6-30-2022

Learning Nonparametric Ordinary Differential Equations: Application to Sparse and Noisy Data

Kamel Lahouel
TGen

Michael Wells
Portland State University

David Lovitz
Portland State University

Victor Rielly
Portland State University

Ethan Lew
Galois, Inc

See next page for additional authors

Citation Details

Authors

Kamel Lahouel, Michael Wells, David Lovitz, Victor Rielly, Ethan Lew, and Bruno Jedynak

LEARNING NONPARAMETRIC ORDINARY DIFFERENTIAL EQUATIONS: APPLICATION TO SPARSE AND NOISY DATA [*]

K. LAHOUEL, M. WELLS, D. LOVITZ, V. RIELLY, E. LEW, AND B. JEDYNAK

Abstract. Learning nonparametric systems of Ordinary Differential Equations (ODEs) $\dot{x} = f(t, x)$ from noisy and sparse data is an emerging machine learning topic. We use the well-developed theory of Reproducing Kernel Hilbert Spaces (RKHS) to define candidates for f for which the solution of the ODE exists and is unique. Learning f consists of solving a constrained optimization problem in an RKHS. We propose a penalty method that iteratively uses the Representer theorem and Euler approximations to provide a numerical solution. We prove a generalization bound for the L^2 distance between x and its estimator. Experiments are provided for the FitzHugh–Nagumo oscillator and for the prediction of the Amyloid level in the cortex of aging subjects. In both cases, we show competitive results when compared with the state of the art.

Key words. learning ODE, RKHS, penalty method, Amyloid accumulation, FHN oscillator.

AMS subject classifications. 62G05, 65L70, 68U99

1. Introduction.

1.1. Description of the problem and related works. Fitting a system of nonparametric ordinary differential equations (ODEs) $\dot{x} = f(t, x)$ to longitudinal data could lead to scientific breakthroughs in disciplines where ODEs or dynamical systems have been used for a long time, including physics, chemistry, and biology, see [12]. By nonparametric, we mean that there is no need to specify the functional form of the vector-field f using a finite dimensional parameter. Instead, this force field can belong to a functional space. This is a great advantage in situations where the from of the vector field is unknown but data is available for learning.

A particular difficulty arises when the data is sparse and noisy. This is often the case for longitudinal healthcare data, obtained during hospital visits. These visits provide measurements that are sparse in time, with a high level of individual variability. The work presented in this paper has been motivated in part by the need to model the accumulation of the Amyloid protein in the brain of aging subjects. It is a challenging but important task due to the significance of the Amyloid in the current understanding of Alzheimer's disease.

Fitting data to nonparametric ODEs is an inverse problem. It requires making assumptions on the initial state of the solution, and on the vector field. Furthermore, one needs to make assumptions on the noise model and provide a tractable optimization algorithm.

We provide now a short bibliographic survey. Further references are provided in the cited papers. First, observe that if the time derivative ($\dot{x}$) was observed, then fitting ODEs to noisy data would reduce to solving a regression problem. This remark has led to the methods known as "gradient matching" and to the earliest success in fitting ODEs to data, see e.g. [8, 3]. It consists in estimating the gradient from the data, then performing nonparametric regression to fit f and eventually, iterating, see [20]. These methods become inefficient when the data is sparse or noisy, or both.

Another approach consists in modeling f with polynomials [14] or with a Deep Neural Network, see [22, 4]. These methods integrate the solution along the vector

---

field from guessed initial conditions and compare the resulting trajectories with the observations. Optimization is used iteratively to refine the estimation of f and the initial conditions. Stochastic gradient descent and backpropagation is used in the latter case. Another modeling approach is to assume that f belongs to a Reproducing Kernel Hilbert Space (RKHS). This idea could be traced back to [18]. It was successfully applied to fluid mechanics in [24] and [3]. This is the conceptual approach pursued here. We believe that this approach is well motivated since there is a tight connection between the regularity (smoothness) properties of a kernel and the regularity properties of f. Specifically, one can choose an RKHS of vector-valued functions for which one is guaranteed the existence and uniqueness of the corresponding initial value problem. This is a necessary step in proving that more data would result in more accurate predictions. In [11], the authors use a Gaussian processes (GP) for the vector field. This is the Bayesian counterpart of the frequentist RKHS modeling, see [15] for a review of the similarities and differences between RKHSs and GPs. Comparisons between a collection of algorithms representative of the state of the art and the proposed algorithm is provided in the experiment section.

For the purpose of providing a visual and easy to understand illustration of the results generated by the algorithms presented in this paper, please see Figure 1. While the details of this experiment are provided in the section 4.6, we see that the proposed algorithm is able to recover a noisy trajectory and to extrapolate the data, contrary to a method that would use a regression model and ignore the ODE.

1.2. Main contributions. The main contributions of this paper are as follows:
1. We present an RKHS model for fitting nonparametric ODEs to observational data. Conditions for existence and uniqueness of the solutions of the corresponding initial value problem are expressed in terms of the regularity of the kernel;
2. We propose a novel algorithm for estimating nonparametric ODEs and the initial condition(s) from noisy data. This algorithm solves a constrained optimization problem using a penalty method;
3. We derive and prove a consistency result for the prediction of the state (interpolation) at unobserved times. This is, up to our knowledge, the first result for the problem of fitting nonparametric ODEs to data.
4. We provide experiments with simulated data. We compare the proposed algorithm to four existing methods representing state of the art for various sparsity and noise levels. We show that our algorithm is competitive.
5. We provide an experiment modeling the accumulation of Amyloid in the cortex of aging subjects. The data is sparse with, on average, three data points per trajectory (subject) and 179 trajectories. We show competitive performance compared to state of the art.

The rest of this paper is organized as follows: Section 2 presents the model and the algorithms. The consistency results are presented in Section 3 and proved in Appendix A. The experiments appear in Section 4 while Section 5 provides concluding remarks. Appendix B provides examples of kernels.

2. Model and algorithm.

2.1. Notations. We assume that we have n observations $y = (y_1, \ldots, y_n)$ where n, for example, is the number of subjects. Each observation y_i consists of m_i d-dimensional data points: $y_i = (y_{i1}(t_{i1}), \ldots, y_{im_i}(t_{im_i}))$ so that $y_{ij} \in \mathbb{R}^d$ is observed at time $t_{ij} \in [0, T]$. Our goal is to make predictions at new time-points for a subject with
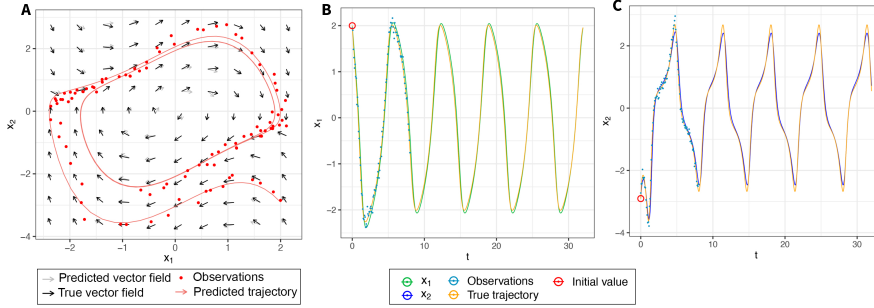
FIG. 1. *The true (grey) and estimated (black) vector fields for the FitzHugh-Nagumo (FHN) oscillator as well as noisy observations and three predicted trajectories.*

one or several known observations. To this end we explore the following nonparametric ODE model:

$$(2.1) \qquad \begin{cases} \dot{x} &= f(t,x) \\ y_{ij}(t_{ij}) &= x(t_{ij}) + \epsilon_{ij} \end{cases}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. The noise ϵ_{ij} is bounded or sub-Gaussian. This model is nonparametric because f is not specified parametrically. We assume that f belongs to a Reproducing Kernel Hilbert Space (RKHS) of smooth functions for which the solution x of the ODE exists and is unique, see Section 2.2. Background material on RKHS can be found in [13] and vector valued RKHS are reviewed in [1]. The rest of the paper is written for the autonomous case, that is when $f(t,x) = f(x)$, to simplify the notation. However, all the statements and algorithms generalize to the non-autonomous case.

2.2. Existence and uniqueness. The following is a classical result, see [25].

THEOREM 2.1. *Consider the initial value problem (IVP):*

$$(2.2) \qquad \dot{x}(t) = f(x(t)) \ and \ x(0) = x_0,$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$. If f is Lipschitz continuous, then the IVP has a unique solution that is defined on the domain $[0, +\infty)$.

Let H be an RKHS of vector-valued functions $\mathbb{R}^d \mapsto \mathbb{R}^d$. Let K be the reproducing kernel of H. K is a (d,d) matrix-valued kernel. It is natural to ask: what is a sufficient condition on K which ensures that all $f \in H$ are Lipschitz continuous? The following corollary provides an answer.

COROLLARY 2.2. *If $f : \mathbb{R}^d \to \mathbb{R}^d$ belongs to an RKHS with kernel K such that:*

$$(2.3) \quad d^2_{K_{ii}}(u,v) := K_{ii}(u,u) - 2K_{ii}(u,v) + K_{ii}(v,v)$$
$$\leq N_K^2 |u-v|^2, \forall u,v \in \mathbb{R}^d, i = 1 \ldots d,$$

for some constant N_K, then the IVP problem (2.2) has a unique solution defined on $[0, +\infty)$.

Proof. Notice that for every $i = 1...d$:

$$|f_i(u) - f_i(v)|^2 = |\langle K(u, \cdot)e_i - K(v, \cdot)e_i, f\rangle_H|^2 \tag{2.4}$$

$$\leq ||K(u, \cdot)e_i - K(v, \cdot)e_i||^2_H ||f||^2_H \tag{2.5}$$

$$= d^2_{K_{ii}}(u, v)||f||^2_H \qquad \square \tag{2.6}$$

where $e = (e_1, \ldots, e_d)$ is the natural basis of $\mathbb{R}^d$. Here we have used the reproducing property of the matrix-valued kernel and the Cauchy-Schwartz inequality. We conclude using theorem 2.1.

Thus, it is straightforward to choose a kernel which guarantees the existence and uniqueness of the solution of the IVP, which will lead to provable asymptotic performances. We believe that this simple result is a good motivator for the proposed modeling approach.

We now provide some examples of kernels satisfying Corollary 2.2. The simplest matrix-valued kernels are separable kernels. They are obtained by choosing a scalar kernel K_1 and a positive semi-definite matrix A. Then,

$$K(x, y) = K_1(x, y)A \tag{2.7}$$

The diagonal elements of K are then multiples of K_1. Thus, if K_1 verifies the regularity condition of Corollary 2.2, then so do all the separable kernels based on K_1. The scalar kernels satisfying the hypothesis of Corollary 2.2 are the linear kernel, Gaussian Kernel, rational quadratic kernel, and the Matérn kernel with $\nu \geq 3/2$. Kernels for which the functions in their corresponding RKHSs are not guaranteed to provide unique solutions to the corresponding IVP due to lack of regularity include polynomial kernels with order at least 2, the Laplacian kernel, and the Matérn Kernel with $\nu \leq 1/2$. Details are provided in Appendix B.

2.3. From constrained to unconstrained optimization. We first construct the optimization algorithm in the case $n = 1$. All the observations are from a single trajectory with the same initial condition. Thus, we temporarily drop the double indexing with subjects and times to simplify the notations.

Assume the observation times are $t_1 < \ldots < t_m$. Consider the following constrained minimization problem:

$$\min_{x,f} \frac{1}{m} \sum_{j=1}^{m} |y_j - x(t_j)|^2 + \lambda ||f - f_0||^2_H, \tag{2.8}$$

under the constraints

$$\begin{cases} f \in H, \text{ the RKHS with matrix-valued kernel } K, \\ x(t) = x(t_1) + \int_{t_1}^{t} f(x(s))ds, \text{ for } t_1 \leq t \leq t_m. \end{cases} \tag{2.9}$$

The function $f_0 \in H$ is an initial guess for f. Section 2.5 describes a gradient matching algorithm for selecting f_0. K is a kernel that satisfies Corollary 2.2.

Consider a regular one-dimensional grid over the interval $[t_1, t_m]$. Specifically, we choose

$$s_l = t_1 + lh \tag{2.10}$$

with $l = 0, \ldots, k$ and we assume that h is small enough so that there are integers $k_1 = 0 < k_2 < \ldots < k_m$, such that the observation times are

$$t_j = t_1 + k_j h, j = 1 \ldots m. \tag{2.11}$$

In practice, the observation times are rounded to fit on this grid. Note that with this notation, $t_j = s_{k_j}$. We now proceed through a series of transformations to rewrite this constrained optimization problem into an unconstrained one.

First, we replace the constraints on x by a finite number of constraints as follows:

(2.12)
$$\begin{cases} f \in H, \text{ the RKHS with kernel } K, \\ x(s_{l+1}) = x(s_l) + \int_{s_l}^{s_{l+1}} f(x(s))ds \\ l = 0 \ldots k - 1. \end{cases}$$

Second, we discretize the constraints using the Euler method of integration:

(2.13)
$$\begin{cases} f \in H, \text{ the RKHS with kernel } K, \\ x(s_{l+1}) = x(s_l) + hf(x(s_l)) \\ \quad \text{for } l = 0 \ldots k - 1. \end{cases}$$

Third, we replace the constrained optimization problem by an unconstrained one using a single Lagrange constant $\gamma > 0$. Notate $z_l = x(s_l)$, $l = 0 \ldots k$,

(2.14)
$$\min_{z \in \mathbb{R}^{d(k+1)}, f \in H} J(z, f, \gamma),$$

with

(2.15) $\quad J(z, f, \gamma) = \dfrac{1}{m} \sum_{j=1}^{m} |y_j - z_{k_j}|^2 + \gamma \dfrac{1}{k} \sum_{l=0}^{k-1} |z_{l+1} - z_l - hf(z_l)|^2 + \lambda ||f - f_0||_H^2.$

2.4. Penalty method. The penalty method is an iterative method which consists of enforcing the constraints by increasing a penalty parameter, in this case γ. The schematic of the method is presented in Algorithm 2.1. At each step, the functional $J(z, f, \gamma)$ in (2.15) is minimized with respect to (z, f), for a fixed value of γ. Then, γ is increased. The optimization for (z, f) is done asynchronously, first optimizing over z for a fixed f, then optimizing over f for the newly updated z.

Let us now describe these optimization steps in more detail. For a fixed γ and f, $J(z, f, \gamma)$ in (2.15) is non-convex in z due to the presence of $f(z_l)$. Therefore we replace f by its first-order Taylor expansion evaluated at the value $z_l^{(s)}$ obtained in the previous iteration s:

(2.16)
$$f(z_l) \approx f(z_l^{(s)}) + (z_l - z_l^{(s)})^T \nabla_{z_l} f(z_l^{(s)})$$

Note that with this approximation, J is convex, quadratic, and sparse in z. This allows the use of a linear solver for this minimization. The number of unknowns is $d(k+1)$.

For a fixed γ and z, minimizing J in f is equivalent to a ridge regression problem. After the change of variable, $g = f - f_0$, and setting

(2.17)
$$u_l = (z_{l+1} - z_l)/h - f_0(z_l), l = 0 \ldots k - 1,$$

we use the representer theorem to show that the minimizer in $f \in H$ of J is of the form

(2.18)
$$f(z) = f_0(z) + \sum_{l=0}^{k} K(z, z_l) w_l,$$

Algorithm 2.1 Penalty method for ODE-RKHS

1: Init: $h, \rho, \lambda, f^{(0)}, \gamma^{(0)}, s = 0$
2: **while** termination condition is not met **do**
3:      $z^{(s+1)} \leftarrow \arg\min_{z \in \mathbb{R}^{d(k+1)}} J(z, f^{(s)}, \gamma^{(s)})$
4:      $f^{(s+1)} \leftarrow \arg\min_{f \in H} J(z^{(s+1)}, f, \gamma^{(s)})$
5:      $\gamma^{(s+1)} \leftarrow \gamma^{(s)}(1 + \rho)$
6:      $s = s + 1$
7:      Check termination condition
8: **end while**

---

where $w_l \in \mathbb{R}^d$. Let $W = (w_1^T, \ldots, w_{k+1}^T)$, be of dimension $(d(k+1), 1)$ and similarly let $U = (u_1^T, \ldots, u_{k+1}^T)$ and K be the matrix with (d, d) block element $K_{kl} = K(x_k, x_l)$. We find that W is a minimizer of the convex quadratic function

$$(2.19) \qquad \frac{\gamma h^2}{k} |U - KW|^2 + \lambda W^T K W$$

and thus W is the solution to the linear system:

$$(2.20) \qquad \left(K + \frac{\lambda k}{\gamma h^2} I \right) W = U$$

2.5. Initial condition and termination criteria. The initial condition f_0 can be chosen using the gradient matching method:
   1. Approximate the time derivatives of x at the observed times $\dot{x}(t_j)$, denoted $\hat{\dot{x}}(t_j)$
   2. Estimating $f_0 \in H$ using ridge regression, i.e. minimizing over H

$$(2.21) \qquad G(f_0) = \frac{1}{m} \sum_{j=1}^{m} |\hat{\dot{x}}(t_j) - f_0(y_j)|^2 + \lambda ||f_0||_H^2$$

There are many possibilities for the approximation in the first step depending on the sparsity of the data and the amount of noise. In the experiments below, we use central differences.

The termination condition of Algorithm 2.1 includes a fixed number of iterations S and a threshold on the quantity $||f^{(s+1)} - f^{(s)}||/||f^{(s)}||$ which allows for early stopping.

2.6. Multiple trajectories. We present here the extension of the method to multiple trajectories, say $n > 1$ subjects. We assume the same number of observations for each subject and regular sampling to simplify the presentation.

First, we replace (2.15) and (2.12) with

$$(2.22) \qquad \min_{x,f} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |y_{ij} - x_i(t_{ij})|^2 + \lambda ||f - f_0||_H^2,$$

under the constraints

$$(2.23) \qquad \begin{cases} f \in H, \text{ the RKHS with matrix-valued kernel K,} \\ x_i(t) = x_i(t_1) + \int_{t_1}^{t} f(x_i(s)) ds, \\ \quad \text{for } t_1 \leq t \leq t_m, i = 1 \ldots n \end{cases}$$

Algorithm 2.2 Multi Trajectories Penalty method for ODE-RKHS

1: Init: $h, \rho, \lambda, f^{(0)}, \gamma^{(0)}, s = 0$
2: **while** termination condition is not met **do**
3:     **for** $i = 1 \ldots n$ **do**
4:         $z_i^{(s+1)} \leftarrow \arg\min_{z_i \in \mathbb{R}^{d(k+1)}} J_{\text{multi}}(z, f^{(s)}, \gamma^{(s)})$
5:     **end for**
6:     $f^{(s+1)} \leftarrow \arg\min_{f \in H} J_{\text{multi}}(z^{(s+1)}, f, \gamma^{(s)})$
7:     $\gamma^{(s+1)} \leftarrow \gamma^{(s)}(1 + \rho)$
8:     $s = s + 1$
9:     Check termination condition
10: **end while**

We then proceed along the same steps as for the single trajectory case, leading to the unconstrained optimization problem, generalizing (2.14) and (2.15).

Notate $z_{il} = x_i(s_l)$, $l = 0 \ldots k$, $i = 1 \ldots n$, and $z = (z_1, \ldots, z_n)$

$$(2.24) \qquad \min_{z \in \mathbb{R}^{nd(k+1)}, f \in H} J_{\text{multi}}(z, f, \gamma),$$

with

$$(2.25) \quad J_{\text{multi}}(z, f, \gamma) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |y_{ij} - z_{ik_j}|^2$$

$$+ \gamma \frac{1}{nk} \sum_{i=1}^{n} \sum_{l=0}^{k-1} |z_{i,l+1} - z_{il} - h f(z_{il})|^2 + \lambda ||f - f_0||_H^2.$$

The key point is that J_{multi} decouples the trajectories such that the optimization over z can be carried out separately for each trajectory. However, all the observations contribute to the estimation of f. The algorithm is presented in Alg 2.2. In Line 6: we use the no-trick formulation using Gaussian quadrature Fourier features as described in [5].

2.7. Computational Complexity. We analyze the complexity of the algorithm Alg 2.2. The key parameters are:
1. d: the dimension of the observed vectors;
2. n: the number of observed trajectories;
3. k: the number of samples in the discretization of the time interval;
4. S: the number of steps in Alg 2.2;
5. n_F: the number of Fourier features.

We use $O(p^3)$ for the time complexity of solving a (dense) linear system with p variables and $O(w^2 p)$ in the case of a band matrix of width w, see [16]. Alg 2.2, line 4 consists in solving a linear system of size dk with a band matrix of bandwidth $w = 3d$, thus $O(kd^3)$ computations. Line 6 consists in solving d full linear systems of dimension n_F, thus $O(dn_F^3)$ computations. In total, we find $O(Snkd^3 + Sdn_F^3)$. Note that k is typically chosen proportional to the average number of data points per trajectory. Thus, overall, the algorithm is linear in the number of observations but cubic in the dimension of the observations.

2.8. Non autonomous systems, covariates, and irregular sampling. Non autonomous systems and covariates are handled by modifying the kernel. The issue

of irregular sampling is addressed by replacing the first term of (2.15) by

$$(2.26) \qquad \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(t_{i,j+1} - t_{ij})|y_{ij} - z_{ik_j}|^2$$

with $t_{i,m_i+1} = T,\ i = 1\ldots,n$

3. Consistency of the solution: A finite sample result. In this section, we assume that our algorithm solves the following optimization problem (where $t_{m+1} = T$ by definition):

$$(3.1) \qquad \min_{\mathbb{R}^{d(k+1)}, f\in H}\sum_{j=1}^{m}(t_{j+1} - t_j)|y_j - z_{k_j}|^2,$$

Under the constraints:
1. $||f - f_0||_H \le R,\ |z_0| \le r$
2. $z_{l+1} = z_l + hf(z_l),\ 0 \le l \le k$

Notice that constraint 2 corresponds to the Euler method for the ODE: $\dot{x} = f(x)$. Therefore, by linearly interpolating between the times of subdivision $s_l,\ 0 \le l \le k$, we can generate a solution $\hat{x}(.)$ defined on $[0, T]$. We denote by $x^*(.)$ the true trajectory generating the noisy observations y_j at each time t_j. The purpose of this section is to present a result controlling (in probability) the L^2 norm squared of $\hat{x} - x^*$:

$$(3.2) \qquad ||\hat{x} - x^*||_{L^2}^2 := \int_0^T |(\hat{x}(t) - x^*(t))|^2 dt$$

Let us make the following assumptions:
- $\mathbf{A_1}$: There exist an $f^* \in H, ||f^* - f_0||_H \le R$ and $|x_0^*| \le r$ such that $x^*(0) = x_0^*$ and $\dot{x}^*(t) = f^*(x^*(t))$ for every $0 \le t \le T$.
- $\mathbf{A_2}$: The noise variables ϵ_{ij} are independent and bounded in absolute value by a constant M_ϵ. (We can assume that the variables are subgaussian instead of bounded if we want to generalize this result)
- $\mathbf{A_3}$: The kernel K is $\mathcal{C}^2(\mathbb{R}^d)$ in it's first argument (this implies that it is also $\mathcal{C}^2(\mathbb{R}^d)$ in its second argument).
- $\mathbf{A_4}$: The kernel K satisfies (2.3).

We refer to section 2.2 for examples of kernels satisfying $\mathbf{A_3}$ and $\mathbf{A_4}$.

These assumptions are sufficient for obtaining the main theorem of this section, controlling $||\hat{x} - x^*||_{L^2}^2$ with high probability.

THEOREM 3.1. *Assuming* $\mathbf{A_1}, \mathbf{A_2}, \mathbf{A_3}$ *and* $\mathbf{A_4}$, *there exist positive constants* K_1, K_2, K_3 *and* K_4, *depending only on* R, r, T, M_ϵ, N_K *and the kernel* K *such that for every* $\epsilon > 0$, *with probability less than* $\exp\left(\frac{-K_2\epsilon^2}{d\sum_{j=1}^{m}(t_{j+1}-t_j)^2}\right)$:

$$(3.3) \qquad ||\hat{x} - x^*||_{L_2}^2 \ge K_1 d\sqrt{\sum_{j=1}^{m}(t_{j+1} - t_j)^2} + h^2 K_3 d + K_4 d\sum_{j=1}^{m}(t_{j+1} - t_j)^2 + \epsilon.$$

For a better understanding of Theorem 3.1, suppose a regular sampling of the interval $[0, T]$ with m points, so that for every j, $t_{j+1} - t_j = \frac{1}{m}$. In that case, under the same hypothesis, for any $\epsilon > 0$, with probability less than $\exp\left(\frac{-K_2 m\epsilon^2}{d}\right)$:

(3.4)
$$||\hat{x} - x^*||^2_{L_2} \geq \frac{K_1 d}{\sqrt{m}} + \frac{K_4 d}{m} + h^2 K_3 d + \epsilon.$$

A proof of Theorem 3.1 is provided in the appendix. We provide here a description of the main ideas. The third term in the right hand side of inequality (3.3) corresponds to the global truncation error between the numerical solution of the ODE and the true solution. The second term corresponds to the error between $||\hat{x} - x^*||^2_{L^2}$ and $\frac{1}{m} \sum_{j=1}^m |x^*(t_j) - \hat{x}(t_j)|^2$. The first term is the leading term, assuming that h is always less than $\frac{1}{m}$. Assume that $\hat{x}$ solves the continuous-constraints optimization problem (without an Euler approximation), i.e:

(3.5)
$$\min_{x,f} \frac{1}{m} \sum_{j=1}^m |y_j - x(t_j)|^2,$$

Under the constraints: $||f - f_0||_H \leq R$, $|x_0| \leq r$ and $x(t) = x_0 + \int_0^t f(x(u))du$, $\forall 0 \leq t \leq T$, we can then consider the "generalization" error:

(3.6)
$$\frac{1}{m} \sum_{j=1}^m |x^*(t_j) - \hat{x}(t_j)|^2.$$

An upper bound of this error is given by the first term. The main tool used to obtain the upper bound is Dudley's chaining inequality, see [27]. We notice that for every $i = 1, \ldots, d$, the set of coordinate functions x_i, where x and f satisfy the constraints of the continuous problem, is included in a set of functions that are uniformly Lipschitz continuous and bounded (the Lipschitz constant and bound does not depend on x_0 and f). Upper bounds of covering numbers of such functions are well-known, see [27], hence the use of Dudley's inequality.

4. Experiments. We report experiments for simulated data as well as for real data. In each case, we compare the performances of the proposed algorithm, generically named ODE-RKHS, with four other algorithms. These algorithms constitute the current state-of-the-art for learning nonparametric ODEs from noisy data. We briefly review these algorithms and provide references below.

4.1. npODE:. Nonparametric Ordinary Differential Equations (npODE) is presented in [11]. The authors use a Bayesian model with Gaussian processes (GP). Arguably, it is the Bayesian counterpart of the frequentist model presented in this paper. Unlike GP regression where the optimization can be computed in closed form, an approximate optimization method is required. The authors use inducing points, see [23] and sensitivity equations, see [17]. The npODE code was downloaded from http://www.github.com/cagatayyildiz/npode in February 2021.

4.2. SINDy:. Sparse Identification of Nonlinear Dynamical systems (SINDy) is a popular technique for identifying nonlinear dynamics from data, see[3]. SINDy predicts governing dynamics equations using gradient matching via sparse regression. In the experiments shown, we test SINDy with two different libraries of possible functions. We test SINDy with polynomials up to order three and a collection of Fourier features. We choose the SR3 sparsity regularization for its superior performance, detailed in [31], which has a threshold value as a hyperparameter. Other hyperparameters in our tests include the polynomial library's degree as

well as the size and lengthscale of the Fourier features library. A grid search tuner was employed to determine the best hyperparameter values, with the same hold-out and evaluation sets as in the competing algorithms. pySINDy v1.6.3 was used for the implementation [6]. The hyperparameter tuning code was downloaded from https://github.com/EthanJamesLew/AutoKoopman in March 2022.

4.3. Koopman Operator Approximation. The Koopman operator is an infinite dimensional linear operator that captures the dynamics of a non-linear dynamical system. Dynamic Mode Decomposition (DMD), described in [24], can approximate the Koopman operator's eigenvalues and eigenvectors based on observations of the system state. DMD can be generalized to extended-DMD (EDMD) by adding a dictionary of observable functions that map the state space to a new space that spans a finite subspace on which the operator can be approximated, see [28]. This dictionary must be selected prior to using EDMD, and can be chosen ad hoc or using library learning methods [29]. For these experiments, we use random Fourier features as the observable functions, specified here [7]. We use grid search optimization to find the best Gaussian kernel lengthscale and DMD rank.

4.4. Gradient descent via optimal control. We implemented a gradient descent algorithm based on the co-state equations derived from optimal control theory, see [21]. Specifically, we compute the gradient of the likelihood function under the constraints provided by the Euler discretization of the ODE. This is an optimization under equality constraints which is performed using a co-state as explained in [30]. The algorithm effectively implements a backpropagation algorithm in a deep neural network with parameters shared among all layers, see also [4].

The FitzHugh-Nagumo (FHN) oscillator data is presented in section 4.6. It is a controlled experiment with known and easy to visualize 2D trajectories. It has helped calibrate the algorithm. It was also demonstrated in [11] for the npODE algorithm.

The Amyloid data is presented in section 4.7. This dataset has motivated the creation of the ODE-RKHS algorithm. It is characterized by a smooth vector field, a large number of trajectories, and few, sparse and noisy observations per trajectory.

4.5. Selection of the hyper-parameters for the ODE-RKHS algorithm. We first select the parameter h, the time discretization. A smaller h provides better accuracy at the cost of a linear increase in computational time. Next, we select the parameter $\gamma^{(0)}$ small enough such that the data term in (2.25) would be the dominant term. Finally, we performed a grid search for the parameters λ and ρ, using a validation set consisting of 20% of the available data in each case.

4.6. Oscillator data. We ran experiments using a simulated dataset generated by the FHN oscillator:

$$\begin{aligned} \dot{v} &= v - v^3/3 - w + 1 \\ \dot{w} &= 0.08(v + 0.7 - 0.8w) \end{aligned}$$

(4.1)

An example of a result with the ODE-RKHS algorithm is presented in figure 1. Intermediate and final results of the running of the ODE-RKHS algorithm are presented in the Fig. 2 for the FHN data. Notice that during the first steps, shown in the top line, the estimated trajectories, shown with solid color lines are rough but tight to the data. During the later steps, shown on the bottom line, the trajectories are smoother, but still fitting the data.

We generated 25 training sets, each with a different level of noise and sparsity in the trajectories. There were 50 trajectories in each training set. We created training
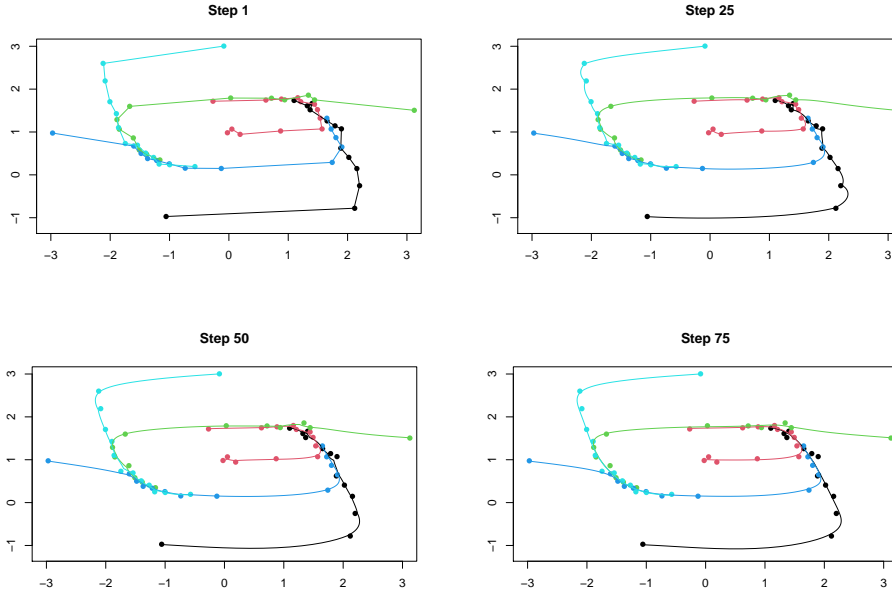
FIG. 2. *Illustration of the ODE-RKHS Algorithm: The observations are shown with dots. The estimated trajectories are shown with lines and curves with corresponding colors. Steps $i=1,25,50,$ and 75 are shown from left to right and from top to bottom*

sets for each combination of noise levels $\epsilon = 0.05, 0.1, 0.15, 0.2, 0.25$ and sparsity levels $N = 100, 75, 50, 25, 12$, where sparsity refers to the number of points on each trajectory in the training set. Next, we generated a single test set consisting of 100 trajectories without noise and 200 observations per trajectory, one for each increment of .1 in time. Testing consisted of computing predicted trajectories starting at the initial condition of the test trajectories and computing the following error measurement

$$(4.2) \qquad \text{Err} = \sqrt{\sum_{i=2}^{n} (t_i - t_{i-1}) \|y_i - \hat{y}_i\|^2}$$

where t_i refers to the ith observation time, y_i to the ith observation of the test trajectory, $\hat{y}_i$ to the ith point of the predicted trajectory and n is the number of observations in the trajectory. We ran experiments with the same training, validation and test sets for all the algorithms. The results are summarized in figure 3. The dots show the mean value of Err defined in (4.2) over all 100 trajectories in the test set. The vertical error bars show the standard error ($= \frac{\sigma}{\sqrt{100}}$), where σ is the standard deviation of Err. Overall the performances decrease with increased sparsity and increased level of noise as expected. We noticed that in most cases, ODE-RKHS or npODE are the best performing algorithms. The FHN is a polynomial system. This might explain why SINDy polynomial (in red) performs better than SINDy Fourier (in green). Koopman is the worst performing. This is likely due to the noise in the data, even at the lowest level of noise considered here. The gradient descent algorithm is constantly in the middle range of performances.
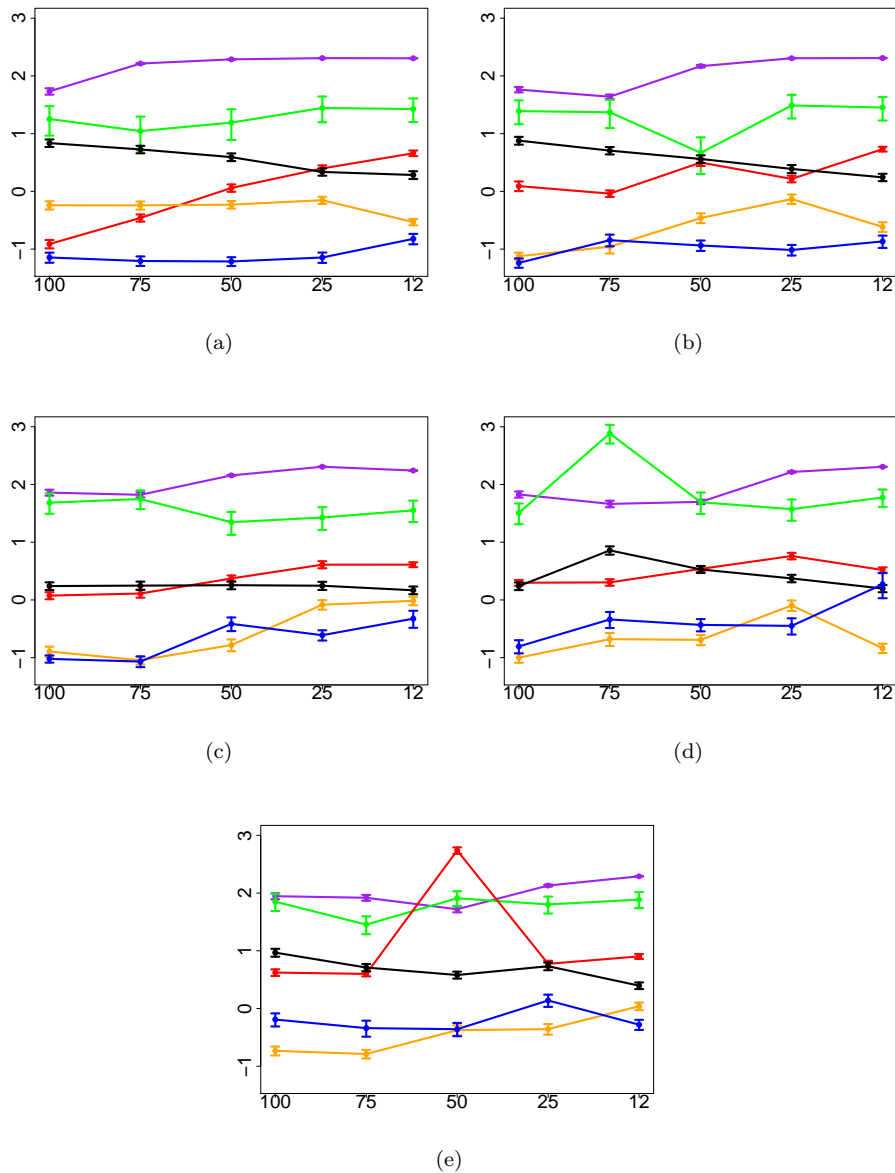
Fig. 3. *Results for the FHN experiment. The y-axis is the log of the mean error. The x-axis is the number of points per trajectory in the training set. Purple: Koopman. Green: SINDy with Fourier features. Red: SINDy with polynomials. Black: Gradient descent. Orange: npODE. Blue: ODE-RKHS. (a) Noise .05 (b) Noise .1 (c) Noise .15 (d) Noise .2 (e) Noise .25.*

4.7. The accumulation of Amyloid in the cortex of aging subjects. The accumulation of Amyloid in the brain is believed to be one of the earliest pathological mechanisms of Alzheimer's disease, beginning more than a decade prior to the onset of clinical symptoms, see [19].

Based on observations from several longitudinal Amyloid positron emission tomography (PET) studies, it is believed that the rate of Amyloid accumulation is

closely associated with the level of Amyloid at the same age, see [26]. We develop a principled mathematical model capturing this phenomenon and use it to predict longitudinally the accumulation of Amyloid across individuals. Note that we know that there are potentially devastating uses of machine learning in healthcare, see [10]. This work is done in collaboration with physicians who are in close contact with patients and ethics committees. These physicians helped us select relevant medical questions.

We used (PiB) PET scans from the Wisconsin Registry for Alzheimer's Prevention (WRAP) to assess global Amyloid burden, measured by the Distribution Volume Ratio (DVR)[1]. The number of subjects in this study is $n = 179$, with 3.06 visits on average, over an average span of 6.84 years. We fit the model in (2.1) to the posterior cingulum, precuneus and gyrus rectus DVRs, averaging the left and right DVR in each case. These regions are known to show Amyloid accumulation early in the disease process. We use the Multi Trajectories Penalty method for ODE-RKHS described in Alg. 2.2 with $d = 3$, and a Gaussian kernel. For each coordinate, we chose a bandwidth equal to 20% of the range of the data. The time step used was $h = .1$ years. We set $\gamma = 1$ and fit λ, ρ using a validation set consisting of 20 percent of the training data. We set a maximum of $S = 500$ iterations and used the early stopping criterion of stopping when the ratio $||f^{(s+1)} - f^{(s)}||/||f^{(s)}||$ was less than $\epsilon = 10^{-3}$. Figure 4 provide a visualization of the trajectories estimated using RKHS-ODE super-imposed (same color) with the data. This shows that the estimated trajectories are qualitatively accurate. We set aside 25 percent (rounded) of the data for testing. Prediction was performed using Euler integration starting at the first observed time-point for this subject. We computed the error for every subject as in the FHN experiment.

We compared with the predictions obtained with the other algorithms in table 1. We found that ODE-RKHS, SINDy polynomial and Fourier, and Gradient descent performed comparably for this data, while Koopman and npODE where not as good. npODE was performing very well for the FHN data and it is the worst performing here. A possible explanation is the dimension of the problem, here 3 instead of 2 for the FHN. It might be that some fine tuning of the npODE algorithm, for example increasing the number of inducing points would increase the performances.

Overall, considering both dataset, the ODE-RKHS performs consistently among the best algorithms.

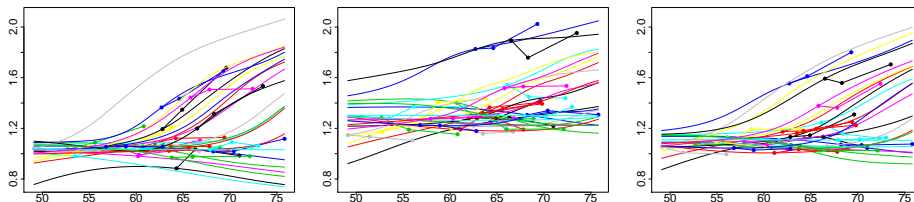FIG. 4. *Amyloid prediction experiment. Horizontal axis is in years. Vertical axis corresponds to DVR. The left-most image corresponds to the gyrus rectus, the middle to the cingulum and the right to the precuneus.*

---

[1] The data used for this experiment has been obtained from the Wisconsin Registry for Alzheimer's Prevention. See https://wrap.wisc.edu/. A request for accessing this data can be initiated from this website.

Alg	Error
ODE-RKHS	**0.38****
npODE	0.87
SINDy polynomial	**0.34****
SINDy Fourier	**0.37****
Gradient descent	**0.39****
Koopman	0.53*

TABLE 1

Results for Amyloid data. Stars indicate the number of methods the result is significantly better than as measured by the Wilcoxon signed-rank test at $\alpha = 5\%$.

5. Discussion. We proposed an algorithm for learning non-parametric ODEs assuming that the function f generating the vector field in $\mathbb{R}^d$ belongs to a vector-valued RKHS with a kernel satisfying certain regularity conditions. The data input of the algorithm consists of noisy observations at different times of multiple trajectories. The algorithm is linear in the number of observations but cubic in their dimension. We proved the consistency of the estimated trajectory, showing that the L^2 squared distance between the estimated trajectory and the true one vanishes as more observations are collected. We assessed the algorithm with simulated and real data and obtained results that compare consistently favorably with the state of the art on a wide range of sparsity and noise levels. Note however that much more theoretical work and experimental verification would be needed in order to truly understand in which situations one algorithm would be preferable over another.

Appendix A. Consistency of the estimator of the trajectory.

A.1. Assuming we solve the problem without Euler approximation. This section gives the proof of theorem presented in section 3 of the main text. We present the proof for $d = 1$ since the generalization to multiple dimensions is straightforward. We also present the proof for the case of autonomous systems. Keeping the notations of the main text, we make the following assumptions:

- **A₁**: There exist an $f^* \in H, ||f^* - f_0||_H \leq R$ and $|x_0^*| \leq r$ such that $x^*(0) = x_0^*$ and $\dot{x}^*(t) = f^*(x^*(t))$ for every $0 \leq t \leq T$.
- **A₂**: The noise variables ϵ_j are independent and bounded by a constant M_ϵ, with a variance denoted by σ^2. (We can assume that the variables are sub-gaussian instead of bounded if we want to generalize this result)
- **A₃**: The kernel K is $\mathcal{C}^2(\mathbb{R})$ in it's first argument (this implies that it is also $\mathcal{C}^2(\mathbb{R})$ in its second argument).
- **A₄**: The kernel K satisfies the hypothesis of Corollary 1.

Without loss of generality, we will assume that $f_0 = 0$ in our proof.

Let H be the RKHS with reproducing kernel K. Let $f \in H$ such that $||f||_H \leq R$. We know using assumption **A₄** and 2.2 that f is uniformly Lipschitz, with a Lipschitz constant that does not depend on f that we denote by L_1. Specifically,

$$(A.1) \qquad |f(x) - f(y)| \leq L_1|x - y|$$

with $L_1 = N_K R$ Using (A.1), we will prove the following lemma:

LEMMA A.1. *Assuming **A₄**, consider the set of solutions to the problem*

$$(A.2) \qquad \frac{\partial x}{\partial t} = \dot{x} = f(x), x(t_0) = x_0$$

where f belongs to the RKHS with kernel K, $|x_0| \leq r$ and $t \in [0, T]$. Then any solution x in this set of solutions is bounded by a uniform constant B_1 that only depends on T, R, L_1 and $L_3^2 := \sup_{||x|| < C} |K(x, x)|$.

Specifically,

$$(A.3) \qquad |x(t) - x(t_0)| \leq B_1 = TL_3Re^{L_1T}$$

Proof. We start by taking f in our class of functions and x_0 such that $|x_0| \leq r$. We therefore can write:

$$(A.4) \qquad |x(t) - x_0| = |\int_0^t (f(x(s)) - f(x_0))ds + tf(x_0)|$$

$$(A.5) \qquad \leq \int_0^t |f(x(s)) - f(x_0)|ds + t||f||_H \sqrt{K(x_0, x_0)}$$

$$(A.6) \qquad \leq L_1 \int_0^t |x(s) - x_0|ds + TL_3R$$

Now denote by $G(t) := |x(t) - x_0|$. If we prove that $G(t)$ is bounded by a constant depending only on T, R, L_1 and L_3, we will be done. So far we have:

$$(A.7) \qquad G(t) \leq L_1 \int_0^t G(s)ds + TL_3R$$

Denote by $V(t) := \int_0^t G(s)ds$. We have that:

$$(A.8) \qquad V'(t) \leq L_1V(t) + TL_3R$$

which implies:

$$(A.9) \qquad e^{-L_1t}V'(t) - L_1e^{-L_1t}V(t) \leq TL_3Re^{-L_1t}$$

Integrating the inequality between 0 and t using the fact that $V(0) = G(0) = 0$, we obtain:

$$(A.10) \qquad \exp(-L_1t)V(t) \leq \frac{TL_3R}{L_1}(1 - e^{-L_1t})$$

or, equivalently,

$$(A.11) \qquad V(t) \leq \frac{TL_3R}{L_1}(e^{L_1t} - 1)$$

Finally since $V'(t) = G(t) \leq L_1V(t) + TL_3R$, we have:

$$(A.12) \qquad G(t) \leq TL_3Re^{L_1t} \leq TL_3Re^{L_1T} \qquad \qquad \square$$

Let us now introduce the following notations:
- We denote by $x(x_0, f, t)$ the solution to the ODE with derivative f and initial condition x_0

- y_i is the observed noisy point from the trajectory at time t_i.
- $x^*(t)$ is the true trajectory evaluated at time t

We now proceed with the following reasoning. We assume that our trajectory minimizes

$$(A.13) \qquad \hat{L}(f, x_0) := \sum_{i=1}^{m} (t_{i+1} - t_i) \left((x(x_0, f, t_i) - y_i)^2 - \sigma^2 \right)$$

over (f, x_0) such that $||f||_H \leq R$, and $|x_0| \leq r$. We denote the minimizer by $(\hat{f}, \hat{x}_0)$.

When x_0 and f are fixed and not data dependent (deterministic), the expected value of $\hat{L}(f, x_0)$ is :

$$(A.14) \qquad L(f, x_0) := \sum_{i=1}^{m} (t_{i+1} - t_i)(x(x_0, f, t_i) - x^*(t_i))^2$$

Notice that $\mathbf{A_1}$ implies:

$$(A.15) \quad min_{||f||_H \leq R, |x_0| \leq r} L(f, x_0) = L(f^*, x_0^*) = \sum_{i=1}^{m} (t_{i+1} - t_i)(x^*(t_i) - x^*(t_i))^2 = 0$$

Our goal is to evaluate $L(\hat{f}, \hat{x}_0)$ and obtain a generalization bound. We have:

$$(A.16) \quad L(\hat{f}, \hat{x}_0) = L(\hat{f}, \hat{x}_0) - \hat{L}(\hat{f}, \hat{x}_0) + \hat{L}(\hat{f}, \hat{x}_0) - \hat{L}(f^*, x_0^*) + \hat{L}(f^*, x_0^*) - L(f^*, x_0^*)$$

And therefore, since the middle term in (A.16): $\hat{L}(\hat{f}, \hat{x}_0) - \hat{L}(f^*, x_0^*) < 0$,

$$(A.17) \qquad L(\hat{f}, \hat{x}_0) \leq \sup_{||f||_H \leq R, |x_0| \leq r} 2|L(f, x_0) - \hat{L}(f, x_0)|$$

We thus consider the following quantity :

$$(A.18) \qquad \text{Err} := \sup_{||f||_H \leq R, |x_0| \leq r} |\hat{L}(f, x_0) - L(f, x_0)|$$

Expanding this quantity we get:

$$(A.19) \quad \sup_{||f||_H \leq R, |x_0| \leq r} |\sum_{i=1}^{m} (t_{i+1} - t_i)(y_i^2 - x^*(t_i)^2 - \sigma^2 - 2x(x_0, f, t_i)(y_i - x^*(t_i))|$$

Notice that if we replace for a given single i, $y_i = x^*(t_i) + \epsilon_i$ by $\tilde{y}_i = x^*(t_i) + \tilde{\epsilon}_i$, the quantity of equation A.19 will change by a quantity bounded by some constant $K_2(t_{i+1} - t_i)$, that we can bound by $4(B_1 + r + M_\epsilon)M_\epsilon + 4(B_1 + r)M_\epsilon$. Therefore, using McDiarmid inequality [9]:

$$(A.20) \qquad \mathbb{P}\left(\text{Err} \geq \mathbb{E}(\text{Err}) + \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{K_2^2 \sum_{i=1}^{m} (t_{i+1} - t_i)^2}\right)$$

We therefore need to provide an upper bound of $\mathbb{E}(\text{Err})$. For that, we are going to view:

(A.21)
$$|\hat{L}(f, x_0) - L(f, x_0)| = |\sum_{i=1}^{m}(t_{i+1} - t_i)(y_i^2 - x^*(t_i)^2 - \sigma^2 - 2x(x_0, f, t_i)(y_i - x^*(t_i)))|$$

as a stochastic process indexed by x, where $x \in \mathcal{X}$: Set of all solutions $x(f, x_0, .)$ for all $||f||_H \leq R$ and $|x_0| \leq r$. In other words, we view the process $|\hat{L}(f, x_0) - L(f, x_0)|$ indexed by f and x_0 as:

(A.22)
$$|\hat{L}(x) - L(x)|$$

where $x \in \mathcal{X}$ is some $x(f, x_0, .)$. Notice that Err is also:

(A.23)
$$\sup_{x \in \mathcal{X}} |\hat{L}(x) - L(x)|$$

Notice that x is a subset of continuous functions defined on $[0, T]$. Therefore we can equip $\mathcal{X}$ with the metric structure $(\mathcal{X}, ||.||_\infty)$. We will apply Dudley's inequality (see for e.g [27], theorem 8.1.3) to bound:

(A.24)
$$\mathbb{E}(\text{Err}) = \mathbb{E}\left(\sup_{||f||_H \leq R, |x_0| \leq r} |\hat{L}(f, x_0) - L(f, x_0)|\right)$$

To apply Dudley's inequality, we are going to use the following lemma.

LEMMA A.2. *The solutions $x \in \mathcal{X}$ are Lipschitz with a Lipschitz constant that is uniform over $\mathcal{X}$, i.e, there exists a constant L_6 such that for every $x \in \mathcal{X}$, $t \in [0, T]$ and $s \in [0, T]$:*

(A.25)
$$|x(t) - x(s)| \leq L_6|t - s|$$

K_6 *depends on R, B_1, r and the kernel K.*

Proof. Let x_0 such that $|x_0| \leq r$ and f such that $||f||_H \leq R$. We have:

(A.26)
$$|\dot{x}(x_0, f, t)| = |f(x(t))|$$

(A.27)
$$\leq R\sqrt{\sup_{|x| \leq B_1 + r} K(x, x)} \qquad \square$$

As a consequence, if we denote by $\mathcal{N}(\mathcal{X}, \epsilon)$ the covering number of $\mathcal{X}$ with a radius ϵ we have the existence of a constant L_7 (L_7 only depends on B_1, r and L_6) such that:

(A.28)
$$\mathcal{N}(\mathcal{X}, \epsilon) \leq \exp\left(\frac{L_7}{\epsilon}\right),$$

where we used a known upper bound that can be found for example in [27] (exercise 8.2.7) on the covering number of uniformly bounded Lipschitz continuous functions defined on a finite interval.

Using this result combined with Dudley's inequality, we obtain the existence of a constant L_8 (depending only on L_7) such that:

PROPOSITION A.3.

$$\text{(A.29)} \qquad \mathbb{E}(\text{Err}) \leq L_8 \sqrt{\sum_{i=1}^{m} (t_{i+1} - t_i)^2}$$

Proof. Apply Dudley's inequality to Err using inequality (A.28) and the fact that the diameter of $\mathcal{X}$ is finite bounded by $2(B_1 + r)$ and that for every $M < \infty$

$$\text{(A.30)} \qquad \int_0^M \sqrt{\log(\mathcal{N}(\mathcal{X}, \epsilon))} d\epsilon \leq \int_0^M \sqrt{\log(\exp{(\frac{K_7}{\epsilon})})} d\epsilon < \infty \qquad \square$$

As a consequence, using (A.20) and theorem (A.3), we obtain the following inequality:

$$\text{(A.31)} \qquad \mathbb{P}\left(\text{Err} \geq L_8 \sqrt{\sum_{i=1}^{m} (t_{i+1} - t_i)^2} + \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{K_2^2 \sum_{i=1}^{m} (t_{i+1} - t_i)^2}\right)$$

Using inequalities (A.17) and (A.31) we finally obtain the following theorem:

THEOREM A.4. *With assumptions* $\mathbf{A_1}, \mathbf{A_2}, \mathbf{A_3}$ *and* $\mathbf{A_4}$*, there exist constants* L_9 *and* K_2 *depending only on* R*,* r*,* T*,* M_ϵ *and the kernel* K *such that for every* ϵ*:*

$$\text{(A.32)} \qquad \mathbb{P}\left(L(\hat{f}, \hat{x}_0) \geq L_9 \sqrt{\sum_{i=1}^{m} (t_{i+1} - t_i)^2} + \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{K_2^2 \sum_{i=1}^{m} (t_{i+1} - t_i)^2}\right)$$

A.2. Including the Euler approximation. In reality, the solution (trajectory) that we propose for every f and x_0 is not $x(x_0, f, .)$ the solution of the ODE but $\tilde{x}(x_0, f, h, .)$, the solution obtained with an Euler's method of time step h. The idea is to use the fact that under some sufficient conditions, we know how to bound the error between Euler's method and the true solution. For example, we know that if f is Lipschitz with a Lipschitz constant K_1 and the solution $x(x_0, f, .)$ is $\mathcal{C}^2$ with a constant K_{11} such that:

$$\text{(A.33)} \qquad x''(x_0, f, t) \leq L_{11}, \forall 0 \leq t \leq T$$

then we have the following global truncation error bound [2]:

$$\text{(A.34)} \qquad \max_{1 \leq i \leq n} |x(x_0, f, t_i) - \tilde{x}(x_0, f, h, t_i)| \leq \frac{hL_{11}}{2L_1} \left(\exp^{L_1 T} - 1\right)$$

We already showed that f is Lipschitz with some constant L_1. To ensure the condition of inequality (A.33), notice that:

$$\text{(A.35)} \qquad x''(x_0, f, t) = f(x(x_0, f, t))f'(x(x_0, f, t))$$

Since we already showed that the solutions $x(x_0, f, .)$ are uniformly bounded by $B_1 + r$, it is sufficient to ensure that f is $\mathcal{C}^1$. This is true if we assume that our kernel

K is $\mathcal{C}^2$ and hence (A.34) will be insured.

Taking into account the Euler approximation and the error bound, the steps of the consistency proof are identical only with the following important difference in equation (A.15) from the previous section

$$(A.36) \qquad \min_{||f||_H \leq R, |x_0| \leq r} L(f, x_0) \leq L(f^*, x_0^*)$$

with
(A.37)
$$L(f^*, x_0^*) = \sum_{i=1}^{m} (t_{i+1} - t_i)(\tilde{x}^*(t_i, h) - x^*(t_i))^2 \leq \frac{h^2 L_{11}{}^2 T}{4L_1^2} \left(\exp^{L_1 T} - 1 \right)^2 := L_{12}$$

With this modification, theorem A.4 becomes:

THEOREM A.5. *Assuming* $\mathbf{A_1, A_2}$, $\mathbf{A_3}$ *and* $\mathbf{A_4}$, *there exist constants* K_2, L_{12} *and* L_{13} *depending only on* R, r, T, M_ϵ *and the kernel* K *such that for every* ϵ:
(A.38)
$$\mathbb{P}\left(L(\hat{f}, \hat{x}_0) \geq L_{13} \sqrt{ \sum_{i=1}^{m} (t_{i+1} - t_i)^2 + h^2 L_{12} } + \epsilon \right) \leq \exp\left(\frac{-2\epsilon^2}{K_2^2 \sum_{i=1}^{m} (t_{i+1} - t_i)^2} \right)$$

A.3. L^2 **squared distance between the true solution and the estimated trajectory.** In reality $L(\hat{f}, \hat{x}_0)$ is an approximation of the L^2 norm squared

$$(A.39) \qquad ||x(\hat{f}, \hat{x}_0, .) - x^*(.)||_{L_2}^2 := \int_0^T \left(x(\hat{f}, \hat{x}_0, t) - x^*(t) \right)^2 dt$$

Since we proved that the solutions are uniformly bounded by $(B_1 + r)$ and $\dot{x}$ is bounded by L_6, we have $t \to \left(x(\hat{f}, \hat{x}_0, t) - x^*(t) \right)^2$ is Lipschitz with Lipschitz constant $8(B_1 + r)L_6$ (we just bound the norm of the derivative). Therefore:

$$(A.40) \qquad |||x(\hat{f}, \hat{x}_0, .) - x^*(.)||_{L_2}^2 - L(\hat{f}, \hat{x}_0)| \leq 8(B_1 + r)L_6 \sum_{i=1}^{m} (t_{i+1} - t_i)^2$$

which proves theorem 2 of the main document.

Appendix B. Kernels.

B.1. Examples of kernels which satisfy the assumptions of Corollary 1.
1. The linear kernel

$$(B.1) \qquad K_1(x, y) = (x^T y + c_0)$$

2. The Gaussian kernel:

$$(B.2) \qquad K_1(x, y) = \exp\left(-\frac{1}{\sigma^2}(|x - y|^2) \right)$$

3. The rational quadratic kernel:

$$(B.3) \qquad K_1(x, y) = 1 - \frac{|x - y|^2}{|x - y|^2 + \theta}, \theta \geq 0$$

4. The Matérn kernel with $\nu \geq 3/2$

B.2. Examples of kernels which do not satisfy the assumptions of Corollary 1:.

1. The polynomial kernels with order larger or equal to 2
2. The Laplacian or exponential kernel

$$
(B.4) \qquad K_1(x,y) = \exp\left(-\frac{|x-y|}{\theta}\right), \theta > 0
$$

3. The Matérn kernel with $\nu \leq 1/2$

REFERENCES

[1] M. A. ALVAREZ, L. ROSASCO, AND N. D. LAWRENCE, *Kernels for vector-valued functions: A review*, arXiv preprint arXiv:1106.6251, (2011).

[2] K. E. ATKINSON, *An introduction to numerical analysis*, John wiley & sons, 2008, ch. 6.

[3] S. L. BRUNTON, J. L. PROCTOR, AND J. N. KUTZ, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proceedings of the national academy of sciences, 113 (2016), pp. 3932–3937.

[4] R. T. CHEN, Y. RUBANOVA, J. BETTENCOURT, AND D. DUVENAUD, *Neural ordinary differential equations*, arXiv preprint arXiv:1806.07366, (2018).

[5] T. DAO, C. DE SA, AND C. RÉ, *Gaussian quadrature for kernel features*, Advances in neural information processing systems, 30 (2017), p. 6109.

[6] B. DE SILVA, K. CHAMPION, M. QUADE, J.-C. LOISEAU, J. KUTZ, AND S. BRUNTON, *Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data*, Journal of Open Source Software, 5 (2020), p. 2104, https://doi.org/10.21105/joss.02104, https://doi.org/10.21105/joss.02104.

[7] A. M. DEGENNARO AND N. M. URBAN, *Scalable extended dynamic mode decomposition using random kernel approximation*, SIAM Journal on Scientific Computing, 41 (2019), pp. A1482–A1499.

[8] F. DONDELINGER, D. HUSMEIER, S. ROGERS, AND M. FILIPPONE, *Ode parameter inference using adaptive gradient matching with gaussian processes*, in Artificial intelligence and statistics, PMLR, 2013, pp. 216–228.

[9] J. L. DOOB, *Regularity properties of certain families of chance variables*, Transactions of the American Mathematical Society, 47 (1940), pp. 455–486.

[10] Y. N. HARARI, *Why technology favors tyranny*, The Atlantic, 322 (2018).

[11] M. HEINONEN, C. YILDIZ, H. MANNERSTRÖM, J. INTOSALMI, AND H. LÄHDESMÄKI, *Learning unknown ode models with gaussian processes*, in International Conference on Machine Learning, PMLR, 2018, pp. 1959–1968.

[12] M. W. HIRSCH, S. SMALE, AND R. L. DEVANEY, *Differential equations, dynamical systems, and an introduction to chaos*, Academic press, 2012.

[13] T. HOFMANN, B. SCHÖLKOPF, AND A. J. SMOLA, *Kernel methods in machine learning*, The annals of statistics, (2008), pp. 1171–1220.

[14] P. HU, W. YANG, Y. ZHU, AND L. HONG, *Revealing hidden dynamics from time-series data by odenet*, arXiv preprint arXiv:2005.04849, (2020).

[15] M. KANAGAWA, P. HENNIG, D. SEJDINOVIC, AND B. K. SRIPERUMBUDUR, *Gaussian processes and kernel methods: A review on connections and equivalences*, arXiv preprint arXiv:1807.02582, (2018).

[16] E. KILIÇ AND P. STANICA, *The inverse of banded matrices*, Journal of Computational and Applied Mathematics, 237 (2013), pp. 126–135.

[17] P. KOKOTOVIC AND J. HELLER, *Direct and adjoint sensitivity equations for parameter optimization*, IEEE Transactions on Automatic Control, 12 (1967), pp. 609–610.

[18] B. O. KOOPMAN, *Hamiltonian systems and transformation in hilbert space*, Proceedings of the national academy of sciences of the united states of america, 17 (1931), p. 315.

[19] M. P. MURPHY AND H. LEVINE III, *Alzheimer's disease and the amyloid-β peptide*, Journal of Alzheimer's disease, 19 (2010), pp. 311–323.

[20] M. NIU, S. ROGERS, M. FILIPPONE, AND D. HUSMEIER, *Fast parameter inference in nonlinear dynamical systems using iterative gradient matching*, in International Conference on Machine Learning, PMLR, 2016, pp. 1699–1707.

[21] L. S. PONTRYAGIN, *Mathematical theory of optimal processes*, CRC press, 1987.

[22] T. QIN, K. WU, AND D. XIU, *Data driven governing equations approximation using deep neural networks*, Journal of Computational Physics, 395 (2019), pp. 620–635.

[23] J. QUINONERO-CANDELA AND C. E. RASMUSSEN, *A unifying view of sparse approximate gaussian process regression*, The Journal of Machine Learning Research, 6 (2005), pp. 1939–1959.

[24] P. J. SCHMID, *Dynamic mode decomposition of numerical and experimental data*, Journal of fluid mechanics, 656 (2010), pp. 5–28.

[25] G. F. SIMMONS, *Differential equations with applications and historical notes*, CRC Press, 2016, ch. 13. Theorem B.

[26] P. VERNHET, M. BILGEL, S. DURRLEMAN, S. M. RESNICK, S. C. JOHNSON, AND B. M. JEDYNAK, *Modeling the early accumulation of amyloid using differential equations in wrap and blsa: Neuroimaging/optimal neuroimaging measures for early detection*, Alzheimer's & Dementia, 16 (2020), p. e039536.

[27] R. VERSHYNIN, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018, ch. 8.

[28] M. O. WILLIAMS, I. G. KEVREKIDIS, AND C. W. ROWLEY, *A data–driven approximation of the koopman operator: Extending dynamic mode decomposition*, Journal of Nonlinear Science, 25 (2015), pp. 1307–1346.

[29] E. YEUNG, S. KUNDU, AND N. HODAS, *Learning deep neural network representations for koopman operators of nonlinear dynamical systems*, in 2019 American Control Conference (ACC), IEEE, 2019, pp. 4832–4839.

[30] L. YOUNES, *Diffeomorphic learning*, Journal of Machine Learning Research, 21 (2020), pp. 1–28, http://jmlr.org/papers/v21/18-415.html.

[31] P. ZHENG, T. ASKHAM, S. L. BRUNTON, J. N. KUTZ, AND A. Y. ARAVKIN, *A unified framework for sparse relaxed regularized regression: Sr3*, IEEE Access, 7 (2018), pp. 1404–1423.