Journal of Machine Learning Research 14 (2013) 1665-1714 Submitted 9/11; Revised 8/12; Published 7/13

Nonparametric Sparsity and Regularization

Lorenzo Rosasco[*] LROSASCO@MIT.EDU
DIBRIS
University of Genova
Via Dodecaneso 35
16146 Genova, ITALY

Silvia Villa SILVIA.VILLA@IIT.IT
Istituto Italiano di Tecnologia
Via Morego 30
16163 Genova, ITALY

Sofia Mosci SOFIA.MOSCI@UNIGE.IT
DIBRIS
University of Genova
Via Dodecaneso 35
16146 Genova, ITALY

Matteo Santoro MATTEO.SANTORO@IIT.IT
Istituto Italiano di Tecnologia
Via Morego 30
16163 Genova, ITALY

Alessandro Verri ALESSANDRO.VERRI@UNIGE.IT
DIBRIS
University of Genova
Via Dodecaneso 35
16146 Genova, ITALY

Editor: John Lafferty

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 a new notion of nonparametric sparsity 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 methods.

Keywords: sparsity, nonparametric, variable selection, regularization, proximal methods, RKHS

∗. Also at Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy and Massachusetts Institute of Technology, Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

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. It is known that learning in such a high dimensional regime is feasible only if the quantity to be estimated satisfies some regularity assumptions (Devroye et al., 1996). 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 *relevant* variables. Detecting such variables is the problem of variable selection.

Largely motivated by recent advances in compressed sensing (Candès and Tao, 2006; Donoho, 2006), 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 and 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. Only a few works, that we discuss in details in Section 2, have considered notions of sparsity beyond additive models.

In this paper, we propose a new approach based on the idea that the importance of a variable, while learning a non-linear functional relation, can be captured by the corresponding partial derivative. This observation suggests a way to define a new notion of nonparametric sparsity and a corresponding regularizer which favors functions where most partial derivatives are essentially zero. The question is how to make this intuition precise and how to derive a feasible computational learning scheme. 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) provides us with tools to answer both questions. In fact, partial derivatives in a RKHS are bounded linear functionals and hence have a suitable representation that allows efficient computations. Moreover, the norm in the RKHS provides a natural further regularizer ensuring stable behavior of the empirical,

derivative based penalty. Our contribution is threefold. First, we propose a new notion of sparsity and discuss a corresponding regularization scheme using concept from the theory of reproducing kernel Hilbert spaces. Second, since the proposed algorithm corresponds to the minimization of a convex, but not differentiable functional, we develop a suitable optimization procedure relying on forward-backward splitting and proximal methods. Third, we study properties of the proposed methods both in theory, in terms of statistical consistency, and in practice, by means of an extensive set of experiments.

Some preliminary results have appeared in a short conference version of this 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 3 we discuss our approach and present the main results in the paper. In Section 4 we discuss the computational aspects of the method. In Section 5 we prove consistency results. In Section 6 we provide an extensive empirical analysis. In Section 7 we conclude with a summary of our study and a discussion of future work. Finally, a list of symbols and notations can be found at the end of the paper.

2. Problem Setting and Previous Work

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 satisfies some suitable prior assumption (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. Estimating the set R_ρ of *relevant* variables is the problem of variable selection.

2.1 Linear and Additive Models

The sparsity requirement 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 approach has been studied extensively and is now 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 the 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 spaces $\mathcal{H}_a$, $a = 1, \ldots, d$. In this case, $\Omega_0(f) = \#\{a \in \{1, \ldots, d\} \ : \ \|f_a\| \neq 0\}$ and $\Omega_1(f) = \sum_{a=1}^d \|f_a\|$ are the analogous of the zero and

ℓ^1 norm, 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. Considering additive models limits the way in which the variables can 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 simplest 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 $\Omega_0(f) = \#\{a = 1,\dots,d \ : \ \|f_a\| \neq 0\} + \#\{(a,b) \ : \ a < b, \ \|f_{a,b}\| \neq 0\}$ 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 number of 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).

2.2 Nonparametric Approaches

The above discussion naturally raises the question:

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

Few papers have considered this question. Here we discuss in some more details Lafferty and Wasserman (2008), Bertin and Lecué (2008), Miller and Hall (2010) and Comminges and Dalalyan (2012), to which we also refer for further references.

The first three papers (Lafferty and Wasserman, 2008; Bertin and Lecué, 2008; Miller and Hall, 2010) follow similar approaches focusing on the pointwise estimation of the regression function and of the relevant variables. The basic idea is to start from a locally linear (or polynomial) pointwise estimator $f_n(x)$ at a point x obtained from the minimizer of

$$\frac{1}{n}\sum_{i=1}^n (y_i - \langle w, x_i - x\rangle_{\mathbb{R}^d})^2 K_H(x_i - x) \tag{1}$$

where K_H is a localizing window function depending on a matrix (or a vector) H of smoothing parameters. Different techniques are used to (locally) select variables. In the RODEO algorithm (Lafferty and Wasserman, 2008), the localizing window function depends on one smoothing parameter per variable and the partial derivative of the local estimator with respect to the smoothing parameter is used to select variables. In Bertin and Lecué (2008), selection is considering a *local lasso*, that is an ℓ_1 to the local empirical risk functional (1). In the LABAVS algorithm discussed in Miller and Hall (2010) several variable selection criterion are discussed including the local lasso, hard thresholding, and backward stepwise approach. The above approaches typically lead to cumbersome computations and do not scale well with the dimensionality of the space and with the number of relevant variables.

Indeed, in all the above works the emphasis is in the theoretical analysis quantifying the estimation error of the proposed methods. It is shown in Lafferty and Wasserman (2008) that the RODEO algorithm is a nearly optimal pointwise estimator of the regression function, under assumption on the marginal distribution and the regression function. These results are further improved in Bertin and Lecué (2008) where optimal rates are derived under milder assumptions and sparsistency (the

recovery of R_ρ) is also studied. Uniform error estimates are derived in Miller and Hall (2010) (see Section 2.6 in Miller and Hall (2010) for further discussions and comparison). More recently, an estimator based on the comparison of some well chosen empirical Fourier coefficients to a prescribed significance level is described and studied in Comminges and Dalalyan (2012) where a careful statistical analysis is proposed considering different regimes for n, d and d^*, where d^* is the cardinality of R_ρ. Finally, in a slightly different context, DeVore et al. (2011) studies the related problem of determining the number of function values at adaptively chosen points that are needed in order to correctly estimate the set of globally relevant variables.

3. Sparsity Beyond Linear Models

In this section we present our approach and summarize our main contributions.

3.1 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]

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

This observation can be developed into a new notion of sparsity and corresponding regularization scheme that we study in the rest of the paper. We note that regularization using derivatives is not new. Indeed, 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 uses derivatives to define regular functions. 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 to define a proper (global) notion of sparsity. If the partial derivative is continuous[2] then a natural idea is to consider

$$\left\|\frac{\partial f}{\partial x^a}\right\|_{\rho_X} = \sqrt{\int_X \left(\frac{\partial f(x)}{\partial x^a}\right)^2 d\rho_X(x)}. \tag{2}$$

where ρ_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 nonparametric sparsity for a smooth, non-linear function f is captured by the following functional

$$\Omega_0^{\mathrm{D}}(f) = \#\left\{a = 1,\ldots,d \;:\; \left\|\frac{\partial f}{\partial x^a}\right\|_{\rho_X} \neq 0\right\}, \tag{3}$$

1. In order for the partial derivatives to be defined at all points we always assume that the closure of X coincides with the closure of its interior.
2. In the following, see Remark 2, we will see that further appropriate regularity properties on f are needed depending on whether the support of ρ_X is connected or not.

and the corresponding relaxation is

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

The above functionals encode the notion of sparsity that we are going to consider. While for linear models, the above definition subsumes the classic notion of sparsity, the above definition is non constrained to any (parametric) additive model.

Second, since ρ_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 (2) by the data-driven regularizer,

$$\widehat{\Omega}_1^D(f) = \sum_{a=1}^{d} \left\| \frac{\partial f}{\partial x^a} \right\|_n. \tag{4}$$

While the above quantity is a natural estimate of (2) 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,

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

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-posedness and generalization, as we discuss in Section 5. 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

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \tau \left(2 \sum_{a=1}^{d} \left\| \frac{\partial f}{\partial x^a} \right\|_n + \nu \|f\|_{\mathcal{H}}^2 \right). \tag{5}$$

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 (Comparison with Derivative Based Regularizers) *It is perhaps useful to remark the difference between the regularizer we propose and other derivative based regularizers. 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

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.

sparsity property. A different derivative based regularizer is given by $\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{d}\left|\frac{\partial f(x_i)}{\partial x^a}\right|$. *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 uses 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*

$$\left\|\frac{\partial f}{\partial x_a}\right\|_{\rho_X}=0 \;\Rightarrow\; f \text{ is constant with respect to } x_a.$$

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* $\mathrm{supp}\rho_X$ *to be connected. If the latter assumption is not satisfied the situation is more complicated, as the following example shows. Suppose that* ρ_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* ρ_X*, although the variable* x *is relevant. To avoid such pathological situations when* $\mathrm{supp}\rho_X$ *is not connected in* $\mathbb{R}^d$ *we need to impose more stringent regularity assumptions that 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.*

3.2 Main Results

We summarize our main contributions.

1. Our main contribution is the analysis of the minimization of (5) and the derivation of a provably convergent iterative optimization procedure. We begin by extending the representer theorem (Wahba, 1990) and show that the minimizer of (5) has the finite dimensional representation
$$\hat{f}^{\tau}(x) = \sum_{i=1}^{n} \frac{1}{n} \alpha_i k(x_i, x) + \sum_{i=1}^{n} \sum_{a=1}^{d} \frac{1}{n} \beta_{ai} \left. \frac{\partial k(s,x)}{\partial s^a} \right|_{s=x_i},$$
with $\alpha, (\beta_{ai})_{i=1}^{n} \in \mathbb{R}^n$ for all $a = 1, \ldots, d$. Then, we show that the coefficients in the expansion can be computed using forwards-backward splitting and proximal methods (Combettes and Wajs, 2005; Beck and Teboulle, 2009). More precisely, we present a fast forward-backward splitting algorithm, in which 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 (and convergence rates) 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
$$\lim_{n \to \infty} \mathrm{P}\big(\mathcal{E}(\hat{f}^{\tau_n}) - \mathcal{E}(f_\rho) > \varepsilon \big) = 0$$
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
$$\lim_{n \to \infty} \mathrm{P}\big(\hat{R}^{\tau_n} \subseteq R_\rho \big) = 1.$$

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 3.2, 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.

4. Computational Analysis

In this section we study the minimization of the functional (5).

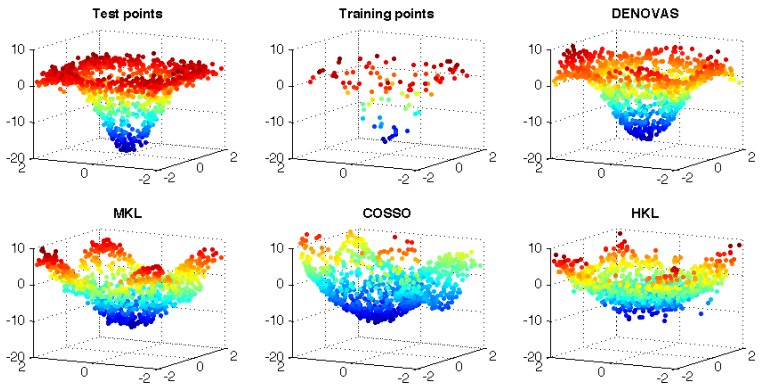

Figure 2: Comparison of predictions for a radial function of 2 out of 20 variables (the 18 irrelevant variables are not shown in the figure). In the upper left plot is depicted the value of the function on the test points (left), the noisy training points (center), the values predicted for the test points 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).

4.1 Basic Assumptions

We first begin by listing some basic conditions that we assume to hold throughout the paper.

We let ρ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. A training set $\mathbf{z}_n = (\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{i=1}^n$ is a sample from ρ^n. We consider a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Aronszajn, 1950) and the associated reproducing kernel Hilbert space $\mathcal{H}$. We assume ρ and k to satisfy the following assumptions.

[A1] *There exists* $\kappa_1 < \infty$ *such that* $\sup_{x \in X} \|t \mapsto k(x,t)\|_{\mathcal{H}} < \kappa_1$.

[A2] *The kernel* k *is* $\mathcal{C}^2(\mathcal{X} \times \mathcal{X})$ *and there exists* $\kappa_2 < \infty$ *such that for all* $a = 1, \ldots, d$ *we have* $\sup_{x \in X} \|t \mapsto \frac{\partial k(s,x)}{\partial s^a}\Big|_{s=t}\|_{\mathcal{H}} < \kappa_2$.

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

4.2 Computing the Regularized Solution

We start our analysis discussing how to compute efficiently a regularized solution of the functional

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

where $\widehat{\Omega}_1^D(f)$ is defined in (4). 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 and uniqueness of the 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 leading to 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 for which we can prove convergence. It is possible to see 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 $a = 1,\dots,d$ k_x is the function $t \mapsto k(x,t)$, and $(\partial_a k)_x$ denotes partial derivatives of the kernel, see (19). The main outcome of our analysis is that the coefficients α and β can be provably computed through an iterative procedure. To describe the algorithm we need some notation. For all $a,b = 1,\dots,d$, 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 $i,j = 1,\dots,n$. 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

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

and the $nd \times nd$ matrix

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

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

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

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

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

- *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: $\alpha^0 = \alpha^1 = 0$, $\beta^0 = \beta^1 = 0$, $s_1 = 1$, $\vec{v}^1 = 0$, $t = 1$
while `convergence not reached` **do**
$t = t + 1$

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

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

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

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

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

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

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

end while
return (α^t, β^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 three remarks. First, the proposed procedure requires the choice of appropriate stopping rules for the inner and outer loops, which will be discussed later, and of the step-sizes σ and η. The simple a priori choice $\sigma = \|K\| + \tau\nu$, $\eta = \|L\|$ ensures convergence, as discussed in the Section 4.5, 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, in Section 4.6 we discuss a principled way to select variable using the norm of the coefficients $(\vec{v}_a^t)_{a=1}^d$.

4.3 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 scheme.

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 $x \in \mathcal{X}$ and

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

for all $f \in \mathcal{H}, x \in \mathcal{X}$. 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 $\alpha_i \in \mathbb{R}, x_i \in \mathcal{X}$ 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. 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, which is true thanks to Assumption (A1).

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 $x \in \mathcal{X}$ and remarkably

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

for $a = 1, \ldots, d$, $x \in X$. 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 $\mathbf{x} = (x_1, \ldots, x_n)$,

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

where $a = 1, \ldots, d$, $i = 1, \ldots, n$. 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, since assumption [A2] is satisfied. 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 3 *The minimizer of (6) can be written as*

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

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

Proposition 3 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.

4.3.1 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 adjoint, that is, the transpose A^T.

4.4 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 by 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) and Jenatton et al. (2010).

4.4.1 FORWARD-BACKWARD SPLITTING ALGORITHMS

The functional $\widehat{\mathcal{E}}^\tau$ is the sum of the two terms $F(\cdot) = \widehat{\mathcal{E}}(\cdot) + \tau\nu\|\cdot\|^2_{\mathcal{H}}$ and $2\tau\widehat{\Omega}^D_1$. 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,

$$f^t = \mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}^D_1}\Big(\tilde{f}^t - \frac{1}{2\sigma}\nabla F(\tilde{f}^t)\Big), \tag{21}$$

$$\tilde{f}^t = c_{1,t} f^{t-1} + c_{2,t} f^{t-2} \tag{22}$$

where $t \geq 2$, $\tilde{f}^0 = f^0 = f^1 \in \mathcal{H}$ is an arbitrary initialization, $c_{1,t}, c_{2,t}$ are suitably chosen positive sequences, and $\mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}^D_1} : \mathcal{H} \to \mathcal{H}$ is the proximity operator (Moreau, 1965) defined by,

$$\mathrm{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}^D_1}(f) = \operatorname*{argmin}_{g\in\mathcal{H}} \left(\frac{\tau}{\sigma}\widehat{\Omega}^D_1(g) + \frac{1}{2}\|f-g\|^2_{\mathcal{H}} \right).$$

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, in our setting the computation of the proximity operator of $\widehat{\Omega}^D_1$ is not trivial and will be discussed in the next paragraph. Here we briefly recall the main properties of the iteration (21), (22) 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 (21), sometimes called FISTA (fast iterative shrinkage thresholding algorithm (Tseng, 2010; Beck and Teboulle, 2009)), which is obtained by setting $s_0 = 1$,

$$s_t = \frac{1}{2}\left(1+\sqrt{1+4s_{t-1}^2}\right), \qquad c_{1,t} = 1+\frac{s_{t-1}-1}{s_t}, \text{ and } \quad c_{2,t} = \frac{1-s_{t-1}}{s_t}. \tag{23}$$

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 the Lipschitz constant L can be non trivial. Theorems 3.1 and 4.4 in Beck and Teboulle (2009) show that the iterative procedure (21) 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 $\widehat{\Omega}_1^D$.

4.4.2 COMPUTING THE PROXIMITY OPERATOR

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

$$\text{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D} = I - \pi_{\frac{\tau}{\sigma}\mathcal{C}_n}, \tag{24}$$

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

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

where

$$\bar{v} \in \operatorname*{argmin}_{v \in \frac{\tau}{\sigma}B_n^d} \|f - \hat{\nabla}^* v\|_{\mathcal{H}}^2. \tag{25}$$

Being a convex constrained problem, (25) 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 (21), again. In fact we can fix an arbitrary initialization $v^0 \in \mathbb{R}^{nd}$ and consider,

$$v^{q+1} = \pi_{\frac{\tau}{\sigma}B_n^d}\left(v^q - \frac{1}{\eta}\hat{\nabla}(\hat{\nabla}^* v^q - f)\right), \tag{26}$$

for a suitable choice of η. In particular, we note 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 (25) on the sequence generated via (26) is guaranteed. Moreover, thanks to the special structure of the minimization problem in (25), it is possible to prove (see Combettes et al., 2010; Mosci et al., 2010) that

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

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

4.5 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

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

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

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

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. (see also Salzo and Villa, 2012) 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 result can be adapted to our setting to prove the following theorem, as shown in Appendix B.

Theorem 4 *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* (27) *with* $\vec{v}^t$ *computed through algorithm* (27). *There exists* $\bar{q}$ *such that if* $\vec{v}^t = v^q$, *for* $q > \bar{q}$, *the following condition is satisfied*

$$\frac{2\tau}{\sigma}\widehat{\Omega}_1^D(f^t) - 2\langle\hat{\nabla}^*\vec{v}^t, f^t\rangle \leq \varepsilon_t^2 . \tag{28}$$

Moreover, 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 $\nu > 0$,

$$\|f^t - \hat{f}^\tau\|_{\mathcal{H}} \leq \frac{2}{t}\sqrt{\frac{C}{\nu\tau}}. \tag{29}$$

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 (29), 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, condition (28) represents an implementable stopping criterion for the inner iteration, once that the representer theorem is proved and is always asymptotically satisfied by sequences minimizing the dual problem (25) (see Proposition 17). Further comments on the stopping rule are given in Section 4.6. 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 $\mathcal{C}_n$. More generally, each algorithm generating a feasible sequence $\{v^q\}$ for the dual problem and minimizing the dual function (25) gives admissible approximations of the proximal point (see Villa et al.).

4.6 Further Algorithmic Considerations

We conclude discussing several practical aspects of the proposed method.

4.6.1 THE FINITE DIMENSIONAL IMPLEMENTATION

We start by showing how the representer theorem can be used, together with the iterations described by (27) and (27), to derive Algorithm 1. This is summarized in the following proposition.

Proposition 5 *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* (27) *is given by*

$$f^t = \frac{1}{n}\sum_{i=1}^{n} \alpha_i^t k_{x_i} + \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{d} \beta_{a,i}^t (\partial_a k)_{x_i} \tag{30}$$

with α^t and β^t defined by the updating rules (14-16), where $\vec{v}^t$ in (16) *can be estimated, starting from any $v^0 \in \mathbb{R}^{nd}$, 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 (29) and (28) suggest the following choices[5]

$$\|f^t - f^{t-1}\|_{\mathcal{H}} \leq \varepsilon^{(\mathrm{ext})} \quad \text{and} \quad \frac{2\tau}{\sigma}\widehat{\Omega}_1^D(f^t) - 2\langle \hat{\nabla}^* v^q, f^t \rangle \leq \varepsilon^{(\mathrm{int})}.$$

As a direct consequence of (30) and using the definition of matrices $\mathrm{K}, \mathrm{Z}, \mathrm{L}$, these quantities can be easily computed as

$$\begin{aligned} \|f^t - f^{t-1}\|_{\mathcal{H}}^2 &= \langle \delta\alpha, \mathrm{K}\delta\alpha \rangle_n + 2\langle \delta\alpha, \mathrm{Z}\delta\beta \rangle_n + \langle \delta\beta, \mathrm{L}\delta\beta \rangle_n, \\ \frac{2\tau}{\sigma}\widehat{\Omega}_1^D(f^t) - 2\langle \hat{\nabla}^* v^q, f^t \rangle &= \sum_{a=1}^{d} \left(\frac{2\tau}{\sigma} \|Z_a\alpha^t + L_a\beta^t\|_n - 2\langle v^q, Z_a^T\alpha^t + L_a\beta^t \rangle_n \right). \end{aligned}$$

where we defined $\delta\alpha = \alpha^t - \alpha^{t-1}$ and $\delta\beta = \beta^t - \beta^{t-1}$. Also note that, according to Theorem 4, $\varepsilon^{(\mathrm{int})}$ must depend on the outer iteration as $\varepsilon^{(\mathrm{int})} = \varepsilon_t^2 \sim t^{-2l}$, $l > 3/2$.

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

4.6.2 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 6 *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

$$\|v_a\|_n < \frac{\tau}{\sigma} \Rightarrow \|\hat{D}_a\hat{f}^\tau\|_n = 0. \tag{31}$$

5. In practice we often use a stopping rule where the tolerance is scaled with the current iterate, $\|f^t - f^{t-1}\|_{\mathcal{H}} \leq \varepsilon^{(\mathrm{ext})}\|f^t\|_{\mathcal{H}}$ and $\frac{2\tau}{\sigma}\widehat{\Omega}_1^D(f^t) - 2\langle \hat{\nabla}^* v^q, f^t \rangle \leq \varepsilon^{(\mathrm{int})}\|\hat{\nabla}^* v^q\|_{\mathcal{H}}$.

Moreover, if $\vec{v}^t$ is given by Algorithm 1 with the inner iteration stopped when the assumptions of Theorem 4 are met, then there exists $\tilde{\varepsilon}_t > 0$ (precisely defined in (37)*) depending on the tolerance ε^t used in the inner iteration and satisfying $\lim_{t\to+\infty} \tilde{\varepsilon}_t = 0$, such that if $m := \min\{\|\hat{D}_a \hat{f}^\tau\|_n : a \in \{1,\dots,d\} \text{s.t.} \|\hat{D}_a \hat{f}^\tau\|_n > 0\}$, then*

$$\|\vec{v}_a^t\|_n \geq \frac{\tau}{\sigma} - \frac{(\tilde{\varepsilon}_t)^2}{2m} \Rightarrow \|\hat{D}_a \hat{f}^\tau\|_n = 0. \tag{32}$$

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 statement follows by observing that, as $\hat{\nabla}^* v$ belongs to the subdifferential of $\widehat{\Omega}_1^D$ at $\hat{f}^\tau$, $\hat{\nabla}^* \vec{v}^t$ 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 $(f^t, \vec{v}^t)$ evaluated via Algorithm 1, we can thus consider to be irrelevant the variables such that $\|\vec{v}_a^t\|_n < \tau/\sigma - (\tilde{\varepsilon}^t)^2/(2m)$. Note that the explicit form of $\tilde{\varepsilon}^t$ is given in (37)).

5. Consistency for Learning and Variable Selection

In this section we study the consistency properties of our method.

5.1 Consistency

As we discussed in Section 3.1, though in practice we consider the regularizer $\widehat{\Omega}_1^D$ defined in (4), ideally we would be interested into $\Omega_1^{\mathrm{D}}(f) = \sum_{a=1}^d \|D_a f\|_{\rho_X}$, $f \in \mathcal{H}$. 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 7 *Let $r < \infty$, then under assumption (A2)*

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

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 (5). 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 8 *Under assumptions (A1), (A2) and (A3), recalling that $\mathcal{E}(f) = \int (y - f(x))^2 \, d\rho(x,y)$,*

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

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

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

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 continuous counterpart. Here we need Theorem 7 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 8 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 lower semicontinuous, 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 9 *Under assumptions (A1), (A2) and (A3), we have*

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

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., 2012) 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). We underline that our result is an asymptotic one, although it would be interesting to get an explicit learning rate, as we discuss in Section 5.3.

5.2 Selection Properties

We next consider the selection properties of our method. Following Equation (3), we start by giving the definition of relevant/irrelevant variables and sparsity in our context.

Definition 10 *We say that a variable $a = 1,\dots,d$ is* irrelevant *with respect to ρ for a differentiable function f, if the corresponding partial derivative $D_a f$ is zero ρ_X-almost everywhere, and relevant otherwise. In other words the set of relevant variables is*

$$R_f := \{a \in \{1,\dots,d\} \mid \|D_a f\|_{\rho_X} > 0\}.$$

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

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

Theorem 11 *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 (sometimes called sparsistency), requires further analysis that we postpone to a future work (see the discussion in the next subsection).

5.3 Learning Rates and Sparsity

The analysis in the previous sections is asymptotic, so it is natural to ask about the finite sample behavior of the proposed method, and in particular about the implication of the sparsity assumption. Indeed, for a variety of additive models it is possible to prove that the sample complexity (the number of samples needed to achieve a given error with a specified probability) depends linearly on the sparsity level and in a much milder way to the total number of variables, for example, logarithmically (Bühlmann and van de Geer, 2011). Proving similar results in our setting is considerably more complex and in this section we discuss the main sources of possible challenges.

Towards this end, it is interesting to contrast the form of our regularizer to that of structured sparsity penalties for which sparsity results can be derived. Inspecting the proof in Appendix C, one can see that it possible to define a suitable family of operators $V_j, \hat{V}_j : \mathcal{H} \to \mathcal{H}$, with $j = 1,\dots,d$, such that

$$\Omega_1^{\mathrm{D}}(f) = \sum_{j=1}^d \|V_j f\|_{\mathcal{H}}, \qquad \widehat{\Omega}_1^D(f) = \sum_{j=1}^d \|\hat{V}_j f\|_{\mathcal{H}}. \tag{33}$$

The operators $(V_j)_j$ are positive and self-adjoint and so are the operators $(\hat{V}_j)_j$. The latter can be shown to be stochastic approximation of the operators $(V_j)_j$, in the sense that the equalities $\mathbb{E}[\hat{V}_j^2] = V_j^2$ hold true for all $j = 1, \dots, d$.

It is interesting to compare the above expression to the one for the group lasso penalty, where for a given linear model, the coefficients are assumed to be divided in groups, only few of which are predictive. More precisely, given a collection of groups of indices $\mathcal{G} = \{G_1, \dots, G_r\}$, which forms a partition of the set $\{1, \dots, p\}$, and a linear model $f(x) = \langle \beta, x \rangle_{\mathbb{R}^p}$, the group lasso penalty is obtained by considering

$$\Omega^{GL}(\beta) = \sum_{\gamma=1}^{r} \|\beta_{|G_\gamma}\|_{\mathbb{R}^{|G_\gamma|}},$$

where, for each γ, $\beta_{|G_\gamma}$ is the $|G_\gamma|$ dimensional vector obtained restricting a vector β to the indices in G_γ. If we let P_γ be the orthogonal projection on the subspace of $\mathbb{R}^d$ corresponding G_γ-th group of indices, we have that $\langle P_\gamma \beta, P'_\gamma \beta' \rangle = 0$ for all $\gamma, \gamma' \in \Gamma$ and $\beta, \beta' \in \mathbb{R}^p$, since the groups form a partition of $\{1, \dots, p\}$. Then it is possible to rewrite the group lasso penalty as

$$\Omega^{GL}(\beta) = \sum_{\gamma=1}^{r} \|P_\gamma \beta\|_{\mathbb{R}^{|G_\gamma|}}.$$

The above idea can be extended to an infinite dimensional setting to obtain multiple kernel learning (MKL). Let $\mathcal{H}$ be a (reproducing kernel) Hilbert space which is the sum of Γ disjoint (reproducing kernel) Hilbert spaces $(\mathcal{H}_\gamma, \|\cdot\|_\gamma)_{\gamma \in \Gamma}$, and $P_\gamma : \mathcal{H} \to \mathcal{H}_\gamma$ the projections of $\mathcal{H}$ onto $\mathcal{H}_\gamma$, then MKL is induced by the penalty

$$\Omega^{MKL}(f) = \sum_{\gamma \in \Gamma} \|P_\gamma f\|_\gamma.$$

When compared to our derivative based penalty, see (33), one can notice at least two source of difficulties:

1. the operators $(V_j)_j$ are not projections and no simple relation exists among their ranges,
2. in practice we have only access to the empirical estimates $(\hat{V}_j)_j$.

Indeed, structured sparsity model induced by more complex index sets have been considered, see, for example, Jenatton et al. (2010), but the penalties are still induced by operators which are orthogonal projections. Interestingly, a class of penalties induced by a (possibly countable) family of bounded operators $\mathcal{V} = \{V_\gamma\}_{\gamma \in \Gamma}$—not necessarily projections—has been considered in Maurer and Pontil (2012). This class of penalties can be written as

$$\Omega^{(\mathcal{V})}(f) = \inf\{\sum_{\gamma \in \Gamma} \|f_\gamma\| \mid f_\gamma \in \mathcal{H}, \sum_{\gamma \in \Gamma} V_\gamma f_\gamma = f\}.$$

It is easy to see that the above penalty does not include the regularizer (33) as a special case.

In conclusion, rewriting our derivative based regularizer as in (33) highlights similarity and differences with respect to previously studied sparsity methods: indeed many of these methods are induced by families of operators. On the other hand, typically, the operators are assumed to satisfy stringent assumptions which do not hold true in our case. Moreover in our case one would have to overcome the difficulties arising from the random estimation of the operators. These interesting questions are outside of the scope of this paper, will be the subject of future work.

6. Empirical Analysis

The content of this section is divided into three parts. First, we describe the choice of tuning parameters. Second, we study the properties of the proposed method on simulated data under different parameter settings, and third, 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.

6.1 Choice of Tuning Parameters

When using Algorithm 1, once the parameters n and ν are fixed, we evaluate the optimal value of the regularization parameter τ via hold out validation on an independent validation set of $nval = n$ samples. The choice of the parameter ν, and its influence on the estimator is discussed in the next section.

Since we use an iterative procedure to compute the solution, the output of our algorithm will not be sparse in general and a selection criterion is needed. In Section 4.6 we discussed a principled way to select variable using the norm of the coefficients $(\vec{v}_a^t)_{a=1}^d$.

When using MKL, ℓ^1 regularization, and RLS we used hold out validation to set the regularization parameters, while for COSSO and HKL we used the choices suggested by the authors.

6.2 Analysis of Our Method

In this subsection we empirically investigate the sensitivity of the proposed algorithm to the model's parameters.

6.2.1 ROLE OF THE SMOOTHNESS ENFORCING PENALTY $\nu\|\cdot\|_{\mathcal{H}}^2$

From a theoretical stand point we have shown that ν has to be nonzero, in order for the proposed regularization problem (5) 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 ρ_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.

Adaption To The Marginal Distribution We consider a toy problem in 2 dimensions, where the support of the marginal distribution $\rho_X(x^1, x^2)$ 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

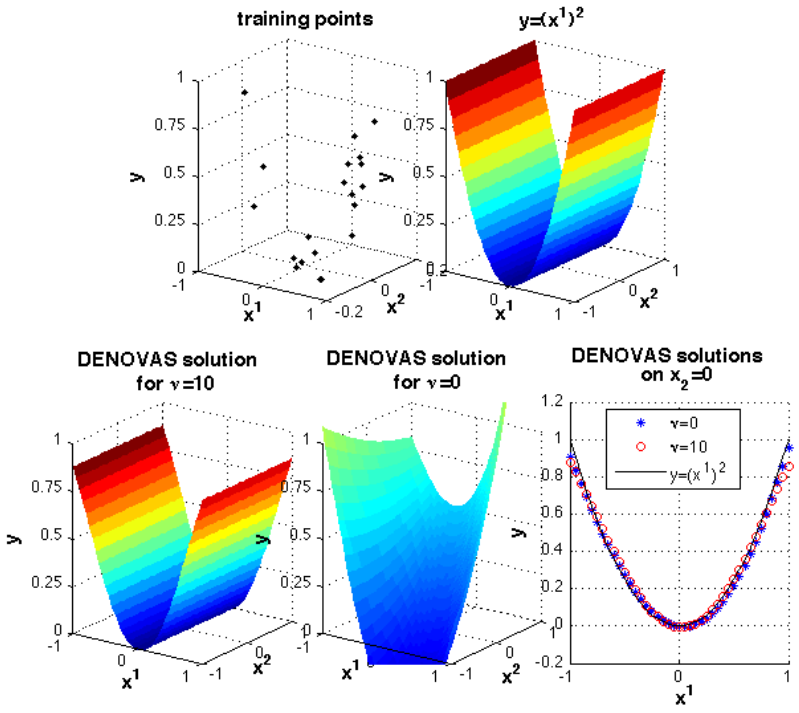

Figure 3: Effect of the combined regularization $\widehat{\Omega}_1^D(\cdot)+\nu\|\cdot\|_{\mathcal{H}}^2$ on a toy problem in $\mathbb{R}^2$ where the support of marginal distribution approximately coincides with the coordinate subspace $x^2=0$. The output labels are drawn from $y=(x^1)^2+w$, with $w\sim\mathcal{N}(0,0.1)$.

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 ρ_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 $\text{supp}(\rho_X)$, and the regularization term $\tau\widehat{\Omega}_1^D$ alone does not penalizes variations out of $\text{supp}(\rho_X)$.

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 50 to 120 with steps of 10. We then apply our method with polynomial kernel of degree $p=4$, letting vary ν in $\{0.5,1,5,20,100\}$. 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 20 repetitions.

In Figure 6.2.1 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 slight decrease in prediction and selection performance. In terms of computing time, the higher the smoothness parameter the better the performance.

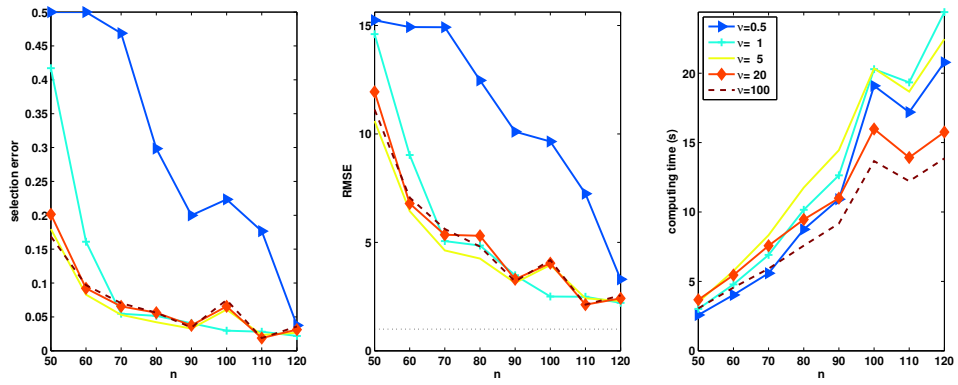

Figure 4: Selection error (left), prediction error (center), and computing time (right) versus n for different values of ν. The points correspond to the mean over the repetitions. The dotted line represents the white noise standard deviation. In the left figure the curves corresponding to $\nu = 5$,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

6.2.2 VARYING THE MODEL'S PARAMETERS

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, that is, $y = \lambda f(x) + w$, 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.

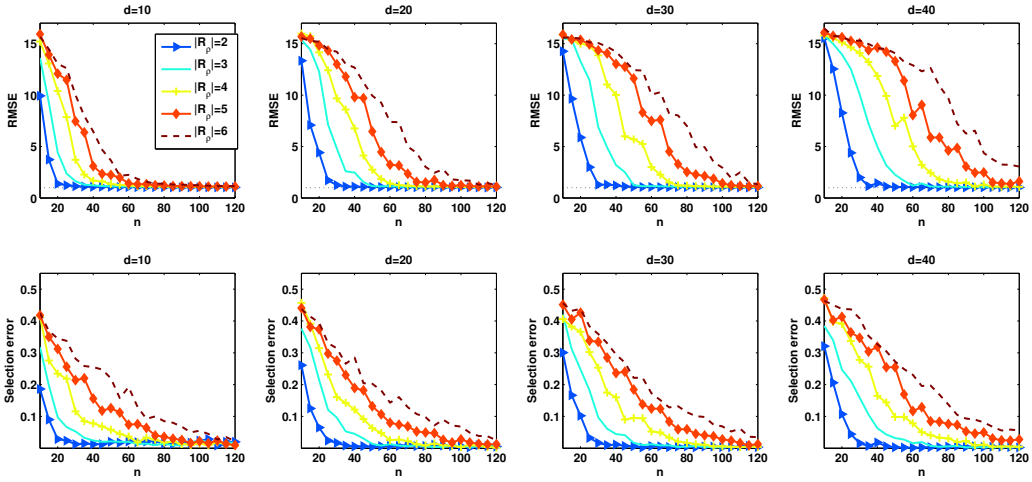

Figure 5: Prediction error (top) and selection error (bottom) versus n for different values of d and number of relevant variables ($|R_\rho|$). The points correspond to the means over the repetitions. The dotted line represents the white noise standard deviation.

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 6.2.2, 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 better visualize the dependence of the selection performance with respect to d and $|R_\rho|$, in Figure 6.2.2 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 6.2.2, 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

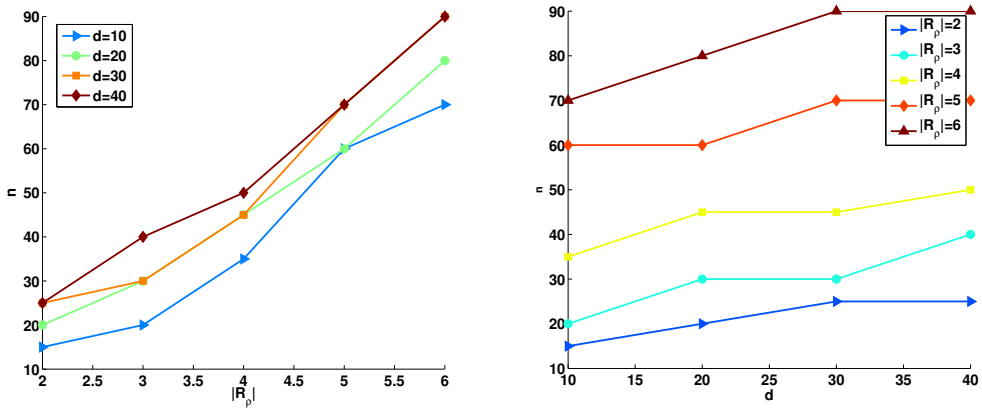

Figure 6: Minimum number of input points (n) necessary to achieve 10% of average selection error versus the number of relevant variables $|R_\rho|$ for different values of d (left), and versus d for different values of $|R_\rho|$ (right).

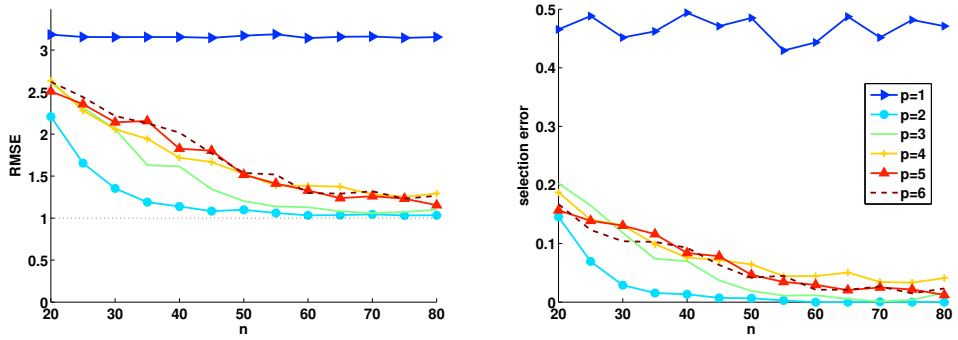

Figure 7: Prediction error (left) and selection error (right) versus n for different values of the polynomial kernel degree (p). The points correspond to the means over the repetitions. The dotted line represents the white noise standard deviation.

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, 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

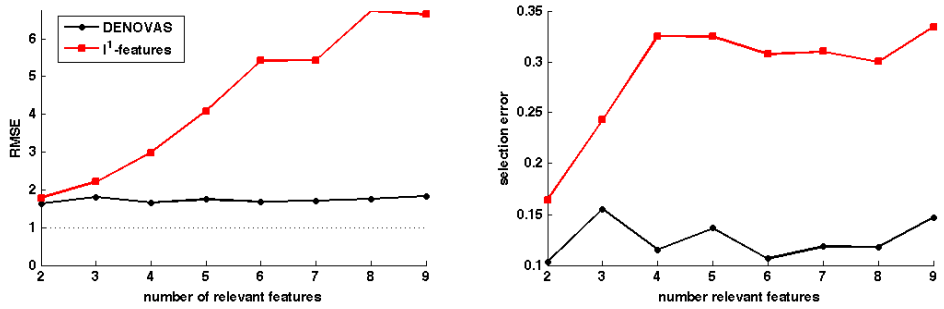

Figure 8: Prediction error (left) selection error (right) versus the number of relevant features. The points correspond to the means over the repetitions. The dotted line represents the white noise standard deviation.

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 6.2.2, 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.

6.3 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).

6.3.1 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 .

6.3.2 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, $y = f(x^1, x^2, x^3, x^4) + \varepsilon$, where epsilon is a white noise, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and σ is chosen so that the signal to noise ratio is 15:1. We consider the 4 models described in Table 6.3.2.

	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	$y = \frac{1}{\pi}((x^1)^2 + (x^2)^2) e^{-((x^1)^2+(x^2)^2)}$

Table 1: Synthetic data design

For model selection and testing we follow the same protocol described at the beginning of Section 6, 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 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

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.

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 6.3.2.

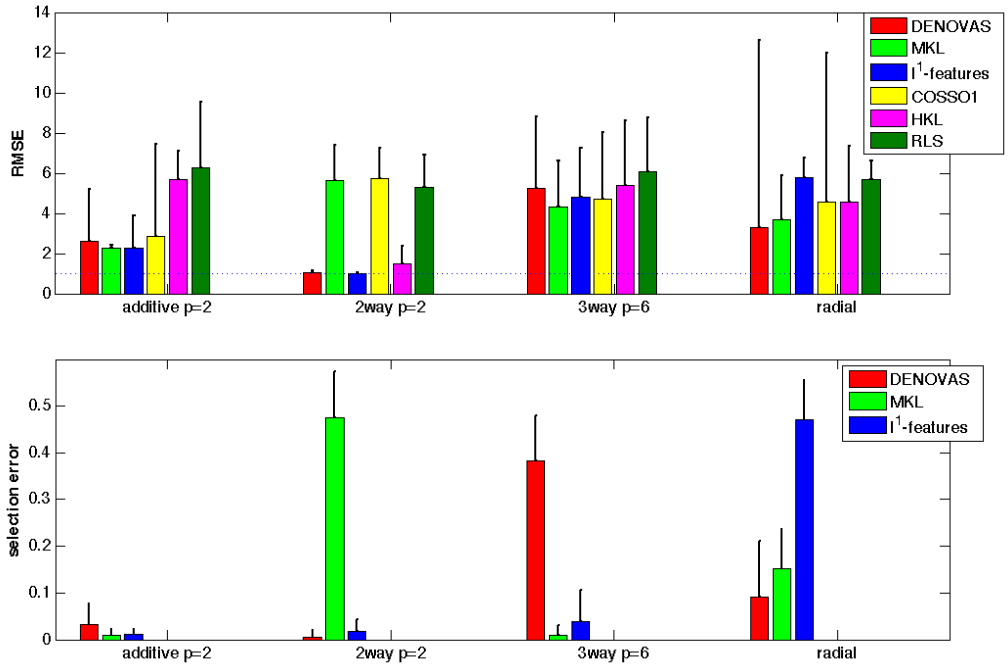

Figure 9: Prediction error (top) and fraction of selected variables (bottom) on synthetic data for the proposed method (DENOVAS), multiple kernel learning for additive models (MKL), ℓ^1 regularization on the feature space associated to a polynomial kernel (ℓ^1-features), COSSO with 1-way interactions (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 3, where we plotted in Figure 3.2 the regression function and its estimates obtained with the different algorithms for one of the 20 repetitions.

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 Section 6.3.3.

6.3.3 REAL DATA

We consider the 7 benchmark data sets described in Table 6.3.3. We build training and validation

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

Table 2: Real data sets

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 Section 6.3.1. with the validation protocol described in Section 6. 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 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.

7. 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

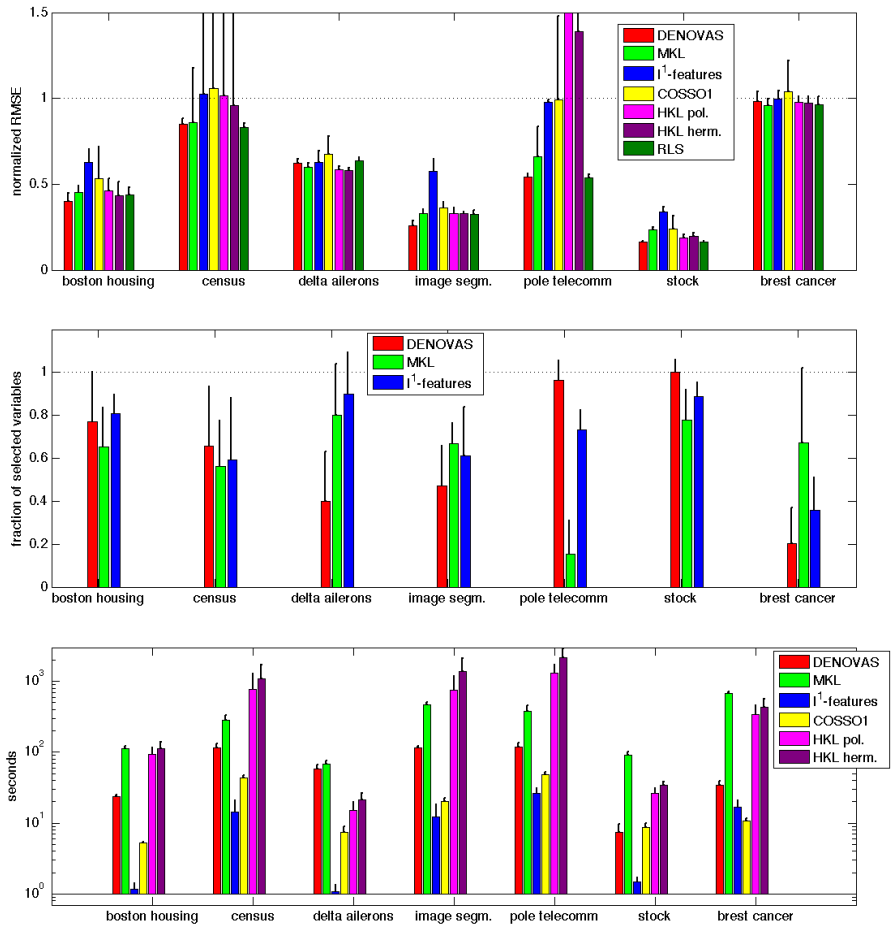

Figure 10: Prediction error (top) and fraction of selected variables (center) and computing time (bottom) on real data for the proposed method (DENOVAS), multiple kernel learning for univariate additive functions (MKL), ℓ^1 regularization on the feature space associated to a polynomial kernel (ℓ^1-features), 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. such errors were still too large to report 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. 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 are 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.

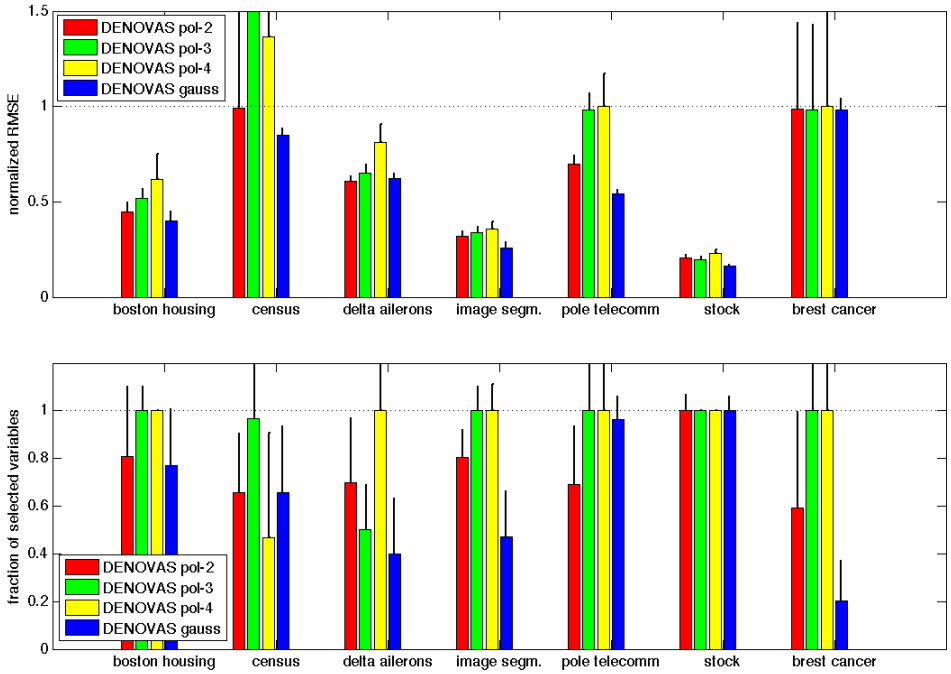

Figure 11: Prediction error (top) and fraction of selected variables (bottom) on real data for our method with different kernels: polynomial kernel of degree $p = 1, 2$ and 3 (DENOVAS pol-p), and Gaussian kernel (DENOVAS gauss). Error bars represent the standard deviations. In order to make the prediction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

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 nonparametric 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. It substantially differs with respect to previous approaches based on local polynomial regression (Lafferty and Wasserman, 2008; Bertin and Lecué, 2008; Miller and Hall, 2010), since it is a regularization scheme directly performing global variable selection. The RKHSs' machinery allows on the one hand to find a computationally efficient algorithmic solution, and on the other hand to consider very general probability distributions ρ, which are not required to have a positive

density with respect to the Lebesgue measure (differently from Comminges and Dalalyan, 2012). Several research directions are yet to be explored.

- From a theoretical point of view it would be interesting to further analyzing 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 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.

Acknowledgments

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 and Guillame Obozinski 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.

Appendix 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 $a = 1,\dots d$, 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 12 *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 13 *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 $a,b = 1,\dots,d$.

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 3 in a slightly more abstract form and give its proof.

Proposition 3 (Extended) *The minimizer of (6) satisfies $\hat{f}^\tau \in Range(\hat{S}^*) + Range(\hat{\nabla}^*)$. Henceforth it satisfies the following representer theorem*

$$\hat{f}^\tau = \hat{S}^*\alpha + \hat{\nabla}^*\beta = \sum_{i=1}^n \frac{1}{n}\alpha_i k_{x_i} + \sum_{i=1}^n \sum_{a=1}^d \frac{1}{n}\beta_{a,i}(\partial_a k)_{x_i} \tag{34}$$

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

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

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

which is clearly minimized by $f^\perp = 0$. The second equality in (34) 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* (30) *are explicitly 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*

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

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

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

Appendix B. Proofs of Section 4

In this appendix we collect the proofs related to the derivation of the iterative procedure given in Algorithm 1. Theorem 4 is a consequence of the general results about convergence of accelerated and inexact FB-splitting algorithms in Villa et al.. 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 14 *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*

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

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

We will need the following results from Villa et al..

Theorem 15 *Consider the following inexact version of the accelerated FB-algorithm in* (21) *with* $c_{1,t}$ *and* $c_{2,t}$ *as in* (23)

$$f^t \approxeq_{\varepsilon_t} prox_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D}\Big(\Big(I-\frac{1}{2\sigma}\nabla F\Big)(c_{1,t}f^{t-1}+c_{2,t}f^{t-2})\Big). \tag{35}$$

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 16 *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* $\partial\omega(0)$ *is bounded. Then for any* $f \in \mathcal{H}$ *and any* $v \in \partial\omega(0)$ *such that*

$$2\lambda\omega(B(f-\lambda B^* v)) - 2\langle \lambda B^* v, f\rangle \leq \varepsilon^2 \tag{36}$$

it holds

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

Proposition 17 *Under the same assumptions of Proposition 16, for any* $f \in \mathcal{H}$*,* $\varepsilon > 0$*, and sequence* $\{v^q\}$ *minimizing the dual problem*

$$\|f - \lambda B^* v\|_{\mathcal{H}} + i_S(v),$$

where i_S *is the indicator function of S, there exists* $\bar{q}$ *such that for any* $q > \bar{q}$*,* $f - \lambda B^* v^q$ *satisfies condition* (36)*.*

Proof [Proof of Theorem 4] Since the the regularizer $\widehat{\Omega}_1^D$ can be written as a composition of $\omega \circ B$, with $B = \hat{\nabla}$ and $\omega : \mathbb{R}^d \to [0,+\infty)$, $\omega(v) = \sum_{a=1}^d \|v_a\|_n$ and $\{v^q\}$ is a minimizing sequence for problem (25), Proposition 17 ensures that v^q satisfies condition (28) if q is big enough. Therefore, the sequence $\hat{\nabla}^* v^q$ obtained from v^q generates, via (27), admissible approximations of $\operatorname{prox}_{\frac{\tau}{\sigma}\widehat{\Omega}_1^D}$.

Therefore, if ε_t is such that $\varepsilon_t \sim 1/t^l$ with $l > 3/2$, Theorem 15 implies that the inexact version of the FB-splitting algorithm in (27) shares the $1/t^2$ convergence rate. Equation (29) directly follows from the definition of strong convexity,

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

■

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

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

Proof [Proof of Proposition 5] 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 13. 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

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

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 (30), first note that, as we have done in Proposition 3 extended, (30) 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 (21), (27), (24) it follows that f^t can be expressed as:

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

and the proposition is proved, letting $\tilde{\alpha}^t$, $\tilde{\beta}^t$ and $\bar{v}^t$ as in Equations (14), (16) and (25).

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 (34) in (27) to obtain (15). ■

Proof [Proof of Proposition 6] Since $\hat{f}^\tau$ is the unique minimizer of the functional $\widehat{\mathcal{E}}^\tau$, it satisfies the Euler equation for $\widehat{\mathcal{E}}^\tau$

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

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

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

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 (31), 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 (32). First, according to Definition 14 (see also Theorem 4.3 in Villa et al. 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 (27) in the above equation, we obtain $\hat{\nabla}^*\vec{v}^t \in \frac{\tau}{\sigma}\partial_{\frac{\sigma}{2\tau}(\varepsilon_t)^2}\widehat{\Omega}_1^D(f^t)$. 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 $\hat{\nabla}^*\vec{v}^t \in \frac{\tau}{\sigma}\partial_{\frac{\sigma}{2\tau}(\tilde{\varepsilon}_t)^2}\widehat{\Omega}_1^D(f^\tau)$. By definition of ε-subdifferential:

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

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

$$\widehat{\Omega}_1^D(f) - \widehat{\Omega}_1^D(\hat{f}^\tau) \geq \langle\frac{\sigma}{\tau}\hat{\nabla}^*\vec{v}^t, f - \hat{f}^\tau\rangle_{\mathcal{H}} - \frac{\sigma}{2\tau}(\tilde{\varepsilon}_t)^2,$$

with

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

From the previous equation, using (29) we have

$$(\tilde{\varepsilon}_t)^2 \leq (\varepsilon_t)^2 + \sqrt{\frac{C}{\nu\tau}}\left(\frac{\tau}{\sigma}\sum_a \sqrt{\|\hat{D}_a^*\hat{D}\|} + 1\right)\frac{4}{t}, \tag{37}$$

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

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

which implies $\frac{\sigma}{\tau}\hat{\nabla}^* \vec{v}^t \in \partial_{\sigma(\tilde{\varepsilon}_t)^2/2\tau}\widehat{\Omega}_1^D(\hat{f}^\tau)$. 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 $\|\vec{v}^t\|_n \geq \frac{\tau}{\sigma}(1 - \frac{(\tilde{\varepsilon}_t)^2}{2\|\hat{D}_a \hat{f}^\tau\|_n})$. ■

Appendix C. Proofs of Section 5

We start proving the following preliminary probabilistic inequalities.

Lemma 18 *For $0 < \eta_1, \eta_2, \eta_3, \eta_4 \leq 1$, $n \in \mathbb{N}$, it holds*

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

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

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

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

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 $\xi_1, \ldots, \xi_n$ are n i.i.d. samples, then

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

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

where $\mathcal{H}S(\mathcal{H}), \|\cdot\|_{\mathcal{H}S(\mathcal{H})}$ 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).

■

C.1 Proofs of the Consistency of the Regularizer

We restate Theorem 7 in an extended form.

Theorem 7 (Extended) *Let $r < \infty$, then under assumption (A2), for any $\eta > 0$,*

$$\mathrm{P}\left(\sup_{\|f\|_{\mathcal{H}} \leq r} |\widehat{\Omega}_1^D(f) - \Omega_1^D(f)| \geq rd\frac{2\sqrt{2}}{(n)^{1/4}}\kappa_2\sqrt{\log\frac{2d}{\eta}}\right) < \eta. \tag{38}$$

Consequently

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

Proof

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

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

that follows from $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$, the definition of $\hat{D}_a, D_a$ and basic inequalities. Then, using d times inequality (d) in Lemma 18 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 easily follows. ■

C.1.1 CONSISTENCY PROOFS

To prove Theorem 8, we need the following lemma.

Lemma 19 *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 probability $1-\eta$.

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

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

The proof follows applying Lemma 18 with probabilities $\eta_1 = \eta_2 = \eta_3 = \eta/3$. ■

We now prove Theorem 8. We use the following standard result in regularization theory (see, for example, Dontchev and Zolezzi, 1993) to control the the approximation error.

Proposition 20 *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 [Proof of Theorem 8] We recall the standard sample/approximation error decomposition

$$\mathcal{E}(\hat{f}^\tau) - \inf_{f\in\mathcal{H}} \mathcal{E}(f) \leq |\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau)| + |\mathcal{E}^\tau(f^\tau) - \inf_{f\in\mathcal{H}} \mathcal{E}(f)| \tag{39}$$

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 $\|f^\tau\|_{\mathcal{H}} \leq (\int_X y^2 d\rho)^{1/2}/\sqrt{\tau\nu} \leq \frac{M}{\sqrt{\tau\nu}}$.

We have the following bound,

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

Let $\eta' \in (0,1]$. Using Lemma 19 with probability $\eta = 3\eta'/(3+d)$, and inequality (38) with $\eta = d\eta'/(3+d)$, and if η' is sufficiently small we obtain

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

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

$$\mathcal{E}(\hat{f}^\tau) - \mathcal{E}^\tau(f^\tau) \leq c \left(\frac{M\kappa_1^2}{n^{1/2}\tau\nu} + \frac{\tau^{1/2} d\kappa_2}{n^{1/4}\sqrt{\nu}} \right) \log\frac{6+2d}{\eta'} \tag{40}$$

where c does not depend on n, τ, ν, d. The proof follows, if we plug (40) in (39) 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 20) and the sample error goes to zero in probability as $n \to \infty$ by (40).

■

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 21 *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 9. 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. (2012) based on Attouch-Wets 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

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

We are now ready to prove Theorem 9, which we present here in an extended form.

Theorem 9 (Extended) *Under assumptions (A1), (A2) and (A3),*

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

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,

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

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

Proof [Proof of Theorem 9]

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

$$\|\hat{f}^\tau - f_\rho^\dagger\|_{\mathcal{H}} \leq \|\hat{f}^\tau - f^\tau\|_{\mathcal{H}} + \|f^\tau - f_\rho^\dagger\|_{\mathcal{H}}. \tag{41}$$

From Theorem 2.6 in Villa et al. (2012) we have that

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

where $\psi^{\diamond}_{\tau\nu}(t) := \inf\{\frac{\tau\nu}{2}s^2 + |t-s| : s \in [0,+\infty)\}$, 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. (2012), 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 7 (equation (38)) and Lemma 19 we obtain with probability $1-\eta'$, if η' is small enough,

$$\begin{aligned}
d_{M/\sqrt{\tau\nu}}(t_{\mathcal{E}^{\tau}}\mathcal{E}^{\tau}, t_{\mathcal{E}^{\tau}}\widehat{\mathcal{E}}^{\tau}) &\leq \frac{2\sqrt{2}}{\sqrt{n}} \left(\kappa_1^2 \frac{4M^2}{\tau\nu} + 4\kappa_1 \frac{M^2}{\sqrt{\tau\nu}} + M^2 \right) \log \frac{6+2d}{\eta'} + \tau \frac{2M}{\sqrt{\tau\nu}} d \frac{2\sqrt{2}}{n^{1/4}} \kappa_2 \sqrt{\log \frac{6+2d}{\eta'}} \\
&\leq 2\sqrt{2}M \left(\frac{4\kappa_1^2 M}{\sqrt{n}\tau\nu} + \frac{4\kappa_1 M}{\sqrt{n}\sqrt{\tau\nu}} + \frac{M}{\sqrt{n}} + \tau \frac{2d\kappa_2}{n^{1/4}\sqrt{\tau\nu}} \right) \log \frac{6+2d}{\eta'}. \qquad (42)
\end{aligned}$$

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

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

Since $\tau = \tau_n \to 0$ by assumption, for sufficiently large n, the bound in (42) is smaller than $1/2\tau\nu$, and we obtain that with probability $1-\eta'$,

$$\|\hat{f}^{\tau} - f^{\tau}\|_{\mathcal{H}} \leq \left(4\sqrt{2}M \left(\frac{4\kappa_1^2 M}{\sqrt{n}\tau^2\nu^2} + \frac{4\kappa_1}{\sqrt{n}\tau\nu\sqrt{\tau\nu}} + \frac{1}{\sqrt{n}\tau\nu} + \frac{2d\kappa_2}{n^{1/4}\nu\sqrt{\tau\nu}} \right) \right)^{1/2} \sqrt{\log \frac{6+2d}{\eta'}}. \qquad (43)$$

If we now plug (43) in (41) 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 21 is satisfied and the bound in (43) goes to 0, so that the limit of the sum of the sample and approximation terms goes to 0.

■

C.1.2 PROOFS OF THE SELECTION PROPERTIES

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

Lemma 22 *Under assumptions (A1), (A2) and (A3) and defining $A(n,\tau)$ as in Theorem 9 extended, we have, for all $a = 1,\dots,d$ and for all $\varepsilon > 0$,*

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

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

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

Using Theorem 9 extended, equation (43), and Lemma 18 with probability $\eta_4 = \eta/(3+d)$, we obtain with probability $1-\eta$

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

We can further write

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

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 21. The proof follows by writing $\varepsilon = a(n,\tau)\log\frac{6+2d}{\eta} + b(\tau)$ and inverting it with respect to η.

■

Finally we can prove Theorem 11.

Proof [Proof of Theorem 11] 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 $C < \min_{a\in R_\rho}\|D_a f_\rho^\dagger\|_{\rho_X}^2$. From Lemma 22, there exist $a(n,\tau)$ and $b(\tau)$ satisfying $\lim_{\tau\to 0} b(\tau) = 0$, such that

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

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

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

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

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

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

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

■

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

H. Attouch and R. Wets. Quantitative stability of variational systems. I. The epigraphical distance. *Transactions of the American Mathematical Society*, 328(2):695–729, 1991.

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

H. Attouch and R. Wets. Quantitative stability of variational systems. III. ε-approximate solutions. *Mathematical 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 *Proceedings of the 21st International Machine Learning Conference*, Banff, Alberta, Canada, 2004.

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

Spaces and distributions	$\mathcal{X} \subseteq \mathbb{R}^d$	input space
	$\mathcal{Y} \subseteq \mathbb{R}$	output space
	ρ	probability distribution on $\mathcal{X} \times \mathcal{Y}$
	$\rho_{\mathcal{X}}$	marginal distribution of ρ
	$L^2(\mathcal{X}, \rho_{\mathcal{X}})$	$\{f : \mathcal{X} \to \mathbb{R} : \text{measurable and s.t. } \int_{\mathcal{X}} f(x)^2 \, d\rho_{\mathcal{X}}(x) < +\infty\}$
	$\mathcal{H}$	RKHS $\subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$
Norms and scalar products	$\|\cdot\|_n$ and $\langle \cdot, \cdot \rangle_n$	$\frac{1}{\sqrt{n}} \cdot$ euclidean norm and scalar product
	$\|\cdot\|_{\rho_{\mathcal{X}}}$ and $\langle \cdot, \cdot \rangle_{\rho_{\mathcal{X}}}$	norm and scalar product in $L^2(\mathcal{X}, \rho_{\mathcal{X}})$
	$\|\cdot\|_{\mathcal{H}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$	norm and scalar product in $\mathcal{H}$
Functionals and Operators	$\Omega_1^{\mathrm{D}} : \mathcal{H} \to [0, +\infty)$	$\Omega_1^{\mathrm{D}}(f) = \sum_{a=1}^{d} \sqrt{\int_{\mathcal{X}} \left(\frac{\partial f(x)}{\partial x^a} \, d\rho_{\mathcal{X}}(x) \right)^2}$
	$\widehat{\Omega}_1^{D} : \mathcal{H} \to [0, +\infty)$	$\widehat{\Omega}_1^{D}(f) = \sum_{a=1}^{d} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial f(x_i)}{\partial x^a} \right)^2}$
	$\mathcal{E} : \mathcal{H} \to [0, +\infty)$	$\mathcal{E}(f) = \int_{\mathcal{X}} (f(x) - y)^2 \, d\rho(x, y)$
	$\mathcal{E}^{\tau} : \mathcal{H} \to [0, +\infty)$	$\mathcal{E}^{\tau}(f) = \int_{\mathcal{X}} (f(x) - y)^2 \, d\rho(x, y) + \tau(2\Omega_1^{\mathrm{D}}(f) + \nu \|f\|_{\mathcal{H}}^2)$
	$\hat{\mathcal{E}} : \mathcal{H} \to [0, +\infty)$	$\hat{\mathcal{E}}(f) = \sum_{i=1}^{n} \frac{1}{n} (f(x_i) - y_i)^2$
	$\hat{\mathcal{E}}^{\tau} : \mathcal{H} \to [0, +\infty)$	$\hat{\mathcal{E}}^{\tau}(f) = \sum_{i=1}^{n} \frac{1}{n} (f(x_i) - y_i)^2 + \tau(2\widehat{\Omega}_1^{D}(f) + \nu \|f\|_{\mathcal{H}}^2)$
	$I_k : \mathcal{H} \to L^2(\mathcal{X}, \rho_{\mathcal{X}})$	$(I_k f)(x) = \langle f, k_x \rangle_{\mathcal{H}}$
	$\hat{S} : \mathcal{H} \to \mathbb{R}^n$	$\hat{S} f = (f(x_1), \ldots, f(x_n))$
	$D_a : \mathcal{H} \to L^2(\mathcal{X}, \rho_{\mathcal{X}})$	$(D_a f)(x) = \langle f, (\partial_a k)_x \rangle$
	$\hat{D}_a : \mathcal{H} \to \mathbb{R}^n$	$\hat{D}_a(f) = \left(\frac{\partial f}{\partial x^a}(x_1), \ldots, \frac{\partial f}{\partial x^a}(x_n) \right)$
	$\nabla : \mathcal{H} \to (L^2(\mathcal{X}, \rho_{\mathcal{X}}))^d$	$\nabla f = (D_a f)_{a=1}^d$
	$\hat{\nabla} : \mathcal{H} \to (\mathbb{R}^n)^d$	$\hat{\nabla} f = (\hat{D}_a f)_{a=1}^d$
Functions	$k_x : \mathcal{X} \to \mathbb{R}$	$t \mapsto k(x, t)$
	$f_{\rho}^{\dagger}$	$\operatorname{argmin}_{f \in \operatorname{argmin} \mathcal{E}} \{\Omega_1^D(f) + \nu \|f\|_{\mathcal{H}}^2\}$
	$(\partial_a k)_x : \mathcal{X} \to \mathbb{R}$	$t \mapsto \left. \frac{\partial k(s,t)}{\partial s^a} \right\|_{s=x}$
	f^{τ}	the minimizer in $\mathcal{H}$ of $\mathcal{E}^{\tau}$
	$\hat{f}^{\tau}$	the minimizer in $\mathcal{H}$ of $\hat{\mathcal{E}}^{\tau}$
Sets	R_{ρ}	$\{a \in \{1, \ldots d\} : \frac{\partial f_{\rho}^{\dagger}}{\partial x^a} \neq 0\}$
	$\hat{R}^{\tau}$	$\{a \in \{1, \ldots d\} : \frac{\partial f^{\tau}}{\partial x^a} \neq 0\}$
	B_n	$\{v \in \mathbb{R}^n : \|v\|_n \leq 1\}$
	B_n^d	$\{v \in \mathbb{R}^n : \|v\|_n \leq 1\}^d$

Table 3: List of symbols and notations

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

J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle. A ℓ^1-unified variational framework for image restoration. In *Proceedings of the 8th European Conference on Computer Vision, Part IV*, pages 1–13, Prague, Czech Republic, 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.

J. Candès, E.and Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

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 Modeling and Simulation*, 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. *Annals of Statistics*, 40(5):2667–2696, 2012.

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. *Constructive Approximation. An International Journal for Approximations and Expansions*, 33(1):125–143, 2011.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

D. Donoho. Compressed sensing. *IEEE. Transactions on Information Theory*, 52(4):1289–1306, 2006.

A. L. Dontchev and T. Zolezzi. *Well-posed Optimization Problems*. Springer-Verlag, Berlin, 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. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

C. Gu. *Smoothing Spline ANOVA Models*. Springer, New York, 2002.

E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for l1-minimization: Methodology and convergence. *SIAM Journal of Optimization*, 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.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 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. *Applied and Computational Harmonic Analysis*, 27(2):247–254, 2009.

A. Maurer and M. Pontil. Structured sparsity and generalization. *Journal Machine Learning Research*, 13:671–690, 2012.

H. Miller and P. Hall. Local polynomial regression and variable selection. In J.O. Berger, T.T. Cai, and I.M. Johnstone, editors, *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown*, volume 6, pages 216–233. Institute of Mathematical Statistics, 2010.

J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. In J.L. Balcàzar, F. Bonchi, A. Gionis, and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6322 of *LNCS*, pages 418–433. Springer Berlin Heidelberg, 2010.

A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, New York, 1983.

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.

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

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory of Probability and its Applications*, 30(1):143–148, 1985.

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.

S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex Analysis*, 19(4), 2012.

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.

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. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

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

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming. A Publication of the Mathematical Programming Society*, 125(2, Ser. B):263–295, 2010.

S. Villa, S. Salzo, L. Baldassarre, and A. Verri. *SIOPT*.

S. Villa, L. Rosasco, S. Mosci, and A. Verri. Consistency of learning algorithms using Attouch-Wets convergence. *Optimization*, 61(3):287–305, 2012.

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. *Annals of Statistics*, 23:1865–1895, 1995.

D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220:456–463, 2008.