An RKHS approach to systematic kernel selection in nonlinear system identification

Yusuf Bhujwalla, Vincent Laurain, Marion Gilson

An RKHS Approach to Systematic Kernel Selection in Nonlinear System Identification

Yusuf Bhujwalla[a,b], 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
{ *yusuf-michael.bhujwalla* | *marion.gilson* | *vincent.laurain*} *@univ-lorraine.fr*

***Abstract*— This paper discusses the problem of kernel selection in Reproducing Kernel Hilbert Spaces (RKHS) for nonlinear system identification and the use of a derivative norm regularization, in place of the traditional functional norm regularization. However in the proposed formulation, an optimal representer for the estimated function cannot be defined. Here, this problem is investigated and a representer for the derivative regularization approach is proposed. Additionally, we show how this permits a fully *a priori* choice of kernel function, determined completely independently of the data distribution and noise level. The advantage of the proposed method is illustrated using two simulation examples, each presenting a scenario where the kernel selection is otherwise highly problematic: non-uniformly distributed data and functions of varying smoothness over the input space.**

Keywords: Nonlinear system identification; nonparametric modeling; RKHS; regularization; derivatives; gradient regularization.

I. Introduction

In system identification, the increasing need for accuracy in complex systems has led to a requirement for models capable of representing the nonlinear structure of many real systems [6], [14]. However, nonlinear system identification is a challenging field, to which many different solutions exist [5], [21], of which almost all benefit from some intuition about the system of interest.

Understandably, such intuition is often scarce. In response to this problem, nonparametric modelling has begun to emerge as a methodology allowing a very flexible representation of a system. Nonetheless, for such approaches the definition of the estimation model class remains a challenging problem. In *Reproducing Kernel Hilbert Space* (RKHS) based approaches [2], [10], this can be interpreted as the kernel selection problem, with the kernel function explicitly defining the model class. In alternative kernel-based framworks, such as *Gaussian Processes* [17], *Kriging* [11] or *Least-Squares Support Vector Machines* [22], the kernel selection is interpreted in different ways - but rests nonetheless problematic. Typically, hyperparameter optimization techniques such as cross-validation are used to determine a suitable configuration - but this can often mean the model selection process is heavily influenced by the noise level and input data characteristics.

Uncertainty concerning the nature of the true system should encourage the user to choose a flexible, unrestrictive model class. However in [4] it was shown that imperfect experimental conditions mean that the user is often forced to accept a compromise between the desired flexibility and the robustness of the model to disturbances. Inspired by RKHS-schemes using derivatives [7], [12], [19] and the analagous spline-smoothing problem [24], a method directly constraining the smoothness of the estimated function was proposed as a way of removing the dependency between smoothness and kernel choice [4]. So far, it has only been shown that the proposed approach could compete with traditional approaches from the literature, whilst determining the kernel function solely based on the input data distribution.

Unfortunately in the analysis of [4], an *a priori* choice of kernel was limited by the suboptimality of the representer [10], [20] in this formulation. In order to overcome this limitation, this paper proposes an *extended representer* as an alternative to the traditional representer. It will be shown that using this extended representer in conjunction with a derivative-norm regularizer allows the definition of an arbitrarily flexible model class for estimation, with significantly reduced sensitivity to the excitation signal and the noise level.

The advantage of the proposed approach with respect to other approaches from the literature will be demonstrated through two examples. Firstly an example using non-uniformly distributed input data will emphasize how traditional kernel optimization can be restrictive under certain conditions. Then, it will be shown how the proposed approach can facilitate the detection of nonlinear strucutres.

This paper is structured as follows. After a brief exposition of the identification problem and the RKHS framework in Section II, the method of [4] is briefly recalled in Section III. Analysis of the representer is presented in Section IV, with how this leads to an *a priori* choice of model structure being shown in Section V. Simulations evaluating the performance of the method are given in Sections VI and VII, with a summary of the principal conclusions in Section VIII.

II. Problem Description

A. The Data-Generating System

Assume N observations of a data-generating system $\mathcal{S}_o$ are measured and recorded:

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

where (u_k, y_k) denotes the input and output of the system at time k. We assume $\mathcal{S}_o$ has an NARX structure such that:

$$\mathcal{S}_o:\ y_{o,k} = f_o(\mathbf{x_k}) + e_{o,k} \tag{2}$$

where the variable $\mathbf{x_k} \in \mathbb{R}^{n_x}$ is the *regressor vector*, composed of the past inputs and ouputs of $\mathcal{S}_o$:

$$\mathbf{x_k} = [y_{k-1}\ \cdots\ y_{k-n_a}\ u_k\ \cdots\ u_{k-n_b}]^\top \tag{3}$$

at each instant $k \in [1 \ldots N]$. Here, $f_o : \mathbb{R}^{n_x} \to \mathbb{R}$ is an unknown nonlinear function and e_o is a white Gaussian noise disturbance at the output.

B. Identification in the RKHS

In the RKHS framework, defining a model class is equivalent to defining a Hilbert space $\mathcal{H}$ of functions such that any $f \in \mathcal{H}$ must have a norm $\|f\|_\mathcal{H} > 0 \quad (\text{for} \quad f \neq 0)$, and an inner-product $\langle f, f\rangle_\mathcal{H}$ [1], [9].

The theory of reproducing kernels states that $\mathcal{H}$ is associated with a unique kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$, spanning the space $\mathcal{H}$ [2]. In other words, the choice of the kernel fully defines $\mathcal{H}$, and hence the model class. From the *Moore-Aronszajn Theorem* [2], this permits the definition of a *reproducing property*, where f is expressed in terms of the kernel function:

$$f(\mathbf{x}) = \langle f, k_\mathbf{x}\rangle_\mathcal{H} = \sum_{i=1}^{\infty} \alpha_i k_{\mathbf{x}_i}(\mathbf{x}), \tag{4}$$

where $\{\alpha_i\}_{i=1}^{\infty} \in \mathbb{R}^\infty$ are weights against the kernel slices $k_{\mathbf{x}_i}$ [20].

C. The Representer Theorem

It can be seen in (4) that f is defined in terms of an infinite sum of kernel slices spanning $\mathcal{X}$. However for certain cost-functions, a finite-dimensional representer of f can be defined, minimizing $\|f\|_\mathcal{H}$. For example, consider:

$$\mathcal{V}_f:\ \mathcal{V}(f) = \|\mathbf{y} - f(\mathbf{x})\|_2^2 + g(\|f\|_\mathcal{H}), \tag{5}$$

where $\mathbf{y} = [y_1 \cdots y_N]^\top$. The term $g(\|f\|_\mathcal{H})$ is the *regularization term*, and must be a *monotonically increasing* function on the norm of f [10], [20]. A popular choice of regularizer is $g(\|f\|_\mathcal{H}) = \lambda\|f\|_\mathcal{H}^2$, which ensures the well-posedness of the solution [23] and will be considered here. In this case, it is proven that the optimal representer of f can be defined as a finite sum around the observations:

$$\mathcal{F}_f:\ f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k_{\mathbf{x}_i}(\mathbf{x}), \quad i = 1, \ldots, N \tag{6}$$

with $\{\alpha_i\}_{i=1}^N \in \mathbb{R}^N$ [10], [20].

Consequently, a closed-form solution can be obtained for $\alpha_f \in \mathbb{R}^N$, the optimal model parameter vector of $\mathcal{V}_f$ of (5), using $\mathcal{F}_f$ of (6):

$$\Rightarrow \quad \alpha_f = (\mathbf{K} + \lambda_f \mathbf{I})^{-1}\mathbf{y}. \tag{7}$$

$\mathbf{K}$ is equivalent to the $N \times N$ *Gram Matrix*, with $\{\mathbf{K}\}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$.

This solution will likely be familiar to the reader, as it is widely used in inverse problems. Determining the optimal parameters α_f of (7) requires the selection of a suitable regularization hyperparameter λ_f and kernel K. Whilst the choice of λ is widely discussed in the literature [15], [17], the question of how to choose the kernel function is more open but equally important.

For example, in case of the well-studied [16] and widely-used Gaussian radial-basis function (RBF):

$$\{\mathbf{K}\}_{i,j} = \exp\left\{-\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{\sigma^2}\right\}, \tag{8}$$

the smaller the kernel width σ, the more flexible the model class. In essence, selecting an excessively large σ defines a model class possibly incapable of fully reconstructing f_o [4].

III. Regularization Using Derivatives of Functions

A. Enforcing Smoothness Through Regularization

The theory of reproducing kernels implies that kernel function explicitly defines the model class for estimation. However in a practical identification context, the effectiveness of $\mathcal{V}_f$ of (5) is undermined if experimental conditions lead to an improper choice of kernel. In [4] the following cost-function was proposed as a solution to this problem:

$$\mathcal{V}_\mathcal{D}:\ \mathcal{V}(f) = \|\mathbf{y} - f(\mathbf{x})\|_2^2 + \lambda\|\mathcal{D}f\|_\mathcal{H}^2, \tag{9}$$

where $\mathcal{D}f$ is the differential operator defining the gradient of f:

$$\mathcal{D}f = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}\ \cdots\ \frac{\partial f(\mathbf{x})}{\partial x_{n_x}}\right]^\top. \tag{10}$$

Here, $\frac{\partial}{\partial x_\mu}\{\cdot\}$ denotes the partial derivative operator with respect to the μ^{th} dimension of $\mathbf{x}$.

In this optimization scheme, the smoothness of $\hat{f}$ can be controlled through the regularization. As a result, the kernel hyperparameter can be removed from the optimization problem, and a flexible kernel function used. This is more likely to ensure that $\mathcal{H}_o \subseteq \mathcal{H}$ (where $\mathcal{H}_o$ is the Hilbert space spanned by f_o).

For a given representer $f(\mathbf{x}) = \sum_{i=1}^P \alpha_i k_{\mathbf{x}_i}(\mathbf{x})$, $\|\mathcal{D}f\|_\mathcal{H}^2$ can be expressed in terms of the derivatives of the kernel [25] as:

$$\|\mathcal{D}f\|_\mathcal{H}^2 = \langle \mathcal{D}f, \mathcal{D}f\rangle = \sum_{i=1}^{P}\sum_{j=1}^{P} \alpha_i\alpha_j \mathcal{D}K(\mathbf{x}_i, \mathbf{x}_j), \tag{11}$$

where $\mathcal{D}K(\mathbf{p}, \mathbf{q}) = \sum_{\mu=1}^{n_x} \frac{\partial^2}{\partial \mathbf{p}_\mu \partial \mathbf{q}_\mu}\{\, K(\mathbf{p}, \mathbf{q})\,\}$.

Under these conditions, the optimal model parameters of $\mathcal{V}_\mathcal{D}$ of (9), $\alpha_\mathcal{D} \in \mathbb{R}^N$, are defined as:

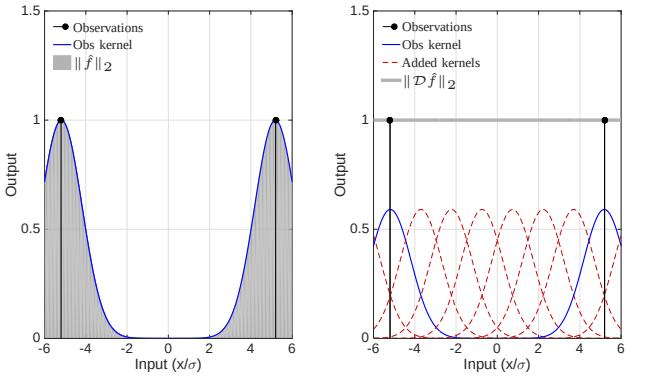
$$\Rightarrow \quad \alpha_\mathcal{D} = \left(\mathbf{K}^\top\mathbf{K} + \lambda_\mathcal{D}\mathcal{D}\mathbf{K}\right)^{-1}\mathbf{K}^\top\mathbf{y} \tag{12}$$

where $\mathcal{D}\mathbf{K}_{\mathbf{i},\mathbf{j}} = \mathcal{D}K(\mathbf{x}_i, \mathbf{x}_j)$ in a similar fashion to (7).

IV. An Extended Representer

In this section, the suboptimality of the representer (6) for solving (9) will be discussed. An extended representer is proposed to cope with this issue, leading to an unconstrained kernel hyperparameter, independently of both the data distribution and the noise. The one-dimensional Gaussian RBF kernel defined in (8) will be used as an example throughout. Note that in this case, extension to the multi-dimensional case is straightforward, by taking the product of the monodimensional kernels [17]. For the Gaussian RBF kernel, the flexibility is defined through the kernel standard deviation σ. Although other choices of kernel function will have different hyperparameters, by identifying the hyperparameter related to the model flexbility the same arguments can be applied.

A. The Suboptimality of $\mathcal{F}_f$ for $\mathcal{V}_\mathcal{D}$



(a) ($\mathcal{V}_f$, $\mathcal{F}_f$): $N = 2$ (b) ($\mathcal{V}_\mathcal{D}$, $\mathcal{F}_\mathcal{D}$): $(N, P) = (2, 8)$

Fig. 1: Reapproaching optimality using added kernels

As pointed out previously, determining $\alpha_\mathcal{D}$ of (12) depends upon choosing a finite representer for $f(\mathbf{x})$. However, the representer (6) is suboptimal here since $\|\mathcal{D}f\|_\mathcal{H}$ is not a *monotically increasing* function on the norm of f. It means here that, in theory, the reproducing property of (4) cannot be truncated to a finite sum over the observations for the cost function (9).

To study this problem, we consider the one-dimensional example of Fig. 1, where $\mathcal{X} \subset \mathbb{R}$ and only $N = 2$ observations of the system are available at $x_1 = -x_2 = -5\sigma$ and $f(x_1) = f(x_2) = 1$. From Fig. 1a, it is clear that $\mathcal{F}_f$ of (6) is the minimal representer of $\mathcal{V}_f$ of (5) as adding kernels along $\mathcal{X}$ would only increase $\|\hat{f}\|_\mathcal{H}$.

However, in the case of $\mathcal{V}_\mathcal{D}$ of (9), $\mathcal{F}_f$ is not optimal. It cannot reproduce the optimal solution of $\mathcal{V}_\mathcal{D}$ in $\mathcal{H}$ (in this case given by the constant function $f_\mathrm{C}(x) = 1$ over the finite interval $x \in [-5\sigma, 5\sigma]$). Furthermore, by adding kernels along $\mathcal{X}$, $\|\mathcal{D}\hat{f}\|_\mathcal{H}$ can be reduced.

Thus, $\mathcal{F}_f$ of (6) is *suboptimal* for $\mathcal{V}_\mathcal{D}$ of (9).

B. An Extended Representer

An infinite amount of kernels would need to be added along $\mathcal{X}$ to minimize $\|\mathcal{D}\hat{f}\|_\mathcal{H}$. Therefore, an optimal representer for $\mathcal{V}_\mathcal{D}$ of (9) cannot be defined. However, the infinite representation of $f(x)$ in (4) can be approximated (see Fig. 1b).

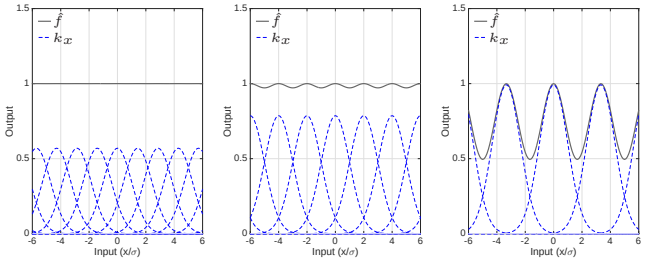
Hence, we propose an *extended* representer of $f(x)$ for $\mathcal{V}_\mathcal{D}$ of (9), based on the true observations and the additional kernels:

$$\mathcal{F}_\mathcal{D} : \; f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k_{\mathbf{x}_i}(\mathbf{x}) + \sum_{j=1}^{P} \alpha_j^* k_{\mathbf{x}_j^*}(\mathbf{x}), \tag{13}$$

where $\mathbf{x_i}^* \in \mathbb{R}^{n_x}, i \in [1 \dots P]$ are points uniformly spanning $\mathcal{X}$.

In this formulation, it is possible to enforce smoothness over $\mathcal{X}$ - through the regularisation - whilst using small kernels.

V. Defining a Maximally Flexible Model Structure



(a) $\rho_k = 0.7$ (b) $\rho_k = 0.5$ (c) $\rho_k = 0.3$

Fig. 2: The effect of the kernel density in $\mathcal{X}$ on $\hat{f}(x)$

Now that the value of σ is unconstrained, the objective is to choose the smallest permissible value, such that the most flexible model structure possible is defined. The only limiting factor is the number of kernels to be added in $\mathcal{F}_\mathcal{D}$ of (13) which is restricted by available computational resources. σ is linked to P through the kernel density, ρ_k:

$$\rho_k = \frac{\sigma}{\Delta_{x^*}}, \quad \Delta_{x^*} = \frac{x^*_{\max} - x^*_{\min}}{P}. \tag{14}$$

However, even though the extended representer is only an approximation of the infinite representation of $f(x)$ in (4), it is now possible to quantify its suboptimality.

Fig. 2 illustrates the dependency of $\hat{f}$ on ρ_k: to ensure smoothness in $\mathcal{X}$, it is required to maintain a sufficiently high kernel density. Hence for a particular P and ρ_k, a minimum value of σ can be determined. To determine ρ_k, let us introduce the smoothness-tolerance parameter ϵ:

$$\epsilon_{\hat{f}} = 100 \; \times \left\{ 1 - \frac{\|\hat{f}\|_\mathrm{inf}}{\mathrm{C}} \right\} \%. \tag{15}$$

ϵ is a user-defined tolerance, quantifying the maximum relative difference between f and $f_\mathrm{C}(x) = \mathrm{C}$, $\forall\, x \in \mathcal{X}$ and $\mathrm{C} \in \mathbb{R}^{+*}$. In practice, the choice of ϵ will depend on the application. To illustrate how this can be used, consider Fig. 3 which approximates ϵ for different values of ρ_k. Here, $\mathbf{x_i}^*$ are evenly distributed along $\mathcal{X} = [-1, 1]$.

Step 1: Define $P_{\max}$. Clearly, increasing P will allow a smaller σ. Hence, P should be chosen as $P_{\max}$, the maximum practical value given computational restrictions (note the requirement for the inversion of $(N+P)\times(N+P)$ matrices). Here, we choose $P = 10^3$.

Step 2: Choose ϵ. A suitably small ϵ should be chosen such that deviations are negligible, but note that this will enfore a larger ρ_k. As an example, we choose $\epsilon_{\max} \approx 10^{-10}\%$, giving $\rho_{k_{\min}} \approx 1.7$.

Step 3: Determine $\sigma_{\min}$. From (14), compute $\sigma_{\min}$. In this example, $\sigma_{\min} \approx 0.003$.

This can be considered a very conservative estimate for σ. For many practical examples, it is sufficient to set $\rho_k = 1$, which ensures $\epsilon \approx 1\%$.

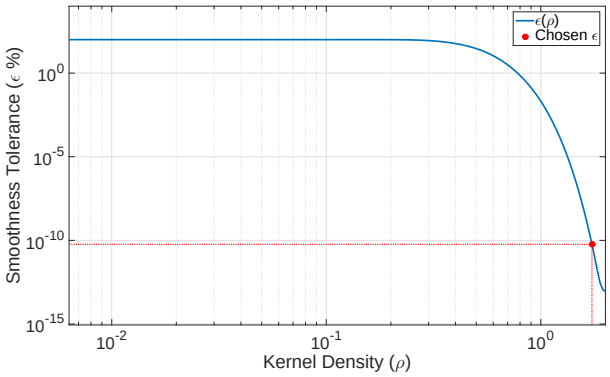


Fig. 3: Selecting an appropriate kernel using ϵ

VI. Simulation Example

A. The Data-Generating System

To illustrate how much the usual kernel optimisation can lead to restrictive model class, the following data generating system $\mathcal{S}_o$ is considered (2):

$$\mathcal{S}_o : \; f_o(x) = 15e^{-5|x|} + 20\sum_{i=0}^{5} e^{-20|x-0.1i|} - 10. \quad (16)$$

with $x \in \mathbb{R}$. Two identification/validation dataset are presented with $N = 50$ points distributed randomly with ($x_i \sim \mathcal{U}(-0.25, 0.25)$, $i = 1, \ldots, N/2$ and $x_i \sim \mathcal{U}(1,1)$, $i = N/2+1, \ldots, N$). The data is corrupted by a white Gaussian noise at the output with SNR $= 20$ dB, where SNR $= 20\log(\sigma_{f_o}/\sigma_e)$. These experimental conditions represent scenarios where the number of available data is low, and unevenly distributed along $\mathcal{X}$.

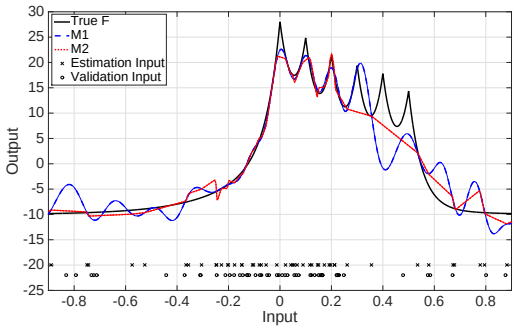
B. Identification Procedure

The identification methods compared in this example and presented in Figure 4 are:

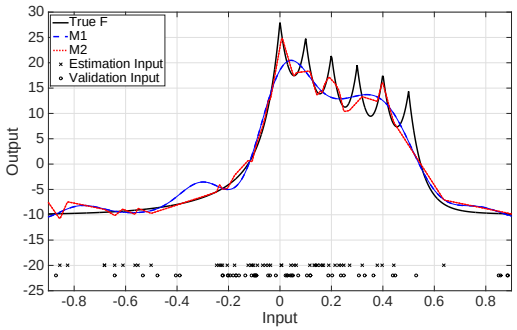
- $\mathcal{M}1$: The regular scheme $\mathcal{V}_f$ using $\mathcal{F}_f$ of (6) with σ optimized by cross validation.
- $\mathcal{M}2$: The proposed scheme $\mathcal{V}_\mathcal{D}$ of (9) using the extended representer $\mathcal{F}_\mathcal{D}$ of (13) and $P_{\max} = 10^3$ leading to $\sigma = \sigma_{\min} = 0.003$.

In both cases, λ is determined using cross-validation against the noisy validation dataset, as would be done in practice.

C. Results



(a) Simulation 1: $x \in [-1, 1]$



(b) Simulation 2: $x \in [-1, 1]$

Fig. 4: Estimation of (16) using $\mathcal{M}1$ and $\mathcal{M}2$

The locations of the identification and validation input data points x are indicated in each figure of Fig. 4. Interestingly, despite the low noise on the data, the optimized kernel of $\mathcal{M}1$ highly depends on the realization: σ was optimized to 0.08 in the first simulation while it was optimized to 0.2 in the second simulation. In simulation 1 (Fig. 4a), both methods can be considered equivalent in terms of model flexibility (even though the behaviour is naturally different in poorly excited regions). However in simulation 2 (Fig 4b), $\mathcal{M}1$ is forced to take an excessively restrictive model class due to an overly large optimized kernel. By contrast, $\mathcal{M}2$ is much better able to cope with the sharpness of f_o in the well excited region (close to $x = 0$), due to the *a priori* choice of a small kernel width.

It should be noted that the fits of both approaches were effectively equivalent in both experiments. However, these results were deemed uninteresting as in this case they provide little insight into the performance of each model on the system.

Despite its apparent simplicity, this example well illustrates a case where the input data distribution prevents a suitable estimation of the kernel through cross validation. In such situations, larger RBF kernels will lead to a poor estimation in densely distributed regions of $\mathcal{X}$. Hence, using $\mathcal{V}_\mathcal{D}$ of (9) and $\mathcal{F}_\mathcal{D}$ of (13), it is possible to define *a priori* a flexible model class, without any consideration of the input distribution or output noise-level.

VII. Locally Optimizing the Smoothness of $\hat{f}$

It has been shown that using $\mathcal{V}_{\mathcal{D}}$ of (9) can permit the definition of a flexible model class. So far however, the ability to characterise different behaviours within a function has been limited by a fixed constraint on the smoothness, determined by the regularization. However, as the smoothness of $\hat{f}$ is now entirely controlled by λ, which appears linearly in the cost function, λ becomes a trade-off between the importance of fitting the data at a given point and the smoothness of the function.

Hence, by introducing linear weights $\omega_k(\mathbf{x}_k)$ into the loss-function, it is possible to locally tune the smoothness at different points in $\mathcal{X}$. The optimization scheme hence becomes:

$$\mathcal{V}_\omega : \mathcal{V}(f) = \sum_{i=1}^{N} \big(\omega_k\big(y_k - f(\mathbf{x}_k)\big)\big)^2 + \lambda \|\mathcal{D}f\|^2_{\mathcal{H}}, \quad (17)$$

which yields the closed-form solution for the optimal model parameters:

$$\alpha_\omega = \left(\mathbf{K}_\omega^\top \mathbf{K}_\omega + \lambda_\omega \mathcal{D}\mathbf{K}\right)^{-1} \mathbf{K}_\omega^\top \mathbf{y}_\omega. \quad (18)$$

with $y_\omega(k) = \omega_k y_k$ and $K_\omega(\mathbf{x}_k, \mathbf{x}) = \omega_k K(\mathbf{x}_k, \mathbf{x})$. Examples of similar approaches can be found in the literature, e.g. [3], [8], [13], [18].

Naturally, the choice of the weights ω_k is open at this point. Nonetheless, a solution to this problem is briefly described here. We propose a direct computation of the weights, based on the second-derivative of $\hat{f}$:

Step 1: Estimate the function $\hat{f}$ using $\mathcal{V}_{\mathcal{D}}$ of (9), which corresponds to (17) with $\omega_k = 1, \forall k \in [1 \dots N]$.

Step 2: Compute the second derivative $\mathcal{D}^2\hat{f}(x)$ by deriving the kernels in $\mathcal{F}_{\mathcal{D}}$ of (13). At this stage it is clear that regions with low values of $\mathcal{D}^2\hat{f}(x)$ indicate smooth regions and vice versa. Therefore, we can compute relative weights $\omega_k \sim \mathcal{D}^2\hat{f}(x_k)$. These weights should be normalised, such that $\bar{\omega} = 1$. This ensures λ does not need to be reoptimized, hence $\lambda_\omega = \lambda_{\mathcal{D}}$.

Step 3: Iterate as long as the fitting score on the validation data improves.

A. A Simulation Example

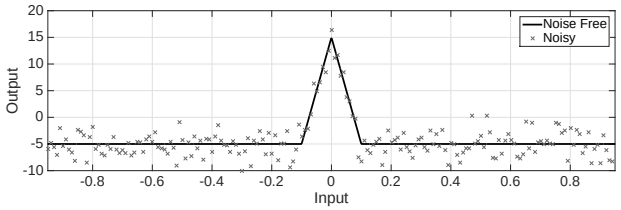
To illustrate the effectiveness of a *non-uniform smoothness* approach, consider the example of Figure 5. N = 1000 measurements of a 1D nonlinear static function f_o are corrupted with white-noise at the output of magnitude 5dB:

$$\mathcal{S}_1 : f_o(x) = \max\{20\,(0.2 - |x|), 0\} - 5. \quad (19)$$

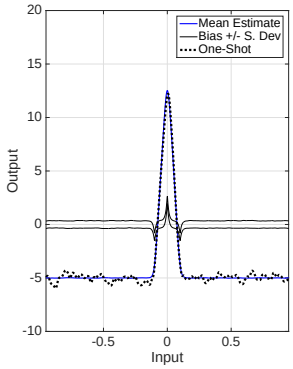
The excitation signal is a uniformly distributed random variable, $x_k \sim U(-1, 1),\ k = 1, \dots, N$.

The approaches of $\mathcal{V}_{\mathcal{D}}$ of (9) and $\mathcal{V}_\omega$ of (17) are used in the estimation. In each case, the kernel hyperparameter was determined *a priori* with $\sigma = 0.01\,(\rho_k \approx 1)$. $\lambda_{\mathcal{D}}$ was determined by cross-validation against a noisy validation dataset.

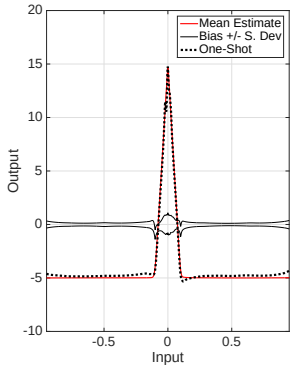
The results of a Monte-Carlo simulation for $n_{mc} = 10^3$ runs are presented in Table 2. The RMSE is calculated as RMSE $= \|\hat{f} - f_o\|_2$, with the mean calculated over n_{mc} trials.



(a) $\mathcal{S}_1$: (19)



(b) $\mathcal{V}_{\mathcal{D}}$: Fixed Smoothness



(c) $\mathcal{V}_\omega$: Variable Smoothness

Fig. 5: Locally applying smoothness using a weighted loss-function

Method	Mean Bias	Mean Std	Mean RMSE
$\mathcal{V}_{\mathcal{D}}$ of (9)	0.094	0.14	15.21
$\mathcal{V}_\omega$ of (17)	0.003	0.11	10.66

TABLE I: Summarized Results of Fig. 5

Table 1 and Fig. 5 show that using $\mathcal{V}_\omega$ of (17), it was possible to better estimate f_o statistically both in terms of bias and variance. Furthermore, examination of the median single-run estimates (i.e. the 500^{th} best estimate in each case) shows that by using $\mathcal{V}_\omega$ of (17) it is also possible to reach a better representation of the true underlying nature of the function f_o, which could be useful as a preliminary step towards an automated parameterization of nonlinear functions.

Finally, as an extension to the approach of $\mathcal{V}_{\mathcal{D}}$ of (9), this proposition still offers the benefits of flexibility in the model class, with the added advantage of better modelling smooth behaviours and sharp changes without any kernel size optimization.

VIII. Conclusions

In this paper, after recalling the recently introduced optimization scheme over derivatives in the RKHS framework, an extended representer has been proposed which allows the *a priori* definition of a maximally flexible model structure, given available computational resources and desired precision.

In practice, this means an arbitrarily small kernel can be used for the estimation of both smooth and nonsmooth functions, irrespective of the input distribution and the noise-level. This greatly reduces the complexity of the hyper-parameter optimization problem, which may be crucial in higher dimensional problems.

Two examples have been presented, illustrating the advantages of the proposed method in cases where experimental conditions otherwise pose undesired limitations on the choice of kernel.

The authors are currently extending this research to dynamical examples, to investigate the application of the theory presented in more common system identification scenarios.

References

[1] N. Akhiezer and I. Glazman. *Theory of Linear Operators in Hilbert Space*. Ungar, New York, 1963.

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

[3] E. Bai and Y. Liu. Recursive direct weight optimization in nonlinear system identification: a minimal probability approach. *IEEE Transactions on Automatic Control*, 52:1218–1231, 2007.

[4] Y. Bhujwalla, V. Laurain, and M. Gilson. The impact of smoothness on model class selection in nonlinear system identification: An application of derivatives in the RKHS. In *Accepted to IEEE American Control Conference (ACC)*, Boston, Massachusetts, USA, July 2016.

[5] S. A. Billings. Identification of nonlinear systems - a survey. *IEEE Proceedings of Control Theory*, 127:272–285, 1980.

[6] 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.

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

[8] M. Gohnen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[9] P. Halmos. *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*. Chelsea, New York, 1957.

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

[11] D. G. Krige. A study of gold and uranium distribution patterns in the klerksdorp gold field. *Geoexploration*, 4(1):43–53, 1966.

[12] F. Lauer, V. Le, and G. Bloch. Learning smooth models of nonsmooth functions via convex optimization. In *IEEE International Workshop on Machine Learning for Signal Processing*, Santander, Spain, 2012.

[13] T. Lin, B. Horne, P. Tino, and C. Giles. Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Automatic Control*, 7:1329–1338, 1996.

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

[15] 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.

[16] G. Pillonetto, M. Quang, and A. Chiuso. A new kernel-based approach for nonlinear system identification. *IEEE Transactions on Automatic Control*, 56(12):2825–2840, 2011.

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

[18] J. Roll, A. Nazin, and L. Ljung. Nonlinear system identification via direct weight optimization. *Automatica*, 41:475–490, 2005.

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

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

[21] 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.

[22] J. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

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

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

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