Research Article

Kernel methods for the approximation of discrete-time linear autonomous and control systems

Boumediene Hamzi[1] (ID) · **Fritz Colonius**[2]

Abstract

Methods from learning theory are used in the state space of linear dynamical and control systems in order to estimate relevant matrices and some relevant quantities such as the topological entropy. An application to stabilization via algebraic Riccati equations is included by viewing a control system as an autonomous system in an extended space of states and control inputs. Kernel methods are the main techniques used in this paper and the approach is illustrated via a series of numerical examples. The advantage of using kernel methods is that they allow to perform function approximation from data and, as illustrated in this paper, allow to approximate linear discrete-time autonomous and control systems from data.

1 Introduction

This paper discusses several problems in dynamical systems and control, where methods from learning theory are used in the state space of linear systems. This is in contrast to previous approaches in the frequency domain [10, 11]. We refer to [11] for a general survey on applications of machine learning to system identification where similar problems have been treated using different techniques.

Basically, learning theory allows to deal with problems when only data from a given system are given. Reproducing Kernel Hilbert spaces (RKHS) allow to work in a very large dimensional space in order to simplify the underlying problem. We will discuss this in the simple case when the matrix A describing a linear discrete-time system is unknown, but a time series from the underlying linear dynamical system is given. We propose a method to estimate the underlying matrix using kernel methods. Applications are given in the stable and unstable case and for estimating the eigenvalues and topological entropy for a

linear map. Furthermore, in the control case, estimation of the relevant matrices for a linear control system is done by viewing a linear control system as a dynamical system in the extended space of states and control inputs. Stabilization via linear-quadratic optimal control is discussed.

The emphasis of the present paper is on the formulation of a number of problems in dynamical systems and control and to illustrate the applicability of our approach via a series of numerical examples. This paper should be viewed as a preliminary step to extend these results to nonlinear discrete-time systems within the spirit of [3, 4] where the authors showed that RKHSs act as "linearizing spaces" and that this approach offers tools for a data-based theory for nonlinear (continuous-time) dynamical systems. The approach used in these papers is based on *embedding a nonlinear system in a high (or infinite) dimensional reproducing kernel Hilbert space (RKHS) where linear theory is applied*. To illustrate this approach, consider a polynomial in $\mathbb{R}$, $p(x) = \alpha + \beta x + \gamma x^2$ where α, β, γ are real numbers. If we consider the map $\phi \colon \mathbb{R} \to \mathbb{R}^3$ defined as $\phi(x) = [1 \ x \ x^2]^T$

✉ Boumediene Hamzi, boumediene.hamzi@gmail.com; Fritz Colonius, fritz.colonius@math.uni-augsburg.de | ¹Department of Mathematics, Imperial College London, London, UK. ²Institut für Mathematik, Universität Augsburg, Augsburg, Germany.

then $p(x) = \alpha \cdot [1\ x\ x^2]^T = \alpha \cdot \phi(x)$ is an affine polynomial in the variable $\phi(x)$. Similarly, consider the nonlinear discrete-time system $x(k+1) = x(k) + x^2(k)$. By rewriting it as $x(k+1) = [1\ 1] \begin{bmatrix} x(k) \\ x(k)^2 \end{bmatrix}$, the nonlinear system becomes linear in the variable $[x(k)\ x(k)^2]$.

The contents are as follows: In Sect. 2, the problem is stated formally and an algorithm based on kernel methods is given for the stable case. In Sect. 3, the algorithm is extended to the unstable case. In particular, the topological entropy of linear maps is computed (which boils down to computing unstable eigenvalues). In Sect. 4, identification of linear control systems is considered and Sect. 5 discusses their stabilization. Every section contains several numerically computed examples (via MATLAB) illustrating the approach. Section 6 draws some conclusions from the numerical experiments. For the reader's convenience we have collected in the appendix basic concepts from learning theory as well as some hints to the relevant literature. A preliminary version of this article appeared in Hamzi and Colonius [9]

2 Statement of the problem

Consider the linear discrete-time system

$$x(k+1) = Ax(k), \tag{1}$$

where $A = [a_{i,j}] \in \mathbb{R}^{n \times n}$. We want to estimate A from the time series $x(1) + \eta_1, \cdots, x(N) + \eta_N$ where the initial condition $x(0)$ is known and η_i are distributed according to a probability measure ρ_x that satisfies the following condition (this is the *Special Assumption* in [18]).

Assumption The measure ρ_x is the marginal on $X = \mathbb{R}^n$ of a Borel measure ρ on $X \times \mathbb{R}$ with zero mean supported on $[-M_x, M_x], M_x > 0$.

One obtains from (1) for the components of the time series that

$$x_i(k+1) = \sum_{j=1}^{n} a_{ij} x_j(k). \tag{2}$$

For every i we want to estimate the coefficients $a_{ij}, j = 1, \ldots, n$. They are determined by the linear maps $f_i^* : \mathbb{R}^n \to \mathbb{R}$ given by

$$(x_1, \ldots, x_n) \mapsto \sum_{j=1}^{n} a_{ij} x_j. \tag{3}$$

This problem can be reformulated as a learning problem as described in the "Appendix" where f_i^* in (3) plays the role of the unknown function (74) and $(x(k), x_i(k+1) + \eta_i)$ are the samples in (76).

We note that in [18], the authors do not consider time series and that we apply their results to time series.

In order to approximate f_i^*, we minimize the criterion in (79). For a positive definite kernel K, let f_i be the kernel expansion of f_i^* in the corresponding RKHS $\mathcal{H}_K$. Then $f_i = \sum_{j=1}^{\infty} c_{i,j} \phi_j$ with certain coefficients $c_{ij} \in \mathbb{R}$ and

$$\|f_i\|_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{c_{i,j}^2}{\lambda_j}, \tag{4}$$

where (λ_j, ϕ_j) are the eigenvalues and eigenfunctions of the integral operator $L_K : \mathcal{L}_\nu^2(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ given by $(L_K f)(x) = \int K(x,t) f(t) d\nu(t)$ with a Borel measure ν on $\mathcal{X}$. Thus $L_K \phi_j = \lambda_j \phi_j$ for $j \in \mathbb{N}^*$ and the eigenvalues $\lambda_j \geq 0$.

Then we consider the problem of minimizing over $(c_{i,1}, \ldots, c_{i,N})$ the functional

$$\mathcal{E}_i = \frac{1}{N} \sum_{k=1}^{N} (y_i(k) - f_i(x(k)))^2 + \gamma_i \|f_i\|_{\mathcal{H}_K}^2, \tag{5}$$

where $y_i(k) := x_i(k+1) + \eta_i = f_i^*(x(k)) + \eta_i$ and γ_i is a regularization parameter.

Since we are dealing with a linear problem, it is natural to choose the linear kernel $k(x,y) = \langle x, y \rangle$. Then the solution of the above optimization problem is given by the kernel expansion of $x_i(k+1), i = 1, \ldots, n$,

$$y_i(k) := x_i(k+1) = \sum_{j=1}^{N} c_{ij} \langle x(j), x(k) \rangle, \tag{6}$$

where the c_{ij} satisfy the following set of equations:

$$\begin{bmatrix} x_i(1) \\ \vdots \\ x_i(N) \end{bmatrix} = \left(N \lambda I_d + \mathbb{K} \right) \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iN} \end{bmatrix}, \tag{7}$$

with

$$\mathbb{K} := \begin{bmatrix} \sum_{\ell=1}^{n} x_\ell(1) x_\ell(0) & \cdots & \sum_{\ell=1}^{n} x_\ell(N) x_\ell(0) \\ \vdots & \cdots & \vdots \\ \sum_{\ell=1}^{n} x_\ell(1) x_\ell(N-1) & \cdots & \sum_{\ell=1}^{n} x_\ell(N) x_\ell(N-1) \end{bmatrix}. \tag{8}$$

This is a consequence of Theorem 2.

From (2), we have

$$x_i(k+1) = \sum_{j=1}^{N} c_{ij} \langle x(j), x(k) \rangle = \sum_{j=1}^{N} c_{ij} x(j)^T \cdot x(k)$$

$$= \sum_{j=1}^{N} \sum_{\ell=1}^{n} c_{ij} x_\ell(j) x_\ell(k)$$

$$= \sum_{\ell=1}^{n} \sum_{j=1}^{N} c_{ij} x_\ell(j) x_\ell(k).$$

Then an estimate of the entries of A is given by

$$\hat{a}_{i\ell} = \sum_{j=1}^{N} c_{i,j} x_\ell(j). \tag{9}$$

This discussion leads us to the following basic algorithm.

Algorithm $\mathcal{A}$ If the eigenvalues of A are all within the unit circle, one proceeds as follows in order to estimate A. Given the time series $x(1), \dots, x(N)$ solve the system of Eq. (7) to find the numbers c_{ij} and then compute $\hat{a}_{i\ell}$ from (9).

Before we present numerical examples and modifications and applications of this algorithm, it is worthwhile to note the following preliminary remarks indicating what may be expected.

The stability assumption in algorithm $\mathcal{A}$ is imposed, since otherwise the time series will diverge exponentially. Then, already for a moderately sized number of data points ($N \approx 10^2$) Eq. (7) will be ill conditioned. Hence for unstable A, modifications of algorithm $\mathcal{A}$ are required.

While for test examples one can compare the entries of the matrix A and its approximation $\hat{A}$, it may appear more realistic to compare the values $x(1), \dots, x(N)$ of the data series and the values $\hat{x}(1), \dots, \hat{x}(N)$ generated by the iteration of the matrix $\hat{A}$.

In general, one should not expect that increasing the number of data points will lead to better approximations of the matrix A. If the matrix A is diagonalizable, for generic initial points $x(0) \in \mathbb{R}^n$ the data points $x(k)$ will approach, for $N \to \infty$, the eigenspace for the eigenvalue with maximal modulus. For general A and generic initial points $x(0) \in \mathbb{R}^n$, the data points $x(N)$ will approach for, $N \to \infty$, the largest Lyapunov space (i.e., the sum of the real generalized eigenspaces for eigenvalues with maximal modulus). Thus in the limit for $N \to \infty$, only part of the matrix can be approximated. A detailed discussion of this (well known) limit behavior is, e.g., given in Colonius and Kliemann [6]. A consequence is that a medium length of the time series should be adequate.

This problem can be overcome by choosing the regularization parameter γ in (5) and (7) using the method of cross validation described in [12]. Briefly, in order to choose γ, we consider a set of values of regularization parameters: we run the learning algorithm over a subset of the samples for each value of the regularization parameter and choose the one that performs the best on the remaining data set. Cross validation helps also in the presence of noise and to improve the results beyond the training set.

A theoretical justification of our algorithm is guaranteed by the error estimates in Theorem 5. In fact, for the linear dynamical system (1), we have that f^* in (74) is the linear map $f^*(x) = f_i(x)$ in (3) and the samples $\mathbf{s}$ in (76) are $(x(k), x_i(k+1) + \eta_i)$. Moreover, by choosing the linear kernel $k(x, y) = \langle x, y \rangle$ we get that $f^* \in \mathcal{H}_K$. In this case, (84) has the form

$$\|\hat{x}_i(k+1) - x_i(k+1)\|^2 \leq 2C_{\bar{x}}\mathcal{E}_{\text{samp}} + 2\|x(k+1)\|_K^2 (\gamma + 8C_{\bar{x}}\Delta), \tag{10}$$

where $\|x_i(k+1)\|_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{c_{ij}^2}{\lambda_j}$. See [3] for more details about error estimates in the general nonlinear case.

The first term in the right hand side of inequality (10) represents the error due to the noise (sampling error) and the second term represents the error due to regularization (regularization error) and the finite-number of samples (integration error).

Next, we discuss several numerical examples, beginning with the following scalar equation.

Example 1 Consider $x(k+1) = \alpha x(k)$ with $\alpha = 0.5$. With the initial condition $x(0) = -0.5$, we generate the time series $x(1), \dots, x(100)$. Applying algorithm $\mathcal{A}$ with the regularization parameter $\gamma = 10^{-6}$ we compute $\hat{\alpha} = 0.4997$. Using cross validation, we get that $\hat{\alpha} = 0.5$ with regularization parameter $\gamma = 1.5259 \times 10^{-5}$. When we introduce an i.i.d perturbation signal $\eta_i \in [-0.1, 0.1]$, the algorithm does not behave well when we fix the regularization parameter. With cross validation, the algorithm works quite well and the regularization parameter adapts to the realization of the signal η_i. Here, for $e(k) = x(k) - \hat{x}(k)$ with $x(k+1) = \alpha x(k)$ and $\hat{x}(k+1) = \hat{\alpha}\hat{x}(k)$, we get that $\|e(300)\| = \sqrt{\sum_{i=1}^{300} e^2(i)} = 0.0914$ and $\sqrt{\sum_{i=100}^{300} e^2(i)} = 1.8218 \times 10^{-30}$.

We observe an analogous behavior of the algorithm when the data are generated from $x(k+1) = \alpha x(k) + \epsilon x(k)^2$ where the algorithm works well in the presence of noise and structural perturbations when using cross validation. When $\epsilon = 0.1$ and with an i.i.d perturbation signal $\eta_i \in [-0.1, 0.1]$, $\hat{\alpha}$ varies between 0.38 and 0.58 depending on the realization of η_i but $\|e(300)\| = \sqrt{\sum_{i=1}^{300} e^2(i)} = 0.2290$ and $\sqrt{\sum_{i=100}^{300} e^2(i)} = 2.8098 \times 10^{-30}$ which shows that the error e decreases exponentially and the generalization properties of the algorithm are quite good.

Example 2 Consider $x(k+1) = Ax(k)$ with matrix A given by

$$A := \begin{bmatrix} -0.5 & 1 & 0 & 0 \\ 0 & 0.6 & 1 & 0 \\ 0 & 0 & 0.7 & 1 \\ 0 & 0 & 0 & -0.8 \end{bmatrix}. \tag{11}$$

For the initial condition $x = [-0.9, 0.1, 15, 0.2]'$ and with $N = 100$ data points, we get

$$\hat{A} = \begin{bmatrix} -0.5000 & 1.0000 & 0.0000 & -0.0000 \\ 0.0000 & 0.6000 & 1.0000 & 0.0000 \\ 0.0000 & -0.0000 & 0.7000 & 0.9994 \\ -0.0000 & 0.0000 & -0.0000 & -0.7995 \end{bmatrix}.$$
(12)

We then simulate $x(k + 1) = Ax(k)$ and $\hat{x}(k + 1) = \hat{A}\hat{x}(k)$ for $k = 0, \dots, 200$ to test the accuracy of our approximation beyond the interval $k = 0, \dots, 100$. Then the norm of the error $e_j(k) = x_j(k) - \hat{x}_j(k)$, for $j = 1, \dots, 4$, $\|e_j(300)\| = \sqrt{\sum_{i=1}^{300} e_j^2(i)}$ is of the order of 10^{-3} and $\sqrt{\sum_{i=100}^{300} e_j^2(i)}$ is of the order of 10^{-11} which shows that the error e decreases exponentially and the generalization properties of the algorithm are quite good. The regularization parameters are $\gamma_i = 0.9313 \times 10^{-9}$ for $i = 1, \dots, 4$.

Also in the presence of small noise $\eta_i \in [-0.01, 0.01]$, the algorithm behaves well and the regularization parameters adapt to the realization of η_i. For example, for a certain realizations of η_i, we obtain the regularization parameters

$$\gamma_1 = 0.0039, \gamma_2 = 2.4114 \times 10^{-4},$$
$$\gamma_3 = 9.3132 \times 10^{-10}, \gamma_4 = 2 \times 10^{-3}$$
(13)

and the error $\|e_j(300)\| = \sqrt{\sum_{i=1}^{300} e_j^2(i)}$ is of the order of 10^{-1} and $\sqrt{\sum_{i=100}^{300} e_j^2(i)}$ is of the order of 10^{-9}.

Suppose that in addition to a small noise $\eta_i \in [-0.01, 0.01]$, there is a quadratic structural perturbation, i.e.,

$$x(k + 1) = Ax(k) + \varepsilon \begin{bmatrix} x_1(k)^2 \\ x_2(k)^2 \\ x_3(k)^2 \\ x_4(k)^2 \end{bmatrix}.$$
(14)

Then with cross validation for $\varepsilon = 0.001$ the algorithm behaves well. For a particular realization of η, the error $\|e_j(300)\| = \sqrt{\sum_{i=1}^{300} e_j^2(i)}$ is between 5 and 15 but $\sqrt{\sum_{i=100}^{300} e_j^2(i)}$ is of the order of 10^{-9} and the regularization parameters are

$$\gamma_1 = 0.5, \gamma_2 = 9.3132 \times 10^{-10}, \gamma_3 = 9.3132 \times 10^{-10},$$
$$\gamma_4 = 9.3132 \times 10^{-10}.$$
(15)

These examples show a very good behavior of the algorithm.

3 Unstable case

Consider

$$x(k + 1) = Ax(k) \text{ with } A \in \mathbb{R}^{n \times n},$$
(16)

where some of the eigenvalues of A are outside the unit circle. Again, we want to estimate A when the following data are given,

$$x(1), x(2), \dots, x(N),$$
(17)

which are generated by system (16), thus $x(k) = A^{k-1}x(1)$.

As remarked above, a direct application of the algorithm $\mathcal{A}$ will not work, since the time series diverges fast. Instead, we construct a new time series from (17) associated to an auxiliary stable system.

For a constant $\sigma > 0$ we define the auxiliary system by

$$y(k + 1) = \tilde{A}y(k) \text{ with } \tilde{A} := \frac{1}{\sigma}A.$$
(18)

Thus

$$y(k) = \left(\frac{A}{\sigma}\right)^{k-1} y(1)$$
(19)

and with $y(1) = x(1)$ one finds

$$y(k) = \frac{1}{\sigma^{k-1}} A^{k-1} x(1) = \frac{1}{\sigma^{k-1}} x(k).$$
(20)

If we choose $\sigma > 0$ such that the eigenvalues of $\frac{A}{\sigma}$ are in the unit circle, we can apply algorithm $\mathcal{A}$ to this stable matrix and hence we would obtain an estimate of $\frac{A}{\sigma}$ and hence of A. However, since the eigenvalues of the matrix A are unknown, we will be content with a somewhat weaker condition than stability of $\frac{A}{\sigma}$.

The data (17) for system (16) yield the following data for system (18):

$$y(1) := x(1), y(2) := \frac{1}{\sigma}x(2), \dots, y(N) := \frac{1}{\sigma^{N-1}}x(N). \quad (21)$$

We propose to choose σ as follows: Define

$$\sigma := \max\left\{\frac{\|x(k+1)\|}{\|x(k)\|}, k \in \{0, 1, \dots, N\}\right\}.$$
(22)

Clearly, the inequality $\sigma \leq \|A\|$ holds. We apply algorithm $\mathcal{A}$ to the time series $y(k)$. This yields an estimate of $\frac{A}{\sigma}$ and hence an estimate $\hat{A}$ of A.

For general A, this choice of σ certainly does not guarantee that the eigenvalues of $\frac{A}{\sigma}$ are within the unit circle. However, as mentioned above, a generic data sequence $x(k), k \in \mathbb{N}$, will converge to the eigenspace of the eigenvalue with maximal modulus. Hence $\frac{\|x(k+1)\|}{\|x(k)\|}$ will approach the maximal modulus of an eigenvalue, thus this choice of σ will lead to a matrix $\frac{A}{\sigma}$ which is not "too unstable".

Example 3 Consider $x(k + 1) = \alpha x(k)$ with $\alpha = 11.46$. With the initial condition $x(0) = -0.5$, we generate the time series $x(1), \ldots, x(100)$. The algorithm above with the regularization parameter $\gamma = 10^{-6}$ yields the estimate $\hat{\alpha} = 11.4086$. Cross validation leads to the regularization parameter $\gamma = 9.5367 \times 10^{-7}$ and the estimate $\hat{\alpha} = 11.4599$. In the presence of a small noise $\eta \in [-0.1, 0.1]$, cross validation yields the regularization parameter $\gamma = 0.002$ and the slightly worse estimate $\hat{\alpha} = 11.1319$.

We observe the same behavior in higher dimensional systems where the eigenvalues are of the same order of magnitude.

Example 4 Consider $x(k + 1) = Ax(k)$ with

$$A = \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & -10 & 0 & 0 \\ 0 & 0 & 15 & 0 \\ 0 & 0 & 0 & -25 \end{bmatrix} \qquad (23)$$

Using cross validation, we get that

$$\hat{A} = \begin{bmatrix} 20.0000 & 0.0000 & 0.0001 & 0.0000 \\ -0.0000 & -10.0000 & 0.0000 & -0.0000 \\ 0.0000 & -0.0000 & 14.9998 & 0.0000 \\ -0.0000 & -0.0000 & -0.0000 & -25.0003 \end{bmatrix} \qquad (24)$$

for $\gamma_i = 0.9313 \times 10^{-9}, i = 1, \ldots, 4$.

For different realizations of a noise η_i of magnitude 0.5×10^{-4}, cross validation gives a good approximation of A and the eigenvalues of $A - \hat{A}$ are all within the unit disk with amplitude of the order of 10^{-3} showing that the dynamics of the error $e(k) = x(k) - \hat{x}(k)$ is asymptotically stable. For example, for a particular realization of η_i of magnitude 0.5×10^{-4}, we get

$$\hat{A} = \begin{bmatrix} 19.9635 & 0.0086 & 0.1365 & -0.0007 \\ -0.0177 & -10.0025 & 0.0379 & -0.0007 \\ -0.0177 & -0.0025 & 15.0376 & -0.0007 \\ -0.0132 & -0.0167 & 0.0065 & -25.0000 \end{bmatrix} \qquad (25)$$

with regularization parameters

$$\gamma_1 = 1.9073 \times 10^{-6}, \gamma_2 = 9.3132 \times 10^{-10}, \gamma_3 = 9.3132 \times 10^{-10},$$
$$\gamma_4 = 1.2207 \times 10^{-4}. \qquad (26)$$

The algorithm fails in the presence of quadratic structural perturbations. This is due to the choice of a linear kernel. A polynomial kernel, for example, would allow for nonlinear perturbations but this would require a complete reformulation of our algorithm. We leave the extension of our algorithm to the nonlinear case for future work.

The next example is an unstable system with a large gap between the eigenvalues.

Example 5 Consider the system $x(k + 1) = Ax(k)$ with

$$A = \begin{bmatrix} 20 & 0 \\ 0 & -0.1 \end{bmatrix}. \qquad (27)$$

With the initial condition $x(0) = [-1.9, 1]$, we generate the time series $x(1), \ldots, x(100)$. The algorithm above yields the (excellent) estimate

$$\hat{A} = \begin{bmatrix} 20.0000 & 0.0000 \\ -0.0000 & -0.1000 \end{bmatrix}, \qquad (28)$$

In the presence of noise of maximal amplitude 10^{-4}, the algorithm approximates well only the large entry $a_{11} = 20$: For a first realization of η_i and with cross validation, we get

$$\hat{A} = \begin{bmatrix} 19.9997 & -0.0111 \\ 0.0000 & -0.1104 \end{bmatrix}, \qquad (29)$$

with $\gamma_1 = 1.5259 \times 10^{-5}$ and $\gamma_2 = 2^{20}$. However another realization of η_i leads to

$$\hat{A} = \begin{bmatrix} 19.9994 & -0.0011 \\ 0.0000 & -0.0000 \end{bmatrix}, \qquad (30)$$

with $\gamma_1 = 3.0518 \times 10^{-5}$ and $\gamma_2 = 2.8147 \times 10^{14}$. This is due to the fact that the data converge to the eigenspace generated by the largest eigenvalue $\lambda = 20$. However, the eigenvalues of $A - \hat{A}$ are within the unit disk with small amplitude which guarantees that the error dynamics of $e(k) = x(k) - \hat{x}(k)$ converges to the origin quite quickly. We observe the same phenomenon with

$$A = \begin{bmatrix} -0.5 & 0 \\ 0 & 25 \end{bmatrix}. \qquad (31)$$

Here, in the absence of noise, we obtain the estimate

$$\hat{A} = \begin{bmatrix} -0.5000 & 0.0000 \\ -0.0000 & 25.0000 \end{bmatrix}, \qquad (32)$$

with $\gamma_1 = \gamma_2 = 0.9313 \times 10^{-9}$. In the presence of noise η_i with amplitude 10^{-4}, the data converge to the eigenspace corresponding to the largest eigenvalue $\lambda = 25$: for some realization of η_i one obtains the estimate

$$\hat{A} = \begin{bmatrix} -0.4809 & 0.0008 \\ 0.0164 & 24.9960 \end{bmatrix}, \tag{33}$$

while for another realization of η

$$\hat{A} = \begin{bmatrix} -0.0000 & -0.0000 \\ -1.0067 & 24.8696 \end{bmatrix}. \tag{34}$$

The regularization parameters γ_1 and γ_2 adapt to the realization of the noise.

As already remarked in the end of Sect. 2, we see that "more data" does not always necessarily lead to better results, since the data sequence converges to the eigenspace generated by the largest eigenvalue. However, whether with or without noise, the approximations of A are good enough to reduce the error between $x(k + 1) = Ax(k)$ and $\hat{x}(k + 1) = \hat{A}\hat{x}(k)$ outside of the training examples, since cross-validation determines a good regularization parameter γ that balances between good fitting and good prediction properties.

The next example has an eigenvalue on the unit circle.

Example 6 Consider $x(k + 1) = Ax(k)$ with

$$A = \begin{bmatrix} 2.2500 & -1.2500 & 1.2500 & -49.5500 \\ 3.7500 & -2.7500 & 13.1500 & -20.6500 \\ 0 & 0 & 10.4000 & -32.3000 \\ 0 & 0 & 0 & -21.9000 \end{bmatrix}. \tag{35}$$

The set of eigenvalues of A is spec $(A) = \{-1.5000, 1.0000, 10.4000, -21.9000\}$. In the absence of noise and initial condition $x = [-0.9, 15, 1.5.2.5]$ with $N = 100$ points, we compute the estimate

$$\hat{A} = \begin{bmatrix} 2.2500 & -1.2500 & 1.2498 & -49.5499 \\ 3.7500 & -2.7500 & 13.1498 & -20.6499 \\ 0.0000 & 0.0000 & 10.3998 & -32.2999 \\ 0.0000 & 0.0000 & -0.0001 & -21.8999 \end{bmatrix}, \tag{36}$$

and regularization parameters $\gamma_1 = \gamma_2 = 0.9313 \times 10^{-9}$. In this case, the set of eigenvalues of $\hat{A}$ is

$$\text{spec}\,(\hat{A}) = \{-21.9000, 10.3999, -1.5000, 1.0000\}. \tag{37}$$

For a given realization of $\eta \in [-10^{-4}, 10^{-4}]$, we obtain the estimate

$$\hat{A} = \begin{bmatrix} 2.2551 & -1.2490 & 1.2187 & -49.5304 \\ 3.7554 & -2.7489 & 13.1175 & -20.6297 \\ 0.0055 & 0.0011 & 10.3669 & -32.2794 \\ 0.0053 & 0.0010 & -0.0325 & -21.8797 \end{bmatrix} \tag{38}$$

with $\gamma_1 = 0.0745 \times 10^{-7}$ and $\gamma_2 = 0.1490 \times 10^{-7}$. The eigenvalues of $A - \hat{A}$ are of the order of 10^{-4} which guarantees that the error dynamics converges quickly to the origin. However, the set of eigenvalues of $\hat{A}$ is

$$\text{spec}\,(\hat{A}) = \{-21.8996, 10.3999, -1.5026, 1.0134\}. \tag{39}$$

Hence an additional unstable eigenvalue occurs.

Example 7 Consider $x(k + 1) = Ax(k)$ with

$$A = \begin{bmatrix} -0.8500 & 0.4500 & -0.4500 & -77.8500 \\ -1.3500 & 0.9500 & 14.3500 & -11.6500 \\ 0 & 0 & 15.3000 & -55.3000 \\ 0 & 0 & 0 & -40.0000 \end{bmatrix}. \tag{40}$$

The eigenvalues of A are given by

$$\text{spec}\,(A) = \{-0.4000, 0.5000, 15.3000, -40.0000\}. \tag{41}$$

For an initial condition $x = [-0.9;15;1.5;2.5]$ and with $N = 100$ data points, we get

$$\hat{A} = \begin{bmatrix} -0.8498 & 0.4501 & -0.4499 & -77.8504 \\ -1.3499 & 0.9500 & 14.3501 & -11.6502 \\ 0.0001 & 0.0001 & 15.3001 & -55.3004 \\ -0.0004 & -0.0002 & -0.0004 & -39.9987 \end{bmatrix} \tag{42}$$

with eigenvalues given by

$$\text{spec}\,(\hat{A}) = \{-40.0000, -0.3974, 0.4982, 15.3008\}. \tag{43}$$

Here we used $\gamma_i = 10^{-12}, i = 1, \dots, 4$. Moreover, the eigenvalues of $A - \hat{A}$ are quite small and such that the error dynamics converges quickly to the origin. In the presence of noise η, the algorithm approximates the largest eigenvalues of A but does not approximate the smaller (stable) ones. For example, for a particular realization of noise with amplitude 10^{-4}, we get the estimate

$$\hat{A} = \begin{bmatrix} -2.1100 & -0.0993 & -1.3259 & -74.4543 \\ -1.7053 & 0.7777 & 13.9397 & -10.5308 \\ -0.8277 & -0.3692 & 14.6466 & -52.9920 \\ -0.8283 & -0.3694 & -0.6539 & -37.6904 \end{bmatrix} \tag{44}$$

and spec $(\hat{A})$ = {− 40.0009, 0.1620 ± 0.8438i, 15.3008}.

For another realization of noise with amplitude 10^{-2}, we get the estimate

$$\hat{A} = \begin{bmatrix} -138.0893 & -60.7052 & -105.8111 & 301.5029 \\ -0.2435 & 0.9101 & 12.9638 & -12.6745 \\ -71.1408 & -31.9557 & -40.3842 & 142.3170 \\ -71.1408 & -31.9557 & -55.6843 & 157.6172 \end{bmatrix} \tag{45}$$

and spec $(\hat{A})$ = {− 40.1391, 3.9326, 0.9601, 15.3002}.

The algorithm introduced above also allows us to compute the topological entropy of linear systems, since it is determined by the unstable eigenvalues. Recall that the topological entropy of a linear map on $\mathbb{R}^n$ is defined in the following way:

Fix a compact subset $K \subset \mathbb{R}^n$, a time $\tau \in \mathbb{N}$ and a constant $\varepsilon > 0$. Then a set $R \subset \mathbb{R}^n$ is called (τ, ε)-spanning for K if for every $y \in K$ there is $x \in R$ with

$$\left\| A^j y - A^j x \right\| < \varepsilon \text{ for all } j = 0, \ldots, \tau. \tag{46}$$

By compactness of K, there are finite (τ, ε)-spanning sets. Let R be a (τ, ε)-spanning set of minimal cardinality $\#R = r_{min}(\tau, \varepsilon, K)$. Then

$$h_{top}(K, A, \varepsilon) := \lim_{\tau \to \infty} \frac{1}{\tau} \log r_{min}(\tau, \varepsilon, K), h_{top}(K, A) := \lim_{\varepsilon \to 0^+} h_{top}(K, \varepsilon). \tag{47}$$

(the limits exist). Finally, the topological entropy of A is

$$h_{top}(A) := \sup_K h_{top}(K, A), \tag{48}$$

where the supremum is taken over all compact subsets K of $\mathbb{R}^n$.

A classical result due to Bowen (cf. [21, Theorem 8.14]) shows that the topological entropy is determined by the sum of the unstable eigenvalues, i.e.,

$$h_{top}(A) = \sum \max(1, |\lambda|), \tag{49}$$

where summation is over all eigenvalues of A counted according to their algebraic multiplicity.

Hence, when we approximate the unstable eigenvalues of A by those of the matrix $\hat{A}$, we also get an approximation of the topological entropy.

Example 8 For Example 6, we get that $h_{top}(A) = 34.80$ while for the estimate $\hat{A}$ one obtains $h_{top}(\hat{A}) = 34.7999$. For Example 7, we get that $h_{top}(A) = 55.30$ and $h_{top}(\hat{A}) = 55.3008$. These estimates appear reasonably good.

4 Identification of linear control systems

Consider the linear control system

$$x(k + 1) = Ax(k) + Bu(k), \tag{50}$$

with $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times 1}$. We want to estimate the matrices A and B from the time series $x(1) + \eta_1$, $\ldots, x(N) + \eta_N$ where η satisfies the Assumption in Sect. 2. The initial condition $x(0)$ and the control sequence $u(0)$, $\ldots, u(N)$ are assumed to be known.

In order to estimate A and B, we will extend algorithm $\mathcal{A}$. The ith component of system (50) is given by

$$x_i(k + 1) = \sum_{j=1}^{n} a_{ij} x_j(k) + b_i u(k). \tag{51}$$

For every i we want to estimate the coefficients b_i and $a_{ij}, j = 1, \ldots, n$. Thus the linear map $f_i : \mathbb{R}^n \to \mathbb{R}$ given by

$$(x_1, \ldots, x_n, u) \mapsto \sum_{j=1}^{n} a_{ij} x_j + b_i u \tag{52}$$

is unknown. To extend algorithm $\mathcal{A}$, we will view system (51) as a system of the form (2) where the state x is the extended state $\underline{x} = (x, u) \in \mathbb{R}^n \times \mathbb{R}$ for (50). Hence, the kernel expansion (6) becomes

$$x_i(k + 1) = \sum_{j=1}^{N} c_{ij} \langle \underline{x}(j), \underline{x}(k) \rangle \tag{53}$$

where $\underline{x}_{n+1} = u$ and the c_{ij} satisfy the following set of equations:

$$\begin{bmatrix} x_i(1) \\ \vdots \\ x_i(N) \end{bmatrix} = \left(N\lambda I_d + \underline{\mathbb{K}} \right) \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iN} \end{bmatrix}, \tag{54}$$

with

$$\underline{\mathbb{K}} = \begin{bmatrix} \sum_{\ell=1}^{n+1} \underline{x}_\ell(1)\underline{x}_\ell(0) & \cdots & \sum_{\ell=1}^{n+1} \underline{x}_\ell(N)\underline{x}_\ell(0) \\ \vdots & \cdots & \vdots \\ \sum_{\ell=1}^{n+1} \underline{x}_\ell(1)\underline{x}_\ell(N-1) & \cdots & \sum_{\ell=1}^{n+1} \underline{x}_\ell(N)\underline{x}_\ell(N-1) \end{bmatrix}. \tag{55}$$

Let us emphasize that $u = x_{n+1}$ does not appear on the left hand side of (53)–(54).

In reference to the case when A has eigenvalues outside the unit circle, we adopt the same method as in Sect. 3 and define

$$\underline{\sigma} := \max \left\{ \frac{\|\underline{x}(k+1)\|}{\|\underline{x}(k)\|}, k \in \{0, 1, \ldots, N\} \right\}. \tag{56}$$

Example 9 (One dimensional case) Consider $x(k+1) = -0.9x(k) + 3.5u$. For an input $u(k) = \sin(k) + \cos(k)$ and for 100 points we obtain the estimate $\hat{A} = -0.9$ and $\hat{B} = 3.5$ when there is no noise η_i. Here cross validation gives $\gamma_1 = 1.5259 \times 10^{-05}$ and $\gamma_2 = 1$. For a certain realization of the noise η_i with amplitude 0.1, we get $\hat{A} = -0.9008$ and $\hat{B} = 3.4983$. Here cross validation gives $\gamma_1 = 0.0078$ and $\gamma_2 = 1$.

Example 10 (Three dimensional stable case) Consider control system (50) with

$$A = \begin{bmatrix} -0.9 & 1 & 0 \\ 0 & -0.1 & 1 \\ 0 & 0 & 0.8 \end{bmatrix} \text{ and } B = \begin{bmatrix} -2.5 \\ -3.5 \\ 4.5 \end{bmatrix}. \tag{57}$$

With the input $u(k) = \sin(k) + \cos(k)$ and 100 points, one computes the estimates

$$\hat{A} = \begin{bmatrix} -0.9000 & 1.0000 & 0.0000 \\ 0.0000 & -0.1000 & 1.0000 \\ -0.0000 & -0.0000 & 0.8000 \end{bmatrix} \text{ and } \hat{B} = \begin{bmatrix} -2.5000 \\ -3.5000 \\ 4.5000 \end{bmatrix}. \tag{58}$$

Here cross validation gives the regularization parameters $\gamma_i = 0.1526 \times 10^{-4}$ for $i = 1, \ldots, 4$. For some realization of perturbations η_i with amplitude 0.1, one computes the estimates

$$\hat{A} = \begin{bmatrix} -0.9047 & 0.9984 & -0.0029 \\ -0.0047 & -0.1016 & 0.9971 \\ -0.0048 & -0.0018 & 0.7971 \end{bmatrix} \text{ and } \hat{B} = \begin{bmatrix} -2.5326 \\ -3.5321 \\ 4.4661 \end{bmatrix}. \tag{59}$$

Here cross validation gives $\gamma_1 = 9.7656 \times 10^{-4}$, $\gamma_2 = 9.7656 \times 10^{-4}$, $\gamma_3 = 1.5259 \times 10^{-5}$, $\gamma_4 = 4$.

Example 11 (Three dimensional unstable case) Consider control system (50) with

$$A = \begin{bmatrix} -20 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 20 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}. \tag{60}$$

The input $u(k) = \sin(k) + \cos(k)$ and 100 points give the estimates

$$\hat{A} = \begin{bmatrix} -19.9945 & 1.0009 & -0.0137 \\ 0.0013 & 0.9995 & 0.9919 \\ 0.0155 & -0.0171 & 19.7835 \end{bmatrix} \text{ and } \hat{B} = \begin{bmatrix} 0.9898 \\ 1.9898 \\ 2.9333 \end{bmatrix}. \tag{61}$$

Here cross validation yields the regularization parameters $\gamma_i = 0.8882 \times 10^{-15}$ for $i = 1, \ldots, 4$. For some realization of perturbations η_i with amplitude 10^{-4}, one computes the estimates

$$\hat{A} = \begin{bmatrix} -20.0000 & 0.9334 & -0.0058 \\ -0.0008 & 0.9382 & 0.9939 \\ -0.0008 & -0.0590 & 19.9937 \end{bmatrix} \text{ and } \hat{B} = \begin{bmatrix} 0.9819 \\ 1.9814 \\ 2.9811 \end{bmatrix}. \tag{62}$$

Here cross validation gives $\gamma_1 = \gamma_2 = 0.2384 \times 10^{-6}$, $\gamma_3 = \gamma_4 = 0.0596 \times 10^{-6}$.

These results show that algorithm $\mathcal{A}$ works quite well in these cases.

5 Stabilization via linear-quadratic optimal control

A basic problem for linear control systems is stabilization by state feedback. A standard method is to use linear quadratic optimal control, where the feedback is computed using the solution of an algebraic Riccati equation. In this section, we propose to replace in the algebraic Riccati equation the system matrix A by the estimate $\hat{A}$ obtained by learning theory.

The linear quadratic optimal control problem has the following form:

Minimize over all (continuous) inputs u

$$J_\infty(x_0; u) = \sum_{k=0}^{\infty} \left[x(k)^\top Q x(k) + u(k)^\top R u(t) \right] \tag{63}$$

with $x(\cdot)$ given by

$$x(k+1) = Ax(k) + Bu(k), \; k \geq 0, \; x(0) = x_0; \tag{64}$$

here $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite and $R \in \mathbb{R}^{m \times m}$ is positive definite, and $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}$.

Consider the discrete algebraic Riccati equation DARE

$$A^\top \left(P - PB(R + B^\top PB)^{-1} B^\top P \right) A + Q = P. \tag{65}$$

Obviously, every solution P is positive semi-definite. We cite the following theorem from [1].

Theorem Suppose that for every $x_0 \in \mathbb{R}^n$ there is an input u, such that $J(x_0, u) < \infty$. Then the following holds:

(i) There is a unique solution P of the DARE.
(ii) For every $x_0 \in \mathbb{R}^n$ one has $J^*(x_0) := \inf\{J(x_0, u) | u \text{ an input}\} = x_0^\top P x_0$ and there is a unique optimal input u^* with $J^*(x_0) = J(x_0, u^*)$. This optimal input is generated by the feedback $F = (R + B^\top PB)^{-1} B^\top PA$ and

$$u(k) = -Fx(k), k \geq 0. \tag{66}$$

In particular, the feedback F stabilizes the system, i.e., $x(k+1) = (A - BF)x(k)$ is stable.

Now we use an estimate $\hat{A}$ and $\hat{B}$ (obtained by kernel methods) instead of A and B in the algebraic Riccati equation and obtain the solution $\hat{P}$. Will the corresponding feedback $u = \hat{F}x := -B^\top \hat{P}x$ also stabilize the system, i.e., is the following system stable:

$$x(k+1) = (A - BB^\top \hat{P})x(k)? \tag{67}$$

Example 12 Consider the one-dimensional system $x(k+1) = -0.9x(k) + 3.5u$ in Example 9. In the absence of noise, we get $\hat{A} = -0.9$ and $\hat{B} = 3.5$. We have that $A - B\hat{F} = \hat{A} - \hat{B}\hat{F} = -0.0643$. When there is noise of amplitude 0.1, we get that $\hat{A} = -0.9002$ and $\hat{B} = 3.4929$ and $A - B\hat{F} = -0.0643$ while $\hat{A} - \hat{B}\hat{F} = -0.0610$. Hence, the controller improves stability.

Example 13 Consider control system (50) with

$$A = \begin{bmatrix} -0.9 & 1 & 0 \\ 0 & -0.1 & 1 \\ 0 & 0 & 0.8 \end{bmatrix} \text{ and } B = \begin{bmatrix} -2.5 \\ -3.5 \\ 4.5 \end{bmatrix}. \tag{68}$$

As illustrated in Example 10, without noise we get excellent approximations of A and B. For both cases, the set of eigenvalues of the closed-loop system is $\{-0.6172, 0.4049, -0.0018\}$. With a noise of maximal amplitude 0.1, the estimates $\hat{A}$ and $\hat{B}$ are given in Example 10. For the feedback system one finds

$$\text{spec}\,(\hat{A} - \hat{B}\hat{F}) = \{-0.6204, 0.4053, -0.0018\},$$

$$\text{spec}\,(A - B\hat{F}) = \{-0.6240, -0.0062, 0.4111\}.$$

In this example the feedback based on the estimate also stabilizes the original system.

Example 14 Consider control system (50) with

$$A = \begin{bmatrix} -20 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 20 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}. \tag{69}$$

As Example 11 illustrates, without noise we get excellent approximations of A and B. For the feedback system one finds

$$\text{spec}\,(\hat{A} - \hat{B}\hat{F}) = \{0.1994, 0.0483, -0.0501\},$$

$$\text{spec}\,(A - B\hat{F}) = \{-0.1234 \pm 2.0777i, 0.5279\}.$$

When there is noise of amplitude 10^{-4}, one computes the estimates

$$\hat{A} = \begin{bmatrix} -19.9805 & 0.7484 & 0.0135 \\ -0.0062 & 0.7969 & 1.0107 \\ -0.0229 & 0.9851 & 19.6776 \end{bmatrix} \text{ and } \hat{B} = \begin{bmatrix} 1.0194 \\ 2.0114 \\ 2.6673 \end{bmatrix}. \tag{70}$$

This are bad approximations for A and B. Furthermore, for the feedback system one finds

$$\text{spec}\,(\hat{A} - \hat{B}\hat{F}) = \{0.1929, 0.0477, -0.0501\},$$

$$\text{spec}\,(A - B\hat{F}) = \{1.4510 \pm 3.0103i, -2.5232\}.$$

Thus the stabilizing controller for the approximate system does not stabilize the true system.

6 Conclusions

This paper has introduced the algorithm $\mathcal{A}$ based on kernel methods to identify a stable linear dynamical system from a time series. The numerical experiments give excellent results in the absence of noise and structural perturbations. In the presence of noise and structural perturbations the algorithm works well in the stable case. In the unstable case, a modified algorithm works quite well in the presence of noise but cannot handle structural perturbations.

Then we have extended algorithm $\mathcal{A}$ to identify linear control systems. In particular, we have used estimates obtained by kernel methods to stabilize linear systems using linear-quadratic control and the algebraic Riccati equation. Here the numerical experiments seem to indicate that the same conclusions on applicability of the algorithm apply.

Extensions of the considered algorithms to nonlinear systems appear feasible and are left to future work.

Compliance with ethical standards

Appendix: Elements of learning theory

In this section, we give a brief overview of Reproducing Kernel Hilbert Spaces (RKHS) as used in statistical learning theory. The discussion here borrows heavily from Cucker and Smale [7], Wahba [20], and Schölkopf and Smola [16]. Early work developing the theory of RKHS was undertaken by Schoenberg [13–15] and then Aronszajn [2]. Historically, RKHS came from the question, when it is possible to embed a metric space into a Hilbert space.

Definition 1 Let $\mathcal{H}$ be a Hilbert space of functions on a set $\mathcal{X}$ which is a closed subset of $\mathbb{R}^n$. Denote by $\langle f, g \rangle$ the inner product on $\mathcal{H}$ and let $\|f\| = \langle f, f \rangle^{1/2}$ be the norm in $\mathcal{H}$, for f and $g \in \mathcal{H}$. We say that $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there exists $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that

(i) K has the reproducing property, i.e., $f(x) = \langle f(\cdot), K(\cdot, x) \rangle$ for all $f \in \mathcal{H}$.
(ii) K spans $\mathcal{H}$, i.e., $\mathcal{H} = \overline{\text{span}} \{ K(x, \cdot) | x \in \mathcal{X} \}$.

K will be called a reproducing kernel of $\mathcal{H}$ and $\mathcal{H}_K$ will denote the RKHS $\mathcal{H}$ with reproducing kernel K.

Definition 2 Given a kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and inputs $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix

$$k := (K(x_i, x_j))_{ij}, \tag{71}$$

is called the *Gram Matrix* of k with respect to $x_1, \ldots, x_n$. If for all $n \in \mathbb{N}$ and distinct $x_i \in \mathcal{X}$ the kernel K gives rise to a strictly positive definite Gram matrix, it is called strictly positive definite.

Definition 3 (Mercer kernel map) A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a Mercer kernel if it is continuous, symmetric and positive definite.

The important properties of reproducing kernels are summarized in the following proposition.

Proposition 1 *If K is a reproducing kernel of a Hilbert space $\mathcal{H}$, then*

(i)   *$K(x, y)$ is unique.*
(ii) *For all $x, y \in \mathcal{X}, K(x, y) = K(y, x)$ (symmetry).*
(iii) *$\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$ for $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$ (positive definitness).*

(iv) $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$.
(v) *The following kernels, defined on a compact domain $\mathcal{X} \subset \mathbb{R}^n$, are Mercer kernels: $K(x, y) = x \cdot y^\top$ (Linear), $K(x, y) = (1 + x \cdot y^\top)^d$, $d \in \mathbb{N}$ (Polynomial), $K(x, y) = e^{-\frac{\|x - y\|^2}{\sigma^2}}$, $\sigma > 0$ (Gaussian).*

Theorem 1 *Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a symmetric and positive definite function. Then there exists a Hilbert space of functions $\mathcal{H}$ defined on $\mathcal{X}$ admitting K as a reproducing Kernel. Moreover, there exists a function $\Phi : X \rightarrow \mathcal{H}$ such that*

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \quad \text{for} \quad x, y \in \mathcal{X}. \tag{72}$$

Φ is called a feature map.

  Conversely, let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, with $\mathcal{X}$ compact, satisfying

For all $x \in \mathcal{X}$ there is $\kappa_x > 0$, such that $|f(x)| \leq \kappa_x \|f\|_{\mathcal{H}}$. (73)

Then $\mathcal{H}$ has a reproducing kernel K.

Remark 1

(i)   The dimension of the RKHS can be infinite and corresponds to the dimension of the eigenspace of the integral operator $L_K : \mathcal{L}^2_\nu(\mathcal{X}) \rightarrow \mathcal{C}(\mathcal{X})$ defined as $(L_K f)(x) = \int K(x, t) f(t) d\nu(t)$ if K is a Mercer kernel, for $f \in \mathcal{L}^2_\nu(\mathcal{X})$ and ν a Borel measure on $\mathcal{X}$.
(ii) In Theorem 1, and using property [iv.] in Proposition 1, we can take $\Phi(x) := K_x := K(x, \cdot)$ in which case $\mathcal{F} = \mathcal{H}$—the "feature space" is the RKHS. This is called the *canonical feature map*.
(iii) The fact that Mercer kernels are positive definite and symmetric shows that kernels can be viewed as generalized Gramians and covariance matrices.
(iv) In practice, we choose a Mercer kernel, such as the ones in [v.] in Proposition 1, and Theorem 1, that guarantees the existence of a Hilbert space admitting such a function as a reproducing kernel.

RKHS play an important role in learning theory whose objective is to find an unknown function

$$f^* : X \rightarrow Y \tag{74}$$

from random samples

$$\mathbf{s} = (x_i, y_i)|_{i=1}^m, \tag{75}$$

In the following we review results from [18] (for a more general setting, cf. [7]) in the special case when the data samples $\mathbf{s}$ are such that the following assumption holds.

Assumption 1 The samples in (75) have the special form

$$\mathcal{S} \neg \quad \mathbf{s} = (x, y_x)|_{x \in \bar{x}}, \tag{76}$$

where $\bar{x} = \{x_i\}|_{i=1}^{d+1}$ and y_x is drawn at random from $f^*(x) + \eta_x$, where η_x is drawn from a probability measure ρ_x.

Here for each $x \in X$, ρ_x is a probability measure with zero mean, and its variance σ_x^2 satisfies $\sigma^2 := \sum_{x \in \bar{x}} \sigma_x^2 < \infty$. Let X be a closed subset of $\mathbb{R}^n$ and $\bar{t} \subset X$ is a discrete subset. Now, consider a kernel $K : X \times X \to \mathbb{R}$ and define a matrix (possibly infinite) $K_{\bar{t},\bar{t}} : \ell^2(\bar{t}) \to \ell^2(\bar{t})$ as

$$(K_{\bar{t},\bar{t}}a)_s = \sum_{t \in \bar{t}} K(s,t)a_t, \quad s \in \bar{t}, a \in \ell^2(\bar{t}), \tag{77}$$

where $\ell^2(\bar{t})$ is the set of sequences $a = (a_t)_{t \in \bar{t}} : \bar{t} \to \mathbb{R}$ with $\langle a,b \rangle = \sum_{t \in \bar{t}} a_t b_t$ defining an inner product. For example, we can take $X = \mathbb{R}$ and $\bar{t} = \{0, 1, \ldots, d\}$.

In the case of a linear dynamical system (1), we are interested in learning the map $x(k) \mapsto x(k+1)$. Here we can apply the following results.

The problem to approximate a function $f^* \in \mathcal{H}_K$ from samples **s** of the form (75) has been studied in [18, 19]. It is reformulated as the minimization problem

$$\bar{f}_{\mathbf{s},\gamma} := \arg \min_{f \in \mathcal{H}_{K,\bar{t}}} \left\{ \sum_{x \in \bar{x}} (f(x) - y_x)^2 + \gamma \|f\|_K^2 \right\}, \tag{78}$$

where $\gamma \geq 0$ is a regularization parameter. Moreover, when $\bar{x}$ is not defined by a uniform grid on X, the authors of [18] introduced a weighting $w := \{w_x\}_{x \in \bar{x}}$ on $\bar{x}$ with $w_x > 0$.[1] Let D_w be the diagonal matrix with diagonal entries $\{w_x\}_{x \in \bar{x}}$. Then, $\|D_w\| \leq \|w\|_\infty$.

In this case, the regularization scheme (78) becomes

$$\bar{f}_{\mathbf{s},\gamma} := \arg \min_{f \in \mathcal{H}_{K,\bar{t}}} \left\{ \sum_{x \in \bar{x}} w_x (f(x) - y_x)^2 + \gamma \|f\|_K^2 \right\}, \tag{79}$$

Theorem 2 *Assume $f^* \in \mathcal{H}_{K,\bar{t}}$ and the standing hypotheses with X, K, $\bar{t}$, ρ as above, y as in (76). Suppose $K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}}$ is invertible. Define $\mathcal{L}$ to be the linear operator $\mathcal{L} = (K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}})^{-1} K_{\bar{t},\bar{x}} D_w$. Then problem (79) has the unique solution*

$$f_{\mathbf{s},\gamma} = \sum_{t \in \bar{t}} (\mathcal{L}y)_t K_t \tag{80}$$

Assumption 2 For each $x \in X$, ρ_x is a probability measure with zero mean supported on $[-M_x, M_x]$ with $\mathcal{B}_w := (\sum_{x \in \bar{x}} w_x M_x^2)^{\frac{1}{2}} < \infty$.

The next theorems give estimates for the different sources of errors.

Theorem 3 (Sample error) [18, Theorem 4, Propositions 2 and 3] *Let Assumptions 1 and 2 be satisfied, suppose that $K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}}$ is invertible and let $f_{\mathbf{s},\gamma} = \sum_{t \in \bar{t}} c_t K_t$ be the solution of (79) given in Theorem 2 by $c = \mathcal{L}y$. Define*

$$\mathcal{L}_w := (K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}})^{-1} K_{\bar{t},\bar{x}} D_w^{1/2}$$
$$\kappa := \|K_{\bar{t},\bar{t}}\| \, \|(K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}})^{-1}\|^2.$$

Then for every $0 < \delta < 1$, with probability at least $1 - \delta$ we have the sample error estimate

$$\|f_{\mathbf{s},\gamma} - f_{\bar{x},\gamma}\|_K^2 \leq \mathcal{E}_{\text{samp}} := \kappa \sigma_w^2 \alpha^{-1} \left(\frac{2\|K_{\bar{t},\bar{t}} \mathcal{L}_w\| \, \|\mathcal{L}_w\| \, \mathcal{B}_w^2}{\kappa \sigma_w^2} \log \frac{1}{\delta} \right), \tag{81}$$

where $\alpha(u) := (u-1)\log u$ for $u > 1$. In particular, $\mathcal{E}_{\text{samp}} \to 0$ when $\gamma \to \infty$ or $\sigma_w^2 \to 0$.

Theorem 4 (Regularization error and integration error) [18, Proposition 4 and Theorem 5] *Let Assumptions 1 and 2 be satisfied and let $\bar{X} = (X_x)_{x \in \bar{x}}$ be the Voronoi cell of X associated with $\bar{x}$ and $w_x = \rho_X(X_x)$. Define the Lipschitz norm on a subset $X' \subset X$ as $\|f\|_{\text{Lip}(X')} := \|f\|_{L^\infty(X')} + \sup_{s,u \in X} \frac{|f(s)-f(u)|}{\|s-u\|_{\ell^\infty(\mathbb{R}^n)}}$ and assume that the inclusion map of $\mathcal{H}_{K,\bar{t}}$ into the Lipschitz space satisfies[2]*

$$C_{\bar{x}} := \sup_{f \in \mathcal{H}_{K,\bar{t}}} \frac{\sum_{x \in \bar{x}} w_x \|f\|_{\text{Lip}(X_x)}^2}{\|f\|_K^2} < \infty. \tag{82}$$

Suppose that $\bar{x}$ is Δ-dense in X, i.e., for each $y \in X$ there is some $x \in \bar{x}$ satisfying $\|x - y\|_{\ell^\infty(\mathbb{R}^n)} \leq \Delta$. Then for $f^* \in \mathcal{H}_{K,\bar{t}}$

$$\|f_{\bar{x},\gamma} - f^*\|^2 \leq \|f^*\|_K^2 (\gamma + 8C_{\bar{x}}\Delta) \tag{83}$$

Theorem 5 (Sample, regularization and integration errors) [18, Corollary 5] *Under the assumptions of Theorems 3 and 4, let $\bar{X} = (X_x)_{x \in \bar{x}}$ be the Voronoi cell of X associated with $\bar{x}$ and $w_x = \rho_x(X_x)$. Suppose that $\bar{x}$ is Δ-dense, $C_{\bar{x}} < \infty$, and $f^* \in \mathcal{H}_{K,\bar{t}}$. Then, for every $0 < \delta < 1$, with probability at least $1 - \delta$ there holds*

$$\|f_{\mathbf{s},\gamma} - f^*\|^2 \leq 2C_{\bar{x}} \mathcal{E}_{\text{samp}} + 2\|f^*\|_K^2 (\gamma + 8C_{\bar{x}}\Delta), \tag{84}$$

where $\mathcal{E}_{\text{samp}}$ is given in (81).

---

[1] A suggestion in [18] is to consider the ρ_X-volume of the Voronoi cell associated with $\bar{x}$. Another example is $w = 1$ or if $|\bar{x}| = m < \infty$, $w = \frac{1}{m}$.

[2] This assumption is true if X is compact and the inclusion map of $\mathcal{H}_{K,\bar{t}}$ into the space of Lipschitz functions on X is bounded which is the case when K is a C^2 Mercer kernel [22]. In fact, if $\|f\|_{\text{Lip}(X)} \leq C_0 \|f\|_K$ for each $f \in \mathcal{H}_{K,\bar{t}}$, then $C_{\bar{x}} \leq C_0^2 \rho_X(X)$.

References

1. Antsaklis PJ, Michel AN (2006) Linear systems. Birkhäuser, Boston
2. Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404
3. Bouvrie J, Hamzi B (2017) Kernel methods for the approximation of nonlinear systems. SIAM J Control Optim 55–4:2460–2492
4. Bouvrie J, Hamzi B (2017) Kernel methods for the approximation of some key quantities of nonlinear systems. J Comput Dyn 4(1—-2):1–19
5. Cheney W, Light W (2009) A course in approximation theory, graduate studies in mathematics, vol 101. American Mathematical Society, Providence
6. Colonius F, Kliemann W (2014) Dynamical systems and linear algebra, graduate studies in mathematics, vol 158. American Mathematical Society, Providence
7. Cucker F, Smale S (2001) On the mathematical foundations of learning. Bull Am Math Soc 39:1–49
8. Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. Adv Comput Math 13(1):1–50
9. Hamzi B, Colonius F (2019)Kernel methods for discrete-time linear equations, springer lecture notes in computer science LNCS, vol 11539
10. Li L, Zhou D-X (2015) Learning theory approach to a system identification problem involving atomic norm. J Fourier Anal Appl 21:734
11. Pillonetto G, Dinuzzo F, Chen T, De Nicolao G, Ljung L (2014) Kernel methods in system identification, machine learning and function estimation: a survey. Automatica 50(3):657–682
12. Rifkin RM, Lippert A (2007) Notes on regularized least squares. Computer science and artificial intelligence laboratory technical report, MIT, MIT-CSAIL-TR-2007-025, CBCL-268
13. Schoenberg IJ (1935) Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". Ann Math 36:724–732
14. Schoenberg IJ (1937) On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space. Ann Math 38:787–793
15. Schoenberg IJ (1938) Metric spaces and positive definite functions. Trans Am Math Soc 44:522–536
16. Schölkopf B, Smola AJ (2002) Learning with kernels. MIT Press, Cambridge
17. Smale S, Zhou D-X (2003) Estimating the approximation error in learning theory. Anal Appl 1(1):17
18. Smale S, Zhou D-X (2004) Shannon sampling and function reconstruction from point values. Bull Am Math Soc 41:279–305
19. Smale S, Zhou D-X (2005) Shannon sampling II: connections to learning theory. Appl Comput Harmon Anal 19(3):285–302
20. Wahba G (1990) Spline models for observational data. In: SIAM CBMS-NSF regional conference series in applied mathematics, vol 59
21. Walters P (1982) An introduction to ergodic theory. Springer, Berlin
22. Zhou D-X (2003) Capacity of reproducing kernel spaces in learning theory. IEEE Trans Inf Theory 49(7):1743–1752