The impact of smoothness on model class selection in nonlinear system identification: An application of derivatives in the RKHS

Yusuf Bhujwalla, Vincent Laurain, Marion Gilson

The Impact of Smoothness on Model Class Selection in Nonlinear System Identification: An Application of Derivatives in the RKHS

Yusuf Bhujwalla[a,b,c], Vincent Laurain[a,b], and Marion Gilson[a,b]

[a] Université de Lorraine, CRAN, UMR 7039, 2 rue Jean Lamour, F-54519, Vandoeuvre-lès-Nancy, France.
[b] CNRS, CRAN, UMR 7039, France
[c] LTER - Zone Atelier du bassin de la Moselle, CRAN
Email Adresses: yusuf-michael.bhujwalla@univ-lorraine.fr, vincent.laurain@univ-lorraine.fr, marion.gilson@univ-lorraine.fr

***Abstract*— In this paper, we discuss the dependency between the kernel choice and the model class it represents. This is typically an undesired relationship, forcing the user to accept a trade-off between an acceptable variance characteristic and flexibility in the underlying function.**

Hence, a method is proposed in this paper that explicitly constrains the smoothness of the model, by regularizing over the derivative of the function. This not only broadens the available model class, but also simplifies the selection of any hyperparameters.

We look at nonparametric models of nonlinear systems, and formulate the problem in the Reproducing Kernel Hilbert Space (RKHS). The proposed method is compared with an equivalent, established scheme by means of a simple example. It is shown that derivative-based regularization can help to extract useful structural information about an underlying system.

Keywords: Nonlinear System Identification; Nonparametric Modeling; RKHS; Regularization; Derivatives; Gradient Regularization.

I. Introduction

Many real systems are nonlinear, or exhibit some nonlinear behavior. As such, nonlinear identification represents a well-studied and important area of system identification [2], [8]. However, attempting to estimate nonlinear models can be very challenging, posing many problems of which the user must be aware [15].

One problem is that nonlinear models represent a vast range of functions. Defining a model class that encompasses the true behavior of a system can be extremely difficult. In the absence of sufficient prior knowledge about the system, how can any useful statements about these nonlinearities be made?

A way of dealing with this is to consider a nonparametric representation of the system. Here, the number of parameters describing a model depends on the amount of data available, offering a very flexible way of describing nonlinear systems. However, this flexible *over-parameterized* approach has certain drawbacks. For a model of this form, its estimated parameters will typically have unacceptably high variance, and be a very poor predictor of future outputs. This is commonly dealt with using *regularization*.

Regularization is an established statistical technique [3], [14], [16], allowing the incorporation of some prior intuition of how the true system may behave into the model. It is typically used as a way of controlling the bias-variance trade-off in a model, in variable selection, or as a way of enforcing some desired properties upon the estimator (such as smoothness [4] or sparsity [12]).

An elegant way of constructing a nonparametric identification problem is to use the theory of Reproducing Kernel Hilbert Spaces (RKHS). It should be mentioned that there is an equivalence between RKHS methods and several other branches of kernel methods, notably the *Gaussian Process* and *Least-Squares Support Vector Machine* approaches in machine learning. This framework is applicable to many different model classes, and allows their implicit definition through an associated kernel function.

Whilst the regularization and its associated hyperparameters are often discussed, the choice of a suitable kernel is frequently overlooked. As the kernel defines the RKHS, it also implicitly defines the model class - and hence the capacity of the model to capture the true nonlinearity and the smoothness of the true function.

In this paper, we will investigate if, by constraining the smoothness of the estimated function through the regularization, it is possible to integrate the kernel selection into the optimization criterion, instead of relying on an *a priori* choice of kernel.

This paper will be structured as follows. In section II, the identification problem will be formulated. Section III discusses the smoothness-flexibility trade-off, and the impact this has on the identification problem. In section IV, a potential solution to this problem is proposed, along with some results regarding derivatives in the RKHS. Finally, in section V, a simulation example will be given, providing an indication of the performance of the proposed scheme and hopefully illustrating its potential advantages.

II. Problem Description

A. The Data-Generating System

We will now introduce the identification setup. Assuming we have N observations of a data-generating system $\mathcal{S}_o$:

$$D_N = \{(u_1, y_1), (u_2, y_2), \cdots, (u_N, y_N)\}. \qquad (1)$$

In System Identification, often a variable $\mathbf{x_k} \in \mathbb{R}^{n_a+n_b+1}$ is introduced. Typically, $\mathbf{x_k}$ is the *regressor vector*, composed of the past and present inputs and outputs of the system, u and y:

$$\mathbf{x_k} = [y_{k-1} \ \cdots \ y_{k-n_a} \ u_k \ \cdots \ u_{k-n_b}]^\top \tag{2}$$

where n_a and n_b are the orders of the output and input respectively.

Using these definitions, we will attempt to reconstruct a general nonlinear model of the unknown true system $\mathcal{S}_o$ describing its behavior:

$$y_k = f_o(\mathbf{x_k}) + e_{o,k} \tag{3}$$

Here, $f_o : \mathbb{R}^{n_a+n_b+1} \to \mathbb{R}$ is an unknown nonlinear function and $e_{o,k}$ is an additive noise term at the output at each state k. To simplify the analysis, here $\mathcal{S}_o$ is assumed to be a single-input single-output (SISO) system, but the extension to the multivariate case is straightforward.

It will also be assumed that e_o is white Gaussian noise (see [6] for a discussion of the effect of colored noise in identification).

B. Hilbert Spaces

The function $f(\cdot)$ is stated to be part of the Hilbert space $\mathcal{H}$ ($f \in \mathcal{H} : \mathcal{X} \to \mathbb{R}$). A Hilbert space represents all possible realizations of some particular class of functions, for example all functions of continuity degree C^k. Moreover, a Hilbert space is a vector space such that any function $f \in \mathcal{H}$ must have a nonnegative norm, $\|f\|_\mathcal{H} > 0 \quad (\text{for} \quad f \neq 0)$. f must be also equipped with an inner-product in $\mathcal{H}$. For example, for real-valued functions $f, g \in \mathcal{L}_2$, the inner-product would have the familiar form:

$$\langle f(x), g(x) \rangle = \int_\mathcal{X} f(x) g(x) \, \mathrm{d}x. \tag{4}$$

The properties of Hilbert spaces have been explored in great detail throughout the literature. For the interested reader, it is recommended to refer to one of the many texts discussing the subject, such as [19].

C. Reproducing Kernels

An extremely useful property of Hilbert Spaces is their equivalence with an associated kernel function [1]. This equivalence allows us the simple definition of a kernel, instead of fully defining the associated vector space. More formally, if a Hilbert space $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS), it will have a unique kernel, $K : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$, spanning the space $\mathcal{H}$.

To simplify the analysis, it will be assumed that $x_k = u_k$, and therefore our observations of the system are now $\{x_i\}_{i=1}^N \in \mathbb{R}^N$.

Importantly, any function in $\mathcal{H}$ can be represented as a infinite weighted linear sum of this kernel evaluated over the space $\mathcal{H}$, as:

$$\begin{aligned} f(\cdot) &= \langle f(x), k_x(\cdot) \rangle_\mathcal{H} \\ &= \sum_{i=1}^{\infty} \alpha_i k_{x_i}(\cdot) \end{aligned} \tag{5}$$

where $\{\alpha_i\}_{i=1}^\infty \in \mathbb{R}^\infty$ are the parameters of the model. This relationship is known as the Reproducing Property and k_{x_i} is known as the kernel slice.

One of the most useful properties of a kernel in the RKHS is that:

$$\begin{aligned} \langle k_{x_i}, k_{x_j} \rangle &= k_{x_i}(x_j) = k_{x_j}(x_i) \\ &= K(x_i, x_j) \end{aligned} \tag{6}$$

that is, the dot-product of a reproducing kernel with itself is itself the kernel. This result is commonly known as the kernel trick, and allows us to write the inner-product as a tractable function which implicitly defines a higher (or even infinite) dimensional space.

For K to be a valid kernel in the RKHS, it must be a *Mercer kernel* [13].

D. The Representer Theorem

In reality any estimation problem will deal only with finite data. But since f is represented by infinitely many parameters, a regularized cost function is commonly used in the form:

$$\begin{aligned} \mathcal{V}(e, f) = c((x_1, y_1, f(x_1)), \cdots, (x_N, y_N, f(x_N))) \\ + \\ g(\|f\|_\mathcal{H}). \end{aligned} \tag{7}$$

For example, $c(\cdot)$ is often a loss-function in $\mathcal{L}_2$, such that

$$c\left(\mathbf{x}, \mathbf{y}, f(\mathbf{x})\right)) = \sum_{k=1}^{N} (y_k - f(x_k))^2 \tag{8}$$

where $\mathbf{x} = [x_1 \cdots x_N]^\top$ and $\mathbf{y} = [y_1 \cdots y_N]^\top$. $g(\cdot)$ is an additional global constraint on the function, independent of the observations. This term is the *regularization term*. Often, $g(\cdot)$ is taken as a constant:

$$g\left(\|f\|_\mathcal{H}\right) = \lambda \|f\|_\mathcal{H} \tag{9}$$

Here λ is the *regularization hyperparameter* - which can be considered as controlling the bias-variance trade-off.

If f is in $\mathcal{H}$, (7) permits a truncated form of the expression given in (5). This means the infinite expansion given in (5) reduces to a finite summation around the observations alone. This is *The Representer Theorem*, for which a proof can be found in [13]. For $\mathcal{V}(e, f)$ given above, f can be written as

$$f(\cdot) = \sum_{i=1}^{N} \alpha_i k_{x_i}(\cdot), \quad i = 1, 2, \ldots, N \quad \alpha_i \in \mathbb{R} \tag{10}$$

Using this expression, the norm $\|f\|_\mathcal{H}$, can now be defined as

$$\begin{aligned} \|f\|_\mathcal{H} &= \langle f, f \rangle_\mathcal{H} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \\ &= \alpha^\top \mathbf{K} \alpha \end{aligned} \tag{11}$$

where $\mathbf{K}$ is the $N \times N$ *Gram Matrix*, $\{\mathbf{K}\}_{i,j} = K(x_i, x_j)$ and $\alpha = [\alpha_1 \cdots \alpha_N]^\top$ is the parameter vector.

Taking (8) and (9) yields the following cost-function:

$$\mathcal{V}(e, f) = \|y - f(x)\|_2^2 + \lambda \|f\|_{\mathcal{H}}. \tag{12}$$

A closed-form solution can now be obtained for α:

$$\Rightarrow \quad \alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}. \tag{13}$$

This solution will likely be familiar to the reader, as it is widely used in inverse problems. Note that this solution will depend on the hyperparameter λ and the choice of kernel K. Typically, determination of hyperparameters is a non-trivial problem, requiring the implementation of methods such as cross-validation or marginal likelihood [9], [10].

III. Model Class Selection

A. The Kernel Selection Dilemma

Whilst the choice of λ is often discussed in the literature, the choice of the kernel K is usually left somewhat open. But in fact, this is a very important part of the identification process. Selecting a kernel usually requires some statement about the nature of the system in question to be made.

A range of kernel functions are available to the user, for example *step, linear, polynomial and spline* functions. A popular choice of kernel for continuous functions, which we will consider here, is the *Gaussian RBF* kernel:

$$k_{x_i}(x) = \exp\left\{ -\frac{\|x - x_i\|^2}{\sigma^2} \right\}. \tag{14}$$

which represents the C^∞ class of functions. Henceforth, the discussion will be continued solely considering RBF kernels. The results developed will nonetheless be valid independent of the choice of kernel.

In the RBF case, the hyperparameter σ defines the width of K, and hence the smoothness of the estimated function. Increasing σ can be considered as smoothing f by progressively filtering out higher-frequencies. As σ characterizes K, and K defines $\mathcal{H}$, σ also influences $\mathcal{H}$. Hence, σ is somehow related to the model class selection. However, the choice of σ will also be influenced by the noise. We will now introduce an example to illustrate this paradox.

B. A Simple Example

A one-dimensional static switching function is corrupted by white noise at the output. A signal-to-noise ratio of SNR $= 10$dB is used in this example, defined by the ratio of the noise-free output to the noise level: SNR $= 20 \log\left(\sigma_{\mathring{y}}/\sigma_e\right)$. No other information about the function, such as the switching points, is provided.

The unknown system $\mathcal{S}_o$ is described by the following equation:

$$y_k = f_o(x_k) + e_{o,k},$$

where $x_k \sim U(-1, 1)$ is a uniformly distributed input-signal and $e_{o,k} \sim N(0, \sigma_N^2)$ is white Gaussian noise. The function f_o is a non-zero mean, non-smooth function:

$$f_o(x) = \begin{cases} 15, & x \leq -0.5, \text{ or } 0 < x \leq 0.5 \\ -5, & 0.5 < x \leq 0, \text{ or } 0.5 < x \end{cases}$$

In Figure 1, each image shows the function reconstructed for different values of σ and λ.

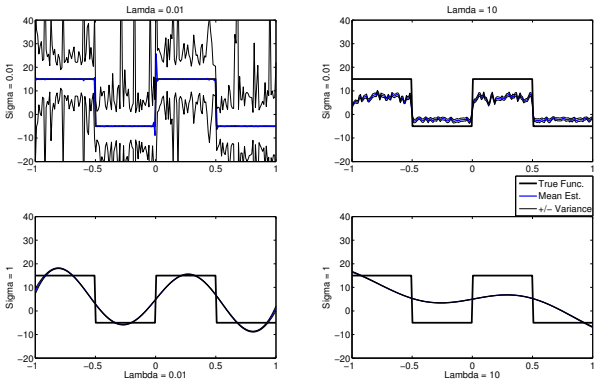

Fig. 1. Estimation of 1D switching signal for different hyperparameter values. The mean value of each function over 1000 Monte-Carlo trials is plotted against the true function, with $\pm$ the variance marked on either side.

As mentioned, an RBF kernel was used, as given in (14). Clearly, the discontinuous $f_o \in C^{-1}$ lies outside the Hilbert space of the RBF kernel - as it would require an infinite range of frequencies to fully reconstruct the function. Using such a function allows us to examine the effect of σ and λ on the limits of the available model class.

The kernel width needs to be sufficiently small not to make any assumptions on the smoothness of f, allowing a large model class. In the top-left figure, it can be seen that choosing such a σ allows the dynamics of the true function to be captured by the estimated function - but it is unacceptably sensitive to noise.

In the bottom-left figure, we see that increasing σ does improve the variance characteristics of the estimate, but at the cost of placing a strong assumption on the smoothness of the function. The effect of reducing the size of the available model class means we have failed to capture the dynamics of the system.

From (9), it is possible to introduce a regularization term to improve the characteristics of the estimator. But, as can be seen in the top-right and bottom-right figures, even when an excessively high λ is used to reduce the variance, such that an undesired bias is introduced, the function is still not smooth.

Optimal values for the hyperparameters can be found, but this does not change that, in practice, we are forced to choose between the flexibility of our model and the smoothness.

Hence, when selecting a kernel, how can we choose the widest class possible in order to avoid making assumptions on f_o, whilst still having a reliable estimator, capable of reproducing any given data and predicting future outputs?

In the following section, we will propose a method to directly enforce smoothness independently of the model class selection by constraining the gradient of the estimated function.

IV. Smoothness in the RKHS

The smoothness of a function f is inextricably linked to the continuity of the space C^k within which it resides. Therefore, a natural way of enforcing smoothness on a

function would be to place constraints on its derivatives:

$$\frac{df(x)}{dx}, \frac{d^2f(x)}{dx^2}, \cdots, \frac{d^kf(x)}{dx^k} \tag{15}$$

There are several ways this could be achieved. Before proposing a solution in section IV-C, we will briefly review several existing methods from the literature.

A. Sobolev Spaces

One relatively well-known method of enforcing smoothness involves the consideration of *Sobolev Spaces*. In a Sobolev space, $S^{k,p}$ of continuity C^k and norm p, a function within the space is quantified in terms of its derivative. For $p = 2$, this Sobolev Space becomes a Hilbert Space, $\mathcal{H}^k$, and we can define:

$$\|f\|_{\mathcal{H}_k} = \sum_{i=0}^{k} \int_{\mathcal{X}} \left(\frac{d^i f(x)}{dx^i} \right)^2 dx \tag{16}$$

Using this definition of the norm will naturally impose continuity constraints on the function up to order k. Under certain conditions, a minimal representation of f in this case can be found based on the *spline* kernel, as discussed in [17]. Though this is a very effective solution, we would like to develop a method valid for a range of kernels, rather than enforcing the choice of kernel.

B. Identification with Observations on the Derivative

Another method discussed in the literature is based on minimization against observations of the derivative [11]. In certain cases it may not be feasible to obtain measurements directly for the output, or it may be of interest to constrain the error against the derivatives. Then, a cost function of the form below may be used:

$$\begin{aligned} \mathcal{V}_{obvs}(f) = \|y - f(x)\|_2^2 + \gamma_1 \left\| \frac{dy}{dx} - \frac{df(x)}{dx} \right\|_2^2 \\ + \cdots \gamma_m \left\| \frac{d^m y}{dx^m} - \frac{d^m f(x)}{dx^m} \right\|_2^2 + \lambda \|f\|_{\mathcal{H}} \end{aligned} \tag{17}$$

where here we have considered an MSE loss. Again, this is capable of enforcing smoothness, but it introduces m additional hyperparameters (denoted here by γ) increasing the complexity of the problem.

C. Regularization using Derivatives

Here, a scheme independent of kernel choice is proposed, with the advantages of neither requiring any additional hyperparameters nor any significant additional computational load (compared with the solution given in (13)). By replacing the regularization term given in (7) with one optimizing against the derivative of the function, we have:

$$\mathcal{V}_{\nabla}(e, f) = \|y - f(x)\|_2^2 + \lambda \|Df\|_{\mathcal{H}} \tag{18}$$

where $D^m f = \dfrac{d^m f(x)}{dx^m}$ is the m^{th} order differential operator. Although not identical, there is a clear parallel with the spline-smoothing problem mentioned in section IV-A. Now λ controls the smoothness to be put on the function, meaning σ is no longer critical in reducing the sensitivity of the model to noise. By explicitly forcing the regularization to act in this way, a greater flexibility in the model class may be permitted.

Note that as λ tends to ∞, for a bounded input this enforces a bounded solution, as $\|f\|_{\mathcal{H}} \to 0$ now becomes $\|Df\|_{\mathcal{H}} \to 0$, which implies that $\forall x \in \mathbb{R}$:

$$\Rightarrow \lim_{\lambda \to \infty} f(x) = c, \quad \text{where } 0 \leq c < \infty \in \mathbb{R} \tag{19}$$

D. The Representer Theorem for Derivative Regularization

As discussed in section II, for the cost-function given in (7), a valid representer is $f(\cdot) = \sum_{i=1}^{N} \alpha_i k_{x_i}(\cdot)$. However, this form is not necessarily valid for the case of (18).

For the proof of the generalized representer theorem given in [13] to hold, it is required that a relationship can be defined such that

$$\|D^m f\|_{\mathcal{H}} = g\left(\|f\|_{\mathcal{H}}\right) \tag{20}$$

where $g(\cdot)$ is a strictly *monotically increasing* function on the norm of f. Whilst a relationship can be defined, it will *not* in general be a monotonic function of $\|f\|_{\mathcal{H}}$. In the case of the Gaussian RBF kernel adopted in section III-B, (20) does not hold.

Hence, a representer of the form stated above will be *suboptimal* for (18). This means that, unlike in the case of (7), the reproducing property of (5) cannot strictly be truncated to a finite expression evaluated solely over the observed data.

Nonetheless, we will proceed using this form, whilst acknowledging that a better-performing solution may exist. Whilst using a suboptimal representer is not ideal, it will allow us to preserve the mathematical simplicity and computational efficiency of the solution given in (13). Furthermore, in addition to being able to control the smoothness of the estimated function, for certain choices of kernels the statistical properties of (7) are loosely preserved:

Lemma 1 (The boundedness of $f(x)$ for (18)). *For a cost-function of the form given in (18), let the representer of the estimated function be $f(x) = \sum_{i=1}^{N} \alpha_i k_{x_i}(x)$. If $f(x) : \mathcal{X} \to \mathbb{R} \in \mathcal{H}$ is bounded such that $\|f\|_{\mathcal{H}} < \infty$, and the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ spanning $\mathcal{H}$ decays such that $k_{x_i}(x) \underset{||x - x_i|| \to \infty}{\longrightarrow} 0$, $f(x)$ is bounded as:*

$$\lim_{\lambda \to \infty} f(x) = 0, \quad \forall x \in \mathbb{R}. \tag{21}$$

Proof: Given that $\|f\|_{\mathcal{H}} < \infty$ and $k_{x_i}(x) \to 0$ as $\|x - x_i\| \to \infty$:

$$\Rightarrow \lim_{\|x - x_i\| \to \infty} f(x) = 0 \tag{22}$$

where $\{x_i\}_{i=1}^{N} \in \mathcal{X}^N$, $\forall i = 1 \cdots N$. Now recall from (19), that for (18):

$$\lim_{\lambda \to \infty} f(x) = c, \quad \forall x \in \mathbb{R} \tag{23}$$

where $0 \leq c < \infty \in \mathbb{R}$.

Observe that (23) applies across $\mathbb{R}$. Hence if $\exists x \in \mathbb{R}$ as in (22), c must be 0. Therefore:

$$\lim_{\lambda \to \infty} f(x) = 0, \quad \forall x \in \mathbb{R}. \tag{24}$$

Hence, under the above-stated conditions, $f(x)$ is bounded by the regularization term of (18) as in (7).

This suboptimal representer thus allows the implicit regularization of $\|f\|_{\mathcal{H}}$ in (18), in addition to enforcing smoothness on $f(x)$.

E. A Closed-Form Solution

From [18], let $f \in \mathcal{H}_k$, where k denotes the order of continuity of the space $\mathcal{H}_k$. Now, for $x \in \mathcal{X}$ where again we consider $\mathcal{X} = \mathbb{R}$, $D^m f(x) \in \mathcal{H}_k$ provided $m < k$. This allows a reproducing property to be defined for derivatives of functions, analogous to that stated in (5):

$$\begin{aligned} D^m f(\cdot) &= \langle f(x), D^{[0\ m]} k_x(\cdot) \rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{\infty} \alpha_i D^{[0\ m]} k_{x_i}(\cdot) \end{aligned} \tag{25}$$

where the operator $D^{[\iota\ \kappa]}$ is defined as differentiation with respect to the first and second variables of the kernel function as:

$$D^{[\iota\ \kappa]} k_{x_i}(x) = \frac{d^{\iota+\kappa} k_{x_i}(x)}{dx^{\kappa}\, dx_i^{\iota}}. \tag{26}$$

Now we can proceed to define $\|D^m f\|_{\mathcal{H}}$. From [18], $\left\langle D^{[0\ m]} K, D^{[0\ m]} K \right\rangle_{\mathcal{H}} = D^{[m\ m]} K$. In (18), $m = 1$, giving:

$$\begin{aligned} \|Df\|_{\mathcal{H}} &= \langle Df, Df \rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \frac{d^2 k_{x_i}(x_j)}{dx_j\, dx_i} \\ &= \alpha^{\top} \mathbf{D}^{[\mathbf{1}\ \mathbf{1}]} \mathbf{K} \alpha \end{aligned} \tag{27}$$

where $\mathbf{D}^{[\mathbf{1}\ \mathbf{1}]}\mathbf{K}$ is the $N \times N$ matrix of derivatives $\{\mathbf{D}^{[\mathbf{1}\ \mathbf{1}]}\mathbf{K}\}_{i,j} = \dfrac{d^2 k_{x_i}(x_j)}{dx_j\, dx_i}$. In the RBF case, the derivative is given by:

$$\frac{d^2 k_{x_i}(x)}{dx\, dx_i} = \frac{2}{\sigma^2} \left(1 - \frac{2}{\sigma^2} (x - x_i)^2 \right) \exp\left\{ - \frac{\|x - x_i\|^2}{\sigma^2} \right\}. \tag{28}$$

From (27), the derivation of a closed-form solution for α_{∇}, the optimal parameters of $\mathcal{V}_{\nabla}$, is trivial:

$$\Rightarrow \quad \alpha_{\nabla} = \left(\mathbf{K}^{\top} \mathbf{K} + \lambda \mathbf{D}^{[\mathbf{1}\ \mathbf{1}]} \mathbf{K} \right)^{-1} \mathbf{K}^{\top} \mathbf{y} \tag{29}$$

Note that this solution has a similar computational complexity to the solution given in (13).

F. A Simple Example

We now demonstrate the effect of regularizing over the gradient on the example from section III-B.

Figure 2 shows the estimation function for a small kernel ($\sigma = 0.01$) over a range of λ values with comparable variances in each case. The effects of the regularization term in the two cases are drastically different. On the left-hand side, the effect of λ on the bias-variance trade-off is clearly visible; the variance is progressively reduced, at the cost of an increasing bias in the model. But, the regularization *does not* act on the smoothness of the function.

On the right-hand side, the variance is reduced by smoothing the function. Eventually the estimated function becomes biased and excessively smooth. In practice, this simply means that λ must be tuned properly.

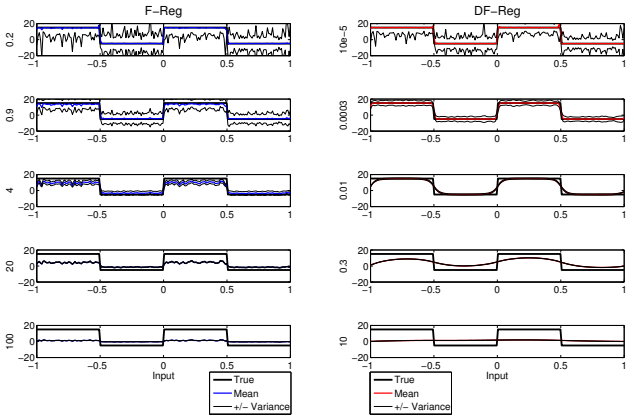

Fig. 2. Monte-Carlo Results: Mean Estimates $\pm$ variance for a small kernel ($\sigma = 0.01$), with λ values given on the y-axis of each plot

V. Simulation Example

A Monte-Carlo study based on a simulation example will now be used to compare the two algorithms discussed in this paper - the $\|f\|_{\mathcal{H}}$-regularized and $\|Df\|_{\mathcal{H}}$-regularized approaches.

A. The Data-Generating System

Two 1-D nonlinear functions are corrupted with white Gaussian noise at the output (SNR = 10dB). Both systems were excited using $N = 1000$ uniformly distributed datapoints, $x_k \sim U(-1, 1)$. The first system $\mathcal{S}_o^1$, a smooth function, is given by the following equation:

$$f_o^1(x) = 10((1 - x) - 2\text{sinc}(4x)^2, \tag{30}$$

and the second system $\mathcal{S}_o^2$, a non-smooth function, by:

$$f_o^2(x) = \begin{cases} 10(1 - x), & |x| > 0.2 \\ 10\left[(1 - x) + 10(|x| - 0.2)\right], & |x| \leq 0.2 \end{cases}$$

Two approaches to the hyperparameterization were used, with 100 Monte-Carlo trials run in each case. For both approaches, the kernel function was chosen to be the Gaussian RBF kernel defined in section III.

1. For $\|f\|_{\mathcal{H}}$-reg: λ and σ were optimized using cross-validation on a separate noisy validation dataset.

2. For $\|Df\|_{\mathcal{H}}$-reg: the kernel width was fixed *a priori* to $\sigma = 0.01$ such that the smallest value capable of enforcing smoothness is chosen ($\sigma \sim \max(x_{i+1} - x_i)$, $i = 1 \ldots N - 1$). λ was determined by cross-validation.

B. Estimation Results

Figure 3 shows the results of the Monte Carlo simulation. The average estimated function and the variance of the output across the input space are plotted for both methods. As an illustration, a sample kernel is given in each case. The central column shows the true nonlinearity and a sample of the noisy estimation data for both functions.

Table 1 shows the hyperparameter values obtained by cross-validation in each case. In addition, the variance of the estimator averaged of the input space and the mean best fit rate against the noise-free function (BFR) over the trials are given (where BFR $= 100\left(1 - \frac{\|\hat{y} - y\|^2}{\|y - \bar{y}\|^2}\right)$).

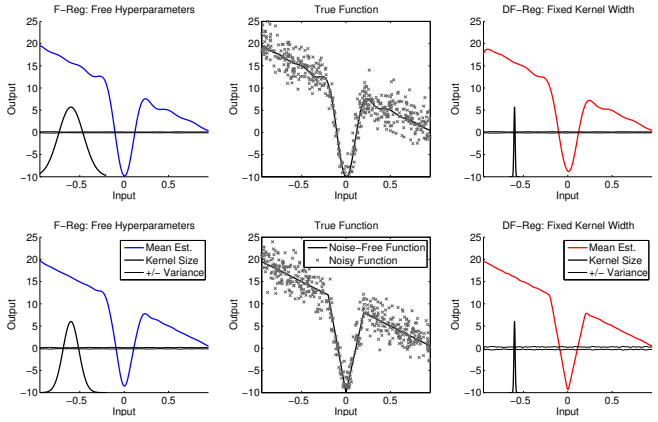

Fig. 3. Mean estimates of the functions, $\pm$ variance, and the kernel size used in each case for the smooth f_1 (top) and non-smooth f_2 (bottom); $\|f\|_{\mathcal{H}}$-Reg (left) and $\|Df\|_{\mathcal{H}}$-Reg (right)

System	$\mathcal{S}_o^1$	$\mathcal{S}_o^1$	$\mathcal{S}_o^2$	$\mathcal{S}_o^2$
Method	$\|f\|_{\mathcal{H}}$	$\|Df\|_{\mathcal{H}}$	$\|f\|_{\mathcal{H}}$	$\|Df\|_{\mathcal{H}}$
Kernel Width (σ)	0.2	0.01	0.13	0.01
Reg. Strength (λ)	0.18	0.01	0.10	5.6×10^{-4}
Mean Fit (%)	99.81	97.98	99.65	99.22
Mean Variance	0.10	0.12	0.14	0.29

TABLE I
SUMMARIZED RESULTS

From Table 1 it can be seen that both approaches estimate the function quite succesfully. However the $\|f\|_{\mathcal{H}}$-Reg scheme outperforms the $\|Df\|_{\mathcal{H}}$-Reg scheme. Despite this, analysis of Figure 3 shows several advantages of the $\|Df\|_{\mathcal{H}}$-Reg approach.

Firstly, using a small kernel, we can still estimate the smooth $\mathcal{S}_o^1$ - even with large levels of noise. In the $\|f\|_{\mathcal{H}}$-Reg approach, a much broader kernel is required. But, as for the $\|Df\|_{\mathcal{H}}$-Reg a small kernel can be chosen, we are never forced to select an overly smooth model class. As such, the structural difference between the smooth $\mathcal{S}_o^1$ and nonsmooth $\mathcal{S}_o^2$ is clearly apparent. For $\|f\|_{\mathcal{H}}$-Reg, it is not.

Furthermore, the results of $\|Df\|_{\mathcal{H}}$-Reg were achieved without having to optimize over the kernel hyperparameter. For $\mathcal{X} = \mathbb{R}$, this is already advantageous. As the dimensionality of the input increases, this becomes increasing attractive.

VI. CONCLUSIONS

In this paper, a derivative-based regularization method has been proposed, attempting to divorce the dependency of the smoothness of the estimated function from the selection of the underlying model class. It has been shown that using this scheme, both smooth and non-smooth functions can be estimated by using a small kernel width and shifting the hyperparameterization problem solely onto λ.

With λ appearing linearly in the parameters, this may simplify the hyperparameter optimization or perhaps permit the introduction of a varying smoothness constraint across the input space. Consideration of the multi-dimensional case and further investigation of the sub-optimality of the representer are intended as future works.

REFERENCES

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

[2] S. A. Billings and S. Chen. Identification of non-linear rational systems using a prediction-error estimation algorithm. *International Journal of Systems Science*, 20:467–494, 1989.

[3] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and gaussian processes - revisited. *Automatica*, 48(8):1525–1535, 2012.

[4] R. Duijkers, R. Tóth, D. Piga, and V. Laurain. Shrinking complexity of scheduling dependencies in ls-svm based lpv system identification. In *Proc of the 53rd IEEE Conference on Decision and Control*, pages 2561 – 2566, Los Angeles, California, USA, Dec. 2014.

[5] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[6] V. Laurain, R. Tóth, D. Piga, and W. Zheng. An instrumental least squares support vector machine for nonlinear systems identification. *Automatica*, 54:340–347, 2015.

[7] L. Ljung. *System Identification, theory for the user*. Prentice Hall, 1999.

[8] O. Nelles. *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer-Verlag, Berlin, 2001.

[9] G. Pillonetto, F. Dinuzzo, T. Chen, G. Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

[10] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[11] L. Rosasco, M. Santoro, S. Mosci, A. Verri, and S. Villa. A regularization approach to nonlinear variable selection. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 9:653–660, 2010.

[12] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric sparsity and regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2014.

[13] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. *Lecture Notes in Computer Science*, 2111:416–426, 2001.

[14] J. Sjöberg, T. McKelvey, and L. Ljung. On the use of regularization in system identification. In *Proc of the 12th IFAC World Congress*, volume 7, pages 318–386, Sydney, Australia, 1993.

[15] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: A unified overview. *Automatica*, 31(12):1691–1724, 1995.

[16] A. Tikhonov and V. Arsenin. *Solutions of Ill-Posed Problems*. Winston/Wiley, 1977.

[17] G. Wahba. Spline models for observational data. In *Proc of the SIAM CBMS-NSF Regional Conference Series in Applied Mathematics*, volume 59, Philadelphia, Pennsylvania, USA, 1990.

[18] D. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, 2008.

[19] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, 1995.