Institute of Technology, USA*

⋆ - *Istituto Italiano di Tecnologia, Italy*

† - *DIMA, University of Genova, Italy*

mosci@disi.unige.it, lrosasco@mit.edu, msantoro@iit.it,
villa@dima.unige.it, verri@disi.unige.it

September 26, 2011

Abstract

In this work we are interested in the problems of supervised learning and variable selection when the input-output dependence is described by a nonlinear function depending on a few variables. Our goal is to consider a sparse nonparametric model, hence avoiding linear or additive models. The key idea is to measure the importance of each variable in the model by making use of partial derivatives. Based on this intuition we propose and study a new regularizer and a corresponding least squares regularization scheme. Using concepts and results from the theory of reproducing kernel Hilbert spaces and proximal methods, we show that the proposed learning algorithm corresponds to a minimization problem which can be provably solved by an iterative procedure. The consistency properties of the obtained estimator are studied both in terms of prediction and selection performance. An extensive empirical analysis shows that the proposed method performs favorably with respect to the state-of-the-art.

1 Introduction

It is now common to see practical applications, for example in bioinformatics and computer vision, where the dimensionality of the data is in the order of hundreds, thousands and even tens of thousands. Learning in such a high dimensional regime is feasible only if the quantity to be estimated satisfies some regularity assumptions. In particular, the idea behind, so called, *sparsity* is that the quantity of interest depends only on a few relevant variables (dimensions). In turn, this latter assumption is often at the basis of the construction of interpretable data models, since the relevant dimensions allow for a compact, hence interpretable, representation. An instance of the above situation is the problem of learning from samples a multivariate function which depends only on a (possibly small) subset of the variables, the *relevant* variables. Detecting such variables is the problem of variable selection.

Largely motivated by recent advances in compressed sensing and related results, the above problem has been extensively studied under the assumption that the function of interest (target function) depends *linearly* to the relevant variables. While a naive approach (trying all possible subsets of variables) would not be computationally feasible it is known that meaningful approximations can be found either by greedy methods (Tropp and Gilbert, 2007), or convex relaxation (ℓ^1 regularization a.k.a. basis pursuit or LASSO (Tibshirani, 1996; Chen et al., 1999; Efron et al., 2004)). In this context efficient algorithms (see Schmidt et al. (2007); Loris (2009) and references therein) as well as theoretical guarantees are now available (see Bühlmann and van de Geer (2011) and references therein). In this paper we are interested into the situation where the target function depends *non-linearly* to the relevant variables. This latter situation is much less understood. Approaches in the literature are mostly restricted to additive models (Hastie and Tibshirani, 1990). In such models the target function is assumed to be a sum of (non-linear) univariate functions. Solutions to the problem of variable selection in this class of models include Ravikumar et al. (2008) and are related to multiple kernel learning (Bach et al., 2004). Higher order additive models can be further considered, encoding explicitly dependence among the variables – for example assuming the target function to be also sum of functions depending on couples, triplets etc. of variables, as in Lin and Zhang (2006) and Bach (2009). Though this approach provides a more interesting, while still interpretable, model, its size/complexity is essentially more than exponential in the initial variables. Among the few papers considering the problem is worth mentioning Lafferty and Wasserman (2008) which proposes a non-parametric estimator (called RODEO).

In this paper we propose a different approach that builds on the idea that the importance of a variable, while learning a non-linear functional relation, can be captured by partial derivatives. This observation suggests to design a regularizer which favors functions where most partial derivatives are essentially zero. A question is how to make this last requirement precise and computational feasible. The first observation is that, while we cannot measure a partial derivative *everywhere*, we can do it at the training set points and hence design a data-dependent regularizer. In order to derive an actual algorithm we have to consider two further issues: How can we estimate reliably partial derivatives in high dimensions? How can we ensure that the data-driven penalty is sufficiently stable? The theory of reproducing kernel Hilbert spaces (RKHSs) provide us with tools to answer both questions. In fact, partial derivatives in a RKHS are bounded linear functionals and hence have a suitable representation allowing for efficient computations. Moreover the norm in the RKHS provides a natural further regularizer ensuring stable behavior of the empirical, derivative based penalty. Our analysis is split into 3 main parts. First we discuss the minimization of the functional defining the algorithm. If we consider the square loss function, the obtained functional is convex but not differentiable. Given the form of the functional, we suggest that proximal methods can be a suitable technique to design a computational solution. The algorithm we derive is an iterative procedure requiring a projection at each step. While such a projection cannot be computed in closed form it can be efficiently computed in an iterative fashion. We prove both in theory and practice that the obtained procedures converge with fast rates. Second, we study the prediction and selection properties of the obtained estimator. Third, we validate the empirical properties of the proposed algorithm on both toy and benchmark datasets. Our method performs favorably and often outperforms other methods.

Some of the results we present have appeared in a shorter conference version of the paper (Rosasco et al., 2010). With respect to the conference version, the current version contains: the detailed discussion of the derivation of the algorithm with all the proofs, the consistency results of Section 4, an augmented set of experiments and several further discussions. The paper is organized as follows. In section 2 we discuss our approach and present the main results in the paper. In Section 3 we discuss the computational aspects of the method. In Section 4 we prove consistency results. In Section 5 we provide an extensive empirical analysis. Finally in Section 6 we conclude with a summary of our study and a discussion of future work.

2 Sparsity Beyond linear Models: Nonparametric and Regularization

Given a training set $\mathbf{z}_n = (\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{i=1}^n$ of input output pairs, with $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathcal{Y} \subseteq \mathbb{R}$, we are interested into learning about the functional relationship between input and output.

More precisely, in statistical learning the data are assumed to be sampled identically and independently from a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$ so that if we measure the error by the square loss function, the regression function $f_\rho(x) = \int y d\rho(x, y)$ minimizes the expected risk $\mathcal{E}(f) = \int (y - f(x))^2 d\rho(x, y)$.

Finding an estimator $\hat{f}$ of f_ρ from finite data is possible, if f_ρ is sufficiently regular (Devroye et al., 1996). In this paper we are interested in the case where the regression function is *sparse* in the sense that it depends only on a subset R_ρ of the possible d variables. Finding such *relevant* variables is the problem of variable selection.

2.1 Linear and Additive Models

The above discussion can be made precise considering linear functions $f(x) = \sum_{a=1}^d \beta_a x^a$ with $x = (x^1, \ldots, x^d)$. In this case the sparsity of a function is quantified by the so called *zero-norm* $\Omega_0(f) = \#\{a = 1, \ldots, d \mid \beta_a \neq 0\}$. The zero norm, while natural for variable selection, does not lead to efficient algorithms and is often replaced by the ℓ^1 norm, that is $\Omega_1(f) = \sum_{a=1}^d |\beta_a|$. This situation is fairly well understood, see (Bühlmann and van de Geer, 2011) and references therein. Regularization with ℓ^1 regularizers, obtained by minimizing

$$\widehat{\mathcal{E}}(f) + \lambda \Omega_1(f), \qquad \widehat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2,$$

can be solved efficiently and, under suitable conditions, provides a solution close to that of a zero-norm regularization.

The above scenario can be generalized to additive models $f(x) = \sum_{a=1}^d f_a(x^a)$, where f_a are univariate functions in some (reproducing kernel) Hilbert space $\mathcal{H}_a$, $a = 1, \dots, d$. In this case the analogous of the zero-norm and the ℓ^1 norm are $\Omega_0(f) = \#\{a = 1, \dots, d \ : \ \|f_a\| \neq 0\}$ and $\Omega_1(f) = \sum_{a=1}^d \|f_a\|$, respectively. This latter setting, related to multiple kernel learning (Bach et al., 2004; Bach, 2008), has been considered for example in Ravikumar et al. (2008), see also Koltchinskii and Yuan (2010) and references therein. The simple additive model limits the way the variables interact. This can be partially alleviated considering higher order terms in the model as it is done in ANOVA decomposition (Wahba et al., 1995; Gu, 2002). More precisely, we can add to the simple additive model functions of couples $f_{a,b}(x^a, x^b)$, triplets $f_{a,b,c}(x^a, x^b, x^c)$, etc. of variables – see Lin and Zhang (2006). For example one can consider functions of the form $f(x) = \sum_{a=1}^d f_a(x^a) + \sum_{a<b} f_{a,b}(x^a, x^b)$. In this case the analogous to the zero and ℓ^1 norms are $\Omega_0(f) = \#\{a = 1, \dots, d \ : \ \|f_a\| \neq 0\} + \#\{(a,b) \ : \ a < b, \ \|f_{b,c}\| \neq 0\}$ and $\Omega_1(f) = \sum_{a=1}^d \|f_a\| + \sum_{a<b} \|f_{a,b}\|$, respectively. Note that in this case sparsity will not be in general with respect to the original variables but rather with respect to the elements in the additive model. Clearly, while this approach provides a more interesting and yet interpretable model, its size/complexity is essentially more than exponential in the initial variables. Some proposed attempts to tackle this problem are based on restricting the set of allowed sparsity patterns and can be found in Bach (2009).

The above discussion raises the following question.

What if we are interested into learning and variable selection when the functions of interest are not described by an additive model?

To the best of our knowledge the only attempt to tackle this question is the one in Lafferty and Wasserman (2008) and is based on local estimators. Similarly, Bertin and Lecué (2008) consider a local polynomial estimator. For a fixed point x, $f(x)$ is modeled as the Taylor polynomial of order 1 at x. The cost functional is thus the empirical risk regularized with the ℓ^1 norm of the Taylor coefficients (scaled w.r.t. a fixed bandwidth), so that sparsity here is over the Taylor expansion. In a different setting, the problem has been considered in DeVore et al. (2011). Recently Comminges and Dalalyan (2011) derived theoretical results on consistent estimation of the sparsity pattern in a nonparametric context, though the proposed learning procedure does not seem to be practical.

2.2 Sparsity and Regularization using Partial Derivatives

Our study starts from the observation that, if a function f is differentiable, the relative importance of a variable at a point x can be captured by the magnitude of the corresponding partial derivative[1]

$$\left| \frac{\partial f}{\partial x^a} \right|.$$

In the rest of the paper we discuss how we can use this observation to define a new notion of sparsity for non linear models and design a regularized algorithm. Regularization using derivatives is not new. The classical splines (Sobolev spaces) regularization (Wahba, 1990), as well as more modern techniques such as manifold regularization (Belkin and Niyogi, 2008) use derivatives to measure the regularity of a function. Similarly total variation regularization utilizes derivatives to define regular function. None of the above methods though allows to capture a notion of sparsity suitable both for learning and variable selection– see Remark 1.

Using partial derivatives to define a new notion of a sparsity and design a regularizer for learning and variable selection requires considering the following two issues. First, we need to quantify the relevance of a variable beyond a single input point. If the partial derivative is continuous [2] then a natural idea is to consider

$$\left\| \frac{\partial f}{\partial x^a} \right\|_{\rho_{\mathcal{X}}} = \sqrt{\int_{\mathcal{X}} \left(\frac{\partial f(x)}{\partial x^a} \right)^2 d\rho_{\mathcal{X}}(x)}. \tag{1}$$

where $\rho_{\mathcal{X}}$ is the marginal probability measure of ρ on X. While considering other L^p norms is possible, in this paper we restrict our attention to L^2. A notion of sparsity for a differentiable, non-linear function f is captured by the following functional

$$\Omega_0^{\mathrm{D}}(f) = \# \left\{ a = 1, \dots, d \ : \ \left\| \frac{\partial f}{\partial x^a} \right\|_{\rho_{\mathcal{X}}} \neq 0 \right\}, \tag{2}$$

[1] In order for the partial derivatives to be defined at all points we always assume that the closure of $\mathcal{X}$ coincides with the closure of its interior.

[2] In the following, see SEction REF, we will see that further appropriate regularity properties on f are needed depending on whether the support of $\rho_{\mathcal{X}}$ is connected or not.

Figure 1: Difference between ℓ^1/ℓ^1 and ℓ^1/ℓ^2 norm for binary matrices (white = 1, black=0), where in the latter case the ℓ^1 norm is taken over the rows (variables) and the ℓ^2 norm over the columns (samples). The two matrices have the same number of nonzero entries, and thus the same ℓ^1/ℓ^1 norm, but the value of the ℓ^1/ℓ^2 norm is smaller for the matrix on the right, where the nonzero entries are positioned to fill a subset of the rows. The situation on the right is thus favored by ℓ^1/ℓ^2 regularization.

and the corresponding relaxation is

$$\Omega_1^{\mathrm{D}}(f) = \sum_{a=1}^{d} \left\| \frac{\partial f}{\partial x^a} \right\|_{\rho_{\mathcal{X}}}.$$

The second issue comes from the fact $\rho_{\mathcal{X}}$ is only known through the training set. To obtain a practical algorithm we start by replacing the L^2 norm with an empirical version

$$\left\| \frac{\partial f}{\partial x^a} \right\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial f(x_i)}{\partial x^a} \right)^2}$$

and by replacing (1) by the data-driven regularizer,

$$\widehat{\Omega}_1^D(f) = \sum_{a=1}^{d} \left\| \frac{\partial f}{\partial x^a} \right\|_n. \tag{3}$$

While the above quantity is a natural estimate of (1) in practice it might not be sufficiently stable to ensure good function estimates where data are poorly sampled. In the same spirit of manifold regularization (Belkin and Niyogi, 2008), we then propose to further consider functions in a reproducing kernel Hilbert space (RKHS) defined by a differentiable kernel and use the penalty

$$\widehat{\Omega}_1^D(f) + \nu \|f\|_{\mathcal{H}}^2$$

where ν is a small positive number. The latter terms ensures stability while making the regularizer strongly convex. This latter property is a key for well-posedeness and generalization as we discuss in Section 4. As we will see in the following, RKHS will also be a key tool allowing computations of partial derivative of potentially high dimensional functions.

The final **learning** algorithm is given by the minimization of the functional

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \tau \left(\sum_{a=1}^{d} \left\| \frac{\partial f}{\partial x^a} \right\|_n + \nu \|f\|_{\mathcal{H}}^2 \right). \tag{4}$$

The remainder of the paper is devoted to the analysis of the above regularization algorithm. Before summarizing our main results we add two remarks.

Remark 1 (Importance of the square root). *It is perhaps useful to remark the importance for sparsity of the square root in the definition of the derivative based regularizer. We start by considering*

$$\sum_{a=1}^{d} \left\| \frac{\partial f}{\partial x^a} \right\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{d} \left(\frac{\partial f(x_i)}{\partial x^a} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f(x_i)\|^2,$$

where $\nabla f(x)$ is the gradient of f at x. This is essentially a data-dependent version of the classical penalty in Sobolev spaces which writes $\int \|\nabla f(x)\|^2 dx$, where the uniform (Lebesgue) measure is considered. It is well known that while this *regularizer measure the smoothness it does not yield any sparsity property. Then we could consider*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{d}\left|\frac{\partial f(x_i)}{\partial x^a}\right|.$$

Though this penalty (which we call ℓ^1/ℓ^1) favors sparsity, it only forces partial derivative at points to be zero. In comparison the regularizer we propose is of the ℓ^1/ℓ^2 type and utilizes the square root to "group" the values of each partial derivative at different points hence favoring functions for which each partial derivative is small at most points. The difference between penalties is illustrated in Figure 1. Finally note that we can also consider $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x_i)\|$. This regularizer, which is akin to the total variation regularizer $\int\|\nabla f(x)\|dx$, groups the partial derivatives differently and favors functions with localized singularities rather than selecting variables.

Remark 2. *As it is clear from the previous discussion, we quantify the importance of a variable based on the norm of the corresponding partial derivative. This approach makes sense only if*

$$\|D_a f\|_{\rho_{\mathcal{X}}} = 0 \;\Rightarrow\; f \text{ is constant with respect to } x_a. \tag{5}$$

The previous fact holds trivially if we assume the function f to be continuously differentiable (so that the derivative is pointwise defined, and is a continuous function) and supp$\rho_{\mathcal{X}}$ to be connected. If the latter assumption is not satisfied the situation is more complicated, as the following example shows. Suppose that $\rho_{\mathcal{X}}$ is the uniform distribution on the disjoint intervals $[-2,-1]$ and $[1,2]$, and $\mathcal{Y}=\{-1,1\}$. Moreover assume that $\rho(y|x)=\delta_{-1}$, if $x\in[-2,-1]$ and $\rho(y|x)=\delta_1$, if $x\in[1,2]$. Then, if we consider the regression function

$$f(x)=\begin{cases}-1 & \text{if } x\in[-2,-1]\\ 1 & \text{if } x\in[1,2]\end{cases}$$

we get that $f'(x)=0$ on the support of $\rho_{\mathcal{X}}$, although the variable x is relevant. To avoid such pathological situations when supp$\rho_{\mathcal{X}}$ is not connected in $\mathbb{R}^d$ we need to impose more stringent regularity assumptions that basically imply that a function which is constant on a open interval is constant everywhere. This is verified when f belongs to the RKHS defined by a polynomial kernel, or, more generally, an analytic kernel such as the Gaussian kernel.

2.3 Main Results

We summarize our main contributions.

1. We extend the representer theorem (Wahba, 1990) and show that the minimizer of (4) has the finite dimensional representation

   $$\hat{f}^{\tau}(x)=\sum_{i=1}^{n}\frac{1}{n}\alpha_i k(x_i,x)+\sum_{i=1}^{n}\sum_{a=1}^{d}\frac{1}{n}\beta_{ai}\;\frac{\partial k(s,x)}{\partial s^a}\bigg|_{s=x_i},$$

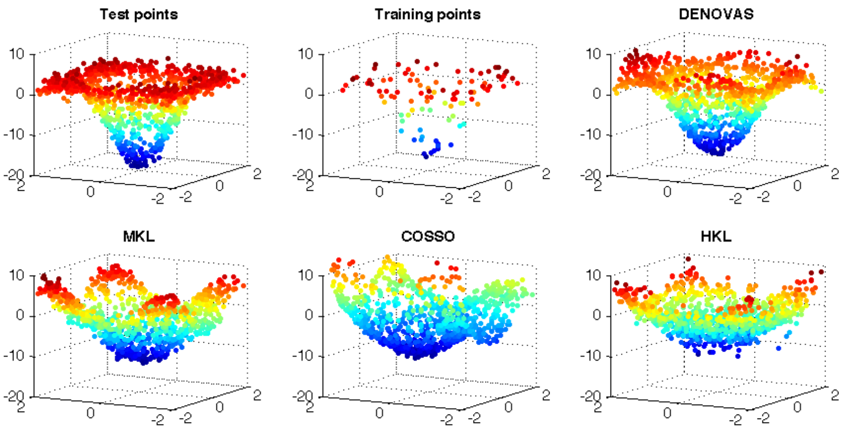
   with $\alpha,(\beta_{ai})_{i=1}^{n}\in\mathbb{R}^n$ for all $a=1,\dots,d$.

   Then we show that the coefficients in the expansion can be computed using proximal methods (Combettes and Wajs, 2005; Beck and Teboulle, 2009). More precisely, we present a fast forward-backward splitting algorithm, where the proximity operator does not admit a closed form and is thus computed in an approximated way. Using recent results for proximal methods with approximate proximity operators, we are able to prove convergence with rate for the overall procedure. The resulting algorithm requires only matrix multiplications and thresholding operations and is in terms of the coefficients α and β and matrices given by the kernel and its first and second derivatives evaluated at the training set points.

2. We study the consistency properties of the obtained estimator. We prove that, if the kernel we use is universal, then there exists a choice of $\tau=\tau_n$ depending on n such that the algorithm is universally consistent (Steinwart and Christmann, 2008), that is

   $$\lim_{n\to\infty}\mathrm{P}\Big(\,\mathcal{E}(\hat{f}^{\tau_n})-\mathcal{E}(f_\rho)>\varepsilon\Big)=0$$

Figure 2: Comparison of predictions for a radial function of 2 out of 20 variables (the 18 irrelevant variables are not shown in the figure). In the upper left plot is depicted the value of the function on the test points (left), the noisy training points (center), the values predicted for the test points by our method (DENOVAS) (right). The bottom plots represent the values predicted for the test points by state-of-the-art algorithms based on additive models. Left: Multiple kernel learning based on additive models using kernels. Center: COSSO, which is a higher order additive model based on ANOVA decomposition (Lin and Zhang, 2006). Right: Hierarchical kernel learning (Bach, 2009).



for all $\varepsilon > 0$. Moreover, we study the selection properties of the algorithm and prove that, if R_ρ is the set of relevant variables and $\hat{R}^{\tau_n}$ the set estimated by our algorithm, then the following consistency result holds

$$\lim_{n\to\infty} \mathrm{P}\left(\hat{R}^{\tau_n} \subseteq R_\rho\right) = 1.$$

3. Finally we provide an extensive empirical analysis both on simulated and benchmark data, showing that the proposed algorithm (DENOVAS) compares favorably and often outperforms other algorithms. This is particularly evident when the function to be estimated is highly non linear. The proposed method can take advantage of working in a rich, possibly infinite dimensional, hypotheses space given by a RKHS, to obtain better estimation and selection properties. This is illustrated in Figure 2, where the regression function is a nonlinear function of 2 of 20 possible input variables. With 100 training samples the algorithms we propose is the only one able to correctly solve the problem among different linear and non linear additive models. On real data our method outperforms other methods on several data sets. In most cases, the performance of our method and regularized least squares (RLS) are similar. However our method brings higher interpretability since it is able to select a smaller subset of relevant variable, while the estimator provided by RLS depends on all variables.

We start our analysis discussing how to compute efficiently a regularized solution.

3 Computing the regularized solution

In this section we study the minimization of the functional (4). We let

$$\widehat{\mathcal{E}}^\tau(f) := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \tau\left(2\widehat{\Omega}_1^D(f) + \nu\|f\|_{\mathcal{H}}^2\right). \tag{6}$$

where $\widehat{\Omega}_1^D(f)$ is defined in (3). We start observing that the term $\|f\|_{\mathcal{H}}^2$ makes the above functional coercive and strongly convex with modulus[3] $\tau\nu/2$, so that standard results (Ekeland and Temam (1976)) ensures existence

[3]We say that a function $\mathcal{E} : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is:

and uniqueness of a minimizer $\hat{f}^\tau$, for any $\nu > 0$.

The rest of this section is divided into two parts. First we show how the theory of RKHS (Aronszajn, 1950) allows to compute derivatives of functions on high dimensional spaces and also to derive a new representer theorem that allows to deal with finite dimensional minimization problems. Second we discuss how to apply proximal methods Combettes and Wajs (2005); Beck and Teboulle (2009) to derive an iterative optimization procedure of which we can prove convergence. The main outcome of our analysis is that the solution of Problem (6) can be written as

$$\hat{f}^\tau(x) = \sum_{i=1}^{n} \frac{1}{n}\alpha_i k_{x_i}(x) + \sum_{i=1}^{n}\sum_{a=1}^{d} \frac{1}{n}\beta_{a,i}(\partial_a k)_{x_i}(x), \tag{7}$$

where $\alpha, (\beta_{a,i})_{i=1}^n \in \mathbb{R}^n$ for all $a = 1, \ldots, d$ and $(\partial_a k)_x$ denotes partial derivatives of the kernel, see (19). The coefficients α and β can be computed through an iterative procedure. To describe the algorithm we need some notation. For all $a, b = 1, \ldots, d$, we define the $n \times n$ matrices $\mathrm{K}, \mathrm{Z}_a, \mathrm{L}_{a,b}$ as

$$\mathrm{K}_{i,j} = \frac{1}{n}k(x_i, x_j), \tag{8}$$

$$[\mathrm{Z}_a]_{i,j} = \frac{1}{n}\left.\frac{\partial k(s, x_j)}{\partial s^a}\right|_{s=x_i}, \tag{9}$$

and

$$[\mathrm{L}_{a,b}]_{i,j} = \frac{1}{n}\left.\frac{\partial^2 k(x, s)}{\partial x^a \partial s^b}\right|_{x=x_i, s=x_j}$$

for all $i, j = 1, \ldots, n$. Clearly the above quantities can be easily computed as soon as we have an explicit expression of the kernel see Example 1 in Appendix A. We introduce also the $n \times nd$ matrices

$$\mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_d)$$
$$\mathrm{L}_a = (\mathrm{L}_{a,1}, \ldots, \mathrm{L}_{a,d}) \qquad \forall a = 1, \ldots, d \tag{10}$$

and the $nd \times nd$ matrix

$$\mathrm{L} = \begin{pmatrix} \mathrm{L}_{1,1} & \ldots & \mathrm{L}_{1,d} \\ \ldots & \ldots & \ldots \\ \mathrm{L}_{d,1} & \ldots & \mathrm{L}_{d,d} \end{pmatrix} = \begin{pmatrix} \mathrm{L}_a \\ \ldots \\ \mathrm{L}_d \end{pmatrix}$$

Denote with B_n the unitary ball in $\mathbb{R}^n$

$$B_n = \{v \in \mathbb{R}^n \,|\|v\|_n \le 1\}. \tag{11}$$

The coefficients in (7) are obtained through Algorithm 1, where β is considered as the nd column vector $\beta = (\beta_{1,1}, \ldots, \beta_{1,n}, \ldots, \beta_{d,1}, \ldots, \beta_{d,n})^T$.

The proposed optimization algorithm consists of two nested iterations, and involves only matrix multiplications and thresholding operations. Before describing its derivation and discussing its convergence properties, we add several remarks. First, the proposed procedure requires the choice of an appropriate stopping rule, which will be discussed later, and of the step sizes σ and η. The simple a priori choice $\sigma = \|\mathrm{K}\| + \tau\nu$, $\eta = \|\mathrm{L}\|$ ensures convergence, as discussed in the Subsection 3.3, and is the one used in our experiments. Second, the computation of the solution for different regularization parameters can be highly accelerated by a simple warm starting procedure, as the one in Hale et al. (2008). Finally, a critical issue is the identification the variables selected by the algorithm. in Subsection 3.4 we discuss a principled way to select variable using on the norm of the coefficients $(\bar{v}_a^t)_{a=1}^d$.

- *coercive* if $\lim_{\|f\|\to+\infty} \mathcal{E}(f)/\|f\| = +\infty$;
- *strongly convex of modulus* μ if $\mathcal{E}(tf + (1-t)g) \le t\mathcal{E}(f) + (1-t)\mathcal{E}(g) - \frac{\mu}{2}t(1-t)\|f-g\|^2$ for all $t \in [0, 1]$.

Algorithm 1

Given: parameters $\tau, \nu > 0$ and step-sizes $\sigma, \eta > 0$
Initialize: $\alpha^0 = \tilde{\alpha}^1 = 0$, $\beta^0 = \tilde{\beta}^1 = 0$, $s_1 = 1$, $\bar{v}^0 = 0$, $t = 1$
while `convergence not reached` **do**
$t = t + 1$

$$s_t = \frac{1}{2}\left(1 + \sqrt{1 + 4s_{t-1}^2}\right) \tag{12}$$

$$\tilde{\alpha}^t = \left(1 + \frac{s_{t-1} - 1}{s_t}\right)\alpha^{t-1} + \frac{1 - s_{t-1}}{s_t}\alpha^{t-2}, \quad \tilde{\beta}^t = \left(1 + \frac{s_{t-1} - 1}{s_t}\right)\beta^{t-1} + \frac{1 - s_{t-1}}{s_t}\beta^{t-2}, \tag{13}$$

$$\alpha^t = \left(1 - \frac{\tau\nu}{\sigma}\right)\tilde{\alpha}^t - \frac{1}{\sigma}\left(\mathrm{K}\tilde{\alpha}^t + \mathrm{Z}\tilde{\beta}^t - \mathbf{y}\right) \tag{14}$$

set $v^0 = \bar{v}^{t-1}$, $q = 0$
while `convergence not reached` **do**
$q = q + 1$
for $a = 1, \ldots d$ **do**

$$v_a^q = \pi_{\frac{\tau}{\sigma}B_n}\left(v_a^{q-1} - \frac{1}{\eta}\left(\mathrm{L}_a v^{q-1} - \left(\mathrm{Z}_a^T\alpha^t + \left(1 - \frac{\tau\nu}{\sigma}\right)\mathrm{L}_a\tilde{\beta}^t\right)\right)\right) \tag{15}$$

end for
end while
set $\bar{v}^t = v^q$

$$\beta^t = \left(1 - \frac{\tau\nu}{\sigma}\right)\tilde{\beta}^t - \bar{v}^t. \tag{16}$$

end while
return (α^t, β^t)

3.1 Kernels, Partial Derivatives and Regularization

We start discussing how (partial) derivatives can be efficiently computed in RKHSs induced by smooth kernels and hence derive a new representer theorem. Practical computation of the derivatives for a differentiable functions is often performed via finite differences. For functions defined on a high dimensional space such a procedure becomes cumbersome and ultimately not-efficient. RKHSs provide an alternative computational tool.

Recall that the RKHS associated to a symmetric positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the unique Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ such that $k_x = k(x, \cdot) \in \mathcal{H}$, for all $x \in X$ and

$$f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \tag{17}$$

for all $f \in \mathcal{H}, x \in X$. Property (17) is called *reproducing property* and k is called reproducing kernel (Aronszajn, 1950). We recall a few basic facts. The functions in $\mathcal{H}$ can be written as pointwise limits of finite linear combinations of the type $\sum_{i=1}^p \alpha_i k_{x_i}$, where $\alpha_i \in \mathbb{R}, x_i \in X$ for all i. One of the most important results for kernel methods, namely the representer theorem (Wahba, 1990), shows that a large class of regularized kernel methods induce estimators that can be written as *finite* linear combinations of kernels centered at the training set points. This result allows to consider potentially infinite dimensional spaces. In the following we will make use of the so called *sampling operator*, which returns the values of a function $f \in \mathcal{H}$ at a set of input points $\mathbf{x} = (x_1, \ldots, x_n)$

$$\hat{S} : \mathcal{H} \to \mathbb{R}^n, \qquad (\hat{S}f)_i = \langle f, k_{x_i} \rangle, \quad i = 1, \ldots, n. \tag{18}$$

The above operator is linear and bounded if the kernel is bounded– see Appendix A, so in the following we always make the following assumption.

Assumption (A1). *There exists* $\kappa_1 < \infty$ *such that* $\sup_{x \in X} \|k_x\|_{\mathcal{H}} < \kappa_1$.

Next we discuss how the theory of RKHS allows efficient derivative computations. Let

$$(\partial_a k)_x := \left.\frac{\partial k(s, \cdot)}{\partial s^a}\right|_{s=x} \tag{19}$$

be the partial derivative of the kernel with respect to the first variable. Then, from Theorem 1 in Zhou (2008) we have that, if k is at least a $\mathcal{C}^2(\mathcal{X} \times \mathcal{X})$, $(\partial_a k)_x$ belongs to $\mathcal{H}$ for all $x \in X$ and most importantly

$$\frac{\partial f(x)}{\partial x^a} = \langle f, (\partial_a k)_x \rangle_{\mathcal{H}},$$

for $a = 1, \dots, d$, $x \in X$. It is useful to define the analogous of the sampling operator for derivatives, which returns the values of the partial derivative of a function $f \in \mathcal{H}$ at a set of input points $\mathbf{x} = (x_1, \dots, x_n)$,

$$\hat{D}_a : \mathcal{H} \to \mathbb{R}^n, \qquad (\hat{D}_a f)_i = \langle f, (\partial_a k)_{x_i} \rangle, \tag{20}$$

where $a = 1, \dots, d$, $i = 1, \dots, n$. It is also useful to define an empirical gradient operator $\hat{\nabla} : \mathcal{H} \to (\mathbb{R}^n)^d$ defined by $\hat{\nabla} f = (\hat{D}_a f)_{a=1}^d$. The above operators are linear and bounded, if the following assumption is satisfied.

Assumption (A2). *The kernel k is $\mathcal{C}^2(\mathcal{X} \times \mathcal{X})$ and there exists $\kappa_2 < \infty$ such that $\sup_{x \in X} \|(\partial_a k)_x\|_{\mathcal{H}} < \kappa_2$, for all $a = 1, \dots, d$.*

We refer to Appendix A for further details and supplementary results.

Provided with the above results we can prove a suitable generalization of the representer theorem.

Proposition 1. *The minimizer of (6) can be written as*

$$\hat{f}^\tau = \sum_{i=1}^n \frac{1}{n} \alpha_i k_{x_i} + \sum_{i=1}^n \sum_{a=1}^d \frac{1}{n} \beta_{a,i} (\partial_a k)_{x_i} \tag{21}$$

with $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^{nd}$.

The above result is proved in Appendix A and shows that the regularized solution is determined by the set of $n + nd$ coefficients $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^{nd}$. We next discuss how such coefficients can be efficiently computed.

Notation. In the following, given an operator A we denote by A^* the corresponding adjoint operator. When A is a matrix we use the standard notation for the transpose $A^T = A^*$.

3.2 Computing the Solution with Proximal Methods

The functional $\widehat{\mathcal{E}}^\tau$ is not differentiable, hence its minimization cannot be done by simple gradient methods. Nonetheless it has a special structure that allows efficient computations using a forward-backward splitting algorithm (Combettes and Wajs, 2005), belonging to the class of the so called proximal methods.

Second order methods, see for example Chan et al. (1999), could also be used to solve similar problems. These methods typically converge quadratically and allows accurate computations. However, they usually have a high cost per iteration and hence are not suitable for large scale problems, as opposed to first order methods having much lower cost per iteration. Furthermore, in the seminal paper Nesterov (1983) first-order methods with optimal convergence rate are proposed (Nemirovski and Yudin, 1983). First order methods have since become a popular tool to solve non-smooth problems in machine learning as well as signal and image processing, see for example **FISTA** – Beck and Teboulle (2009) and references therein. These methods have proved to be fast and accurate (Becker et al., 2011), both for ℓ^1-based regularization – see Combettes and Wajs (2005), Daubechies et al. (2007), Figueiredo et al. (2007), Loris et al. (2009) – and more general regularized learning methods – see for example Duchi and Singer (2009), Mosci et al. (2010), Jenatton et al. (2010) –.

Forward-backward splitting algorithms The functional $\widehat{\mathcal{E}}^\tau$ is the sum of the two terms $F(\cdot) = \widehat{\mathcal{E}}(\cdot) + \tau\nu \|\cdot\|_{\mathcal{H}}^2$ and $2\tau\widehat{\Omega}_1^D$. The first term is strongly convex of modulus $\tau\nu$ and differentiable, while the second term is convex but not differentiable. The minimization of this class of functionals can be done iteratively using the forward-backward (FB) splitting algorithm,

$$f^t = \operatorname{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D}\left(\tilde{f}^t - \frac{1}{2\sigma}\nabla F(\tilde{f}^t)\right) \tag{22}$$

$$\tilde{f}^t = c_{1,t} f^{t-1} + c_{2,t} f^{t-2} \tag{23}$$

where $f^0 = f^1 \in \mathcal{H}$ is an arbitrary initialization, $c_{1,t}, c_{2,t}$ are suitably chosen positive sequences, and $\mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D} : \mathcal{H} \to \mathcal{H}$ is the proximity operator (Moreau, 1965) defined by,

$$\mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D}(f) = \operatorname*{argmin}_{g\in\mathcal{H}} \left(\frac{\tau}{\sigma}\widehat{\Omega}_1^D(g) + \frac{1}{2}\|f-g\|^2 \right).$$

The above approach decouples the contribution of the differentiable and not differentiable terms. Unlike other simpler penalties used in additive models, such as the ℓ^1 norm in the lasso, the computation of the proximity operator of $\widehat{\Omega}_1^D$ is not trivial and will be discussed in the next paragraph. Here we briefly recall the main properties of the iteration (22), (23) depending on the choice of $c_{1,t}, c_{2,t}$ and σ. The basic version of the algorithm (Combettes and Wajs, 2005), sometimes called ISTA (iterative shrinkage thresholding algorithm (Beck and Teboulle, 2009)), is obtained setting $c_{1,t} = 1$ and $c_{2,t} = 0$ for all $t > 0$, so that each step depends only on the previous iterate. The convergence of the algorithm for both the objective function values and the minimizers is extensively studied in Combettes and Wajs (2005), but a convergence rate is not provided. In Beck and Teboulle (2009) it is shown that the convergence of the objective function values is of order $O(1/t)$ provided that the step size σ satisfies $\sigma \geq L$, where L is the Lipschitz constant of $\nabla F/2$. An alternative choice of $c_{1,t}$ and $c_{2,t}$ leads to an accelerated version of the algorithm (22), sometimes called FISTA (fast iterative shrinkage thresholding algorithm (Tseng, 2010; Beck and Teboulle, 2009)), which is obtained by setting $s_0 = 1$,

$$s_t = \frac{1}{2}\left(1+\sqrt{1+4s_{t-1}^2}\right), \qquad c_{1,t} = 1 + \frac{s_{t-1}-1}{s_t}, \text{ and } c_{2,t} = \frac{1-s_{t-1}}{s_t}. \tag{24}$$

The algorithm is analyzed in (Beck and Teboulle, 2009) and in (Tseng, 2010) where it is proved that the objective values generated by such a procedure have convergence of order $O(1/t^2)$, if the step size satisfies $\sigma \geq L$.

Computing L can be non trivial. Theorems 3.1 and 4.4 in Beck and Teboulle (2009) show that the iterative procedure (22) with an adaptive choice for the step size, called *backtracking*, which does not require the computation of L, shares the same rate of convergence of the corresponding procedure with fixed step-size. Finally, it is well known that, if the functional is strongly convex with a positive modulus, the convergence rate of both the basic and accelerated scheme is indeed linear for both the function values and the minimizers (Nesterov, 1983; Mosci et al., 2010; Nesterov, 2007).

In our setting we use FISTA to tackle the minimization of $\widehat{\mathcal{E}}^\tau$ but, as we mentioned before, we have to deal with the computation of the proximity operator associated to $\widehat{\Omega}_1^D$.

Computing the proximity operator. Since $\widehat{\Omega}_1^D$ is one-homogeneus, i.e. $\widehat{\Omega}_1^D(\lambda f) = \lambda\widehat{\Omega}_1^D(f)$ for $\lambda > 0$, the Moreau identity, see (Combettes and Wajs, 2005), gives a useful alternative formulation for the proximity operator, that is

$$\mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D} = I - \pi_{\frac{\tau}{\sigma}\mathcal{C}_n}, \tag{25}$$

where $\mathcal{C}_n = (\partial\widehat{\Omega}_1^D)(0)$ is the subdifferential [4] of $\widehat{\Omega}_1^D$ at the origin, and $\pi_{\frac{\tau}{\sigma}\mathcal{C}_n} : \mathcal{H} \to \mathcal{H}$ is the projection on $\frac{\tau}{\sigma}\mathcal{C}_n$– which is well defined since $\mathcal{C}_n$ is a closed convex subset of $\mathcal{H}$. To describe how to practically compute such a projection, we start observing that the DENOVAS penalty $\widehat{\Omega}_1^D$ is the sum of d norms in $\mathbb{R}^n$. Then following Section 3.2 in Mosci et al. (2010) (see also Ekeland and Temam (1976)) we have

$$\mathcal{C}_n = \partial\widehat{\Omega}_1^D(0) = \left\{ f \in \mathcal{H} \mid f = \hat{\nabla}^* v \text{ with } v \in B_n^d \right\},$$

where B_n^d is the cartesian product of d unitary balls in $\mathbb{R}^n$,

$$B_n^d = \underbrace{B_n \times \cdots \times B_n}_{d \text{ times}} = \{v = (v_1, \ldots, v_d) \,|v_a \in \mathbb{R}^n, \ \|v_a\|_n \leq 1, \ a = 1, \ldots, d\},$$

[4] Recall that the subdifferential of a convex functional $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is denoted with $\partial\Omega(f)$ and is defined as the set

$$\partial\Omega(f) := \{h \in \mathcal{H} \,:\, \Omega(g) - \Omega(f) \geq \langle h, g-f\rangle_{\mathcal{H}}, \ \forall g \in \mathcal{H}\}.$$

with B_n defined in (11). Then, by definition, the projection is given by

$$\pi_{\frac{\tau}{\sigma}\mathcal{C}_n}(f) = \hat{\nabla}^* \bar{v},$$

where

$$\bar{v} \in \operatorname*{argmin}_{v \in \frac{\tau}{\sigma} B_n^d} \|f - \hat{\nabla}^* v\|_{\mathcal{H}}^2. \tag{26}$$

Being a convex constrained problem, (26) can be seen as the sum of the smooth term $\|f - \hat{\nabla}^* v\|_{\mathcal{H}}^2$ and the indicator function of the convex set B_n^d. We can therefore use (22), again. In fact we can fix an arbitrary initialization $v^0 \in \mathbb{R}^{nd}$ and consider,

$$v^{q+1} = \pi_{\frac{\tau}{\sigma} B_n^d}\left(v^q - \frac{1}{\eta}\hat{\nabla}(\hat{\nabla}^* v^q - f)\right), \tag{27}$$

for a suitable choice of η. In particular, we notice that $\pi_{\frac{\tau}{\sigma} B_n^d}$ can be easily computed in closed form, and corresponds to the proximity operator associated to the indicator function of B_n^d. Applying the results mentioned above, if $\eta \geq \|\hat{\nabla}\hat{\nabla}^*\|$, convergence of the function values of problem (26) on the sequence generated via (27) is guaranteed. However, since we are interested in the computation of the proximity operator, this is not enough. Thanks to the special structure of the minimization problem in (26), it is possible to prove (see (Combettes et al., 2010; Mosci et al., 2010)) that

$$\|\hat{\nabla}^* v^q - \hat{\nabla}^* \bar{v}\|_{\mathcal{H}} \to 0, \quad \text{or, equivalently} \quad \|\hat{\nabla}^* v^q - \pi_{\frac{\tau}{\sigma}\mathcal{C}_n}(f)\|_{\mathcal{H}} \to 0. \tag{28}$$

A similar first-order method to compute convergent approximations of $\hat{\nabla}^* \bar{v}$ has been proposed in (Bect et al., 2004).

3.3 Overall Procedure and Convergence analysis

To compute the minimizer of $\widehat{\mathcal{E}}^\tau$ we consider the combination of the accelerated FB-splitting algorithm (outer iteration) and the basic FB-splitting algorithm for computing the proximity operator (inner iteration). The overall procedure is given by

$$\begin{aligned}
s_t &= \frac{1}{2}\left(1 + \sqrt{1 + 4s_{t-1}^2}\right) \\
\tilde{f}^t &= \left(1 + \frac{s_{t-1} - 1}{s_t}\right) f^{t-1} + \frac{1 - s_{t-1}}{s_t} f^{t-2} \\
f^t &= \left(1 - \frac{\tau\nu}{\sigma}\right)\tilde{f}^t - \frac{1}{\sigma}\hat{S}^*\left(\hat{S}\tilde{f}^t - \mathbf{y}\right) - \hat{\nabla}^* \bar{v}^t,
\end{aligned} \tag{29}$$

for $t = 2, 3, \ldots$, where $\bar{v}^t$ is computed through the iteration

$$v^q = \pi_{\frac{\tau}{\sigma} B_n^d}\left(v^{q-1} - \frac{1}{\eta}\hat{\nabla}\left(\hat{\nabla}^* v^{q-1} - \left(1 - \frac{\tau\nu}{\sigma}\right)\tilde{f}^t - \frac{1}{\sigma}\hat{S}^*\left(\hat{S}\tilde{f}^t - \mathbf{y}\right)\right)\right), \tag{30}$$

for given initializations.

The above algorithm is an *inexact* accelerated FB-splitting algorithm, in the sense that the proximal or backward step is computed only approximately. The above discussion on the convergence of FB-splitting algorithms was limited to the case where computation of the proximity operator is done exactly (we refer to this case as the *exact* case). The convergence of the inexact FB-splitting algorithm does not follow from this analysis. For the basic – not accelerated – FB-splitting algorithm, convergence in the inexact case is still guaranteed (without a rate) (Combettes and Wajs, 2005), if the computation of the proximity operator is sufficiently accurate. The convergence of the inexact accelerated FB-splitting algorithm is studied in Villa et al. (2011) where it is shown that the same convergence rate of the exact case can be achieved, again provided that the accuracy in the computation of the proximity operator can be suitably controlled. Such a result can be adapted to our setting to prove the following theorem, as shown in Appendix B.

Theorem 1. *Let $\varepsilon^t \sim t^{-l}$ with $l > 3/2$, $\sigma \geq \|\hat{S}^*\hat{S}\| + \tau\nu$, $\eta \geq \|\hat{\nabla}\hat{\nabla}^*\|$, and f^t given by* (29) *with $\bar{v}^t$ computed through* (30)*. Define $g^t = \left(1 - \frac{\tau\nu}{\sigma}\right)\tilde{f}^t - \frac{1}{\sigma}\hat{S}^*\left(\hat{S}\tilde{f}^t - \mathbf{y}\right)$. If $\bar{v}^t = v^q$, for q such that the following condition is satisfied*

$$\|\hat{\nabla}^* v^q - \pi_{\frac{\tau}{\sigma}C_n}(g^t)\|_{\mathcal{H}} \leq \frac{(\varepsilon^t)^2}{2dist\big(g^t, \frac{\tau}{\sigma}\mathcal{C}_n\big) + 4\dfrac{d\tau}{\sigma}\|\hat{\nabla}^*\|}, \tag{31}$$

where $dist\big(g^t, \frac{\tau}{\sigma}\mathcal{C}_n\big)$ denotes the distance in $\mathcal{H}$ of g^t from the set $\frac{\tau}{\sigma}\mathcal{C}_n$, then there exists a constant $C > 0$ such that

$$\widehat{\mathcal{E}}^\tau(f^t) - \widehat{\mathcal{E}}^\tau(\hat{f}^\tau) \leq \frac{C}{t^2},$$

and thus, if $\nu > 0$,

$$\|f^t - \hat{f}^\tau\|_{\mathcal{H}} \leq \frac{2}{t}\sqrt{\frac{C}{\nu\tau}}. \tag{32}$$

As for the exact accelerated FB-splitting algorithm, the step size of the outer iteration has to be greater than or equal to $L = \|\hat{S}^*\hat{S}\| + \tau\nu$. In particular, we choose $\sigma = \|\hat{S}^*\hat{S}\| + \tau\nu$ and, similarly, $\eta = \|\hat{\nabla}\hat{\nabla}^*\|$.

We add few remarks. First, as it is evident from (32), the choice of $\nu > 0$ allows to obtain convergence of f^t to $\hat{f}^\tau$ with respect to the norm in $\mathcal{H}$, and positively influences the rate of convergence. This is a crucial property in variable selection, where it is necessary to accurately estimate the minimizer of the expected risk $f_\rho^\dagger$ and not only its minimum $\mathcal{E}(f_\rho^\dagger)$. Second, even if condition (31) requires the knowledge of the asymptotic solution $\pi_{\frac{\tau}{\sigma}C_n}(f^t)$, the same condition is always satisfied if the number of inner iterations is large enough. The discussion of an implementable stopping rule is given in Section 3.4. Third, we remark that for proving convergence of the inexact procedure, it is essential that the specific algorithm proposed to compute the proximal step generates a sequence belonging to $\mathcal{C}_n$ and satisfying (28). This is not true for the accelerated version (FISTA), as it does not satisfy (28).

3.4 Further Algorithmic Considerations

We conclude discussing several practical aspects of the proposed method.

The finite dimensional implementation. We start by showing how the representer theorem can be used, together with the iterations described by (29) and (30), to derive Algorithm 1. This is summarized in the following proposition.

Proposition 2. *For $\nu > 0$ and $f^0 = \frac{1}{n}\sum_i \alpha_i^0 k_{x_i} + \frac{1}{n}\sum_i\sum_a \beta_{a,i}^0(\partial_a k)_{x_i}$ for any $\alpha^0 \in \mathbb{R}^n, \beta^0 \in \mathbb{R}^{nd}$, the solution at step t for the updating rule* (29) *is given by*

$$f^t = \frac{1}{n}\sum_{i=1}^n \alpha_i^t k_{x_i} + \frac{1}{n}\sum_{i=1}^n\sum_{a=1}^d \beta_{a,i}^t(\partial_a k)_{x_i} \tag{33}$$

with α^t and β^t defined by the updating rules (14-13), where $\bar{v}^t$ in (16) *can be estimated, starting from any $v^0 \in \mathbb{R}^{nd}$, and using the iterative rule* (15)*.*

The proof of the above proposition can be found in Appendix B, and is based on the observation that $\mathrm{K}, \mathrm{Z}_a, \mathrm{Z}, \mathrm{L}_a$ defined at the beginning of this Section are the matrices associated to the operators $\hat{S}\hat{S}^* : \mathbb{R}^n \to \mathbb{R}^n$, $\hat{S}\hat{D}_a^* : \mathbb{R}^n \to \mathbb{R}^n$, $\hat{S}\hat{\nabla}^* : \mathbb{R}^{nd} \to \mathbb{R}^n$ and $\hat{D}_a\hat{\nabla}^* : \mathbb{R}^{nd} \to \mathbb{R}^n$, respectively.

Using the same reasoning we can make the following two further observations. First, one can compute the step sizes σ and η as $\sigma = \|\mathrm{K}\| + \tau\nu$, and $\eta = \|\mathrm{L}\|$. Second, since in practice we have to define suitable stopping rules, Equations (32) and (28) suggest the following choices [5]

$$\|f^t - f^{t-1}\|_{\mathcal{H}} \leq \varepsilon^{(\text{ext})} \quad \text{and} \quad \|\hat{\nabla}^*(v^q - v^{q-1})\|_{\mathcal{H}} \leq \varepsilon^{(\text{int})}.$$

[5] In practice we often use a stopping rule where the tolerance is scaled with the current iterate, $\|f^t - f^{t-1}\|_{\mathcal{H}} \leq \varepsilon^{(\text{ext})}\|f^t\|_{\mathcal{H}}$ and $\|\hat{\nabla}^*(v^q - v^{q-1})\|_{\mathcal{H}} \leq \varepsilon^{(\text{int})}\|\hat{\nabla}^* v^q\|_{\mathcal{H}}$.

As a direct consequence of (33) and using the definition of matrices $\mathrm{K}, \mathrm{Z}, \mathrm{L}$, these quantities can be easily computed as

$$\begin{aligned} \|f^t - f^{t-1}\|_{\mathcal{H}}^2 &= \langle \delta\alpha, \mathrm{K}\delta\alpha\rangle_n + 2\langle \delta\alpha, \mathrm{Z}\delta\beta\rangle_n + \langle \delta\beta, \mathrm{L}\delta\beta\rangle_n, \\ \|\hat{\nabla}^*(v^q - v^{q-1})\|_{\mathcal{H}}^2 &= \langle (v^q - v^{q-1}), \mathrm{L}(v^q - v^{q-1})\rangle_n. \end{aligned}$$

where we defined $\delta\alpha = \alpha^t - \alpha^{t-1}$ and $\delta\beta = \beta^t - \beta^{t-1}$. Also note that, according to Theorem 1, $\varepsilon^{(\mathrm{int})}$ must depend on the outer iteration as $\varepsilon^{(\mathrm{int})} = \varepsilon^t \sim t^{-2l}$, $l > 3/2$.

Finally we discuss a criterion for identifying the variables selected by the algorithm.

Selection. Note that in the linear case $f(x) = \beta \cdot x$ the coefficients $\beta^1, \ldots, \beta^d$ coincide with the partial derivatives, and the coefficient vector β given by ℓ^1 regularization is sparse (in the sense that it has zero entries), so that it is easy to detect which variables are to be considered relevant. For a general non-linear function, we then expect the vector $(\|\hat{D}_a f\|_n^2)_{a=1}^d$ of the norms of the partial derivatives evaluated on the training set points, to be sparse as well. In practice since the projection $\pi_{\tau/\sigma B_n^d}$ is computed only approximately, the norms of the partial derivatives will be small but typically not zero. The following proposition elaborates on this point.

Proposition 3. *Let $v = (v_a)_{a=1}^d \in B_n^d$ such that, for any $\sigma > 0$*

$$\hat{\nabla}^* v = -\frac{1}{\sigma}\nabla(\widehat{\mathcal{E}}(\hat{f}^\tau) + \tau\nu\|\hat{f}^\tau\|_{\mathcal{H}}^2),$$

then

$$\|v_a\|_n < \frac{\tau}{\sigma} \Rightarrow \|\hat{D}_a \hat{f}^\tau\|_n = 0. \tag{34}$$

Moreover, if $\bar{v}^t$ is given by Algorithm 1 with the inner iteration stopped when the assumptions of Theorem 1 are met, then there exists $\tilde{\varepsilon}^t > 0$ (precisely defined in (39)*) depending on the tolerance ε^t used in the inner iteration and satisfying $\lim_{t\to 0} \tilde{\varepsilon}^t = 0$, such that if $m := \min\{\|\hat{D}_a \hat{f}^\tau\|_n \,:\, a \in \{1, \ldots, d\} \text{s.t.} \|\hat{D}_a \hat{f}^\tau\|_n > 0\}$.*

$$\|\bar{v}_a^t\|_n \geq \frac{\tau}{\sigma} - \frac{(\tilde{\varepsilon}^t)^2}{2m} \Rightarrow \|\hat{D}_a \hat{f}^\tau\|_n = 0. \tag{35}$$

The above result, whose proof can be found in Appendix B, is a direct consequence of the Euler equation for $\widehat{\mathcal{E}}^\tau$ and of the characterization of the subdifferential of $\widehat{\Omega}_1^D$. The second part of the proof follows by observing that, as $\hat{\nabla}^* v$ belongs to the subdifferential of $\widehat{\Omega}_1^D$ at $\hat{f}^\tau$,$\hat{\nabla}^*\bar{v}^t$ belongs to the *approximate* subdifferential of $\widehat{\Omega}_1^D$ at $\hat{f}^\tau$, where the approximation of the subdifferential is controlled by the precision used in evaluating the projection. Given the pair $(f^t, \bar{v}^t)$ evaluated via Algorithm 1, we can thus consider to be irrelevant the variables such that $\|\bar{v}_a^t\|_n < \tau/\sigma - (\tilde{\varepsilon}^t)^2/(2m)$. Note that the explicit form of $\tilde{\varepsilon}^t$ is given in (39)).

4 Consistency for Learning and Variable Selection

In this section we study the consistency properties of our method. Throughout this section we make the following assumption which is quite standard in machine learning and trivially verified in classification.

Assumption (A3). *There exists $M < \infty$ such that $\mathcal{Y} \subseteq [-M, M]$.*

4.1 Consistency

As we discussed in Section 2.2, though in practice we consider the regularizer $\widehat{\Omega}_1^D$ defined in (3), ideally we would be interested into $\Omega_1^{\mathrm{D}}(f) = \sum_{a=1}^d \|D_a f\|_{\rho_{\mathcal{X}}}$, $f \in \mathcal{H}$. The following preliminary result shows that indeed $\widehat{\Omega}_1^D$ is a consistent estimator of Ω_1^{D} when considering functions in $\mathcal{H}$ having uniformly bounded norm.

Theorem 2. *Let $r < \infty$, then under assumption (A2)*

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{\|f\|_{\mathcal{H}} \leq r} |\widehat{\Omega}_1^D(f) - \Omega_1^D(f)| > \epsilon\right) = 0 \qquad \forall \epsilon > 0.$$

The restriction to functions such that $\|f\|_{\mathcal{H}} \leq r$ is natural and is required since the penalty $\widehat{\Omega}_1^D$ forces the partial derivatives to be zero only on the training set points. To guarantee that a partial derivative, which is zero on the training set, is also close to zero on the rest of the input space, we must control the smoothness of the function class where the derivatives are computed. This motivates constraining the function class by adding the (squared) norm in $\mathcal{H}$ into (4). This is in the same spirit of the manifold regularization proposed in Belkin and Niyogi (2008).

The above result on the consistency of the derivative based regularizer is at the basis of the following consistency result.

Theorem 3. *Under assumptions A1, A2 and A3, recalling that* $\mathcal{E}(f) = \int (y - f(x))^2 \, d\rho(x, y)$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mathcal{E}(\hat{f}^{\tau_n}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \geq \epsilon \right) = 0 \qquad \forall \epsilon > 0,$$

for any τ_n *satisfying*

$$\tau_n \to 0 \qquad (\sqrt{n}\tau_n)^{-1} \to 0.$$

The proof is given in the appendix and is based on a sample/approximation error decomposition

$$\mathcal{E}(\hat{f}^\tau) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq \underbrace{|\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau)|}_{\text{sample error}} + \underbrace{|\mathcal{E}^\tau(f^\tau) - \inf_{f\in\mathcal{H}} \mathcal{E}(f)|}_{\text{approximation error}},$$

where

$$\mathcal{E}^\tau(f) := \mathcal{E}(f) + 2\tau\Omega_1^{\mathrm{D}}(f) + \tau\nu\|f\|_{\mathcal{H}}^2, \qquad f^\tau := \operatorname*{argmin}_{\mathcal{H}} \mathcal{E}^\tau.$$

The control of both terms allows to find a suitable parameter choice which gives consistency. When estimating the sample error one has typically to control only the deviation of the empirical risk from its continuos counterpart. Here we need Theorem 2 to also control the deviation of $\widehat{\Omega}_1^D$ from Ω_1^{D}. Note that, if the kernel is universal (Steinwart and Christmann, 2008), then $\inf_{f\in\mathcal{H}} \mathcal{E}(f) = \mathcal{E}(f_\rho)$ and Theorem 3 gives the universal consistency of the estimator $\hat{f}^{\tau_n}$.

To study the selection properties of the estimator $\hat{f}^{\tau_n}$– see next section– it useful to study the distance of $\hat{f}^{\tau_n}$ to f_ρ in the $\mathcal{H}$-norm. Since in general f_ρ might not belong to $\mathcal{H}$, for the sake of generality here we compare $\hat{f}^{\tau_n}$ to a minimizer of $\inf_{f\in\mathcal{H}} \mathcal{E}(f)$ which we always assume to exist. Since the minimizers might be more then one we further consider a suitable minimal norm minimizer $f_\rho^\dagger$– see below. More precisely given the set

$$\mathcal{F}_{\mathcal{H}} := \{ f \in \mathcal{H} \mid \mathcal{E}(f) = \inf_{f\in\mathcal{H}} \mathcal{E}(f) \}$$

(which we assume to be not empty), we define

$$f_\rho^\dagger := \operatorname*{argmin}_{f\in\mathcal{F}_{\mathcal{H}}} \{ \Omega_1^{\mathrm{D}}(f) + \nu\|f\|_{\mathcal{H}}^2 \}.$$

Note that $f_\rho^\dagger$ is well defined and unique, since $\Omega_1^{\mathrm{D}}(\cdot) + \nu\|\cdot\|_{\mathcal{H}}^2$ is strongly convex and $\mathcal{E}$ is convex and lower semi-continuous on $\mathcal{H}$, which implies that $\mathcal{F}_{\mathcal{H}}$ is closed and convex in $\mathcal{H}$. Then, we have the following result.

Theorem 4. *Under assumptions A1, A2 and A3, we have*

$$\lim_{n\to\infty} \mathrm{P}\left(\|\hat{f}^{\tau_n} - f_\rho^\dagger\|_{\mathcal{H}} \geq \epsilon \right) = 0, \qquad \forall \epsilon > 0,$$

for any τ_n *such that* $\tau_n \to 0$ *and* $(\sqrt{n}\tau_n^2)^{-1} \to 0$.

The proof, given in Appendix C, is based on the decomposition in sample error, $\|\hat{f}^\tau - f^\tau\|_{\mathcal{H}}$, and approximation error, $\|f^\tau - f_\rho^\dagger\|_{\mathcal{H}}$. To bound the sample error we use recent results (Villa et al., 0) that exploit Attouch-Wets convergence (Attouch and Wets, 1991, 1993a,b) and coercivity of the penalty (ensured by the RKHS norm) to control the distance between the minimizers $\hat{f}^\tau, f^\tau$ by the distance the minima $\widehat{\mathcal{E}}^\tau(\hat{f}^\tau)$ and $\mathcal{E}^\tau(f^\tau)$. Convergence of the approximation error is again guaranteed by standard results in regularization theory (Dontchev and Zolezzi, 1993).

4.2 Selection properties

We next consider the selection properties of our method. Following Equation (2), we start by giving the definition of relevant/irrelevant variables and sparsity in our context.

Definition 1. *We say that a variable $a = 1, \ldots, d$ is* irrelevant *with respect to ρ for a differentiable function f, if the corresponding partial derivative $D_a f$ is zero $\rho_{\mathcal{X}}$-almost everywhere, and relevant otherwise. In other words the set of relevant variables is*

$$R_f := \{a \in \{1, \ldots, d\} \mid \|D_a f\|_{\rho_{\mathcal{X}}} > 0\}.$$

We say that a differentiable function f is sparse if $\Omega_0^D(f) := |R_f| < d$.

The goal of variable selection is to correctly estimate the set of relevant variables, $R_\rho := R_{f_\rho^\dagger}$. In the following we study how this can be achieved by the empirical set of relevant variables, $\hat{R}^{\tau_n}$, defined as

$$\hat{R}^{\tau_n} := \{a \in \{1, \ldots, d\} | \|\hat{D}_a \hat{f}^{\tau_n}\|_n > 0\}.$$

Theorem 5. *Under assumptions A1, A2 and A3*

$$\lim_{n \to \infty} \mathrm{P}\Big(R_\rho \subseteq \hat{R}^{\tau_n} \Big) = 1$$

for any τ_n satisfying

$$\tau_n \to 0 \qquad (\sqrt{n}\tau_n^2)^{-1} \to 0.$$

The above result shows that the proposed regularization scheme is a safe filter for variable selection, since it does not discard relevant variables, in fact, for a sufficiently large number of training samples, the set of truly relevant variables, R_ρ, is contained with high probability in the set of relevant variables identified by the algorithm, $\hat{R}^{\tau_n}$. The proof of the converse inclusion, giving consistency for variable selection, requires further analysis that we postpone to a future work.

5 Empirical Analysis

The content of this section is divided into two parts. First we study the properties of the proposed method on simulated data under different parameter settings. Second we compare our method to related regularization methods for learning and variable selection.

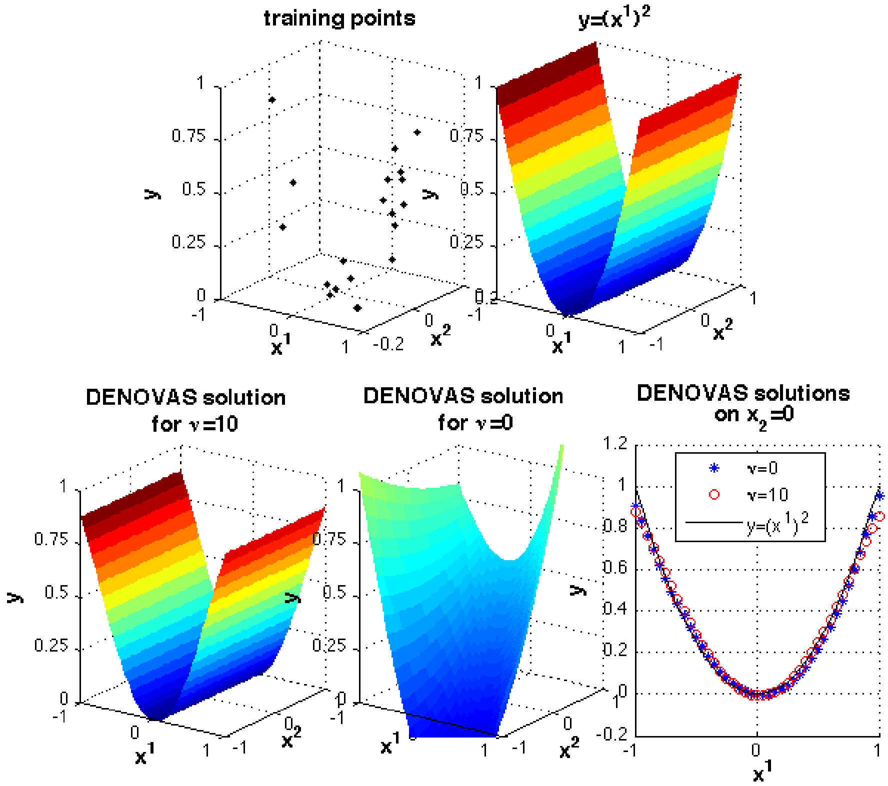
When we refer to our method we always consider a two-step procedure based on variable selection via Algorithm 1 and regression on the selected variable via (kernel) Regularized Least Squares (RLS). The kernel used in both steps is the same. Where possible, we applied the same reweighting procedure to the methods we compared with.

5.1 Analysis of Our Method

5.1.1 Role of the smoothness enforcing penalty $\tau\nu\|\cdot\|_{\mathcal{H}}^2$

From a theoretical stand point we have shown that ν has to be nonzero, in order for the proposed regularization problem (4) to be well-posed. We also mentioned that the combination of the two penalties $\widehat{\Omega}_1^D$ and $\|\cdot\|_{\mathcal{H}}^2$ ensures that the regularized solution will not depend on variables that are irrelevant for two different reasons. The first is irrelevance with respect to the output. The second type of irrelevance is meant in an unsupervised sense. This happens when one or more variables are (approximately) constant with respect to the marginal distribution $\rho_{\mathcal{X}}$, so that the support of the marginal distribution is (approximately) contained in a coordinate subspace. Here we present two experiments aimed at empirically assessing the role of the smoothness enforcing penalty $\|\cdot\|_{\mathcal{H}}^2$ and of the parameter ν. We first present an experiment where the support of the marginal distribution approximately coincides with a coordinate subspace $x^2 = 0$. Then we systematically investigate the stabilizing effect of the smoothness enforcing penalty also when the marginal distribution is not degenerate.

Figure 3: Effect of the combined regularization $\widehat{\Omega}_1^D(\cdot) + \nu\|\cdot\|_{\mathcal{H}}^2$ on a toy problem in $\mathbb{R}^2$ where the support of marginal distribution approximately coincides with the coordinate subspace $x^2 = 0$. The output labels are drawn from $y = (x^1)^2 + w$, with $w \sim \mathcal{N}(0, 0.1)$.

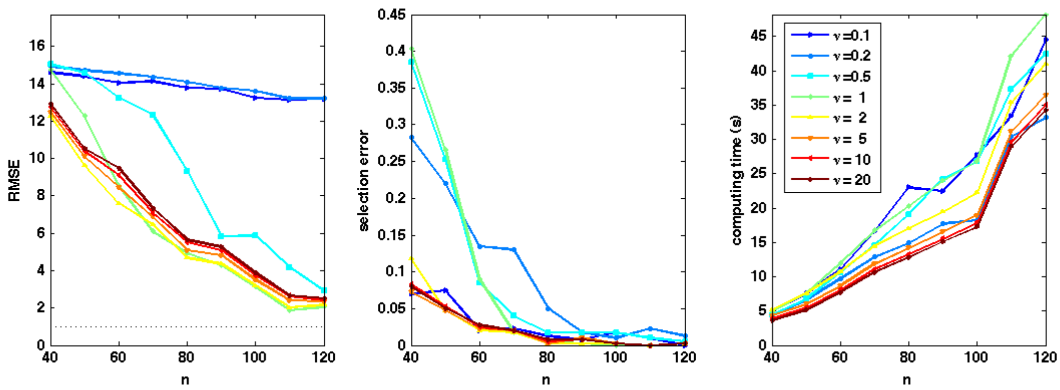


Adaption to the Marginal Distribution We consider a toy problem in 2 dimensions, where the support of the marginal distribution $\rho_{\mathcal{X}}(x^1, x^2)$ approximately coincides with the coordinate subspace $x^2 = 0$. Precisely x^1 is uniformly sampled from $[-1, 1]$, whereas x^2 is drawn from a normal distribution $x^2 \sim \mathcal{N}(0, 0.05)$. The output labels are drawn from $y = (x^1)^2 + w$, where w is a white noise, sampled from a normal distribution with zero mean and variance 0.1. Given a training set of $n = 20$ samples i.i.d. drawn from the above distribution (Figure 3 top-left), we evaluate the optimal value of the regularization parameter τ via hold out validation on an independent validation set of $n_{\text{val}} = n = 20$ samples. We repeat the process for $\nu = 0$ and $\nu = 10$. In both cases the reconstruction accuracy on the support of $\rho_{\mathcal{X}}$ is high, see Figure 3 bottom-right . However, while $\nu = 10$ our method correctly selects the only relevant variable x^1 (Figure 3 bottom-left), when $\nu = 0$ both variables are selected (Figure 3 bottom-center), since functional $\widehat{\mathcal{E}}^{\tau,0}$ is insensible to errors out of supp$(\rho_{\mathcal{X}})$, and the regularization term $\tau\widehat{\Omega}_1^D$ alone does not penalizes variations out of supp$(\rho_{\mathcal{X}})$.

Effect of varying ν Here we empirically investigate the stabilizing effect of the smoothness enforcing penalty when the marginal distribution is not degenerate. The input variables $x = (x^1, \ldots, x^{20})$ are uniformly drawn from $[-1, 1]^{20}$. The output labels are i.i.d. drawn from $y = \lambda \sum_{a=1}^4 \sum_{b=a+1}^4 x^a x^b + w$, where $w \sim \mathcal{N}(0, 1)$, and λ is a rescaling factor that determines the signal to noise ratio to be 15:1. We extract training sets of size n which varies from 40 to 120 with steps of 10. We then apply our method with polynomial kernel of degree $p = 4$, letting vary ν in $\{0.1, 0.2, 0.5, 1, 2, 5, 10, 20\}$. For fixed n and ν we evaluate the optimal value of the regularization parameter τ via hold out validation on an independent validation set of $n_{\text{val}} = n$ samples. We measure the selection error as the mean of the false negative rate (fraction of relevant variables that were not selected) and false positive rate (fraction of irrelevant variables that were selected). Then, we evaluate the prediction error as the root mean square error (RMSE) error of the selected model on an independent test set of $n_{\text{test}} = 500$ samples. Finally we average over 50 repetitions.

In Figure 4 we display the prediction error, selection error, and computing time, versus n for different values of ν. Clearly, if ν is too small, both prediction and selection are poor. For $\nu \geq 1$ the algorithm is quite stable with respect to small variations of ν. However, excessive increase of the smoothness parameter leads to a decrease in prediction and selection performance. In terms of computing time, the higher the smoothness parameter the better the performance.

Figure 4: Selection error (left), prediction error (center), and computing time (right) versus n for different values of ν. The points correspond to the mean over the repetitions. The dotted line represents the white noise standard deviation. In the left figure the curves corresponding to $\nu = 5$, $\nu = 10$, and $\nu = 20$ are overlapping.



5.1.2 Varying the parameters setting

We present 3 sets of experiments where we evaluated the performance of our method (DENOVAS) when varying part of the inputs parameters and leaving the others unchanged. The parameters we take into account are the following

- n, training set size
- d, input space dimensionality
- $|R_\rho|$, number of relevant variables
- p, size of the hypotheses space, measured as the degree of the polynomial kernel.

In all the following experiments the input variables $x = (x^1, \ldots, x^d)$ are uniformly drawn from $[-1, 1]^d$. The output labels are computed using a noise-corrupted regression function f that depends nonlinearly from a set of the input variables, i.e. $y = \lambda f(x) + w$, where w is a white noise, sampled from a normal distribution with zero mean and variance 1, and λ is a rescaling factor that determines the signal to noise ratio. For fixed n, d, and $|R_\rho|$ we evaluate the optimal value of the regularization parameter τ via hold out validation on an independent validation set of $n_{\text{val}} = n$ samples.

Varying n, d, **and** $|R_\rho|$ In this experiment we want to empirically evaluate the effect of the input space dimensionality, d, and the number of relevant variables, $|R_\rho|$, when the other parameters are left unchanged. In particular we use $d = 10, 20, 30, 40$ and $|R_\rho| = 2, 3, 4, 5, 6$. For each value of $|R_\rho|$ we use a different regression function, $f(x) = \lambda \sum_{a=1}^{|R_\rho|} \sum_{b=a+1}^{|R_\rho|} c_{ab} x^a x^b$, so that for fixed $|R_\rho|$ all 2-way interaction terms are present, and the polynomial degree of the regression function is always 2. The coefficients c_{ab} are randomly drawn from $[.5, 1]$ And λ is determined in order to set the signal to noise ratio as 15:1. We then apply our method with polynomial kernel of degree $p = 2$. The regression function thus always belongs to the hypotheses space.
In Figure 5, we display the selection error, and the prediction error, respectively, versus n for different values of d and number of relevant variables $|R_\rho|$. Both errors decrease with n and increase with d and $|R_\rho|$. In order to

Figure 5: Prediction error (top) and selection error (bottom) versus n for different values of d and number of relevant variables ($|R_\rho|$). The points correspond to the means over the repetitions. The dotted line represents the white noise standard deviation.

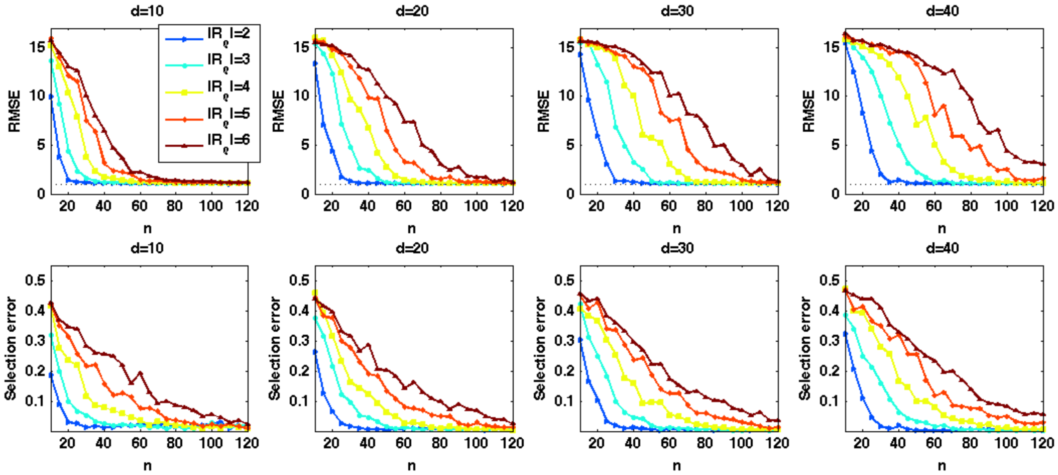
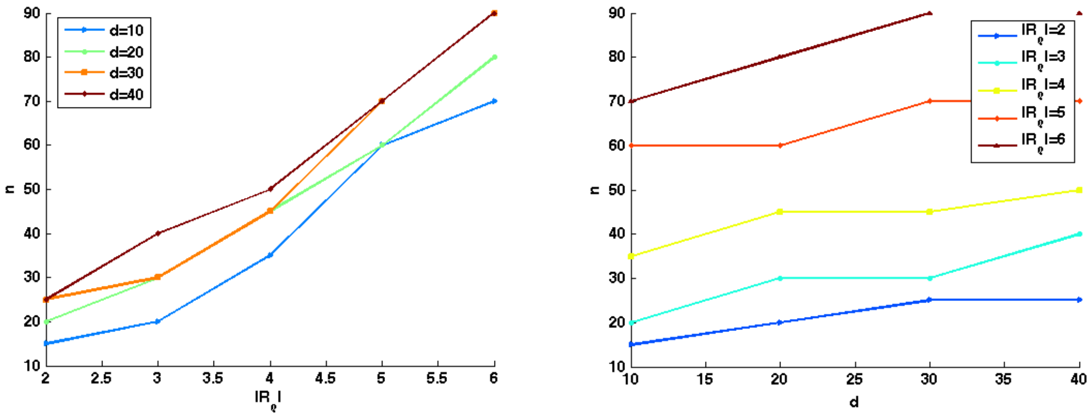


Figure 6: Minimum number of input points (n) necessary to achieve 10% of average selection error versus the number of relevant variables $|R_\rho|$ for different values of d (left), and versus d for different values of $|R_\rho|$ (right).



better visualize the dependance of the selection performance with respect to d and $|R_\rho|$, in Figure 6 we plotted the minimum number of input points that are necessary in order to achieve 10% of average selection error. It is clear by visual inspection that $|R_\rho|$ has a higher influence than d on the selection performance of our method.

Varying n and p In this experiment we want to empirically evaluate the effect of the size of the hypotheses space on the performance of our method. We therefore leave unchanged the data generation setting, made exception for the number of training samples, and vary the polynomial kernel degree as $p = 1, 2, 3, 4, 5, 6$. We let $d = 20$, $R_\rho = \{1, 2\}$, and $f(x) = x^1 x^2$, and let vary n from 20 to 80 with steps of 5. The signal to noise ratio is 3:1.

In Figure 7, we display the prediction and selection error, versus n, for different values of p. For $p \geq 2$, that is when the hypotheses space contains the regression function, both errors decrease with n and increase with p. Nevertheless the effect of p decreases for large p, in fact for $p = 4, 5$, and 6, the performance is almost the same. On the other hand, when the hypotheses space is too small to include the regression function, as for the set of linear functions ($p = 1$), the selection error coincides with chance (0.5), and the prediction error is very high,

Figure 7: Prediction error (left) and selection error (right) versus n for different values of the polynomial kernel degree (p). The points correspond to the means over the repetitions. The dotted line represents the white noise standard deviation.

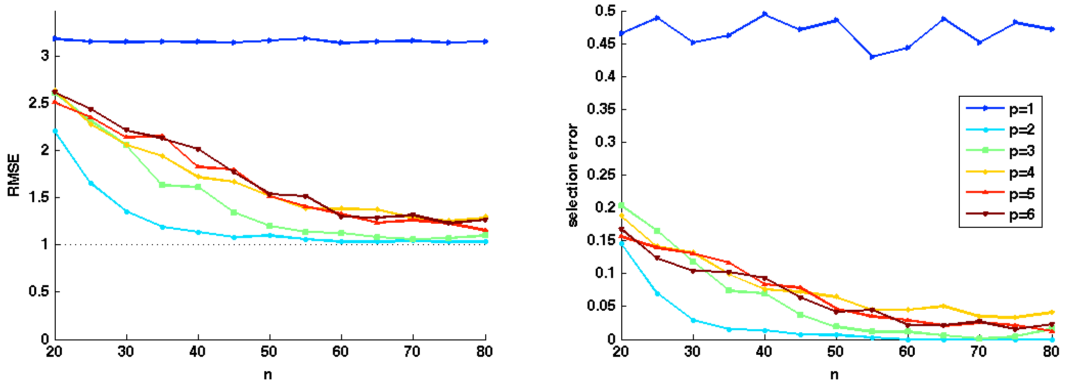
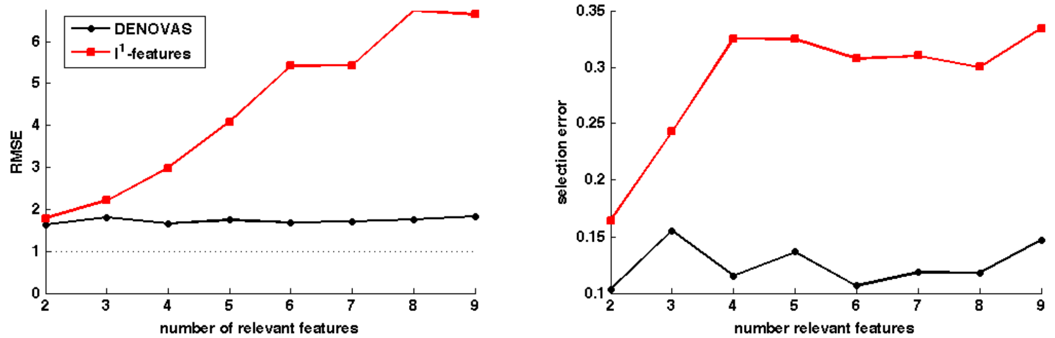


Figure 8: Prediction error (left) selection error (right) versus the number of relevant features. The points correspond to the means over the repetitions. The dotted line represents the white noise standard deviation.



even for large numbers of samples.

Varying the number of relevant features, for fixed $|R_\rho|$: comparison with ℓ^1 regularization on the feature space In this experiment we want to empirically evaluate the effect of the number of features involved in the regression function (that is the number of monomials constituting the polynomial) on the performance of our method when $|R_\rho|$ remains the same as well as all other input parameters. Note that while $|R_\rho|$ is the number of relevant variables, here we vary the number of relevant features (not variables!), which, in theory, has nothing to do with $|R_\rho|$. Furthermore we compare the performance of our method to that of ℓ^1 regularization on the feature space (ℓ^1-features). We therefore leave unchanged the data generation setting, made exception for the regression function. We set $d = 10$, $R_\rho = \{1, 2, 3\}$, $n = 30$, and then use a polynomial kernel of degree 2. The signal to noise ratio is this time 3:1. Note that with this setting the size of the features space is 66. For fixed number of relevant features the regression function is set to be a randomly chosen linear combination of the features involving one or two of the first three variables ($x^1, (x^1)^2, x^1x^2, x^1x^3$, etc.), with the constraint that the combination must be a polynomial of degree 2, involving all 3 variables.
In Figure 8, we display the prediction and selection error, versus the number of relevant features. While the performance of ℓ^1-features fades when the number of relevant features increases, our method presents stable performance both in terms of selection and prediction error. From our simulation it appears that, while our method depends on the number of relevant variables, it is indeed independent of the number of features.

5.2 Comparison with Other Methods

In this section we present numerical experiments aimed at comparing our method with state-of-the-art algorithms. In particular, since our method is a regularization method, we focus on alternative regularization approaches to the problem of nonlinear variable selection. For comparisons with more general techniques for nonlinear variable selection we refer the interested reader to Bach (2009).

5.2.1 Compared algorithms

We consider the following regularization algorithms:

- Additive models with multiple kernels, that is Multiple Kernel Learning (MKL)
- ℓ^1 regularization on the feature space associated to a polynomial kernel (ℓ^1-features)
- COSSO (Lin and Zhang, 2006) with 1-way interaction (COSSO1) and 2-way interactions (COSSO2) [6]
- Hierarchical Kernel Learning (Bach, 2009) with polynomial (HKL pol.) and hermite (HKL herm.) kernel
- Regularized Least Squares (RLS).

Note that, differently from the first 4 methods, RLS is not a variable selection algorithm, however we consider it since it is typically a good benchmark for the prediction error.

For ℓ^1-features and MKL we use our own Matlab implementation based on proximal methods (for details see Mosci et al. (2010)). For COSSO we used the Matlab code available at `www.stat.wisc.edu/~yilin` or `www4.stat.ncsu.edu/~hzhang` which can deal with 1 and 2-way interactions. For HKL we used the code available online at `http://www.di.ens.fr/~fbach/hkl/index.html`. While for MKL and ℓ^1-features we are able to identify the set of selected variables, for COSSO and HKL extracting the sparsity patterns from the available code is not straightforward. We therefore compute the selection errors only for ℓ^1-features, MKL, and our method .

5.2.2 Synthetic data

We simulated data with d input variables, where each variable is uniformly sampled from [-2,2]. The output y is a nonlinear function of the first 4 variables, $y = f(x^1, x^2, x^3, x^4) + \epsilon$, where epsilon is a white noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and σ is chosen so that the signal to noise ratio is 15:1. We consider the 4 models described in Table 1.

Table 1: Synthetic data design

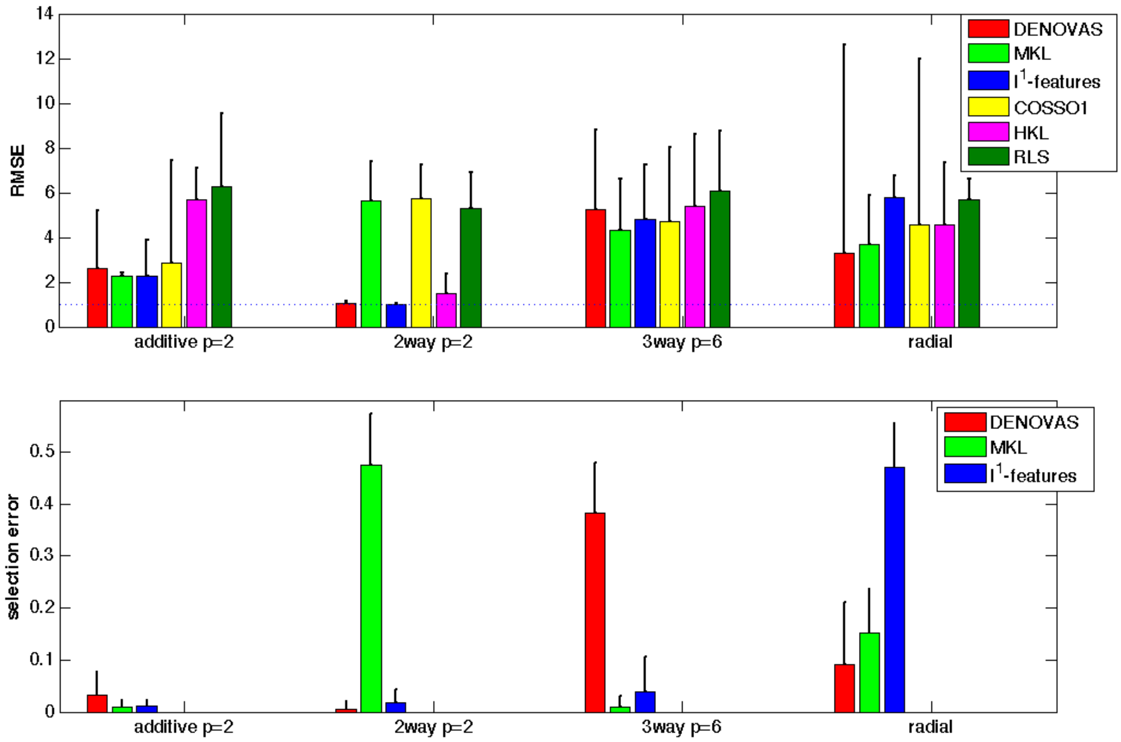
	d	**number of relevant variables**	**n**	**model (f)**
additive p=2	40	4	100	$y = \sum_{a=1}^{4} (x^a)^2$
2way p=2	40	4	100	$y = \sum_{a=1}^{4} \sum_{b=a+1}^{4} x^a x^b$
3way p=6	40	3	100	$y = (x^1 x^2 x^3)^2$
radial	20	2	100	$y = \frac{1}{\pi}((x^1)^2 + (x^2)^2) e^{-((x^1)^2 + (x^2)^2)}$

For model selection and testing we follow the same protocol described at the beginning of Section 5, with $n = 100, 100$ and 1000 for training, validation and testing, respectively. Finally we average over 20 repetitions. In the first 3 models, for MKL, HKL, RLS and our method we employed the polynomial kernel of degree p, where p is the polynomial degree of the regression function f. For ℓ^1-features we used the polynomial kernel with degree chosen as the minimum between the polynomial degree of f and 3. This was due to computational

[6] In all toy data, and in part of the real data, the following warning message was displayed:
```
Maximum number of iterations exceeded; increase options.MaxIter.
To continue solving the problem with the current solution as the starting point,
set x0 = x before calling lsqlin.
```
In those cases the algorithm did not reach convergence in a reasonable amount of time, therefore the error bars corresponding to COSSO2 were omitted.

reasons, in fact, with $p = 4$ and $d = 40$, the number of features is highly above $100,000$. For the last model, we used the polynomial kernel of degree 4 for MKL, ℓ^1-features and HKL, and the Gaussian kernel with kernel parameter $\sigma = 2$ for RLS and our method [7]. COSSO2 never reached convergence. Results in terms of prediction and selection errors are reported in Figure 9.

Figure 9: Prediction error (top) and fraction of selected variables (bottom) on synthetic data for the proposed method (DENOVAS), multiple kernel learning for additive models (MKL), ℓ^1 regularization on the feature space associated to a polynomial kernel (ℓ^1-features), COSSO with 1-way interactions (COSSO1), hierarchical kernel learning with polynomial kernel (HKL pol.), and regularized least squares (RLS). The dotted line in the upper plot corresponds to the white noise standard deviation. Selection errors for COSSO, and HKL are not reported because they are not straightforwardly computable from the available code.



When the regression function is simple (low interaction degree or low polynomial degree) more tailored algorithms, such as MKL–which is additive by design– for the experiment "additive p=2", or ℓ^1-features for experiments "2way p=4" – in this case the dictionary size is less than 1000–, compare favorably with respect to our method. However, when the nonlinearity of the regression function favors the use of a large hypotheses space, our method significantly outperforms the other methods. This is particularly evident in the experiment "radial", which was anticipated in Section 2, where we plotted in Figure 2 the regression function and its estimates obtained with the different algorithms for one of the 20 repetitions.

5.2.3 Real data

We consider the 7 benchmark data sets described in Table 2. We build training and validation sets by randomly drawing n_{train} and n_{val} samples, and using the remaining samples for testing. For the first 6 data sets we let $n_{\text{train}} = n_{\text{val}} = 150$, whereas for breast cancer data we let $n_{\text{train}} = n_{\text{val}} = 60$. We then apply the algorithms described in Subsection 5.2.1. with the validation protocol described in Section 5. For our method and RLS we used the gaussian kernel with the kernel parameter σ chosen as the mean over the samples of the euclidean

[7] Note that here we are interested in evaluating the ability of our method of dealing with a general kernel like the Gaussian kernel, not in the choice of the kernel parameter itself. Nonetheless, a data driven choice for σ will be presented in the real data experiments in Subsection 5.2.3.

Table 2: Real data sets

data name	number of input variables	number of instances	source	task
boston housing	13	506	LIACC[8]	regression
census	16	22784	LIACC	regression
delta ailerons	5	7129	LIACC	regression
stock	10	950	LIACC	regression
image segmentation	18	2310	IDA[9]	classification
pole telecomm	26[10]	15000	LIACC	regression
breast cancer	32	198	UCI [11]	regression

distance form the 20-th nearest neighbor. Since the other methods cannot be run with the gaussian kernel we used a polynomial kernel of degree $p = 3$ for MKL and ℓ^1-features. For HKL we used both the polynomial kernel and the hermite kernel, both with $p = 3$. Results in terms of prediction and selection error are reported in Figure 10.

Some of the data, such as the stock data, seem not to be variable selection problem, in fact the best performance is achieved by our method though selecting (almost) all variables, or, equivalently by RLS. Our method outperforms all other methods on several data sets. In most cases, the performance of our method and RLS are similar. Nonetheless our method brings higher interpretability since it is able to select a smaller subset of relevant variable, while the estimate provided by RLS depends on all variables.

We also run experiments on the same 7 data sets with different kernel choices for our method . We consider the polynomial kernel with degree $p = 2, 3$ and 4, and the gaussian kernel. Comparisons among the different kernels in terms of prediction and selection accuracy are plotted in Figure 11. Interestingly the choice of the gaussian kernel seems to be the preferable choice in most data sets.

6 Discussion

Sparsity based method has recently emerged as way to perform learning and variable selection from high dimensional data. So far, compared to other machine learning techniques, this class of methods suffers from strong modeling assumptions and is in fact limited to parametric or semi-parametric models (additive models). In this paper we discuss a possible way to circumvent this shortcoming and exploit sparsity ideas in a non-parametric context.

We propose to use partial derivatives of functions in a RKHS to design a new sparsity penalty and a corresponding regularization scheme. Using results from the theory of RKHS and proximal methods we show that the regularized estimator can be provably computed through an iterative procedure. The consistency property of the proposed estimator are studied. Exploiting the non-parametric nature of the method we can prove universal consistency. Moreover we study selection properties and show that that the proposed regularization scheme represents a safe filter for variable selection, as it does not discard relevant variables. Extensive simulations on synthetic data demonstrate the prediction and selection properties of the proposed algorithm. Finally, comparisons to state-of-the-art algorithms for nonlinear variable selection on toy data as well as on a cohort of benchmark data sets, show that our approach often leads to better prediction and selection performance.

Our work can be considered as a first step towards understanding the role of sparsity beyond additive models. Several research directions are yet to be explored.

- From a theoretical point of view it would be interesting to further analyze the sparsity property of the obtained estimator in terms of finite sample estimates for the prediction and the selection error.
- From a computational point of view the main question is whether our method can be scaled to work in very high dimensions. Current computations are limited by memory constraints. A variety of method for large scale optimization can be considered towards this end.
- A natural by product of computational improvements would be the possibility of considering a semi-supervised setting which is naturally suggested by our approach. More generally we plan to investigate

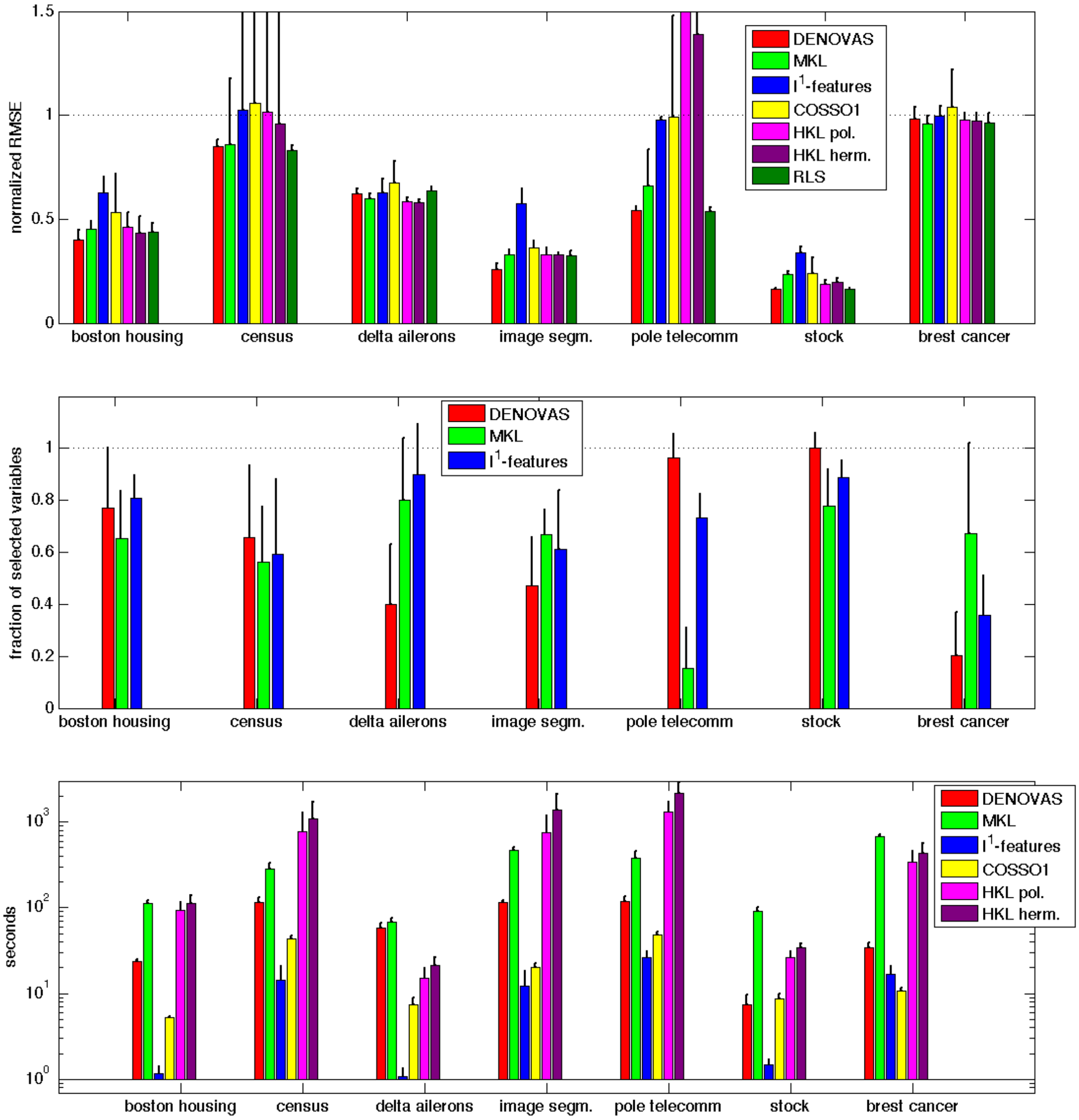


Figure 10: Prediction error (top) and fraction of selected variables (center) and computing time (bottom) on real data for the proposed method (DENOVAS), multiple kernel learning for univariate additive functions (MKL), ℓ^1 regularization on the feature space associated to a polynomial kernel (ℓ^1-features), COSSO with 1-way interactions (COSSO1), hierarchical kernel learning with polynomial kernel (HKL pol.), hierarchical kernel learning with hermite kernel (HKL herm.) and regularized least squares (RLS). Prediction errors for COSSO2 are not reported because it is always outperformed by COSSO1. Furthermore prediction errors for COSSO2 were largely out of scale in the first three data sets, and were not available since the algorithm did not reach convergence for image segmentation, pole telecomm and breast cancer data. In order to make the prediction errors comparable among experiments, root mean squared errors (RMSE) were divided by the outputs standard deviation, which corresponds to the dotted line. Error bars represent the standard deviations of the normalized RMSE. Though the largest normalized RMSE appear out of the figure axis, we preferred to display the prediction errors with the current axes limits in order to allow the reader to appreciate the difference between the smallest, and thus most significant, errors. Selection errors for COSSO, and HKL are not reported because they are not straightforwardly computable from the available code. The computing time corresponds to the entire model selection and testing protocol. Computing time for RLS is not reported because it was always negligible with respect to the other methods.

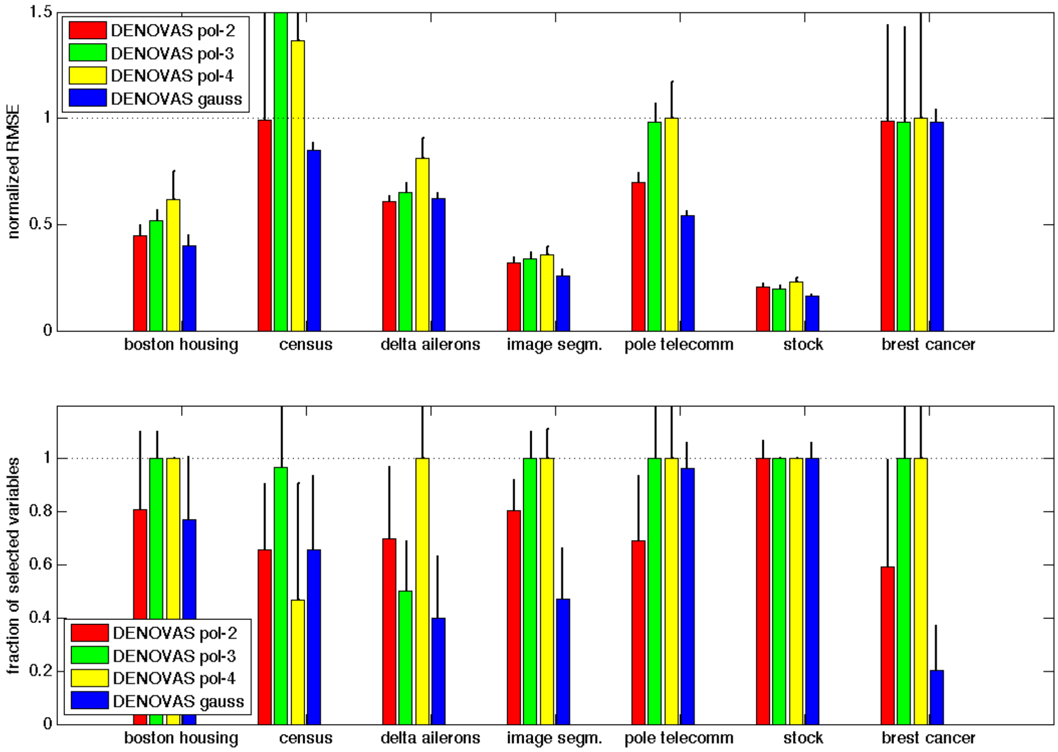


Figure 11: Prediction error (top) and fraction of selected variables (bottom) on real data for our method with different kernels: polynomial kernel of degree $p = 1, 2$ and 3 (DENOVAS pol-p), and Guassian kernel (DENOVAS gauss). Error bars represent the standard deviations. In order to make the prediction errors comparable among experiments, root mean square errors were divided by the outputs standard deviation, which corresponds to the dotted line.

the application of the RKHS representation for differential operators in unsupervised learning.

- More generally, our study begs the question of whether there are alternative/better ways to perform learning and variable selection beyond additive models and using non parametric models.

Acknowledgements

The authors would like to thank Ernesto De Vito for many useful discussions and suggesting the proof of **Lemma 4**. SM and LR would like to thank Francis Bach, Guillame Obozinski and Tomaso Poggio for useful discussions. This paper describes a joint research work done at and at the Departments of Computer Science and Mathematics of the University of Genoa and at the IIT@MIT lab hosted in the Center for Biological and Computational Learning (within the McGovern Institute for Brain Research at MIT), at the Department of Brain and Cognitive Sciences (affiliated with the Computer Sciences and Artificial Intelligence Laboratory), The authors have been partially supported by the Integrated Project Health-e-Child IST-2004-027749 and by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), and Compagnia di San Paolo, Torino. Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

H. Attouch and R. Wets. Quantitative stability of variational systems. I. The epigraphical distance. *Trans. Amer. Math. Soc.*, 328(2):695–729, 1991.

H. Attouch and R. Wets. Quantitative stability of variational systems. II. A framework for nonlinear conditioning. *SIAM J. Optim.*, 3(2):359–381, 1993a.

H. Attouch and R. Wets. Quantitative stability of variational systems. III. ϵ-approximate solutions. *Math. Programming*, 61(2, Ser. A):197–214, 1993b.

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9: 1179–1225, 2008.

F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical Report HAL 00413473, INRIA, 2009.

F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, volume 69 of *ACM International Conference Proceeding Series*, 2004.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

S. Becker, J. Bobin, and E. Candès. NESTA: a fast and accurate first-order method for sparse recovery. *SIAM J. Imaging Sci.*, 4(1):1–39, 2011. ISSN 1936-4954.

J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle. A ℓ^1-unified variational framework for image restoration. In T. Pajdla and J. Matas, editors, *ECCV 2004*, volume 3024 of *Lecture Notes in Computer Science*, pages 1–13. Springer, Berlin, 2004.

M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

K. Bertin and G. Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241, 2008.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, Berlin, 2011.

T. Chan, G. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *Siam Journal on Scientific Computing*, 20, 1999.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

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

P. L. Combettes, D. Dũng, and B. C. Vũ. Dualization of signal recovery problems. *Set-Valued and Variational Analysis*, 18(3-4):373–404, 2010.

L. Comminges and A. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. In *Proceeding of the 24th Annual Conference on Learning Theory*, 2011.

I. Daubechies, G. Teschke, and L. Vese. Iteratively solving linear inverse problems under general convex constraints. *Inverse Problems and Imaging*, 1(1):29–46, 2007.

E. De Vito, L. Rosasco, A. Caponnetto, U. De giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.

R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constr. Approx.*, 33(1):125–143, 2011.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

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

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, December 2009.

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

I. Ekeland and R. Temam. *Convex analysis and variational problems*. North-Holland Publishing Co., Amsterdam, 1976.

M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):586–597, 2007.

C. Gu. *Smoothing spline ANOVA models*. Springer series in statistics. Springer, 2002.

E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for l1-minimization: Methodology and convergence. *SIOPT*, 19(3):1107–1130, 2008.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms: Part I: Fundamentals*. Springer, Berlin, 1993.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Intearnational Annual Conference on Machine Learning*, pages 433–440, 2009.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceeding of ICML 2010*, 2010.

V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*, 38(6):3660–3695, 2010.

J. Lafferty and L. Wasserman. Rodeo: Sparse, greedy nonparametric regression. *Annals of Statistics*, 36(1):28–63, 2008.

Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34:2272, 2006.

I. Loris. On the performance of algorithms for the minimization of l_1-penalized functionals. *Inverse Problems*, 25(3):035008, 16, 2009.

I. Loris, M. Bertero, C. De Mol, R. Zanella, and L. Zanni. Accelerating gradient projection methods for l_1-constrained signal recovery by steplength selection rules. *Appl. Comput. Harmon. Anal.*, 27(2):247–254, 2009. ISSN 1063-5203. doi: 10.1016/j.acha.2009.02.003. URL `http://dx.doi.org/10.1016/j.acha.2009.02.003`.

J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. *LNCS*, 6322:418–433, 2010.

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*. NIPS Foundation, 2009.

A.S. Nemirovski and D.B. Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience Series in Discrete Mathematics*, 1983.

Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Paper 2007/76, Catholic University of Louvain, September 2007.

Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Doklady AN SSSR*, 269(3):543–547, 1983.

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985. ISSN 0040-361X.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*. NIPS Foundation, 2008.

L. Rosasco, S. Mosci, M. S. Santoro, A. Verri, and S. Villa. A regularization approach to nonlinear variable selection. In *Proceedings of the 13 International Conference on Artificial Intelligence and Statistics*, 2010.

M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 286–297, 2007. doi: 10.1007/978-3-540-74958-5_28.

I. Steinwart and A. Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53:4655–4666, 2007.

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program.*, 125(2, Ser. B):263–295, 2010. doi: 10.1007/s10107-010-0394-2.

S. Villa, L. Rosasco, S. Mosci, and A. Verri. Consistency of learning algorithms using attouchwets convergence. *Optimization*, 0(0):1–19, 0. doi: 10.1080/02331934.2010.511671. URL `http://www.tandfonline.com/doi/abs/10.1080/02331934.2010.511671`.

S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *Optimization Online*, E-Print 2011 08 3132, 2011. URL `http://www.optimization-online.org`.

G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 1990.

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist*, 23:1865–1895, 1995.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.

D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *J. Comput. Appl. Math.*, 220:456–463, 2008.

A Derivatives in RKHS and Representer Theorem

Consider $L^2(\mathcal{X}, \rho_{\mathcal{X}}) = \{f : \mathcal{X} \to \mathbb{R} \text{ measurable} | \int |f(x)|^2 d\rho_{\mathcal{X}}(x) < \infty\}$ and $\mathbb{R}^n$ with inner product normalized by a factor $1/n$, $\|\cdot\|_n$.

The operator $I_k : \mathcal{H} \to L^2(\mathcal{X}, \rho_{\mathcal{X}})$ defined by $(I_k f)(x) = \langle f, k_x \rangle_{\mathcal{H}}$, for almost all $x \in X$, is well-defined and bounded thanks to assumption A1. The sampling operator (18) can be seen as its empirical counterpart. Similarly $D_a : \mathcal{H} \to L^2(\mathcal{X}, \rho_{\mathcal{X}})$ defined by $(D_a f)(x) = \langle f, (\partial_a k)_x \rangle$, for almost all $x \in X$ and $a = 1, \ldots d$, is well-defined and bounded thanks to assumption A2. The operator (20) can be seen as its empirical counterpart. Several properties of such operators and related quantities are given by the following two propositions.

Proposition 4. *If assumptions A1 and A2 are met, the operator I_k and the continuous partial derivative D_a are Hilbert-Schmidt operators from $\mathcal{H}$ to $L^2(\mathcal{X}, \rho_{\mathcal{X}})$, and*

$$I_k^* g(t) = \int_{\mathcal{X}} k_x(t) g(x) d\rho_{\mathcal{X}}(x), \qquad I_k^* I_k f(t) = \int_{\mathcal{X}} k_x(t) \langle f, k_x \rangle_{\mathcal{H}} d\rho_{\mathcal{X}}(x)$$
$$D_a^* g(t) = \int_{\mathcal{X}} (\partial_a k)_x(t) g(x) d\rho_{\mathcal{X}}(x), \qquad D_a^* D_b f(t) = \int_{\mathcal{X}} (\partial_a k)_x(t) \langle f, (\partial_b k)_x \rangle_{\mathcal{H}} d\rho_{\mathcal{X}}(x)$$

Proposition 5. *If assumptions A1 and A2 are met, the sampling operator $\hat{S}$ and the empirical partial derivative $\hat{D}_a$ are Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathbb{R}^n$, and*

$$\hat{S}^* v = \frac{1}{n} \sum_{i=1}^n k_{x_i} v_i, \qquad \hat{S}^* \hat{S} f = \frac{1}{n} \sum_{i=1}^n k_{x_i} \langle f, k_{x_i} \rangle_{\mathcal{H}}$$
$$\hat{D}_a^* v = \frac{1}{n} \sum_{i=1}^n (\partial_a k)_{x_i} v_i, \qquad \hat{D}_a^* \hat{D}_b f = \frac{1}{n} \sum_{i=1}^n (\partial_a k)_{x_i} \langle f, (\partial_b k)_{x_i} \rangle_{\mathcal{H}}$$

where $a, b = 1, \ldots, d$.

The proof can be found in De Vito et al. (2005) for I_k and $\hat{S}$, where assumption A1 is used. The proof for D_a and $\hat{D}_a$ is based on the same tools and on assumption A2. Furthermore, a similar result can be obtained for the continuous and empirical gradient

$$\nabla : \mathcal{H} \to (L^2(\mathcal{X}, \rho_{\mathcal{X}}))^d, \qquad \nabla f = (D_a f)_{a=1}^d$$
$$\hat{\nabla} : \mathcal{H} \to (\mathbb{R}^n)^d, \qquad \hat{\nabla} f = (\hat{D}_a f)_{a=1}^d,$$

which can be shown to be Hilbert-Schmidt operators from $\mathcal{H}$ to $(L^2(\mathcal{X}, \rho_{\mathcal{X}}))^d$ and from $\mathcal{H}$ to $(\mathbb{R}^n)^d$, respectively.

We next restate Proposition 1 in a slightly more abstract form and give its proof.

Proposition (Proposition 1 Extended)**.** *The minimizer of (6) satisfies $\hat{f}^\tau \in$ Range$(\hat{S}^*)$ + Range$(\hat{\nabla}^*)$. Henceforth it satisfies the following representer theorem*

$$\hat{f}^\tau = \hat{S}^* \alpha + \hat{\nabla}^* \beta = \sum_{i=1}^n \frac{1}{n} \alpha_i k_{x_i} + \sum_{i=1}^n \sum_{a=1}^d \frac{1}{n} \beta_{a,i} (\partial_a k)_{x_i} \tag{36}$$

with $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^{nd}$.

Proof. Being Range$(\hat{S}^*)$ + Range$(\hat{\nabla}^*)$ a closed subspace of $\mathcal{H}$, we can write any function $f \in \mathcal{H}$ as $f = f^{//} + f^\perp$, where $f^{//} \in$ Range$(\hat{S}^*)$ + Range$(\hat{\nabla}^*)$ and $\langle f^\perp, g \rangle_{\mathcal{H}}$ for all $g \in$ Range$(\hat{S}^*)$ + Range$(\hat{\nabla}^*)$. Now if we plug the decomposition $f = f^{//} + f^\perp$ in the variational problem (6), we obtain

$$\hat{f}^\tau = \underset{f \in \mathcal{H},\, f = f^{//} + f^\perp}{\operatorname{argmin}} \left\{ \widehat{\mathcal{E}}(f^{//}) + 2\tau \widehat{\Omega}_1^D(f^{//}) + \tau\nu \|f^{//}\|_{\mathcal{H}}^2 + \tau\nu \|f^\perp\|_{\mathcal{H}}^2 \right\}$$

which is clearly minimized by $f^\perp = 0$. The second equality in (36) then derives directly from definition of $\hat{S}^*$ and $\hat{\nabla}^*$. □

We conclude with the following example on how to compute derivatives and related quantities for the Gaussian Kernel.

Example 1. *Note that all the terms involved in* (33) *are explictly computable. As an example we show how to compute them when* $k(x,s) = e^{-\frac{\|x-s\|^2}{2\gamma^2}}$ *is the gaussian kernel on* $\mathbb{R}^d$*. By definition*

$$(\partial_a k)_{x_i}(x) = \langle \frac{\partial k(s,\cdot)}{\partial s^a}\Big|_{s=x_i}, k_x\rangle_{\mathcal{H}}.$$

Given $x \in \mathcal{X}$ *it holds*

$$\frac{\partial k(s,x)}{\partial s^a} = e^{-\frac{\|x-s\|^2}{2\gamma^2}} \cdot \left(-\frac{s^a - x^a}{\gamma^2}\right) \quad \Longrightarrow \quad (\partial_a k)_{x_i}(x) = e^{-\frac{\|x-x_i\|^2}{2\gamma^2}} \cdot \left(-\frac{x_i^a - x^a}{\gamma^2}\right).$$

Moreover, as we mentioned above, the computation of $\beta_{a,i}^t$ *and* α_i^t *requires the knowledge of matrices* $\mathrm{K}, \mathrm{Z}_a, \mathrm{Z}, \mathrm{L}_a$*. Also their entries are easily found starting from the kernel and the training points. We only show how the entries of* Z *and* L_a *look like. Using the previous computations we immediately get*

$$[\mathrm{Z}_a]_{i,j} = e^{-\frac{\|x_j-x_i\|^2}{2\gamma^2}} \cdot \left(-\frac{x_i^a - x_j^a}{\gamma^2}\right).$$

In order to compute L_a *we need the second partial derivatives of the kernel:*

$$\frac{\partial k(s,x)}{\partial x^b \partial s^a} = \begin{cases} -e^{-\frac{\|x-s\|^2}{2\gamma^2}} \cdot \frac{s^a-x^a}{\gamma^2} \cdot \frac{s^b-x^b}{\gamma^2} & \text{if } a \neq b \\ -e^{-\frac{\|x-s\|^2}{2\gamma^2}} \cdot \left(\frac{(s^a-x^a)^2}{\gamma^2} - \frac{1}{\gamma^2}\right) & \text{if } a = b. \end{cases}$$

so that

$$[\mathrm{L}_{a,b}]_{i,j} = \begin{cases} -e^{-\frac{\|x_j-x_i\|^2}{2\gamma^2}} \cdot \frac{x_i^a-x_j^a}{\gamma^2} \cdot \frac{x_i^b-x_j^b}{\gamma^2} & \text{if } a \neq b \\ -e^{-\frac{\|x_j-x_i\|^2}{2\gamma^2}} \cdot \left(\frac{(x_i^a-x_j^a)^2}{\gamma^2} - \frac{1}{\gamma^2}\right) & \text{if } a = b. \end{cases}$$

B Proofs of Section 3

In this appendix we collect the proofs related to the derivation of the iterative procedure given in Algorithm 1. Theorem 1 is a consequence of the general results about convergence of accelerated and inexact FB-splitting algorithms in Villa et al. (2011). In that paper it is shown that inexact schemes converge only when the errors in the computation of the proximity operator are of a suitable type and satisfy a sufficiently fast decay condition. We first introduce the notion of admissible approximations.

Definition 2. *Let* $\varepsilon \geq 0$ *and* $\lambda > 0$*. We say that* $h \in \mathcal{H}$ *is an* approximation of $\mathrm{prox}_{\lambda\widehat{\Omega}_1^D}(f)$ with ε-precision *and we write* $h \approxeq_\varepsilon \mathrm{prox}_{\lambda\widehat{\Omega}_1^D}(f)$*, if and only if*

$$\frac{f-h}{\lambda} \in \partial_{\frac{\varepsilon^2}{2\lambda}} \widehat{\Omega}_1^D(h), \tag{37}$$

where $\partial_{\frac{\varepsilon^2}{2\lambda}}$ *denotes the* ε*-subdifferential.*[12]

We will need the following results from Villa et al. (2011).

[12]Recall that the ε-subdifferential ∂_ϵ of a convex functional $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is defined as the set

$$\partial_\varepsilon \Omega(f) := \{h \in \mathcal{H} \,:\, \Omega(g) - \Omega(f) \geq \langle h, g-f\rangle_{\mathcal{H}} - \epsilon, \;\; \forall g \in \mathcal{H}\}, \qquad \forall f \in \mathcal{H}.$$

Theorem 6. *Consider the following inexact version of the accelerated FB-algorithm in* (22) *with* $c_{1,t}$ *and* $c_{2,t}$ *as in* (24)

$$f^t \approxeq_{\varepsilon^t} prox_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D}\Big(\Big(I - \frac{1}{2\sigma}\nabla F\Big)(c_{1,t}f^{t-1} + c_{2,t}f^{t-2})\Big). \tag{38}$$

Then, if $\varepsilon^t \sim 1/t^l$ *with* $l > 3/2$*, there exists a constant* $C > 0$ *such that*

$$\widehat{\mathcal{E}}^\tau(f^t) - \inf \widehat{\mathcal{E}}^\tau \leq \frac{C}{t^2}.$$

Proposition 6. *Suppose that* $\Omega : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ *can be written as* $\Omega = \omega\circ B$*, where* $B : \mathcal{H} \to \mathcal{G}$ *is a linear and bounded operator between Hilbert spaces and* $\omega : \mathcal{G} \to \mathbb{R}\cup\{+\infty\}$ *is a one-homogeneous function such that* $S := \partial\omega(0)$ *is bounded. Then for any* $f \in \mathcal{H}$ *and any* $v \in S$ *such that* $\|\lambda B^*v - \pi_{\lambda B^*(S)}(f)\| \leq \varepsilon^2/(2\lambda\|B^*\|\mathrm{diam}(S) + 2\|f - \pi_{\lambda B^*(S)}(g)\|)$ *it holds*

$$f - \lambda B^*v \approxeq_\varepsilon \mathrm{prox}_{\lambda\Omega}(f).$$

Proof of Theorem 1. Since the the regularizer $\widehat{\Omega}_1^D$ can be written as a composition of $\omega \circ B$, with $B = \hat{\nabla}$ and $\omega : \mathbb{R}^d \to [0, +\infty)$, $\omega(v) = \sum_{a=1}^d \|v_a\|_n$ Proposition 6 applied with $\lambda = \tau/\sigma$, ensures that each sequence of the type $\hat{\nabla}^*v^q$ which meets the condition (28) generates, via (29), admissible approximations of $\mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D}$. Therefore, if ε^t is such that $\varepsilon^t \sim 1/t^l$ with $l > 3/2$, Theorem 6 implies that the inexact version of the FB-splitting algorithm in (29) shares the $1/t^2$ convergence rate. Equation (32) directly follows from the definition of strong convexity,

$$\frac{\tau\nu}{8}\|f^t - \hat{f}^\tau\|^2 \leq \widehat{\mathcal{E}}^\tau(f^t)/2 + \widehat{\mathcal{E}}^\tau(\hat{f}^\tau)/2 - \widehat{\mathcal{E}}^\tau(f^t/2 + \hat{f}^\tau/2) \leq \frac{1}{2}(\widehat{\mathcal{E}}^\tau(f^t) - \widehat{\mathcal{E}}^\tau(\hat{f}^\tau))$$

□

Proof of Proposition 2. We first show that the matrices $\mathrm{K}, \mathrm{Z}_a, \mathrm{L}_a$ defined in (8),(9), and (10), are the matrices associated to the operators $\hat{S}\hat{S}^* : \mathbb{R}^n \to \mathbb{R}^n$, $\hat{S}\hat{D}_a^* : \mathbb{R}^n \to \mathbb{R}^n$ and $\hat{D}_a\hat{\nabla}^* : \mathbb{R}^{nd} \to \mathbb{R}^n$, respectively. For K, the proof is trivial and derives directly from the definition of adjoint of $\hat{S}$– see Proposition 5. For Z_a and Z, from the definition of $\hat{D}_a^*$ we have that

$$\left(\hat{S}\hat{D}_a^*\alpha\right)_i = \frac{1}{n}\sum_{j=1}^n \alpha_j(\partial_a k)_{x_j}(x_i) = \sum_{j=1}^n (\mathrm{Z}_a)_{i,j}\alpha_j = (\mathrm{Z}_a\alpha)_i,$$

so that $\hat{S}\hat{\nabla}^*\beta = \sum_{a=1}^d \hat{S}\hat{D}_a^*(\beta_{a,i})_{i=1}^n = \sum_{a=1}^d \mathrm{Z}_a(\beta_{a,i})_{i=1}^n = \mathrm{Z}\beta$. For L_a, we first observe that

$$\langle(\partial_a k)_x, (\partial_b k)_{x'}\rangle_{\mathcal{H}} = \left.\frac{\partial(\partial_b k)_{x'}(t)}{\partial t^a}\right|_{t=x} = \left.\frac{\partial^2 k(s,t)}{\partial t^a \partial s^b}\right|_{t=x, s=x'},$$

so that operator $\hat{D}_a\hat{D}_b^* : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$\left(\left(\hat{D}_a\hat{D}_b^*\right)v\right)_i = \langle(\partial_a k)_{x_i}, \hat{D}_b^*v\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{j=1}^n \langle(\partial_a k)_{x_i}, (\partial_b k)_{x_j}\rangle_{\mathcal{H}} v_j = (\mathrm{L}_{a,b})_{i,j}v_j$$

for $i = 1, \ldots, n$, for all $v \in \mathbb{R}^n$. Then, since $\hat{D}_a\hat{\nabla}^*\beta = \sum_{a=1}^d \hat{D}_a\hat{D}_b^*(\beta_{a,i})_{i=1}^n$, we have that L_a is the matrix associated to the operator $\hat{D}_a\hat{\nabla}^* : \mathbb{R}^{nd} \to \mathbb{R}^n$, that is

$$(\hat{D}_a\hat{\nabla}^*\beta)_i = \sum_{j=1}^n \sum_{b=1}^d (\mathrm{L}_{a,b})_{i,j}\beta_{b,j},$$

for $i = 1, \ldots, n$, for all $\beta \in \mathbb{R}^{nd}$. To prove equation (33), first note that, as we have done in Proposition 1 extended, (33) can be equivalently rewritten as $f^t = \hat{S}^*\alpha^t + \hat{\nabla}^*\beta^t$. We now proceed by induction. The base case, namely the representation for $t = 0$ and $t = 1$, is clear. Then, by the inductive hypothesis we have that $f^{t-1} = \hat{S}^*\alpha^{t-1} + \hat{\nabla}^*\beta^{t-1}$, and $f^{t-2} = \hat{S}^*\alpha^{t-2} + \hat{\nabla}^*\beta^{t-2}$ so that $\tilde{f}^t = \hat{S}^*\tilde{\alpha}^t + \hat{\nabla}^*\tilde{\beta}^t$ with $\tilde{\alpha}^t$ and $\tilde{\beta}^t$ defined by (12) and (13). Therefore, using (22), (29), (25) it follows that f^t can be expressed as:

$$(I - \pi_{\frac{\tau}{\sigma}\mathcal{C}_n})\left(\hat{S}^*\left(\left(1 - \frac{\tau\nu}{\sigma}\right)\tilde{\alpha}^t - \frac{1}{\sigma}\left(\mathrm{K}\tilde{\alpha}^t + \mathrm{Z}\tilde{\beta}^t - \mathbf{y}\right)\right) + \left(1 - \frac{\tau\nu}{\sigma}\right)\hat{\nabla}^*\tilde{\beta}^t\right)$$

and the proposition is proved, letting $\tilde{\alpha}^t$, $\tilde{\beta}^t$ and $\bar{v}^t$ as in Equations (14), (16) and (26).

For the projection, we first observe that operator $\hat{D}_a\hat{S}^* : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$(\hat{D}_a\hat{S}^*\alpha)_i = (\langle \hat{S}^*\alpha, (\partial_a k)_{x_i}\rangle_{\mathcal{H}}) = \frac{1}{n}\sum_{j=1}^n \alpha_j(\partial_a k)_{x_i}(x_j) = \sum_{j=1}^n \alpha_j (\mathrm{Z}_a)_{j,i} = \mathrm{Z}_a^T\alpha.$$

Then, we can plug the representation (36) in (30) to obtain (15). □

Proof of Proposition 3. Since $\hat{f}^\tau$ is the unique minimizer of the functional $\widehat{\mathcal{E}}^\tau$, it satisfies the Euler equation for $\widehat{\mathcal{E}}^\tau$

$$0 \in \partial(\widehat{\mathcal{E}}(\hat{f}^\tau) + 2\tau\widehat{\Omega}_1^D(\hat{f}^\tau) + \tau\nu\|\hat{f}^\tau\|^2_{\mathcal{H}}).$$

where, for an arbitrary $\lambda > 0$, the subdifferential of $\lambda\widehat{\Omega}_1^D$ at f is given by

$$\partial\lambda\widehat{\Omega}_1^D(f) = \{\hat{\nabla}^* v, v = (v_a)_{a=1}^d \in (\mathbb{R}^n)^d \mid v_a = \lambda\hat{D}_a f/\|\hat{D}_a f\|_n \text{ if } \|\hat{D}_a f\|_n > 0, \text{ and } \|v_a\|_n \leq \lambda \text{ otherwise}, \forall a = 1,\ldots,d\}$$

Using the above characterization and the fact that $\widehat{\mathcal{E}} + \tau\nu\|\cdot\|^2_{\mathcal{H}}$ is differentiable, the Euler equation is equivalent to

$$\hat{\nabla}^* v = -\frac{1}{2\sigma}\nabla(\widehat{\mathcal{E}} + \tau\nu\|\cdot\|^2_{\mathcal{H}})(\hat{f}^\tau),$$

for any $\sigma > 0$, and for some $v = (v_a)_{a=1}^d \in (\mathbb{R}^n)^d$ with $v = (v_a)_{a=1}^d \in (\mathbb{R}^n)^d$ such that

$$\begin{aligned} v_a &= \frac{\tau}{\sigma}\frac{\hat{D}_a\hat{f}^\tau}{\|\hat{D}_a\hat{f}^\tau\|_n} \quad \text{if} \quad \|\hat{D}_a\hat{f}^\tau\|_n > 0, \\ v_a &\in \frac{\tau}{\sigma}B_n \qquad\qquad \text{otherwise.} \end{aligned}$$

In order to prove (34), we proceed by contradiction and assume that $\|\hat{D}_a\hat{f}^\tau\|_n > 0$. This would imply $\|v_a\|_n = \tau/\sigma$, which contradicts the assumption, hence $\|\hat{D}_a\hat{f}^\tau\|_n = 0$.

We now prove (35). First, according to Definition 2 (see also Theorem 4.3 in Villa et al. (2011) and Beck and Teboulle (2009) for the case when the proximity operator is evaluated exactly), the algorithm generates by construction sequences $\tilde{f}^t$ and f^t such that

$$\tilde{f}^t - f^t - \frac{1}{2\sigma}\nabla F(\tilde{f}^t) \in \frac{1}{2\sigma}\partial_{\sigma(\varepsilon^t)^2}2\tau\widehat{\Omega}_1^D(f^t) = \frac{\tau}{\sigma}\partial_{\frac{\sigma}{2\tau}(\varepsilon^t)^2}\widehat{\Omega}_1^D(f^t).$$

where ∂_ε denotes the ε-subdifferential[13]. Plugging the definition of f^t from (29) in the above equation, we obtain $\hat{\nabla}^*\bar{v}^t \in \frac{\tau}{\sigma}\partial_{\frac{\sigma}{2\tau}(\varepsilon^t)^2}\widehat{\Omega}_1^D(f^t)$. Now, we can use a kind of *transportation formula* (Hiriart-Urruty and Lemaréchal, 1993) for the ε-subdifferential to find $\tilde{\varepsilon}^t$ such that $\hat{\nabla}^*\bar{v}^t \in \frac{\tau}{\sigma}\partial_{\frac{\sigma}{2\tau}(\tilde{\varepsilon}^t)^2}\widehat{\Omega}_1^D(f^\tau)$. By definition of ε-subdifferential:

$$\widehat{\Omega}_1^D(f) - \widehat{\Omega}_1^D(f^t) \geq \langle\frac{\sigma}{\tau}\hat{\nabla}^*\bar{v}^t, f - f^t\rangle_{\mathcal{H}} - \frac{\sigma}{2\tau}(\varepsilon^t)^2, \qquad \forall f \in \mathcal{H}.$$

Adding and subtracting $\widehat{\Omega}_1^D(\hat{f}^\tau)$ and $\langle\frac{\sigma}{\tau}\hat{\nabla}^* v^t, \hat{f}^\tau\rangle$ to the previous inequality we obtain

$$\widehat{\Omega}_1^D(f) - \widehat{\Omega}_1^D(\hat{f}^\tau) \geq \langle\frac{\sigma}{\tau}\hat{\nabla}^*\bar{v}^t, f - \hat{f}^\tau\rangle_{\mathcal{H}} - \frac{\sigma}{2\tau}(\tilde{\varepsilon}^t)^2, \quad \text{with } (\tilde{\varepsilon}^t)^2 = (\varepsilon^t)^2 + \frac{2\tau}{\sigma}\Big(\widehat{\Omega}_1^D(f^t) - \widehat{\Omega}_1^D(\hat{f}^\tau)\Big) + \langle 2\hat{\nabla}^*\bar{v}^t, f^t - \hat{f}^\tau\rangle_{\mathcal{H}}.$$

From the previous equation, using (32) we have

$$(\tilde{\varepsilon}^t)^2 = (\varepsilon^t)^2 + \sqrt{\frac{C}{\nu\tau}}\left(\frac{\tau}{\sigma}\sum_a\sqrt{\|\hat{D}_a^*\hat{D}\|} + 1\right)\frac{4}{t}, \tag{39}$$

which implies $\frac{\sigma}{\tau}\hat{\nabla}^*\bar{v}^t \in \partial_{\sigma(\tilde{\varepsilon^t})^2/2\tau}\widehat{\Omega}_1^D(\hat{f}^\tau)$. Now, relying on the structure of $\widehat{\Omega}_1^D$, it is easy to see that

$$\partial_\varepsilon\widehat{\Omega}_1^D(f) \subseteq \{\hat{\nabla}^* v, v = (v_a)_{a=1}^d \in (\mathbb{R}^n)^d \mid \|v_a\|_n \geq 1 - \varepsilon/\|\hat{D}_a f\|_n \text{ if } \|\hat{D}_a f\|_n > 0\}.$$

Thus, if $\|\hat{D}_a\hat{f}^\tau\|_n > 0$ we have $\|\bar{v}^t\| \geq \frac{\tau}{\sigma}(1 - \frac{(\tilde{\varepsilon^t})^2}{2\|\hat{D}_a\hat{f}^\tau\|_n})$. □

[13]Recall that the ε-subdifferential, ∂_ϵ, of a convex functional $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is defined as the set

$$\partial_\varepsilon\Omega(f) := \{h \in \mathcal{H} \,:\, \Omega(g) - \Omega(f) \geq \langle h, g - f\rangle_{\mathcal{H}} - \epsilon, \ \forall g \in \mathcal{H}\}, \qquad \forall f \in \mathcal{H}.$$

C Proofs of Section 4

We start proving the following preliminary probabilistic inequalities.

Lemma 1. *For $0 < \eta_1, \eta_2, \eta_3, \eta_4 \leq 1$, $n \in \mathbb{N}$, it holds*

(1) $\quad \mathrm{P}\Big(\Big| \|\mathbf{y}\|_n^2 - \int_{\mathcal{X}\times\mathcal{Y}} y^2 d\rho(x,y) \Big| \leq \epsilon(n,\eta_1) \Big) \geq 1-\eta_1$ *with* $\epsilon(n,\eta_1) = \frac{2\sqrt{2}}{\sqrt{n}} M^2 \log \frac{2}{\eta_1}$,

(2) $\quad \mathrm{P}\Big(\|\hat{S}^*\mathbf{y} - I_k^* f_\rho\|_{\mathcal{H}} \leq \epsilon(n,\eta_2) \Big) \geq 1-\eta_2$ *with* $\epsilon(n,\eta_2) = \frac{2\sqrt{2}}{\sqrt{n}} \kappa_1 M \log \frac{2}{\eta_2}$,

(3) $\quad \mathrm{P}\Big(\|\hat{S}^*\hat{S} - I_k^* I_k\| \leq \epsilon(n,\eta_3) \Big) \geq 1-\eta_3$ *with* $\epsilon(n,\eta_3) = \frac{2\sqrt{2}}{\sqrt{n}} \kappa_1^2 \log \frac{2}{\eta_3}$,

(4) $\quad \mathrm{P}\Big(\|\hat{D}_a^*\hat{D}_a - D_a^* D_a\| \leq \epsilon(n,\eta_4) \Big) \geq 1-\eta_4$ *with* $\epsilon(n,\eta_4) = \frac{2\sqrt{2}}{\sqrt{n}} \kappa_2^2 \log \frac{2}{\eta_4}$.

Proof. From standard concentration inequalities for Hilbert space valued random variables – see for example (Pinelis and Sakhanenko, 1985)– we have that, if ξ is a random variable with values in a Hilbert space $\mathcal{H}$ bounded by L and $\xi_1, \dots, \xi_n$ are n i.i.d. samples, then

$$\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}(\xi)\| \leq \epsilon(n,\eta) = \frac{2\sqrt{2}}{\sqrt{n}} L \log \frac{2}{\eta}$$

with probability at least $1-\eta$, $\eta \in [0,1]$. The proof is a direct application of the above inequalities to the random variables,

(1) $\xi = y^2$ $\quad \xi \in \mathbb{R}$ $\quad$ with $\sup_{\mathbf{z}_n} \|\xi\| \leq M^2$,
(2) $\xi = k_x y$ $\quad \xi \in \mathcal{H}\otimes\mathbb{R}$ $\quad$ with $\sup_{\mathbf{z}_n} \|\xi\| \leq \kappa_1 M$,
(3) $\xi = \langle \cdot, k_x \rangle_{\mathcal{H}} k_x$ $\quad \xi \in \mathcal{HS}(\mathcal{H})$ $\quad$ with $\sup_{\mathbf{z}_n} \|\xi\|_{\mathcal{HS}}(\mathcal{H}) \leq \kappa_1^2$,
(4) $\xi = \langle \cdot, (\partial_a k)_x \rangle_{\mathcal{H}} (\partial_a k)_x$ $\quad \xi \in \mathcal{HS}(\mathcal{H})$ $\quad$ with $\sup_{\mathbf{z}_n} \|\xi\|_{\mathcal{HS}}(\mathcal{H}) \leq \kappa_2^2$.

where $\mathcal{HS}(\mathcal{H})$, $\|\cdot\|_{\mathcal{HS}(\mathcal{H})}$ are the space of Hilbert-Schmidt operators on $\mathcal{H}$ and the corresponding norm, respectively (note that in the final bound we upper-bound the operator norm by the Hilbert-Schmidt norm). □

Proofs of the Consistency of the Regularizer. We restate Theorem 2 in an extended form.

Theorem (Theorem 2 Extended)**.** *Let $r < \infty$, then under assumption (A2), for any $\eta > 0$,*

$$\mathrm{P}\left(\sup_{\|f\|_{\mathcal{H}} \leq r} |\widehat{\Omega}_1^D(f) - \Omega_1^D(f)| \geq rd \frac{2\sqrt{2}}{(n)^{1/4}} \kappa_2 \sqrt{\log \frac{2d}{\eta}} \right) < \eta. \tag{40}$$

Consequently

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{\|f\|_{\mathcal{H}} \leq r} |\widehat{\Omega}_1^D(f) - \Omega_1^D(f)| > \epsilon \right) = 0, \qquad \forall \epsilon > 0.$$

Proof. For $f \in \mathcal{H}$ consider the following chain of inequalities,

$$\begin{aligned} |\widehat{\Omega}_1^D(f) - \Omega_1^{\mathrm{D}}(f)| &\leq \sum_{a=1}^d \Big| \|\hat{D}_a f\|_n - \|D_a f\|_{\rho_\mathcal{X}} \Big| \\ &\leq \sum_{a=1}^d \Big(\Big| \|\hat{D}_a f\|_n^2 - \|D_a f\|_{\rho_\mathcal{X}}^2 \Big| \Big)^{1/2} \\ &= \sum_{a=1}^d \Big(\Big| \langle f, (\hat{D}_a^*\hat{D}_a - {D_a}^* D_a) f \rangle_{\mathcal{H}} \Big| \Big)^{1/2} \\ &\leq \sum_{a=1}^d \|\hat{D}_a^*\hat{D}_a - {D_a}^* D_a\|^{1/2} \|f\|_{\mathcal{H}}, \end{aligned}$$

that follows from from $|\sqrt{x}-\sqrt{y}| \leq \sqrt{|x-y|}$, the definition of $\hat{D}_a, D_a$ and basic inequalities. Then, using d times inequality (d) in Lemma 1 with η/d in place of η_4, and taking the supremum on $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \leq r$, we have with probability $1-\eta$,

$$\sup_{\|f\|_{\mathcal{H}} \leq r} |\widehat{\Omega}_1^D(f) - \Omega_1^{\mathrm{D}}(f)| \leq rd\frac{2\sqrt{2}}{(n)^{1/4}}\kappa_2\sqrt{\log\frac{2d}{\eta}}.$$

The last statement of the theorem follows easily. □

Consistency Proofs. To prove Theorem 3, we need the following lemma.

Lemma 2. *Let $\eta \in (0,1]$. Under assumptions A1 and A3, we have*

$$\sup_{\|f\|_{\mathcal{H}} \leq r} |\widehat{\mathcal{E}}(f) - \mathcal{E}(f)| \leq \frac{2\sqrt{2}}{\sqrt{n}}\left(\kappa_1^2 r^2 + 2\kappa_1 M r + M^2\right)\log\frac{6}{\eta},$$

with probabilty $1-\eta$.

Proof. Recalling the definition of I_k we have that,

$$\begin{aligned}
\mathcal{E}(f) &= \int_{\mathcal{X}\times\mathcal{Y}} (I_k f(x) - y)^2 d\rho(x,y) \\
&= \int_{\mathcal{X}} (I_k f(x))^2 d\rho_{\mathcal{X}}(x) + \int_{\mathcal{X}\times\mathcal{Y}} y^2 d\rho(x,y) - 2\int_{\mathcal{X}\times\mathcal{Y}} I_k f(x) y d\rho(x,y) \\
&= \int_{\mathcal{X}} (I_k f(x))^2 d\rho_{\mathcal{X}}(x) + \int_{\mathcal{X}\times\mathcal{Y}} y^2 d\rho(x,y) - 2\int_{\mathcal{X}} I_k f(x) f_\rho(x) d\rho_{\mathcal{X}}(x) \\
&= \langle f, I_k^* I_k f\rangle_{\mathcal{H}} + \int_{\mathcal{X}\times\mathcal{Y}} y^2 d\rho(x,y) - 2\langle f, I_k^* f_\rho\rangle_{\mathcal{H}}.
\end{aligned}$$

Similarly $\widehat{\mathcal{E}}(f) = \langle f, \hat{S}^*\hat{S} f\rangle_{\mathcal{H}} + \|\mathbf{y}\|_n^2 - 2\langle f, \hat{S}^* f_\rho\rangle_{\mathcal{H}}$. Then, for all $f \in \mathcal{H}$, we have the bound

$$|\widehat{\mathcal{E}}(f) - \mathcal{E}(f)| \leq \|\hat{S}^*\hat{S} - I_k^* I_k\|\|f\|_{\mathcal{H}}^2 + 2\|\hat{S}^*\mathbf{y} - I_k^* f_\rho\|_{\mathcal{H}}\|f\|_{\mathcal{H}} + \left|\|\mathbf{y}\|_n^2 - \int_{\mathcal{X}\times\mathcal{Y}} y^2 d\rho(x,y)\right|$$

The proof follows applying Lemma 1 with probabilities $\eta_1 = \eta_2 = \eta_3 = \eta/3$. □

We now prove Theorem 3. We use the following standard result in regularization theory (see for example (Dontchev and Zolezzi, 1993)) to control the the approximation error.

Proposition 7. *Let $\tau_n \to 0$, be a positive sequence. Then we have that*

$$\mathcal{E}^{\tau_n}(f^{\tau_n}) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \to 0.$$

Proof of Theorem 3. We recall the standard sample/approximation error decomposition

$$\mathcal{E}(\hat{f}^\tau) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq |\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau)| + |\mathcal{E}^\tau(f^\tau) - \inf_{f\in\mathcal{H}} \mathcal{E}(f)| \tag{41}$$

where $\mathcal{E}^\tau(f) = \mathcal{E}(f) + 2\tau\Omega_1^{\mathrm{D}}(f) + \tau\nu\|f\|_{\mathcal{H}}^2$.

We first consider the sample error. Toward this end, we note that

$$\tau\nu\|\hat{f}^\tau\|_{\mathcal{H}}^2 \leq \widehat{\mathcal{E}}^\tau(\hat{f}^\tau) \leq \widehat{\mathcal{E}}^\tau(0) = \|\mathbf{y}\|_n^2 \Longrightarrow \|\hat{f}^\tau\|_{\mathcal{H}} \leq \frac{\|\mathbf{y}\|_n}{\sqrt{\tau\nu}} \leq \frac{M}{\sqrt{\tau\nu}},$$

and similarly $\|f^\tau\|_{\mathcal{H}} \le (\int_{\mathcal{X}} y^2 d\rho)^{1/2}/\sqrt{\tau\nu} \le \frac{M}{\sqrt{\tau\nu}}$.
We have the following bound,

$$\begin{aligned}
\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau) &\le (\mathcal{E}(\hat{f}^\tau) - \widehat{\mathcal{E}}(\hat{f}^\tau)) + \widehat{\mathcal{E}}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau) \\
&\le (\mathcal{E}(\hat{f}^\tau) - \widehat{\mathcal{E}}(\hat{f}^\tau)) + \widehat{\mathcal{E}}^\tau(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau) \\
&\le (\mathcal{E}(\hat{f}^\tau) - \widehat{\mathcal{E}}(\hat{f}^\tau)) + \widehat{\mathcal{E}}^\tau(f^\tau) - \mathcal{E}^\tau(f^\tau) \\
&\le (\mathcal{E}(\hat{f}^\tau) - \widehat{\mathcal{E}}(\hat{f}^\tau)) + (\widehat{\mathcal{E}}(f^\tau) - \mathcal{E}(f^\tau)) + \tau(\widehat{\Omega}_1^D(f^\tau) - \Omega_1^{\mathrm{D}}(f^\tau)) \\
&\le 2 \sup_{\|f\|_{\mathcal{H}} \le \frac{M}{\sqrt{\tau\nu}}} |\widehat{\mathcal{E}}(f) - \mathcal{E}(f)| + \tau \sup_{\|f\|_{\mathcal{H}} \le \frac{M}{\sqrt{\tau\nu}}} |\widehat{\Omega}_1^D(f) - \Omega_1^{\mathrm{D}}(f)|.
\end{aligned}$$

Let $\eta' \in (0,1]$. Using Lemma 2 with probability $\eta = 3\eta'/(3+d)$, and inequality (40) with $\eta = d\eta'/(3+d)$, and if η' is sufficiently small we obtain

$$\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau) \le \frac{4\sqrt{2}}{\sqrt{n}} M^2 \left(\frac{\kappa_1^2}{\tau\nu} + \frac{2\kappa_1}{\sqrt{\tau\nu}} + 1 \right) \log \frac{6+2d}{\eta'} + \tau \frac{2\sqrt{2}}{(n)^{1/4}} d \frac{M}{\sqrt{\tau\nu}} \kappa_2 \sqrt{\log \frac{6+2d}{\eta'}}.$$

with probability $1-\eta'$. Furthermore, we have the bound

$$\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau) \le c \left(\frac{M\kappa_1^2}{n^{1/2}\tau\nu} + \frac{\tau^{1/2} d\kappa_2}{n^{1/4}\sqrt{\nu}} \right) \log \frac{6+2d}{\eta'} \tag{42}$$

where c does not depend on n, τ, ν, d. The proof follows, if we plug (42) in (41) and take $\tau = \tau_n$ such that $\tau_n \to 0$ and $(\tau_n \sqrt{n})^{-1} \to 0$, since the approximation error goes to zero (using Proposition 7) and the sample error goes to zero in probability as $n \to \infty$ by (42).

□

We next consider convergence in the RKHS norm. The following result on the convergence of the approximation error is standard (Dontchev and Zolezzi, 1993).

Proposition 8. *Let $\tau_n \to 0$, be a positive sequence. Then we have that*

$$\|f_\rho^\dagger - f^{\tau_n}\|_{\mathcal{H}} \to 0.$$

We can now prove Theorem 4. The main difficulty is to control the sample error in the $\mathcal{H}$-norm. This requires showing that controlling the distance between the minima of two functionals, we can control the distance between their minimizers. Towards this end it is critical to use the results in Villa et al. (0) based on Attouch-Wetts convergence. We need to recall some useful quantities. Given two subsets A and B in a metric space $(\mathcal{H}, d)$, the excess of A on B is defined as $e(A,B) := \sup_{f\in A} d(f,B)$, with the convention that $e(\emptyset, B) = 0$ for every B. Localizing the definition of the excess we get the quantity $e_r(A,B) : e(A \cap B(0,r), B)$ for each ball $B(0,r)$ of radius r centered at the origin. The *r-epi-distance* between two subsets A and B of $\mathcal{H}$, is denoted by $d_r(A,B)$ and is defined as

$$d_r(A,B) := \max\{e_r(A,B), e_r(B,A)\}.$$

The notion of epi-distance can be extended to any two functionals $F, G : \mathcal{H} \to \mathbb{R}$ by

$$d_r(G,F) := d_r(\mathrm{epi}(G), \mathrm{epi}(F)),$$

where for any $F : \mathcal{H} \to \mathbb{R}$, epi(F) denotes the epigraph of F defined as

$$\mathrm{epi}(F) := \{(f,\alpha), F(f) \le \alpha\}.$$

We are now ready to prove Theorem 4, which we present here in an extended form

Theorem (Theorem 4 Extended)**.** *Under assumptions A1, A2 and A3,*

$$\mathrm{P}\Big(\|\hat{f}^\tau - f_\rho^\dagger\|_{\mathcal{H}} \ge A(n,\tau)^{1/2} + \|f^\tau - f_\rho^\dagger\|_{\mathcal{H}}, \Big) < \eta \tag{43}$$

where

$$A(n,\tau) = 4\sqrt{2}M\left(\frac{4\kappa_1^2 M}{\sqrt{n}\tau^2\nu^2} + \frac{4\kappa_1}{\sqrt{n}\tau\nu\sqrt{\tau\nu}} + \frac{1}{\sqrt{n}\tau\nu} + \frac{2d\kappa_2}{n^{1/4}\nu\sqrt{\tau\nu}}\right)$$

for $0 < \eta \leq 1$. *Moreover,*

$$\lim_{n\to\infty} \mathrm{P}\Big(\|\hat{f}^{\tau_n} - f_\rho^\dagger\|_{\mathcal{H}} \geq \epsilon\Big) = 0, \qquad \forall \epsilon > 0,$$

for any τ_n *such that* $\tau_n \to 0$ *and* $(\sqrt{n}\tau_n^2)^{-1} \to 0$.

Proof of Theorem 4. We consider the decomposition of $\|\hat{f}^\tau - f_\rho^\dagger\|_{\mathcal{H}}$ into a sample and approximation term,

$$\|\hat{f}^\tau - f_\rho^\dagger\|_{\mathcal{H}} \leq \|\hat{f}^\tau - f^\tau\|_{\mathcal{H}} + \|f^\tau - f_\rho^\dagger\|_{\mathcal{H}}. \tag{44}$$

From Theorem 2.6 in Villa et al. (0) we have that

$$\psi_{\tau\nu}^\diamond(\|\hat{f}^\tau - f^\tau\|_{\mathcal{H}}) \leq 4d_{M/\sqrt{\tau\nu}}(t_{\mathcal{E}^\tau}\mathcal{E}^\tau, t_{\mathcal{E}^\tau}\widehat{\mathcal{E}}^\tau)$$

where $\psi_{\tau\nu}^\diamond(t) := \inf\{\frac{\tau\nu}{2}s^2 + |t-s| : s \in [0,+\infty)\}$, and $t_{\mathcal{E}^\tau}$ is the translation map defined as

$$t_{\mathcal{E}^\tau}G(f) = G(f + f^\tau) - \mathcal{E}^\tau(f^\tau)$$

for all $G : \mathcal{H} \to \mathbb{R}$.

From Theorem 2.7 in Villa et al. (0), we have that

$$d_{M/\sqrt{\tau\nu}}(t_{\mathcal{E}^\tau}\mathcal{E}^\tau, t_{\mathcal{E}^\tau}\widehat{\mathcal{E}}^\tau) \leq \sup_{\|f\|_{\mathcal{H}} \leq M/\sqrt{\tau\nu}} |t_{\mathcal{E}^\tau}\mathcal{E}^\tau(f) - t_{\mathcal{E}^\tau}\widehat{\mathcal{E}}^\tau(f)|.$$

We have the bound,

$$\begin{aligned}\sup_{\|f\|_{\mathcal{H}} \leq M/\sqrt{\tau\nu}} |t_{\mathcal{E}^\tau}\mathcal{E}^\tau(f) - t_{\mathcal{E}^\tau}\widehat{\mathcal{E}}^\tau(f)| &\leq \sup_{\|f\|_{\mathcal{H}} \leq M/\sqrt{\tau\nu} + \|f^\tau\|_{\mathcal{H}}} |\mathcal{E}^\tau(f) - \widehat{\mathcal{E}}^\tau(f)| \\ &\leq \sup_{\|f\|_{\mathcal{H}} \leq 2M/\sqrt{\tau\nu}} |\mathcal{E}(f) - \widehat{\mathcal{E}}(f)| + \tau \sup_{\|f\|_{\mathcal{H}} \leq 2M/\sqrt{\tau\nu}} |\Omega_1^{\mathrm{D}}(f) - \widehat{\Omega}_1^D(f)|.\end{aligned}$$

Using Theorem 2 (equation (40)) and Lemma 2 we obtain with probability $1-\eta'$, if η' is small enough,

$$\begin{aligned}d_{M/\sqrt{\tau\nu}}(t_{\mathcal{E}^\tau}\mathcal{E}^\tau, t_{\mathcal{E}^\tau}\widehat{\mathcal{E}}^\tau) &\leq \frac{2\sqrt{2}}{\sqrt{n}}\left(\kappa_1^2\frac{4M^2}{\tau\nu} + 4\kappa_1\frac{M^2}{\sqrt{\tau\nu}} + M^2\right)\log\frac{6+2d}{\eta'} + \tau\frac{2M}{\sqrt{\tau\nu}}d\frac{2\sqrt{2}}{n^{1/4}}\kappa_2\sqrt{\log\frac{6+2d}{\eta'}} \\ &\leq 2\sqrt{2}M\left(\frac{4\kappa_1^2 M}{\sqrt{n}\tau\nu} + \frac{4\kappa_1 M}{\sqrt{n}\sqrt{\tau\nu}} + \frac{M}{\sqrt{n}} + \tau\frac{2d\kappa_2}{n^{1/4}\sqrt{\tau\nu}}\right)\log\frac{6+2d}{\eta'}.\end{aligned} \tag{45}$$

From the definition of $\psi_{\tau\nu}^\diamond$ it is possible to see that we can write explicitly $(\psi_{\tau\nu}^\diamond)^{-1}$ as

$$(\psi_{\tau\nu}^\diamond)^{-1}(y) = \begin{cases}\sqrt{\frac{2y}{\tau\nu}} & \text{if } y < \frac{1}{2\tau\nu} \\ y + \frac{1}{2\tau} & \text{otherwise.}\end{cases}$$

Since $\tau = \tau_n \to 0$ by assumption, for sufficiently large n, the bound in (45) is smaller than $1/2\tau\nu$, and we obtain that with probability $1-\eta'$,

$$\|\hat{f}^\tau - f^\tau\|_{\mathcal{H}} \leq \left(4\sqrt{2}M\left(\frac{4\kappa_1^2 M}{\sqrt{n}\tau^2\nu^2} + \frac{4\kappa_1}{\sqrt{n}\tau\nu\sqrt{\tau\nu}} + \frac{1}{\sqrt{n}\tau\nu} + \frac{2d\kappa_2}{n^{1/4}\nu\sqrt{\tau\nu}}\right)\right)^{1/2}\sqrt{\log\frac{6+2d}{\eta'}}. \tag{46}$$

If we now plug (46) in (44) we obtain the first part of the proof. The rest of the proof follows by taking the limit $n \to \infty$, and by observing that, if one chooses $\tau = \tau_n$ such that $\tau_n \to 0$ and $(\tau_n^2\sqrt{n})^{-1} \to 0$, the assumption of Proposition 8 is satisfied and the bound in (46) goes to 0, so that the limit of the sum of the sample and approximation terms goes to 0.

□

Proofs of the Selection properties. In order to prove our main selection result, we will need the following lemma.

Lemma 3. *Under assumptions A1, A2 and A3 and defining $A(n,\tau)$ as in Theorem 4 extended, we have, for all $a = 1,\dots,d$ and for all $\epsilon > 0$,*

$$\mathrm{P}\Big(\Big|\|\hat{D}_a\hat{f}^\tau\|_n^2 - \|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2\Big| \geq \epsilon\Big) < (6+2d)\mathrm{exp}\Big(-\frac{\epsilon - b(\tau)}{a(n,\tau)}\Big),$$

where $a(n,\tau) = 2\max\{\frac{2\sqrt{2}M^2\kappa_2^2}{\sqrt{n}\tau\nu}, 2\kappa_2^2 A(n,\tau)\}$ and $\lim_{\tau\to 0} b(\tau) = 0$.

Proof. We have the following set of inequalities

$$\begin{aligned}\Big|\|\hat{D}_a\hat{f}^\tau\|_n^2 - \|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2\Big| &= |\langle\hat{f}^\tau, \hat{D}_a^*\hat{D}_a\hat{f}^\tau\rangle_{\mathcal{H}} - \langle f_\rho^\dagger, D_a^* D_a f_\rho^\dagger\rangle_{\mathcal{H}} + \\ &\quad \langle\hat{f}^\tau, D_a^* D_a\hat{f}^\tau\rangle_{\mathcal{H}} - \langle\hat{f}^\tau, D_a^* D_a\hat{f}^\tau\rangle_{\mathcal{H}} + \\ &\quad \langle f_\rho^\dagger, D_a^* D_a\hat{f}^\tau\rangle_{\mathcal{H}} - \langle f_\rho^\dagger, D_a^* D_a\hat{f}^\tau\rangle_{\mathcal{H}}| \\ &= \Big|\langle\hat{f}^\tau, (\hat{D}_a^*\hat{D}_a - D_a^* D_a)\hat{f}^\tau\rangle_{\mathcal{H}} + \langle\hat{f}^\tau - f_\rho^\dagger, D_a^* D_a(\hat{f}^\tau - f_\rho^\dagger)\rangle_{\mathcal{H}}\Big| \\ &\leq \|\hat{D}_a^*\hat{D}_a - D_a^* D_a\|\frac{M^2}{\tau\nu} + \kappa_2^2\|\hat{f}^\tau - f_\rho^\dagger\|_{\mathcal{H}}^2 \\ &\leq \|\hat{D}_a^*\hat{D}_a - D_a^* D_a\|\frac{M^2}{\tau\nu} + 2\kappa_2^2\|\hat{f}^\tau - f^\tau\|_{\mathcal{H}}^2 + 2\kappa_2^2\|f^\tau - f_\rho^\dagger\|_{\mathcal{H}}^2.\end{aligned}$$

Using Theorem 4 extended, equation (46), and Lemma 1 with probability $\eta_4 = \eta/(3+d)$, we obtain with probability $1-\eta$

$$\Big|\|\hat{D}_a\hat{f}^\tau\|_n^2 - \|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2\Big| \leq \frac{2\sqrt{2}M^2\kappa_2^2}{\sqrt{n}\tau\nu}\log\frac{6+2d}{\eta} + 2\kappa_2^2 A(n,\tau)\log\frac{6+2d}{\eta} + 2\kappa_2^2\|f^\tau - f_\rho^\dagger\|_{\mathcal{H}}^2.$$

We can further write

$$\Big|\|\hat{D}_a\hat{f}^\tau\|_n^2 - \|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2\Big| \leq a(n,\tau)\log\frac{6+2d}{\eta} + b(\tau),$$

where $a(n,\tau) = 2\max\{\frac{2\sqrt{2}M^2\kappa_2^2}{\sqrt{n}\tau\nu}, 2\kappa_2^2 A(n,\tau)\}$ and $\lim_{\tau\to 0} b(\tau) = 0$ according to Proposition 8. The proof follows by writing $\epsilon = a(n,\tau)\log\frac{6+2d}{\eta} + b(\tau)$ and inverting it with respect to η.

□

Finally we can prove Theorem 5.

Proof of Theorem 5. We have

$$\mathrm{P}\Big(R_\rho \subseteq \hat{R}^\tau\Big) = 1 - \mathrm{P}\Big(R_\rho \not\subseteq \hat{R}^\tau\Big) = 1 - \mathrm{P}\left(\bigcup_{a\in R_\rho}\{a \notin \hat{R}^\tau\}\right) \geq 1 - \sum_{a\in R_\rho}\mathrm{P}\Big(a \notin \hat{R}^\tau\Big)$$

Let us now estimate $\mathrm{P}\Big(a \notin \hat{R}^\tau\Big)$ or equivalently $\mathrm{P}\Big(a \in \hat{R}^\tau\Big) = \mathrm{P}\Big(\|\hat{D}_a\hat{f}^\tau\|_n^2 > 0\Big)$, for $a \in R_\rho$. Let $C < \min_{a\in R_\rho}\|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2$. From Lemma 3, there exist $a(n,\tau)$ and $b(\tau)$ satisfying $\lim_{\tau\to 0} b(\tau) = 0$, such that

$$\Big|\|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2 - \|\hat{D}_a\hat{f}^\tau\|_n^2\Big| \leq \epsilon$$

with probability $1 - (6+2d)\mathrm{exp}\Big(-\frac{\epsilon - b(\tau)}{a(n,\tau)}\Big)$, for all $a = 1,\dots,d$. Therefore, for $\epsilon = C$, for $a \in R_\rho$, it holds

$$\|\hat{D}_a\hat{f}^\tau\|_n^2 \geq \|D_a f_\rho^\dagger\|_{\rho\mathcal{X}}^2 - C \gneq 0.$$

with probability $1 - (6+2d)\mathrm{exp}\Big(-\frac{C - b(\tau)}{a(n,\tau)}\Big)$. We than have

$$\mathrm{P}\Big(a \in \hat{R}^\tau\Big) = \mathrm{P}\Big(\|\hat{D}_a\hat{f}^\tau\|_n^2 > 0\Big) \geq 1 - (6+2d)\mathrm{exp}\left(-\frac{C - b(\tau)}{a(n,\tau)}\right),$$

so that $\mathrm{P}\Big(a \notin \hat{R}^{\tau} \Big) \leq (6+2d)\exp\left(-\frac{C-b(\tau)}{a(n,\tau)}\right)$. Finally, if we let $\tau = \tau_n$ satisfying the assumption, we have $\lim_n b(\tau_n) \to 0$, $\lim_n a(n,\tau_n) \to 0$, so that

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{P}\Big(R_\rho \subseteq \hat{R}^{\tau_n} \Big) &\geq \lim_{n\to\infty} \left[1 - |R_\rho|(6+2d)\exp\left(-\frac{C-c(\tau_n)}{a(n,\tau_n)}\right)\right] \\
&= 1 - |R_\rho|(6+2d) \lim_{n\to\infty} \exp\left(-\frac{C-b(\tau_n)}{a(n,\tau_n)}\right) \\
&= 1.
\end{aligned}$$

□

D Some Connections with Structured Sparsity Regularization Approaches

As already mentioned in the Introduction, a number of authors consider a dictionary of nonlinear features $(\phi_\gamma(x))_{\gamma\in\Gamma}$ – so that any function can be identified with a coefficient vector defining its expansion with respect to the dictionary, $f_\beta = \sum_{\gamma\in\Gamma} \beta_\gamma \phi_\gamma$ – and use a (structured) sparsity based algorithm to select the relevant elements of the dictionary. We recall that given a finite dictionary $(\phi_\gamma(x))_{\gamma\in\Gamma}$ and a function $f_\beta(x) = \sum_{\gamma\in\Gamma} \beta_\gamma \phi_\gamma(x)$, a structured sparsity penalty is of the form:

$$\Omega^{\text{Str.Sp.}}(\beta) = \sum_{a=1}^{p} \|A^a \beta\|_{\mathbb{R}^{d_a}}$$

where $A^a : \mathbb{R}^{|\Gamma|} \to \mathbb{R}^{d_a}$ are linear bounded operators for $a = 1, \ldots, p$. For more details see Zhao et al. (2009), Jacob et al. (2009), Negahban et al. (2009), Jenatton et al. (2010), Mosci et al. (2010) and references therein.

ℓ^1 regularization on the feature space The simplest of such approaches simply amounts to penalizing the ℓ^1 norm of the coefficient vector,

$$\Omega^{\ell^1\text{-features}}(\beta) = \sum_{\gamma\in\Gamma} |\beta_\gamma|$$

and thus induces selection with respect to the features rather than on the input variables.

Additive models Another approach is that of considering additive models where the hypotheses space is the space of linear combinations of nonlinear functions, each one depending on a single variable only, $\mathcal{H} = \{f = \sum_{a=1}^{d} f_a(x_a), f_a \in \mathcal{H}_a\}$. Variable selection can thus be enforced via Multiple Kernel Learning algorithms (see for example Bach et al., 2004; Bach, 2008; Ravikumar et al., 2008), with the penalty

$$\Omega^{\text{MKL}}(f) = \sum_{a=1}^{d} \|f_a\|_{\mathcal{H}_a}.$$

If each space $\mathcal{H}_a$ is finite dimensional, it admits a finite set of features $(\phi^a_\gamma)_{\gamma\in\Gamma^{(a)}}$, so that each function in $\mathcal{H}$ is identified with the set of coefficients $\beta = (\beta^1, \ldots, \beta^d) \in \mathbb{R}^{\sum_a |\Gamma_a|}$ since $f_\beta = \sum_{a=1}^{d} \sum_{\gamma\in\Gamma^{(a)}} \beta^a_\gamma \phi^a_\gamma$, and the MKL penalty can be viewed as a structured sparsity penalty

$$\Omega^{\text{MKL}}(f_\beta) = \sum_{a=1}^{d} \|\beta^a\|_{\mathbb{R}^{|\Gamma^{(a)}|}} = \sum_{a=1}^{d} \|A^a \beta\|_{\mathbb{R}^{|\Gamma^{(a)}|}}$$

where $A^a : \mathbb{R}^{|\Gamma|} \to \mathbb{R}^{|\Gamma^{(a)}|}$ extracts the elements of β corresponding to space $\mathcal{H}_a$, that is $[A^a]_{\gamma,\gamma'} = \delta_{\gamma,\gamma'}$ if $\gamma, \gamma' \in \Gamma^{(a)}$ and 0 otherwise.

Hierarchical kernel learning A more advanced approach would be to consider dictionaries encoding more complex interactions among the variables (Lin and Zhang, 2006). For example, one could consider the features defined by a second-degree polynomial kernel in order to take into account 2-way interactions among variables. The shortcoming of this approach is that the size of the dictionary grows more than exponentially as one considers higher orders of interactions. For instance, for 40 variables with a polynomial kernel of degree 4 – in this case the interaction degree is at most 4 –, the dictionary size is already highly above 100,000. Recently, Bach (2009) showed that it is still possible to learn with such dictionaries if the atoms of the dictionary are embedded into a directed acyclic graph, and the admissible sparsity patterns coincides with suitably chosen subnetworks of the graph. The penalty proposed in Bach (2009) for nonlinear variable selection is given by

$$\Omega^{\mathrm{HKL}}(\beta) = \sum_{a=1}^{p} \|A^a \beta\|_{\mathbb{R}^{d_a}}$$

where p is the number of admissible subnetworks of the graph, and A^a extracts the d_a atoms belonging to the a-th subnetwork.

Our approach differs from these latter works since we do not try to design dictionaries encoding variables interactions but we use partial derivatives to derive a new regularizer that induces a different form of sparsity. Interestingly it is possible to show that, if RKHS $\mathcal{H}$ admits a *finite* set of nonlinear features, the proposed penalty can be written as a structured sparsity penalty with suitable sparsity patterns and data dependent weights.

In fact, if the hypotheses space $\mathcal{H}$ is a space of sufficiently smooth functions the partial derivative is a bounded linear operator in $\mathcal{H}$. Therefore, given a finite set of nonlinear features $(\phi_\gamma(x))_{\gamma\in\Gamma}$ associate to $\mathcal{H}$, there exists $\mathrm{D}^a \in \mathbb{R}^{|\Gamma|} \otimes \mathbb{R}^{|\Gamma|}$ such that

$$\frac{\partial \phi_\gamma(x)}{\partial x^a} = \sum_{\gamma'\in\Gamma} \mathrm{D}^a_{\gamma\gamma'} \phi_{\gamma'}(x),$$

and, for all $f_\beta(x) = \sum_{\gamma\in\Gamma} \beta_\gamma \phi_\gamma(x)$ we can write

$$\left.\frac{\partial f_\beta(x)}{\partial x^a}\right|_{x=x_i} = \sum_{\gamma,\in\Gamma} \beta_\gamma \left.\frac{\partial \phi_\gamma(x)}{\partial x^a}\right|_{x=x_i} = \sum_{\gamma,\in\Gamma} \beta_\gamma \underbrace{\sum_{\gamma'\in\Gamma} \mathrm{D}^a_{\gamma\gamma'} \phi_{\gamma'}(x_i)}_{[A^a_{\mathbf{z}_n}]_{i,\gamma}} := (A^a_{\mathbf{z}_n}\beta)_i$$

where $A^a_{\mathbf{z}_n} : \mathbb{R}^{|\Gamma|} \to \mathbb{R}^n$ are linear bounded operators for all $a = 1, \ldots, d$. The DENOVAS penalty thus writes

$$\widehat{\Omega}^D_1(f_\beta) = \frac{1}{\sqrt{n}} \sum_{a=1}^{d} \|A^a_{\mathbf{z}_n}\beta\|_{\mathbb{R}^n}.$$

The above expression for the penalty also highlights the differences of the proposed approach with respect to the regularization approaches to nonlinear variable selection described above. In fact, the ℓ^1 penalty on the feature space is totally unaware of the dependancy of f on the original variables, and therefore performs selection with respect to the features, not the variables. The MKL penalty differs from $\widehat{\Omega}^D_1$ in that it does not couple the features corresponding to different variables, since the dictionary has no interaction terms by construction. Finally, the HKL penalty proposed in Bach (2009) is substantially different from $\widehat{\Omega}^D_1$, since, while for HKL $[A^a]_{\gamma,\gamma'}$ belongs to $\{0,1\}$ and is thus independent of $\mathbf{z}_n$, in $\widehat{\Omega}^D_1$ the $A^a_{\mathbf{z}_n}$ are data dependent, precisely they depend on the value of the partial derivatives of the kernel evaluated in the training set points. Furthermore both ℓ^1 regularization and HKL can only deal with finite dimensional spaces, and therefore the corresponding regularized solution are not universally consistent. The MKL penalty is the only one that can potentially deal with infinite dimensional spaces, however the hypotheses space is never dense in $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ due to its univariate nature, and universal consistency is thus precluded.