Automatica 93 (2018) 321–332

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

System identification using kernel-based regularization: New insights on stability and consistency issues☆

Gianluigi Pillonetto

Department of Information Engineering, University of Padova, Padova, Italy

ARTICLE INFO

Article history:
Received 28 December 2016
Received in revised form 6 February 2018
Accepted 26 February 2018

Keywords:
Learning from examples
System identification
Reproducing kernel Hilbert spaces of dynamic systems
Kernel-based regularization
BIBO stability
Regularization networks
Generalization and consistency

ABSTRACT

Learning from examples is one of the key problems in science and engineering. It deals with function reconstruction from a finite set of direct and noisy samples. *Regularization in reproducing kernel Hilbert spaces* (RKHSs) is widely used to solve this task and includes powerful estimators such as regularization networks. Recent achievements include the proof of the statistical consistency of these kernel-based approaches. Parallel to this, many different system identification techniques have been developed but the interaction with machine learning does not appear so strong yet. One reason is that the RKHSs usually employed in machine learning do not embed the information available on dynamic systems, e.g. BIBO stability. In addition, in system identification the independent data assumptions routinely adopted in machine learning are never satisfied in practice. This paper provides some new results which strengthen the connection between system identification and machine learning. Our starting point is the introduction of RKHSs of dynamic systems. They contain functionals over spaces defined by system inputs and allow to interpret system identification as learning from examples. In both linear and nonlinear settings, it is shown that this perspective permits to derive in a relatively simple way conditions on RKHS stability (i.e. the property of containing only BIBO stable systems or predictors), also facilitating the design of new kernels for system identification. Furthermore, we prove the convergence of the regularized estimator to the optimal predictor under conditions typical of dynamic systems.

1. Introduction

Learning from examples is key in science and engineering, considered at the core of intelligence's understanding (Poggio & Shelton, 1999). In mathematical terms, it can be described as follows. We are given a finite set of training data (x_i, y_i), where x_i is the so called input location while y_i is the corresponding output measurement. The goal is then the reconstruction of a function with good prediction capability on future data: for a new pair (x, y), the prediction $g(x)$ should be close to y.

To solve this task, nonparametric techniques have been extensively used in the last years. Within this paradigm, instead of assigning to the unknown function a specific parametric structure, g is searched over a possibly infinite-dimensional functional space. The modern approach uses Tikhonov regularization theory (Bertero, 1989; Tikhonov & Arsenin, 1977) in conjunction with Reproducing Kernel Hilbert Spaces (RKHSs) (Aronszajn, 1950; Bergman, 1950). RKHSs possess many important properties, being in one to one correspondence with the class of positive definite kernels. Their connection with Gaussian processes is also described in Aravkin, Bell, Burke, and Pillonetto (2015), Bell and Pillonetto (2004), Kimeldorf and Wahba (1970) and Lukic and Beder (2001).

While applications of RKHSs in statistics, approximation theory and computer vision trace back to Bertero, Poggio, and Torre (1988), Poggio and Girosi (1990) and Wahba (1990), these spaces were introduced to the machine learning community in Girosi (1997). RKHSs permit to treat in a unified way many different regularization methods. The so called kernel-based methods (Evgeniou, Pontil, & Poggio, 2000; Schölkopf & Smola, 2001) include smoothing splines (Wahba, 1990), regularization networks (Poggio & Girosi, 1990), Gaussian regression (Rasmussen & Williams, 2006), and support vector machines (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997; Vapnik, 1998). In particular, a regularization network (RN) has the structure

$$\hat{g} = \arg\min_{f \in \mathscr{H}} \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{N} + \gamma \|f\|^2_{\mathscr{H}} \qquad \text{RN} \tag{1}$$

where $\mathscr{H}$ denotes a RKHS with norm $\|\cdot\|_{\mathscr{H}}$.

☆ This research has been partially supported by the Progetto di Ateneo CPDA147754/14-New statistical learning approach for multi-agents adaptive estimation and coverage control. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the direction of Editor Torsten Söderström. The author would like to thank Isabella Vendramini for helpful discussions on the subject.

E-mail address: giapi@dei.unipd.it.

https://doi.org/10.1016/j.automatica.2018.03.065
0005-1098/

Thus, the function estimate minimizes an objective sum of two contrasting terms: the quadratic loss which measures the adherence to experimental data and the regularizer (the RKHS squared norm) which restores the well-posedness. Finally, the positive scalar γ is the regularization parameter which has to suitably trade off these two components.

The use of (1) has significant advantages. The choice of an appropriate RKHS, often obtained just including function smoothness information (Schölkopf & Smola, 2001), and a careful tuning of γ, e.g. by the empirical Bayes approach (Aravkin, Burke, Chiuso, & Pillonetto, 2012, 2014; Maritz & Lwin, 1989), can well balance bias and variance. One can thus obtain favorable mean squared error properties. Furthermore, even if $\mathscr{H}$ is infinite-dimensional, the solution $\hat{g}$ is always unique, belongs to a finite-dimensional subspace and is available in closed-form. This result comes from the representer theorem (Argyriou & Dinuzzo, 2014; Argyriou, Micchelli, & Pontil, 2009; Kimeldorf & Wahba, 1970; Schölkopf, Herbrich, & Smola, 2001). Building upon the work (Wahba, 1977), many new results have been also recently obtained on the statistical consistency of (1). In particular, the property of $\hat{g}$ to converge to the optimal predictor as the data set size grows to infinity is discussed e.g. in Mukherjee, Niyogi, Poggio, and Rifkin (2006), Poggio, Rifkin, Mukherjee, and Niyogi (2004), Smale and Zhou (2007), Wu, Ying, and Zhou (2006) and Yuan and Tony Cai (2010). Parallel to this, many system identification techniques have been developed in the last decades. In linear contexts, the first regularized approaches trace back to Akaike (1979), Kitagawa and Gersch (1996) and Schiller (1979), see also Goodwin, Gevers, and Ninness (1992) and Ljung, Goodwin, and Agero (2014) where model error is described via a nonparametric structure. More recent approaches, also inspired by nuclear and atomic norms (Chandrasekaran, Recht, Parrilo, & Willsky, 2012), can instead be found in Grossmann, Jones, and Morari (2009), Liu and Vandenberghe (2009), Mohan and Fazel (2010), Pillonetto, Chen, Chiuso, De Nicolao, and Ljung (2016) and Rojas, Toth, and Hjalmarsson (2014). In the last years, many nonparametric techniques have been proposed also for nonlinear system identification. They exploit e.g. neural networks (Lin, Horne, Tino, & Giles, 1996; Shun-Feng & Yang, 2002), Volterra theory (Franz & Schölkopf, 2006), kernel-type estimators (Leithead, Solak, & Leith, 2003; Pillonetto, Chiuso, & Quang, 2011b; Zhao, Chen, Bai, & Li, 2015) which include also weights optimization to control the mean squared error (Bai & Liu, 2007; Bai, 2010; Roll, Nazin, & Ljung, 2005). Important connections between kernel-based regularization and nonlinear system identification have been also obtained by the least squares support vector machines (Suykens, Alzate, & Pelckmans, 2010; Suykens, Van Gestel, Brabanter, De Moor, & Vandewalle, 2002) and using Gaussian regression for state space models (Frigola, Lindsten, Schon, & Rasmussen, 2013; Frigola & Rasmussen, 2013). Most of these approaches are inspired by machine learning, a fact not surprising since predictor estimation is at the core of the machine learning philosophy. Indeed, a black-box relationship can be obtained through (1) using past inputs and outputs to define the input locations (regressors). However, the kernels currently used for system identification are those conceived by the machine learning community for the reconstruction of *static* maps. RKHSs suited to linear system identification have been proposed only recently, e.g. in computer vision exploiting compound matrices built from system trajectories (Vishwanathan, Smola, & Vidal, 2007) or by introducing *stable spline kernels* which embed information on impulse response regularity and stability (Chen, Ohlsson, & Ljung, 2012; Pillonetto, Chiuso, & De Nicolao, 2011a; Pillonetto & De Nicolao, 2010). Furthermore, while stability of a RKHS (i.e. its property of containing only stable systems or predictors) is treated in Carmeli, De Vito, and Toigo (2006), Dinuzzo (2015) and Pillonetto, Dinuzzo, Chen, Nicolao, and Ljung (2014), the nonlinear scenario still appears unexplored. Beyond stability, we also notice that the most used kernels for nonlinear regression, e.g. the Gaussian (Schölkopf & Smola, 2001), do not include other important information on dynamic systems like the fact that output energy is expected to increase if input energy augments.

Another aspect that weakens the interaction between system identification and machine learning stems also from the (apparently) different contexts these disciplines are applied to. In machine learning one typically assumes that data (x_i, y_i) are i.i.d. random vectors assuming values on a bounded subset of the Euclidean space. But in system identification, even when the system input is white noise, the input locations are not mutually independent. Already in the classical Gaussian noise setting, the outputs are not even bounded, i.e. there is no compact set containing them with probability one. Remarkably, this implies that none of the aforementioned consistency results developed for kernel-based methods can be applied. Some extensions to the case of correlated samples can be found in Guo and Zhou (2013), Steinwart, Hush, and Scovel (2009), Vidyasagar (1997) and Wang and Zhou (2011) but still under conditions far from the system identification setting.

In this paper we provide some new insights on the interplay between system identification and machine learning in a RKHS setting. Our starting point is the introduction of what we call *RKHSs of dynamic systems* which contain functionals over input spaces $\mathscr{X}$ induced by system inputs u. More specifically, each input location $x \in \mathscr{X}$ contains a piece of the trajectory of u so that any $g \in \mathscr{H}$ can be associated to a dynamic system. When u is a stationary stochastic process, its distribution then defines the probability measure on $\mathscr{X}$ from which the input locations are drawn. Again, we stress that this framework has been (at least implicitly) used in previous works on nonlinear system identification, see e.g. Lin et al. (1996), Pillonetto, Chiuso, Quang, et al. (2011b), Sjöberg et al. (1995), Shun-Feng and Yang (2002) and Suykens et al. (2002). However, it has never been cast and studied in its full generality under a RKHS setting.

Even if in this context the input space can turn out unbounded (e.g. when the system input is Gaussian) and complex (e.g. $\mathscr{X}$ is a function space itself in continuous-time), it will be shown that our perspective is key to obtain the following achievements:

- linear and nonlinear system identification can be treated in a unified way in both discrete- and continuous-time. Thus, the estimator (1) can be used in many different contexts, relevant for the control community, just changing the RKHS. This is important for the development of a general theory which links regularization in RKHS and system identification;
- system input's role in determining the nature of the RKHS is made explicit. This will be also described in more detail in the linear system context, illustrating the distinction between the concept of RKHSs $\mathscr{H}$ of dynamic systems and that of RKHSs $\mathscr{I}$ of impulse responses;
- for linear systems we provide a new and simple derivation of the necessary and sufficient condition for RKHS stability (Carmeli et al., 2006; Dinuzzo, 2015; Pillonetto et al., 2014) that relies just on basic RKHS theory;
- in the nonlinear scenario, we obtain a sufficient condition for RKHS stability which has wide applicability. We also derive a new stable kernel for nonlinear system identification;
- consistency of the RN (1) is proved under assumptions suited to system identification, revealing the link between consistency and RKHS stability.

The paper is organized as follows. In Section 2 we provide a brief overview on RKHSs. In Section 3, the concept of RKHSs of dynamic systems is defined by introducing input spaces $\mathscr{X}$ induced by system inputs. The case of linear dynamic systems is then detailed via its relationship with linear kernels. The difference between the