Kernel Methods for Discrete-Time Linear Equations

Boumediene Hamzi[1,2](✉) and Fritz Colonius[3]

[1] Department of Mathematics, Imperial College London, London, UK
boumediene.hamzi@gmail.com
[2] Department of Mathematics, AlFaisal University,
Riyadh, Kingdom of Saudi Arabia
[3] Institut für Mathematik, Universität Augsburg, Augsburg, Germany

Abstract. Methods from learning theory are used in the state space of linear dynamical systems in order to estimate the system matrices and some relevant quantities such as a the topological entropy.

The approach is illustrated via a series of numerical examples.

Keywords: Reproducing Kernel Hilbert spaces · Linear discrete-time equations · Parameter estimation

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 [8,21]. We refer to [8] for a general survey on applications of machine learning to system identification.

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 topological entropy for a linear map. Furthermore, in the control case, 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

BH thanks the European Commission and the Scientific and the Technological Research Council of Turkey (Tubitak) for financial support received through Marie Curie Fellowships.

within the spirit of [3,4] where the authors showed that RKHSs act as “linearizing spaces” and 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 $\phi : \mathbb{R} \to \mathbb{R}^3$ defined as $\phi(x) = [1\, x\, x^2]^T$ 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 is 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). Section4 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.

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 [12]).

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 $a_{ij}, j = 1, \cdots, n$. They are determined by the linear maps $f_i^* : \mathbb{R}^n \to \mathbb{R}$ given by

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

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 (36) and $(x(k), x_i(k+1) + \eta_i)$ are the samples in (38).

We note that in [12], 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 (41). 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 $(c_{i,1}, \cdots, c_{i,N})$ 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 $x_i(k+1)$, $i = 1, \cdots, n$,

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

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

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

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

$$\begin{aligned} x_i(k+1) &= \sum_{j=1}^{N} c_{ij}\langle x(j), x(k)\rangle = \sum_{j=1}^{N} c_{ij}x(j)^T \cdot x(k) = \sum_{j=1}^{N}\sum_{\ell=1}^{n} c_{ij}x_\ell(j)x_\ell(k) \\ &= \sum_{\ell=1}^{n}\sum_{j=1}^{N} c_{ij}x_\ell(j)x_\ell(k). \end{aligned}$$

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 $x(1), \cdots, x(N)$ 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 $x(1), \cdots, x(N)$ of the data series and the values $\hat{x}(1), \cdots, \hat{x}(N)$ 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 [10]. 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 (36) is the linear map $f^*(x) = f_i(x)$ in (3) and the samples $\mathbf{s}$ in (38) 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, (46) 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 $||x_i(k+1)||_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{c_{i,j}^2}{\lambda_j}$.

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 $x(1), \cdots, x(100)$. 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 $\gamma = 1.5259 \cdot 10^{-5}$. 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 $\sqrt{\sum_{i=100}^{300} e^2(i)} = 1.8218 \cdot 10^{-30}$.

We observe an analogous behavior of the algorithm when the data are generated from $x(k+1) = \alpha x(k) + \varepsilon x(k)^2$ where the algorithm works well in the presence of noise and structural perturbations when using cross validation. When $\varepsilon = 0.1$ 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 $\sqrt{\sum_{i=100}^{300} e^2(i)} = 2.8098 \cdot 10^{-30}$ which shows that the error e decreases exponentially and the generalization properties of the algorithm are quite good.

3 Unstable Case

Consider

$$x(k+1) = Ax(k) \text{ with } A \in \mathbb{R}^{n \times n}, \tag{11}$$

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

$$x(1), x(2), ..., x(N), \tag{12}$$

which are generated by system (11), 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 (12) associated to an auxiliary stable system.

For a constant $\sigma > 0$ we define the auxiliary system by $y(k+1) = \tilde{A}y(k)$ (13) with $\tilde{A} := \frac{1}{\sigma}A$. Thus $y(k) = \left(\frac{A}{\sigma}\right)^{k-1} y(1)$ and with $y(1) = x(1)$ one finds $y(k) = \frac{1}{\sigma^{k-1}} A^{k-1} x(1) = \frac{1}{\sigma^{k-1}} x(k)$. 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 (12) for system (11) yield the following data for system (3): $y(1) := x(1), y(2) := \frac{1}{\sigma}x(2), ..., y(N) := \frac{1}{\sigma^{N-1}}x(N)$. We propose to choose σ as follows: Define

$$\sigma := \max\left\{\frac{\|x(k+1)\|}{\|x(k)\|}, k \in \{0, 1, ..., N\}\right\}. \tag{13}$$

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 2. Consider $x(k+1) = \alpha x(k)$ with $\alpha = 11.46$. With the initial condition $x(0) = -0.5$, we generate the time series $x(1), \cdots, x(100)$. The algorithm above with the regularization parameter $\gamma = 10^{-6}$ yields the estimate $\hat{\alpha} = 11.4086$. Cross validation leads to the regularization parameter $\gamma = 9.5367 \cdot 10^{-7}$ 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.

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

Example 3. Consider the system $x(k+1) = Ax(k)$ with $A = \begin{bmatrix} 20 & 0 \\ 0 & -0.1 \end{bmatrix}$. With the initial condition $x(0) = [-1.9, 1]$, we generate the time series $x(1), \cdots, x(100)$. The algorithm above yields the (excellent) estimate $\hat{A} = \begin{bmatrix} 20.0000 & 0.0000 \\ -0.0000 & -0.1000 \end{bmatrix}$. 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 $\hat{A} = \begin{bmatrix} 19.9997 & -0.0111 \\ 0.0000 & -0.1104 \end{bmatrix}$, with $\gamma_1 = 1.5259 \cdot 10^{-5}$ and $\gamma_2 = 2^{20}$. However another realization of η_i leads to $\hat{A} = \begin{bmatrix} 19.9994 & -0.0011 \\ 0.0000 & -0.0000 \end{bmatrix}$, with $\gamma_1 = 3.0518 \cdot 10^{-5}$ and $\gamma_2 = 2.8147 \cdot 10^{14}$. 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

$$A = \begin{bmatrix} -0.5 & 0 \\ 0 & 25 \end{bmatrix}. \tag{14}$$

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

$$\hat{A} = \begin{bmatrix} -0.5000 & 0.0000 \\ -0.0000 & 25.0000 \end{bmatrix}, \tag{15}$$

with $\gamma_1 = \gamma_2 = 0.9313 \cdot 10^{-9}$. 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

$$\hat{A} = \begin{bmatrix} -0.4809 & 0.0008 \\ 0.0164 & 24.9960 \end{bmatrix}, \tag{16}$$

while for another realization of η

$$\hat{A} = \begin{bmatrix} -0.0000 & -0.0000 \\ -1.0067 & 24.8696 \end{bmatrix}. \tag{17}$$

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

$$A = \begin{bmatrix} 2.2500 & -1.2500 & 1.2500 & -49.5500 \\ 3.7500 & -2.7500 & 13.1500 & -20.6500 \\ 0 & 0 & 10.4000 & -32.3000 \\ 0 & 0 & 0 & -21.9000 \end{bmatrix}. \tag{18}$$

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

$$\hat{A} = \begin{bmatrix} 2.2500 & -1.2500 & 1.2498 & -49.5499 \\ 3.7500 & -2.7500 & 13.1498 & -20.6499 \\ 0.0000 & 0.0000 & 10.3998 & -32.2999 \\ 0.0000 & 0.0000 & -0.0001 & -21.8999 \end{bmatrix}, \tag{19}$$

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

$$\mathrm{spec}(\hat{A}) = \{-21.9000, 10.3999, -1.5000, 1.0000\}. \tag{20}$$

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

$$\hat{A} = \begin{bmatrix} 2.2551 & -1.2490 & 1.2187 & -49.5304 \\ 3.7554 & -2.7489 & 13.1175 & -20.6297 \\ 0.0055 & 0.0011 & 10.3669 & -32.2794 \\ 0.0053 & 0.0010 & -0.0325 & -21.8797 \end{bmatrix} \tag{21}$$

with $\gamma_1 = 0.0745 \cdot 10^{-7}$ and $\gamma_2 = 0.1490 \cdot 10^{-7}$. 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

$$\mathrm{spec}(\hat{A}) = \{-21.8996, 10.3999, -1.5026, 1.0134\}. \tag{22}$$

Hence an additional unstable eigenvalue occurs.

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

$$A = \begin{bmatrix} -0.8500 & 0.4500 & -0.4500 & -77.8500 \\ -1.3500 & 0.9500 & 14.3500 & -11.6500 \\ 0 & 0 & 15.3000 & -55.3000 \\ 0 & 0 & 0 & -40.0000 \end{bmatrix}. \tag{23}$$

The eigenvalues of A are given by

$$\mathrm{spec}(A) = \{-0.4000, 0.5000, 15.3000, -40.0000\}. \tag{24}$$

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

$$\hat{A} = \begin{bmatrix} -0.8498 & 0.4501 & -0.4499 & -77.8504 \\ -1.3499 & 0.9500 & 14.3501 & -11.6502 \\ 0.0001 & 0.0001 & 15.3001 & -55.3004 \\ -0.0004 & -0.0002 & -0.0004 & -39.9987 \end{bmatrix} \tag{25}$$

with eigenvalues given by

$$\mathrm{spec}(\hat{A}) = \{-40.0000, -0.3974, 0.4982, 15.3008\}. \tag{26}$$

Here we used $\gamma_i = 10^{-12}$, $i = 1, \cdots, 4$. 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

$$\hat{A} = \begin{bmatrix} -2.1100 & -0.0993 & -1.3259 & -74.4543 \\ -1.7053 & 0.7777 & 13.9397 & -10.5308 \\ -0.8277 & -0.3692 & 14.6466 & -52.9920 \\ -0.8283 & -0.3694 & -0.6539 & -37.6904 \end{bmatrix} \tag{27}$$

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

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

$$\hat{A} = \begin{bmatrix} -138.0893 & -60.7052 & -105.8111 & 301.5029 \\ -0.2435 & 0.9101 & 12.9638 & -12.6745 \\ -71.1408 & -31.9557 & -40.3842 & 142.3170 \\ -71.1408 & -31.9557 & -55.6843 & 157.6172 \end{bmatrix} \tag{28}$$

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

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

$$\left\| A^j y - A^j x \right\| < \varepsilon \text{ for all } j = 0, ..., \tau. \tag{29}$$

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

$$h_{top}(K, A, \varepsilon) := \lim_{\tau \to \infty} \frac{1}{\tau} \log r_{\min}(\tau, \varepsilon, K), h_{top}(K, A) := \lim_{\varepsilon \to 0^+} h_{top}(K, \varepsilon). \tag{30}$$

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

$$h_{top}(A) := \sup_K h_{top}(K, A), \tag{31}$$

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

A classical result due to Bowen (cf. [19, Theorem 8.14]) shows that the topological entropy is determined by the sum of the unstable eigenvalues, i.e.,

$$h_{top}(A) = \sum \max(1, |\lambda|), \tag{32}$$

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 6. For Example 4, we get that $h_{top}(A) = 34.80$ while for the estimate $\hat{A}$ one obtains $h_{top}(\hat{A}) = 34.7999$. For Example 5, we get that $h_{top}(A) = 55.30$ and $h_{top}(\hat{A}) = 55.3008$. These estimates appear reasonably good.

4 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.

A 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 [18], and Schölkopf and Smola [17]. Early work developing the theory of RKHS was undertaken by Schoenberg [14–16] 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 $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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., $\mathcal{H} = \overline{span\{K(x, \cdot)|x \in \mathcal{X}\}}$.

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 $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and inputs $x_1, \cdots, x_n \in \mathcal{X}$, the $n \times n$ matrix*

$$k := (K(x_i, x_j))_{ij}, \tag{33}$$

is called the Gram Matrix *of k with respect to $x_1, \cdots, x_n$. 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 $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 definiteness).*
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),* $K(x,y) = (1 + x \cdot y^\top)^d, \quad d \in \mathbb{N}$ *(Polynomial),* $K(x,y) = e^{-\frac{||x-y||^2}{\sigma^2}}, \quad \sigma > 0$ *(Gaussian).*

Theorem 1. *Let* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *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* $\Phi : X \to \mathcal{H}$ *such that*

$$K(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}} \quad \textit{for} \quad x, y \in \mathcal{X}. \tag{34}$$

Φ *is called a feature map.*

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

$$\textit{For all } x \in \mathcal{X} \textit{ there is } \kappa_x > 0, \textit{ such that } |f(x)| \leq \kappa_x ||f||_{\mathcal{H}}. \tag{35}$$

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 $L_K : \mathcal{L}^2_\nu(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ 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

$$f^* : X \to Y \tag{36}$$

from random samples

$$\mathbf{s} = (x_i, y_i)|_{i=1}^{m}, \tag{37}$$

In the following we review results from [12] (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 (37) have the special form

$$\mathcal{S}: \quad \mathbf{s} = (x, y_x)|_{x \in \bar{x}}, \tag{38}$$

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

$$(K_{\bar{t},\bar{t}} a)_s = \sum_{t \in \bar{t}} K(s,t) a_t, \quad s \in \bar{t}, a \in \ell^2(\bar{t}), \tag{39}$$

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 $\bar{t} = \{0, 1, \cdots, d\}$.

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 $\mathbf{s}$ of the form (37) has been studied in [12,13]. It is reformulated as the minimization problem

$$\bar{f}_{\mathbf{s},\gamma} := \text{argmin}_{f \in \mathcal{H}_{K,\bar{t}}} \left\{ \sum_{x \in \bar{x}} (f(x) - y_x)^2 + \gamma ||f||_K^2 \right\}, \tag{40}$$

where $\gamma \geq 0$ is a regularization parameter. Moreover,when $\bar{x}$ is not defined by a uniform grid on X, the authors of [12] 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 (40) becomes

$$\bar{f}_{\mathbf{s},\gamma} := \text{argmin}_{f \in \mathcal{H}_{K,\bar{t}}} \left\{ \sum_{x \in \bar{x}} w_x (f(x) - y_x)^2 + \gamma ||f||_K^2 \right\}, \tag{41}$$

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 (38). 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 (41) has the unique solution*

$$f_{\mathbf{s},\gamma} = \sum_{t \in \bar{t}} (\mathcal{L} y)_t K_t \tag{42}$$

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.

[1] A suggestion in [12] is to consider the ρ_X-volume of the Voronoi cell associated with $\bar{x}$. Another example is $w = 1$ or if $|\bar{x}| = m < \infty$, $w = \frac{1}{m}$.

Theorem 3. *(Sample Error) [12, 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 (41) 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

$$||f_{\mathbf{s},\gamma} - f_{\bar{x},\gamma}||_K^2 \leq \mathcal{E}_{samp} := \kappa \sigma_w^2 \alpha^{-1} \left(\frac{2||K_{\bar{t},\bar{t}} \mathcal{L}_w||\ ||\mathcal{L}_w||\ \mathcal{B}_w^2}{\kappa \sigma_w^2} \log \frac{1}{\delta} \right), \tag{43}$$

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

Theorem 4. *(Regularization Error and Integration Error) [12, 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 $||f||_{Lip(X')} := ||f||_{L^\infty(X')} + \sup_{s,u \in X} \frac{|f(s)-f(u)|}{||s-u||_{\ell^\infty(\mathbb{R}^n)}}$ and assume that the inclusion map of $\mathcal{H}_{K,\bar{t}}$ into the Lipschitz space satisfies*[2]

$$C_{\bar{x}} := \sup_{f \in \mathcal{H}_{K,\bar{t}}} \frac{\sum_{x \in \bar{x}} w_x ||f||_{Lip(X_x)}^2}{||f||_K^2} < \infty. \tag{44}$$

Suppose that $\bar{x}$ is $\Delta-$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 $f^* \in \mathcal{H}_{K,\bar{t}}$

$$||f_{\bar{x},\gamma} - f^*||^2 \leq ||f^*||_K^2 (\gamma + 8 C_{\bar{x}} \Delta) \tag{45}$$

Theorem 5. *(Sample, Regularization and Integration Errors) [12, 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 $\Delta-$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*

$$||f_{\mathbf{s},\gamma} - f^*||^2 \leq 2 C_{\bar{x}} \mathcal{E}_{samp} + 2||f^*||_K^2 (\gamma + 8 C_{\bar{x}} \Delta), \tag{46}$$

where $\mathcal{E}_{samp}$ is given in (43).

[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 [20]. In fact, if $||f||_{\mathrm{Lip}(X)} \leq C_0 ||f||_K$ for each $f \in \mathcal{H}_{K,\bar{t}}$, then $C_{\bar{x}} \leq C_0^2 \rho_X(X)$.

References

1. Antsaklis, P.J., Michel, A.N.: Linear Systems. Birkhäuser, Boston (2006)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
3. Bouvrie, J., Hamzi, B.: Kernel methods for the approximation of nonlinear systems. SIAM J. Control Optim. **55-4**, 2460–2492 (2017)
4. Bouvrie, J., Hamzi, B.: Kernel methods for the approximation of some key quantities of nonlinear systems. J. Comput. Dyn. **4**(1&2), 1–19 (2017)
5. Cheney, W., Light, W.: A Course in Approximation Theory. Graduate Studies in Mathematics, vol. 101. American Mathematical Society, Providence (2009)
6. Colonius, F., Kliemann, W.: Dynamical Systems and Linear Algebra. Graduate Studies in Mathematics, vol. 158. American Mathematical Society, Providence (2014)
7. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. **39**, 1–49 (2001)
8. Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., Ljung, L.: Kernel methods in system identification, machine learning and function estimation: a survey. Automatica **50**(3), 657–682 (2014)
9. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**(1), 1–50 (2000)
10. Rifkin, R.M., Lippert, A.: Notes on regularized least squares. Computer Science and Artificial Intelligence Laboratory Technical repor, MIT, MIT-CSAIL-TR-2007-025, CBCL-268 (2007)
11. Smale, S., Zhou, D.-X.: Estimating the approximation error in learning theory. Anal. Appl. **1**(1), 17–41 (2003)
12. Smale, S., Zhou, D.-X.: Shannon sampling and function reconstruction from point values. Bull. Am. Math. Soc. **41**, 279–305 (2004)
13. Smale, S., Zhou, D.-X.: Shannon sampling II: connections to learning theory. Appl. Comput. Harmonic Anal. **19**(3), 285–302 (2005)
14. Schoenberg, I.J.: 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 (1935)
15. Schoenberg, I.J.: On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in Hilbert space. Ann. Math. **38**, 787–793 (1937)
16. Schoenberg, I.J.: Metric spaces and positive definite functions. Trans. Am. Math. Soc. **44**, 522–536 (1938)
17. Schölkopf, B., Smola, A.J.: Learning with Kernels. The MIT Press, Cambridge (2002)
18. Wahba, G.: Spline Models for Observational Data. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59 (1990)
19. Walters, P.: An Introduction to Ergodic Theory. Springer, New York (1982)
20. Zhou, D.-X.: Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inf. Theory **49**(7), 1743–1752 (2003)
21. Li, L., Zhou, D.-X.: Learning theory approach to a system identification problem involving atomic norm. J. Fourier Anal. Appl. **21**, 734–753 (2015)